TiDB

TiDB 运维指南（企业版）

产品版本：v5.4 | 文档版本：20220331



PingCAP

北京平凯星辰科技发展有限公司

法律声明

PingCAP 提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过 PingCAP 网站或 PingCAP 提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为 PingCAP 的保密信息，您应当严格遵守保密义务；未经 PingCAP 事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经 PingCAP 事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。PingCAP 保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在 PingCAP 授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过 PingCAP 网站或 PingCAP 授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用 PingCAP 产品及服务的参考性指引，PingCAP 以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。PingCAP 在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但 PingCAP 在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证，本文档所引用的性能数据和程序示例仅用于说明目的，实际的性能结果可能因特定配置和操作条件而异。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的， PingCAP 不承担任何法律责任。在任何情况下，PingCAP 均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使 PingCAP 已被告知该等损失的可能性）。

5. 本文档中及PingCAP 网站上所有内容，包括但不限于作品、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计等，均由 PingCAP 和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经 PingCAP 和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表 PingCAP 网站、产品程序或包括本文档在内的内容。此外，未经 PingCAP 事先书面同意，任何人不得以任何目的使用 PingCAP 商标（包括但不限于单独为或以组合形式包含“PingCAP”等 PingCAP 和/或其关联公司品牌，上述品牌的附属标志及图案或任何

© 2022 PingCAP 公司保留所有权利。除非版权法允许，否则在未得到本公司事先给出的书面许可情况下，严禁复制、改编或翻译本文。

类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别 PingCAP 和/或其关联公司）。

6. PingCAP 可能拥有涵盖本文档中描述的主题的专利或者专利申请，未经 PingCAP 事先书面许可，本文档并不授予您关于这些专利或专利申请的任何许可。

7. 本文档中如有关于 PingCAP 未来方向或意图的声明，仅表示目标或者目的， 如有更改或撤销，恕不另行通知。

8. 如若发现本文档存在任何错误，请与 PingCAP 取得直接联系。PingCAP 可能会以它认为合适的任何方式使用或分发您提供的任何信息，而无需对您承担任何义务。

© 2022 PingCAP 公司保留所有权利。除非版权法允许，否则在未得到本公司事先给出的书面许可情况下，严禁复制、改编或翻译本文。

目录

1 关于 TiDB 16

1.1 TiDB 简介 16

1.1.1 五大核心特性 16

1.1.2 四大核心应用场景 16

1.1.3 另请参阅 17

1.2 TiDB 整体架构 17

2 部署方案 19

2.1 TiDB 软件和硬件环境建议配置 19

2.1.1 Linux 操作系统版本要求 19

2.1.2 软件配置要求 19

2.1.3 服务器建议配置 20

2.1.4 网络要求 21

2.1.5 客户端 Web 浏览器要求 22

2.2 TiDB 环境与系统配置检查 22

2.2.1 在 TiKV 部署目标机器上添加数据盘 EXT4 文件系统挂载参数 22

2.2.2 检测及关闭系统 swap 24

2.2.3 检测及关闭目标部署机器的防火墙 24

2.2.4 检测及安装 NTP 服务 25

2.2.5 检查和配置操作系统优化参数 27

2.2.6 手动配置 SSH 互信及 sudo 免密码 32

2.2.7 安装 numactl 工具 32

2.3 规划集群拓扑 . . . 33
2.3.1 最小拓扑架构 . . . 33
2.3.2 TiFlash 部署拓扑 . . . 34
2.3.3 TiCDC 部署拓扑 . . . 35
2.3.4 TiDB Binlog 部署拓扑 . . . 36
2.3.5 TiSpark 部署拓扑 . . . 39
2.3.6 跨数据中心部署拓扑 . . . 42
2.3.7 混合部署拓扑 . . . 45
2.4 安装与启动 . . . 48
2.4.1 使用 TiUP 部署 TiDB 集群 . . . 48
2.4.2 在 Kubernetes 上部署 TiDB . . . 62
2.5 验证集群运行状态 . . . 63
2.5.1 通过 TiUP 检查集群状态 . . . 63
2.5.2 通过 TiDB Dashboard 和 Grafana 检查集群状态 . . . 63
2.5.3 登录数据库执行简单 DML/DDL 操作和查询 SQL 语句 . . . 64
2.6 测试集群性能 . . . 67
2.6.1 如何用 Sysbench 测试 TiDB . . . 67
2.6.2 如何对 TiDB 进行 TPC-C 测试 . . . 71

3 数据迁移 73
3.1 数据迁移场景综述 . . . 73
3.1.1 迁移 Aurora MySQL 到 TiDB . . . 73
3.1.2 迁移 MySQL 到 TiDB . . . 73
3.1.3 分库分表 MySQL 合并迁移到 TiDB . . . 74
3.1.4 从文件迁移数据到 TiDB . . . 74
3.1.5 复杂迁移场景 . . . 74
3.2 数据迁移工具概览 . . . 74
3.2.1 使用 TiUP 快速安装 . . . 82
3.2.2 探索更多 . . . 83

3.3 数据迁移场景 84
3.3.1 从 Amazon Aurora 迁移数据到 TiDB 84
3.3.2 从小数据量 MySQL 迁移数据到 TiDB 91
3.3.3 从大数据量 MySQL 迁移数据到 TiDB 95
3.3.4 从小数据量分库分表 MySQL 合并迁移数据到 TiDB 101
3.3.5 从大数据量分库分表 MySQL 合并迁移数据到 TiDB 106
3.3.6 从 CSV 文件迁移数据到 TiDB 114
3.3.7 从 SQL 文件迁移数据到 TiDB 118
3.3.8 将 TiDB 集群的增量数据实时同步到另一集群 120
3.4 复杂迁移场景 122
3.4.1 上游使用 pt-osc/gh-ost 工具的持续同步场景 122
3.4.2 下游存在更多列的迁移场景 124
3.4.3 如何过滤 binlog 事件 132
3.4.4 如何通过 SQL 表达式过滤 DML 134

4 升级方案 136
4.0.1 使用 TiUP 升级 TiDB 136

5 集群扩缩容 141
5.1 使用 TiUP 扩容缩容 TiDB 集群 141
5.1.1 扩容 TiDB/PD/TiKV 节点 143
5.1.2 扩容 TiFlash 节点 145
5.1.3 扩容 TiCDC 节点 146
5.1.4 缩容 TiDB/PD/TiKV 节点 147
5.1.5 缩容 TiFlash 节点 149
5.1.6 缩容 TiCDC 节点 152

6 例行维护 153
6.1 日常巡检 153
6.1.1 TiDB Dashboard 关键指标 153
6.2 时区支持 158
6.3 TiFlash 集群运维 160
6.3.1 查看 TiFlash 版本 160
6.3.2 TiFlash 重要日志介绍 160
6.3.3 TiFlash 系统表 161

6.4 TiUP 常见运维操作 · · · · · 161
6.4.1 查看集群列表 · · · · · 162
6.4.2 启动集群 · · · · · 162
6.4.3 查看集群状态 · · · · · 162
6.4.4 修改配置参数 · · · · · 163
6.4.5 Hotfix 版本替换 · · · · · 164
6.4.6 重命名集群 · · · · · 164
6.4.7 关闭集群 · · · · · 165
6.4.8 清除集群数据 · · · · · 165
6.4.9 销毁集群 · · · · · 165
6.5 在线修改集群配置 · · · · · 166
6.5.1 常用操作 · · · · · 166
6.6 Online Unsafe Recovery 使用文档 · · · · · 189
6.6.1 功能说明 · · · · · 189
6.6.2 适用场景 · · · · · 190
6.6.3 使用步骤 · · · · · 190
7 安全加固 191
7.1 为 TiDB 客户端服务端间通信开启加密传输 · · · · · 191
7.1.1 配置 TiDB 服务端启用安全连接 · · · · · 192
7.1.2 配置 MySQL Client 使用安全连接 · · · · · 192
7.1.3 配置启用身份验证 · · · · · 193
7.1.4 检查当前连接是否是加密连接 · · · · · 193
7.1.5 支持的 TLS 版本及密钥交换协议和加密算法 · · · · · 194
7.1.6 重加载证书、密钥和 CA · · · · · 195
7.1.7 监控 · · · · · 195
7.1.8 另请参阅 · · · · · 195
7.2 为 TiDB 组件间通信开启加密传输 · · · · · 195
7.2.1 配置开启加密传输 · · · · · 196
7.2.2 认证组件调用者身份 · · · · · 197
7.2.3 证书重加载 · · · · · 199
7.2.4 另请参阅 · · · · · 199

7.3 为 DM 的连接开启加密传输 ····· 199
7.3.1 为 DM-master，DM-worker，dmctl 组件之间的连接开启加密传输 ····· 199
7.3.2 DM 组件与上下游数据库之间的连接开启加密传输 ····· 200
7.4 生成自签名证书 ····· 201
7.4.1 安装 OpenSSL ····· 202
7.4.2 生成 CA 证书 ····· 202
7.4.3 签发各个组件的证书 ····· 202
7.5 静态加密 ····· 204
7.5.1 各 TiDB 组件支持的加密方式 ····· 204
7.5.2 TiKV 静态加密 ····· 205
7.5.3 BR S3 服务端加密 ····· 208
7.6 为 TiDB 落盘文件开启加密 ····· 209
7.6.1 配置方法 ····· 209
7.7 日志脱敏 ····· 209
7.7.1 TiDB 组件日志脱敏 ····· 209
7.7.2 TiKV 组件日志脱敏 ····· 210
7.7.3 PD 组件日志脱敏 ····· 210
7.7.4 TiFlash 组件日志脱敏 ····· 210

8 备份与恢复 210
8.1 备份与恢复工具 BR 简介 ····· 210
8.1.1 工作原理 ····· 210
8.1.2 部署使用 BR 工具 ····· 212
8.1.3 BR 相关文档 ····· 219
8.2 使用 BR 命令行进行备份恢复 ····· 219
8.2.1 BR 命令行描述 ····· 219
8.2.2 使用 BR 命令行备份集群数据 ····· 221
8.2.3 使用 BR 命令行恢复集群数据 ····· 225
8.3 BR 备份与恢复场景示例 ····· 230
8.3.1 目标读者 ····· 230
8.3.2 环境准备 ····· 231
8.3.3 使用场景 ····· 231
8.3.4 备份过程中的异常处理 ····· 248

8.4 外部存储 · · · · · 248
8.4.1 外部存储 · · · · · 248
8.4.2 在 Azure Blob Storage 备份恢复 · · · · · 253
8.5 BR 特性 · · · · · 257
8.5.1 BR 自动调节从 v5.4 版本开始引入 · · · · · 257
8.6 Backup & Restore 常见问题 · · · · · 259
8.6.1 在 TiDB v5.4.0 及后续版本中，当在有负载的集群进行备份时，备份速度为什么会变得很慢？ · 259
8.6.2 恢复的时候，报错 `could not read local://...:download sst failed`，该如何处理？ · · · · 260
8.6.3 BR 备份时，对集群影响多大？ · · · · · 260
8.6.4 BR 会备份系统表吗？在数据恢复的时候，这些系统表会冲突吗？ · · · · · 260
8.6.5 BR 遇到 Permission denied 或者 No such file or directory 错误，即使用 root 运行 BR 也无法解决，该如何处理？ · · · · · 261
8.6.6 BR 遇到错误信息 `Io(Os...)`，该如何处理？ · · · · · 262
8.6.7 BR 遇到错误信息 `rpc error: code = Unavailable desc =...`，该如何处理？ · · · · · 262
8.6.8 使用 local storage 的时候，BR 备份的文件会存在哪里？ · · · · · 263
8.6.9 备份数据有多大，备份会有副本吗？ · · · · · 263
8.6.10 BR 恢复到 TiCDC / Drainer 的上游集群时，要注意些什么？ · · · · · 263
8.6.11 BR 会备份表的 `SHARD_ROW_ID_BITS` 和 `PRE_SPLIT_REGIONS` 信息吗？恢复出来的表会有多个 Region 吗？ · · · · · 263
8.6.12 使用 BR 恢复备份数据后，SQL 查询报错 `region is unavailable` · · · · · 263
8.6.13 BR 是否支持就地 (in-place) 全量恢复某个历史备份？ · · · · · 263
8.6.14 在 Kubernetes 环境中如何使用 BR 进行增量备份？ · · · · · 263
8.6.15 BR backupTS 如何转化成 Unix 时间？ · · · · · 263
8.6.16 BR 恢复存档后是否需要对表执行 ANALYZE 以更新 TiDB 在表和索引上留下的统计信息？ · · · · · 264
8.6.17 是否可以同时使用多个 BR 进程对单个集群进行恢复？ · · · · · 264
9 监控与告警 264
9.1 TiDB 监控框架概述 · · · · · 264
9.1.1 Prometheus 在 TiDB 中的应用 · · · · · 264
9.1.2 Grafana 在 TiDB 中的应用 · · · · · 265
9.2 TiDB 集群监控 API · · · · · 267
9.2.1 使用状态接口 · · · · · 267
9.2.2 使用 metrics 接口 · · · · · 270

9.3 TiDB 集群监控部署 · · · · · 270
9.3.1 部署 Prometheus 和 Grafana · · · · · 270
9.3.2 配置 Grafana · · · · · 273
9.3.3 查看组件 metrics · · · · · 274
9.4 将 Grafana 监控数据导出成快照 · · · · · 275
9.4.1 使用方法 · · · · · 276
9.4.2 FAQs · · · · · 276
9.5 TiDB 集群报警规则 · · · · · 277
9.5.1 TiDB 报警规则 · · · · · 277
9.5.2 PD 报警规则 · · · · · 281
9.5.3 TiKV 报警规则 · · · · · 285
9.5.4 TiFlash 报警规则 · · · · · 293
9.5.5 TiDB Binlog 报警规则 · · · · · 293
9.5.6 TiCDC 报警规则 · · · · · 293
9.5.7 Node_exporter 主机报警规则 · · · · · 293
9.5.8 Blackbox_exporter TCP、 ICMP 和 HTTP 报警规则 · · · · · 296
9.6 TiFlash 报警规则 · · · · · 300
9.6.1 TiFlash_schema_error · · · · · 300
9.6.2 TiFlash_schema_apply_duration · · · · · 301
9.6.3 TiFlash_raft_read_index_duration · · · · · 301
9.6.4 TiFlash_raft_wait_index_duration · · · · · 301
9.7 处理告警 · · · · · 301
9.7.1 高可用告警 · · · · · 302
9.7.2 任务状态告警 · · · · · 302
9.7.3 relay log 告警 · · · · · 302
9.7.4 Dump/Load 告警 · · · · · 303
9.7.5 Binlog replication 告警 · · · · · 303
9.8 Grafana Dashboard 监控指标 · · · · · 304
9.8.1 Overview 面板重要监控指标详解 · · · · · 304
9.8.2 TiDB 重要监控指标详解 · · · · · 306
9.8.3 PD 重要监控指标详解 · · · · · 310
9.8.4 TiKV 监控指标详解 · · · · · 319

9.8.5 TiFlash 集群监控 · · · · · 339
9.8.6 TiCDC 重要监控指标详解 · · · · · 341
9.8.7 DM 监控指标 · · · · · 346
9.9 TiDB Dashboard 监控指标 · · · · · 355
9.9.1 TiDB Dashboard 介绍 · · · · · 355
9.9.2 运维 · · · · · 357
9.9.3 访问 TiDB Dashboard · · · · · 370
9.9.4 TiDB Dashboard 概况页面 · · · · · 374
9.9.5 TiDB Dashboard 集群信息页面 · · · · · 380
9.9.6 Top SQL · · · · · 383
9.9.7 TiDB Dashboard 流量可视化页面 · · · · · 385
9.9.8 TiDB Dashboard 监控关系图 · · · · · 398
9.9.9 SQL 语句分析 · · · · · 403
9.9.10 TiDB Dashboard 慢查询页面 · · · · · 417
9.9.11 集群诊断页面 · · · · · 421
9.9.12 TiDB Dashboard 日志搜索页面 · · · · · 443
9.9.13 实例性能分析 · · · · · 447
9.9.14 会话管理与配置 · · · · · 457
9.9.15 TiDB Dashboard 常见问题 · · · · · 481
10 性能调优 486
10.1 系统调优 · · · · · 486
10.1.1 操作系统性能参数调优 · · · · · 486
10.2 软件调优 · · · · · 490
10.2.1 配置 · · · · · 490
10.2.2 下推计算结果缓存 · · · · · 518
10.3 SQL 性能调优 · · · · · 520
10.3.1 SQL 性能调优 · · · · · 520
10.3.2 理解 TiDB 执行计划 · · · · · 520
10.3.3 SQL 优化流程 · · · · · 581
10.3.4 控制执行计划 · · · · · 634
11 最佳实践 655

11.1 TiDB 最佳实践 · · · · · 655
11.1.1 前言 · · · · · 655
11.1.2 基本概念 · · · · · 655
11.1.3 场景与实践 · · · · · 658
11.1.4 TiDB 的最佳适用场景 · · · · · 660
11.2 三节点混合部署的最佳实践 · · · · · 660
11.2.1 部署环境和测试方法 · · · · · 660
11.2.2 参数调整 · · · · · 661
11.3 TiDB 高并发写入场景最佳实践 · · · · · 665
11.3.1 目标读者 · · · · · 665
11.3.2 高并发批量插入场景 · · · · · 665
11.3.3 TiDB 数据分布原理 · · · · · 665
11.3.4 热点产生的实例 · · · · · 666
11.3.5 热点问题产生的原因 · · · · · 669
11.3.6 热点问题的规避方法 · · · · · 671
11.3.7 参数配置 · · · · · 674
11.4 在三数据中心下就近读取数据 · · · · · 674
11.4.1 部署三数据中心的 TiDB 集群 · · · · · 675
11.4.2 使用 Stale Read 就近读取数据 · · · · · 675
11.5 使用 Grafana 监控 TiDB 的最佳实践 · · · · · 675
11.5.1 监控架构 · · · · · 675
11.5.2 监控数据的来源与展示 · · · · · 676
11.5.3 Grafana 使用技巧 · · · · · 678
11.5.4 总结 · · · · · 685
11.6 PD 调度策略最佳实践 · · · · · 685
11.6.1 PD 调度原理 · · · · · 686
11.6.2 查询调度状态 · · · · · 688
11.6.3 调度策略控制 · · · · · 689
11.6.4 典型场景分析与处理 · · · · · 690
11.7 海量 Region 集群调优最佳实践 · · · · · 693
11.7.1 Raftstore 的工作流程 · · · · · 693
11.7.2 性能问题 · · · · · 694
11.7.3 性能优化方法 · · · · · 695
11.7.4 其他问题和解决方案 · · · · · 696

11.8 通过拓扑 label 进行副本调度 · · · · · 697
11.8.1 根据集群拓扑配置 labels · · · · · 698
11.8.2 基于拓扑 label 的 PD 调度策略 · · · · · 700
11.9 高可用容灾方案 · · · · · 701
11.9.1 同城多数据中心部署 TiDB · · · · · 701
11.9.2 两地三中心部署 · · · · · 707
11.9.3 同城两中心自适应同步模式部署 · · · · · 713

12 运维工具 719
12.1 TiDB 工具功能概览 · · · · · 719
12.1.1 部署运维工具 · · · · · 719
12.1.2 数据管理工具 · · · · · 720
12.1.3 OLAP 分析工具 · · · · · 722
12.2 TiDB 工具适用场景 · · · · · 722
12.2.1 在物理机或虚拟机上部署运维 TiDB · · · · · 722
12.2.2 在 Kubernetes 上部署运维 TiDB · · · · · 723
12.2.3 从 CSV 导入数据到 TiDB · · · · · 723
12.2.4 从 MySQL/Aurora 导入全量数据 · · · · · 723
12.2.5 从 MySQL/Aurora 迁移数据 · · · · · 723
12.2.6 TiDB 集群备份与恢复 · · · · · 723
12.2.7 迁出数据到 TiDB · · · · · 723
12.2.8 TiDB 增量数据订阅 · · · · · 723
12.3 TiDB 工具下载 · · · · · 723
12.3.1 TiUP · · · · · 724
12.3.2 TiDB Operator · · · · · 724
12.3.3 TiDB Binlog · · · · · 724
12.3.4 TiDB Lightning · · · · · 725
12.3.5 备份和恢复 (BR) 工具 · · · · · 725
12.3.6 TiDB DM (Data Migration) · · · · · 726
12.3.7 Dumpling · · · · · 727
12.3.8 sync-diff-inspector · · · · · 728
12.3.9 TiCDC · · · · · 729

12.4 TiUP · · · · · 729
12.4.1 TiUP 文档地图 · · · · · 729
12.4.2 TiUP 简介 · · · · · 729
12.4.3 TiUP 术语及核心概念 · · · · · 732
12.4.4 使用 TiUP 命令管理组件 · · · · · 733
12.4.5 TiUP FAQ · · · · · 736
12.4.6 TiUP 故障排查 · · · · · 738
12.4.7 TiUP 命令参考手册 · · · · · 739
12.4.8 通过 TiUP 部署 TiDB 集群的拓扑文件配置 · · · · · 808
12.4.9 通过 TiUP 部署 DM 集群的拓扑文件配置 · · · · · 824
12.4.10 TiUP 镜像参考指南 · · · · · 832
12.4.11 TiUP 组件文档 · · · · · 839
12.5 TiDB Operator · · · · · 861
12.6 使用 Dumpling 导出数据 · · · · · 861
12.6.1 从 TiDB/MySQL 导出数据 · · · · · 863
12.6.2 Dumpling 主要选项表 · · · · · 868
12.7 TiDB Lightning · · · · · 882
12.7.1 TiDB Lightning 简介 · · · · · 882
12.7.2 前置需求 · · · · · 884
12.7.3 主要功能 · · · · · 894
12.7.4 TiDB Lightning 教程 · · · · · 933
12.7.5 TiDB Lightning 部署与执行 · · · · · 936
12.7.6 TiDB Lightning 配置参数 · · · · · 939
12.7.7 TiDB Lightning 监控告警 · · · · · 953
12.7.8 TiDB Lightning 常见问题 · · · · · 963
12.7.9 TiDB Lightning 术语表 · · · · · 971
12.8 TiDB Data Migration · · · · · 975
12.8.1 Data Migration 简介 · · · · · 975
12.8.2 TiDB Data Migration 快速上手指南 · · · · · 978
12.8.3 部署 DM 集群 · · · · · 983
12.8.4 入门指南 · · · · · 998
12.8.5 进阶教程 · · · · · 1032
12.8.6 运维管理 · · · · · 1068
12.8.7 参考手册 · · · · · 1113

12.9 Backup & Restore (BR) · · · · · 1187
12.9.1 备份与恢复工具 BR 简介 · · · · · 1187
12.9.2 使用 BR 命令行进行备份恢复 · · · · · 1196
12.9.3 BR 备份与恢复场景示例 · · · · · 1207
12.9.4 BR 特性 · · · · · 1225
12.9.5 Backup & Restore 常见问题 · · · · · 1227
12.10 TiDB Binlog · · · · · 1232
12.10.1 TiDB Binlog 简介 · · · · · 1232
12.10.2 TiDB Binlog 教程 · · · · · 1234
12.10.3 TiDB Binlog 集群部署 · · · · · 1243
12.10.4 TiDB Binlog 集群运维 · · · · · 1251
12.10.5 TiDB Binlog 配置说明 · · · · · 1254
12.10.6 TiDB Binlog 版本升级方法 · · · · · 1269
12.10.7 TiDB Binlog 集群监控 · · · · · 1272
12.10.8 Reparo 使用文档 · · · · · 1275
12.10.9 binlogctl 工具 · · · · · 1278
12.10.10 Binlog Consumer Client 用户文档 · · · · · 1281
12.10.11 TiDB Binlog Relay Log · · · · · 1284
12.10.12 集群间双向同步 · · · · · 1285
12.10.13 TiDB Binlog 术语表 · · · · · 1289
12.10.14 故障诊断 · · · · · 1290
12.10.15 TiDB Binlog 常见问题 · · · · · 1291
12.11 TiCDC · · · · · 1297
12.11.1 TiCDC 简介 · · · · · 1297
12.11.2 TiCDC 安装部署 · · · · · 1305
12.11.3 TiCDC 运维操作及任务管理 · · · · · 1306
12.11.4 TiCDC 常见问题和故障处理 · · · · · 1322
12.11.5 TiCDC 集群监控报警规则 · · · · · 1334
12.11.6 TiCDC OpenAPI · · · · · 1337
12.11.7 TiCDC Open Protocol · · · · · 1349
12.11.8 TiCDC Canal-JSON Protocol · · · · · 1359
12.11.9 TiDB 集成 Confluent Platform 快速上手指南 · · · · · 1369
12.11.10 TiCDC 术语表 · · · · · 1371

12.12 sync-diff-inspector · · · · · 1372
12.12.1 sync-diff-inspector 用户文档 · · · · · 1372
12.12.2 不同库名或表名的数据校验 · · · · · 1379
12.12.3 分库分表场景下的数据校验 · · · · · 1380
12.12.4 TiDB 主从集群的数据校验 · · · · · 1384
12.12.5 基于 DM 同步场景下的数据校验 · · · · · 1385
12.13 TiSpark · · · · · 1386
12.13.1 TiSpark 快速上手 · · · · · 1386
12.13.2 TiSpark 用户指南 · · · · · 1389

13 故障诊断 1398
13.1 集群性能问题 · · · · · 1398
13.1.1 使用 PLAN REPLAYER 保存和恢复集群现场信息 · · · · · 1398
13.1.2 TiDB 集群故障诊断 · · · · · 1400
13.1.3 TiDB 集群问题导图 · · · · · 1402
13.1.4 TiDB 热点问题处理 · · · · · 1416
13.1.5 读写延迟增加 · · · · · 1426
13.1.6 乐观事务模型下写写冲突问题排查 · · · · · 1430
13.1.7 TiDB 磁盘 I/O 过高的处理办法 · · · · · 1433
13.1.8 TiDB 锁冲突问题处理 · · · · · 1435
13.2 SQL 性能问题 · · · · · 1445
13.2.1 慢查询日志 · · · · · 1445
13.2.2 分析慢查询 · · · · · 1457
13.2.3 SQL 诊断 · · · · · 1464
13.2.4 定位消耗系统资源多的查询 · · · · · 1465
13.2.5 Statement Summary Tables · · · · · 1467
13.3 TiFlash 常见问题 · · · · · 1474
13.3.1 TiFlash 未能正常启动 · · · · · 1474
13.3.2 TiFlash 副本始终处于不可用状态 · · · · · 1475
13.3.3 TiFlash 查询时间不稳定，同时错误日志中打印出大量的 Lock Exception · · · · · 1476
13.3.4 部分查询返回 Region Unavailable 的错误 · · · · · 1476
13.3.5 数据文件损坏 · · · · · 1476
13.3.6 TiFlash 分析慢 · · · · · 1476

13.3.7 TiFlash 数据不同步 ····· 1477
13.3.8 TiFlash 数据同步卡住 ····· 1479
13.3.9 数据同步慢 ····· 1479
13.4 错误码与故障诊断 ····· 1480
13.4.1 错误码 ····· 1480
13.4.2 故障诊断 ····· 1486
14 常见问题解答 (FAQ) 1486
14.1 安装部署 FAQ ····· 1486
14.1.1 环境准备 FAQ ····· 1486
14.1.2 硬件要求 FAQ ····· 1486
14.1.3 安装部署 FAQ ····· 1487
14.2 迁移常见问题 ····· 1493
14.2.1 全量数据导出导入 ····· 1493
14.2.2 在线数据同步 ····· 1495
14.2.3 业务流量迁入 ····· 1496
14.3 升级与升级后常见问题 ····· 1497
14.3.1 升级常见问题 ····· 1497
14.3.2 升级后常见问题 ····· 1498
14.4 监控 FAQ ····· 1502
14.4.1 目前的监控使用方式及主要监控指标，有没有更好看的监控？ ····· 1503
14.4.2 Prometheus 监控数据默认 15 天自动清除一次，可以自己设定成 2 个月或者手动删除吗？ ····· 1503
14.4.3 Region Health 监控项 ····· 1503
14.4.4 Statement Count 监控项中的 selectsimplefull 是什么意思？ ····· 1503
14.4.5 监控上的 QPS 和 Statement OPS 有什么区别？ ····· 1503
14.5 集群管理 FAQ ····· 1503
14.5.1 集群日常管理 ····· 1503
14.5.2 PD 管理 ····· 1506
14.5.3 TiDB server 管理 ····· 1507
14.5.4 TiKV 管理 ····· 1508
14.5.5 TiDB 测试 ····· 1511
14.5.6 TiDB 备份恢复 ····· 1512
15 术语表 1512

15.1 A · · · · · 1512
15.1.1 ACID · · · · · 1512
15.2 C · · · · · 1513
15.2.1 Continuous Profiling · · · · · 1513
15.3 I · · · · · 1513
15.3.1 Index Merge · · · · · 1513
15.4 L · · · · · 1513
15.4.1 Leader/Follower/Learner · · · · · 1513
15.5 O · · · · · 1513
15.5.1 Old value · · · · · 1513
15.5.2 Operator · · · · · 1513
15.5.3 Operator Step · · · · · 1513
15.6 P · · · · · 1514
15.6.1 Pending/Down · · · · · 1514
15.6.2 Predicate columns · · · · · 1514
15.7 R · · · · · 1514
15.8 Raft Engine · · · · · 1514
15.8.1 Region/Peer/Raft Group · · · · · 1514
15.8.2 Region Split · · · · · 1514
15.8.3 Restore · · · · · 1514
15.9 S · · · · · 1515
15.9.1 Scheduler · · · · · 1515
15.9.2 Store · · · · · 1515
15.10 T · · · · · 1515
15.10.1 Top SQL · · · · · 1515
15.10.2 TSO · · · · · 1515

1 关于 TiDB

1.1 TiDB 简介

TiDB 是 PingCAP 公司自主设计、研发的开源分布式关系型数据库，是一款同时支持在线事务处理与在线分析处理 (Hybrid Transactional and Analytical Processing, HTAP) 的融合型分布式数据库产品，具备水平扩容或者缩容、金融级高可用、实时 HTAP、云原生的分布式数据库、兼容 MySQL 5.7 协议和 MySQL 生态等重要特性。目标是为用户提供一站式 OLTP (Online Transactional Processing)、OLAP (Online Analytical Processing)、HTAP 解决方案。TiDB 适合高可用、强一致要求较高、数据规模较大等各种应用场景。

关于 TiDB 的关键技术创新，请观看以下视频。

1.1.1 五大核心特性

- 一键水平扩容或者缩容

  得益于 TiDB 存储计算分离的架构的设计，可按需对计算、存储分别进行在线扩容或者缩容，扩容或者缩容过程中对应用运维人员透明。

- 金融级高可用

  数据采用多副本存储，数据副本通过 Multi-Raft 协议同步事务日志，多数派写入成功事务才能提交，确保数据强一致性且少数副本发生故障时不影响数据的可用性。可按需配置副本地理位置、副本数量等策略满足不同容灾级别的要求。

- 实时 HTAP

  提供行存储引擎 TiKV、列存储引擎 TiFlash 两款存储引擎，TiFlash 通过 Multi-Raft Learner 协议实时从 TiKV 复制数据，确保行存储引擎 TiKV 和列存储引擎 TiFlash 之间的数据强一致。TiKV、TiFlash 可按需部署在不同的机器，解决 HTAP 资源隔离的问题。

- 云原生的分布式数据库

  专为云而设计的分布式数据库，通过 TiDB Operator 可在公有云、私有云、混合云中实现部署工具化、自动化。

- 兼容 MySQL 5.7 协议和 MySQL 生态

  兼容 MySQL 5.7 协议、MySQL 常用的功能、MySQL 生态，应用无需或者修改少量代码即可从 MySQL 迁移到 TiDB。提供丰富的数据迁移工具帮助应用便捷完成数据迁移。

1.1.2 四大核心应用场景

- 对数据一致性及高可靠、系统高可用、可扩展性、容灾要求较高的金融行业属性的场景

  众所周知，金融行业对数据一致性及高可靠、系统高可用、可扩展性、容灾要求较高。传统的解决方案是同城两个机房提供服务、异地一个机房提供数据容灾能力但不提供服务，此解决方案存在以下缺点：资源利用率低、维护成本高、RTO (Recovery Time Objective) 及 RPO (Recovery Point Objective) 无法真实达到企业所期望的值。TiDB 采用多副本 + Multi-Raft 协议的方式将数据调度到不同的机房、机架、机器，当部分机器出现故障时系统可自动进行切换，确保系统的 RTO <= 30s 及 RPO = 0。

- 对存储容量、可扩展性、并发要求较高的海量数据及高并发的 OLTP 场景

  随着业务的高速发展，数据呈现爆炸性的增长，传统的单机数据库无法满足因数据爆炸性的增长对数据库的容量要求，可行方案是采用分库分表的中间件产品或者 NewSQL 数据库替代、采用高端的存储设备等，其中性价比最大的是 NewSQL 数据库，例如：TiDB。TiDB 采用计算、存储分离的架构，可对计算、存储分别进行扩容和缩容，计算最大支持 512 节点，每个节点最大支持 1000 并发，集群容量最大支持 PB 级别。

- Real-time HTAP 场景

  随着 5G、物联网、人工智能的高速发展，企业所生产的数据会越来越多，其规模可能达到数百 TB 甚至 PB 级别，传统的解决方案是通过 OLTP 型数据库处理在线联机交易业务，通过 ETL 工具将数据同步到 OLAP 型数据库进行数据分析，这种处理方案存在存储成本高、实时性差等多方面的问题。TiDB 在 4.0 版本中引入列存储引擎 TiFlash 结合行存储引擎 TiKV 构建真正的 HTAP 数据库，在增加少量存储成本的情况下，可以同一个系统中做联机交易处理、实时数据分析，极大地节省企业的成本。

- 数据汇聚、二次加工处理的场景

  当前绝大部分企业的业务数据都分散在不同的系统中，没有一个统一的汇总，随着业务的发展，企业的决策层需要了解整个公司的业务状况以便及时做出决策，故需要将分散在各个系统的数据汇聚在同一个系统并进行二次加工处理生成 T+0 或 T+1 的报表。传统常见的解决方案是采用 ETL + Hadoop 来完成，但 Hadoop 体系太复杂，运维、存储成本太高无法满足用户的需求。与 Hadoop 相比，TiDB 就简单得多，业务通过 ETL 工具或者 TiDB 的同步工具将数据同步到 TiDB，在 TiDB 中可通过 SQL 直接生成报表。

关于 TiDB 典型应用场景和用户案例的介绍，请观看以下视频。

1.1.3 另请参阅

- TiDB 整体架构
- TiDB 数据库的存储
- TiDB 数据库的计算
- TiDB 数据库的调度

1.2 TiDB 整体架构

推荐先观看以下视频（时长约 14 分钟），快速了解 TiDB 的整体架构。

与传统的单机数据库相比，TiDB 具有以下优势：

- 纯分布式架构，拥有良好的扩展性，支持弹性的扩缩容
- 支持 SQL，对外暴露 MySQL 的网络协议，并兼容大多数 MySQL 的语法，在大多数场景下可以直接替换 MySQL
- 默认支持高可用，在少数副本失效的情况下，数据库本身能够自动进行数据修复和故障转移，对业务透明
- 支持 ACID 事务，对于一些有强一致需求的场景友好，例如：银行转账
- 具有丰富的工具链生态，覆盖数据迁移、同步、备份等多种场景

在内核设计上，TiDB 分布式数据库将整体架构拆分成了多个模块，各模块之间互相通信，组成完整的 TiDB 系统。对应的架构图如下：

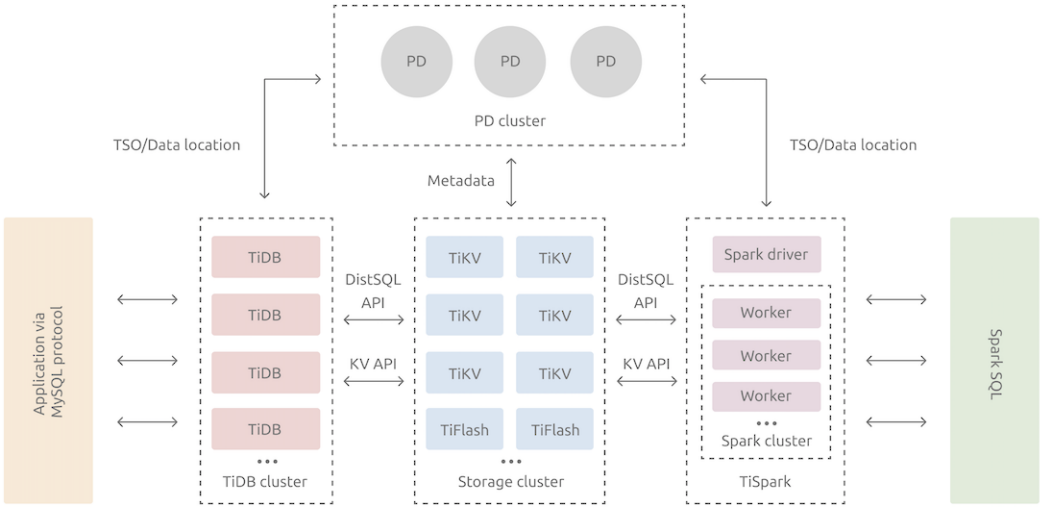


图 1: architecture

- TiDB Server：SQL 层，对外暴露 MySQL 协议的连接 endpoint，负责接受客户端的连接，执行 SQL 解析和优化，最终生成分布式执行计划。TiDB 层本身是无状态的，实践中可以启动多个 TiDB 实例，通过负载均衡组件（如 LVS、HAProxy 或 F5）对外提供统一的接入地址，客户端的连接可以均匀地分摊在多个 TiDB 实例上以达到负载均衡的效果。TiDB Server 本身并不存储数据，只是解析 SQL，将实际的数据读取请求转发给底层的存储节点 TiKV（或 TiFlash）。
- PD (Placement Driver) Server：整个 TiDB 集群的元信息管理模块，负责存储每个 TiKV 节点实时的数据分布情况和集群的整体拓扑结构，提供 TiDB Dashboard 管控界面，并为分布式事务分配事务 ID。PD 不仅存储元信息，同时还会根据 TiKV 节点实时上报的数据分布状态，下发数据调度命令给具体的 TiKV 节点，可以说是整个集群的“大脑”。此外，PD 本身也是由至少 3 个节点构成，拥有高可用的能力。建议部署奇数个 PD 节点。
- 存储节点
  - TiKV Server：负责存储数据，从外部看 TiKV 是一个分布式的提供事务的 Key-Value 存储引擎。存储数据的基本单位是 Region，每个 Region 负责存储一个 Key Range（从 StartKey 到 EndKey 的左闭右开区间）的数据，每个 TiKV 节点会负责多个 Region。TiKV 的 API 在 KV 键值对层面提供对分布式事务的原生支持，默认提供了 SI (Snapshot Isolation) 的隔离级别，这也是 TiDB 在 SQL 层面支持分布式事务的核心。TiDB 的 SQL 层做完 SQL 解析后，会将 SQL 的执行计划转换为对 TiKV API 的实际调用。所以，数据都存储在 TiKV 中。另外，TiKV 中的数据都会自动维护多副本（默认为三副本），天然支持高可用和自动故障转移。
  - TiFlash：TiFlash 是一类特殊的存储节点。和普通 TiKV 节点不一样的是，在 TiFlash 内部，数据是以列式的形式进行存储，主要的功能是为分析型的场景加速。

2 部署方案

2.1 TiDB 软件和硬件环境建议配置

TiDB 作为一款开源分布式 NewSQL 数据库，可以很好地部署和运行在 Intel 架构服务器环境、ARM 架构的服务器环境及主流虚拟化环境，并支持绝大多数的主流硬件网络。作为一款高性能数据库系统，TiDB 支持主流的 Linux 操作系统环境。

2.1.1 Linux 操作系统版本要求

Linux 操作系统平台	版本
Red Hat Enterprise Linux	7.3 及以上的 7.x 版本
CentOS	7.3 及以上的 7.x 版本
Oracle Enterprise Linux	7.3 及以上的 7.x 版本
Ubuntu LTS	16.04 及以上的版本

注意：

- 目前尚不支持 Red Hat Enterprise Linux 8.0、 CentOS 8 Stream 和 Oracle Enterprise Linux 8.0，因为目前对这些平台的测试还在进行中。
- 不计划支持 CentOS 8 Linux，因为 CentOS 的上游支持将于 2021 年 12 月 31 日终止。
- TiDB 将不再支持 Ubuntu 16.04。强烈建议升级到 Ubuntu 18.04 或更高版本。

其他 Linux 操作系统版本（例如 Debian Linux 和 Fedora Linux）也许可以运行 TiDB，但尚未得到 TiDB 官方支持。

2.1.2 软件配置要求

2.1.2.1 中控机软件配置

软件	版本
sshpass	1.06 及以上
TiUP	1.5.0 及以上

注意：

中控机需要部署TiUP 软件来完成 TiDB 集群运维管理。

2.1.2.2 目标主机建议配置软件

软件	版本
sshpass	1.06 及以上
numa	2.0.12 及以上
tar	任意

2.1.3 服务器建议配置

TiDB 支持部署和运行在 Intel x86-64 架构的 64 位通用硬件服务器平台或者 ARM 架构的硬件服务器平台。对于开发、测试及生产环境的服务器硬件配置（不包含操作系统 OS 本身的占用）有以下要求和建议：

2.1.3.1 开发及测试环境

组件	CPU	内存	本地存储	网络	实例数量 (最低要求)
TiDB	8 核 +	16 GB+	无特殊要求	千兆网卡	1（可与 PD 同机器）
PD	4 核 +	8 GB+	SAS, 200 GB+	千兆网卡	1（可与 TiDB 同机器）
TiKV	8 核 +	32 GB+	SSD, 200 GB+	千兆网卡	3
TiFlash	32 核 +	64 GB+	SSD, 200 GB+	千兆网卡	1
TiCDC	8 核 +	16 GB+	SAS, 200 GB+	千兆网卡	1

注意：

- 验证测试环境中的 TiDB 和 PD 可以部署在同一台服务器上。
- 如进行性能相关的测试，避免采用低性能存储和网络硬件配置，防止对测试结果的正确性产生干扰。
- TiKV 的 SSD 盘推荐使用 NVME 接口以保证读写更快。
- 如果仅验证功能，建议使用 TiDB 数据库快速上手指南进行单机功能测试。
- TiDB 对于磁盘的使用以存放日志为主，因此在测试环境中对于磁盘类型和容量并无特殊要求。

2.1.3.2 生产环境

组件	CPU	内存	硬盘类型	网络	实例数量 (最低要求)
TiDB	16 核 +	32 GB+	SAS	万兆网卡（2 块最佳）	2
PD	4 核 +	8 GB+	SSD	万兆网卡（2 块最佳）	3
TiKV	16 核 +	32 GB+	SSD	万兆网卡（2 块最佳）	3
TiFlash	48 核 +	128 GB+	1 or more SSDs	万兆网卡（2 块最佳）	2
TiCDC	16 核 +	64 GB+	SSD	万兆网卡（2 块最佳）	2

组件	CPU	内存	硬盘类型	网络	实例数量 (最低要求)
监控	8 核 +	16 GB+	SAS	千兆网卡	1

注意：

- 生产环境中的 TiDB 和 PD 可以部署和运行在同一台服务器上，如对性能和可靠性有更高的要求，应尽可能分开部署。
- 生产环境强烈推荐使用更高的配置。
- TiKV 硬盘大小配置建议 PCI-E SSD 不超过 2 TB，普通 SSD 不超过 1.5 TB。
- TiFlash 支持多盘部署。
- TiFlash 数据目录的第一块磁盘推荐用高性能 SSD 来缓冲 TiKV 同步数据的实时写入，该盘性能应不低于 TiKV 所使用的磁盘，比如 PCI-E SSD。并且该磁盘容量建议不小于总容量的 10%，否则它可能成为这个节点的能承载的数据量的瓶颈。而其他磁盘可以根据需求部署多块普通 SSD，当然更好的 PCI-E SSD 硬盘会带来更好的性能。
- TiFlash 推荐与 TiKV 部署在不同节点，如果条件所限必须将 TiFlash 与 TiKV 部署在相同节点，则需要适当增加 CPU 核数和内存，且尽量将 TiFlash 与 TiKV 部署在不同的磁盘，以免互相干扰。
- TiFlash 硬盘总容量大致为：整个 `TiKV` 集群的需同步数据容量 `/` `TiKV` 副本数 `*` `TiFlash` 副本数。例如整体 TiKV 的规划容量为 1 TB、TiKV 副本数为 3、TiFlash 副本数为 2，则 TiFlash 的推荐总容量为 `1024 GB / 3 * 2`。用户可以选择同步部分表数据而非全部，具体容量可以根据需要同步的表的数据量具体分析。
- TiCDC 硬盘配置建议 1 TB+ PCIE-SSD。

2.1.4 网络要求

TiDB 作为开源分布式 NewSQL 数据库，其正常运行需要网络环境提供如下的网络端口配置要求，管理员可根据实际环境中 TiDB 组件部署的方案，在网络侧和主机侧开放相关端口：

组件	默认端口	说明
TiDB	4000	应用及 DBA 工具访问通信端口
TiDB	10080	TiDB 状态信息上报通信端口
TiKV	20160	TiKV 通信端口
TiKV	20180	TiKV 状态信息上报通信端口
PD	2379	提供 TiDB 和 PD 通信端口
PD	2380	PD 集群节点间通信端口
TiFlash	9000	TiFlash TCP 服务端口
TiFlash	8123	TiFlash HTTP 服务端口
TiFlash	3930	TiFlash RAFT 服务和 Coprocessor 服务端口
TiFlash	20170	TiFlash Proxy 服务端口
TiFlash	20292	Prometheus 拉取 TiFlash Proxy metrics 端口
TiFlash	8234	Prometheus 拉取 TiFlash metrics 端口

组件	默认端口	说明
Pump	8250	Pump 通信端口
Drainer	8249	Drainer 通信端口
CDC	8300	CDC 通信接口
Prometheus	9090	Prometheus 服务通信端口
Node_exporter	9100	TiDB 集群每个节点的系统信息上报通信端口
Blackbox_exporter	9115	Blackbox_exporter 通信端口，用于 TiDB 集群端口监控
Grafana	3000	Web 监控服务对外服务和客户端 (浏览器) 访问端口
Alertmanager	9093	告警 web 服务端口
Alertmanager	9094	告警通信端口

2.1.5 客户端 Web 浏览器要求

TiDB 提供了基于 Grafana 的技术平台，对数据库集群的各项指标进行可视化展现。采用支持 Javascript 的微软 IE、Google Chrome、 Mozilla Firefox 的较新版本即可访问监控入口。

2.2 TiDB 环境与系统配置检查

本文介绍部署 TiDB 前的环境检查操作，以下各项操作按优先级排序。

2.2.1 在 TiKV 部署目标机器上添加数据盘 EXT4 文件系统挂载参数

生产环境部署，建议使用 EXT4 类型文件系统的 NVME 类型的 SSD 磁盘存储 TiKV 数据文件。这个配置方案为最佳实施方案，其可靠性、安全性、稳定性已经在大量线上场景中得到证实。

使用 root 用户登录目标机器，将部署目标机器数据盘格式化成 ext4 文件系统，挂载时添加 nodelalloc 和 noatime 挂载参数。nodelalloc 是必选参数，否则 TiUP 安装时检测无法通过；noatime 是可选建议参数。

> **注意：**
>
> 如果你的数据盘已经格式化成 ext4 并挂载了磁盘，可先执行 umount /dev/nvme0n1p1 命令卸载，从编辑 /etc/fstab 文件步骤开始执行，添加挂载参数重新挂载即可。

以 /dev/nvme0n1 数据盘为例，具体操作步骤如下：

1. 查看数据盘。

```
fdisk -l
```

```
Disk /dev/nvme0n1: 1000 GB
```

2. 创建分区。

```
parted -s -a optimal /dev/nvme0n1 mklabel gpt -- mkpart primary ext4 1 -1
```

> **注意：**
>
> 使用 `lsblk` 命令查看分区的设备号：对于 nvme 磁盘，生成的分区设备号一般为 `nvme0n1p1`；对于普通磁盘（例如 `/dev/sdb`），生成的分区设备号一般为 `sdb1`。

3. 格式化文件系统。

```
mkfs.ext4 /dev/nvme0n1p1
```

4. 查看数据盘分区 UUID。

   本例中 `nvme0n1p1` 的 UUID 为 `c51eb23b-195c-4061-92a9-3fad812cc12f`。

```
lsblk -f
```

```
NAME    FSTYPE LABEL UUID                                 MOUNTPOINT
sda
├─sda1  ext4         237b634b-a565-477b-8371-6dff0c41f5ab /boot
├─sda2  swap         f414c5c0-f823-4bb1-8fdf-e531173a72ed
└─sda3  ext4         547909c1-398d-4696-94c6-03e43e317b60 /
sr0
nvme0n1
└─nvme0n1p1 ext4         c51eb23b-195c-4061-92a9-3fad812cc12f
```

5. 编辑 `/etc/fstab` 文件，添加 `nodelalloc` 挂载参数。

```
vi /etc/fstab
```

```
UUID=c51eb23b-195c-4061-92a9-3fad812cc12f /data1 ext4 defaults,nodelalloc,noatime 0 2
```

6. 挂载数据盘。

```
mkdir /data1 && \
mount -a
```

7. 执行以下命令，如果文件系统为 ext4，并且挂载参数中包含 `nodelalloc`，则表示已生效。

```
mount -t ext4
```

```
/dev/nvme0n1p1 on /data1 type ext4 (rw,noatime,nodelalloc,data=ordered)
```

2.2.2 检测及关闭系统 swap

TiDB 运行需要有足够的内存。如果内存不足，不建议使用 swap 作为内存不足的缓冲，因为这会降低性能。建议永久关闭系统 swap。

要永久关闭 swap，可执行以如下命令:

```
echo "vm.swappiness = 0">> /etc/sysctl.conf
swapoff -a && swapon -a
sysctl -p
```

> 注意：
>
> - 一起执行 `swapoff -a` 和 `swapon -a` 命令是为了刷新 swap，将 swap 里的数据转储回内存，并清空 swap 里的数据。不可省略 swappiness 设置而只执行 `swapoff -a`；否则，重启后 swap 会再次自动打开，使得操作失效。
> - 执行 `sysctl -p` 命令是为了在不重启的情况下使配置生效。

2.2.3 检测及关闭目标部署机器的防火墙

本段介绍如何关闭目标主机防火墙配置，因为在 TiDB 集群中，需要将节点间的访问端口打通才可以保证读写请求、数据心跳等信息的正常的传输。在普遍线上场景中，数据库到业务服务和数据库节点的网络联通都是在安全域内完成数据交互。如果没有特殊安全的要求，建议将目标节点的防火墙进行关闭。否则建议按照端口使用规则，将端口信息配置到防火墙服务的白名单中。

1. 检查防火墙状态（以 CentOS Linux release 7.7.1908 (Core) 为例）

```
sudo firewall-cmd --state
sudo systemctl status firewalld.service
```

2. 关闭防火墙服务

```
sudo systemctl stop firewalld.service
```

3. 关闭防火墙自动启动服务

```
sudo systemctl disable firewalld.service
```

4. 检查防火墙状态

```
sudo systemctl status firewalld.service
```

2.2.4 检测及安装 NTP 服务

TiDB 是一套分布式数据库系统，需要节点间保证时间的同步，从而确保 ACID 模型的事务线性一致性。目前解决授时的普遍方案是采用 NTP 服务，可以通过互联网中的 pool.ntp.org 授时服务来保证节点的时间同步，也可以使用离线环境自己搭建的 NTP 服务来解决授时。

采用如下步骤检查是否安装 NTP 服务以及与 NTP 服务器正常同步：

1. 执行以下命令，如果输出 running 表示 NTP 服务正在运行：

```
sudo systemctl status ntpd.service
```

```
ntpd.service - Network Time Service
Loaded: loaded (/usr/lib/systemd/system/ntpd.service; disabled; vendor preset: disabled)
Active: active (running) since 一 2017-12-18 13:13:19 CST; 3s ago
```

   - 若返回报错信息 Unit ntpd.service could not be found.，请尝试执行以下命令，以查看与 NTP 进行时钟同步所使用的系统配置是 chronyd 还是 ntpd：

```
sudo systemctl status chronyd.service
```

```
chronyd.service - NTP client/server
Loaded: loaded (/usr/lib/systemd/system/chronyd.service; enabled; vendor preset:
    ↪ enabled)
Active: active (running) since Mon 2021-04-05 09:55:29 EDT; 3 days ago
```

   若发现系统既没有配置 chronyd 也没有配置 ntpd，则表示系统尚未安装任一服务。此时，应先安装其中一个服务，并保证它可以自动启动，默认使用 ntpd。

如果你使用的系统配置是 `chronyd`，请直接执行步骤 3。

2. 执行 ntpstat 命令检测是否与 NTP 服务器同步：

> **注意：**
>
> Ubuntu 系统需安装 ntpstat 软件包。

```
ntpstat
```

   - 如果输出 synchronised to NTP server，表示正在与 NTP 服务器正常同步：

```
synchronised to NTP server (85.199.214.101) at stratum 2
time correct to within 91 ms
polling server every 1024 s
```

   - 以下情况表示 NTP 服务未正常同步：

```
unsynchronised
```

- 以下情况表示 NTP 服务未正常运行：

```
Unable to talk to NTP daemon. Is it running?
```

3. 执行 `chronyc tracking` 命令查看 Chrony 服务是否与 NTP 服务器同步。

> **注意：**
> 该操作仅适用于使用 Chrony 的系统，不适用于使用 NTPd 的系统。

```
chronyc tracking
```

- 如果该命令返回结果为 `Leap status : Normal`，则代表同步过程正常。

```
Reference ID    : 5EC69F0A (ntp1.time.nl)
Stratum         : 2
Ref time (UTC)  : Thu May 20 15:19:08 2021
System time     : 0.000022151 seconds slow of NTP time
Last offset     : -0.000041040 seconds
RMS offset      : 0.000053422 seconds
Frequency       : 2.286 ppm slow
Residual freq   : -0.000 ppm
Skew            : 0.012 ppm
Root delay      : 0.012706812 seconds
Root dispersion : 0.000430042 seconds
Update interval : 1029.8 seconds
Leap status     : Normal
```

- 如果该命令返回结果如下，则表示同步过程出错：

```
Leap status    : Not synchronised
```

- 如果该命令返回结果如下，则表示 Chrony 服务未正常运行：

```
506 Cannot talk to daemon
```

如果要使 NTP 服务尽快开始同步，执行以下命令。可以将 `pool.ntp.org` 替换为你的 NTP 服务器：

```
sudo systemctl stop ntpd.service && \
sudo ntpdate pool.ntp.org && \
sudo systemctl start ntpd.service
```

如果要在 CentOS 7 系统上手动安装 NTP 服务，可执行以下命令：

```
sudo yum install ntp ntpdate && \
sudo systemctl start ntpd.service && \
sudo systemctl enable ntpd.service
```

2.2.5 检查和配置操作系统优化参数

在生产系统的 TiDB 中，建议对操作系统进行如下的配置优化：

1. 关闭透明大页（即 Transparent Huge Pages，缩写为 THP）。数据库的内存访问模式往往是稀疏的而非连续的。当高阶内存碎片化比较严重时，分配 THP 页面会出现较高的延迟。
2. 将存储介质的 I/O 调度器设置为 noop。对于高速 SSD 存储介质，内核的 I/O 调度操作会导致性能损失。将调度器设置为 noop 后，内核不做任何操作，直接将 I/O 请求下发给硬件，以获取更好的性能。同时，noop 调度器也有较好的普适性。
3. 为调整 CPU 频率的 cpufreq 模块选用 performance 模式。将 CPU 频率固定在其支持的最高运行频率上，不进行动态调节，可获取最佳的性能。

采用如下步骤检查操作系统的当前配置，并配置系统优化参数：

1. 执行以下命令查看透明大页的开启状态。

```
cat /sys/kernel/mm/transparent_hugepage/enabled
```

```
[always] madvise never
```

> 注意：
> `[always] madvise never` 表示透明大页处于启用状态，需要关闭。

2. 执行以下命令查看数据目录所在磁盘的 I/O 调度器。假设在 sdb、sdc 两个磁盘上创建了数据目录。

```
cat /sys/block/sd[bc]/queue/scheduler
```

```
noop [deadline] cfq
noop [deadline] cfq
```

> 注意：
> `noop [deadline] cfq` 表示磁盘的 I/O 调度器使用 `deadline`，需要进行修改。

3. 执行以下命令查看磁盘的唯一标识 `ID_SERIAL`。

```
udevadm info --name=/dev/sdb | grep ID_SERIAL
```

```
E: ID_SERIAL=36d0946606d79f90025f3e09a0c1f9e81
E: ID_SERIAL_SHORT=6d0946606d79f90025f3e09a0c1f9e81
```

> **注意：**
> 如果多个磁盘都分配了数据目录，需要多次执行以上命令，记录所有磁盘各自的唯一标识。

4. 执行以下命令查看 cpufreq 模块选用的节能策略。

```
cpupower frequency-info --policy
```

```
analyzing CPU 0:
current policy: frequency should be within 1.20 GHz and 3.10 GHz.
              The governor "powersave" may decide which speed to use within this range.
```

> **注意：**
> The governor "powersave" 表示 cpufreq 的节能策略使用 powersave，需要调整为 performance 策略。如果是虚拟机或者云主机，则不需要调整，命令输出通常为 Unable to determine current policy。

5. 配置系统优化参数

- 方法一：使用 tuned（推荐）
  1. 执行 tuned-adm list 命令查看当前操作系统的 tuned 策略。

```
tuned-adm list
```

```
Available profiles:
- balanced                    - General non-specialized tuned profile
- desktop                     - Optimize for the desktop use-case
- hpc-compute                 - Optimize for HPC compute workloads
- latency-performance         - Optimize for deterministic performance at the cost
    ↪ of increased power consumption
- network-latency             - Optimize for deterministic performance at the cost
    ↪ of increased power consumption, focused on low latency network performance
- network-throughput          - Optimize for streaming network throughput,
    ↪ generally only necessary on older CPUs or 40G+ networks
- powersave                   - Optimize for low power consumption
- throughput-performance      - Broadly applicable tuning that provides excellent
    ↪ performance across a variety of common server workloads
- virtual-guest               - Optimize for running inside a virtual guest
- virtual-host                - Optimize for running KVM guests
Current active profile: balanced
```

`Current active profile: balanced` 表示当前操作系统的 tuned 策略使用 balanced，建议在当前策略的基础上添加操作系统优化配置。

2. 创建新的 tuned 策略。

```
mkdir /etc/tuned/balanced-tidb-optimal/
vi /etc/tuned/balanced-tidb-optimal/tuned.conf
```

```
[main]
include=balanced

[cpu]
governor=performance

[vm]
transparent_hugepages=never

[disk]
devices_udev_regex=(ID_SERIAL=36d0946606d79f90025f3e09a0c1fc035)|(ID_SERIAL=36
    ↪ d0946606d79f90025f3e09a0c1f9e81)
elevator=noop
```

`include=balanced` 表示在现有的 balanced 策略基础上添加操作系统优化配置。

3. 应用新的 tuned 策略。

```
tuned-adm profile balanced-tidb-optimal
```

- 方法二：使用脚本方式。如果已经使用 tuned 方法，请跳过本方法。

1. 执行 grubby 命令查看默认内核版本。

> 注意：
> 需安装 grubby 软件包。

```
grubby --default-kernel
```

```
/boot/vmlinuz-3.10.0-957.el7.x86_64
```

2. 执行 `grubby --update-kernel` 命令修改内核配置。

```
grubby --args="transparent_hugepage=never" --update-kernel /boot/vmlinuz
    ↪ -3.10.0-957.el7.x86_64
```

> 注意：
> `--update-kernel` 后需要使用实际的默认内核版本。

3. 执行 `grubby --info` 命令查看修改后的默认内核配置。

```
grubby --info /boot/vmlinuz-3.10.0-957.el7.x86_64
```

> **注意：**
> `--info` 后需要使用实际的默认内核版本。

```
index=0
kernel=/boot/vmlinuz-3.10.0-957.el7.x86_64
args="ro crashkernel=auto rd.lvm.lv=centos/root rd.lvm.lv=centos/swap rhgb quiet
    ↪ LANG=en_US.UTF-8 transparent_hugepage=never"
root=/dev/mapper/centos-root
initrd=/boot/initramfs-3.10.0-957.el7.x86_64.img
title=CentOS Linux (3.10.0-957.el7.x86_64) 7 (Core)
```

4. 修改当前的内核配置立即关闭透明大页。

```
echo never > /sys/kernel/mm/transparent_hugepage/enabled
echo never > /sys/kernel/mm/transparent_hugepage/defrag
```

5. 配置 udev 脚本应用 IO 调度器策略。

```
vi /etc/udev/rules.d/60-tidb-schedulers.rules
```

```
ACTION=="add|change", SUBSYSTEM=="block", ENV{ID_SERIAL}=="36
    ↪ d0946606d79f90025f3e09a0c1fc035", ATTR{queue/scheduler}="noop"
ACTION=="add|change", SUBSYSTEM=="block", ENV{ID_SERIAL}=="36
    ↪ d0946606d79f90025f3e09a0c1f9e81", ATTR{queue/scheduler}="noop"
```

6. 应用 udev 脚本。

```
udevadm control --reload-rules
udevadm trigger --type=devices --action=change
```

7. 创建 CPU 节能策略配置服务。

```
cat  >> /etc/systemd/system/cpupower.service << EOF
[Unit]
Description=CPU performance
[Service]
Type=oneshot
ExecStart=/usr/bin/cpupower frequency-set --governor performance
[Install]
WantedBy=multi-user.target
EOF
```

8. 应用 CPU 节能策略配置服务。

```
systemctl daemon-reload
systemctl enable cpupower.service
systemctl start cpupower.service
```

6. 执行以下命令验证透明大页的状态。

```
cat /sys/kernel/mm/transparent_hugepage/enabled
```

```
always madvise [never]
```

7. 执行以下命令验证数据目录所在磁盘的 I/O 调度器。

```
cat /sys/block/sd[bc]/queue/scheduler
```

```
[noop] deadline cfq
[noop] deadline cfq
```

8. 执行以下命令查看 cpufreq 模块选用的节能策略。

```
cpupower frequency-info --policy
```

```
analyzing CPU 0:
current policy: frequency should be within 1.20 GHz and 3.10 GHz.
              The governor "performance" may decide which speed to use within this range.
```

9. 执行以下命令修改 sysctl 参数。

```
echo "fs.file-max = 1000000">> /etc/sysctl.conf
echo "net.core.somaxconn = 32768">> /etc/sysctl.conf
echo "net.ipv4.tcp_tw_recycle = 0">> /etc/sysctl.conf
echo "net.ipv4.tcp_syncookies = 0">> /etc/sysctl.conf
echo "vm.overcommit_memory = 1">> /etc/sysctl.conf
sysctl -p
```

10. 执行以下命令配置用户的 limits.conf 文件。

```
cat << EOF >>/etc/security/limits.conf
tidb           soft    nofile          1000000
tidb           hard    nofile          1000000
tidb           soft    stack          32768
tidb           hard    stack          32768
EOF
```

2.2.6 手动配置 SSH 互信及 sudo 免密码

对于有需求，通过手动配置中控机至目标节点互信的场景，可参考本段。通常推荐使用 TiUP 部署工具会自动配置 SSH 互信及免密登录，可忽略本段内容。

1. 以 root 用户依次登录到部署目标机器创建 tidb 用户并设置登录密码。

```
useradd tidb && \
passwd tidb
```

2. 执行以下命令，将 tidb ALL=(ALL)NOPASSWD: ALL 添加到文件末尾，即配置好 sudo 免密码。

```
visudo
```

```
tidb ALL=(ALL) NOPASSWD: ALL
```

3. 以 tidb 用户登录到中控机，执行以下命令。将 10.0.1.1 替换成你的部署目标机器 IP，按提示输入部署目标机器 tidb 用户密码，执行成功后即创建好 SSH 互信，其他机器同理。新建的 tidb 用户下没有 .ssh 目录，需要执行生成 rsa 密钥的命令来生成 .ssh 目录。如果要在中控机上部署 TiDB 组件，需要为中控机和中控机自身配置互信。

```
ssh-keygen -t rsa
ssh-copy-id -i ~/.ssh/id_rsa.pub 10.0.1.1
```

4. 以 tidb 用户登录中控机，通过 ssh 的方式登录目标机器 IP。如果不需要输入密码并登录成功，即表示 SSH 互信配置成功。

```
ssh 10.0.1.1
```

```
[tidb@10.0.1.1 ~]$
```

5. 以 tidb 用户登录到部署目标机器后，执行以下命令，不需要输入密码并切换到 root 用户，表示 tidb 用户 sudo 免密码配置成功。

```
sudo -su root
```

```
[root@10.0.1.1 tidb]#
```

2.2.7 安装 numactl 工具

本段主要介绍如果安装 NUMA 工具。在生产环境中，因为硬件机器配置往往高于需求，为了更合理规划资源，会考虑单机多实例部署 TiDB 或者 TiKV。NUMA 绑核工具的使用，主要为了防止 CPU 资源的争抢，引发性能衰退。

> **注意：**

- NUMA 绑核是用来隔离 CPU 资源的一种方法，适合高配置物理机环境部署多实例使用。
- 通过 `tiup cluster deploy` 完成部署操作，就可以通过 `exec` 命令来进行集群级别管理工作。

1. 登录到目标节点进行安装（以 CentOS Linux release 7.7.1908 (Core) 为例）

```
sudo yum -y install numactl
```

2. 通过 TiUP 的 cluster 执行完 exec 命令来完成批量安装

```
tiup cluster exec --help
```

```
Run shell command on host in the tidb cluster

Usage:
cluster exec <cluster-name> [flags]

Flags:
    --command string   the command run on cluster host (default "ls")
-h, --help             help for exec
    --sudo             use root permissions (default false)
```

将 tidb-test 集群所有目标主机通过 sudo 权限执行安装命令

```
tiup cluster exec tidb-test --sudo --command "yum -y install numactl"
```

2.3 规划集群拓扑

2.3.1 最小拓扑架构

本文档介绍 TiDB 集群最小部署的拓扑架构。

2.3.1.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	3	16 VCore 32GB * 1	10.0.1.1 10.0.1.2 10.0.1.3	默认端口全局目录配置
PD	3	4 VCore 8GB * 1	10.0.1.4 10.0.1.5 10.0.1.6	默认端口全局目录配置
TiKV	3	16 VCore 32GB 2TB (nvme ssd) * 1	10.0.1.7 10.0.1.8 10.0.1.9	默认端口全局目录配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.10	默认端口全局目录配置

2.3.1.1.1 拓扑模版

简单最小配置模板

详细最小配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

> 注意：
>
> - 无需手动创建配置文件中的 `tidb` 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
> - 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.2 TiFlash 部署拓扑

本文介绍在部署最小拓扑集群的基础上，部署 TiFlash 的拓扑结构。TiFlash 是列式的存储引擎，已经成为集群拓扑的标配，适合 Real-Time HTAP 业务。

2.3.2.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	3	16 VCore 32GB * 1	10.0.1.1 10.0.1.2 10.0.1.3	默认端口全局目录配置
PD	3	4 VCore 8GB * 1	10.0.1.4 10.0.1.5 10.0.1.6	默认端口全局目录配置
TiKV	3	16 VCore 32GB 2TB (nvme ssd) * 1	10.0.1.1 10.0.1.2 10.0.1.3	默认端口全局目录配置
TiFlash	1	32 VCore 64 GB 2TB (nvme ssd) * 1	10.0.1.10	默认端口全局目录配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.10	默认端口全局目录配置

2.3.2.1.1 拓扑模版

- 简单 TiFlash 配置模版
- 详细 TiFlash 配置模版

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

2.3.2.1.2 关键参数介绍

- 需要将配置模板中 `replication.enable-placement-rules` 设置为 `true`，以开启 PD 的Placement Rules 功能。
- `tiflash_servers` 实例级别配置 `"-host"` 目前只支持 IP，不支持域名。
- TiFlash 具体的参数配置介绍可参考 TiFlash 参数配置。

注意：

- 无需手动创建配置文件中的 tidb 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
- 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.3 TiCDC 部署拓扑

注意：

TiCDC 从 v4.0.6 起成为正式功能，可用于生产环境。

本文介绍TiCDC 部署的拓扑，以及如何在最小拓扑的基础上同时部署 TiCDC。TiCDC 是 4.0 版本开始支持的 TiDB 增量数据同步工具，支持多种下游 (TiDB/MySQL/MQ)。相比于 TiDB Binlog，TiCDC 有延迟更低、天然高可用等优点。

2.3.3.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	3	16 VCore 32GB * 1	10.0.1.1 10.0.1.2 10.0.1.3	默认端口全局目录配置
PD	3	4 VCore 8GB * 1	10.0.1.4 10.0.1.5 10.0.1.6	默认端口全局目录配置
TiKV	3	16 VCore 32GB 2TB (nvme ssd) * 1	10.0.1.7 10.0.1.8 10.0.1.9	默认端口全局目录配置
CDC	3	8 VCore 16GB * 1	10.0.1.11 10.0.1.12 10.0.1.13	默认端口全局目录配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.11	默认端口全局目录配置

2.3.3.1.1 拓扑模版

简单 TiCDC 配置模板

详细 TiCDC 配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

注意：

- 无需手动创建配置文件中的 tidb 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
- 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.4 TiDB Binlog 部署拓扑

本文介绍在部署最小拓扑集群的基础上，同时部署TiDB Binlog。TiDB Binlog 是目前广泛使用的增量同步组件，可提供准实时备份和同步功能。

2.3.4.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	3	16 VCore 32 GB	10.0.1.1 10.0.1.2 10.0.1.3	默认端口配置；开启 enable_binlog；开启 ignore-error
PD	3	4 VCore 8 GB	10.0.1.4 10.0.1.5 10.0.1.6	默认端口配置
TiKV	3	16 VCore 32 GB	10.0.1.7 10.0.1.8 10.0.1.9	默认端口配置

实例	个数	物理机配置	IP	配置
Pump	3	8 VCore 16GB	10.0.1.1 10.0.1.7 10.0.1.8	默认端口配置；设置 GC 时间 7 天

实例	个数	物理机配置	IP	配置
Drainer	1	8 VCore 16GB	10.0.1.11	默认端口配置；设置默认初始化 commitTS -1 为最近的时间戳配置下游目标 TiDB 10.0.1.12:4000

2.3.4.1.1 拓扑模版

简单 TiDB Binlog 配置模板（下游为 MySQL）

简单 TiDB Binlog 配置模板（下游为 file）

详细 TiDB Binlog 配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

2.3.4.1.2 关键参数介绍

拓扑配置模版的关键参数如下：

- `server_configs.tidb.binlog.enable: true`

  开启 TiDB Binlog 服务，默认为 false。

- `server_configs.tidb.binlog.ignore-error: true`

  高可用场景建议开启，如果设置为 true，发生错误时，TiDB 会停止写入 TiDB Binlog，并且在监控项 `tidb_server_critical_error_total` 上计数加 1；如果设置为 false，一旦写入 TiDB Binlog 失败，会停止整个 TiDB 的服务。

- `drainer_servers.config.syncer.db-type`

  TiDB Binlog 的下游类型，目前支持 `mysql`、`tidb`、`kafka` 和 `file`。

- `drainer_servers.config.syncer.to`

  TiDB Binlog 的下游配置。根据 `db-type` 的不同，该选项可配置下游数据库的连接参数、Kafka 的连接参数、文件保存路径。详细说明可参见TiDB Binlog 配置说明。

> 注意：
>
> - 编辑配置文件模版时，如无需自定义端口或者目录，仅修改 IP 即可。
> - 无需手动创建配置文件中的 `tidb` 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
> - 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.5 TiSpark 部署拓扑

本文介绍 TiSpark 部署的拓扑，以及如何在最小拓扑的基础上同时部署 TiSpark。TiSpark 是 PingCAP 为解决用户复杂 OLAP 需求而推出的产品。它借助 Spark 平台，同时融合 TiKV 分布式集群的优势，和 TiDB 一起为用户一站式解决 HTAP (Hybrid Transactional/Analytical Processing) 的需求。

关于 TiSpark 的架构介绍与使用，参见TiSpark 用户指南。

> 警告：
>
> TiUP Cluster 的 TiSpark 支持目前为实验特性，不建议在生产环境中使用。

2.3.5.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	3	16 VCore 32GB * 1	10.0.1.1 10.0.1.2 10.0.1.3	默认端口全局目录配置
PD	3	4 VCore 8GB * 1	10.0.1.4 10.0.1.5 10.0.1.6	默认端口全局目录配置
TiKV	3	16 VCore 32GB 2TB (nvme ssd) * 1	10.0.1.7 10.0.1.8 10.0.1.9	默认端口全局目录配置

实例	个数	物理机配置	IP	配置
TiSpark	3	8 VCore 16GB * 1	10.0.1.21 (master) 10.0.1.22 (worker) 10.0.1.23 (worker)	默认端口全局目录配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.11	默认端口全局目录配置

2.3.5.1.1 拓扑模版

简单 TiSpark 配置模板

详细 TiSpark 配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

> 注意：
>
> - 无需手动创建配置文件中的 `tidb` 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
> - 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.5.2 环境要求

由于 TiSpark 基于 Apache Spark 集群，在启动包含 TiSpark 组件的 TiDB 集群前，需要在部署了 TiSpark 组件的服务器上安装 Java 运行时环境 (JRE) 8，否则将无法启动相关组件。

TiUP 不提供自动安装 JRE 的支持，该操作需要用户自行完成。JRE 8 的安装方法可以参考 OpenJDK 的文档说明。

如果部署服务器上已经安装有 JRE 8，但不在系统的默认包管理工具路径中，可以通过在拓扑配置中设置 java_home 参数来指定要使用的 JRE 环境所在的路径。该参数对应系统环境变量 JAVA_HOME。

2.3.6 跨数据中心部署拓扑

本文以典型的两地三中心为例，介绍跨数据中心部署的拓扑以及关键参数。本文示例所涉及的城市是上海（即 sha）和北京（即 bja 和 bjb）。

2.3.6.1 拓扑信息

实例	个数	物理机配置	BJ IP	SH IP	配置
TiDB	5	16 VCore 32GB * 1	10.0.1.1 10.0.1.2 10.0.1.3 10.0.1.4	10.0.1.5	默认端口全局目录配置
PD	5	4 VCore 8GB * 1	10.0.1.6 10.0.1.7 10.0.1.8 10.0.1.9	10.0.1.10	默认端口全局目录配置

实例	个数	物理机配置	BJ IP	SH IP	配置
TiKV	5	16 VCore 32GB 2TB (nvme ssd) * 1	10.0.1.11 10.0.1.12 10.0.1.13 10.0.1.14	10.0.1.15	默认端口全局目录配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.16		默认端口全局目录配置

2.3.6.1.1 拓扑模版

跨机房配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

2.3.6.1.2 关键参数配置

本节介绍跨数据中心部署 TiDB 集群的关键参数配置。

TiKV 参数

- 设置 gRPC 的压缩格式，默认为 none。为提高跨机房部署场景的目标节点间 gRPC 包的传输速度，建议设置为 gzip 格式。

```
server.grpc-compression-type: gzip
```

- label 配置

  由于采用跨机房部署 TiKV，为了避免物理机宕机导致 Raft Group 默认的 5 副本中丢失 3 副本，使集群不可用的问题，可以通过 label 来实现 PD 智能调度，保证同中心、同机柜、同机器 TiKV 实例不会出现 Raft Group 有 3 副本的情况。

- TiKV 配置

  相同物理机配置相同的 host 级别 label 信息：

```
config:
  server.labels:
    zone: bj
    dc: bja
    rack: rack1
    host: host2
```

- 防止异地 TiKV 节点发起不必要的 Raft 选举，需要将异地 TiKV 节点发起选举时经过最少的 tick 个数和最多经过的 tick 个数都调大，这两个参数默认设置均为 0。

```
raftstore.raft-min-election-timeout-ticks: 1000
raftstore.raft-max-election-timeout-ticks: 1020
```

PD 参数

- PD 元数据信息记录 TiKV 集群的拓扑信息，根据四个维度调度 Raft Group 副本。

```
replication.location-labels: ["zone","dc","rack","host"]
```

- 调整 Raft Group 的副本数据量为 5 ，保证集群的高可用性。

```
replication.max-replicas: 5
```

- 拒绝异地机房 TiKV 的 Raft 副本选举为 Leader。

```
label-property:
      reject-leader:
        - key: "dc"
          value: "sha"
```

> 注意：
>
> TiDB 5.2 及以上版本默认不支持 `label-property` 配置。若要设置副本策略，请使用Placement Rules。

有关 Label 的使用和 Raft Group 副本数量，详见通过拓扑 label 进行副本调度。

> 注意：
>
> - 无需手动创建配置文件中的 `tidb` 用户，TiUP cluster 组件会在目标主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
> - 如果部署目录配置为相对路径，会部署在用户家目录下。

2.3.7 混合部署拓扑

本文介绍 TiDB 集群的 TiKV 和 TiDB 混合部署拓扑以及主要参数。常见的场景为，部署机为多路 CPU 处理器，内存也充足，为提高物理机资源利用率，可单机多实例部署，即 TiDB、TiKV 通过 numa 绑核，隔离 CPU 资源。PD 和 Prometheus 混合部署，但两者的数据目录需要使用独立的文件系统。

2.3.7.1 拓扑信息

实例	个数	物理机配置	IP	配置
TiDB	6	32 VCore 64GB	10.0.1.1 10.0.1.2 10.0.1.3	配置 numa 绑核操作
PD	3	16 VCore 32 GB	10.0.1.4 10.0.1.5 10.0.1.6	配置 location_labels 参数

实例	个数	物理机配置	IP	配置
TiKV	6	32 VCore 64GB	10.0.1.7 10.0.1.8 10.0.1.9	1. 区分实例级别的 port、status_port； 2. 配置全局参数 readpool、storage 以及 raftstore； 3. 配置实例级别 host 维度的 labels； 4. 配置 numa 绑核

实例	个数	物理机配置	IP	配置
Monitoring & Grafana	1	4 VCore 8GB * 1 500GB (ssd)	10.0.1.11	默认配置

2.3.7.1.1 拓扑模版

简单混部配置模板

详细混部配置模板

以上 TiDB 集群拓扑文件中，详细的配置项说明见通过 TiUP 部署 TiDB 集群的拓扑文件配置。

2.3.7.1.2 混合部署的关键参数介绍

本节介绍单机多实例的关键参数，主要用于 TiDB、TiKV 的单机多实例部署场景。你需要按照提供的计算公式，将结果填写至上一步的配置模板中。

- TiKV 进行配置优化
  - readpool 线程池自适应，配置 `readpool.unified.max-thread-count` 参数可以使 `readpool.storage` 和 `readpool.coprocessor` 共用统一线程池，同时要分别设置自适应开关。
    * 开启 `readpool.storage` 和 `readpool.coprocessor`：

```
readpool.storage.use-unified-pool: true
readpool.coprocessor.use-unified-pool: true
```

    * 计算公式如下：

```
readpool.unified.max-thread-count = cores * 0.8 / TiKV 数量
```

  - storage CF (all RocksDB column families) 内存自适应，配置 `storage.block-cache.capacity` 参数即可实现 CF 之间自动平衡内存使用。
    * `storage.block-cache` 默认开启 CF 自适应，无需修改。

```
storage.block-cache.shared: true
```

    * 计算公式如下：

```
storage.block-cache.capacity = (MEM_TOTAL * 0.5 / TiKV 实例数量)
```

- 如果多个 TiKV 实例部署在同一块物理磁盘上，需要在 tikv 配置中添加 capacity 参数：

```
raftstore.capacity = 磁盘总容量 / TiKV 实例数量
```

- label 调度配置

  由于采用单机多实例部署 TiKV，为了避免物理机宕机导致 Region Group 默认 3 副本的 2 副本丢失，导致集群不可用的问题，可以通过 label 来实现 PD 智能调度，保证同台机器的多 TiKV 实例不会出现 Region Group 只有 2 副本的情况。

  - TiKV 配置

    相同物理机配置相同的 host 级别 label 信息：

```
config:
  server.labels:
    host: tikv1
```

  - PD 配置

    PD 需要配置 labels 类型来识别并调度 Region：

```
pd:
  replication.location-labels: ["host"]
```

- numa_node 绑核

  - 在实例参数模块配置对应的 numa_node 参数，并添加对应的物理 CPU 的核数；
  - numa 绑核使用前，确认已经安装 numactl 工具，以及物理机对应的物理机 CPU 的信息后，再进行参数配置；
  - numa_node 这个配置参数与 `numactl --membind` 配置对应。

注意：

- 编辑配置文件模版时，注意修改必要参数、IP、端口及目录。
- 各个组件的 deploy_dir，默认会使用 global 中的 <deploy_dir>/<components_name>-<port>。例如 tidb 端口指定 4001，则 deploy_dir 默认为 ‘/tidb-deploy/tidb-4001’。因此，在多实例场景下指定非默认端口时，无需再次指定目录。
- 无需手动创建配置文件中的 tidb 用户，TiUP cluster 组件会在部署主机上自动创建该用户。可以自定义用户，也可以和中控机的用户保持一致。
- 如果部署目录配置为相对路径，会部署在用户家目录下。

2.4 安装与启动

2.4.1 使用 TiUP 部署 TiDB 集群

TiUP 是 TiDB 4.0 版本引入的集群运维工具，TiUP cluster 是 TiUP 提供的使用 Golang 编写的集群管理组件，通过 TiUP cluster 组件就可以进行日常的运维工作，包括部署、启动、关闭、销毁、弹性扩缩容、升级 TiDB 集群，以及管

理 TiDB 集群参数。

目前 TiUP 可以支持部署 TiDB、 TiFlash、 TiDB Binlog、 TiCDC 以及监控系统。本文将介绍不同集群拓扑的具体部署步骤。

> 注意：
>
> TiDB、 TiUP 及 TiDB Dashboard 默认会收集使用情况信息，并将这些信息分享给 PingCAP 用于改善产品。若要了解所收集的信息详情及如何禁用该行为，请参见遥测。

2.4.1.1 第 1 步：软硬件环境需求及前置检查

软硬件环境需求

环境与系统配置检查

2.4.1.2 第 2 步：在中控机上部署 TiUP 组件

在中控机上部署 TiUP 组件有两种方式：在线部署和离线部署。

2.4.1.2.1 在线部署

以普通用户身份登录中控机。以 `tidb` 用户为例，后续安装 TiUP 及集群管理操作均通过该用户完成：

1. 执行如下命令安装 TiUP 工具：

```
curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
```

2. 按如下步骤设置 TiUP 环境变量：

    1. 重新声明全局环境变量：

    ```
    shell source .bash_profile
    ```

    2. 确认 TiUP 工具是否安装：

    ```
    which tiup
    ```

3. 安装 TiUP cluster 组件：

```
tiup cluster
```

4. 如果已经安装，则更新 TiUP cluster 组件至最新版本：

```
tiup update --self && tiup update cluster
```

预期输出 “`Update successfully!`” 字样。

5. 验证当前 TiUP cluster 版本信息。执行如下命令查看 TiUP cluster 组件版本：

```
tiup --binary cluster
```

2.4.1.2.2 离线部署

离线部署 TiUP 组件的操作步骤如下。

准备 TiUP 离线组件包

方式一：在官方下载页面选择对应版本的 TiDB server 离线镜像包（包含 TiUP 离线组件包）。

方式二：使用 `tiup mirror clone` 命令手动打包离线组件包。步骤如下：

1. 在在线环境中安装 TiUP 包管理器工具

    1. 执行如下命令安装 TiUP 工具：

    ```
    curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
    ```

    2. 重新声明全局环境变量：

    ```
    source .bash_profile
    ```

    3. 确认 TiUP 工具是否安装：

    ```
    which tiup
    ```

2. 使用 TiUP 制作离线镜像

    1. 在一台和外网相通的机器上拉取需要的组件：

    ```
    tiup mirror clone tidb-community-server-${version}-linux-amd64 ${version} --os=linux --
        ↪ arch=amd64
    ```

    该命令会在当前目录下创建一个名叫 `tidb-community-server-${version}-linux-amd64` 的目录，里面包含 TiUP 管理的组件包。

    2. 通过 tar 命令将该组件包打包然后发送到隔离环境的中控机：

    ```
    tar czvf tidb-community-server-${version}-linux-amd64.tar.gz tidb-community-server-${
        ↪ version}-linux-amd64
    ```

    此时，`tidb-community-server-${version}-linux-amd64.tar.gz` 就是一个独立的离线环境包。

3. 自定义制作的离线镜像，或调整已有离线镜像中的内容

    如果从官网下载的离线镜像不满足你的具体需求，或者希望对已有的离线镜像内容进行调整，例如增加某个组件的新版本等，可以采取以下步骤进行操作：

    1. 在制作离线镜像时，可通过参数指定具体的组件和版本等信息，获得不完整的离线镜像。例如，要制作一个只包括 v1.9.0 版本 TiUP 和 TiUP Cluster 的离线镜像，可执行如下命令：

    ```
    tiup mirror clone tiup-custom-mirror-v1.9.0 --tiup v1.9.0 --cluster v1.9.0
    ```

    如果只需要某一特定平台的组件，也可以通过 `--os` 和 `--arch` 参数来指定。

    2. 参考上文“使用 TiUP 制作离线镜像”第 2 步的方式，将此不完整的离线镜像传输到隔离环境的中控机。

3. 在隔离环境的中控机上，查看当前使用的离线镜像路径。较新版本的 TiUP 可以直接通过命令获取当前的镜像地址：

```
tiup mirror show
```

以上命令如果提示 show 命令不存在，可能当前使用的是较老版本的 TiUP。此时可以通过查看 $HOME/.tiup/tiup.toml 获得正在使用的镜像地址。将此镜像地址记录下来，后续步骤中将以变量 ${base_mirror} 指代此镜像地址。

4. 将不完整的离线镜像合并到已有的离线镜像中：

首先将当前离线镜像中的 keys 目录复制到 $HOME/.tiup 目录中：

```
cp -r ${base_mirror}/keys $HOME/.tiup/
```

然后使用 TiUP 命令将不完整的离线镜像合并到当前使用的镜像中：

```
tiup mirror merge tiup-custom-mirror-v1.9.0
```

5. 上述步骤完成后，通过 tiup list 命令检查执行结果。在本文例子中，使用 tiup list tiup 和 tiup list cluster 均应能看到对应组件的 v1.9.0 版本出现在结果中。

部署离线环境 TiUP 组件

将离线包发送到目标集群的中控机后，执行以下命令安装 TiUP 组件：

```
tar xzvf tidb-community-server-${version}-linux-amd64.tar.gz && \
sh tidb-community-server-${version}-linux-amd64/local_install.sh && \
source /home/tidb/.bash_profile
```

local_install.sh 脚本会自动执行 tiup mirror set tidb-community-server-${version}-linux-amd64 命令将当前镜像地址设置为 tidb-community-server-${version}-linux-amd64。

若需将镜像切换到其他目录，可以通过手动执行 tiup mirror set <mirror-dir> 进行切换。如果需要切换到在线环境，可执行 tiup mirror set https://tiup-mirrors.pingcap.com。

2.4.1.3 第 3 步：初始化集群拓扑文件

执行如下命令，生成集群初始化配置文件：

```
tiup cluster template > topology.yaml
```

针对两种常用的部署场景，也可以通过以下命令生成建议的拓扑模板：

- 混合部署场景：单台机器部署多个实例，详情参见混合部署拓扑架构 。

```
tiup cluster template --full > topology.yaml
```

- 跨机房部署场景：跨机房部署 TiDB 集群，详情参见跨机房部署拓扑架构。

```
tiup cluster template --multi-dc > topology.yaml
```

执行 vi topology.yaml，查看配置文件的内容：

```
global:
  user: "tidb"
  ssh_port: 22
  deploy_dir: "/tidb-deploy"
  data_dir: "/tidb-data"
server_configs: {}
pd_servers:
  - host: 10.0.1.4
  - host: 10.0.1.5
  - host: 10.0.1.6
tidb_servers:
  - host: 10.0.1.7
  - host: 10.0.1.8
  - host: 10.0.1.9
tikv_servers:
  - host: 10.0.1.1
  - host: 10.0.1.2
  - host: 10.0.1.3
monitoring_servers:
  - host: 10.0.1.4
grafana_servers:
  - host: 10.0.1.4
alertmanager_servers:
  - host: 10.0.1.4
```

下表列出了常用的 7 种场景，请根据链接中的拓扑说明以及配置文件模板配置`topology.yaml`。如果有其他组合场景的需求，请根据多个模板自行调整。

场景	配置任务	配置文件模板	拓扑说明
OLTP 业务	部署最小拓扑架构	简单最小配置模板详细最小配置模板	最小集群拓扑，包括 tidb-server、tikv-server、pd-server。

场景	配置任务	配置文件模板	拓扑说明
HTAP 业务	部署 Ti-Flash 拓扑架构	简单 Ti-Flash 配置模版 详细 Ti-Flash 配置模版	在最小拓扑的基础上部署 Ti-Flash。Ti-Flash 是列式存储引擎，已经逐步成为集群拓扑的标配。

场景	配置任务	配置文件模板	拓扑说明
使用 TiCDC 进行增量同步	部署 TiCDC 拓扑架构	简单 TiCDC 配置模板 详细 TiCDC 配置模板	在最小拓扑的基础上部署 TiCDC。TiCDC 支持多种下游 (TiDB/MySQL/MQ)。

场景	配置任务	配置文件模板	拓扑说明
使用 TiDB Binlog 进行增量同步	部署 TiDB Binlog 拓扑架构	简单 TiDB Binlog 配置模板（下游为 MySQL） 简单 TiDB Binlog 配置模板（下游为 file） 详细 TiDB Binlog 配置模板	在最小拓扑的基础上部署 TiDB Binlog。

场景	配置任务	配置文件模板	拓扑说明
使用 Spark 的 OLAP 业务	部署 TiSpark 拓扑架构	简单 TiSpark 配置模板 详细 TiSpark 配置模板	在最小拓扑的基础上部署 TiSpark 组件。TiSpark 是 PingCAP 为解决用户复杂 OLAP 需求而推出的产品。TiUP cluster 组件对 TiSpark 的支持

场景	配置任务	配置文件模板	拓扑说明
单台机器，多个实例	混合部署拓扑架构	简单混部配置模板详细混部配置模板	也适用于单机多实例需要额外增加目录、端口、资源配比、label 等配置的场景。

场景	配置任务	配置文件模板	拓扑说明
跨机房部署 TiDB 集群	跨机房部署拓扑架构	跨机房配置模板	以典型的两地三中心架构为例，介绍跨机房部署架构，以及需要注意的关键设置。

注意：

- 对于需要全局生效的参数，请在配置文件中 `server_configs` 的对应组件下配置。

- 对于需要某个节点生效的参数，请在具体节点的 config 中配置。
- 配置的层次结构使用 . 表示。如：log.slow-threshold。更多格式参考 TiUP 配置参数模版。
- 如果需要指定在目标机创建的用户组名，可以参考这个例子。

更多参数说明，请参考 TiDB config.toml.example、TiKV config.toml.example、PD config.toml.example 和 TiFlash 配置参数。

2.4.1.4 第 4 步：执行部署命令

注意：

通过 TiUP 进行集群部署可以使用密钥或者交互密码方式来进行安全认证：

- 如果是密钥方式，可以通过 -i 或者 –identity_file 来指定密钥的路径。
- 如果是密码方式，可以通过 -p 进入密码交互窗口。
- 如果已经配置免密登录目标机，则不需填写认证。

一般情况下 TiUP 会在目标机器上创建 topology.yaml 中约定的用户和组，以下情况例外：

- topology.yaml 中设置的用户名在目标机器上已存在。
- 在命令行上使用了参数 –skip-create-user 明确指定跳过创建用户的步骤。

执行部署命令前，先使用 check 及 check --apply 命令检查和自动修复集群存在的潜在风险：

1. 检查集群存在的潜在风险：

```
tiup cluster check ./topology.yaml --user root [-p] [-i /home/root/.ssh/gcp_rsa]
```

2. 自动修复集群存在的潜在风险：

```
tiup cluster check ./topology.yaml --apply --user root [-p] [-i /home/root/.ssh/gcp_rsa]
```

3. 部署 TiDB 集群：

```
tiup cluster deploy tidb-test v5.4.0 ./topology.yaml --user root [-p] [-i /home/root/.ssh/
    ↪ gcp_rsa]
```

以上部署示例中：

- tidb-test 为部署的集群名称。

- v5.4.0 为部署的集群版本，可以通过执行 `tiup list tidb` 来查看 TiUP 支持的最新可用版本。
- 初始化配置文件为 `topology.yaml`。
- `--user root` 表示通过 root 用户登录到目标主机完成集群部署，该用户需要有 ssh 到目标机器的权限，并且在目标机器有 sudo 权限。也可以用其他有 ssh 和 sudo 权限的用户完成部署。
- [-i] 及 [-p] 为可选项，如果已经配置免密登录目标机，则不需填写。否则选择其一即可，[-i] 为可登录到目标机的 root 用户（或 –user 指定的其他用户）的私钥，也可使用 [-p] 交互式输入该用户的密码。

预期日志结尾输出 `Deployed cluster \`tidb-test\` successfully` 关键词，表示部署成功。

2.4.1.5 第 5 步：查看 TiUP 管理的集群情况

```
tiup cluster list
```

TiUP 支持管理多个 TiDB 集群，该命令会输出当前通过 TiUP cluster 管理的所有集群信息，包括集群名称、部署用户、版本、密钥信息等。

2.4.1.6 第 6 步：检查部署的 TiDB 集群情况

例如，执行如下命令检查 `tidb-test` 集群情况：

```
tiup cluster display tidb-test
```

预期输出包括 `tidb-test` 集群中实例 ID、角色、主机、监听端口和状态（由于还未启动，所以状态为 Down/inactive）、目录信息。

2.4.1.7 第 7 步：启动集群

安全启动是 TiUP cluster 从 v1.9.0 起引入的一种新的启动方式，采用该方式启动数据库可以提高数据库安全性。推荐使用安全启动。

安全启动后，TiUP 会自动生成 TiDB root 用户的密码，并在命令行界面返回密码。

> 注意：
>
> - 使用安全启动方式后，不能通过无密码的 root 用户登录数据库，你需要记录命令行返回的密码进行后续操作。
> - 该自动生成的密码只会返回一次，如果没有记录或者忘记该密码，请参照忘记 root 密码修改密码。

方式一：安全启动

```
tiup cluster start tidb-test --init
```

预期结果如下，表示启动成功。

```
Started cluster `tidb-test` successfully.
The root password of TiDB database has been changed.
The new password is: 'y_+3Hwp=*AWz8971s6'.
Copy and record it to somewhere safe, it is only displayed once, and will not be stored.
The generated password can NOT be got again in future.
```

方式二：普通启动

```
tiup cluster start tidb-test
```

预期结果输出 `Started cluster `tidb-test` successfully`，表示启动成功。使用普通启动方式后，可通过无密码的 root 用户登录数据库。

2.4.1.8 第 8 步：验证集群运行状态

```
tiup cluster display tidb-test
```

预期结果输出：各节点 Status 状态信息为 Up 说明集群状态正常。

2.4.1.9 探索更多

如果你已同时部署了 TiFlash，接下来可参阅以下文档：

- 使用 TiFlash
- TiFlash 集群运维
- TiFlash 报警规则与处理方法
- TiFlash 常见问题

如果你已同时部署了TiCDC，接下来可参阅以下文档：

- TiCDC 任务管理
- TiCDC 常见问题

2.4.2 在 Kubernetes 上部署 TiDB

你可以使用 TiDB Operator 在 Kubernetes 上部署 TiDB。TiDB Operator 是 Kubernetes 上的 TiDB 集群自动运维系统，提供包括部署、升级、扩缩容、备份恢复、配置变更的 TiDB 全生命周期管理。借助 TiDB Operator，TiDB 可以无缝运行在公有云或私有部署的 Kubernetes 集群上。

TiDB Operator 的文档目前独立于 TiDB 文档。要查看如何在 Kubernetes 上部署 TiDB 的详细步骤，请参阅对应版本的 TiDB Operator 文档：

- TiDB in Kubernetes 用户文档

2.5 验证集群运行状态

在部署完一套 TiDB 集群后，需要检查集群是否正常运行。本文介绍如何通过 TiUP 命令、TiDB Dashboard 和 Grafana 检查集群状态，以及如何登录 TiDB 数据库执行简单的 SQL 操作。

2.5.1 通过 TiUP 检查集群状态

检查集群状态的命令是 `tiup cluster display <cluster-name>`，例如：

```
tiup cluster display tidb-test
```

预期结果输出：各节点 Status 状态信息为 Up 说明集群状态正常。

2.5.2 通过 TiDB Dashboard 和 Grafana 检查集群状态

本节介绍如何通过TiDB Dashboard 和 Grafana 检查集群状态。

2.5.2.1 查看 TiDB Dashboard 检查 TiDB 集群状态

1. 通过 `{pd-ip}:{pd-port}/dashboard` 登录 TiDB Dashboard，登录用户和口令为 TiDB 数据库 `root` 用户和口令。如果你修改过数据库的 `root` 密码，则以修改后的密码为准，默认密码为空。

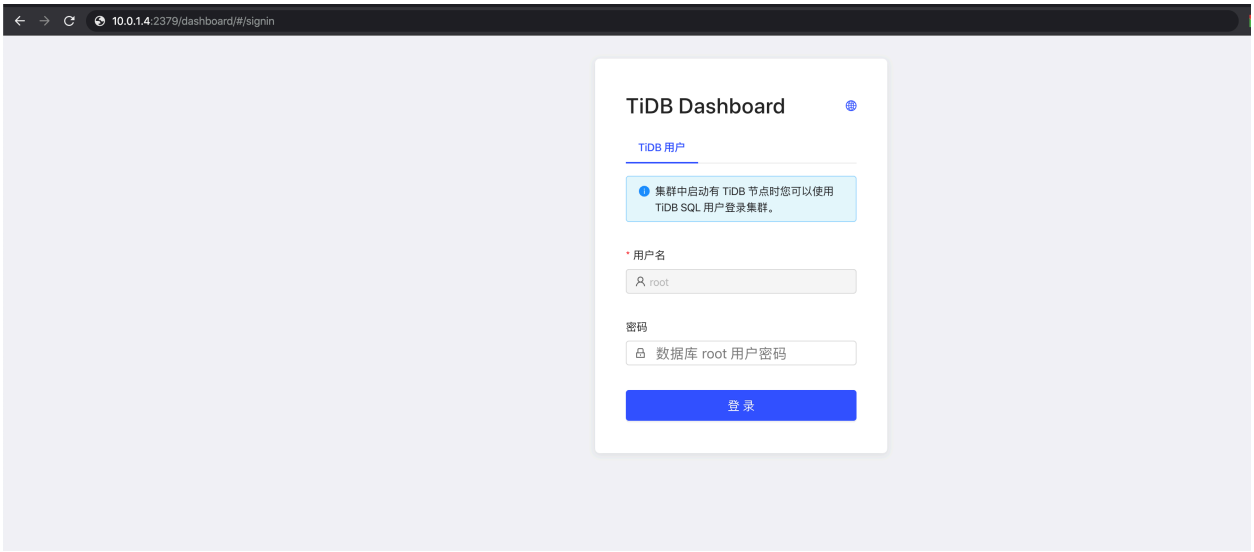


图 2: TiDB-Dashboard

2. 主页面显示 TiDB 集群中节点信息

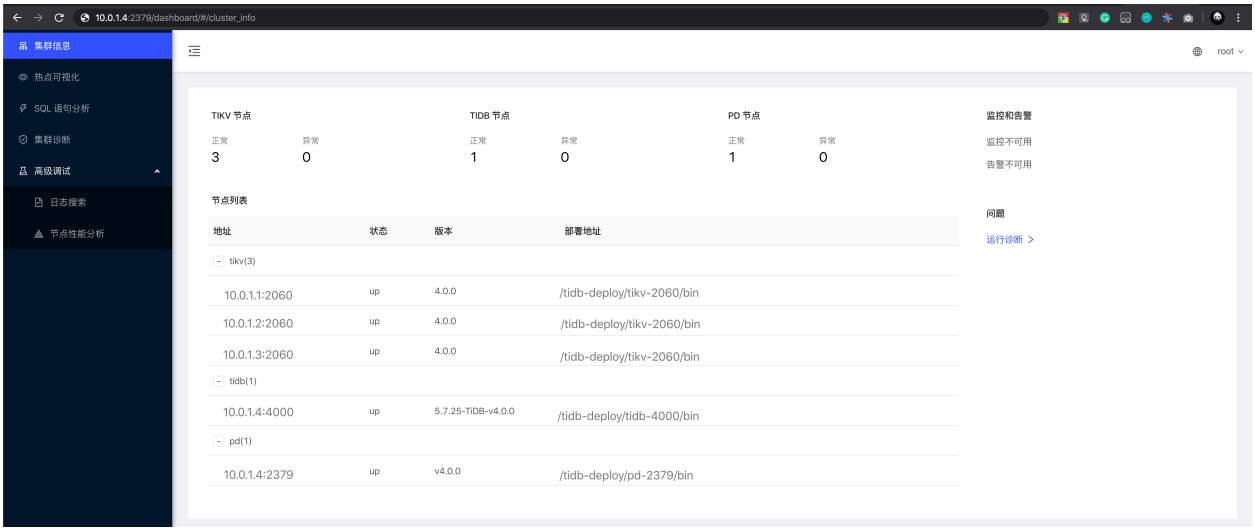


图 3: TiDB-Dashboard-status

2.5.2.2 查看 Grafana 监控 Overview 页面检查 TiDB 集群状态

- 通过 `{Grafana-ip}:3000` 登录 Grafana 监控，默认用户名及密码为 `admin/admin`。
- 点击 Overview 监控页面检查 TiDB 端口和负载监控信息。

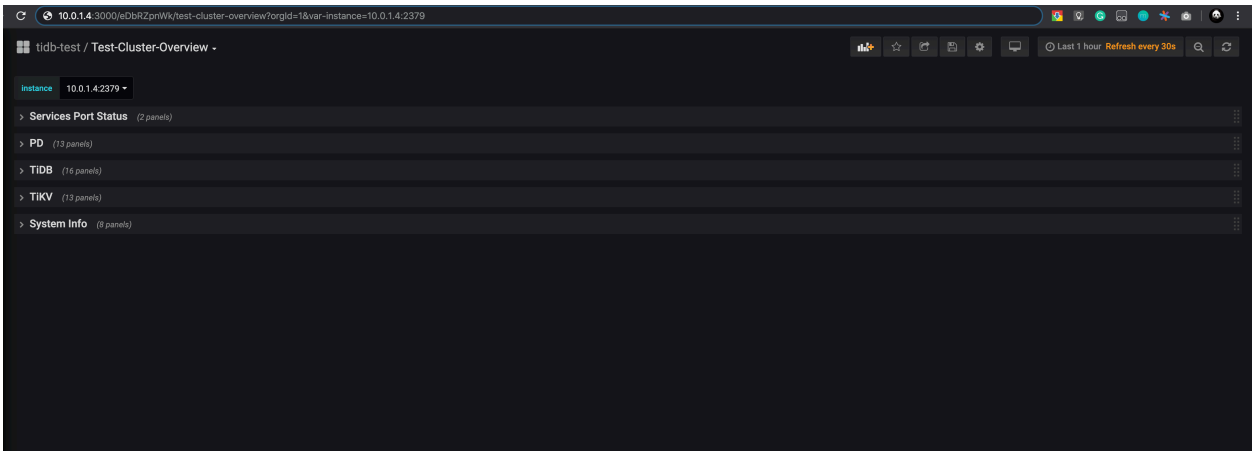


图 4: Grafana-overview

2.5.3 登录数据库执行简单 DML/DDL 操作和查询 SQL 语句

> **注意：**
>
> 登录数据库前，你需要安装 MySQL 客户端。

执行以下命令登录数据库：

```
mysql -u root -h ${tidb_server_host_IP_address} -P 4000
```

其中，${tidb_server_host_IP_address} 是在初始化集群拓扑文件时为 tidb_servers 配置的 IP 地址之一，例如 10.0.1.7。

输出下列信息表示登录成功：

```
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 3
Server version: 5.7.25-TiDB-v5.0.0 TiDB Server (Apache License 2.0) Community Edition, MySQL 5.7
    ↪ compatible

Copyright (c) 2000, 2015, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

2.5.3.1 数据库操作

- 检查 TiDB 版本

```
select tidb_version()\G
```

预期结果输出：

```
*************************** 1. row ***************************
tidb_version(): Release Version: v5.0.0
Edition: Community
Git Commit Hash: 689a6b6439ae7835947fcaccf329a3fc303986cb
Git Branch: HEAD
UTC Build Time: 2020-05-28 11:09:45
GoVersion: go1.13.4
Race Enabled: false
TiKV Min Version: v3.0.0-60965b006877ca7234adaced7890d7b029ed1306
Check Table Before Drop: false
1 row in set (0.00 sec)
```

- 创建 PingCAP database

```
create database pingcap;
```

```
Query OK, 0 rows affected (0.10 sec)
```

```
use pingcap;
```

预期输出

```
Database changed
```

- 创建 tab_tidb 表

```
CREATE TABLE `tab_tidb` (
`id` int(11) NOT NULL AUTO_INCREMENT,
`name` varchar(20) NOT NULL DEFAULT '',
`age` int(11) NOT NULL DEFAULT 0,
`version` varchar(20) NOT NULL DEFAULT '',
PRIMARY KEY (`id`),
KEY `idx_age` (`age`));
```

预期输出

```
Query OK, 0 rows affected (0.11 sec)
```

- 插入数据

```
insert into `tab_tidb` values (1,'TiDB',5,'TiDB-v5.0.0');
```

预期输出

```
Query OK, 1 row affected (0.03 sec)
```

- 查看 tab_tidb 结果

```
select * from tab_tidb;
```

预期输出

```
+----+------+-----+-------------+
| id | name | age | version     |
+----+------+-----+-------------+
|  1 | TiDB |   5 | TiDB-v5.0.0 |
+----+------+-----+-------------+
1 row in set (0.00 sec)
```

- 查看 TiKV store 状态、store_id、存储情况以及启动时间

```
select STORE_ID,ADDRESS,STORE_STATE,STORE_STATE_NAME,CAPACITY,AVAILABLE,UPTIME from
    ↪ INFORMATION_SCHEMA.TIKV_STORE_STATUS;
```

预期输出

```
+--
    ↪ --------+--------------------+-------------+-----------------+----------+-----------+--------------
    ↪
| STORE_ID | ADDRESS            | STORE_STATE | STORE_STATE_NAME | CAPACITY | AVAILABLE |
    ↪ UPTIME              |
+--
    ↪ --------+--------------------+-------------+-----------------+----------+-----------+--------------
    ↪
|        1 | 10.0.1.1:20160 |           0 | Up               | 49.98GiB | 46.3GiB   | 5
    ↪ h21m52.474864026s |
|        4 | 10.0.1.2:20160 |           0 | Up               | 49.98GiB | 46.32GiB  | 5
    ↪ h21m52.522669177s |
|        5 | 10.0.1.3:20160 |           0 | Up               | 49.98GiB | 45.44GiB  | 5
    ↪ h21m52.713660541s |
+--
    ↪ --------+--------------------+-------------+-----------------+----------+-----------+--------------
    ↪
3 rows in set (0.00 sec)
```

- 退出

```
exit
```

预期输出

```
Bye
```

2.6 测试集群性能

2.6.1 如何用 Sysbench 测试 TiDB

建议使用 Sysbench 1.0 或之后的更新版本，可在 Sysbench Release 1.0.14 页面下载。

2.6.1.1 测试方案

2.6.1.1.1 TiDB 配置

升高日志级别，可以减少打印日志数量，对 TiDB 的性能有积极影响。开启 TiDB 配置中的 `prepared plan cache`，以减少优化执行计划的开销。具体在 TiUP 配置文件中加入：

```
server_configs:
  tidb:
    log.level: "error"
    prepared-plan-cache.enabled: true
```

2.6.1.1.2 TiKV 配置

升高 TiKV 的日志级别同样有利于提高性能表现。

TiKV 集群存在多个 Column Family，包括 Default CF、Write CF 和 LockCF，主要用于存储不同类型的数据。对于 Sysbench 测试，需要关注 Default CF 和 Write CF，导入数据的 Column Family 在 TiDB 集群中的比例是固定的。这个比例是：

Default CF : Write CF = 4 : 1

在 TiKV 中需要根据机器内存大小配置 RocksDB 的 block cache，以充分利用内存。以 40 GB 内存的虚拟机部署一个 TiKV 为例，其 block cache 建议配置如下：

```
server_configs:
  tikv:
    log-level: "error"
    rocksdb.defaultcf.block-cache-size: "24GB"
    rocksdb.writecf.block-cache-size: "6GB"
```

还可以使用共享 block cache 的方式进行设置：

```
server_configs:
  tikv:
    storage.block-cache.capacity: "30GB"
```

更详细的 TiKV 参数调优请参考TiKV 内存参数性能调优。

2.6.1.2 测试过程

> **注意：**
>
> 此文档中的测试并没有使用如 HAproxy 等负载均衡工具。在 TiDB 单一节点上进行 Sysbench 测试，并把结果相加。负载均衡工具和不同版本参数也会影响性能表现。

2.6.1.2.1 Sysbench 配置

以下为 Sysbench 配置文件样例：

```
mysql-host={TIDB_HOST}
mysql-port=4000
mysql-user=root
mysql-password=password
mysql-db=sbtest
time=600
threads={8, 16, 32, 64, 128, 256}
report-interval=10
db-driver=mysql
```

可根据实际需求调整其参数，其中 TIDB_HOST 为 TiDB server 的 IP 地址（配置文件中不能写多个地址），threads 为测试中的并发连接数，可在“8, 16, 32, 64, 128, 256”中调整，导入数据时，建议设置 threads = 8 或者 16。调整后，将该文件保存为名为 config 的文件。

配置文件参考示例如下：

```
mysql-host=172.16.30.33
mysql-port=4000
mysql-user=root
mysql-password=password
mysql-db=sbtest
time=600
threads=16
report-interval=10
db-driver=mysql
```

2.6.1.2.2 数据导入

> 注意：
>
> 如果 TiDB 启用了乐观事务模型（默认为悲观锁模式），当发现并发冲突时，会回滚事务。将 tidb_disable_txn_auto_retry 设置为 off 会开启事务冲突后的自动重试机制，可以尽可能避免事务冲突报错导致 Sysbench 程序退出的问题。

在数据导入前，需要对 TiDB 进行简单设置。在 MySQL 客户端中执行如下命令：

```
set global tidb_disable_txn_auto_retry = off;
```

然后退出客户端。

重新启动 MySQL 客户端执行以下 SQL 语句，创建数据库 sbtest：

```
create database sbtest;
```

调整 Sysbench 脚本创建索引的顺序。Sysbench 按照“建表 -> 插入数据 -> 创建索引”的顺序导入数据。对于 TiDB 而言，该方式会花费更多的导入时间。你可以通过调整顺序来加速数据的导入。

假设使用的 Sysbench 版本为 1.0.14，可以通过以下两种方式来修改：

1. 直接下载为 TiDB 修改好的 oltp_common.lua 文件，覆盖 /usr/share/sysbench/oltp_common.lua 文件。
2. 将 /usr/share/sysbench/oltp_common.lua 的第 235 行到第 240 行移动到第 198 行以后。

> 注意：
>
> 此操作为可选操作，仅节约了数据导入时间。

命令行输入以下命令，开始导入数据，config 文件为上一步中配置的文件：

```
sysbench --config-file=config oltp_point_select --tables=32 --table-size=10000000 prepare
```

2.6.1.2.3 数据预热与统计信息收集

数据预热可将磁盘中的数据载入内存的 block cache 中，预热后的数据对系统整体的性能有较大的改善，建议在每次重启集群后进行一次数据预热。

Sysbench 1.0.14 没有提供数据预热的功能，因此需要手动进行数据预热。如果使用更新的 Sysbench 版本，可以使用自带的预热功能。

以 Sysbench 中某张表 sbtest7 为例，执行如下 SQL 语句进行数据预热：

```
SELECT COUNT(pad) FROM sbtest7 USE INDEX (k_7);
```

统计信息收集有助于优化器选择更为准确的执行计划，可以通过 analyze 命令来收集表 sbtest 的统计信息，每个表都需要统计。

```
ANALYZE TABLE sbtest7;
```

2.6.1.2.4 Point select 测试命令

```
sysbench --config-file=config oltp_point_select --tables=32 --table-size=10000000 run
```

2.6.1.2.5 Update index 测试命令

```
sysbench --config-file=config oltp_update_index --tables=32 --table-size=10000000 run
```

2.6.1.2.6 Read-only 测试命令

```
sysbench --config-file=config oltp_read_only --tables=32 --table-size=10000000 run
```

2.6.1.3 常见问题

2.6.1.3.1 在高并发压力下，TiDB、TiKV 的配置都合理，为什么整体性能还是偏低？

这种情况可能与使用了 proxy 有关。可以尝试直接对单个 TiDB 加压，将求和后的结果与使用 proxy 的情况进行对比。

以 HAproxy 为例。nbproc 参数可以增加其最大启动的进程数，较新版本的 HAproxy 还支持 nbthread 和 cpu-map 等。这些都可以减少对其性能的不利影响。

2.6.1.3.2 在高并发压力下，为什么 TiKV 的 CPU 利用率依然很低？

TiKV 虽然整体 CPU 偏低，但部分模块的 CPU 可能已经达到了很高的利用率。

TiKV 的其他模块，如 storage readpool、 coprocessor 和 gRPC 的最大并发度限制是可以通过 TiKV 的配置文件进行调整的。

通过 Grafana 的 TiKV Thread CPU 监控面板可以观察到其实际使用率。如出现多线程模块瓶颈，可以通过增加该模块并发度进行调整。

2.6.1.3.3 在高并发压力下，TiKV 也未达到 CPU 使用瓶颈，为什么 TiDB 的 CPU 利用率依然很低？

在某些高端设备上，使用的是 NUMA 架构的 CPU，跨 CPU 访问远端内存将极大降低性能。TiDB 默认将使用服务器所有 CPU，goroutine 的调度不可避免地会出现跨 CPU 内存访问。

因此，建议在 NUMA 架构服务器上，部署 n 个 TiDB（n = NUMA CPU 的个数），同时将 TiDB 的 `max-procs` 变量的值设置为与 NUMA CPU 的核数相同。

2.6.2 如何对 TiDB 进行 TPC-C 测试

本文介绍如何对 TiDB 进行 TPC-C 测试。

TPC-C 是一个对 OLTP（联机交易处理）系统进行测试的规范，使用一个商品销售模型对 OLTP 系统进行测试，其中包含五类事务：

- NewOrder – 新订单的生成
- Payment – 订单付款
- OrderStatus – 最近订单查询
- Delivery – 配送
- StockLevel – 库存缺货状态分析

在测试开始前，TPC-C Benchmark 规定了数据库的初始状态，也就是数据库中数据生成的规则，其中 ITEM 表中固定包含 10 万种商品，仓库的数量可进行调整，假设 WAREHOUSE 表中有 W 条记录，那么：

- STOCK 表中应有 W * 10 万条记录（每个仓库对应 10 万种商品的库存数据）
- DISTRICT 表中应有 W * 10 条记录（每个仓库为 10 个地区提供服务）
- CUSTOMER 表中应有 W * 10 * 3000 条记录（每个地区有 3000 个客户）
- HISTORY 表中应有 W * 10 * 3000 条记录（每个客户一条交易历史）
- ORDER 表中应有 W * 10 * 3000 条记录（每个地区 3000 个订单），并且最后生成的 900 个订单被添加到 NEW-ORDER 表中，每个订单随机生成 5 ~ 15 条 ORDER-LINE 记录。

我们将以 1000 WAREHOUSE 为例进行测试。

TPC-C 使用 tpmC 值 (Transactions per Minute) 来衡量系统最大有效吞吐量 (MQTh, Max Qualified Throughput)，其中 Transactions 以 NewOrder Transaction 为准，即最终衡量单位为每分钟处理的新订单数。

本文使用 go-tpc 作为 TPC-C 测试实现，可以通过TiUP 命令下载测试程序：

```
tiup install bench
```

关于 TiUP Bench 组件的详细用法可参考TiUP Bench。

2.6.2.1 导入数据

导入数据通常是整个 TPC-C 测试中最耗时，也是最容易出问题的阶段。

在 shell 中运行 TiUP 命令：

```
tiup bench tpcc -H 172.16.5.140 -P 4000 -D tpcc --warehouses 1000 prepare
```

基于不同的机器配置，这个过程可能会持续几个小时。如果是小型集群，可以使用较小的 WAREHOUSE 值进行测试。

数据导入完成后，可以通过命令 `tiup bench tpcc -H 172.16.5.140 -P 4000 -D tpcc --warehouses 4 check` 验证数据正确性。

2.6.2.2 运行测试

运行测试的命令是：

```
tiup bench tpcc -H 172.16.5.140 -P 4000 -D tpcc --warehouses 1000 run
```

运行过程中控制台上会持续打印测试结果：

```
[Current] NEW_ORDER - Takes(s): 4.6, Count: 5, TPM: 65.5, Sum(ms): 4604, Avg(ms): 920, 90th(ms):
    ↪ 1500, 99th(ms): 1500, 99.9th(ms): 1500
[Current] ORDER_STATUS - Takes(s): 1.4, Count: 1, TPM: 42.2, Sum(ms): 256, Avg(ms): 256, 90th(ms)
    ↪ : 256, 99th(ms): 256, 99.9th(ms): 256
[Current] PAYMENT - Takes(s): 6.9, Count: 5, TPM: 43.7, Sum(ms): 2208, Avg(ms): 441, 90th(ms):
    ↪ 512, 99th(ms): 512, 99.9th(ms): 512
[Current] STOCK_LEVEL - Takes(s): 4.4, Count: 1, TPM: 13.8, Sum(ms): 224, Avg(ms): 224, 90th(ms):
    ↪  256, 99th(ms): 256, 99.9th(ms): 256
...
```

运行结束后，会打印测试统计结果：

```
[Summary] DELIVERY - Takes(s): 455.2, Count: 32, TPM: 4.2, Sum(ms): 44376, Avg(ms): 1386, 90th(ms
    ↪ ): 2000, 99th(ms): 4000, 99.9th(ms): 4000
[Summary] DELIVERY_ERR - Takes(s): 455.2, Count: 1, TPM: 0.1, Sum(ms): 953, Avg(ms): 953, 90th(ms
    ↪ ): 1000, 99th(ms): 1000, 99.9th(ms): 1000
[Summary] NEW_ORDER - Takes(s): 487.8, Count: 314, TPM: 38.6, Sum(ms): 282377, Avg(ms): 899, 90th
    ↪ (ms): 1500, 99th(ms): 1500, 99.9th(ms): 1500
[Summary] ORDER_STATUS - Takes(s): 484.6, Count: 29, TPM: 3.6, Sum(ms): 8423, Avg(ms): 290, 90th(
    ↪ ms): 512, 99th(ms): 1500, 99.9th(ms): 1500
[Summary] PAYMENT - Takes(s): 490.1, Count: 321, TPM: 39.3, Sum(ms): 144708, Avg(ms): 450, 90th(
    ↪ ms): 512, 99th(ms): 1000, 99.9th(ms): 1500
[Summary] STOCK_LEVEL - Takes(s): 487.6, Count: 41, TPM: 5.0, Sum(ms): 9318, Avg(ms): 227, 90th(
    ↪ ms): 512, 99th(ms): 1000, 99.9th(ms): 1000
```

测试完成之后，也可以运行 `tiup bench tpcc -H 172.16.5.140 -P 4000 -D tpcc --warehouses 4 check` 进行数据正确性验证。

2.6.2.3 清理测试数据

```
tiup bench tpcc -H 172.16.5.140 -P 4000 -D tpcc --warehouses 4 cleanup
```

3 数据迁移

3.1 数据迁移场景综述

本文档总体介绍可用于 TiDB 的数据迁移方案。数据迁移方案如下：

- 全量数据迁移。
  - 数据导入：使用 TiDB Lightning 将 Aurora Snapshot，CSV 文件或 Mydumper SQL 文件的数据全量导入到 TiDB 集群。
  - 数据导出：使用 Dumpling 将 TiDB 集群的数据全量导出为 CSV 文件或 Mydumper SQL 文件，从而更好地配合从 MySQL 数据库或 MariaDB 数据库进行数据迁移。
  - TiDB DM (Data migration) 也提供了适合小规模数据量数据库（例如小于 1 TiB）的全量数据迁移功能。
- 快速初始化 TiDB 集群：TiDB Lightning 提供的快速导入功能可以实现快速初始化 TiDB 集群的指定表的效果。请注意，使用快速初始化 TiDB 集群的功能对 TiDB 集群的影响极大，在进行初始化的过程中，TiDB 集群不支持对外访问。
- 增量数据迁移：使用 TiDB DM 从 MySQL，MariaDB 或 Aurora 同步 Binlog 到 TiDB，该功能可以极大降低业务迁移过程中停机窗口时间。
- TiDB 集群复制：TiDB 支持备份恢复功能，该功能可以实现将 TiDB 的某个快照初始化到另一个全新的 TiDB 集群。

根据迁移数据所在数据库类型、部署位置、业务数据规模大小、业务需求等因素，会有不同数据迁移选择。下面展示一些常用的数据迁移场景，方便用户依据这些线索选择到最适合自己的数据迁移方案。

3.1.1 迁移 Aurora MySQL 到 TiDB

从 Aurora 迁移数据到部署在 AWS 的 TiDB 集群，数据迁移可以分为全量迁移和增量迁移两个步骤进行。请根据你的业务需求选择相应的步骤。

- 从 Aurora 迁移数据到 TiDB

3.1.2 迁移 MySQL 到 TiDB

如果你没有使用 Cloud storage (S3) 服务，而且网络联通和延迟情况良好，那么从 MySQL 迁移数据到 TiDB 时可以参照下面的方案。

- 从小数据量 MySQL 迁移数据到 TiDB

如果你对数据迁移速度有要求，或者数据规模特别大（例如大于 1 TiB），并且禁止 TiDB 集群在迁移期间有其他业务写入，那么你可以先使用 Lightning 进行快速导入，然后根据业务需要选择是否使用 DM 进行增量数据 (Binlog) 同步。

- 从大数据量 MySQL 迁移数据到 TiDB

3.1.3 分库分表 MySQL 合并迁移到 TiDB

如果你的业务使用了基于 MySQL 分库的方案来存储数据，业务数据从 MySQL 迁移到 TiDB 后，合并这些分表数据到一张合并，那么你可以使用 DM 进行分表合并迁移。

- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB

如果分表数据总规模特别大（例如大于 1 TiB），并且禁止 TiDB 集群在迁移期间有其他业务写入，那么你可以使用 Lightning 对分表数据进行快速合并导入，然后根据业务需要选择是否使用 DM 进行增量数据 (Binlog) 的分表同步。

- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

3.1.4 从文件迁移数据到 TiDB

- 从 CSV 文件迁移数据到 TiDB
- 从 SQL 文件迁移数据到 TiDB

3.1.5 复杂迁移场景

DM 在实时同步过程中，多个已有特性可以使得同步过程更加灵活，适应各类业务需求：

- 上游使用 pt/gh-ost 工具的持续同步场景
- 下游存在更多列的迁移场景
- 如何过滤 binlog 事件
- 如何通过 SQL 表达式过滤 binlog

3.2 数据迁移工具概览

TiDB 提供了丰富的数据迁移相关的工具，用于全量迁移、增量迁移、备份恢复、数据同步等多种场景。

本文介绍了使用这些工具的场景、优势和相关限制等信息。请根据你的需求选择合适的工具。

下图显示了各迁移工具的使用场景。

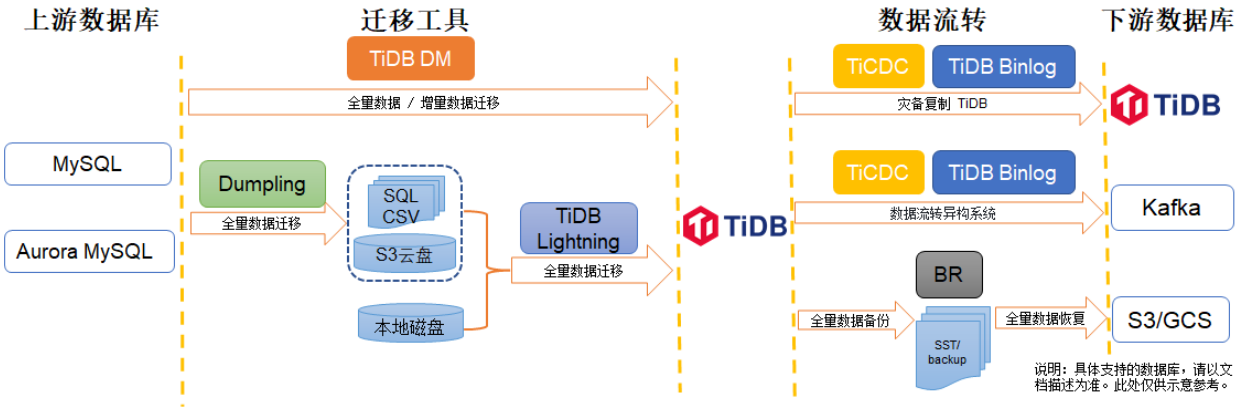


图 5: TiDB Migration Tools

下表介绍了迁移工具的使用场景、支持的上下游等信息。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
TiDB DM	用于将数据从与 MySQL 协议兼容的数据库迁移到 TiDB。	MySQL，MariaDB，Aurora，MySQL	TiDB	一体化的数据迁移任务管理工具，支持全量迁移和增量同步；支持对表与操作进行过滤；支持分库分表的合并迁移。	建议用于 1TB 以内的存量数据迁移。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
Dumpling	用于将数据从 MySQL/TiDB 进行全量导出。	MySQL，TiDB	SQL，CSV	支持全新的 table-filter，筛选数据更加方便；支持导出到 Amazon S3 云盘	如果导出后计划往非 TiDB 的数据库恢复，建议使用 Dumpling；如果是往另一个 TiDB 恢复，建议使用 BR。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
TiDB Lightning	用于将数据全量导入到 TiDB。	Dumpling 输出的文件；CSV 文件；从本地盘或 Amazon S3 云盘读取数据。	TiDB	支持迅速导入大量新数据，实现快速初始化 TiDB 集群的指定表；支持断点续传；支持数据过滤。	如果使用 Local-backend 进行数据导入，TiDB Lightning 运行后，TiDB 集群将无法正常对外提供服务。如果你不希望 TiDB 集群的对外服务受到影响，可以参考 TiDB Lightning TiDB-backend 中的硬件需求与部署方式进行数据导

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
Backup & Restore (BR)	用于对大数据量的 TiDB 集群进行数据备份和恢复。	TiDB	SST；backup.meta 文件；backup.lock 文件	适用于向另一个 TiDB 恢复数据。支持数据冷备份到外部存储，可以用于灾备恢复。	BR 恢复到 TiCDC / Drainer 的上游集群时，恢复数据无法由 TiCDC / Drainer 同步到下游。BR 只支持在 new_collations_enabled_on_first_bootstrap 开关值相同的集群之间进行操作。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
TiCDC	通过拉取 TiKV 变更日志实现的 TiDB 增量数据同步工具，具有将数据还原到与上游任意 TSO 一致状态的能力，支持其他系统订阅数据变更。	TiDB	TiDB，MySQL，Apache Pulsar，Kafka，Confluent	提供开放数据协议 (TiCDC Open Protocol)。	TiCDC 只能同步至少存在一个有效索引的表。暂不支持以下场景：暂不支持单独使用 RawKV 的 TiKV 集群。暂不支持在 TiDB 中创建 SEQUENCE 的 DDL 操作和 SEQUENCE 函数。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
TiDB Binlog	用于 TiDB 集群间的增量数据同步，如将其中一个 TiDB 集群作为另一个 TiDB 集群的从集群。	TiDB	TiDB，MySQL，Kafka，增量备份文件	支持实时备份和恢复。备份 TiDB 集群数据，同时可以用于 TiDB 集群故障时恢复。	与部分 TiDB 版本不兼容，不能一起使用。

名称	使用场景	上游（或输入源文件）	下游（或输出文件）	主要优势	使用限制
sync-diff-inspector	用于校验 MySQL/TiDB 中两份数据的一致性。	TiDB，MySQL	TiDB，MySQL	提供了修复数据的功能，适用于修复少量不一致的数据。	对于 MySQL 和 TiDB 之间的数据同步不支持在线校验。不支持 JSON、BIT、BINARY、BLOB 等类型的数据。

3.2.1 使用 TiUP 快速安装

从 TiDB 4.0 开始，TiUP 作为软件包管理器，帮助你轻松管理 TiDB 生态系统中的不同集群组件。现在你可以只用一个 TiUP 命令行来管理任何组件。

3.2.1.1 第 1 步：安装 TiUP

```
curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
```

重新声明全局环境变量

```
source ~/.bash_profile
```

3.2.1.2 第 2 步：安装组件

你可以通过以下命令查看所有可用组件:

```
tiup list
```

以下输出为所有可用组件：

```
Available components:
Name            Owner    Description
----            -----    -----------
bench           pingcap  Benchmark database with different workloads
br              pingcap  TiDB/TiKV cluster backup restore tool
cdc             pingcap  CDC is a change data capture tool for TiDB
client          pingcap  Client to connect playground
cluster         pingcap  Deploy a TiDB cluster for production
ctl             pingcap  TiDB controller suite
dm              pingcap  Data Migration Platform manager
dmctl           pingcap  dmctl component of Data Migration Platform
errdoc          pingcap  Document about TiDB errors
pd-recover      pingcap  PD Recover is a disaster recovery tool of PD, used to recover the PD
    ↪ cluster which cannot start or provide services normally
playground      pingcap  Bootstrap a local TiDB cluster for fun
tidb            pingcap  TiDB is an open source distributed HTAP database compatible with the
    ↪ MySQL protocol
tidb-lightning  pingcap  TiDB Lightning is a tool used for fast full import of large amounts of
    ↪ data into a TiDB cluster
tiup            pingcap  TiUP is a command-line component management tool that can help to
    ↪ download and install TiDB platform components to the local system
```

选择所需要的组件进行安装

```
tiup install dumpling tidb-lightning
```

> Note:
>
> 如果需要安装特定版本，可以使用 `tiup install <component>[:version]` 命令.

3.2.1.3 第 3 步：更新 TiUP 及组件 (可选)

建议先查看新版本的更新日志及兼容性说明

```
tiup update --self && tiup update dm
```

3.2.2 探索更多

- 离线方式安装 TiUP
- 以二进制包形式安装各工具

3.3 数据迁移场景

3.3.1 从 Amazon Aurora 迁移数据到 TiDB

本文档介绍如何从 Amazon Aurora 迁移数据到 TiDB，迁移过程采用 DB snapshot，可以节约大量的空间和时间成本。整个迁移包含两个过程：

- 使用 Lightning 导入全量数据到 TiDB
- 使用 DM 持续增量同步到 TiDB（可选）

3.3.1.1 前提条件

- 安装 Dumpling 和 Lightning。
- 获取 Dumpling 所需上游数据库权限。
- 获取 Lightning 所需下游数据库权限。

3.3.1.2 导入全量数据到 TiDB

3.3.1.2.1 第 1 步：导出 Aurora 快照文件到 Amazon S3

1. 在 Aurora 上，执行以下命令，查询并记录当前 binlog 位置：

```
mysql> SHOW MASTER STATUS;
```

你将得到类似以下的输出，请记录 binlog 名称和位置，供后续步骤使用：

```
+------------------+----------+--------------+------------------+-------------------+
| File             | Position | Binlog_Do_DB | Binlog_Ignore_DB | Executed_Gtid_Set |
+------------------+----------+--------------+------------------+-------------------+
| mysql-bin.000002 |    52806 |              |                  |                   |
+------------------+----------+--------------+------------------+-------------------+
1 row in set (0.012 sec)
```

2. 导出 Aurora 快照文件。具体方式请参考 Aurora 的官方文档：Exporting DB snapshot data to Amazon S3.

请注意，上述两步的时间间隔建议不要超过 5 分钟，否则记录的 binlog 位置过旧可能导致增量同步时产生数据冲突。

完成上述两步后，你需要准备好以下信息：

- 创建快照点时，Aurora binlog 的名称及位置。
- 快照文件的 S3 路径，以及具有访问权限的 SecretKey 和 AccessKey。

3.3.1.2.2 第 2 步：导出 schema

因为 Aurora 生成的快照文件并不包含建表语句文件，所以你需要使用 Dumpling 自行导出 schema 并使用 Lightning 在下游创建 schema。你也可以跳过此步骤，并以手动方式在下游自行创建 schema。

运行以下命令，建议使用 `--filter` 参数仅导出所需表的 schema：

```
tiup dumpling --host ${host} --port 3306 --user root --password ${password} --filter 'my_db1.
    ↪ table[12]' --no-data --output 's3://my-bucket/schema-backup?region=us-west-2' --filter "
    ↪ mydb.*"
```

命令中所用参数描述如下。如需更多信息可参考Dumpling overview。

参数	说明
-u 或 –user	Aurora MySQL 数据库的用户
-p 或 –password	MySQL 数据库的用户密码
-P 或 –port	MySQL 数据库的端口
-h 或 –host	MySQL 数据库的 IP 地址
-t 或 –thread	导出的线程数
-o 或 –output	存储导出文件的目录，支持本地文件路径或外部存储 URL 格式
-r 或 –row	单个文件的最大行数
-F	指定单个文件的最大大小，单位为 MiB, 建议值 256 MiB
-B 或 –database	导出指定数据库
-T 或 –tables-list	导出指定数据表
-d 或 –no-data	不导出数据，仅导出 schema
-f 或 –filter	导出能匹配模式的表，不可用 -T 一起使用，语法可参考table filter

3.3.1.2.3 第 3 步：编写 Lightning 配置文件

根据以下内容创建`tidb-lightning.toml` 配置文件：

```
vim tidb-lightning.toml
```

```
[tidb]

### 目标 TiDB 集群信息.
host = ${host}                # 例如：172.16.32.1
port = ${port}                # 例如：4000
user = "${user_name}"         # 例如："root"
password = "${password}"      # 例如："rootroot"
status-port = ${status-port}  # 表结构信息在从 TiDB 的“状态端口”获取例如：10080
pd-addr = "${ip}:${port}"     # 集群 PD 的地址，lightning 通过 PD 获取部分信息，例如
    ↪ 172.16.31.3:2379。当 backend = "local" 时 status-port 和 pd-addr 必须正确填写，
    ↪ 否则导入将出现异常。

[tikv-importer]
### "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
```

```
### "tidb": TB 级以下数据量也可以采用`tidb`后端模式，下游 TiDB 可正常提供服务。
    ↪ 关于后端模式更多信息请参阅：https://docs.pingcap.com/tidb/stable/tidb-lightning-backends
backend = "local"

### 设置排序的键值对的临时存放地址，目标路径需要是一个空目录，至少需要数据源最大单表的空间，
    ↪ 建议与 `data-source-dir` 不同磁盘目录并使用闪存介质，独占 IO 会获得更好的导入性能。
sorted-kv-dir = "${path}"

[mydumper]
### 快照文件的地址
data-source-dir = "${s3_path}"  # eg: s3://my-bucket/sql-backup?region=us-west-2

[[mydumper.files]]
### 解析 parquet 文件所需的表达式
pattern = '(?i)^(?:[^/]*/)*([a-z0-9_]+)\.([a-z0-9_]+)/(?:[^/]*/)*(?:[a-z0-9\-_.]+\.(parquet))$'
schema = '$1'
table = '$2'
type = '$3'
```

如果需要在 TiDB 开启 TLS ，请参考TiDB Lightning Configuration。

3.3.1.2.4 第 4 步：导入全量数据到 TiDB

1. 使用 Lightning 在下游 TiDB 建表:

   ```
   tiup tidb-lightning -config tidb-lightning.toml -d 's3://my-bucket/schema-backup?region=us-
       ↪ west-2' -no-schema=false
   ```

2. 运行 tidb-lightning。如果直接在命令行中启动程序，可能会因为 SIGHUP 信号而退出，建议配合nohup
   ↪ 或screen等工具，如：

   将有权限访问该 Amazon S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 Lightning 节点。同时还支持从 ~/.aws/credentials 读取凭证文件。

   ```
   export AWS_ACCESS_KEY_ID=${access_key}
   export AWS_SECRET_ACCESS_KEY=${secret_key}
   nohup tiup tidb-lightning -config tidb-lightning.toml -no-schema=true > nohup.out 2>&1 &
   ```

3. 导入开始后，可以采用以下任意方式查看进度：

   - 通过 grep 日志关键字 progress 查看进度，默认 5 分钟更新一次。
   - 通过监控面板查看进度，请参考TiDB Lightning 监控。
   - 通过 Web 页面查看进度，请参考Web 界面。

4. 导入完毕后，TiDB Lightning 会自动退出。查看日志的最后 5 行中会有 the whole procedure completed，则表示导入成功。

> **注意：**
>
> 无论导入成功与否，最后一行都会显示 `tidb lightning exit`。它只是表示 TiDB Lightning 正常退出，不代表任务完成。

如果导入过程中遇到问题，请参见TiDB Lightning 常见问题。

3.3.1.3 持续增量同步数据到 TiDB（可选）

3.3.1.3.1 前提条件

- 安装 DM 集群
- 获取 DM 所需上下游数据库权限

3.3.1.3.2 第 1 步：创建数据源

1. 新建`source1.yaml`文件, 写入以下内容：

```
# 唯一命名，不可重复。
source-id: "mysql-01"

# DM-worker 是否使用全局事务标识符（GTID）拉取 binlog。使用前提是上游 MySQL 已开启 GTID 模式
    ↪ 。若上游存在主从自动切换，则必须使用 GTID 模式。
enable-gtid: false

from:
  host: "${host}"         # 例如：172.16.10.81
  user: "root"
  password: "${password}" # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用
  port: 3306
```

2. 在终端中执行下面的命令，使用 `tiup dmctl` 将数据源配置加载到 DM 集群中:

```
tiup dmctl --master-addr ${advertise-addr} operate-source create source1.yaml
```

该命令中的参数描述如下：

参数	描述
`--master-addr`	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
`operate-source create`	向 DM 集群加载数据源

3.3.1.3.3 第 2 步：创建迁移任务

新建 task1.yaml 文件，写入以下内容：

```
### 任务名，多个同时运行的任务不能重名。
name: "test"
### 任务模式，可设为
### full：只进行全量数据迁移
### incremental： binlog 实时同步
### all： 全量 + binlog 迁移
task-mode: "incremental"
### 下游 TiDB 配置信息。
target-database:
  host: "${host}"                  # 例如：172.16.10.83
  port: 4000
  user: "root"
  password: "${password}"          # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用

### 黑白名单全局配置，各实例通过配置项名引用。
block-allow-list:                    # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  listA:                             # 名称
    do-tables:                       # 需要迁移的上游表的白名单。
    - db-name: "test_db"             # 需要迁移的表的库名。
      tbl-name: "test_table"         # 需要迁移的表的名称。

### 配置数据源
mysql-instances:
  - source-id: "mysql-01"              # 数据源 ID，即 source1.yaml 中的 source-id
    block-allow-list: "listA"          # 引入上面黑白名单配置。
###       syncer-config-name: "global"    # 引用上面的 syncers 增量数据配置。
    meta:                              # task-mode 为 incremental 且下游数据库的 checkpoint
        ↪ 不存在时 binlog 迁移开始的位置；如果 checkpoint 存在，则以 checkpoint 为准。
      binlog-name: "mysql-bin.000004"   # “Step 1. 导出 Aurora 快照文件到 Amazon S3”
          ↪ 中记录的日志位置，当上游存在主从切换时，必须使用 gtid。
      binlog-pos: 109227
      # binlog-gtid: "09bec856-ba95-11ea-850a-58f2b4af5188:1-9"

   # 【可选配置】 如果增量数据迁移需要重复迁移已经在全量数据迁移中完成迁移的数据，则需要开启 safe
       ↪  mode 避免增量数据迁移报错。
   ##  该场景多见于以下情况：全量迁移的数据不属于数据源的一个一致性快照，
       ↪ 随后从一个早于全量迁移数据之前的位置开始同步增量数据。
   # syncers:            # sync 处理单元的运行配置参数。
   #  global:           # 配置名称。
   #    safe-mode: true # 设置为 true，会将来自数据源的 INSERT 改写为 REPLACE，将 UPDATE 改写为
       ↪ DELETE 与 REPLACE，从而保证在表结构中存在主键或唯一索引的条件下迁移数据时可以重复导入
```

```
    ↪ DML。在启动或恢复增量复制任务的前 1 分钟内 TiDB DM 会自动启动 safe mode。
```

以上内容为执行迁移的最小任务配置。关于任务的更多配置项，可以参考DM 任务完整配置文件介绍

3.3.1.3.4 第 3 步：启动任务

在你启动数据迁移任务之前，建议使用 check-task 命令检查配置是否符合 DM 的配置要求，以降低后期报错的概率：

```
tiup dmctl --master-addr ${advertise-addr} check-task task.yaml
```

使用 tiup dmctl 执行以下命令启动数据迁移任务。

```
tiup dmctl --master-addr ${advertise-addr} start-task task.yaml
```

该命令中的参数描述如下：

参数	描述
--master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261

参数	描述
start-task	命令用于创建数据迁移任务

如果任务启动失败，可根据返回结果的提示进行配置变更后，再次执行上述命令，重新启动任务。遇到问题请参考故障及处理方法以及常见问题。

3.3.1.3.5 第 4 步：查看任务状态

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用 `tiup dmctl` 执行 `query-status` 命令进行查询：

```
tiup dmctl --master-addr ${advertise-addr} query-status ${task-name}
```

关于查询结果的详细解读，请参考查询状态。

3.3.1.3.6 第 5 步：监控任务与查看日志

要查看迁移任务的历史状态以及更多的内部运行指标，可参考以下步骤。

如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，则使用部署时填写的 IP 及端口进入 Grafana，选择 DM 的 dashboard 查看 DM 相关监控项。

DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息。各组件的日志目录如下：

- DM-master 日志目录：通过 DM-master 进程参数 `--log-file` 设置。如果使用 TiUP 部署 DM，则日志目录默认位于 `/dm-deploy/dm-master-8261/log/`。
- DM-worker 日志目录：通过 DM-worker 进程参数 `--log-file` 设置。如果使用 TiUP 部署 DM，则日志目录默认位于 `/dm-deploy/dm-worker-8262/log/`。

3.3.1.4 探索更多

- 暂停数据迁移任务
- 恢复数据迁移任务

- 停止数据迁移任务
- 导出和导入集群的数据源和任务配置
- 处理出错的 DDL 语句

3.3.2 从小数据量 MySQL 迁移数据到 TiDB

本文档介绍如何使用 TiDB DM （以下简称 DM ）以全量 + 增量的模式数据到 TiDB。本文所称“小数据量”通常指 TiB 级别以下。

一般而言，受到表结构索引数目等信息、硬件以及网络环境影响，迁移速率在 30 ~ 50GB/h 不等。使用 TiDB DM 迁移的流程如下图所示。

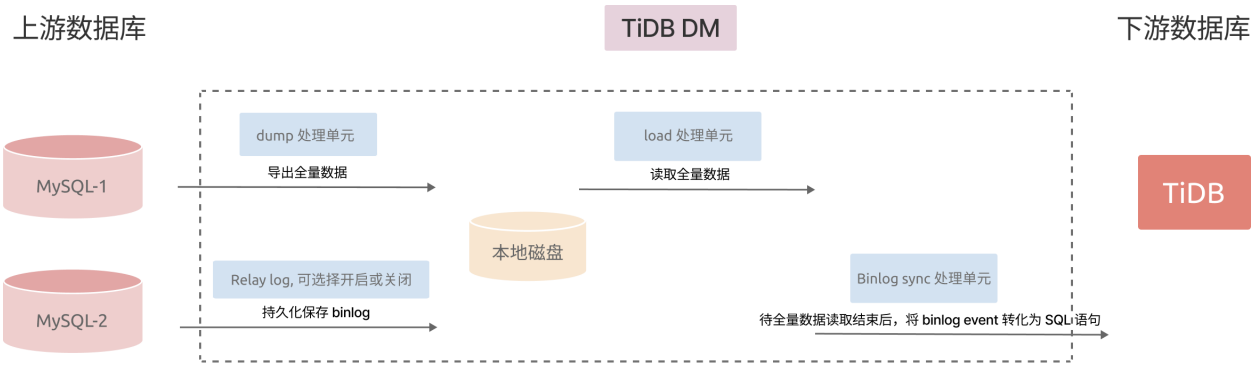


图 6: dm

3.3.2.1 前提条件

- 使用 TiUP 安装 DM 集群
- DM 所需上下游数据库权限

3.3.2.2 第 1 步：创建数据源

首先，新建 source1.yaml 文件，写入以下内容：

```
### 唯一命名，不可重复。
source-id: "mysql-01"

### DM-worker 是否使用全局事务标识符（GTID）拉取 binlog。使用前提是上游 MySQL 已开启 GTID 模式。
    ↪ 若上游存在主从自动切换，则必须使用 GTID 模式。
enable-gtid: true

from:
  host: "${host}"         # 例如：172.16.10.81
  user: "root"
  password: "${password}" # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用
```

```
  port: 3306
```

其次，在终端中执行下面的命令后，使用 `tiup dmctl` 将数据源配置加载到 DM 集群中:

```
tiup dmctl --master-addr ${advertise-addr} operate-source create source1.yaml
```

该命令中的参数描述如下：

参数	描述
`--master-addr`	`dmctl` 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
`operate-source create`	向 DM 集群加载数据源

3.3.2.3 第 2 步：创建迁移任务

新建 `task1.yaml` 文件, 写入以下内容：

```
### 任务名，多个同时运行的任务不能重名。
name: "test"
### 任务模式，可设为
### full：只进行全量数据迁移
### incremental： binlog 实时同步
### all： 全量 + binlog 迁移
task-mode: "all"
### 下游 TiDB 配置信息。
target-database:
  host: "${host}"                  # 例如：172.16.10.83
  port: 4000
  user: "root"
  password: "${password}"          # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用

### 当前数据迁移任务需要的全部上游 MySQL 实例配置。
mysql-instances:
-
  # 上游实例或者复制组 ID。
  source-id: "mysql-01"
  # 需要迁移的库名或表名的黑白名单的配置项名称，用于引用全局的黑白名单配置，全局配置见下面的 `
      ↪ block-allow-list` 的配置。
  block-allow-list: "listA"


### 黑白名单全局配置，各实例通过配置项名引用。
block-allow-list:
  listA:                           # 名称
    do-tables:                     # 需要迁移的上游表的白名单。
    - db-name: "test_db"           # 需要迁移的表的库名。
```

```
    tbl-name: "test_table"          # 需要迁移的表的名称。
```

以上内容为执行迁移的最小任务配置。关于任务的更多配置项，可以参考DM 任务完整配置文件介绍。

3.3.2.4 第 3 步：启动任务

在你启动数据迁移任务之前，建议使用 `check-task` 命令检查配置是否符合 DM 的配置要求，以避免后期报错。

```
tiup dmctl --master-addr ${advertise-addr} check-task task.yaml
```

使用 `tiup dmctl` 执行以下命令启动数据迁移任务。

```
tiup dmctl --master-addr ${advertise-addr} start-task task.yaml
```

该命令中的参数描述如下：

参数	描述
--	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
master	
-	
addr	

参数	描述
start-task	参数用于启动数据迁移任务

如果任务启动失败，可根据返回结果的提示进行配置变更后执行 start-task task.yaml 命令重新启动任务。遇到问题请参考故障及处理方法 以及常见问题。

3.3.2.5 第 4 步：查看任务状态

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用 `tiup dmctl` 执行 `query-status` 命令进行查询：

```
tiup dmctl --master-addr ${advertise-addr} query-status ${task-name}
```

关于查询结果的详细解读，请参考查询状态。

3.3.2.6 第 5 步：监控任务与查看日志（可选）

要查看迁移任务的历史状态以及更多的内部运行指标，可参考以下步骤。

如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，则使用部署时填写的 IP 及端口进入 Grafana，选择 DM 的 Dashboard 查看 DM 相关监控项。

DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息。各组件的日志目录如下：

- DM-master 日志目录：通过 DM-master 进程参数 `--log-file`设置。如果使用 TiUP 部署 DM，则日志目录默认位于 `/dm-deploy/dm-master-8261/log/`。
- DM-worker 日志目录：通过 DM-worker 进程参数 `--log-file` 设置。如果使用 TiUP 部署 DM，则日志目录默认位于 `/dm-deploy/dm-worker-8262/log/`。

3.3.2.7 探索更多

- 暂停数据迁移任务
- 恢复数据迁移任务

- 停止数据迁移任务
- 导出和导入集群的数据源和任务配置
- 处理出错的 DDL 语句

3.3.3 从大数据量 MySQL 迁移数据到 TiDB

通常数据量较低时，使用 DM 进行迁移较为简单，可直接完成全量 + 持续增量迁移工作。但当数据量较大时，DM 较低的数据导入速度 (30~50 GiB/h) 可能令整个迁移周期过长。本文所称“大数据量”通常指 TiB 级别以上。

因此，本文档介绍使用 Dumpling 和 TiDB Lightning 进行全量数据迁移，其本地导入 (local backend) 模式导入速度可达每小时 500 GiB。完成全量数据迁移后，再使用 DM 完成增量数据迁移。

3.3.3.1 前提条件

- 部署 DM 集群。
- 安装 Dumpling 和 TiDB Lightning。
- 配置 DM 所需上下游数据库权限。
- 获取 TiDB Lightning 所需下游数据库权限。
- 获取 Dumpling 所需上游数据库权限。

3.3.3.2 资源要求

操作系统：本文档示例使用的是若干新的、纯净版 CentOS 7 实例，你可以在本地虚拟化一台主机，或在供应商提供的平台上部署一台小型的云虚拟主机。TiDB Lightning 运行过程中，默认会占满 CPU，建议单独部署在一台主机上。如果条件不允许，你可以将 TiDB Lightning 和其他组件（比如 `tikv-server`）部署在同一台机器上，然后设置 `region-concurrency` 配置项的值为逻辑 CPU 数的 75%，以限制 TiDB Lightning 对 CPU 资源的使用。

内存和 CPU：因为 TiDB Lightning 对计算机资源消耗较高，建议分配 64 GB 以上的内存以及 32 核以上的 CPU，而且确保 CPU 核数和内存（GB）比为 1:2 以上，以获取最佳性能。

磁盘空间：

- Dumpling 需要足够储存整个数据源的存储空间。
- TiDB Lightning 导入期间，需要临时空间来存储排序键值对，磁盘空间需要至少能存储数据源的最大单表。
- 若全量数据量较大，可适当加长上游 binlog 保存时间，以避免增量同步时缺必要 binlog 导致重做。

说明：目前无法精确计算 Dumpling 从 MySQL 导出的数据大小，但你可以用下面 SQL 语句统计信息表的 `data_length` 字段估算数据量：

```
/* 统计所有 schema 大小，单位 MiB，注意修改 ${schema_name} */
SELECT table_schema,SUM(data_length)/1024/1024 AS data_length,SUM(index_length)/1024/1024 AS
    ↪ index_length,SUM(data_length+index_length)/1024/1024 AS SUM FROM information_schema.
    ↪ tables WHERE table_schema = "${schema_name}" GROUP BY table_schema;

/* 统计最大单表，单位 MiB，注意修改 ${schema_name} */
```

```
SELECT table_name,table_schema,SUM(data_length)/1024/1024 AS data_length,SUM(index_length)
    ↪ /1024/1024 AS index_length,SUM(data_length+index_length)/1024/1024 AS SUM from
    ↪ information_schema.tables WHERE table_schema = "${schema_name}" GROUP BY table_name,
    ↪ table_schema ORDER BY SUM DESC LIMIT 5;
```

3.3.3.2.1 目标 TiKV 集群的磁盘空间要求

目标 TiKV 集群必须有足够空间接收新导入的数据。除了标准硬件配置以外，目标 TiKV 集群的总存储空间必须大于数据源大小 × 副本数量 × 2。例如，集群默认使用 3 副本，那么总存储空间需为数据源大小的 6 倍以上。公式中的 2 倍可能难以理解，其依据是以下因素的估算空间占用：

- 索引会占据额外的空间。
- RocksDB 的空间放大效应。

3.3.3.3 第 1 步：从 MySQL 导出全量数据

1. 运行以下命令，从 MySQL 导出全量数据：

    ```
    tiup dumpling -h ${ip} -P 3306 -u root -t 16 -r 200000 -F 256MiB -B my_db1 -f 'my_db1.table
        ↪ [12]' -o 's3://my-bucket/sql-backup?region=us-west-2'
    ```

    Dumpling 默认导出数据格式为 SQL 文件，你也可以通过设置 `--filetype` 指定导出文件的类型。

    以上命令行中用到的参数描述如下。要了解更多 Dumpling 参数，请参考Dumpling 使用文档。

参数	说明
-u 或 –user	MySQL 数据库的用户
-p 或 –password	MySQL 数据库的用户密码
-P 或 –port	MySQL 数据库的端口
-h 或 –host	MySQL 数据库的 IP 地址
-t 或 –thread	导出的线程数。增加线程数会增加 Dumpling 并发度提高导出速度，但也会加大数据库内存消耗，因此不宜设
-o 或 –output	存储导出文件的目录，支持本地文件路径或外部存储 URL 格式
-r 或 –row	用于指定单个文件的最大行数，指定该参数后 Dumpling 会开启表内并发加速导出，同时减少内存使用。
-F	指定单个文件的最大大小，单位为 MiB。强烈建议使用-F参数以避免单表过大导致备份过程中断
-B 或 –database	导出指定数据库
-f 或 –filter	导出能匹配模式的表，语法可参考table-filter。

    请确保 `${data-path}` 拥有足够的空间。强烈建议使用 `-F` 参数以避免单表过大导致备份过程中断。

2. 查看在 `${data-path}` 目录下的 `metadata` 文件，这是 Dumpling 自动生成的元信息文件，请记录其中的 binlog 位置信息，这将在第 3 步增量同步的时候使用。

    ```
    SHOW MASTER STATUS:
    Log: mysql-bin.000004
    Pos: 109227
    GTID:
    ```

3.3.3.4 第 2 步：导入全量数据到 TiDB

1. 编写配置文件 tidb-lightning.toml：

```
[lightning]
# 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
# "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
# "tidb"：TB 级以下数据量也可以采用`tidb`后端模式，下游 TiDB 可正常提供服务。
    ↪ 关于后端模式更多信息请参阅：https://docs.pingcap.com/tidb/stable/tidb-lightning-
    ↪ backends
backend = "local"
# 设置排序的键值对的临时存放地址，目标路径需要是一个空目录，至少需要数据源最大单表的空间，
    ↪ 建议与 `data-source-dir` 不同磁盘目录并使用闪存介质，独占 IO 会获得更好的导入性能
sorted-kv-dir = "${sorted-kv-dir}"

[mydumper]
# 源数据目录，即第 1 步中 Dumpling 保存数据的路径。
data-source-dir = "${data-path}" # 本地或 S3 路径，例如：'s3://my-bucket/sql-backup?region=
    ↪ us-west-2'

[tidb]
# 目标集群的信息
host = ${host}                # 例如：172.16.32.1
port = ${port}                # 例如：4000
user = "${user_name}"         # 例如："root"
password = "${password}"      # 例如："rootroot"
status-port = ${status-port}  # 导入过程 Lightning 需要在从 TiDB 的“状态端口”
    ↪ 获取表结构信息，例如：10080
pd-addr = "${ip}:${port}"     # 集群 PD 的地址，Lightning 通过 PD 获取部分信息，例如
    ↪ 172.16.31.3:2379。当 backend = "local" 时 status-port 和 pd-addr 必须正确填写，
    ↪ 否则导入将出现异常。
```

关于更多 Lightning 的配置，请参考TiDB Lightning 配置参数。

2. 运行 tidb-lightning。如果直接在命令行中启动程序，可能会因为 SIGHUP 信号而退出，建议配合nohup
   ↪ 或screen等工具，如：

   若从 S3 导入，则需将有权限访问该 Amazon S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 Lightning 节点。同时还支持从 ~/.aws/credentials 读取凭证文件。

```
export AWS_ACCESS_KEY_ID=${access_key}
export AWS_SECRET_ACCESS_KEY=${secret_key}
nohup tiup tidb-lightning -config tidb-lightning.toml  > nohup.out 2>&1 &
```

3. 导入开始后，可以采用以下任意方式查看进度：

- 通过 grep 日志关键字 progress 查看进度，默认 5 分钟更新一次。
- 通过监控面板查看进度，请参考TiDB Lightning 监控。
- 通过 Web 页面查看进度，请参考Web 界面。

4. 导入完毕后，TiDB Lightning 会自动退出。查看日志的最后 5 行中会有 the whole procedure completed，则表示导入成功。

> 注意：
>
> 无论导入成功与否，最后一行都会显示 tidb lightning exit。它只是表示 TiDB Lightning 正常退出，不代表任务完成。

如果导入过程中遇到问题，请参见TiDB Lightning 常见问题。

3.3.3.5 第 3 步：使用 DM 持续复制增量数据到 TiDB

3.3.3.5.1 添加数据源

1. 新建 source1.yaml 文件，写入以下内容：

```
# 唯一命名，不可重复。
source-id: "mysql-01"

# DM-worker 是否使用全局事务标识符（GTID）拉取 binlog。使用前提是上游 MySQL 已开启 GTID 模式
    ↪ 。若上游存在主从自动切换，则必须使用 GTID 模式。
enable-gtid: true

from:
  host: "${host}"           # 例如：172.16.10.81
  user: "root"
  password: "${password}"   # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用
  port: 3306
```

2. 在终端中执行下面的命令，使用 tiup dmctl 将数据源配置加载到 DM 集群中:

```
tiup dmctl --master-addr ${advertise-addr} operate-source create source1.yaml
```

该命令中的参数描述如下：

参数	描述
--master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
operate-source create	向 DM 集群加载数据源

3.3.3.5.2 添加同步任务

1. 编辑 task.yaml，配置增量同步模式，以及每个数据源的同步起点：

```
name: task-test                     # 任务名称，需要全局唯一。
task-mode: incremental              # 任务模式，设为 "incremental" 即只进行增量数据迁移。

# 配置下游 TiDB 数据库实例访问信息
target-database:                    # 下游数据库实例配置。
  host: "${host}"                   # 例如：127.0.0.1
  port: 4000
  user: "root"
  password: "${password}"           # 推荐使用经过 dmctl 加密的密文。

#  使用黑白名单配置需要同步的表
block-allow-list:                   # 数据源数据库实例匹配的表的 block-allow-list
    ↪ 过滤规则集，如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  bw-rule-1:                        # 黑白名单配置项 ID。
    do-dbs: ["${db-name}"]          # 迁移哪些库。

# 配置数据源
mysql-instances:
  - source-id: "mysql-01"           # 数据源 ID，即 source1.yaml 中的 source-id
    block-allow-list: "bw-rule-1"   # 引入上面黑白名单配置。
    # syncer-config-name: "global"    # 引用下面的 syncers 增量数据配置。
    meta:                           # task-mode 为 incremental 且下游数据库的 checkpoint
        ↪ 不存在时 binlog 迁移开始的位置；如果 checkpoint 存在，则以 checkpoint 为准。
      # binlog-name: "mysql-bin.000004"  # 第 1 步中记录的日志位置，当上游存在主从切换时，
          ↪ 必须使用 gtid。
      # binlog-pos: 109227
      binlog-gtid: "09bec856-ba95-11ea-850a-58f2b4af5188:1-9"

# 【可选配置】 如果增量数据迁移需要重复迁移已经在全量数据迁移中完成迁移的数据，则需要开启
    ↪ safe mode 避免增量数据迁移报错。
#  该场景多见于以下情况：全量迁移的数据不属于数据源的一个一致性快照，
    ↪ 随后从一个早于全量迁移数据之前的位置开始同步增量数据。
# syncers:            # sync 处理单元的运行配置参数。
#  global:           # 配置名称。
#    safe-mode: true # 设置为 true，会将来自数据源的 INSERT 改写为 REPLACE，将 UPDATE 改写为
    ↪  DELETE 与 REPLACE，
```

```
    ↪ 从而保证在表结构中存在主键或唯一索引的条件下迁移数据时可以重复导入 DML。
    ↪ 在启动或恢复增量复制任务的前 1 分钟内 TiDB DM 会自动启动 safe mode。
```

以上内容为执行迁移的最小任务配置。关于任务的更多配置项，可以参考DM 任务完整配置文件介绍

在你启动数据迁移任务之前，建议使用check-task命令检查配置是否符合 DM 的配置要求，以降低后期报错的概率。

```
tiup dmctl --master-addr ${advertise-addr} check-task task.yaml
```

2. 使用 tiup dmctl 执行以下命令启动数据迁移任务：

```
tiup dmctl --master-addr ${advertise-addr} start-task task.yaml
```

该命令中的参数描述如下：

参数	描述
--master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}, 例如：172.16.10.71:8261
start-task	命令用于创建数据迁移任务

如果任务启动失败，可根据返回结果的提示进行配置变更，再执行上述命令重新启动任务。遇到问题请参考故障及处理方法 以及常见问题。

3.3.3.5.3 查看任务状态

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用 tiup dmctl 执行 query-status 命令进行查询：

```
tiup dmctl --master-addr ${advertise-addr} query-status ${task-name}
```

关于查询结果的详细解读，请参考查询状态。

3.3.3.5.4 监控任务与查看日志

要查看迁移任务的历史状态以及更多的内部运行指标，可参考以下步骤。

如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，则使用部署时填写的 IP 及端口进入 Grafana，选择 DM 的 dashboard 查看 DM 相关监控项。

DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息。各组件的日志目录如下：

- DM-master 日志目录：通过 DM-master 进程参数 --log-file 设置。如果使用 TiUP 部署 DM，则日志目录默认位于 /dm-deploy/dm-master-8261/log/。
- DM-worker 日志目录：通过 DM-worker 进程参数 --log-file 设置。如果使用 TiUP 部署 DM，则日志目录默认位于 /dm-deploy/dm-worker-8262/log/。

3.3.3.6 探索更多

- 暂停数据迁移任务
- 恢复数据迁移任务
- 停止数据迁移任务
- 导出和导入集群的数据源和任务配置
- 处理出错的 DDL 语句

3.3.4 从小数据量分库分表 MySQL 合并迁移数据到 TiDB

如果你想把上游多个 MySQL 数据库实例合并迁移到下游的同一个 TiDB 数据库中，且数据量较小，你可以使用 DM 工具进行分库分表的合并迁移。本文所称“小数据量”通常指 TiB 级别以下。本文举例介绍了合并迁移的操作步骤、注意事项、故障排查等。本文档适用于:

- TiB 级以内的分库分表数据合并迁移
- 基于 MySQL binlog 的增量、持续分库分表合并迁移

若要迁移分表总和 1 TiB 以上的数据，则 DM 工具耗时较长，可参考从大数据量分库分表 MySQL 合并迁移数据到 TiDB。

本文以一个简单的场景为例，示例中的两个数据源 MySQL 实例的分库和分表数据迁移至下游 TiDB 集群。示意图如下。

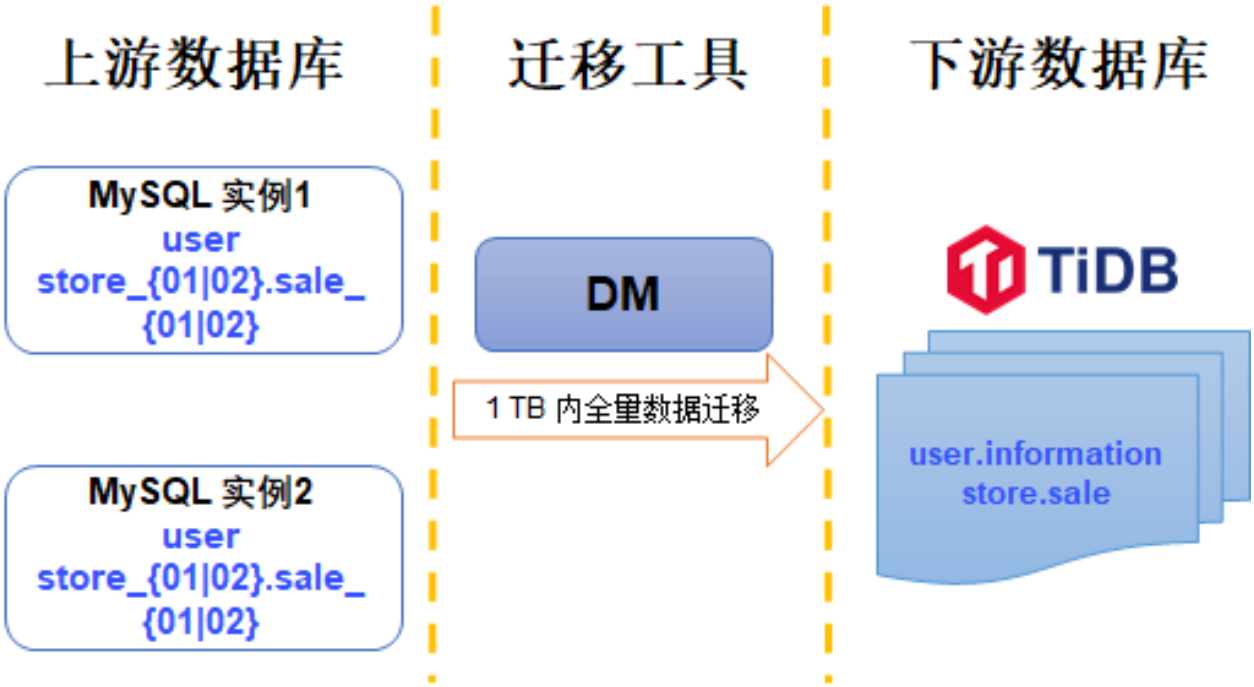


图 7: migrate-01

数据源 MySQL 实例 1 和实例 2 均使用以下表结构，计划将 store_01 和 store_02 中 sale 开头的表合并导入下游 store.sale 表

Schema	Tables
store_01	sale_01, sale_02
store_02	sale_01, sale_02

迁移目标库的结构如下：

Schema	Tables
store	sale

3.3.4.1 前提条件

- 使用 TiUP 安装 DM 集群
- DM 所需上下游数据库权限

3.3.4.2 分表数据冲突检查

迁移中如果涉及合库合表，来自多张分表的数据可能引发主键或唯一索引的数据冲突。因此在迁移之前，需要检查各分表数据的业务特点。详情请参考跨分表数据在主键或唯一索引冲突处理

在本示例中：`sale_01` 和 `sale_02` 具有相同的表结构如下：

```
CREATE TABLE `sale_01` (
  `id` bigint(20) NOT NULL AUTO_INCREMENT,
  `sid` bigint(20) NOT NULL,
  `pid` bigint(20) NOT NULL,
  `comment` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `sid` (`sid`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

其中`id`列为主键，`sid`列为分片键，具有全局唯一性。`id`列具有自增属性，多个分表范围重复会引发数据冲突。`sid`可以保证全局满足唯一索引，因此可以按照参考去掉自增主键的主键属性中介绍的操作绕过`id`列。在下游创建`sale`表时移除`id`列的唯一键属性

```
CREATE TABLE `sale` (
  `id` bigint(20) NOT NULL,
  `sid` bigint(20) NOT NULL,
  `pid` bigint(20) NOT NULL,
  `comment` varchar(255) DEFAULT NULL,
  INDEX (`id`),
  UNIQUE KEY `sid` (`sid`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

3.3.4.3 第 1 步：创建数据源

新建source1.yaml文件, 写入以下内容：

```
### 唯一命名，不可重复。
source-id: "mysql-01"

### DM-worker 是否使用全局事务标识符（GTID）拉取 binlog。使用前提是上游 MySQL 已开启 GTID 模式。
    ↪ 若上游存在主从自动切换，则必须使用 GTID 模式。
enable-gtid: true

from:
  host: "${host}" # 例如：172.16.10.81
  user: "root"
  password: "${password}" # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用
  port: 3306
```

在终端中执行下面的命令，使用tiup dmctl将数据源配置加载到 DM 集群中:

```
tiup dmctl --master-addr ${advertise-addr} operate-source create source1.yaml
```

该命令中的参数描述如下：| 参数 | 描述 | |- |-| |–master-addr |dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261| |operate-source create| 向 DM 集群加载数据源 |

重复以上操作直至所有数据源均添加完成。

3.3.4.4 第 2 步：创建迁移任务

新建task1.yaml文件, 写入以下内容：

```
name: "shard_merge"
### 任务模式，可设为
### full：只进行全量数据迁移
### incremental： binlog 实时同步
### all： 全量 + binlog 迁移
task-mode: all
### 分库分表合并任务则需要配置 shard-mode。默认使用悲观协调模式 "pessimistic"，
    ↪ 在深入了解乐观协调模式的原理和使用限制后，也可以设置为乐观协调模式 "optimistic"
### 详细信息可参考：https://docs.pingcap.com/zh/tidb/dev/feature-shard-merge/
shard-mode: "pessimistic"
meta-schema: "dm_meta"                          # 将在下游数据库创建 schema 用于存放元数据
ignore-checking-items: ["auto_increment_ID"]    # 本示例中上游存在自增主键，
    ↪ 因此需要忽略掉该检查项

target-database:
  host: "${host}"                               # 例如：192.168.0.1
  port: 4000
```

```
  user: "root"
  password: "${password}"                       # 支持但不推荐使用明文密码，建议使用 dmctl
      ↪ encrypt 对明文密码进行加密后使用

mysql-instances:
  -
    source-id: "mysql-01"                                   # 数据源 ID，即 source1.yaml 中的
        ↪ source-id
    route-rules: ["sale-route-rule"]                        # 应用于该数据源的 table route 规则
    filter-rules: ["store-filter-rule", "sale-filter-rule"] # 应用于该数据源的 binlog event
        ↪ filter 规则
    block-allow-list:  "log-bak-ignored"                    # 应用于该数据源的 Block & Allow
        ↪ Lists 规则
  -
    source-id: "mysql-02"
    route-rules: ["sale-route-rule"]
    filter-rules: ["store-filter-rule", "sale-filter-rule"]
    block-allow-list:  "log-bak-ignored"

### 分表合并配置
routes:
  sale-route-rule:
    schema-pattern: "store_*"
    table-pattern: "sale_*"
    target-schema: "store"
    target-table:  "sale"

### 过滤部分 DDL 事件
filters:
  sale-filter-rule:
    schema-pattern: "store_*"
    table-pattern: "sale_*"
    events: ["truncate table", "drop table", "delete"]
    action: Ignore
  store-filter-rule:
    schema-pattern: "store_*"
    events: ["drop database"]
    action: Ignore

### 黑白名单
block-allow-list:
  log-bak-ignored:
    do-dbs: ["store_*"]
```

以上内容为执行迁移的最小任务配置。关于任务的更多配置项，可以参考DM 任务完整配置文件介绍

若想了解配置文件中`routes`，`filters`等更多用法，请参考：

- Table routing
- Block & Allow Table Lists
- 如何过滤 binlog 事件
- 如何通过 SQL 表达式过滤 DML

3.3.4.5 第 3 步：启动任务

在你启动数据迁移任务之前，建议使用`check-task`命令检查配置是否符合 DM 的配置要求，以降低后期报错的概率。

```
tiup dmctl --master-addr ${advertise-addr} check-task task.yaml
```

使用 `tiup dmctl` 执行以下命令启动数据迁移任务。

```
tiup dmctl --master-addr ${advertise-addr} start-task task.yaml
```

该命令中的参数描述如下：

参数	描述
–master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
start-task	命令用于创建数据迁移任务

如果任务启动失败，可根据返回结果的提示进行配置变更后执行 start-task task.yaml 命令重新启动任务。遇到问题请参考故障及处理方法 以及常见问题

3.3.4.6 第 4 步：查看任务状态

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用`tiup dmctl`执行`query-status`命令进行查询：

```
tiup dmctl --master-addr ${advertise-addr} query-status ${task-name}
```

关于查询结果的详细解读，请参考查询状态

3.3.4.7 第 5 步：监控任务与查看日志 (可选)

你可以通过 Grafana 或者日志查看迁移任务的历史状态以及各种内部运行指标。

- 通过 Grafana 查看

  如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，则使用部署时填写的 IP 及端口进入 Grafana，选择 DM 的 dashboard 查看 DM 相关监控项。

- 通过日志查看

  DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息，其中包含迁移任务的相关信息。各组件的日志目录如下：

- DM-master 日志目录：通过 DM-master 进程参数`--log-file`设置。如果使用 TiUP 部署 DM，则日志目录默认位于`/dm-deploy/dm-master-8261/log/`。
- DM-worker 日志目录：通过 DM-worker 进程参数`--log-file`设置。如果使用 TiUP 部署 DM，则日志目录默认位于`/dm-deploy/dm-worker-8262/log/`。

3.3.4.8 探索更多

- 分库分表合并中的悲观/乐观模式
- 分表合并数据迁移最佳实践
- 故障及处理方法
- 性能问题及处理方法
- 常见问题

3.3.5 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

如果分表数据总规模特别大（例如大于 1 TiB），并且允许 TiDB 集群在迁移期间无其他业务写入，那么你可以使用 TiDB Lightning 对分表数据进行快速合并导入，然后根据业务需要选择是否使用 TiDB DM 进行增量数据的分表同步。本文所称“大数据量”通常指 TiB 级别以上。本文档举例介绍了导入数据的操作步骤。

如果分库分表合并迁移在 1 TiB 以内，请参考从小数据量分库分表 MySQL 合并迁移数据到 TiDB，支持全量和增量且更为简单。

使用 TiDB Lightning 快速合并导入的原理如下图所示。

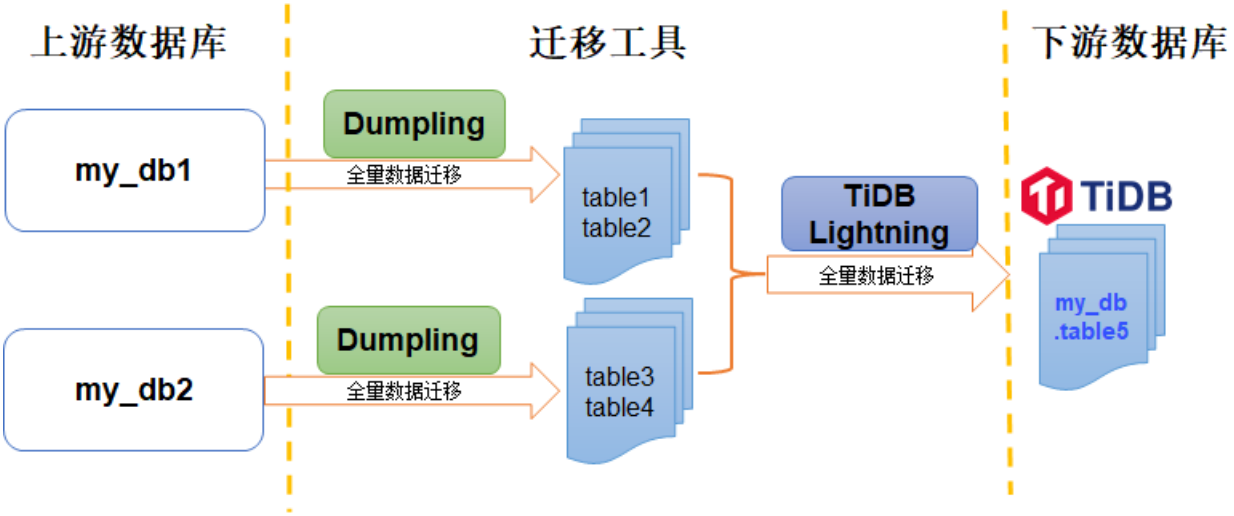


图 8: 使用 Dumpling 和 TiDB Lightning 合并导入分表数据

在这个示例中，假设有两个数据库 my_db1 和 my_db2，使用 Dumpling 分别从 my_db1 中导出 table1 和 table2 两个表，从 my_db2 中导出 table3 和 table4 两个表，然后再用 TiDB Lightning 把导出的 4 个表合并导入到下游 TiDB 中的同一个库 my_db 的同一个表格 table5 中。

本文将以三个步骤演示导入流程：

1. 使用 Dumpling 导出全量数据备份。在本文档示例中，分别从 2 个源数据库中各导出 2 个表：
    - 从实例 1 MySQL 的 my_db1 导出 table1、 table2
    - 从实例 2 MySQL 的 my_db2 导出 table3、 table4
2. 启动 Lightning 执行导入 TiDB 中的 mydb.table5
3. 使用 DM 进行增量数据迁移（可选）

3.3.5.1 前提条件

- 使用 TiUP 安装 DM 集群
- 使用 TiUP 安装 Dumpling 和 Lightning
- Dumpling 所需上游数据库权限
- TiDB Lightning 所需下游数据库权限
- TiDB Lightning 下游数据库所需空间
- DM 所需上下游数据库权限

3.3.5.1.1 分表数据冲突检查

迁移中如果涉及合库合表，来自多张分表的数据可能引发主键或唯一索引的数据冲突。因此在迁移之前，需要检查各分表数据的业务特点。详情请参考跨分表数据在主键或唯一索引冲突处理，这里做简要描述：

假设 table1~4 具有相同的表结构如下：

```
CREATE TABLE `table1` (
  `id` bigint(20) NOT NULL AUTO_INCREMENT,
  `sid` bigint(20) NOT NULL,
  `pid` bigint(20) NOT NULL,
  `comment` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`id`),
  UNIQUE KEY `sid` (`sid`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

其中id列为主键，具有自增属性，多个分表范围重复会引发数据冲突。sid列为分片键，可以保证全局满足唯一索引。因此可以移除下游table5表id列的唯一键属性

```
CREATE TABLE `table5` (
  `id` bigint(20) NOT NULL,
  `sid` bigint(20) NOT NULL,
  `pid` bigint(20) NOT NULL,
  `comment` varchar(255) DEFAULT NULL,
  INDEX (`id`),
  UNIQUE KEY `sid` (`sid`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

3.3.5.2 第 1 步：用 Dumpling 导出全量数据备份

如果需要导出的多个分表属于同一个上游 MySQL 实例，建议直接使用 Dumpling 的 `-f` 参数一次导出多个分表的结果。如果多个分表分布在不同的 MySQL 实例，可以使用 Dumpling 分两次导出，并将两次导出的结果放置在相同的父目录下即可。下面的例子中同时用到了上述两种方式，然后将导出的数据存放在同一父目录下。

首先使用 Dumpling 从 my_db1 中导出表 table1 和 table2，如下：

```
tiup dumpling -h ${ip} -P 3306 -u root -t 16 -r 200000 -F 256MB -B my_db1 -f 'my_db1.table[12]' -
    ↪ o ${data-path}/my_db1
```

以上命令行中用到的参数描述如下。要了解更多 Dumpling 参数，请参考Dumpling 使用文档。

参数	描述
-u 或 –user	MySQL 数据库的用户
-p 或 –password	MySQL 数据库的用户密码
-P 或 –port	MySQL 数据库的端口
-h 或 –host	MySQL 数据库的 IP 地址
-t 或 –thread	导出的线程数。增加线程数会增加 Dumpling 并发度提高导出速度，但也会加大数据库内存消耗，因此不宜设
-o 或 –output	存储导出文件的目录，支持本地文件路径或外部存储 URL 格式
-r 或 –row	用于指定单个文件的最大行数，指定该参数后 Dumpling 会开启表内并发加速导出，同时减少内存使用。
-F	指定单个文件的最大大小，单位为 MiB。强烈建议使用 `-F`参数以避免单表过大导致备份过程中断
-B 或 –database	导出指定数据库
-f 或 –filter	导出能匹配模式的表，语法可参考table-filter。

然后使用同样的方式从 my_db2 中导出表 table3 和 table4。注意路径在相同`${data-path}`下的不同子目录`my_db2`。

```
tiup dumpling -h ${ip} -P 3306 -u root -t 16 -r 200000 -F 256MB -B my_db2 -f 'my_db2.table[34]' -
    ↪ o ${data-path}/my_db2
```

这样所需的全量备份数据就全部导出到了 `${data-path}` 目录中。将所有源数据表格存储在一个目录中，是为了后续方便用 TiDB Lightning 导入。

第 3 步增量同步的时候所需的起始位点信息，在`${data-path}`目录下,`my_db1`和`my_db2`的`metadata`文件中，这是 Dumpling 自动生成的元信息文件，请记录其中的 binlog 位置信息。

3.3.5.3 第 2 步：启动 TiDB Lightning 进行导入

在启动 TiDB Lightning 进行迁移之前，建议先了解如何处理检查点，然后根据需要选择合适的方式进行迁移。

3.3.5.3.1 断点续传

大量数据导入一般耗时数小时甚至数天，长时间运行的进程会有一定机率发生非正常中断。如果每次重启都从头开始，之前已成功导入的数据就会前功尽弃。为此，TiDB Lightning 提供了断点续传的功能，即使 TiDB Lightning 崩溃，在重启时仍然从断点开始继续工作。

若 TiDB Lightning 因不可恢复的错误而退出，例如数据出错，在重启时不会使用断点，而是直接报错离开。为保证已导入的数据安全，必须先解决掉这些错误才能继续。你可以使用`tidb-lightning-ctl` 命令控制导入出错后的行为。该命令的选项有：

- –checkpoint-error-destroy：出现错误后，让失败的表从头开始整个导入过程。
- –checkpoint-error-ignore：如果导入表曾经出错，该命令会清除出错状态，如同错误没有发生过一样。
- –checkpoint-remove：无论是否有出错，把表的断点清除。

关于断点续传的更多信息，请参考TiDB Lightning 断点续传。

3.3.5.3.2 执行导入操作

启动 tidb-lightning 的步骤如下：

1. 编写配置文件tidb-lightning.toml。

```
[lightning]
# 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
# "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
# "tidb"：TB 级以下数据量也可以采用`tidb`后端模式，下游 TiDB 可正常提供服务。
    ↪ 关于后端模式更多信息请参阅：https://docs.pingcap.com/tidb/stable/tidb-lightning-
    ↪ backends
backend = "local"
# 设置排序的键值对的临时存放地址，目标路径需要是一个空目录，至少需要数据源最大单表的空间，
    ↪ 建议与 `data-source-dir` 不同磁盘目录，独占 IO 会获得更好的导入性能
sorted-kv-dir = "${sorted-kv-dir}"

# 设置分库分表合并规则，将 my_db1 中的 table1、table2 两个表,以及 my_db2 中的 table3、table4
    ↪  两个表，共计 2 个数据库中的 4 个表都导入到目的数据库 my_db 中的 table5 表中。
[[routes]]
schema-pattern = "my_db1"
table-pattern = "table[1-2]"
target-schema = "my_db"
target-table = "table5"

[[routes]]
schema-pattern = "my_db2"
table-pattern = "table[3-4]"
target-schema = "my_db"
target-table = "table5"

# 目标集群的信息，示例仅供参考。请把 IP 地址等信息替换成真实的信息。
[tidb]
# 目标集群的信息
host = ${host}              # 例如：172.16.32.1
```

```
port = ${port}              # 例如：4000
user = "${user_name}"       # 例如："root"
password = "${password}"    # 例如："rootroot"
status-port = ${status-port} # 导入过程 Lightning 需要在从 TiDB 的“状态端口”获取表结构信息
    ↪ ，例如：10080
# PD 集群的地址，Lightning 通过 PD 获取部分信息。
pd-addr = "${ip}:${port}"   # 例如 172.16.31.3:2379。当 backend = "local" 时 status-port 和
    ↪ pd-addr 必须正确填写，否则导入将出现异常。
```

2. 运行 tidb-lightning。如果直接在命令行中启动程序，可能会因为 SIGHUP 信号而退出，建议配合nohup ↪ 或screen等工具，如：

   若从 S3 导入，则需将有权限访问该 Amazon S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 Lightning 节点。同时还支持从 ~/.aws/credentials 读取凭证文件。

```
export AWS_ACCESS_KEY_ID=${access_key}
export AWS_SECRET_ACCESS_KEY=${secret_key}
nohup tiup tidb-lightning -config tidb-lightning.toml  > nohup.out 2>&1 &
```

3. 导入开始后，可以采用以下任意方式查看进度：

- 通过 grep 日志关键字 progress 查看进度，默认 5 分钟更新一次。
- 通过监控面板查看进度，请参考TiDB Lightning 监控。
- 通过 Web 页面查看进度，请参考Web 界面。

4. 导入完毕后，TiDB Lightning 会自动退出。查看日志的最后 5 行中会有 the whole procedure completed，则表示导入成功。

> 注意：
>
> 无论导入成功与否，最后一行都会显示 tidb lightning exit。它只是表示 TiDB Lightning 正常退出，不代表任务完成。

如果导入过程中遇到问题，请参见TiDB Lightning 常见问题。

3.3.5.4 第 3 步：使用 DM 持续复制增量数据到 TiDB (可选)

基于 binlog 从指定位置同步数据库到 TiDB，可以使用 DM 来执行增量复制

3.3.5.4.1 添加数据源

新建source1.yaml文件, 写入以下内容：

```
### 唯一命名，不可重复。
source-id: "mysql-01"

### DM-worker 是否使用全局事务标识符 (GTID) 拉取 binlog。使用前提是上游 MySQL 已开启 GTID 模式。
    ↪ 若上游存在主从自动切换，则必须使用 GTID 模式。
enable-gtid: true

from:
  host: "${host}"           # 例如：172.16.10.81
  user: "root"
  password: "${password}"   # 支持但不推荐使用明文密码，建议使用 dmctl encrypt
      ↪ 对明文密码进行加密后使用
  port: 3306
```

在终端中执行下面的命令，使用tiup dmctl将数据源配置加载到 DM 集群中:

```
tiup dmctl --master-addr ${advertise-addr} operate-source create source1.yaml
```

该命令中的参数描述如下：

参数	描述
–master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}
operate-source create	向 DM 集群加载数据源

重复以上步骤直至所有 MySQL 实例被加入 DM。

3.3.5.4.2 添加同步任务

编辑task.yaml，配置增量同步模式，以及每个数据源的同步起点：

```
name: task-test               # 任务名称，需要全局唯一。
task-mode: incremental        # 任务模式，设为 "incremental" 即只进行增量数据迁移。
### 分库分表合并任务则需要配置 shard-mode。默认使用悲观协调模式 "pessimistic"，
    ↪ 在深入了解乐观协调模式的原理和使用限制后，也可以设置为乐观协调模式 "optimistic"
### 详细信息可参考：https://docs.pingcap.com/zh/tidb/dev/feature-shard-merge/
shard-mode: "pessimistic"

#### 配置下游 TiDB 数据库实例访问信息
target-database:              # 下游数据库实例配置。
  host: "${host}"             # 例如：127.0.0.1
  port: 4000
  user: "root"
  password: "${password}"     # 推荐使用经过 dmctl 加密的密文。
```

```
####  使用黑白名单配置需要同步的表
block-allow-list:             # 数据源数据库实例匹配的表的 block-allow-list 过滤规则集，如果 DM
    ↪ 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  bw-rule-1:                  # 黑白名单配置项 ID。
    do-dbs: ["my_db1"]        # 迁移哪些库。这里将实例1的 my_db1 和 实例2的 my_db2 分别配置为两条
        ↪  rule。以示例如何避免实例1的 my_db2 被同步。
  bw-rule-2:
    do-dbs: ["my_db2"]

routes:                               # 上游和下游表之间的路由 table routing 规则集
  route-rule-1:                       # 配置名称。将 my_db1 中的 table1 和 table2 合并导入下游
      ↪ my_db.table5
    schema-pattern: "my_db1"          # 库名匹配规则，支持通配符 "*" 和 "?"
    table-pattern: "table[1-2]"       # 表名匹配规则，支持通配符 "*" 和 "?"
    target-schema: "my_db"            # 目标库名称
    target-table: "table5"            # 目标表名称
  route-rule-2:                       # 配置名称。将 my_db2 中的 table3 和 table4 合并导入下游
      ↪ my_db.table5
    schema-pattern: "my_db2"
    table-pattern: "table[3-4]"
    target-schema: "my_db"
    target-table: "table5"

#### 配置数据源，以两个数据源为例
mysql-instances:
  - source-id: "mysql-01"             # 数据源 ID，即 source1.yaml 中的 source-id
    block-allow-list: "bw-rule-1"     # 引入上面黑白名单配置。同步实例1的 my_db1
    route-rules: ["route-rule-1"]     # 引入上面表合并配置。
###       syncer-config-name: "global"  # 引用后面的 syncers 增量数据配置。
    meta:                             # task-mode 为 incremental 且下游数据库的 checkpoint
        ↪ 不存在时 binlog 迁移开始的位置；如果 checkpoint 存在，则以 checkpoint 为准。
      binlog-name: "${binlog-name}"   # 第 1 步中 ${data-path}/my_db1/metadata 记录的日志位置，
          ↪ 当上游存在主从切换时，必须使用 gtid。
      binlog-pos: ${binlog-position}
      # binlog-gtid:                  " 例如：09bec856-ba95-11ea-850a-58f2b4af5188:1-9"
  - source-id: "mysql-02"             # 数据源 ID，即 source1.yaml 中的 source-id
    block-allow-list: "bw-rule-2"     # 引入上面黑白名单配置。实例2的 my_db2
    route-rules: ["route-rule-2"]     # 引入上面表合并配置。

###       syncer-config-name: "global"  # 引用后面的 syncers 增量数据配置。
    meta:                             # task-mode 为 incremental 且下游数据库的 checkpoint
        ↪ 不存在时 binlog 迁移开始的位置；如果 checkpoint 存在，则以 checkpoint 为准。
      # binlog-name: "${binlog-name}"   # 第 1 步中 ${data-path}/my_db2/metadata 记录的日志位置，
          ↪ 当上游存在主从切换时，必须使用 gtid。
```

```
      # binlog-pos: ${binlog-position}
      binlog-gtid: "09bec856-ba95-11ea-850a-58f2b4af5188:1-9"

#### 【可选配置】 如果增量数据迁移需要重复迁移已经在全量数据迁移中完成迁移的数据，则需要开启 safe
    ↪  mode 避免增量数据迁移报错。
####  该场景多见于以下情况：全量迁移的数据不属于数据源的一个一致性快照，
    ↪ 随后从一个早于全量迁移数据之前的位置开始同步增量数据。
### syncers:            # sync 处理单元的运行配置参数。
###  global:           # 配置名称。
###    safe-mode: true # 设置为 true，会将来自数据源的 INSERT 改写为 REPLACE，将 UPDATE 改写为
    ↪ DELETE 与 REPLACE，从而保证在表结构中存在主键或唯一索引的条件下迁移数据时可以重复导入 DML
    ↪ 。在启动或恢复增量复制任务的前 1 分钟内 TiDB DM 会自动启动 safe mode。
```

关于任务的更多配置项，可以参考DM 任务完整配置文件介绍

在你启动数据迁移任务之前，建议使用check-task命令检查配置是否符合 DM 的配置要求，以降低后期报错的概率。

```
tiup dmctl --master-addr ${advertise-addr} check-task task.yaml
```

使用 tiup dmctl 执行以下命令启动数据迁移任务。

```
tiup dmctl --master-addr ${advertise-addr} start-task task.yaml
```

该命令中的参数描述如下：

参数	描述
–master-addr	dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}，例如：172.16.10.71:8261
start-task	命令用于创建数据迁移任务

如果任务启动失败，可根据返回结果的提示进行配置变更后执行 start-task task.yaml 命令重新启动任务。遇到问题请参考故障及处理方法 以及常见问题

3.3.5.4.3 查看任务状态

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用tiup dmctl执行query-status命令进行查询：

```
tiup dmctl --master-addr ${advertise-addr} query-status ${task-name}
```

关于查询结果的详细解读，请参考查询状态

3.3.5.4.4 监控任务与查看日志

你可以通过 Grafana 或者日志查看迁移任务的历史状态以及各种内部运行指标。

- 通过 Grafana 查看

如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，则使用部署时填写的 IP 及端口进入 Grafana，选择 DM 的 dashboard 查看 DM 相关监控项。

- 通过日志查看

  DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息，其中包含迁移任务的相关信息。各组件的日志目录如下：

  - DM-master 日志目录：通过 DM-master 进程参数`--log-file`设置。如果使用 TiUP 部署 DM，则日志目录默认位于`/dm-deploy/dm-master-8261/log/`。
  - DM-worker 日志目录：通过 DM-worker 进程参数`--log-file`设置。如果使用 TiUP 部署 DM，则日志目录默认位于`/dm-deploy/dm-worker-8262/log/`。

3.3.5.5 探索更多

- 关于 Dumpling
- 关于 Lightning
- 分库分表合并中的悲观/乐观模式
- 暂停数据迁移任务
- 恢复数据迁移任务
- 停止数据迁移任务
- 导出和导入集群的数据源和任务配置
- 处理出错的 DDL 语句
- 故障及处理方法

3.3.6 从 CSV 文件迁移数据到 TiDB

本文档介绍如何从 CSV 文件迁移数据到 TiDB。

TiDB Lightning 支持读取 CSV 格式的文件，以及其他定界符格式，如 TSV（制表符分隔值）。对于其他“平面文件”类型的数据导入，也可以参考本文档进行。

3.3.6.1 前提条件

- 安装 TiDB Lightning。
- 获取 Lightning 所需下游数据库权限。

3.3.6.2 第 1 步：准备 CSV 文件

将所有要导入的 CSV 文件放在同一目录下，若要 TiDB Lightning 识别所有 CSV 文件，文件名必须满足以下格式：

- 包含整张表数据的 CSV 文件，需命名为 `${db_name}.${table_name}.csv`。
- 如果一张表分布于多个 CSV 文件，这些 CSV 文件命名需加上文件编号的后缀，如 `${db_name}.${table_name}.003.csv`。数字部分不需要连续，但必须递增，并且需要用零填充数字部分，保证后缀为同样长度。

3.3.6.3 第 2 步：创建目标表结构

CSV 文件自身未包含表结构信息。要将 CSV 数据导入 TiDB，就必须为数据提供表结构。可以通过以下任一方法创建表结构：

- 方法一：使用 TiDB Lightning 创建表结构。

    1. 编写包含 DDL 语句的 SQL 文件。

        - 文件名格式为 `${db_name}-schema-create.sql`，其内容需包含 `CREATE DATABASE` 语句。
        - 文件名格式为 `${db_name}.${table_name}-schema.sql`，其内容需包含 `CREATE TABLE` 语句。

- 方法二：手动在下游 TiDB 建库和表。

3.3.6.4 第 3 步：编写配置文件

新建文件 `tidb-lightning.toml`，包含以下内容：

```
[lightning]
### 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
### "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
### "tidb"：TB 级以下数据量也可以采用`tidb`后端模式，下游 TiDB 可正常提供服务。
    ↪ 关于后端模式更多信息请参阅：https://docs.pingcap.com/tidb/stable/tidb-lightning-backends
backend = "local"
### 设置排序的键值对的临时存放地址，目标路径需要是一个空目录，至少需要数据源最大单表的空间，
    ↪ 建议与 `data-source-dir` 不同磁盘目录并使用闪存介质，独占 IO 会获得更好的导入性能
sorted-kv-dir = "/mnt/ssd/sorted-kv-dir"

[mydumper]
### 源数据目录。
data-source-dir = "${data-path}" # 本地或 S3 路径，例如：'s3://my-bucket/sql-backup?region=us-
    ↪ west-2'

### 定义 CSV 格式
[mydumper.csv]
### 字段分隔符，必须不为空。如果源文件中包含非字符串或数值类型的字段（如 binary, blob, bit 等），
    ↪ 则不建议源文件使用默认的“,”简单分隔符，推荐“|+|”等非常见字符组合
separator = ','
### 引用定界符，可以为零或多个字符。
delimiter = '"'
### CSV 文件是否包含表头。
### 如果为 true，则 lightning 会使用首行内容解析字段的对应关系。
header = true
### CSV 是否包含 NULL。
```

```
### 如果为 true，CSV 文件的任何列都不能解析为 NULL。
not-null = false
### 如果 `not-null` 为 false（即 CSV 可以包含 NULL），
### 为以下值的字段将会被解析为 NULL。
null = '\N'
### 是否将字符串中包含的反斜杠（'\'）字符作为转义字符处理
backslash-escape = true
### 是否移除行尾的最后一个分隔符。
trim-last-separator = false

[tidb]
### 目标集群的信息
host = ${host}                # 例如：172.16.32.1
port = ${port}                # 例如：4000
user = "${user_name}"         # 例如："root"
password = "${password}"      # 例如："rootroot"
status-port = ${status-port}  # 导入过程 Lightning 需要在从 TiDB 的“状态端口”获取表结构信息，
    ↪ 例如：10080
pd-addr = "${ip}:${port}"     # 集群 PD 的地址，Lightning 通过 PD 获取部分信息，例如
    ↪ 172.16.31.3:2379。当 backend = "local" 时 status-port 和 pd-addr 必须正确填写，
    ↪ 否则导入将出现异常。
```

关于配置文件更多信息，可参阅TiDB Lightning 配置参数。

3.3.6.5 第 4 步：导入性能优化（可选）

导入文件的大小统一约为 256 MiB 时，TiDB Lightning 可达到最佳工作状态。如果导入单个 CSV 大文件，TiDB Lightning 在默认配置下只能使用一个线程来处理，这会降低导入速度。

要解决此问题，可先将 CSV 文件分割为多个文件。对于通用格式的 CSV 文件，在没有读取整个文件的情况下，无法快速确定行的开始和结束位置。因此，默认情况下 TiDB Lightning 不会自动分割 CSV 文件。但如果你确定待导入的 CSV 文件符合特定的限制要求，则可以启用 `strict-format` 模式。启用后，TiDB Lightning 会将单个 CSV 大文件分割为单个大小为 256 MiB 的多个文件块进行并行处理。

> **注意：**
>
> 如果 CSV 文件不是严格格式，但 `strict-format` 被误设为 `true`，跨多行的单个完整字段会被分割成两部分，导致解析失败，甚至不报错地导入已损坏的数据。

严格格式的 CSV 文件中，每个字段仅占一行，即必须满足以下条件之一：

- delimiter 为空；
- 每个字段不包含 CR（\r）或 LF（\n）。

如果你确认满足条件，可按如下配置开启 strict-format 模式以加快导入速度。

```
[mydumper]
strict-format = true
```

3.3.6.6 第 5 步：执行导入

运行 tidb-lightning。如果直接在命令行中启动程序，可能会因为 SIGHUP 信号而退出，建议配合 nohup 或 screen 等工具，如：

```
nohup tiup tidb-lightning -config tidb-lightning.toml > nohup.out 2>&1 &
```

导入开始后，可以采用以下任意方式查看进度：

- 通过 grep 日志关键字 progress 查看进度，默认 5 分钟更新一次。
- 通过监控面板查看进度，请参考TiDB Lightning 监控。
- 通过 Web 页面查看进度，请参考Web 界面。

导入完毕后，TiDB Lightning 会自动退出。查看日志的最后 5 行中会有 the whole procedure completed，则表示导入成功。

> 注意：
>
> 无论导入成功与否，最后一行都会显示 tidb lightning exit。它只是表示 TiDB Lightning 正常退出，不代表任务完成。

如果导入过程中遇到问题，请参见TiDB Lightning 常见问题。

3.3.6.7 其他格式的文件

若数据源为其他格式，除文件名仍必须以 .csv 结尾外，配置文件 tidb-lightning.toml 的 [mydumper.csv] 格式定义同样需要做相应修改。常见的格式修改如下：

TSV：

```
### 格式示例
### ID    Region    Count
### 1     East      32
### 2     South     NULL
### 3     West      10
### 4     North     39

### 格式配置
[mydumper.csv]
separator = "\t"
```

```
delimiter = ''
header = true
not-null = false
null = 'NULL'
backslash-escape = false
trim-last-separator = false
```

TPC-H DBGEN：

```
### 格式示例
### 1|East|32|
### 2|South|0|
### 3|West|10|
### 4|North|39|

### 格式配置
[mydumper.csv]
separator = '|'
delimiter = ''
header = false
not-null = true
backslash-escape = false
trim-last-separator = true
```

3.3.6.8 探索更多

- CSV 支持与限制

3.3.7 从 SQL 文件迁移数据到 TiDB

本文介绍如何使用 TiDB Lightning 从 MySQL SQL 文件迁移数据到 TiDB。关于如何生成 MySQL SQL 文件，请参考 Dumpling 文档中的导出为 SQL 文件。

3.3.7.1 前提条件

- 使用 TiUP 安装 TiDB Lightning
- Lightning 所需下游数据库权限

3.3.7.2 第 1 步：准备 SQL 文件

将所有 SQL 文件放到统一目录下，例如 `/data/my_datasource/` 或 `s3://my-bucket/sql-backup?region=us-west-2`。Lightning 将递归地寻找该目录下及其子目录内的所有 `.sql` 文件。

3.3.7.3 第 2 步：定义目标表结构

要导入 TiDB，就必须为 SQL 文件提供表结构。

如果使用 Dumpling 工具导出数据，则会自动导出表结构文件。此外，其他方式导出的数据可以通过以下任一方法创建表结构：

- 方法一：使用 TiDB Lightning 创建表结构。
    1. 编写包含 DDL 语句的 SQL 文件。
        - 文件名格式为 `${db_name}-schema-create.sql`，其内容需包含 `CREATE DATABASE` 语句。
        - 文件名格式为 `${db_name}.${table_name}-schema.sql`，其内容需包含 `CREATE TABLE` 语句。
    2. 后续导入过程中，在 `tidb-lightning.toml` 中添加如下设置：
- 方法二：手动在下游 TiDB 建库和表。

3.3.7.4 第 3 步：编写配置文件

新建文件 `tidb-lightning.toml`，包含以下内容：

```
[lightning]
### 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
### "local"：默认使用该模式，适用于 TiB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
backend = "local"
### # "tidb"：TiB 级以下数据量也可以采用 `tidb` 后端模式，下游 TiDB 可正常提供服务。
    ↪ 关于后端模式更多信息请参考 https://docs.pingcap.com/zh/tidb/stable/tidb-lightning-
    ↪ backends。
### 设置排序的键值对的临时存放地址。目标路径需要是一个空目录，至少需要数据源最大单表的空间，
    ↪ 建议与 `data-source-dir` 不同磁盘目录并使用闪存介质，独占 I/O 会获得更好的导入性能。
sorted-kv-dir = "${sorted-kv-dir}"

[mydumper]
### 源数据目录
data-source-dir = "${data-path}" # 本地或 S3 路径，例如：'s3://my-bucket/sql-backup?region=us-
    ↪ west-2'

[tidb]
### 目标集群的信息
host = ${host}               # 例如：172.16.32.1
port = ${port}               # 例如：4000
user = "${user_name}"        # 例如："root"
password = "${password}"     # 例如："rootroot"
status-port = ${status-port}  # 导入过程 Lightning 需要在从 TiDB 的“状态端口”获取表结构信息，
    ↪ 例如：10080
```

```
pd-addr = "${ip}:${port}"     # 集群 PD 的地址，Lightning 通过 PD 获取部分信息，例如
    ↪ 172.16.31.3:2379。当 backend = "local" 时 status-port 和 pd-addr 必须正确填写，
    ↪ 否则导入将出现异常。
```

关于配置文件更多信息，可参阅TiDB Lightning Configuration.

3.3.7.5 第 4 步：执行导入

运行 tidb-lightning。如果直接在命令行中启动程序，可能会因为 SIGHUP 信号而退出，建议配合 nohup 或 screen 等工具。

若从 S3 导入，则需将有权限访问该 Amazon S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 Lightning 节点。

```
export AWS_ACCESS_KEY_ID=${access_key}
export AWS_SECRET_ACCESS_KEY=${secret_key}
nohup tiup tidb-lightning -config tidb-lightning.toml  > nohup.out 2>&1 &
```

同时，TiDB Lightning 还支持从 ~/.aws/credentials 读取凭证文件。

导入开始后，可以采用以下任意方式查看进度：

- 通过 grep 日志关键字 progress 查看进度，默认 5 分钟更新一次。
- 通过监控面板查看进度，请参考TiDB Lightning 监控。
- 通过 Web 页面查看进度，请参考Web 界面。

导入完毕后，TiDB Lightning 会自动退出。查看日志的最后 5 行中会有 the whole procedure completed，则表示导入成功。

> 注意：
>
> 无论导入成功与否，最后一行都会显示 tidb lightning exit。它只是表示 TiDB Lightning 正常退出，不代表任务完成。

如果导入过程中遇到问题，请参见TiDB Lightning 常见问题。

3.3.8 将 TiDB 集群的增量数据实时同步到另一集群

本文档介绍如何配置一个 TiDB 集群以及该集群的 TiDB 或 MySQL 从集群，并将增量数据实时从主集群同步到从集群。

如果你需要配置一个运行中的 TiDB 集群和其从集群，以进行实时增量数据同步，可使用Backup & Restore (BR)、Dumpling 和TiDB Binlog。

3.3.8.1 实现原理

任何写入 TiDB 的事务均被指定了唯一的 commit TS（commit 时间戳），可以通过该 TS 获取一个 TiDB 集群全局一致的状态。

通过 BR 或者 Dumpling 将集群的数据从一个全局一致的时间点导出，然后再用 TiDB Binlog 衔接这个全局一致的时间点开启增量同步，即同步过程分为全量同步和增量同步两个阶段：

1. 执行全量备份恢复，并且获取到备份数据的 commit TS。
2. 执行增量同步，确保增量同步的起始点是备份数据的 commit TS。

> **注意：**
>
> 导出备份数据时的 commit TS 是闭区间，使用 TiDB Binlog 开启同步的 initial-commit-ts 是开区间。

3.3.8.2 同步操作步骤

假设现有 A 集群正常运行，你需要先新建一个 B 集群作为 A 集群的从集群，并将 A 集群的数据实时增量进行同步到 B 集群，可以通过以下步骤来实现。

3.3.8.2.1 步骤一：开启 TiDB Binlog

确保 A 集群已经部署并且开启了 TiDB Binlog。

3.3.8.2.2 步骤二：全量导出集群数据

1. 使用以下任一工具将集群 A 的数据全局一致地导出到指定路径下：
    - 使用BR 全量备份
    - 使用Dumpling 全量导出
2. 获取全局一致的时间戳 COMMIT_TS：
    - 使用 BR `validate` 指令获取备份的时间戳，示例如下：

```
COMMIT_TS=`br validate decode --field="end-version" -s local:///home/tidb/backupdata |
    ↪ tail -n1`
```

    - 或者查阅 Dumpling metadata 获取 Pos (COMMIT_TS)。

```
cat metadata
```

```
Started dump at: 2020-11-10 10:40:19
SHOW MASTER STATUS:
        Log: tidb-binlog
        Pos: 420747102018863124

Finished dump at: 2020-11-10 10:40:20
```

3. 将集群 A 的数据导入 B 集群。

3.3.8.2.3 步骤三：同步增量数据

修改 TiDB Binlog 的 `drainer.toml` 配置文件，增加如下配置，从指定的位置 `COMMIT_TS` 开始将数据同步至 B 集群。

```
initial-commit-ts = COMMIT_TS
[syncer.to]
host = {B 集群的 IP 地址}
port = 3306
```

3.4 复杂迁移场景

3.4.1 上游使用 pt-osc/gh-ost 工具的持续同步场景

在生产业务中执行 DDL 时，产生的锁表操作会一定程度阻塞数据库的读取或者写入。为了把对读写的影响降到最低，用户往往会选择 online DDL 工具执行 DDL。常见的 Online DDL 工具有 gh-ost 和 pt-osc。

在使用 DM 完成 MySQL 到 TiDB 的数据迁移时，可以开启`online-ddl`配置，实现 DM 工具与 gh-ost 或 pt-osc 的协同。

具体迁移操作可参考已有数据迁移场景：

- 从小数据量 MySQL 迁移数据到 TiDB
- 从大数据量 MySQL 迁移数据到 TiDB
- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB
- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

3.4.1.1 开启 DM 的 online-ddl 特性

配置 DM 的任务配置文件时，将全局参数的`online-ddl`设置为 true，具体配置示例如下图：

```
### ----------- 全局配置 -----------
#### ********* 基本信息配置 *********
name: test                      # 任务名称，需要全局唯一
task-mode: all                  # 任务模式，可设为 "full"、"incremental"、"all"
shard-mode: "pessimistic"       # 如果为分库分表合并任务则需要配置该项。默认使用悲观协调模式 "
    ↪ pessimistic"，在深入了解乐观协调模式的原理和使用限制后，也可以设置为乐观协调模式 "
    ↪ optimistic"
meta-schema: "dm_meta"          # 下游储存 `meta` 信息的数据库
online-ddl: true                # 开启 DM 的 online DDL 支持特性。兼容上游使用 gh-ost 、pt-osc
    ↪ 两种工具的自动处理
```

3.4.1.2 开启 online-ddl 的影响

当开启`online-ddl`特性后，DM 同步 gh-ost 或 pt-osc 工具所产生的 DDL 语句将会发生一些变化：

上游 gh-ost 或 pt-osc 工具工作流如下：

- 根据 DDL 目标表 (real table) 的表结构新建一张镜像表 (ghost table)；
- 在镜像表上应用变更 DDL；
- 将 DDL 目标表的数据同步到镜像表；
- 在目标表与镜像表数据一致后，通过 RENAME 语句使镜像表替换掉目标表。

`online-ddl=true` 时 DM 的同步方式：

- 不在下游新建镜像表 (ghost table)；
- 记录变更 DDL；
- 仅从镜像表同步数据；
- 在下游执行 DDL 变更。

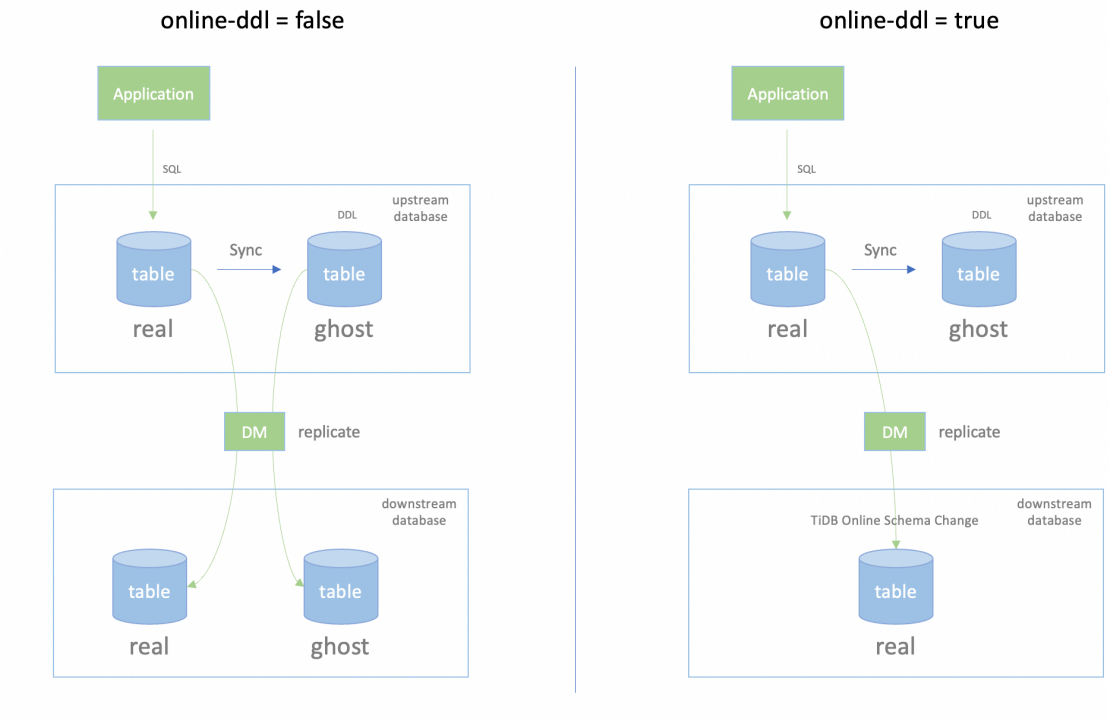


图 9: dm-online-ddl

这些变化将带来一些好处：

- 下游 TiDB 无需创建和同步镜像表，节约相应存储空间和网络传输等开销；
- 在分库分表合并场景下，自动忽略各分表镜像表的 RENAME 操作，保证同步正确性。

如果您想深入了解其实现原理，请阅读以下两篇技术博客：

- DM 源码阅读系列文章（八）Online Schema Change 迁移支持
- TiDB Online Schema Change 原理

3.4.1.3 探索更多

- DM 与 online DDL 工具协作细节

3.4.2 下游存在更多列的迁移场景

本文档介绍数据同步时，下游存在更多列的迁移场景需要的注意事项。具体迁移操作可参考已有数据迁移场景：

- 从小数据量 MySQL 迁移数据到 TiDB
- 从大数据量 MySQL 迁移数据到 TiDB
- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB
- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

3.4.2.1 使用 DM 迁移至存在更多列的下游

DM 同步上游的 binlog 时，会尝试使用下游当前的表结构来解析 binlog 并生成相应的 DML 语句。如果上游的 binlog 里数据表的列数与下游表结构的列数不一致，则会产生如下错误：

```
"errors": [
    {
        "ErrCode": 36027,
        "ErrClass": "sync-unit",
        "ErrScope": "internal",
        "ErrLevel": "high",
        "Message": "startLocation: [position: (mysql-bin.000001, 2022), gtid-set:09bec856-ba95-11
            ↪ ea-850a-58f2b4af5188:1-9 ], endLocation: [position: (mysql-bin.000001, 2022),
            ↪ gtid-set: 09bec856-ba95-11ea-850a-58f2b4af5188:1-9]: gen insert sqls failed,
            ↪ schema: log, table: messages: Column count doesn't match value count: 3 (columns)
            ↪  vs 2 (values)",
        "RawCause": "",
        "Workaround": ""
    }
]
```

例如上游表结构为：

```
### 上游表结构
CREATE TABLE `messages` (
  `id` int(11) NOT NULL,
  PRIMARY KEY (`id`)
)
```

例如下游表结构为：

```
### 下游表结构
CREATE TABLE `messages` (
  `id` int(11) NOT NULL,
  `message` varchar(255) DEFAULT NULL, # 下游比上游多出的列。
  PRIMARY KEY (`id`)
)
```

当 DM 尝试使用下游表结构解析上游产生的 binlog event 时，DM 会报出上述 `Column count doesn't match` 错误。

此时，你可以使用 `operate-schema` 命令来为数据源中需要迁移的表指定表结构，表结构需要对应 DM 将要开始同步的 binlog event 的数据。如果你在进行分表合并的数据迁移，那么需要为每个分表按照如下步骤在 DM 中设置用于解析 binlog event 的表结构。具体操作为：

1. 在 DM 中，新建一个 `.sql` 文件，并将上游表结构对应的 `CREATE TABLE` 语句添加到该文件。例如，将以下表结构保存到 `log.messages.sql` 中。

    ```
    # 上游表结构
    CREATE TABLE `messages` (
    `id` int(11) NOT NULL,
    PRIMARY KEY (`id`)
    )
    ```

2. 使用 `operate-schema` 命令为数据源中需要迁移的表设置表结构（此时数据迁移任务应该由于上述 `Column count doesn't match` 错误而处于 Paused 状态）。

    ```
    tiup dmctl --master-addr ${advertise-addr} operate-schema set -s ${source-id} ${task-name} -
        ↪ d ${database-name} -t ${table-name} ${schema-file}
    ```

    该命令中的参数描述如下：

参数	描述

参数	描述
–master-addr	指定 dm-ctl 要连接的集群的任意 DM-master 节点的 ${ advertise-addr }。${ advertise-addr } 表示 DM-master 向外界宣告的地址。

参数	描述
operate-schema set	手动设置 schema 信息

参数	描述
-s	指定 source。$ ↪ { ↪ source ↪ - ↪ id ↪ } ↪ 表示 MySQL 数据源 ID。$ ↪ { ↪ task ↪ - ↪ name ↪ } ↪ 表示数据同步任务配置文件 task ↪ . ↪ yaml ↪ 中定义的同步任务

参数	描述
-d	指定 database。$ ↪ { ↪ database ↪ - ↪ name ↪ } ↪ 表示上游数据库名。

参数	描述
-t	指定 table。
	${table-name}
	表示上游数据表名。
	${schema-file}
	表示将被设置的表结构文件。

例如：

```
tiup dmctl --master-addr 172.16.10.71:8261 operate-schema set -s mysql-01 task-test -d log -
    ↪ t message log.message.sql
```

3. 使用 `resume-task` 命令恢复处于 Paused 状态的同步任务。

```
tiup dmctl --master-addr ${advertise-addr} resume-task ${task-name}
```

4. 使用 `query-status` 命令确认数据迁移任务是否运行正常。

```
tiup dmctl --master-addr ${advertise-addr} query-status resume-task ${task-name}
```

3.4.3 如何过滤 binlog 事件

本文档介绍使用 DM 持续增量数据同步时，如何过滤 binlog 事件。具体迁移操作可参考已有数据迁移场景：

- 从小数据量 MySQL 迁移数据到 TiDB
- 从大数据量 MySQL 迁移数据到 TiDB
- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB
- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

3.4.3.1 配置方式

配置 DM 的任务配置文件时，增加如下`filter`，具体配置示例如下图：

```
filters:
  rule-1:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    events: ["truncate table", "drop table"]
    sql-pattern: ["^DROP\\s+PROCEDURE", "^CREATE\\s+PROCEDURE"]
    action: Ignore
```

- `schema-pattern/table-pattern`：对匹配上的 schema 或 table 进行过滤
- `events`：binlog events，支持的 Event 如下表所示:

Event	分类	说明
all		匹配所有 events
all dml		匹配所有 DML events
all ddl		匹配所有 DDL events
none		不匹配任何 events
none ddl		不包含任何 DDL events
none dml		不包含任何 DML events
insert	DML	匹配 insert DML event
update	DML	匹配 update DML event

Event	分类	说明
delete	DML	匹配 delete DML event
create database	DDL	匹配 create database event
drop database	DDL	匹配 drop database event
create table	DDL	匹配 create table event
create index	DDL	匹配 create index event
drop table	DDL	匹配 drop table event
truncate table	DDL	匹配 truncate table event
rename table	DDL	匹配 rename table event
drop index	DDL	匹配 drop index event
alter table	DDL	匹配 alter table event

- `sql-pattern`：匹配指定的 DDL SQL 语句，支持正则表达式匹配。
- `action`：可取值 Do 或 Ignore。
  - Do：白名单。binlog event 如果满足下面两个条件之一将会被同步：
    * 符合 events 条件；
    * sql-pattern 不为空，且对应的 SQL 可以匹配上 sql-pattern 中任意一项。
  - Ignore：黑名单。如果满足下面两个条件之一就会被过滤掉：
    * 符合 events 条件；
    * sql-pattern 不为空，且对应的 SQL 可以匹配上 sql-pattern 中任意一项

注意：如果同时配置 Do/Ignore，则 Ignore 优先级更高。binlog event 不匹配白名单或者匹配黑名单都将被直接过滤。

3.4.3.2 使用场景举例

3.4.3.2.1 过滤分库分表的所有删除操作

设置 filter-table-rule 和 filter-schema-rule 两个过滤规则，具体如下：

```
filters:
  filter-table-rule:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    events: ["truncate table", "drop table", "delete"]
    action: Ignore
  filter-schema-rule:
    schema-pattern: "test_*"
    events: ["drop database"]
    action: Ignore
```

3.4.3.2.2 只迁移分库分表的 DML 操作

设置两个 Binlog event filter rule：

```
filters:
  do-table-rule:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    events: ["create table", "all dml"]
    action: Do
  do-schema-rule:
    schema-pattern: "test_*"
    events: ["create database"]
    action: Do
```

3.4.3.2.3 过滤 TiDB 不支持的 SQL 语句

```
filters:
  filter-procedure-rule:
    schema-pattern: "*"
    sql-pattern: [".*\\s+DROP\\s+PROCEDURE", ".*\\s+CREATE\\s+PROCEDURE", "ALTER\\s+TABLE[\\s\\S
        ↪ ]*ADD\\s+PARTITION", "ALTER\\s+TABLE[\\s\\S]*DROP\\s+PARTITION"]
    action: Ignore
```

> 注意：
>
> 全局过滤规则的设置必须尽可能严格，以避免过滤掉需要迁移的数据。

3.4.3.3 探索更多

- 如何通过 SQL 表达式过滤 binlog

3.4.4 如何通过 SQL 表达式过滤 DML

本文档介绍使用 DM 持续增量数据同步时，如何更加精细的过滤 binlog 事件。具体迁移操作可参考已有数据迁移场景：

- 从小数据量 MySQL 迁移数据到 TiDB
- 从大数据量 MySQL 迁移数据到 TiDB
- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB
- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

在进行增量数据迁移时，可以通过如何过滤 binlog 事件功能过滤某些类型的 binlog event，例如不向下游迁移 DELETE 事件以达到归档、审计等目的。但是 binlog event filter 无法以更细粒度判断某一行的 DELETE 事件是否要被过滤。

为了解决上述问题，从 v2.0.5 起，DM 支持在增量数据同步阶段使用binlog value filter过滤迁移数据。DM 支持的 ROW 格式的 binlog 中，binlog event 带有所有列的值。你可以基于这些值配置 SQL 表达式。如果该表达式对于某条行变更的计算结果是 TRUE，DM 就不会向下游迁移该条行变更。

与如何过滤 binlog 事件类似，表达式过滤需要在数据迁移任务配置文件里配置，详见下面配置样例。完整的配置及意义，可以参考DM 完整配置文件示例：

```
name: test
task-mode: all

mysql-instances:
  - source-id: "mysql-replica-01"
    expression-filters: ["even_c"]

expression-filter:
  even_c:
    schema: "expr_filter"
    table: "tbl"
    insert-value-expr: "c % 2 = 0"
```

上面的示例配置了 even_c 规则，并让 source ID 为 mysql-replica-01 的数据源引用了该规则。even_c 规则的含义是：对于 expr_filter 库下的 tbl 表，当插入的 c 的值为偶数 (c % 2 = 0) 时，不将这条插入语句迁移到下游。下面展示该规则的使用效果。

在上游数据源增量插入以下数据：

```
INSERT INTO tbl(id, c) VALUES (1, 1), (2, 2), (3, 3), (4, 4);
```

随后在下游查询 tbl 表，可见只有 c 的值为单数的行迁移到了下游：

```
MySQL [test]> select * from tbl;

+------+------+
| id   | c    |
+------+------+
|    1 |    1 |
|    3 |    3 |
+------+------+

2 rows in set (0.001 sec)
```

3.4.4.1 配置参数及规则说明

- schema：要匹配的上游数据库库名，不支持通配符匹配或正则匹配。

- table: 要匹配的上游表名，不支持通配符匹配或正则匹配。
- insert-value-expr：配置一个表达式，对 INSERT 类型的 binlog event (WRITE_ROWS_EVENT) 带有的值生效。不能与 update-old-value-expr、update-new-value-expr、delete-value-expr 出现在一个配置项中。
- update-old-value-expr：配置一个表达式，对 UPDATE 类型的 binlog event (UPDATE_ROWS_EVENT) 更新对应的旧值生效。不能与 insert-value-expr、delete-value-expr 出现在一个配置项中。
- update-new-value-expr：配置一个表达式，对 UPDATE 类型的 binlog event (UPDATE_ROWS_EVENT) 更新对应的新值生效。不能与 insert-value-expr、delete-value-expr 出现在一个配置项中。
- delete-value-expr：配置一个表达式，对 DELETE 类型的 binlog event (DELETE_ROWS_EVENT) 带有的值生效。不能与 insert-value-expr、update-old-value-expr、update-new-value-expr 出现在一个配置项中。

> 注意：
>
> - update-old-value-expr 可以与 update-new-value-expr 同时配置。
> - 当二者同时配置时，会将 “更新 + 旧值 “满足update-old-value-expr 且” 更新 + 新值 “满足 update-new-value-expr 的行过滤掉。
> - 当只配置一者时，配置的这条表达式会决定是否过滤整个行变更，即旧值的删除和新值的插入会作为一个整体被过滤掉。

SQL 表达式可以涉及一列或多列，也可使用 TiDB 支持的 SQL 函数，例如 `c % 2 = 0`、`a*a + b*b = c*c`、`ts > NOW()`。

TIMESTAMP 类型的默认时区是任务配置文件中指定的时区，默认值是下游时区。可以使用 `c_timestamp = '2021-01-01 12:34:56.5678+08:00'` 的方式显式指定时区。

配置项 expression-filter 下可以定义多条过滤规则，上游数据源在其 expression-filters 配置项中引用需要的规则使其生效。当有多条规则生效时，匹配到任意一条规则即会导致某个行变更被过滤。

> 注意：
>
> 为某张表设置过多的表达式过滤会增加 DM 的计算开销，可能导致数据迁移速度变慢。

4 升级方案

4.0.1 使用 TiUP 升级 TiDB

本文档适用于以下升级路径：

- 使用 TiUP 从 TiDB 4.0 版本升级至 TiDB 5.4 及后续修订版本。
- 使用 TiUP 从 TiDB 5.0 版本升级至 TiDB 5.4 及后续修订版本。
- 使用 TiUP 从 TiDB 5.1 版本升级至 TiDB 5.4 及后续修订版本。
- 使用 TiUP 从 TiDB 5.2 版本升级至 TiDB 5.4 及后续修订版本。
- 使用 TiUP 从 TiDB 5.3 版本升级至 TiDB 5.4 及后续修订版本。

> **注意：**
>
> 如果原集群是 3.0 或 3.1 或更早的版本，不支持直接升级到 5.4 及后续修订版本。你需要先从早期版本升级到 4.0 后，再从 4.0 升级到 5.4 及后续修订版本。

4.0.1.1 1. 升级兼容性说明

- TiDB 目前暂不支持版本降级或升级后回退。
- 使用 TiDB Ansible 管理的 4.0 版本集群，需要先按照 4.0 版本文档的说明将集群导入到 TiUP (`tiup cluster`) 管理后，再按本文档说明升级到 5.4 版本及后续修订版本。
- 若要将 3.0 之前的版本升级至 5.4 版本：
    1. 首先通过 TiDB Ansible 升级到 3.0 版本。
    2. 然后按照 4.0 版本文档的说明，使用 TiUP (`tiup cluster`) 将 TiDB Ansible 配置导入。
    3. 将集群升级至 4.0 版本。
    4. 按本文档说明将集群升级到 5.4 版本。
- 支持 TiDB Binlog，TiCDC，TiFlash 等组件版本的升级。
- 具体不同版本的兼容性说明，请查看各个版本的 Release Note。请根据各个版本的 Release Note 的兼容性更改调整集群的配置。
- 升级 v5.3 之前版本的集群到 v5.3 及后续版本时，默认部署的 Prometheus 会从 v2.8.1 升级到 v2.27.1，v2.27.1 提供更多的功能并解决了安全风险。Prometheus v2.27.1 相对于 v2.8.1 存在 Alert 时间格式变化，详情见 Prometheus commit。

4.0.1.2 2. 升级前准备

本部分介绍实际开始升级前需要进行的更新 TiUP 和 TiUP Cluster 组件版本等准备工作。

> **警告：**
>
> - 在升级 TiDB 集群的过程中，请勿执行 DDL 语句，否则可能会出现行为未定义的问题。
> - 集群中有 DDL 语句正在被执行时（通常为 `ADD INDEX` 和列类型变更等耗时较久的 DDL 语句），请勿进行升级操作。在升级前，建议使用 `ADMIN SHOW DDL` 命令查看集群中是否有正在进行的 DDL Job。如需升级，请等待 DDL 执行完成或使用 `ADMIN CANCEL DDL` 命令取消该 DDL Job 后再进行升级。

4.0.1.2.1 2.1 升级 TiUP 或更新 TiUP 离线镜像

升级 TiUP 和 TiUP Cluster

> **注意：**
>
> 如果原集群中控机不能访问 `https://tiup-mirrors.pingcap.com` 地址，可跳到步骤 2.2 使用离线升级方式。

1. 先升级 TiUP 版本（建议 `tiup` 版本不低于 `1.9.0`）：

```
tiup update --self
tiup --version
```

2. 再升级 TiUP Cluster 版本（建议 `tiup cluster` 版本不低于 `1.9.0`）：

```
tiup update cluster
tiup cluster --version
```

更新 TiUP 离线镜像

> **注意：**
>
> 如果原集群不是通过离线部署方式部署的，可忽略此步骤。

可以参考使用 TiUP 部署 TiDB 集群的步骤下载部署新版本的 TiUP 离线镜像，上传到中控机。在执行 `local_install.sh` 后，TiUP 会完成覆盖升级。

```
tar xzvf tidb-community-server-${version}-linux-amd64.tar.gz
sh tidb-community-server-${version}-linux-amd64/local_install.sh
source /home/tidb/.bash_profile
```

覆盖升级完成后，执行下列命令升级 Cluster 组件：

```
tiup update cluster
```

此时离线镜像已经更新成功。如果覆盖后发现 TiUP 运行报错，可能是 manifest 未更新导致，可尝试 `rm -rf ~/.tiup/manifests/*` 后再使用。

4.0.1.2.2 2.2 编辑 TiUP Cluster 拓扑配置文件

> **注意：**
>
> 以下情况可跳过此步骤：
>
> - 原集群没有修改过配置参数，或通过 tiup cluster 修改过参数但不需要调整。
> - 升级后对未修改过的配置项希望使用 `5.4` 默认参数。

1. 进入拓扑文件的 vi 编辑模式：

```
tiup cluster edit-config <cluster-name>
```

2. 参考 topology 配置模板的格式，将希望修改的参数填到拓扑文件的 server_configs 下面。

修改完成后 :wq 保存并退出编辑模式，输入 Y 确认变更。

> 注意：
>
> 升级到 5.4 版本前，请确认已在 4.0 修改的参数在 5.4 版本中是兼容的，可参考 TiKV 配置文件描述。
>
> 以下 TiKV 参数在 TiDB v5.0 已废弃。如果在原集群配置过以下参数，需要通过 edit-config 编辑模式删除这些参数：
>
> - pessimistic-txn.enabled
> - server.request-batch-enable-cross-command
> - server.request-batch-wait-duration

4.0.1.2.3 2.3 检查当前集群的健康状况

为避免升级过程中出现未定义行为或其他故障，建议在升级前对集群当前的 region 健康状态进行检查，此操作可通过 check 子命令完成。

```
tiup cluster check <cluster-name> --cluster
```

执行结束后，最后会输出 region status 检查结果。如果结果为 “All regions are healthy”，则说明当前集群中所有 region 均为健康状态，可以继续执行升级；如果结果为 “Regions are not fully healthy: m miss-peer, n pending-peer” 并提示 “Please fix unhealthy regions before other operations.”，则说明当前集群中有 region 处在异常状态，应先排除相应异常状态，并再次检查结果为 “All regions are healthy” 后再继续升级。

4.0.1.3 3. 升级 TiDB 集群

本部分介绍如何滚动升级 TiDB 集群以及如何进行升级后的验证。

4.0.1.3.1 3.1 将集群升级到指定版本

升级的方式有两种：不停机升级和停机升级。TiUP Cluster 默认的升级 TiDB 集群的方式是不停机升级，即升级过程中集群仍然可以对外提供服务。升级时会对各节点逐个迁移 leader 后再升级和重启，因此对于大规模集群需要较长时间才能完成整个升级操作。如果业务有维护窗口可供数据库停机维护，则可以使用停机升级的方式快速进行升级操作。

不停机升级

```
tiup cluster upgrade <cluster-name> <version>
```

以升级到 5.4.0 版本为例：

```
tiup cluster upgrade <cluster-name> v5.4.0
```

> 注意：
>
> - 滚动升级会逐个升级所有的组件。升级 TiKV 期间，会逐个将 TiKV 上的所有 leader 切走再停止该 TiKV 实例。默认超时时间为 5 分钟（300 秒），超时后会直接停止该实例。
> - 如果不希望驱逐 leader，而希望快速升级集群至新版本，可以在上述命令中指定 `--force`，该方式会造成性能抖动，不会造成数据损失。
> - 如果希望保持性能稳定，则需要保证 TiKV 上的所有 leader 驱逐完成后再停止该 TiKV 实例，可以指定 `--transfer-timeout` 为一个更大的值，如 `--transfer-timeout 3600`，单位为秒。

停机升级

在停机升级前，首先需要将整个集群关停。

```
tiup cluster stop <cluster-name>
```

之后通过 `upgrade` 命令添加 `--offline` 参数来进行停机升级。

```
tiup cluster upgrade <cluster-name> <version> --offline
```

升级完成后集群不会自动启动，需要使用 `start` 命令来启动集群。

```
tiup cluster start <cluster-name>
```

4.0.1.3.2 3.2 升级后验证

执行 `display` 命令来查看最新的集群版本 `TiDB Version`：

```
tiup cluster display <cluster-name>
```

```
Cluster type:       tidb
Cluster name:       <cluster-name>
Cluster version:    v5.4.0
```

> 注意：
>
> TiUP 及 TiDB 默认会收集使用情况信息，并将这些信息分享给 PingCAP 用于改善产品。若要了解所收集的信息详情及如何禁用该行为，请参见遥测。

4.0.1.4 4. 升级 FAQ

本部分介绍使用 TiUP 升级 TiDB 集群遇到的常见问题。

4.0.1.4.1 4.1 升级时报错中断，处理完报错后，如何继续升级

重新执行 `tiup cluster upgrade` 命令进行升级，升级操作会重启之前已经升级完成的节点。如果不希望重启已经升级过的节点，可以使用 `replay` 子命令来重试操作，具体方法如下：

1. 使用 `tiup cluster audit` 命令查看操作记录：

```
tiup cluster audit
```

   在其中找到失败的升级操作记录，并记下该操作记录的 ID，下一步中将使用 `<audit-id>` 表示操作记录 ID 的值。

2. 使用 `tiup cluster replay <audit-id>` 命令重试对应操作：

```
tiup cluster replay <audit-id>
```

4.0.1.4.2 4.2 升级过程中 evict leader 等待时间过长，如何跳过该步骤快速升级

可以指定 `--force`，升级时会跳过 `PD transfer leader` 和 `TiKV evict leader` 过程，直接重启并升级版本，对线上运行的集群性能影响较大。命令如下：

```
tiup cluster upgrade <cluster-name> <version> --force
```

4.0.1.4.3 4.3 升级完成后，如何更新 pd-ctl 等周边工具版本

可通过 TiUP 安装对应版本的 `ctl` 组件来更新相关工具版本：

```
tiup install ctl:v5.4.0
```

4.0.1.5 5. TiDB 5.4 兼容性变化

- 兼容性变化请参考 5.4 Release Notes。
- 请避免在对使用 TiDB Binlog 的集群进行滚动升级过程中新创建聚簇索引表。

5 集群扩缩容

5.1 使用 TiUP 扩容缩容 TiDB 集群

TiDB 集群可以在不中断线上服务的情况下进行扩容和缩容。

本文介绍如何使用 TiUP 扩容缩容集群中的 TiDB、TiKV、PD、TiCDC 或者 TiFlash 节点。如未安装 TiUP，可参考部署文档中的步骤。

你可以通过 `tiup cluster list` 查看当前的集群名称列表。

例如，集群原拓扑结构如下所示：

主机 IP	服务
10.0.1.3	TiDB + TiFlash
10.0.1.4	TiDB + PD
10.0.1.5	TiKV + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.1 扩容 TiDB/PD/TiKV 节点

如果要添加一个 TiDB 节点，IP 地址为 10.0.1.5，可以按照如下步骤进行操作。

> **注意：**
>
> 添加 PD 节点和添加 TiDB 节点的步骤类似。添加 TiKV 节点前，建议预先根据集群的负载情况调整 PD 调度参数。

5.1.1.1 1. 编写扩容拓扑配置

> **注意：**
>
> - 默认情况下，可以不填写端口以及目录信息。但在单机多实例场景下，则需要分配不同的端口以及目录，如果有端口或目录冲突，会在部署或扩容时提醒。
> - 从 TiUP v1.0.0 开始，扩容配置会继承原集群配置的 global 部分。

在 scale-out.yaml 文件添加扩容拓扑配置：

```
vi scale-out.yaml
```

```
tidb_servers:
  - host: 10.0.1.5
    ssh_port: 22
    port: 4000
    status_port: 10080
    deploy_dir: /data/deploy/install/deploy/tidb-4000
    log_dir: /data/deploy/install/log/tidb-4000
```

TiKV 配置文件参考：

```
tikv_servers:
  - host: 10.0.1.5
```

```
    ssh_port: 22
    port: 20160
    status_port: 20180
    deploy_dir: /data/deploy/install/deploy/tikv-20160
    data_dir: /data/deploy/install/data/tikv-20160
    log_dir: /data/deploy/install/log/tikv-20160
```

PD 配置文件参考：

```
pd_servers:
  - host: 10.0.1.5
    ssh_port: 22
    name: pd-1
    client_port: 2379
    peer_port: 2380
    deploy_dir: /data/deploy/install/deploy/pd-2379
    data_dir: /data/deploy/install/data/pd-2379
    log_dir: /data/deploy/install/log/pd-2379
```

可以使用 `tiup cluster edit-config <cluster-name>` 查看当前集群的配置信息，因为其中的 `global` 和 `server_configs` 参数配置默认会被 `scale-out.yaml` 继承，因此也会在 `scale-out.yaml` 中生效。

5.1.1.2 2. 执行扩容命令

执行 scale-out 命令前，先使用 check 及 `check --apply` 命令，检查和自动修复集群存在的潜在风险：

> **注意：**
>
> 针对 scale-out 命令的检查功能在 tiup cluster v1.9.0 及后续版本中支持，请操作前先升级 tiup cluster 版本。

（1）检查集群存在的潜在风险：

```
shell tiup cluster check <cluster-name> scale-out.yaml --cluster --user root [-p] [-i /home/root
↪ /.ssh/gcp_rsa]
```

（2）自动修复集群存在的潜在风险：

```
shell tiup cluster check <cluster-name> scale-out.yaml --cluster --apply --user root [-p] [-i /
↪ home/root/.ssh/gcp_rsa]
```

（3）执行 scale-out 命令扩容 TiDB 集群：

```
shell tiup cluster scale-out <cluster-name> scale-out.yaml [-p] [-i /home/root/.ssh/gcp_rsa]
```

以上操作示例中：

- 扩容配置文件为 `scale-out.yaml`。

- `--user root` 表示通过 root 用户登录到目标主机完成集群部署，该用户需要有 ssh 到目标机器的权限，并且在目标机器有 sudo 权限。也可以用其他有 ssh 和 sudo 权限的用户完成部署。
- [-i] 及 [-p] 为可选项，如果已经配置免密登录目标机，则不需填写。否则选择其一即可，[-i] 为可登录到目标机的 root 用户（或 –user 指定的其他用户）的私钥，也可使用 [-p] 交互式输入该用户的密码。

预期日志结尾输出 `Scaled cluster `<cluster-name>` out successfully` 信息，表示扩容操作成功。

5.1.1.3 3. 检查集群状态

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群和新增节点的状态。

扩容后，集群拓扑结构如下所示：

主机 IP	服务
10.0.1.3	TiDB + TiFlash
10.0.1.4	TiDB + PD
10.0.1.5	TiDB + TiKV + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.2 扩容 TiFlash 节点

如果要添加一个 TiFlash 节点，其 IP 地址为 `10.0.1.4`，可以按照如下步骤进行操作。

> **注意：**
>
> 在原有 TiDB 集群上新增 TiFlash 组件需要注意：
>
> 1. 首先确认当前 TiDB 的版本支持 TiFlash，否则需要先升级 TiDB 集群至 v5.0 以上版本。
> 2. 执行 `tiup ctl:<cluster-version> pd -u http://<pd_ip>:<pd_port> config set enable-placement-rules true` 命令，以开启 PD 的 Placement Rules 功能。或通过 pd-ctl 执行对应的命令。

5.1.2.1 1. 添加节点信息到 scale-out.yaml 文件

编写 scale-out.yaml 文件，添加该 TiFlash 节点信息（目前只支持 ip，不支持域名）：

```
tiflash_servers:
  - host: 10.0.1.4
```

5.1.2.2 2. 运行扩容命令

```
tiup cluster scale-out <cluster-name> scale-out.yaml
```

> **注意：**
>
> 此处假设当前执行命令的用户和新增的机器打通了互信，如果不满足已打通互信的条件，需要通过 -p 来输入新机器的密码，或通过 -i 指定私钥文件。

5.1.2.3 3. 查看集群状态

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群和新增节点的状态。

扩容后，集群拓扑结构如下所示：

主机 IP	服务
10.0.1.3	TiDB + TiFlash
10.0.1.4	TiDB + PD + TiFlash
10.0.1.5	TiDB+ TiKV + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.3 扩容 TiCDC 节点

如果要添加 TiCDC 节点，IP 地址为 10.0.1.3、10.0.1.4，可以按照如下步骤进行操作。

5.1.3.1 1. 添加节点信息到 scale-out.yaml 文件

编写 scale-out.yaml 文件：

```
cdc_servers:
  - host: 10.0.1.3
    gc-ttl: 86400
    data_dir: /data/deploy/install/data/cdc-8300
  - host: 10.0.1.4
    gc-ttl: 86400
    data_dir: /data/deploy/install/data/cdc-8300
```

5.1.3.2 2. 运行扩容命令

```
tiup cluster scale-out <cluster-name> scale-out.yaml
```

> **注意：**
>
> 此处假设当前执行命令的用户和新增的机器打通了互信，如果不满足已打通互信的条件，需要通过 -p 来输入新机器的密码，或通过 -i 指定私钥文件。

5.1.3.3 3. 查看集群状态

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群和新增节点的状态。

扩容后，集群拓扑结构如下所示：

主机 IP	服务
10.0.1.3	TiDB + TiFlash + TiCDC
10.0.1.4	TiDB + PD + TiFlash + TiCDC
10.0.1.5	TiDB+ TiKV + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.4 缩容 TiDB/PD/TiKV 节点

如果要移除 IP 地址为 10.0.1.5 的一个 TiKV 节点，可以按照如下步骤进行操作。

> **注意：**
>
> - 移除 TiDB、 PD 节点和移除 TiKV 节点的步骤类似。
> - 由于 TiKV、 TiFlash 和 TiDB Binlog 组件是异步下线的，且下线过程耗时较长，所以 TiUP 对 TiKV、 TiFlash 和 TiDB Binlog 组件做了特殊处理，详情参考下线特殊处理。

> **注意：**
>
> TiKV 中的 PD Client 会缓存 PD 节点的列表。当前版本的 TiKV 有定期自动更新 PD 节点的机制，可以降低 TiKV 缓存的 PD 节点列表过旧这一问题出现的概率。但你应尽量避免在扩容新 PD 后直接一次性缩容所有扩容前就已经存在的 PD 节点。如果需要，请确保在下线所有之前存在的 PD 节点前将 PD 的 leader 切换至新扩容的 PD 节点。

5.1.4.1 1. 查看节点 ID 信息

```
tiup cluster display <cluster-name>
```

```
Starting /root/.tiup/components/cluster/v1.9.0/cluster display <cluster-name>

TiDB Cluster: <cluster-name>

TiDB Version: v5.4.0

ID       Role         Host    Ports                            Status  Data Dir        Deploy Dir

--       ----         ----      -----                            ------  --------
    ↪ ----------

10.0.1.3:8300  cdc          10.0.1.3    8300                           Up      data/cdc-8300
    ↪      deploy/cdc-8300

10.0.1.4:8300  cdc          10.0.1.4    8300                           Up      data/cdc-8300
    ↪      deploy/cdc-8300

10.0.1.4:2379  pd           10.0.1.4    2379/2380                       Healthy data/pd-2379
    ↪      deploy/pd-2379

10.0.1.1:20160 tikv         10.0.1.1    20160/20180                     Up      data/tikv-20160
    ↪     deploy/tikv-20160

10.0.1.2:20160 tikv         10.0.1.2    20160/20180                     Up      data/tikv-20160
    ↪     deploy/tikv-20160

10.0.1.5:20160 tikv        10.0.1.5    20160/20180                    Up      data/tikv-20160
    ↪     deploy/tikv-20160

10.0.1.3:4000  tidb        10.0.1.3    4000/10080                     Up      -
    ↪ deploy/tidb-4000

10.0.1.4:4000  tidb        10.0.1.4    4000/10080                     Up      -
    ↪ deploy/tidb-4000

10.0.1.5:4000  tidb         10.0.1.5    4000/10080                      Up      -
    ↪ deploy/tidb-4000

10.0.1.3:9000   tiflash      10.0.1.3    9000/8123/3930/20170/20292/8234  Up      data/tiflash
    ↪ -9000       deploy/tiflash-9000
```

```
10.0.1.4:9000   tiflash      10.0.1.4    9000/8123/3930/20170/20292/8234  Up      data/tiflash
    ↪ -9000       deploy/tiflash-9000

10.0.1.5:9090  prometheus   10.0.1.5    9090                             Up      data/prometheus
    ↪ -9090  deploy/prometheus-9090

10.0.1.5:3000  grafana      10.0.1.5    3000                             Up      -
    ↪ deploy/grafana-3000

10.0.1.5:9093  alertmanager 10.0.1.5    9093/9094                        Up      data/
    ↪ alertmanager-9093 deploy/alertmanager-9093
```

5.1.4.2 2. 执行缩容操作

```
tiup cluster scale-in <cluster-name> --node 10.0.1.5:20160
```

其中 --node 参数为需要下线节点的 ID。

预期输出 Scaled cluster <cluster-name> in successfully 信息，表示缩容操作成功。

5.1.4.3 3. 检查集群状态

下线需要一定时间，下线节点的状态变为 Tombstone 就说明下线成功。

执行如下命令检查节点是否下线成功：

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群的状态。

调整后，拓扑结构如下：

Host IP	Service
10.0.1.3	TiDB + TiFlash + TiCDC
10.0.1.4	TiDB + PD + TiFlash + TiCDC
10.0.1.5	TiDB + Monitor（TiKV 已删除）
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.5 缩容 TiFlash 节点

如果要缩容 IP 地址为 10.0.1.4 的一个 TiFlash 节点，可以按照如下步骤进行操作。

5.1.5.1 1. 根据 TiFlash 剩余节点数调整数据表的副本数

在下线节点之前，确保 TiFlash 集群剩余节点数大于等于所有数据表的最大副本数，否则需要修改相关表的

TiFlash 副本数。

1. 在 TiDB 客户端中针对所有副本数大于集群剩余 TiFlash 节点数的表执行：

```
alter table <db-name>.<table-name> set tiflash replica 0;
```

2. 等待相关表的 TiFlash 副本被删除（按照查看表同步进度一节操作，查不到相关表的同步信息时即为副本被删除）。

5.1.5.2 2. 执行缩容操作

接下来，请任选下列方案其一进行缩容。

5.1.5.2.1 方案一：通过 TiUP 缩容 TiFlash 节点

1. 通过以下命令确定需要下线的节点名称：

```
tiup cluster display <cluster-name>
```

2. 执行 scale-in 命令来下线节点，假设步骤 1 中获得该节点名为 10.0.1.4:9000

```
tiup cluster scale-in <cluster-name> --node 10.0.1.4:9000
```

5.1.5.2.2 方案二：手动缩容 TiFlash 节点

在特殊情况下（比如需要强制下线节点），或者 TiUP 操作失败的情况下，可以使用以下方法手动下线 TiFlash 节点。

1. 使用 pd-ctl 的 store 命令在 PD 中查看该 TiFlash 节点对应的 store id。

    - 在 pd-ctl（tidb-ansible 目录下的 `resources/bin` 包含对应的二进制文件）中输入 store 命令。
    - 若使用 TiUP 部署，可以调用以下命令代替 `pd-ctl`：

```
tiup ctl:<cluster-version> pd -u http://<pd_ip>:<pd_port> store
```

> **注意：**
> 如果集群中有多个 PD 实例，只需在以上命令中指定一个活跃 PD 实例的 `IP:端口` 即可。

2. 在 pd-ctl 中下线该 TiFlash 节点。

    - 在 pd-ctl 中输入 `store delete <store_id>`，其中 `<store_id>` 为上一步查到的该 TiFlash 节点对应的 store id。
    - 若通过 TiUP 部署，可以调用以下命令代替 `pd-ctl`：

```
tiup ctl:<cluster-version> pd -u http://<pd_ip>:<pd_port> store delete <store_id>
```

> **注意：**
> 如果集群中有多个 PD 实例，只需在以上命令中指定一个活跃 PD 实例的 IP:端口即可。

3. 等待该 TiFlash 节点对应的 store 消失或者 state_name 变成 Tombstone 再关闭 TiFlash 进程。
4. 手动删除 TiFlash 的数据文件，具体位置可查看在集群拓扑配置文件中 TiFlash 配置部分下的 data_dir 目录。
5. 手动更新 TiUP 的集群配置文件，在编辑模式中手动删除我们已经下线的 TiFlash 节点信息：

```
tiup cluster edit-config <cluster-name>
```

> **注意：**
> 如果在集群中所有的 TiFlash 节点停止运行之前，没有取消所有同步到 TiFlash 的表，则需要手动在 PD 中清除同步规则，否则无法成功完成 TiFlash 节点的下线。

手动在 PD 中清除同步规则的步骤如下：

1. 查询当前 PD 实例中所有与 TiFlash 相关的数据同步规则。

```
curl http://<pd_ip>:<pd_port>/pd/api/v1/config/rules/group/tiflash
```

```
[
  {
    "group_id": "tiflash",
    "id": "table-45-r",
    "override": true,
    "start_key": "7480000000000000FF2D5F720000000000FA",
    "end_key": "7480000000000000FF2E00000000000000F8",
    "role": "learner",
    "count": 1,
    "label_constraints": [
      {
        "key": "engine",
        "op": "in",
        "values": [
          "tiflash"
        ]
      }
    ]
  }
]
```

2. 删除所有与 TiFlash 相关的数据同步规则。以 id 为 table-45-r 的规则为例，通过以下命令可以删除该规则。

```
curl -v -X DELETE http://<pd_ip>:<pd_port>/pd/api/v1/config/rule/tiflash/table-45-r
```

5.1.5.3 3. 查看集群状态

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群的状态。

调整后，拓扑结构如下：

Host IP	Service
10.0.1.3	TiDB + TiFlash + TiCDC
10.0.1.4	TiDB + PD + TiCDC （ TiFlash 已删除 ）
10.0.1.5	TiDB + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

5.1.6 缩容 TiCDC 节点

如果要缩容 IP 地址为 10.0.1.4 的一个 TiCDC 节点，可以按照如下步骤进行操作。

5.1.6.1 1. 下线该 TiCDC 节点

```
tiup cluster scale-in <cluster-name> --node 10.0.1.4:8300
```

5.1.6.2 2. 查看集群状态

```
tiup cluster display <cluster-name>
```

打开浏览器访问监控平台 http://10.0.1.5:3000，监控整个集群的状态。

调整后，拓扑结构如下：

Host IP	Service
10.0.1.3	TiDB + TiFlash + TiCDC
10.0.1.4	TiDB + PD + (TiCDC 已删除)
10.0.1.5	TiDB + Monitor
10.0.1.1	TiKV
10.0.1.2	TiKV

6 例行维护

6.1 日常巡检

TiDB 作为分布式数据库，对比单机数据库机制更加复杂，其自带的监控项也很丰富。为了更便捷地运维 TiDB，本文介绍了运维 TiDB 集群需要常关注的关键性能指标。



图 10: 实例面板

以上实例面板的各指标说明如下：

- 状态：用于查看实例状态是否正常，如果在线可忽略此项。
- 启动时间：是关键指标。如果有发现启动时间有变动，那么需要排查组件重启的原因。
- 版本、部署路径、Git 哈希值：通过查看这三个指标可以避免有部署路径和版本不一致或者错误的情况。

6.1.1.2 主机面板



图 11: 主机面板

通过主机面板可以查看 CPU、内存、磁盘使用率。当任何资源的使用率超过 80% 时，推荐扩容对应组件。

6.1.1.3 SQL 分析面板



图 12: SQL 分析面板

通过 SQL 分析面板可以分析对集群影响较大的慢 SQL，然后进行对应的 SQL 优化。

6.1.1.4 Region 信息面板

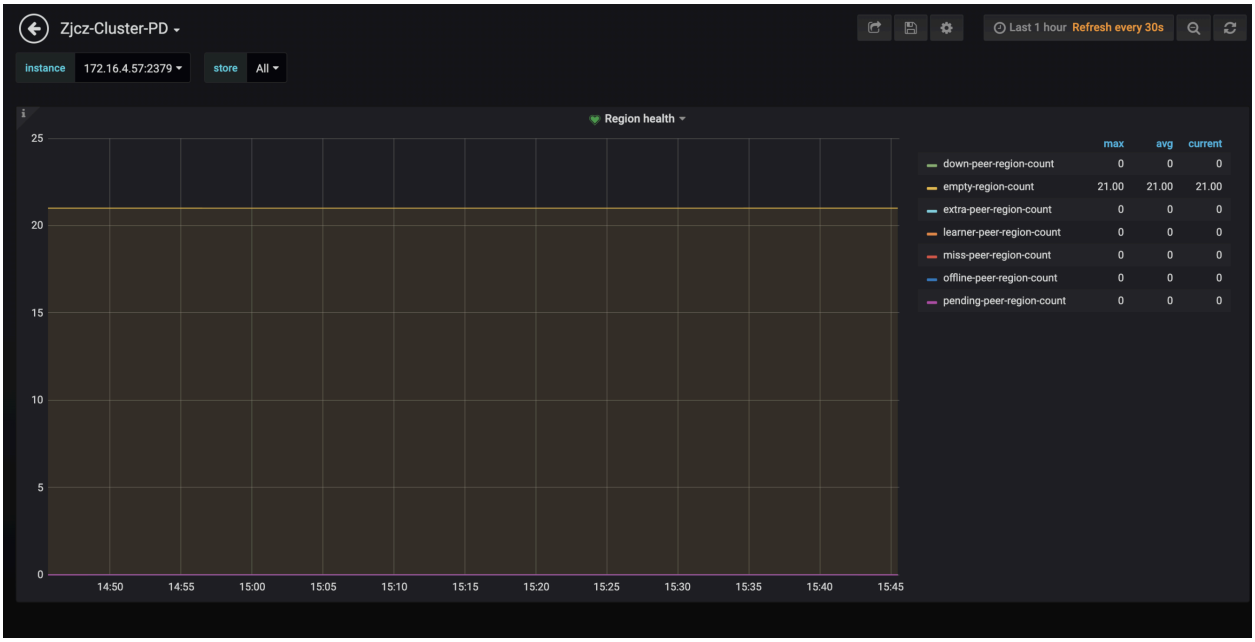


图 13: Region 信息面板

以上 Region 信息面板说明如下：

- `miss-peer-region-count`：缺副本的 Region 数量，不会一直大于 0。
- `extra-peer-region-count`：多副本的 Region 数量，调度过程中会产生。
- `empty-region-count`：空 Region 的数量，一般是 `TRUNCATE TABLE/DROP TABLE` 语句导致。如果数量较多，可以考虑开启跨表 Region merge。
- `pending-peer-region-count`：Raft log 落后的 Region 数量。由于调度产生少量的 pending peer 是正常的，但是如果持续很高，可能有问题。
- `down-peer-region-count`：Raft leader 上报有不响应 peer 的 Region 数量。
- `offline-peer-region-count`：peer 下线过程中的 Region 数量。

原则上来说，该监控面板偶尔有数据是符合预期的。但长期有数据，需要排查是否存在问题。

6.1.1.5 KV Request Duration

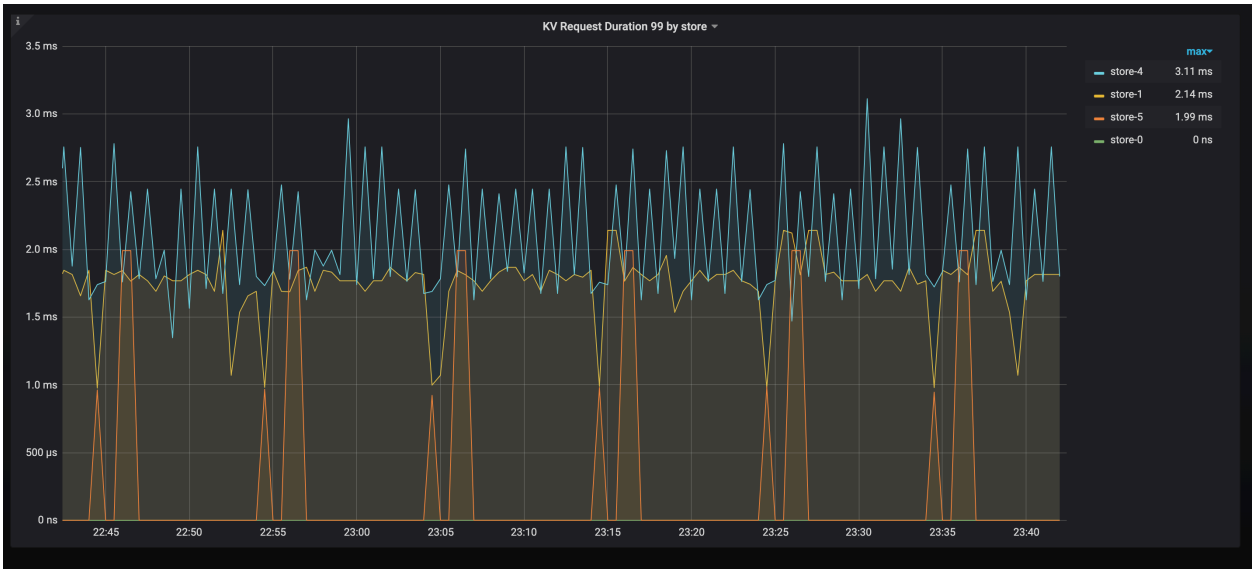


图 14: TiKV 相应时间

TiKV 当前.99（百分位）的响应时间。如果发现有明显高的节点，可以排查是否有热点，或者相关节点性能较差。

6.1.1.6 PD TSO Wait Duration

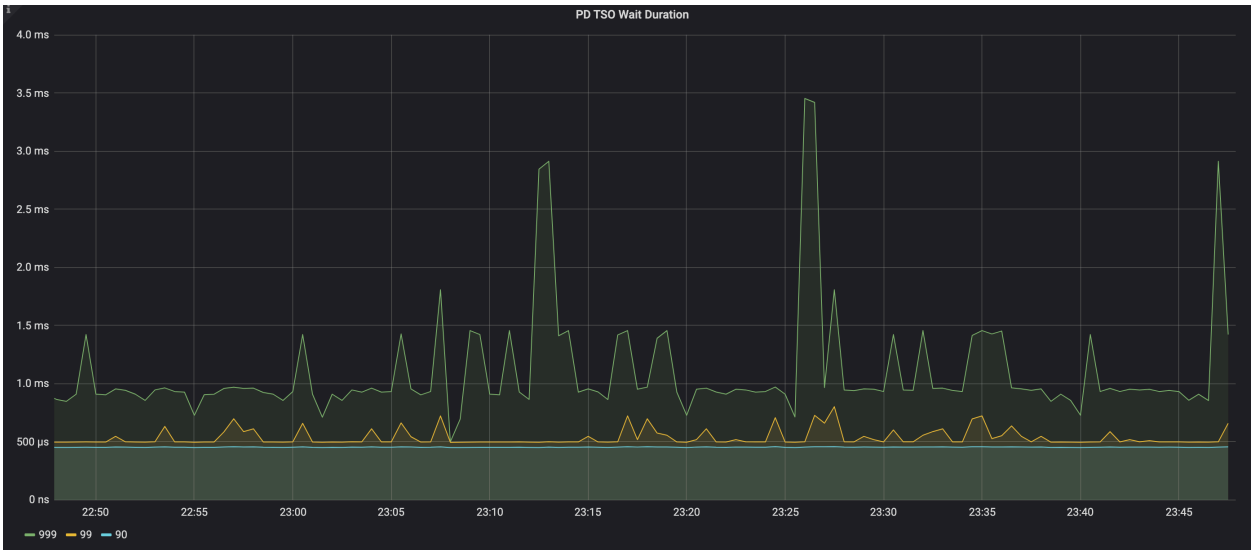


图 15: TiDB 从 PD 获取 TSO 的时间

TiDB 从 PD 获取 TSO 的时间。如果相关响应时间较高，一般常见原因如下：

- TiDB 到 PD 的网络延迟较高，可以手动 Ping 一下网络延迟。

- TiDB 压力较高，导致获取较慢。
- PD 服务器压力较高，导致获取较慢。

6.1.1.7 Overview 面板

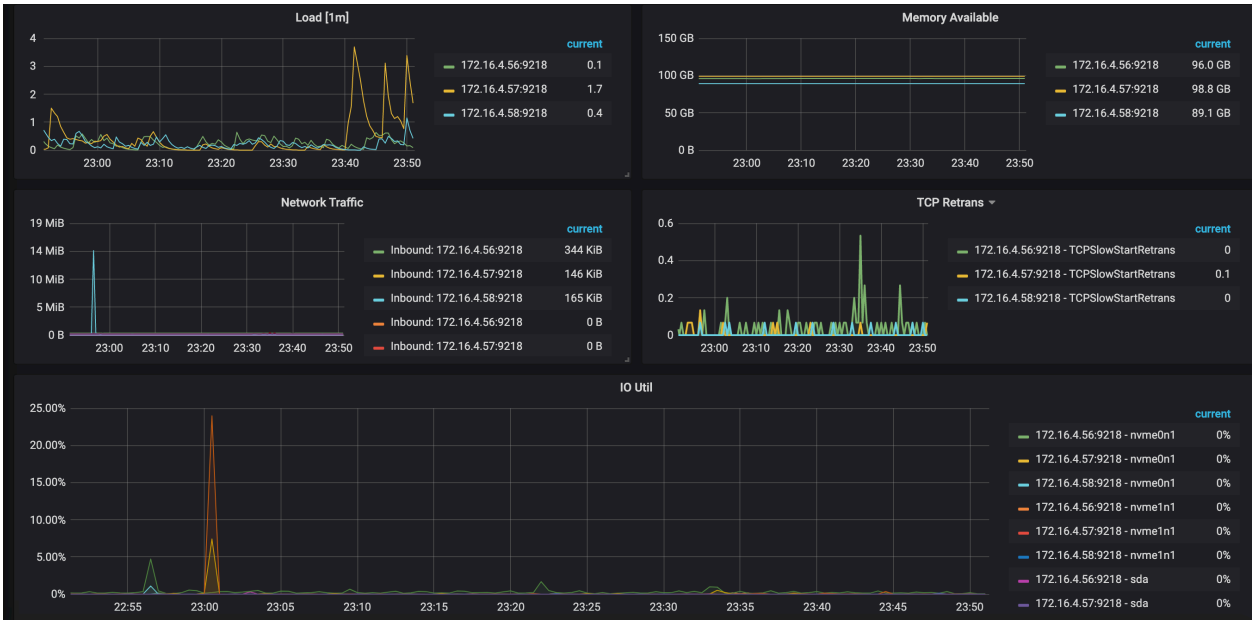


图 16: Overview 面板

以上面板展示常见的负载、内存、网络、IO 监控。发现有瓶颈时，推荐扩容或者优化集群拓扑，优化 SQL、集群参数等。

6.1.1.8 异常监控面板

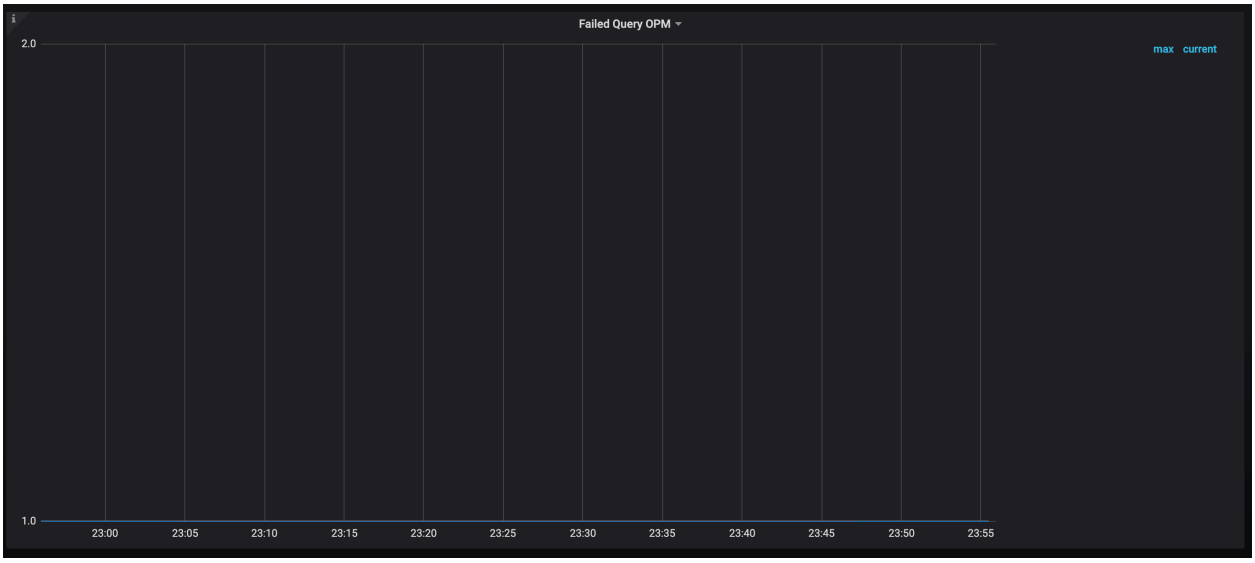


图 17: 异常监控面板

以上面板展示每个 TiDB 实例上，执行 SQL 语句发生的错误，并按照错误类型进行统计，例如语法错误、主键冲突等。

6.1.1.9 GC 状态面板

```
(root@127.0.0.1) [(none)]>select * from mysql.tidb;
+-----------------------+-----------------------------------------------------------------------------------------------------------+-----------------------------------------------------------------------------------------------------------+
| VARIABLE_NAME         | VARIABLE_VALUE                                                                                            | COMMENT                                                                                                   |
+-----------------------+-----------------------------------------------------------------------------------------------------------+-----------------------------------------------------------------------------------------------------------+
| bootstrapped          | True                                                                                                      | Bootstrap flag. Do not delete.                                                                            |
| tidb_server_version   | 44                                                                                                        | Bootstrap version. Do not delete.                                                                         |
| system_tz             | Asia/Shanghai                                                                                             | TiDB Global System Timezone.                                                                              |
| new_collation_enabled | False                                                                                                     | If the new collations are enabled. Do no                                                                  |
| tikv_gc_leader_uuid   | 5c8c8865ca80006                                                                                           | Current GC worker leader UUID. (DO NOT                                                                    |
| tikv_gc_leader_desc   | host:wangjundeMacBook-Pro.local, pid:81322, start at 2020-05-20 20:49:21.735312 +0800 CST m=+7.904021075 | Host name and pid of current GC leader.                                                                   |
| tikv_gc_leader_lease  | 20200522-00:06:30 +0800                                                                                   | Current GC worker leader lease. (DO NOT                                                                   |
| tikv_gc_enable        | true                                                                                                      | Current GC enable status                                                                                  |
| tikv_gc_run_interval  | 10m0s                                                                                                     | GC run interval, at least 10m, in Go fo                                                                   |
| tikv_gc_life_time     | 10m0s                                                                                                     | All versions within life time will not                                                                    |
| tikv_gc_last_run_time | 20200521-23:55:30 +0800                                                                                   | The time when last GC starts. (DO NOT E                                                                   |
| tikv_gc_safe_point    | 20200521-23:45:30 +0800                                                                                   | All versions after safe point can be ac                                                                   |
| tikv_gc_auto_concurrency | true                                                                                                   | Let TiDB pick the concurrency automatica                                                                  |
| tikv_gc_mode          | distributed                                                                                               | Mode of GC, "central" or "distributed"                                                                    |
+-----------------------+-----------------------------------------------------------------------------------------------------------+-----------------------------------------------------------------------------------------------------------+
14 rows in set (0.01 sec)
```

图 18: GC 状态面板

以上面板展示最后 GC（垃圾清理）的时间，观察 GC 是否正常。如果 GC 发生异常，可能会造成历史数据存留过多，影响访问效率。

6.2 时区支持

TiDB 使用的时区由 `time_zone` 全局变量和 session 变量决定。`time_zone` 的默认值是 `System`，`System` 对应的实际时区在 TiDB 集群 bootstrap 初始化时设置。具体逻辑如下：

- 优先使用 TZ 环境变量
- 如果失败，则从 /etc/localtime 的实际软链地址提取。
- 如果上面两种都失败则使用 UTC 作为系统时区。

在运行过程中可以修改全局时区：

```
SET GLOBAL time_zone = timezone;
```

TiDB 还可以通过设置 session 变量 time_zone 为每个连接维护各自的时区。默认条件下，这个值取的是全局变量 time_zone 的值。修改 session 使用的时区：

```
SET time_zone = timezone;
```

使用以下 SQL 语句查看当前全局时区、客户端时区和系统时区的值：

```
SELECT @@global.time_zone, @@session.time_zone, @@global.system_time_zone;
```

设置 time_zone 的值的格式：

- ‘SYSTEM’ 表明使用系统时间
- 相对于 UTC 时间的偏移，比如 ‘+10:00’ 或者 ‘-6:00’
- 某个时区的名字，比如 ‘Europe/Helsinki’，‘US/Eastern’ 或 ‘MET’

NOW() 和 CURTIME() 的返回值都受到时区设置的影响。

> 注意：
>
> 只有 Timestamp 数据类型的值是受时区影响的。可以理解为，Timestamp 数据类型的实际表示使用的是（字面值 + 时区信息）。其它时间和日期类型，比如 Datetime/Date/Time 是不包含时区信息的，所以也不受到时区变化的影响。

```
create table t (ts timestamp, dt datetime);
```

```
Query OK, 0 rows affected (0.02 sec)
```

```
set @@time_zone = 'UTC';
```

```
Query OK, 0 rows affected (0.00 sec)
```

```
insert into t values ('2017-09-30 11:11:11', '2017-09-30 11:11:11');
```

```
Query OK, 1 row affected (0.00 sec)
```

```
set @@time_zone = '+8:00';
```

```
Query OK, 0 rows affected (0.00 sec)
```

```
select * from t;
```

```
+---------------------|---------------------+
| ts                  | dt                  |
+---------------------|---------------------+
| 2017-09-30 19:11:11 | 2017-09-30 11:11:11 |
+---------------------|---------------------+
1 row in set (0.00 sec)
```

上面的例子中，无论怎么调整时区的值，Datetime 类型字段的值是不受影响的，而 Timestamp 则随着时区改变，显示的值会发生变化。其实 Timestamp 持久化到存储的值始终没有变化过，只是根据时区的不同显示值不同。

Timestamp 类型和 Datetime 等类型的值，两者相互转换的过程中，会涉及到时区。这种情况一律基于 session 的当前 `time_zone` 时区处理。

另外，用户在导数据的过程中，也要需注意主库和从库之间的时区设定是否一致。

6.3 TiFlash 集群运维

本文介绍 TiFlash 集群运维的一些常见操作，包括查看 TiFlash 版本、TiFlash 重要日志及系统表。

6.3.1 查看 TiFlash 版本

查看 TiFlash 版本有以下两种方法：

- 假设 TiFlash 的二进制文件名为 `tiflash`，则可以通过 `./tiflash version` 方式获取 TiFlash 版本。

  但是由于 TiFlash 的运行依赖于动态库 `libtiflash_proxy.so`，因此需要将包含动态库 `libtiflash_proxy.so` 的目录路径添加到环境变量 `LD_LIBRARY_PATH` 后，上述命令才能正常执行。

  例如，当 `tiflash` 和 `libtiflash_proxy.so` 在同一个目录下时，切换到该目录后，可以通过如下命令查看 TiFlash 版本：

  ```
  LD_LIBRARY_PATH=./ ./tiflash version
  ```

- 在 TiFlash 日志（日志路径见配置文件 tiflash.toml [logger] 部分）中查看 TiFlash 版本，例如：

  ```
  <information>: TiFlash version: TiFlash 0.2.0 master-375035282451103999f3863c691e2fc2
  ```

6.3.2 TiFlash 重要日志介绍

日志信息	日志含义
[INFO] [<unknown>] [“KVStore: Start to persist [region 47, applied: term 6 index 10]”] [thread_id=23]	在 TiFlash 中看到类似日志代表数据开始同步
[DEBUG] [<unknown>] [“CoprocessorHandler: grpc::Status DB::CoprocessorHandler::execute(): Handling DAG request”] [thread_id=30]	该日志代表 TiFlash 开始处理一个 Coprocessor 请求
[DEBUG] [<unknown>] [“CoprocessorHandler: grpc::Status DB::CoprocessorHandler::execute(): Handle DAG request done”] [thread_id=30]	该日志代表 TiFlash 完成 Coprocessor 请求的处理

你可以找到一个 Coprocessor 请求的开始或结束，然后通过日志前面打印的线程号找到该 Coprocessor 请求的其他相关日志。

6.3.3 TiFlash 系统表

`information_schema.tiflash_replica` 系统表的列名及含义如下：

列名	含义
TABLE_SCHEMA	数据库名
TABLE_NAME	表名
TABLE_ID	表 ID
REPLICA_COUNT	TiFlash 副本数
AVAILABLE	是否可用（0/1）
PROGRESS	同步进度 [0.0~1.0]

6.4 TiUP 常见运维操作

本文介绍了使用 TiUP 运维 TiDB 集群的常见操作，包括查看集群列表、启动集群、查看集群状态、修改配置参数、关闭集群、销毁集群等。

6.4.1 查看集群列表

TiUP cluster 组件可以用来管理多个 TiDB 集群，在每个 TiDB 集群部署完毕后，该集群会出现在 TiUP 的集群列表里，可以使用 list 命令来查看。

```
tiup cluster list
```

6.4.2 启动集群

启动集群操作会按 PD -> TiKV -> Pump -> TiDB -> TiFlash -> Drainer -> TiCDC -> Prometheus -> Grafana -> Alertmanager 的顺序启动整个 TiDB 集群所有组件：

```
tiup cluster start ${cluster-name}
```

> 注意：
>
> 你需要将 `${cluster-name}` 替换成实际的集群名字，若忘记集群名字，可通过 `tiup cluster list` 查看。

该命令支持通过 `-R` 和 `-N` 参数来只启动部分组件。

例如，下列命令只启动 PD 组件：

```
tiup cluster start ${cluster-name} -R pd
```

下列命令只启动 `1.2.3.4` 和 `1.2.3.5` 这两台机器上的 PD 组件：

```
tiup cluster start ${cluster-name} -N 1.2.3.4:2379,1.2.3.5:2379
```

> 注意：
>
> 若通过 `-R` 和 `-N` 启动指定组件，需要保证启动顺序正确（例如需要先启动 PD 才能启动 TiKV），否则可能导致启动失败。

6.4.3 查看集群状态

集群启动之后需要检查每个组件的运行状态，以确保每个组件工作正常。TiUP 提供了 display 命令，节省了登录到每台机器上去查看进程的时间。

```
tiup cluster display ${cluster-name}
```

6.4.4 修改配置参数

集群运行过程中，如果需要调整某个组件的参数，可以使用 `edit-config` 命令来编辑参数。具体的操作步骤如下：

1. 以编辑模式打开该集群的配置文件：

```
tiup cluster edit-config ${cluster-name}
```

2. 设置参数：

   首先确定配置的生效范围，有以下两种生效范围：

   - 如果配置的生效范围为该组件全局，则配置到 `server_configs`。例如：

```
server_configs:
  tidb:
    log.slow-threshold: 300
```

   - 如果配置的生效范围为某个节点，则配置到具体节点的 `config` 中。例如：

```
tidb_servers:
- host: 10.0.1.11
    port: 4000
    config:
        log.slow-threshold: 300
```

   参数的格式参考 TiUP 配置参数模版。

   配置项层次结构使用 . 表示。

   关于组件的更多配置参数说明，可参考 tidb `config.toml.example`、tikv `config.toml.example` 和 pd `config.toml.example`。

3. 执行 `reload` 命令滚动分发配置、重启相应组件：

```
tiup cluster reload ${cluster-name} [-N <nodes>] [-R <roles>]
```

6.4.4.1 示例

如果要调整 tidb-server 中事务大小限制参数 `txn-total-size-limit` 为 1G，该参数位于 performance 模块下，调整后的配置如下：

```
server_configs:
  tidb:
    performance.txn-total-size-limit: 1073741824
```

然后执行 `tiup cluster reload ${cluster-name} -R tidb` 命令滚动重启。

6.4.5 Hotfix 版本替换

常规的升级集群请参考升级文档，但是在某些场景下（例如 Debug），可能需要用一个临时的包替换正在运行的组件，此时可以用 patch 命令：

```
tiup cluster patch --help
```

```
Replace the remote package with a specified package and restart the service

Usage:
  tiup cluster patch <cluster-name> <package-path> [flags]

Flags:
  -h, --help                    帮助信息
  -N, --node strings            指定被替换的节点
      --overwrite               在未来的 scale-out 操作中使用当前指定的临时包
  -R, --role strings            指定被替换的服务类型
      --transfer-timeout int    transfer leader 的超时时间

Global Flags:
      --native-ssh        使用系统默认的 SSH 客户端
      --wait-timeout int  等待操作超时的时间
      --ssh-timeout int   SSH 连接的超时时间
  -y, --yes               跳过所有的确认步骤
```

例如，有一个 TiDB 实例的 hotfix 包放在 /tmp/tidb-hotfix.tar.gz 目录下。如果此时想要替换集群上的所有 TiDB 实例，则可以执行以下命令：

```
tiup cluster patch test-cluster /tmp/tidb-hotfix.tar.gz -R tidb
```

或者只替换其中一个 TiDB 实例：

```
tiup cluster patch test-cluster /tmp/tidb-hotfix.tar.gz -N 172.16.4.5:4000
```

6.4.6 重命名集群

部署并启动集群后，可以通过 tiup cluster rename 命令来对集群重命名：

```
tiup cluster rename ${cluster-name} ${new-name}
```

注意：

- 重命名集群会重启监控（Prometheus 和 Grafana）。
- 重命名集群之后 Grafana 可能会残留一些旧集群名的面板，需要手动删除这些面板。

6.4.7 关闭集群

关闭集群操作会按 Alertmanager -> Grafana -> Prometheus -> TiCDC -> Drainer -> TiFlash -> TiDB -> Pump -> TiKV -> PD 的顺序关闭整个 TiDB 集群所有组件（同时也会关闭监控组件）：

```
tiup cluster stop ${cluster-name}
```

和 `start` 命令类似，`stop` 命令也支持通过 `-R` 和 `-N` 参数来只停止部分组件。

例如，下列命令只停止 TiDB 组件：

```
tiup cluster stop ${cluster-name} -R tidb
```

下列命令只停止 `1.2.3.4` 和 `1.2.3.5` 这两台机器上的 TiDB 组件：

```
tiup cluster stop ${cluster-name} -N 1.2.3.4:4000,1.2.3.5:4000
```

6.4.8 清除集群数据

此操作会关闭所有服务，并清空其数据目录或/和日志目录，并且无法恢复，需要谨慎操作。

清空集群所有服务的数据，但保留日志：

```
tiup cluster clean ${cluster-name} --data
```

清空集群所有服务的日志，但保留数据：

```
tiup cluster clean ${cluster-name} --log
```

清空集群所有服务的数据和日志：

```
tiup cluster clean ${cluster-name} --all
```

清空 Prometheus 以外的所有服务的日志和数据：

```
tiup cluster clean ${cluster-name} --all --ignore-role prometheus
```

清空节点 `172.16.13.11:9000` 以外的所有服务的日志和数据：

```
tiup cluster clean ${cluster-name} --all --ignore-node 172.16.13.11:9000
```

清空部署在 `172.16.13.12` 以外的所有服务的日志和数据：

```
tiup cluster clean ${cluster-name} --all --ignore-node 172.16.13.12
```

6.4.9 销毁集群

销毁集群操作会关闭服务，清空数据目录和部署目录，并且无法恢复，需要谨慎操作。

```
tiup cluster destroy ${cluster-name}
```

6.5 在线修改集群配置

> **警告：**
>
> 该功能目前是实验性阶段，不建议在生产环境中使用。

在线配置变更主要是通过利用 SQL 对包括 TiDB、TiKV 以及 PD 在内的各组件的配置进行在线更新。用户可以通过在线配置变更对各组件进行性能调优而无需重启集群组件。但目前在线修改 TiDB 实例配置的方式和修改其他组件 (TiKV, PD) 的有所不同。

6.5.1 常用操作

6.5.1.1 查看实例配置

可以通过 SQL 语句 `show config` 来直接查看集群所有实例的配置信息，结果如下：

```
show config;
```

```
+--
    ↪ ----+-----------------+-----------------------------------------------------------------+-----------------------
    ↪
| Type | Instance        | Name                                                            | Value
    ↪
    ↪ |
+--
    ↪ ----+-----------------+-----------------------------------------------------------------+-----------------------
    ↪
| tidb | 127.0.0.1:4001  | advertise-address                                               | 127.0.0.1
    ↪
    ↪ |
| tidb | 127.0.0.1:4001  | binlog.binlog-socket                                            |
    ↪
    ↪ |
| tidb | 127.0.0.1:4001  | binlog.enable                                                   | false
    ↪
    ↪ |
| tidb | 127.0.0.1:4001  | binlog.ignore-error                                             | false
    ↪
    ↪ |
| tidb | 127.0.0.1:4001  | binlog.strategy                                                 | range
    ↪
    ↪ |
```

```
| tidb | 127.0.0.1:4001  | binlog.write-timeout                                    | 15s
    ↪
    ↪ |
| tidb | 127.0.0.1:4001  | check-mb4-value-in-utf8                                 | true
    ↪
    ↪ |

...
```

还可以根据对应的字段进行过滤，如：

```
show config where type='tidb'
show config where instance in (...)
show config where name like '%log%'
show config where type='tikv' and name='log.level'
```

6.5.1.2 在线修改 TiKV 配置

> 注意：
>
> 在线修改 TiKV 配置项后，同时会自动修改 TiKV 的配置文件。但还需要使用 `tiup edit-config` 命令来修改对应的配置项，否则 `upgrade` 和 `reload` 等运维操作会将在线修改配置后的结果覆盖。修改配置的操作请参考：使用 TiUP 修改配置。执行 `tiup edit-config` 后不需要执行 `tiup reload` 操作。

执行 SQL 语句 `set config`，可以结合实例地址或组件类型来修改单个实例配置或全部实例配置，如：

修改全部 TiKV 实例配置：

> 注意：
>
> 建议使用反引号包裹变量名称。

```
set config tikv `split.qps-threshold`=1000
```

修改单个 TiKV 实例配置：

```
set config "127.0.0.1:20180" `split.qps-threshold`=1000
```

设置成功会返回 `Query OK`：

```
Query OK, 0 rows affected (0.01 sec)
```

在批量修改时如果有错误发生，会以 warning 的形式返回：

```
set config `tikv log-level`='warn';
```

```
Query OK, 0 rows affected, 1 warning (0.04 sec)
```

```
show warnings;
```

```
+--
    ↪ -------+------+-----------------------------------------------------------------------------------------------
    ↪
| Level   | Code | Message
    ↪
    ↪ |
+--
    ↪ -------+------+-----------------------------------------------------------------------------------------------
    ↪
| Warning | 1105 | bad request to http://127.0.0.1:20180/config: fail to update, error: "config
    ↪ log-level can not be changed" |
+--
    ↪ -------+------+-----------------------------------------------------------------------------------------------
    ↪
1 row in set (0.00 sec)
```

批量修改配置不保证原子性，可能出现某些实例成功，而某些失败的情况。如使用 `set tikv key=val` 命令修改整个 TiKV 集群配置时，可能有部分实例失败，请执行 `show warnings` 进行查看。

如遇到部分修改失败的情况，需要重新执行对应的修改语句，或通过修改单个实例的方式完成修改。如果因网络或者机器故障等原因无法访问到的 TiKV，需要等到恢复后再次进行修改。

针对 TiKV 可在线修改的配置项，如果成功修改后，修改的结果会被持久化到配置文件中，后续以配置文件中的配置为准。某些配置项名称可能和 TiDB 预留关键字冲突，如 `limit`、`key` 等，对于此类配置项，需要用反引号 ` 包裹起来，如 `` `raftstore.raft-log-gc-size-limit` ``。

支持的配置项列表如下：

配置项	简介
raftstore.raft-entry-max-size	单个日志最大大小

配置项	简介
raftstore.raft-log-gc-tick-interval	删除 Raft 日志的轮询任务调度间隔时间
raftstore.raft-log-gc-threshold	允许残余的 Raft 日志个数，软限制
raftstore.raft-log-gc-count-limit	允许残余的 Raft 日志个数，硬限制
raftstore.raft-log-gc-size-limit	允许残余的 Raft 日志大小，硬限制
raftstore.raft-entry-cache-life-time	内存中日志 cache 允许的最长残留时间

配置项	简介
raftstore.region-check-tick-interval	检查 Region 是否需要分裂的时间间隔
raftstore.region-split-check-diff	允许 Region 数据超过指定大小的最大值
raftstore.region-compact-check-interval	检查是否需要人工触发 RocksDB compaction 的时间间隔
raftstore.region-compact-check-step	每轮校验人工 compaction 时，一次性检查的 Region 个数

配置项	简介
raftstore.region-compact-min-tombstones	触发 RocksDB compaction 需要的 tombstone 个数
raftstore.region-compact-tombstones-percent	触发 RocksDB compaction 需要的 tombstone 所占比例
raftstore.pd-heartbeat-tick-interval	触发 Region 对 PD 心跳的时间间隔
raftstore.pd-store-heartbeat-tick-interval	触发 store 对 PD 心跳的时间间隔

配置项	简介
raftstore.snap-mgr-gc-tick-interval	触发回收过期 snapshot 文件的时间间隔
raftstore.snap-gc-timeout	snapshot 文件的最长保存时间
raftstore.lock-cf-compact-interval	触发对 lock CF 进行 compact 检查的时间间隔
raftstore.lock-cf-compact-bytes-threshold	触发对 lock CF 进行 compact 的大小
raftstore.messages-per-tick	每轮处理的消息最大个数

配置项	简介
raftstore.max-peer-down-duration	副本允许的最长未响应时间
raftstore.max-leader-missing-duration	允许副本处于无主状态的最长时间，超过将会向 PD 校验自己是否已经被删除
raftstore.abnormal-leader-missing-duration	允许副本处于无主状态的时间，超过将视为异常，标记在 metrics 和日志中

配置项	简介
raftstore.peer-stale-state-check-interval	触发检验副本是否处于无主状态的时间间隔
raftstore.consistency-check-interval	触发一致性检查的时间间隔
raftstore.region-store-max-leader-lease	主可信任期的最长时间
raftstore.merge-check-tick-interval	触发 Merge 完成检查的时间间隔
raftstore.cleanup-import-sst-interval	触发检查过期 SST 文件的时间间隔

配置项	简介
raftstore.local-read-batch-size	一轮处理读请求的最大个数
raftstore.hibernate-timeout	启动后进入静默状态前需要等待的最短时间，在该时间段内不会进入静默状态（未 release）
raftstore.apply-pool-size	apply 线程池大小
raftstore.store-pool-size	store 线程池大小

配置项	简介
coprocessor.split-region-on-table	开启按 table 分裂 Region 的开关
coprocessor.batch-split-limit	批量分裂 Region 的阈值
coprocessor.region-max-size	Region 容量空间的最大值
coprocessor.region-split-size	分裂后新 Region 的大小
coprocessor.region-max-keys	Region 最多允许的 key 的个数
coprocessor.region-split-keys	分裂后新 Region 的 key 的个数

配置项	简介
pessimistic-txn.wait-for-lock-timeout	悲观事务遇到锁后的最长等待时间
pessimistic-txn.wake-up-delay-duration	悲观事务被重新唤醒的时间
pessimistic-txn.pipelined	是否开启流水线式加悲观锁流程
gc.ratio-threshold	跳过 Region GC 的阈值（GC 版本个数/key 个数）
gc.batch-keys	一轮处理 key 的个数

配置项	简介
gc.max-write-bytes-per-sec	一秒可写入 RocksDB 的最大字节数
gc.enable-compaction-filter	是否使用 compaction filter
gc.compaction-filter-skip-version-check	是否跳过 compaction filter 的集群版本检查（未 release）
{db-name}.max-total-wal-size	WAL 总大小限制
{db-name}.max-background-jobs	RocksDB 后台线程个数
{db-name}.max-background-flushes	RocksDB flush 线程个数

配置项	简介
{db-name}.max-open-files	RocksDB 可以打开的文件总数
{db-name}.compaction-readahead-size	Compaction 时候 readahead 的大小
{db-name}.bytes-per-sync	异步同步的限速速率
{db-name}.wal-bytes-per-sync	WAL 同步的限速速率
{db-name}.writable-file-max-buffer-size	WritableFileWrite 所使用的最大的 buffer 大小
{db-name}.{cf-name}.block-cache-size	block cache 大小
{db-name}.{cf-name}.write-buffer-size	memtable 大小

配置项	简介
{db-name}.{cf-name}.max-write-buffer-number	最大 memtable 个数
{db-name}.{cf-name}.max-bytes-for-level-base	base level (L1) 最大字节数
{db-name}.{cf-name}.target-file-size-base	base level 的目标文件大小
{db-name}.{cf-name}.level0-file-num-compaction-trigger	触发 compaction 的 L0 文件最大个数
{db-name}.{cf-name}.level0-slowdown-writes-trigger	触发 write stall 的 L0 文件最大个数
{db-name}.{cf-name}.level0-stop-writes-trigger	完全阻停写入的 L0 文件最大个数

配置项	简介
{db-name}.{cf-name}.max-compaction-bytes	一次 compaction 最大写入字节数
{db-name}.{cf-name}.max-bytes-for-level-multiplier	每一层的默认放大倍数
{db-name}.{cf-name}.disable-auto-compactions	自动 compaction 的开关
{db-name}.{cf-name}.soft-pending-compaction-bytes-limit	pending compaction bytes 的软限制
{db-name}.{cf-name}.hard-pending-compaction-bytes-limit	pending compaction bytes 的硬限制
{db-name}.{cf-name}.titan.blob-run-mode	处理 blob 文件的模式

配置项	简介
storage.block-cache.capacity	共享 block cache 的大小（自 v4.0.3 起支持）
backup.num-threads	backup 线程的数量（自 v4.0.3 起支持）

配置项	简介
split.qps-threshold	对 Region 执行 load-base-split 的阈值。如果连续一段时间内，某个 Region 的读请求的 QPS 超过 qps-threshold，则切分该 Region

配置项	简介
split.byte-threshold	对 Region 执行 load-base-split 的阈值。如果连续一段时间内，某个 Region 的读请求的流量超过 byte-threshold，则切分该 Region

配置项	简介
split.split-balance-score	load-base-split 的控制参数，确保 Region 切分后左右访问尽量均匀，数值越小越均匀，但也可能导致无法切分
split.split-contained-score	load-base-split 的控制参数，数值越小，Region 切分后跨 Region 的访问越少

配置项	简介
cdc.min-ts-interval	定期推进 Resolved TS 的时间间隔
cdc.old-value-cache-memory-quota	缓存在内存中的 TiCDC Old Value 的条目占用内存的上限
cdc.sink-memory-quota	缓存在内存中的 TiCDC 数据变更事件占用内存的上限
cdc.incremental-scan-speed-limit	增量扫描历史数据的速度上限

配置项	简介
cdc.incremental-scan-concurrency	增量扫描历史数据任务的最大并发执行个数

上述前缀为 {db-name} 或 {db-name}.{cf-name} 的是 RocksDB 相关的配置项。db-name 的取值可为 rocksdb 或 raftdb。

- 当 db-name 为 rocksdb 时，cf-name 的可取值有：defaultcf、writecf、lockcf、raftcf；
- 当 db-name 为 raftdb 时，cf-name 的可取值有：defaultcf。

具体配置项的意义可参考 TiKV 配置文件描述

6.5.1.3 在线修改 PD 配置

PD 暂不支持单个实例拥有独立配置。所有实例共享一份配置，可以通过下列方式修改 PD 的配置项：

```
set config pd `log.level`='info'
```

设置成功会返回 Query OK：

```
Query OK, 0 rows affected (0.01 sec)
```

针对 PD 可在线修改的配置项，成功修改后则会持久化到 etcd 中，不会对配置文件进行持久化，后续以 etcd 中的配置为准。同上，若和 TiDB 预留关键字冲突，需要用反引号 ` 包裹此类配置项，例如 `schedule.leader-schedule-limit`。

支持配置项列表如下：

配置项	简介
log.level	日志级别
cluster-version	集群的版本
schedule.max-merge-region-size	控制 Region Merge 的 size 上限（单位是 MB）
schedule.max-merge-region-keys	控制 Region Merge 的 key 数量上限
schedule.patrol-region-interval	控制 replicaChecker 检查 Region 健康状态的运行频率
schedule.split-merge-interval	控制对同一个 Region 做 split 和 merge 操作的间隔
schedule.max-snapshot-count	控制单个 store 最多同时接收或发送的 snapshot 数量

配置项	简介
schedule.max-pending-peer-count	控制单个 store 的 pending peer 上限
schedule.max-store-down-time	PD 认为失联 store 无法恢复的时间
schedule.leader-schedule-policy	用于控制 leader 调度的策略
schedule.leader-schedule-limit	可以控制同时进行 leader 调度的任务个数
schedule.region-schedule-limit	可以控制同时进行 Region 调度的任务个数
schedule.replica-schedule-limit	可以控制同时进行 replica 调度的任务个数
schedule.merge-schedule-limit	控制同时进行的 Region Merge 调度的任务
schedule.hot-region-schedule-limit	可以控制同时进行的热点调度的任务个数
schedule.hot-region-cache-hits-threshold	用于设置 Region 被视为热点的阈值
schedule.high-space-ratio	用于设置 store 空间充裕的阈值
schedule.low-space-ratio	用于设置 store 空间不足的阈值
schedule.tolerant-size-ratio	控制 balance 缓冲区大小
schedule.enable-remove-down-replica	用于开启自动删除 DownReplica 的特性
schedule.enable-replace-offline-replica	用于开启迁移 OfflineReplica 的特性
schedule.enable-make-up-replica	用于开启补充副本的特性
schedule.enable-remove-extra-replica	用于开启删除多余副本的特性
schedule.enable-location-replacement	用于开启隔离级别检查
schedule.enable-cross-table-merge	用于开启跨表 Merge
schedule.enable-one-way-merge	用于开启单向 Merge（只允许和下一个相邻的 Region Merge）
replication.max-replicas	用于设置副本的数量
replication.location-labels	用于设置 TiKV 集群的拓扑信息
replication.enable-placement-rules	开启 Placement Rules
replication.strictly-match-label	开启 label 检查
pd-server.use-region-storage	开启独立的 Region 存储
pd-server.max-gap-reset-ts	用于设置最大的重置 timestamp 的间隔（BR）
pd-server.key-type	用于设置集群 key 的类型
pd-server.metric-storage	用于设置集群 metrics 的存储地址
pd-server.dashboard-address	用于设置 dashboard 的地址
replication-mode.replication-mode	备份的模式

具体配置项意义可参考 PD 配置文件描述。

6.5.1.4 在线修改 TiDB 配置

在线修改 TiDB 配置的方式和 TiKV/PD 有所不同，用户通过系统变量来完成修改。

下面例子展示了如何通过变量 `tidb_slow_log_threshold` 在线修改配置项 `slow-threshold`。

`slow-threshold` 默认值是 300 毫秒，可以通过设置系统变量 `tidb_slow_log_threshold` 将其修改为 200 毫秒：

```
set tidb_slow_log_threshold = 200;
```

```
Query OK, 0 rows affected (0.00 sec)
```

```
select @@tidb_slow_log_threshold;
```

```
+---------------------------+
| @@tidb_slow_log_threshold |
+---------------------------+
| 200                       |
+---------------------------+
1 row in set (0.00 sec)
```

支持在线修改的配置项和相应的 TiDB 系统变量如下：

配置项	对应变量	简介
mem-quota-query	tidb_mem_quota_query	查询语句的内存使用限制
log.enable-slow-log	tidb_enable_slow_log	慢日志的开关
log.slow-threshold	tidb_slow_log_threshold	慢日志阈值
log.expensive-threshold	tidb_expensive_query_time_threshold	expensive 查询阈值

6.6 Online Unsafe Recovery 使用文档

> 警告：
>
> - 此功能为有损恢复，无法保证数据和数据索引完整性。
> - 此功能为实验特性，其接口、策略和内部实现在 GA 前可能会有所变化。虽然已通过部分场景的测试，但尚未经过广泛验证，使用此功能可能导致系统不可用，不建议在生产环境中使用。
> - 建议在 TiDB 团队支持下进行相关操作，操作不当可能导致集群难以恢复。

当多数副本的永久性损坏造成部分数据不可读写时，可以使用 Online Unsafe Recovery 功能进行数据有损恢复。

6.6.1 功能说明

在 TiDB 中，根据用户定义的多种副本规则，一份数据可能会同时存储在多个节点中，从而保证在单个或少数节点暂时离线或损坏时，读写数据不受任何影响。但是，当一个 Region 的多数或全部副本在短时间内全部下线时，为了保证数据的完整性，该 Region 会处于暂不可用的状态。

如果一段数据的多数副本发生了永久性损坏（如磁盘损坏）等问题，从而导致节点无法上线时，此段数据会一直保持暂不可用的状态。这时，如果用户希望集群恢复正常使用，在用户能够容忍数据回退或数据丢失的前提下，用户理论上可以通过手动移除不可用副本的方式，使 TiDB 重新形成多数派，进而让上层业务可以写入和读取（可能是 stale 的，或者为空）这一段数据分片。

在这个情况下，当存有可容忍丢失的数据的部分节点受到永久性损坏时，用户可以通过使用 Online Unsafe Recovery，快速简单地进行有损恢复。使用 Online Unsafe Recovery 时，PD 会收集全部节点内的数据分片元信息，用 PD 的全局视角生成一份更实时、更完整的恢复计划后，将其计划下发给各个存活的节点，使各节点执行数据恢复任务。另外，下发恢复计划后，PD 还会定期查看恢复进度，并在需要时，重新向各节点分发恢复计划。

6.6.2 适用场景

Online Unsafe Recovery 功能适用于以下场景：

- 部分节点受到永久性损坏，导致节点无法重启，造成业务端的部分数据不可读、不可写。
- 可以容忍数据丢失，希望受影响的数据恢复读写。
- 希望在线一站式恢复数据。

6.6.3 使用步骤

6.6.3.1 前提条件

在使用 Online Unsafe Recovery 功能进行数据有损恢复前，请确认以下事项：

- 离线节点导致部分数据确实不可用。
- 离线节点确实无法自动恢复或重启。

6.6.3.2 第 1 步：关闭各类调度

暂时关闭负载均衡等各类内部调度。关闭后，建议等待约 10 分钟，使已经触发的调度能有充分的时间完成调度任务。

> 注意：
>
> 关闭调度后，系统将无法处理系统数据热点问题，请在恢复后尽快重新开启调度。

1. 使用 PD Control 执行`config show`命令，获取当前的配置信息。
2. 使用 PD Control 关闭各类调度：
    - `config set region-schedule-limit 0`
    - `config set replica-schedule-limit 0`
    - `config set merge-schedule-limit 0`

6.6.3.3 第 2 步：移除无法自动恢复的节点

使用 PD Control 执行`unsafe remove-failed-stores <store_id>[,<store_id>,...]`命令，移除无法自动恢复的节点。

> 注意：
>
> 此命令成功返回，仅表示请求已被接受，而不代表恢复成功。节点实际会在后台进行恢复。

6.6.3.4 第 3 步：查看进度

节点移除命令运行成功后，使用 PD Control 执行`unsafe remove-failed-stores show`命令，查看移除进度。当命令执行结果显示 “Last recovery has finished” 时，系统恢复完成。

6.6.3.5 第 4 步：测试读写任务

进度命令提示任务已完成后，可以尝试运行一些简单 SQL 查询或写入操作确保数据可以读写。示例如下：

```
select count(*) from table_that_suffered_from_group_majority_failure;
```

> 注意：
>
> 数据可以读写并不代表没有数据丢失。

6.6.3.6 第 5 步：重新开启调度

把在第 1 步中修改的配置 `config set region-schedule-limit 0`，`config set replica-schedule-limit 0`，`config set merge-schedule-limit 0` 调整回初始值，重新开启调度功能。至此，整个流程结束。

7 安全加固

7.1 为 TiDB 客户端服务端间通信开启加密传输

TiDB 服务端与客户端之间默认采用非加密连接，因而具备监视信道流量能力的第三方可以知悉 TiDB 服务端与客户端之间发送和接受的数据，包括但不限于查询语句内容、查询结果等。若信道是不可信的，例如客户端是通过公网连接到 TiDB 服务端的，则非加密连接容易造成信息泄露，建议使用加密连接确保安全性。

TiDB 服务端支持启用基于 TLS（传输层安全）协议的加密连接，协议与 MySQL 加密连接一致，现有 MySQL Client 如 MySQL Shell 和 MySQL 驱动等能直接支持。TLS 的前身是 SSL，因而 TLS 有时也被称为 SSL，但由于 SSL 协议有已知安全漏洞，TiDB 实际上并未支持。TiDB 支持的 TLS/SSL 协议版本为 TLS 1.0、TLS 1.1、TLS 1.2、TLS 1.3。

使用加密连接后，连接将具有以下安全性质：

- 保密性：流量明文被加密，无法被窃听；
- 完整性：流量明文无法被篡改；
- 身份验证（可选）：客户端和服务端能验证双方身份，避免中间人攻击。

要为 TiDB 客户端与服务端间的通信开启 TLS 加密传输，首先需要在 TiDB 服务端通过配置开启 TLS 加密连接的支持，然后通过配置客户端应用程序使用 TLS 加密连接。一般情况下，如果服务端正确配置了 TLS 加密连接支持，客户端库都会自动启用 TLS 加密传输。

另外，与 MySQL 相同，TiDB 也支持在同一 TCP 端口上开启 TLS 连接或非 TLS 连接。对于开启了 TLS 连接支持的 TiDB 服务端，客户端既可以选择通过加密连接安全地连接到该 TiDB 服务端，也可以选择使用普通的非加密连接。如需使用加密连接，你可以通过以下方式进行配置：

- 通过在启动参数中配置 `--require-secure-transport` 要求所有用户必须使用加密连接来连接到 TiDB。
- 通过在创建用户 (`create user`)，或修改已有用户 (`alter user`) 时指定 `REQUIRE SSL` 要求指定用户必须使用加密连接来连接 TiDB。以创建用户为例：

```
CREATE USER 'u1'@'%' IDENTIFIED BY 'my_random_password' REQUIRE SSL;
```

> 注意：
>
> 如果登录用户已配置使用 TiDB 证书鉴权功能校验用户证书，也会隐式要求对应用户必须使用加密连接连接 TiDB。

7.1.1 配置 TiDB 服务端启用安全连接

要启用安全连接，请参考以下相关参数说明：

- `auto-tls`: 启用证书自动生成功能（从 v5.2.0 开始）
- `ssl-cert`：指定 SSL 证书文件路径
- `ssl-key`：指定证书文件对应的私钥
- `ssl-ca`：可选，指定受信任的 CA 证书文件路径

`auto tls` 支持安全连接，但不提供客户端证书验证。有关证书验证和控制证书生成方式的说明，请参考下面配置 `ssl-cert`，`ssl-key` 和 `ssl-ca` 变量的建议：

- 在启动 TiDB 时，至少需要在配置文件中同时指定 `ssl-cert` 和 `ssl-key` 参数，才能在 TiDB 服务端开启安全连接。还可以指定 `ssl-ca` 参数进行客户端身份验证（请参见配置启用身份验证章节）。
- 参数指定的文件都为 PEM 格式。另外目前 TiDB 尚不支持加载有密码保护的私钥，因此必须提供一个没有密码的私钥文件。若提供的证书或私钥无效，则 TiDB 服务端将照常启动，但并不支持客户端加密连接到 TiDB 服务端。
- 若证书参数无误，则 TiDB 在启动时将会输出 `secure connection is enabled`，否则 TiDB 会输出 `secure connection is NOT ENABLED`。

> 注意：
>
> v5.2.0 版本之前，你可以使用 `mysql_ssl_rsa_setup --datadir=./certs` 生成证书。`mysql_ssl_rsa_setup` 工具是 MySQL Server 的一部分。

7.1.2 配置 MySQL Client 使用安全连接

MySQL 5.7 及以上版本自带的客户端默认尝试使用安全连接，若服务端不支持安全连接则自动退回到使用非安全连接；MySQL 5.7 以下版本自带的客户端默认采用非安全连接。

可以通过命令行参数修改客户端的连接行为，包括：

- 强制使用安全连接，若服务端不支持安全连接则连接失败 (--ssl-mode=REQUIRED)
- 尝试使用安全连接，若服务端不支持安全连接则退回使用不安全连接
- 使用不安全连接 (--ssl-mode=DISABLED)

除此参数外，MySQL 8.0 客户端有两种 SSL 模式：

- --ssl-mode=VERIFY_CA: 根据 --ssl-ca 签发的 CA 验证来自服务器的证书。
- --ssl-mode=VERIFY_IDENTITY: 与 VERIFY_CA 相同，但也验证所连接的主机名是否与证书匹配。

详细信息请参阅 MySQL 文档中关于客户端配置安全连接的部分。

7.1.3 配置启用身份验证

若在 TiDB 服务端或 MySQL Client 中未指定 ssl-ca 参数，则默认不会进行客户端或服务端身份验证，无法抵御中间人攻击，例如客户端可能会 “安全地” 连接到了一个伪装的服务端。可以在服务端和客户端中配置 ssl-ca 参数进行身份验证。一般情况下只需验证服务端身份，但也可以验证客户端身份进一步增强安全性。

- 若要使 MySQL Client 验证 TiDB 服务端身份，TiDB 服务端需至少配置 ssl-cert 和 ssl-key 参数，客户端需至少指定 --ssl-ca 参数，且 --ssl-mode 至少为 VERIFY_CA。必须确保 TiDB 服务端配置的证书 (ssl-cert) 是由客户端 --ssl-ca 参数所指定的 CA 所签发的，否则身份验证失败。
- 若要使 TiDB 服务端验证 MySQL Client 身份，TiDB 服务端需配置 ssl-cert、ssl-key、ssl-ca 参数，客户端需至少指定 --ssl-cert、--ssl-key 参数。必须确保服务端配置的证书和客户端配置的证书都是由服务端配置指定的 ssl-ca 签发的。
- 若要进行双向身份验证，请同时满足上述要求。

默认情况，服务端对客户端的身份验证是可选的。若客户端在 TLS 握手时未出示自己的身份证书，也能正常建立 TLS 连接。但也可以通过在创建用户 (create user)，赋予权限 (grant) 或修改已有用户 (alter user) 时指定 require 509 要求客户端需进行身份验证，以创建用户为例：

```
create user 'u1'@'%'  require x509;
```

> 注意：
>
> 如果登录用户已配置使用 TiDB 证书鉴权功能校验用户证书，也会隐式要求对应用户需进行身份验证。

7.1.4 检查当前连接是否是加密连接

可以通过 SHOW STATUS LIKE "%Ssl%"; 了解当前连接的详细情况，包括是否使用了安全连接、安全连接采用的加密协议、TLS 版本号等。

以下是一个安全连接中执行该语句的结果。由于客户端支持的 TLS 版本号和加密协议会有所不同，执行结果相应地也会有所变化。

```
SHOW STATUS LIKE "%Ssl%";
```

```
......
| Ssl_verify_mode | 5                           |
| Ssl_version     | TLSv1.2                     |
| Ssl_cipher      | ECDHE-RSA-AES128-GCM-SHA256  |
......
```

除此以外，对于 MySQL 自带客户端，还可以使用 STATUS 或 \s 语句查看连接情况：

```
\s
```

```
...
SSL: Cipher in use is ECDHE-RSA-AES128-GCM-SHA256
...
```

7.1.5 支持的 TLS 版本及密钥交换协议和加密算法

TiDB 支持的 TLS 版本及密钥交换协议和加密算法由 Golang 官方库决定。

你使用的客户端库和操作系统加密策略也可能会影响到支持的协议版本和密码套件。

7.1.5.1 支持的 TLS 版本

- TLS 1.0
- TLS 1.1
- TLS 1.2
- TLS 1.3

7.1.5.2 支持的密钥交换协议及加密算法

- TLS_RSA_WITH_AES_128_CBC_SHA
- TLS_RSA_WITH_AES_256_CBC_SHA
- TLS_RSA_WITH_AES_128_CBC_SHA256
- TLS_RSA_WITH_AES_128_GCM_SHA256
- TLS_RSA_WITH_AES_256_GCM_SHA384
- TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA
- TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA
- TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA
- TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA
- TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA256
- TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
- TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256

- TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256
- TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384
- TLS_ECDHE_ECDSA_WITH_AES_256_GCM_SHA384
- TLS_ECDHE_RSA_WITH_CHACHA20_POLY1305_SHA256
- TLS_ECDHE_ECDSA_WITH_CHACHA20_POLY1305_SHA256
- TLS_AES_128_GCM_SHA256
- TLS_AES_256_GCM_SHA384
- TLS_CHACHA20_POLY1305_SHA256

7.1.6 重加载证书、密钥和 CA

在需要替换证书、密钥或 CA 时，可以在完成对应文件替换后，对运行中的 TiDB 实例执行 `ALTER INSTANCE RELOAD TLS` 语句从原配置的证书 (`ssl-cert`)、密钥 (`ssl-key`) 和 CA (`ssl-ca`) 的路径重新加证书、密钥和 CA，而无需重启 TiDB 实例。

新加载的证书密钥和 CA 将在语句执行成功后对新建立的连接生效，不会影响语句执行前已建立的连接。

7.1.7 监控

自 TiDB v5.2.0 版本起，你可以使用 `Ssl_server_not_after` 和 `Ssl_server_not_before` 状态变量监控证书有效期的起止时间。

```
SHOW GLOBAL STATUS LIKE 'Ssl\_server\_not\_%';
```

```
+----------------------+--------------------------+
| Variable_name        | Value                    |
+----------------------+--------------------------+
| Ssl_server_not_after  | Nov 28 06:42:32 2021 UTC |
| Ssl_server_not_before | Aug 30 06:42:32 2021 UTC |
+----------------------+--------------------------+
2 rows in set (0.0076 sec)
```

7.1.8 另请参阅

- 为 TiDB 组件间通信开启加密传输

7.2 为 TiDB 组件间通信开启加密传输

本部分介绍如何为 TiDB 集群内各部组件间开启加密传输，一旦开启以下组件间均将使用加密传输：

- TiDB 与 TiKV、PD
- TiKV 与 PD
- TiDB Control 与 TiDB，TiKV Control 与 TiKV，PD Control 与 PD
- TiKV、PD、TiDB 各自集群内内部通讯

目前暂不支持只开启其中部分组件的加密传输。

7.2.1 配置开启加密传输

1. 准备证书。

   推荐为 TiDB、TiKV、PD 分别准备一个 Server 证书，并保证可以相互验证，而它们的 Control 工具则可选择共用 Client 证书。

   有多种工具可以生成自签名证书，如 openssl，easy-rsa，cfssl。

   这里提供一个使用 openssl 生成证书的示例：生成自签名证书。

2. 配置证书。

   - TiDB

     在 config 文件或命令行参数中设置：

     ```
     [security]
     # Path of file that contains list of trusted SSL CAs for connection with cluster
         ↪ components.
     cluster-ssl-ca = "/path/to/ca.pem"
     # Path of file that contains X509 certificate in PEM format for connection with cluster
         ↪  components.
     cluster-ssl-cert = "/path/to/tidb-server.pem"
     # Path of file that contains X509 key in PEM format for connection with cluster
         ↪ components.
     cluster-ssl-key = "/path/to/tidb-server-key.pem"
     ```

   - TiKV

     在 config 文件或命令行参数中设置，并设置相应的 URL 为 https：

     ```
     [security]
     # set the path for certificates. Empty string means disabling secure connectoins.
     ca-path = "/path/to/ca.pem"
     cert-path = "/path/to/tikv-server.pem"
     key-path = "/path/to/tikv-server-key.pem"
     ```

   - PD

     在 config 文件或命令行参数中设置，并设置相应的 URL 为 https：

     ```
     [security]
     # Path of file that contains list of trusted SSL CAs. if set, following four settings
         ↪ shouldn't be empty
     cacert-path = "/path/to/ca.pem"
     # Path of file that contains X509 certificate in PEM format.
     cert-path = "/path/to/pd-server.pem"
     # Path of file that contains X509 key in PEM format.
     key-path = "/path/to/pd-server-key.pem"
     ```

   - TiFlash（从 v4.0.5 版本开始引入）

     在 tiflash.toml 文件中设置，将 http_port 项改为 https_port:

```
[security]
# Path of file that contains list of trusted SSL CAs. if set, following four settings
    ↪ shouldn't be empty
ca_path = "/path/to/ca.pem"
# Path of file that contains X509 certificate in PEM format.
cert_path = "/path/to/tiflash-server.pem"
# Path of file that contains X509 key in PEM format.
key_path = "/path/to/tiflash-server-key.pem"
```

在 tiflash-learner.toml 文件中设置，

```
[security]
# Sets the path for certificates. The empty string means that secure connections are
    ↪ disabled.
ca-path = "/path/to/ca.pem"
cert-path = "/path/to/tiflash-server.pem"
key-path = "/path/to/tiflash-server-key.pem"
```

- TiCDC

  在启动命令行中设置，并设置相应的 URL 为 https：

```
cdc server --pd=https://127.0.0.1:2379 --log-file=ticdc.log --addr=0.0.0.0:8301 --
    ↪ advertise-addr=127.0.0.1:8301 --ca=/path/to/ca.pem --cert=/path/to/ticdc-cert.
    ↪ pem --key=/path/to/ticdc-key.pem
```

此时 TiDB 集群各个组件间已开启加密传输。

> 注意：
>
> 若 TiDB 集群各个组件间开启加密传输后，在使用 tidb-ctl、tikv-ctl 或 pd-ctl 工具连接集群时，需要指定 client 证书，示例：

```
./tidb-ctl -u https://127.0.0.1:10080 --ca /path/to/ca.pem --ssl-cert /path/to/client.pem --
    ↪ ssl-key /path/to/client-key.pem
```

```
./pd-ctl -u https://127.0.0.1:2379 --cacert /path/to/ca.pem --cert /path/to/client.pem --key
    ↪  /path/to/client-key.pem
```

```
./tikv-ctl --host="127.0.0.1:20160" --ca-path="/path/to/ca.pem" --cert-path="/path/to/client
    ↪ .pem" --key-path="/path/to/clinet-key.pem"
```

7.2.2 认证组件调用者身份

通常被调用者除了校验调用者提供的密钥、证书和 CA 有效性外，还需要校验调用方身份以防止拥有有效证书的非法访问者进行访问（例如：TiKV 只能被 TiDB 访问，需阻止拥有合法证书但非 TiDB 的其他访问者访问 TiKV）。

如希望进行组件调用者身份认证，需要在生证书时通过 Common Name 标识证书使用者身份，并在被调用者配置检查证书 Common Name 列表来检查调用者身份。

- TiDB

  在 config 文件或命令行参数中设置：

```
[security]
cluster-verify-cn = [
  "TiDB-Server",
  "TiKV-Control",
]
```

- TiKV

  在 config 文件或命令行参数中设置：

```
[security]
cert-allowed-cn = [
    "TiDB-Server", "PD-Server", "TiKV-Control", "RawKvClient1",
]
```

- PD

  在 config 文件或命令行参数中设置：

```
[security]
cert-allowed-cn = ["TiKV-Server", "TiDB-Server", "PD-Control"]
```

- TiCDC

  在启动命令行中设置：

```
cdc server --pd=https://127.0.0.1:2379 --log-file=ticdc.log --addr=0.0.0.0:8301 --advertise-
    ↪ addr=127.0.0.1:8301 --ca=/path/to/ca.pem --cert=/path/to/ticdc-cert.pem --key=/path/
    ↪ to/ticdc-key.pem --cert-allowed-cn="client1,client2"
```

- TiFlash（从 v4.0.5 版本开始引入）

  在 tiflash.toml 文件中设置：

```
[security]
cert_allowed_cn = ["TiKV-Server", "TiDB-Server"]
```

  在 tiflash-learner.toml 文件中设置：

```
[security]
cert-allowed-cn = ["PD-Server", "TiKV-Server", "TiFlash-Server"]
```

7.2.3 证书重加载

TiDB、PD 和 TiKV 和各种 Client 都会在每次新建相互通讯的连接时重新读取当前的证书和密钥文件内容，实现证书和密钥的重加载。目前暂不支持 CA 的重加载。

7.2.4 另请参阅

- 为 TiDB 客户端服务端间通信开启加密传输

7.3 为 DM 的连接开启加密传输

本文介绍如何为 DM 的连接开启加密传输，包括 DM-master，DM-worker，dmctl 组件之间的连接以及 DM 组件与上下游数据库之间的连接。

7.3.1 为 DM-master，DM-worker，dmctl 组件之间的连接开启加密传输

本节介绍如何为 DM-master，DM-worker，dmctl 组件之间的连接开启加密传输。

7.3.1.1 配置开启加密传输

1. 准备证书。

   推荐为 DM-master、DM-worker 分别准备一个 Server 证书，并保证可以相互验证，而 dmctl 工具则可选择共用 Client 证书。

   有多种工具可以生成自签名证书，如 `openssl`，`cfssl` 及 `easy-rsa` 等基于 `openssl` 的工具。

   这里提供一个使用 `openssl` 生成证书的示例：生成自签名证书。

2. 配置证书。

> **注意：**
>
> DM-master、DM-worker 与 dmctl 三个组件可使用同一套证书。

- DM-master

  在 DM-master 配置文件或命令行参数中设置：

```
ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/master-cert.pem"
ssl-key = "/path/to/master-key.pem"
```

- DM-worker

  在 DM-worker 配置文件或命令行参数中设置：

```
ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/worker-cert.pem"
ssl-key = "/path/to/worker-key.pem"
```

- dmctl

  若 DM 集群各个组件间开启加密传输后，在使用 dmctl 工具连接集群时，需要指定 Client 证书，示例如下：

```
./dmctl --master-addr=127.0.0.1:8261 --ssl-ca /path/to/ca.pem --ssl-cert /path/to/
    ↪ client-cert.pem --ssl-key /path/to/client-key.pem
```

7.3.1.2 认证组件调用者身份

通常被调用者除了校验调用者提供的密钥、证书和 CA 有效性外，还需要校验调用者身份以防止拥有有效证书的非法访问者进行访问（例如：DM-worker 只能被 DM-master 访问，需阻止拥有合法证书但非 DM-master 的其他访问者访问 DM-worker）。

如希望进行组件调用者身份认证，需要在生成证书时通过 Common Name (CN) 标识证书使用者身份，并在被调用者配置检查证书 Common Name 列表时检查调用者身份。

- DM-master

  在 config 文件或命令行参数中设置：

```
cert-allowed-cn = ["dm"]
```

- DM-worker

  在 config 文件或命令行参数中设置：

```
cert-allowed-cn = ["dm"]
```

7.3.1.3 证书重加载

DM-master、DM-worker 和 dmctl 都会在每次新建相互通讯的连接时重新读取当前的证书和密钥文件内容，实现证书和密钥的重加载。

当 ssl-ca、ssl-cert 或 ssl-key 的文件内容更新后，可通过重启 DM 组件使其重新加载证书与密钥内容并重新建立连接。

7.3.2 DM 组件与上下游数据库之间的连接开启加密传输

本节介绍如何为 DM 组件与上下游数据库之间的连接开启加密传输。

7.3.2.1 为上游数据库连接开启加密传输

1. 配置上游数据库，启用加密连接支持并设置 Server 证书，具体可参考 Using encrypted connections
2. 在 source 配置文件中设置 MySQL Client 证书：

> **注意：**
>
> 请确保所有 DM-master 与 DM-worker 组件能通过指定路径读取到证书与密钥文件的内容。

```
from:
    security:
        ssl-ca: "/path/to/mysql-ca.pem"
        ssl-cert: "/path/to/mysql-client-cert.pem"
        ssl-key: "/path/to/mysql-client-key.pem"
```

7.3.2.2 为下游 TiDB 连接开启加密传输

1. 配置下游 TiDB 启用加密连接支持，具体可参考配置 TiDB 启用加密连接支持
2. 在 task 配置文件中设置 TiDB Client 证书：

> **注意：**
>
> 请确保所有 DM-master 与 DM-worker 组件能通过指定路径读取到证书与密钥文件的内容。

```
target-database:
    security:
        ssl-ca: "/path/to/tidb-ca.pem"
        ssl-cert: "/path/to/tidb-client-cert.pem"
        ssl-key: "/path/to/tidb-client-key.pem"
```

7.4 生成自签名证书

> **注意：**
>
> 要在 TiDB 客户端与服务端间通信开启加密传输，你只需配置 `auto-tls` 参数。

本文档提供使用 `openssl` 生成自签名证书的一个示例，用户也可以根据自己的需求生成符合要求的证书和密钥。

假设实例集群拓扑如下：

Name	Host IP	Services
node1	172.16.10.11	PD1, TiDB1
node2	172.16.10.12	PD2

Name	Host IP	Services
node3	172.16.10.13	PD3
node4	172.16.10.14	TiKV1
node5	172.16.10.15	TiKV2
node6	172.16.10.16	TiKV3

7.4.1 安装 OpenSSL

对于 Debian 或 Ubuntu 操作系统：

```
apt install openssl
```

对于 RedHat 或 CentOS 操作系统：

```
yum install openssl
```

也可以参考 OpenSSL 官方的下载文档进行安装。

7.4.2 生成 CA 证书

CA 的作用是签发证书。实际情况中，请联系你的管理员签发证书或者使用信任的 CA 机构。CA 会管理多个证书对，这里只需生成原始的一对证书，步骤如下：

1. 生成 root 密钥：

    ```
    openssl genrsa -out root.key 4096
    ```

2. 生成 root 证书：

    ```
    openssl req -new -x509 -days 1000 -key root.key -out root.crt
    ```

3. 验证 root 证书：

    ```
    openssl x509 -text -in root.crt -noout
    ```

7.4.3 签发各个组件的证书

7.4.3.1 集群中可能使用到的证书

- tidb certificate 由 TiDB 使用，为其他组件和客户端验证 TiDB 身份。
- tikv certificate 由 TiKV 使用，为其他组件和客户端验证 TiKV 身份。
- pd certificate 由 PD 使用，为其他组件和客户端验证 PD 身份。
- client certificate 用于 PD、TiKV、TiDB 验证客户端。例如 `pd-ctl`，`tikv-ctl` 等。

7.4.3.2 给 TiKV 实例签发证书

给 TiKV 实例签发证书的步骤如下：

1. 生成该证书对应的私钥：

```
openssl genrsa -out tikv.key 2048
```

2. 拷贝一份 OpenSSL 的配置模板文件。

   模板文件可能存在多个位置，请以实际位置为准：

```
cp /usr/lib/ssl/openssl.cnf .
```

   如果不知道实际位置，请在根目录下查找：

```
find / -name openssl.cnf
```

3. 编辑 `openssl.cnf`，在 `[ req ]` 字段下加入 `req_extensions = v3_req`，然后在 `[ v3_req ]` 字段下加入 `subjectAltName = @alt_names`。最后新建一个字段，并编辑 SAN 的信息：

```
[ alt_names ]
IP.1 = 127.0.0.1
IP.2 = 172.16.10.14
IP.3 = 172.16.10.15
IP.4 = 172.16.10.16
```

4. 保存 `openssl.cnf` 文件后，生成证书请求文件（在这一步也可以为该证书指定 Common Name，其作用是让服务端验证接入的客户端的身份，各个组件默认不会开启验证，需要在配置文件中启用该功能才生效）：

```
openssl req -new -key tikv.key -out tikv.csr -config openssl.cnf
```

5. 签发生成证书：

```
openssl x509 -req -days 365 -CA root.crt -CAkey root.key -CAcreateserial -in tikv.csr -out
    ↪ tikv.crt -extensions v3_req -extfile openssl.cnf
```

6. 验证证书携带 SAN 字段信息（可选）：

```
openssl x509 -text -in tikv.crt -noout
```

7. 确认在当前目录下得到如下文件：

```
root.crt
tikv.crt
tikv.key
```

为其它 TiDB 组件签发证书的过程类似，此文档不再赘述。

7.4.3.3 为客户端签发证书

为客户端签发证书的步骤如下。

1. 生成该证书对应的私钥：

```
openssl genrsa -out client.key 2048
```

2. 生成证书请求文件（在这一步也可以为该证书指定 Common Name，其作用是让服务端验证接入的客户端的身份，默认不会开启对各个组件的验证，需要在配置文件中启用该功能才生效）

```
openssl req -new -key client.key -out client.csr
```

3. 签发生成证书：

```
openssl x509 -req -days 365 -CA root.crt -CAkey root.key -CAcreateserial -in client.csr -out
    ↪  client.crt
```

7.5 静态加密

> **注意：**
>
> 如果集群部署在 AWS 上并在使用 EBS (Elastic Block Store) 存储数据，建议使用 EBS 加密，详细信息请参考 AWS 文档 - EBS 加密。如果集群部署在 AWS 上，但未使用 EBS 存储（例如使用本地 NVMe 存储），则建议使用本文中介绍的静态加密方式。

静态加密 (encryption at rest) 即在存储数据时进行数据加密。对于数据库，静态加密功能也叫透明数据加密 (TDE)，区别于传输数据加密 (TLS) 或使用数据加密（很少使用）。SSD 驱动器、文件系统、云供应商等都可进行静态加密。但区别于这些加密方式，若 TiKV 在存储数据前就进行数据加密，攻击者则必须通过数据库的身份验证才能访问数据。例如，即使攻击者获得物理机的访问权限时，也无法通过复制磁盘上的文件来访问数据。

7.5.1 各 TiDB 组件支持的加密方式

在一个 TiDB 集群中，不同的组件使用不同的加密方式。本节介绍 TiKV、TiFlash、PD 和 Backup & Restore (BR) 等不同 TiDB 组件支持的加密方式。

部署好一个 TiDB 集群后，大部分用户数据会存储在 TiKV 和 TiFlash 节点上。此外，一些元数据会存储在 PD 节点上，例如用作为 TiKV Region 边界的二级索引密钥。为了能够充分发挥静态加密的优势，所有组件都需要启用加密功能。此外，进行加密时，也需要对备份、日志文件和通过网络传输的数据进行加密。

7.5.1.1 TiKV

TiKV 支持静态加密，即在 CTR 模式下使用 AES 对数据文件进行透明加密。要启用静态加密，用户需提供一个加密密钥，即主密钥。TiKV 自动轮换 (rotate) 用于加密实际数据文件的密钥，主密钥则可以由用户手动轮换。请注意，静态加密仅加密静态数据（即磁盘上的数据），而不加密网络传输中的数据。建议 TLS 与静态加密一起使用。

可以选择将 AWS KMS (Key Management Service) 用于云上部署或本地部署，也可以指定将密钥以明文形式存储在文件中。

TiKV 当前不从核心转储 (core dumps) 中排除加密密钥和用户数据。建议在使用静态加密时禁用 TiKV 进程的核心转储，该功能目前无法由 TiKV 独立处理。

TiKV 使用文件的绝对路径来跟踪已加密的数据文件。一旦 TiKV 节点开启了加密功能，用户就不应更改数据文件的路径配置，例如 `storage.data-dir`，`raftstore.raftdb-path`，`rocksdb.wal-dir` 和 `raftdb.wal-dir`。

7.5.1.2 TiFlash

TiFlash 支持静态加密。数据密钥由 TiFlash 生成。TiFlash（包括 TiFlash Proxy）写入的所有文件，包括数据文件、Schema 文件、临时文件等，均由当前数据密钥加密。TiFlash 支持的加密算法、加密配置方法（配置项在 `tiflash-learner.toml` 中）和监控项含义等均与 TiKV 一致。

如果 TiFlash 中部署了 Grafana 组件，可以查看 TiFlash-Proxy-Details -> Encryption。

7.5.1.3 PD

PD 的静态加密为实验特性，其配置方式与 TiKV 相同。

7.5.1.4 BR 备份

BR 支持对备份到 S3 的数据进行 S3 服务端加密 (SSE)。BR S3 服务端加密也支持使用用户自行创建的 AWS KMS 密钥进行加密，详细信息请参考BR S3 服务端加密。

7.5.1.5 日志

TiKV，TiDB 和 PD 信息日志中可能包含用于调试的用户数据。信息日志不会被加密，建议开启日志脱敏功能。

7.5.2 TiKV 静态加密

TiKV 当前支持的加密算法包括 AES128-CTR、AES192-CTR 和 AES256-CTR。TiKV 使用信封加密 (envelop encryption)，所以启用加密后，TiKV 使用以下两种类型的密钥：

- 主密钥 (master key)：主密钥由用户提供，用于加密 TiKV 生成的数据密钥。用户在 TiKV 外部进行主密钥的管理。
- 数据密钥 (data key)：数据密钥由 TiKV 生成，是实际用于加密的密钥。

多个 TiKV 实例可共用一个主密钥。在生产环境中，推荐通过 AWS KMS 提供主密钥。首先通过 AWS KMS 创建用户主密钥 (CMK)，然后在配置文件中将 CMK 密钥的 ID 提供给 TiKV。TiKV 进程在运行时可以通过 IAM 角色访问 KMS CMK。如果 TiKV 无法访问 KMS CMK，TiKV 就无法启动或重新启动。详情参阅 AWS 文档中的 KMS and IAM。

用户也可以通过文件形式提供主密钥。该文件须包含一个用十六进制字符串编码的 256 位（32 字节）密钥，并以换行符结尾（即 \n），且不包含其他任何内容。将密钥存储在磁盘上会泄漏密钥，因此密钥文件仅适合存储在 RAM 内存的 tempfs 中。

数据密钥由 TiKV 传递给底层存储引擎（即 RocksDB）。RocksDB 写入的所有文件，包括 SST 文件，WAL 文件和 MANIFEST 文件，均由当前数据密钥加密。TiKV 使用的其他临时文件（可能包括用户数据）也由相同的数据密钥加密。默认情况下，TiKV 每周自动轮换数据密钥，但是该时间段是可配置的。密钥轮换时，TiKV 不会重写全部现有文件来替换密钥，但如果集群的写入量恒定，则 RocksDB compaction 会将使用最新的数据密钥对数据重新加密。TiKV 跟踪密钥和加密方法，并使用密钥信息对读取的内容进行解密。

无论用户配置了哪种数据加密方法，数据密钥都使用 AES256-GCM 算法进行加密，以方便对主密钥进行验证。所以当使用文件而不是 KMS 方式指定主密钥时，主密钥必须为 256 位（32 字节）。

7.5.2.1 创建密钥

如需在 AWS 上创建一个密钥，请执行以下步骤：

1. 进入 AWS 控制台的 AWS KMS。
2. 确保在控制台的右上角选择正确的区域。
3. 点击创建密钥，选择对称 (Symmetric) 作为密钥类型。
4. 为钥匙设置一个别名。

你也可以使用 AWS CLI 执行该操作：

```
aws --region us-west-2 kms create-key
aws --region us-west-2 kms create-alias --alias-name "alias/tidb-tde" --target-key-id 0987dcba-09
    ↪ fe-87dc-65ba-ab0987654321
```

需要在第二条命令中输入的 --target-key-id 是第一条命令的结果。

7.5.2.2 配置加密

要启用加密，你可以在 TiKV 和 PD 的配置文件中添加加密部分：

```
[security.encryption]
data-encryption-method = "aes128-ctr"
data-key-rotation-period = "168h" # 7 days
```

data-encryption-method 的可选值为 "aes128-ctr"、"aes192-ctr"、"aes256-ctr" 和 "plaintext"。默认值为 "plaintext"，即默认不开启加密功能。data-key-rotation-period 指定 TiKV 轮换密钥的频率。可以为新 TiKV 群集或现有 TiKV 群集开启加密，但只有启用后写入的数据才保证被加密。要禁用加密，请在配置文件中删除 data-encryption-method，或将该参数值为 "plaintext"，然后重启 TiKV。

如果启用了加密（即 data-encryption-method 的值不是 "plaintext"），则必须指定主密钥。要使用 AWS KMS 方式指定为主密钥，请在 [security.encryption] 部分之后添加 [security.encryption.master-key] 部分：

```
[security.encryption.master-key]
type = "kms"
key-id = "0987dcba-09fe-87dc-65ba-ab0987654321"
```

```
region = "us-west-2"
endpoint = "https://kms.us-west-2.amazonaws.com"
```

key-id 指定 KMS CMK 的密钥 ID。region 为 KMS CMK 的 AWS 区域名。endpoint 通常无需指定，除非你在使用非 AWS 提供的 AWS KMS 兼容服务或需要使用 KMS VPC endpoint。

你也可以使用 AWS 多区域键。为此，你需要在一个特定的区域设置一个主键，并在需要的区域中添加副本密钥。

若要使用文件方式指定主密钥，主密钥配置应如下所示：

```
[security.encryption.master-key]
type = "file"
path = "/path/to/key/file"
```

以上示例中，path 为密钥文件的路径。该文件须包含一个 256 位（32 字节）的十六进制字符串，并以换行符结尾（即 \n），且不包含其他任何内容。密钥文件示例如下：

```
3b5896b5be691006e0f71c3040a29495ddcad20b14aff61806940ebd780d3c62
```

7.5.2.3 轮换主密钥

要轮换主密钥，你必须在配置中同时指定新的主密钥和旧的主密钥，然后重启 TiKV。用 security.encryption.master-key 指定新的主密钥，用 security.encryption.previous-master-key 指定旧的主密钥。security.encryption.previous-master-key 配置的格式与 encryption.master-key 相同。重启时，TiKV 必须能同时访问旧主密钥和新主密钥，但一旦 TiKV 启动并运行，就只需访问新密钥。此后，可以将 encryption.previous-master-key 项保留在配置文件中。即使重启时，TiKV 也只会在无法使用新的主密钥解密现有数据时尝试使用旧密钥。

TiKV 当前不支持在线轮换主密钥，因此你需要重启 TiKV 进行主密钥轮换。建议对运行中的、提供在线查询的 TiKV 集群进行滚动重启。

轮换 KMS CMK 的配置示例如下：

```
[security.encryption.master-key]
type = "kms"
key-id = "50a0c603-1c6f-11e6-bb9e-3fadde80ce75"
region = "us-west-2"

[security.encryption.previous-master-key]
type = "kms"
key-id = "0987dcba-09fe-87dc-65ba-ab0987654321"
region = "us-west-2"
```

7.5.2.4 监控和调试

要监控静态加密（如果 TiKV 中部署了 Grafana 组件），可以查看 TiKV-Details -> Encryption 面板中的监控项：

- Encryption initialized：如果在 TiKV 启动期间初始化了加密，则为 1，否则为 0。进行主密钥轮换时可通过该监控项确认主密钥轮换是否已完成。
- Encryption data keys：现有数据密钥的数量。每次轮换数据密钥后，该数字都会增加 1。通过此监控指标可以监测数据密钥是否按预期轮换。
- Encrypted files：当前的加密数据文件数量。为先前未加密的群集启用加密时，将此数量与数据目录中的当前数据文件进行比较，可通过此监控指标估计已经被加密的数据量。
- Encryption meta file size：加密元数据文件的大小。
- Read/Write encryption meta duration：对用于加密的元数据进行操作带来的额外开销。

在调试方面，可使用 `tikv-ctl` 命令查看加密元数据（例如使用的加密方法和数据密钥列表）。该操作可能会暴露密钥，因此不推荐在生产环境中使用。详情参阅 TiKV Control。

7.5.2.5 TiKV 版本间兼容性

为了减少 TiKV 管理加密元数据造成的 I/O 操作和互斥锁竞争引发的开销，TiKV v4.0.9 对此进行了优化。此优化可以通过 TiKV 配置文件中的 `security.encryption.enable-file-dictionary-log` 参数启用或关闭。此配置参数仅在 TiKV v4.0.9 或更高版本中生效。

该配置项默认开启，此时 TiKV v4.0.8 或更早期的版本无法识别加密元数据的数据格式。例如，假设你正在使用 TiKV v4.0.9 或更高版本，开启了静态加密和默认开启了 `enable-file-dictionary-log` 配置。如果将集群降级到 TiKV v4.0.8 或更早版本，TiKV 将无法启动，并且信息日志中会有类似如下的报错：

```
[2020/12/07 07:26:31.106 +08:00] [ERROR] [mod.rs:110] ["encryption: failed to load file
    ↪ dictionary."]
[2020/12/07 07:26:33.598 +08:00] [FATAL] [lib.rs:483] ["called `Result::unwrap()` on an `Err`
    ↪ value: Other(\"[components/encryption/src/encrypted_file/header.rs:18]: unknown version
    ↪ 2\")"]
```

为了避免上面所示错误，你可以首先将 `security.encryption.enable-file-dictionary-log` 设置为 `false`，然后启动 TiKV v4.0.9 或更高版本。TiKV 成功启动后，加密元数据的数据格式将降级为 TiKV 早期版本可以识别的格式。此时，你可再将 TiKV 集群降级到较早的版本。

7.5.3 BR S3 服务端加密

使用 BR 备份数据到 S3 时，若要启用 S3 服务端加密，需要传递 `--s3.sse` 参数并将参数值设置为 `aws:kms`。S3 将使用自己的 KMS 密钥进行加密。示例如下：

```
./br backup full --pd <pd-address> --storage "s3://<bucket>/<prefix>" --s3.sse aws:kms
```

若要使用用户创建和拥有的自定义 AWS KMS CMK，需另外传递 `--s3.sse-kms-key-id` 参数。此时，BR 进程和集群中的所有 TiKV 节点都需访问该 KMS CMK（例如，通过 AWS IAM），并且该 KMS CMK 必须与存储备份的 S3 bucket 位于同一 AWS 区域。建议通过 AWS IAM 向 BR 进程和 TiKV 节点授予对 KMS CMK 的访问权限。参见 AWS 文档中的 IAM。示例如下：

```
./br backup full --pd <pd-address> --storage "s3://<bucket>/<prefix>" --s3.region <region> --s3.
    ↪ sse aws:kms --s3.sse-kms-key-id 0987dcba-09fe-87dc-65ba-ab0987654321
```

恢复备份时，不需要也不可指定 --s3.sse 和 --s3.sse-kms-key-id 参数。S3 将自动相应进行解密。用于恢复备份数据的 BR 进程和集群中的 TiKV 节点也需要访问 KMS CMK，否则恢复将失败。示例如下：

```
./br restore full --pd <pd-address> --storage "s3://<bucket>/<prefix> --s3.region <region>"
```

7.6 为 TiDB 落盘文件开启加密

当配置项 oom-use-tmp-storage 为 true 时，如果单条 SQL 语句的内存使用超出 mem-quota-query 的限制，某些算子可以将执行时的中间结果作为临时文件落盘保存，直到查询执行完成之后将它们删除。

用户可以开启落盘文件加密功能，防止攻击者通过读取临时文件来访问数据。

7.6.1 配置方法

要启用落盘文件加密功能，可以在 TiDB 配置文件中的 [security] 部分，配置spilled-file-encryption-method 选项：

```
[security]
spilled-file-encryption-method = "aes128-ctr"
```

spilled-file-encryption-method 的可选值为 aes128-ctr 和 plaintext。默认值为 plaintext，表示不启用加密。

7.7 日志脱敏

TiDB 在提供详细的日志信息时，可能会把数据库敏感的数据（例如用户数据）打印出来，造成数据安全方面的风险。因此 TiDB、TiKV、PD 等组件各提供了一个配置项开关，开关打开后，会隐藏日志中包含的用户数据值。

7.7.1 TiDB 组件日志脱敏

TiDB 侧的日志脱敏需要将global.tidb_redact_log 的值设为 1。该变量值默认为 0，即关闭脱敏。

可以通过 set 语法，设置全局系统变量 tidb_redact_log，示例如下：

```
set @@global.tidb_redact_log=1;
```

设置后，所有新 session 产生的日志都会脱敏：

```
create table t (a int, unique key (a));
Query OK, 0 rows affected (0.00 sec)

insert into t values (1),(1);
ERROR 1062 (23000): Duplicate entry '1' for key 'a'
```

打印出的错误日志如下：

```
[2020/10/20 11:45:49.539 +08:00] [INFO] [conn.go:800] ["command dispatched failed"] [conn=5] [
    ↪ connInfo="id:5, addr:127.0.0.1:57222 status:10, collation:utf8_general_ci,  user:root"] [
    ↪ command=Query] [status="inTxn:0, autocommit:1"] [sql="insert into t values ( ? ) , ( ? )
    ↪ "] [txn_mode=OPTIMISTIC] [err="[kv:1062]Duplicate entry '?' for key 'a'"]
```

从以上报错日志可以看到，开启 tidb_redact_log 后，报错信息里的敏感内容被隐藏掉了（目前是用问号替代）。TiDB 日志中会把敏感信息隐藏掉，以此规避数据安全风险。

7.7.2 TiKV 组件日志脱敏

TiKV 侧的日志脱敏需要将security.redact-info-log 的值设为 true。该配置项值默认为 false，即关闭脱敏。

7.7.3 PD 组件日志脱敏

PD 侧的日志脱敏需要将security.redact-info-log 的值设为 true。该配置项值默认为 false，即关闭脱敏。

7.7.4 TiFlash 组件日志脱敏

TiFlash 侧的日志脱敏需要将 tiflash-server 中security.redact_info_log 配置项的值以及 tiflash-learner 中security
↪ .redact-info-log 配置项的值均设为 true。两配置项默认值均为 false，即关闭脱敏。

8 备份与恢复

8.1 备份与恢复工具 BR 简介

BR 全称为 Backup & Restore，是 TiDB 分布式备份恢复的命令行工具，用于对 TiDB 集群进行数据备份和恢复。

相比 dumpling，BR 更适合大数据量的场景。

BR 除了可以用来进行常规备份恢复外，也可以在保证兼容性前提下用来做大规模的数据迁移。

本文介绍了 BR 的工作原理、推荐部署配置、使用限制以及几种使用方式。

8.1.1 工作原理

BR 将备份或恢复操作命令下发到各个 TiKV 节点。TiKV 收到命令后执行相应的备份或恢复操作。

在一次备份或恢复中，各个 TiKV 节点都会有一个对应的备份路径，TiKV 备份时产生的备份文件将会保存在该路径下，恢复时也会从该路径读取相应的备份文件。

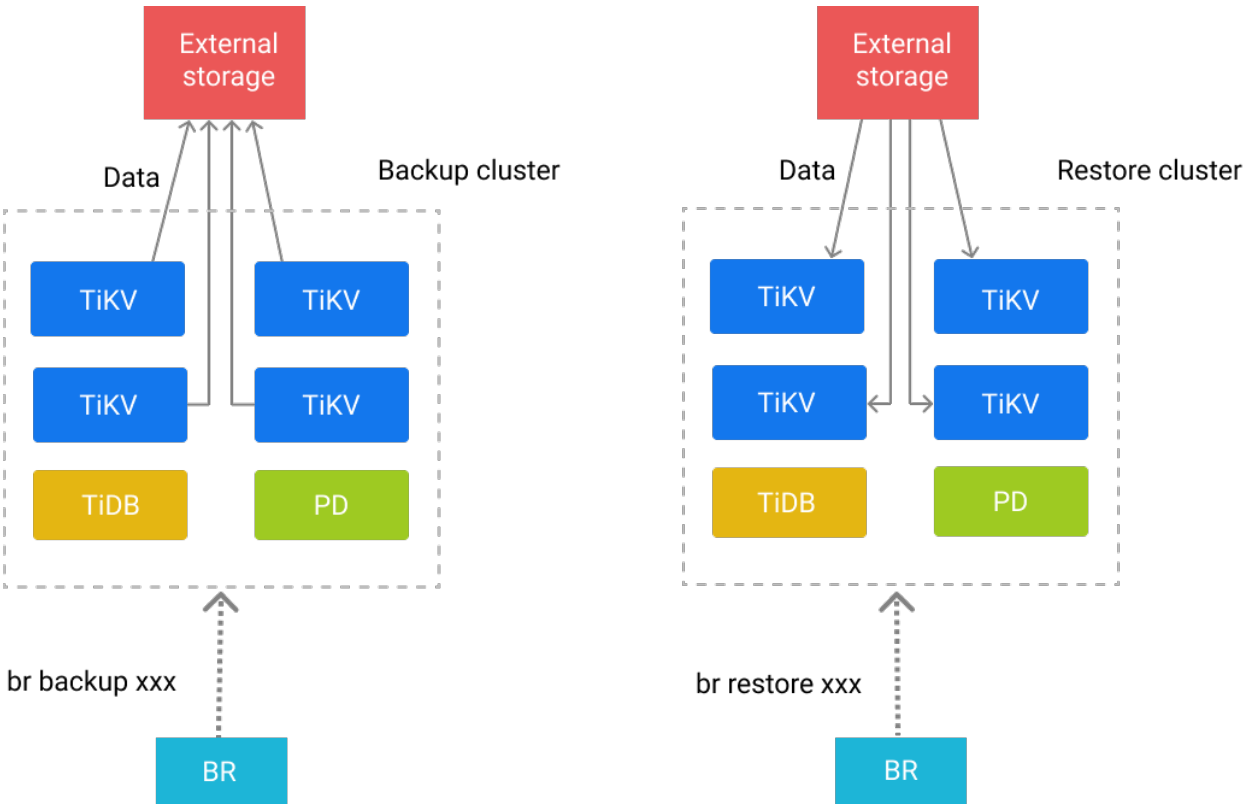


图 19: br-arch

更多信息请参阅备份恢复设计方案。

8.1.1.1 备份文件类型

备份路径下会生成以下两种类型文件：

- SST 文件：存储 TiKV 备份下来的数据信息
- `backupmeta` 文件：存储本次备份的元信息，包括备份文件数、备份文件的 Key 区间、备份文件大小和备份文件 Hash (sha256) 值
- `backup.lock` 文件：用于防止多次备份到同一目录

8.1.1.2 SST 文件命名格式

SST 文件以 `storeID_regionID_regionEpoch_keyHash_cf` 的格式命名。格式名的解释如下：

- storeID：TiKV 节点编号
- regionID：Region 编号
- regionEpoch：Region 版本号
- keyHash：Range startKey 的 Hash (sha256) 值，确保唯一性
- cf：RocksDB 的 ColumnFamily（默认为 `default` 或 `write`）

8.1.2 部署使用 BR 工具

8.1.2.1 推荐部署配置

- 推荐 BR 部署在 PD 节点上。
- 推荐使用一块高性能 SSD 网盘，挂载到 BR 节点和所有 TiKV 节点上，网盘推荐万兆网卡，否则带宽有可能成为备份恢复时的性能瓶颈。

> 注意：
>
> - 如果没有挂载网盘或者使用其他共享存储，那么 BR 备份的数据会生成在各个 TiKV 节点上。由于 BR 只备份 leader 副本，所以各个节点预留的空间需要根据 leader size 来预估。
> - 同时由于 TiDB 默认使用 leader count 进行平衡，所以会出现 leader size 差别大的问题，导致各个节点备份数据不均衡。

8.1.2.2 使用限制

下面是使用 BR 进行备份恢复的几条限制：

- BR 恢复到 TiCDC / Drainer 的上游集群时，恢复数据无法由 TiCDC / Drainer 同步到下游。
- BR 只支持在 new_collations_enabled_on_first_bootstrap 开关值相同的集群之间进行操作。这是因为 BR 仅备份 KV 数据。如果备份集群和恢复集群采用不同的排序规则，数据校验会不通过。所以恢复集群时，你需要确保 `select VARIABLE_VALUE from mysql.tidb where VARIABLE_NAME='new_collation_enabled'` 语句的开关值查询结果与备份时的查询结果相一致，才可以进行恢复。

8.1.2.3 兼容性

BR 和 TiDB 集群的兼容性问题分为以下两方面：

- BR 部分版本和 TiDB 集群的接口不兼容

BR 在 v5.4.0 之前不支持恢复 `charset=GBK` 的表。并且，任何版本的 BR 都不支持恢复 `charset=GBK` 的表到 5.4.0 之前的 TiDB 集群。

- 某些功能在开启或关闭状态下，会导致 KV 格式发生变化，因此备份和恢复期间如果没有统一开启或关闭，就会带来不兼容的问题

下表整理了会导致 KV 格式发生变化的功能。

功能	相关 issue	解决方式
聚簇索引	#565	确保备份时 `tidb_enable_clustered_index` 全局变量和恢复时一致，否则会导致数据不一致的问题，例如 `default not found` 和数据索引不一致。
New collation	#352	确保恢复时集群的 `new_collations_enabled_on_first_bootstrap` 变量值和备份时的一致，否则会导致数据索引不一致和 checksum 通不过。

功能	相关 issue	解决方式
恢复集群开启 TiCDC 同步	#364	TiKV 暂不能将 BR ingest 的 SST 文件下推到 TiCDC，因此使用 BR 恢复时候需要关闭 TiCDC。
全局临时表		确保使用 BR v5.3.0 及以上版本进行备份和恢复，否则会导致全局临时表的表定义错误。

在上述功能确保备份恢复一致的前提下，BR 和 TiKV/TiDB/PD 还可能因为版本内部协议不一致/接口不一致出现不兼容的问题，因此 BR 内置了版本检查。

8.1.2.3.1 版本检查

BR 内置版本会在执行备份和恢复操作前，对 TiDB 集群版本和自身版本进行对比检查。如果大版本不匹配（比如 BR v4.x 和 TiDB v5.x 上），BR 会提示退出。如要跳过版本检查，可以通过设置 `--check-requirements=false` 强行跳过版本检查。

需要注意的是，跳过检查可能会遇到版本不兼容的问题。BR 和 TiDB 各版本兼容情况如下表所示：

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR nightly 备份 TiDB nightly	✓	✓	✓（如果恢复了使用非整数类型聚簇主键的表到 v4.0 的 TiDB 集群，BR 会无任何警告地导致数据错误）

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR v5.0 备份 TiDB v5.0	✓	✓	✓（如果恢复了使用非整数类型聚簇主键的表到 v4.0 的 TiDB 集群，BR 会无任何警告地导致数据错误）

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR v4.0 备份 TiDB v4.0	□	□	□（如果 TiKV >= v4.0.0-rc.1，BR 包含 #233 Bug 修复，且 TiKV 不包含 #7241 Bug 修复，那么 BR 会导致 TiKV 节点重启）
用 BR nightly 或 v5.0 备份 TiDB v4.0	□（当 TiDB 版本小于 v4.0.9 时会出现 #609 问题）	□（当 TiDB 版本小于 v4.0.9 会出现 #609 问题）	□（当 TiDB 版本小于 v4.0.9 会出现 #609 问题）

8.1.2.4 备份和恢复 mysql 系统库下的表数据（实验特性）

> **警告：**
>
> 该功能目前为实验性功能，尚未经过充分测试。不建议在生产环境中使用该功能。

在 v5.1.0 之前，BR 备份时会过滤掉系统库 `mysql.*` 的表数据。自 v5.1.0 起，BR 默认备份集群内的全部数据，包括系统库 `mysql.*` 中的数据。但由于恢复 `mysql.*` 中系统表数据的技术实现尚不完善，因此 BR 默认不恢复系统库 `mysql` 中的表数据。

如果你希望将系统表（如 `mysql.usertable1`）的数据恢复到系统库 `mysql` 中，可以设置`filter` 参数 来过滤表名 (`-f "mysql.usertable1"`)。完成设置后，系统库下的表数据会恢复到临时库中，然后通过对临时库表进行重命名的方式恢复到系统库。

由于技术原因，以下系统表不能被正确恢复。即使你指定了 `-f "mysql.*"`，以下表也不会被恢复：

- 统计信息相关的表：“stats_buckets”，“stats_extended”，“stats_feedback”，“stats_fm_sketch”，“stats_histograms”，“stats_meta”， “stats_top_n”
- 权限或系统相关的表：“tidb”，“global_variables”，“columns_priv”，“db”，“default_roles”，“global_grants”，“global_priv”， “role_edges”， “tables_priv”， “user”， “gc_delete_range”， “gc_delete_range_done”，“schema_index_usage”

8.1.2.5 运行 BR 的最低机型配置要求

运行 BR 的最低机型配置要求如下：

CPU	内存	硬盘类型	网络
1 核	4 GB	HDD	千兆网卡

一般场景下（备份恢复的表少于 1000 张），BR 在运行期间的 CPU 消耗不会超过 200%，内存消耗不会超过 4 GB。但在备份和恢复大量数据表时，BR 的内存消耗可能会上升到 4 GB 以上。在实际测试中，备份 24000 张表大概需要消耗 2.7 GB 内存，CPU 消耗维持在 100% 以下。

8.1.2.6 最佳实践

下面是使用 BR 进行备份恢复的几种推荐操作：

- 推荐在业务低峰时执行备份操作，这样能最大程度地减少对业务的影响。
- BR 支持在不同拓扑的集群上执行恢复，但恢复期间对在线业务影响很大，建议低峰期或者限速 (`rate-limit`) 执行恢复。
- BR 备份最好串行执行。不同备份任务并行会导致备份性能降低，同时也会影响在线业务。
- BR 恢复最好串行执行。不同恢复任务并行会导致 Region 冲突增多，恢复的性能降低。
- 推荐在 `-s` 指定的备份路径上挂载一个共享存储，例如 NFS。这样能方便收集和管理备份文件。
- 在使用共享存储时，推荐使用高吞吐的存储硬件，因为存储的吞吐会限制备份或恢复的速度。
- BR 默认会分别在备份、恢复完成后，进行一轮数据校验，将文本数据同集群数据比较，来保证正确性。在迁移场景下，为了减少迁移耗时，推荐备份时关闭校验 (`--checksum=false`)，只在恢复时开启校验。

8.1.2.7 使用方式

目前支持以下几种方式来运行 BR 工具，分别是通过 SQL 语句、命令行工具或在 Kubernetes 环境下进行备份恢复。

8.1.2.7.1 通过 SQL 语句

TiDB 支持使用 SQL 语句BACKUP 和RESTORE 进行备份恢复。如果要查看备份恢复的进度，你可以使用 SHOW BACKUPS|RESTORES 语句。

8.1.2.7.2 通过命令行工具

TiDB 支持使用 BR 命令行工具进行备份恢复（需手动下载）。关于 BR 命令行工具的具体使用方法，请参阅使用备份与恢复工具 BR。

8.1.2.7.3 在 Kubernetes 环境下

目前支持使用 BR 工具备份 TiDB 集群数据到兼容 S3 的存储、 Google Cloud Storage 以及持久卷，并作恢复：

> **注意：**
>
> Amazon S3 和 Google Cloud Storage (GCS) 参数描述见外部存储文档。

- 备份 TiDB 集群数据到兼容 S3 的存储
- 恢复 S3 兼容存储上的备份数据
- 备份 TiDB 集群到 Google Cloud Storage
- 恢复 Google Cloud Storage 上的备份数据
- 备份 TiDB 集群到持久卷
- 恢复持久卷上的备份数据

8.1.3 BR 相关文档

- 使用 BR 命令行备份恢复
- BR 备份与恢复场景示例
- BR 常见问题
- 外部存储

8.2 使用 BR 命令行进行备份恢复

本文介绍如何 BR 命令行进行 TiDB 集群数据的备份和恢复。

在阅读本文前，请确保你已通读备份与恢复工具 BR 简介，尤其是使用限制和最佳实践这两节。

8.2.1 BR 命令行描述

一条 br 命令是由子命令、选项和参数组成的。子命令即不带 - 或者 -- 的字符。选项即以 - 或者 -- 开头的字符。参数即子命令或选项字符后紧跟的、并传递给命令和选项的字符。

以下是一条完整的 br 命令行：

```
br backup full --pd "${PDIP}:2379" -s "local:///tmp/backup"
```

命令行各部分的解释如下：

- backup：br 的子命令
- full：backup 的子命令
- -s 或 --storage：备份保存的路径
- "local:///tmp/backup"：-s 的参数，保存的路径为各个 TiKV 节点本地磁盘的 /tmp/backup
- --pd：PD 服务地址
- "${PDIP}:2379"：--pd 的参数

> 注意：
>
> 在使用 local storage 的时候，备份数据会分散在各个节点的本地文件系统中。
>
> 不建议在生产环境中备份到本地磁盘，因为在日后恢复的时候，必须手动聚集这些数据才能完成恢复工作（见恢复集群数据）。
>
> 聚集这些备份数据可能会造成数据冗余和运维上的麻烦，而且在不聚集这些数据便直接恢复的时候会遇到颇为迷惑的 SST file not found 报错。
>
> 建议在各个节点挂载 NFS 网盘，或者直接备份到 S3 对象存储中。

8.2.1.1 命令和子命令

BR 由多层命令组成。目前，BR 包含 backup、restore 和 version 三个子命令：

- br backup 用于备份 TiDB 集群
- br restore 用于恢复 TiDB 集群

以上三个子命令可能还包含这些子命令：

- full：可用于备份或恢复全部数据。
- db：可用于备份或恢复集群中的指定数据库。
- table：可用于备份或恢复集群指定数据库中的单张表。

8.2.1.2 常用选项

- --pd：用于连接的选项，表示 PD 服务地址，例如 "${PDIP}:2379"。
- -h/--help：获取所有命令和子命令的使用帮助。例如 br backup --help。
- -V（或 --version）：检查 BR 版本。
- --ca：指定 PEM 格式的受信任 CA 的证书文件路径。
- --cert：指定 PEM 格式的 SSL 证书文件路径。
- --key：指定 PEM 格式的 SSL 证书密钥文件路径。
- --status-addr：BR 向 Prometheus 提供统计数据的监听地址。

8.2.2 使用 BR 命令行备份集群数据

使用 `br backup` 命令来备份集群数据。可选择添加 `full` 或 `table` 子命令来指定备份的范围：全部集群数据或单张表的数据。

8.2.2.1 备份全部集群数据

要备份全部集群数据，可使用 `br backup full` 命令。该命令的使用帮助可以通过 `br backup full -h` 或 `br backup full --help` 来获取。

用例：将所有集群数据备份到各个 TiKV 节点的 `/tmp/backup` 路径，同时也会将备份的元信息文件 `backupmeta` 写到该路径下。

> 注意：
>
> - 经测试，在全速备份的情况下，如果备份盘和服务盘不同，在线备份会让只读线上服务的 QPS 下降 15%~25% 左右。如果希望降低影响，请参考 `--ratelimit` 进行限速。
> - 假如备份盘和服务盘相同，备份将会和服务争夺 I/O 资源，这可能会让只读线上服务的 QPS 骤降一半以上。请尽量禁止将在线服务的数据备份到 TiKV 的数据盘。

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
```

以上命令中，`--ratelimit` 选项限制了每个 TiKV 执行备份任务的速度上限（单位 MiB/s）。`--log-file` 选项指定把 BR 的 log 写到 `backupfull.log` 文件中。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。进度条效果如下：

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
Full Backup <---------/.....................................................> 17.12%.
```

8.2.2.2 备份单个数据库的数据

要备份集群中指定单个数据库的数据，可使用 `br backup db` 命令。同样可通过 `br backup db -h` 或 `br backup db --help` 来获取子命令 db 的使用帮助。

用例：将数据库 `test` 备份到各个 TiKV 节点的 `/tmp/backup` 路径，同时也会将备份的元信息文件 `backupmeta` 写到该路径下。

```
br backup db \
    --pd "${PDIP}:2379" \
    --db test \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backuptable.log
```

db 子命令的选项为 --db，用来指定数据库名。其他选项的含义与备份全部集群数据相同。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。

8.2.2.3 备份单张表的数据

要备份集群中指定单张表的数据，可使用 br backup table 命令。同样可通过 br backup table -h 或 br backup table --help 来获取子命令 table 的使用帮助。

用例：将表 test.usertable 备份到各个 TiKV 节点的 /tmp/backup 路径，同时也会将备份的元信息文件 backupmeta 写到该路径下。

```
br backup table \
    --pd "${PDIP}:2379" \
    --db test \
    --table usertable \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backuptable.log
```

table 子命令有 --db 和 --table 两个选项，分别用来指定数据库名和表名。其他选项的含义与备份全部集群数据相同。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。

8.2.2.4 使用表库过滤功能备份多张表的数据

如果你需要以更复杂的过滤条件来备份多个表，执行 br backup full 命令，并使用 --filter 或 -f 来指定表库过滤规则。

用例：以下命令将所有 db*.tbl* 形式的表格数据备份到每个 TiKV 节点上的 /tmp/backup 路径，并将 backupmeta 文件写入该路径。

```
br backup full \
    --pd "${PDIP}:2379" \
    --filter 'db*.tbl*' \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
```

8.2.2.5 备份数据到 Amazon S3 后端存储

如果备份的存储并不是在本地，而是在 Amazon 的 S3 后端存储，那么需要在 storage 子命令中指定 S3 的存储路径，并且赋予 BR 节点和 TiKV 节点访问 Amazon S3 的权限。

这里可以参照 AWS 官方文档在指定的 Region 区域中创建一个 S3 桶 Bucket，如果有需要，还可以参照 AWS 官方文档在 Bucket 中创建一个文件夹 Folder。

> 注意：
>
> 要完成一次备份，通常 TiKV 和 BR 需要的最小权限为 s3:ListBucket，s3:PutObject 和 s3:AbortMultipartUpload。

将有权限访问该 S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 BR 节点，并且通过 BR 将权限传给 TiKV 节点。

```
export AWS_ACCESS_KEY_ID=${AccessKey}
export AWS_SECRET_ACCESS_KEY=${SecretKey}
```

在进行 BR 备份时，显示指定参数 --s3.region 和 --send-credentials-to-tikv, --s3.region 表示 S3 存储所在的区域，--send-credentials-to-tikv 表示将 S3 的访问权限传递给 TiKV 节点。

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "s3://${Bucket}/${Folder}" \
    --s3.region "${region}" \
    --send-credentials-to-tikv=true \
    --ratelimit 128 \
    --log-file backuptable.log
```

8.2.2.6 增量备份

如果想要备份增量，只需要在备份的时候指定上一次的备份时间戳 --lastbackupts 即可。

注意增量备份有以下限制：

- 增量备份需要与前一次全量备份在不同的路径下
- GC safepoint 必须在 lastbackupts 之前

```
br backup full\
    --pd ${PDIP}:2379 \
    --ratelimit 128 \
    -s local:///home/tidb/backupdata/incr \
    --lastbackupts ${LAST_BACKUP_TS}
```

以上命令会备份 (LAST_BACKUP_TS, current PD timestamp] 之间的增量数据。

你可以使用 validate 指令获取上一次备份的时间戳，示例如下：

```
LAST_BACKUP_TS=`br validate decode --field="end-version" -s local:///home/tidb/backupdata | tail
    ↪ -n1`
```

示例备份的增量数据记录 (LAST_BACKUP_TS, current PD timestamp] 之间的数据变更，以及这段时间内的 DDL。在恢复的时候，BR 会先把所有 DDL 恢复，而后才会恢复数据。

8.2.2.7 加密备份数据（实验性功能）

自 TiDB v5.3.0 起，TiDB 开始支持备份加密功能，你可配置下列参数在备份过程中到达加密数据的效果：

- --crypter.method：加密算法，支持 aes128-ctr/aes192-ctr/aes256-ctr 三种算法，缺省值为 plaintext，表示不加密
- --crypter.key：加密密钥，十六进制字符串格式，aes128-ctr 对应 128 位（16 字节）密钥长度，aes192-ctr 为 24 字节，aes256-ctr 为 32 字节
- --crypter.key-file：密钥文件，可直接将存放密钥的文件路径作为参数传入，此时 crypter.key 不需要传入

> 警告：
>
> - 当前该功能为实验特性，不建议在生产环境中使用。
> - 密钥丢失，备份的数据将无法恢复到集群中。
> - 加密功能需在 br 工具和 TiDB 集群都不低于 v5.3.0 的版本上使用，且加密备份得到的数据无法在低于 v5.3.0 版本的集群上恢复。

备份加密的示例如下：

```
br backup full\
    --pd ${PDIP}:2379 \
    -s local:///home/tidb/backupdata/incr \
    --crypter.method aes128-ctr \
    --crypter.key 0123456789abcdef0123456789abcdef
```

8.2.2.8 Raw KV 备份（实验性功能）

> 警告：
>
> Raw KV 备份功能还在实验中，没有经过完备的测试。暂时请避免在生产环境中使用该功能。

在某些使用场景下，TiKV 可能会独立于 TiDB 运行。考虑到这点，BR 也提供跳过 TiDB 层，直接备份 TiKV 中数据的功能：

```
br backup raw --pd $PD_ADDR \
    -s "local://$BACKUP_DIR" \
    --start 31 \
    --ratelimit 128 \
    --end 3130303030303030 \
    --format hex \
    --cf default
```

以上命令会备份 default CF 上 `[0x31, 0x3130303030303030)` 之间的所有键到 `$BACKUP_DIR` 去。

这里，`--start` 和 `--end` 的参数会先依照 `--format` 指定的方式解码，再被送到 TiKV 上去，目前支持以下解码方式：

- “raw”：不进行任何操作，将输入的字符串直接编码为二进制格式的键。
- “hex”：将输入的字符串视作十六进制数字。这是默认的编码方式。
- “escape”：对输入的字符串进行转义之后，再编码为二进制格式。

8.2.3 使用 BR 命令行恢复集群数据

使用 `br restore` 命令来恢复备份数据。可选择添加 `full`、`db` 或 `table` 子命令来指定恢复操作的范围：全部集群数据、某个数据库或某张数据表。

> 注意：
>
> 如果使用本地存储，在恢复前必须将所有备份的 SST 文件复制到各个 TiKV 节点上 `--storage` 指定的目录下。
>
> 即使每个 TiKV 节点最后只需要读取部分 SST 文件，这些节点也需要有所有 SST 文件的完全访问权限。原因如下：
>
> - 数据被复制到了多个 Peer 中。在读取 SST 文件时，这些文件必须要存在于所有 Peer 中。这与数据的备份不同，在备份时，只需从单个节点读取。
> - 在数据恢复的时候，每个 Peer 分布的位置是随机的，事先并不知道哪个节点将读取哪个文件。
>
> 使用共享存储可以避免这些情况。例如，在本地路径上安装 NFS，或使用 S3。利用这些网络存储，各个节点都可以自动读取每个 SST 文件，此时上述注意事项不再适用。
>
> 同时，请注意同一时间对同一个集群只能运行一个恢复任务，否则可能会出现非预期的行为，详见FAQ。

8.2.3.1 恢复全部备份数据

要将全部备份数据恢复到集群中来，可使用 `br restore full` 命令。该命令的使用帮助可以通过 `br restore full -h` 或 `br restore full --help` 来获取。

用例：将 `/tmp/backup` 路径中的全部备份数据恢复到集群中。

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file restorefull.log
```

以上命令中，`--ratelimit` 选项限制了每个 TiKV 执行恢复任务的速度上限（单位 MiB/s）。`--log-file` 选项指定把 BR 的 log 写到 `restorefull.log` 文件中。

恢复期间还有进度条会在终端中显示，当进度条前进到 100% 时，说明恢复已完成。在完成恢复后，BR 为了确保数据安全性，还会校验恢复数据。进度条效果如下：

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file restorefull.log
Full Restore <---------/.....................................................> 17.12%.
```

8.2.3.2 恢复单个数据库的数据

要将备份数据中的某个数据库恢复到集群中，可以使用 `br restore db` 命令。该命令的使用帮助可以通过 `br restore db -h` 或 `br restore db --help` 来获取。

用例：将 `/tmp/backup` 路径中备份数据中的某个数据库恢复到集群中。

```
br restore db \
    --pd "${PDIP}:2379" \
    --db "test" \
    --ratelimit 128 \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

以上命令中 `--db` 选项指定了需要恢复的数据库名字。其余选项的含义与恢复全部备份数据相同。

> **注意：**
>
> 恢复备份数据的时候，`--db` 选项指定的数据库名必须与执行备份时候 `--db`选项指定的数据库名相同，否则无法恢复成功。由于备份数据的元文件`backupmeta` 记录了该数据库名，因此只能将数据恢复到同名的数据库。推荐做法是把备份文件恢复到另一个集群的同名数据库中。

8.2.3.3 恢复单张表的数据

要将备份数据中的某张数据表恢复到集群中，可以使用 `br restore table` 命令。该命令的使用帮助可通过 `br restore table -h` 或 `br restore table --help` 来获取。

用例：将 `/tmp/backup` 路径下的备份数据中的某个数据表恢复到集群中。

```
br restore table \
    --pd "${PDIP}:2379" \
    --db "test" \
    --table "usertable" \
    --ratelimit 128 \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

8.2.3.4 使用表库功能过滤恢复数据

如果你需要用复杂的过滤条件来恢复多个表，执行 `br restore full` 命令，并用 `--filter` 或 `-f` 指定使用表库过滤。

用例：以下命令将备份在 `/tmp/backup` 路径的表的子集恢复到集群中。

```
br restore full \
    --pd "${PDIP}:2379" \
    --filter 'db*.tbl*' \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

8.2.3.5 从 Amazon S3 后端存储恢复数据

如果需要恢复的数据并不是存储在本地，而是在 Amazon 的 S3 后端，那么需要在 `storage` 子命令中指定 S3 的存储路径，并且赋予 BR 节点和 TiKV 节点访问 Amazon S3 的权限。

> **注意：**
>
> 要完成一次恢复，通常 TiKV 和 BR 需要的最小权限为 `s3:ListBucket` 和 `s3:GetObject`。

将有权限访问该 S3 后端存储的账号的 `SecretKey` 和 `AccessKey` 作为环境变量传入 BR 节点，并且通过 BR 将权限传给 TiKV 节点。

```
export AWS_ACCESS_KEY_ID=${AccessKey}
export AWS_SECRET_ACCESS_KEY=${SecretKey}
```

在进行 BR 恢复时，显示指定参数 `--s3.region` 和 `--send-credentials-to-tikv`, `--s3.region` 表示 S3 存储所在的区域，`--send-credentials-to-tikv` 表示将 S3 的访问权限传递给 TiKV 节点。`--storage`参数中的 `Bucket` 和 `Folder` 分别代表需要恢复的数据所在的 S3 存储桶和文件夹。

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "s3://${Bucket}/${Folder}" \
    --s3.region "${region}" \
    --ratelimit 128 \
    --send-credentials-to-tikv=true \
    --log-file restorefull.log
```

以上命令中 `--table` 选项指定了需要恢复的表名。其余选项的含义与恢复单个数据库相同。

8.2.3.6 增量恢复

增量恢复的方法和使用 BR 进行全量恢复的方法并无差别。需要注意，恢复增量数据的时候，需要保证备份时指定的 `last backup ts` 之前备份的数据已经全部恢复到目标集群。

8.2.3.7 恢复创建在 mysql 数据库下的表（实验性功能）

BR 可以并且会默认备份 `mysql` 数据库下的表。

在执行恢复时，`mysql` 下的表默认不会被恢复。如果需要恢复 `mysql` 下的用户创建的表，可以通过table filter 来显式地包含目标表。以下示例中要恢复目标用户表为 `mysql.usertable`；该命令会在执行正常的恢复的同时恢复 `mysql.usertable`。

```
br restore full -f '*.*' -f '!mysql.*' -f 'mysql.usertable' -s $external_storage_url --ratelimit
    ↪ 128
```

在如上的命令中，`-f '*.*'` 用于覆盖掉默认的规则，`-f '!mysql.*'` 指示 BR 不要恢复 `mysql` 中的表，除非另有指定。`-f 'mysql.usertable'` 则指定需要恢复 `mysql.usertable`。具体原理请参考table filter 的文档。

如果只需要恢复 `mysql.usertable`，而无需恢复其他表，可以使用以下命令：

```
br restore full -f 'mysql.usertable' -s $external_storage_url --ratelimit 128
```

> **警告：**
>
> 系统表（例如 `mysql.tidb`）可以通过 BR 进行备份。但恢复系统表存在限制。即便是使用了 `-filter` 设置，也不能通过 BR 恢复以下系统表：
>
> - 统计信息表（`mysql.stat_*`）
> - 系统变量表（`mysql.tidb`，`mysql.global_variables`）
> - 用户信息表（`mysql.user`，`mysql.columns_priv`，等等）
> - 其他系统表
>
> 恢复系统表可能还存在更多兼容性问题。为了防止意外发生，请避免在生产环境中恢复系统表。

8.2.3.8 解密恢复数据（实验性功能）

> **警告：**
>
> - 当前该功能为实验特性，不建议在生产环境中使用。

在对数据做加密备份后，恢复操作需传入相应的解密参数，解密算法或密钥不正确则无法恢复，解密参数和加密参数一致即可。解密恢复的示例如下：

```
br restore full\
    --pd ${PDIP}:2379 \
    -s local:///home/tidb/backupdata/incr \
    --crypter.method aes128-ctr \
    --crypter.key 0123456789abcdef0123456789abcdef
```

8.2.3.9 Raw KV 恢复（实验性功能）

> **警告：**
>
> Raw KV 恢复功能还在实验中，没有经过完备的测试。暂时请避免在生产环境中使用该功能。

和Raw KV 备份相似地，恢复 Raw KV 的命令如下：

```
br restore raw --pd $PD_ADDR \
    -s "local://$BACKUP_DIR" \
    --start 31 \
    --end 3130303030303030 \
    --ratelimit 128 \
    --format hex \
    --cf default
```

以上命令会将范围在 `[0x31, 0x3130303030303030)` 的已备份键恢复到 TiKV 集群中。这里键的编码方式和备份时相同。

8.2.3.10 在线恢复（实验性功能）

> **警告：**
>
> 在线恢复功能还在实验中，没有经过完备的测试，同时还依赖 PD 的不稳定特性 Placement Rules。暂时请避免在生产环境中使用该功能。

在恢复的时候，写入过多的数据会影响在线集群的性能。为了尽量避免影响线上业务，BR 支持通过Placement rules 隔离资源。让下载、导入 SST 的工作仅仅在指定的几个节点（下称“恢复节点”）上进行，具体操作如下：

1. 配置 PD，启动 Placement rules：

```
echo "config set enable-placement-rules true" | pd-ctl
```

2. 编辑恢复节点 TiKV 的配置文件，在 server 一项中指定：

```
[server]
labels = { exclusive = "restore" }
```

3. 启动恢复节点的 TiKV，使用 BR 恢复备份的文件，和非在线恢复相比，这里只需要加上 --online 标志即可：

```
br restore full \
    -s "local://$BACKUP_DIR" \
    --ratelimit 128 \
    --pd $PD_ADDR \
    --online
```

8.3 BR 备份与恢复场景示例

BR 是一款分布式的快速备份和恢复工具。

本文展示了以下几种备份和恢复场景下的 BR 操作过程：

- 将单表数据备份到网络盘（推荐生产环境使用）
- 从网络磁盘恢复备份数据（推荐生产环境使用）
- 将单表数据备份到本地磁盘（推荐测试环境试用）
- 从本地磁盘恢复备份数据（推荐测试环境试用）

以帮助读者达到以下目标：

- 正确使用网络盘或本地盘进行备份或恢复
- 通过相关监控指标了解备份或恢复的状态
- 了解在备份或恢复时如何调优性能
- 处理备份时可能发生的异常

8.3.1 目标读者

你需要对 TiDB 和 TiKV 有一定的了解。

在阅读本文前，请确保你已通读备份与恢复工具 BR 简介，尤其是使用限制和最佳实践这两节。

8.3.2 环境准备

本节介绍 TiDB 的推荐部署方式、BR 使用示例中的集群版本、TiKV 集群硬件信息和集群配置。

你可以根据自己的硬件和配置来预估备份恢复的性能。

8.3.2.1 部署方式

推荐使用TiUP 部署 TiDB 集群，再下载TiDB Toolkit 获取 BR 工具。

8.3.2.2 集群版本

- TiDB: v5.0.0
- TiKV: v5.0.0
- PD: v5.0.0
- BR: v5.0.0

> **注意：**
>
> v5.0.0 为编写本文档时的最新版本。推荐使用最新版本 TiDB/TiKV/PD/BR，同时需要确保 BR 版本和 TiDB 相同。

8.3.2.3 TiKV 集群硬件信息

- 操作系统：CentOS Linux release 7.6.1810 (Core)
- CPU：16-Core Common KVM processor
- RAM：32GB
- 硬盘：500G SSD * 2
- 网卡：万兆网卡

8.3.2.4 配置

BR 可以直接将命令下发到 TiKV 集群来执行备份和恢复，不依赖 TiDB server 组件，因此无需对 TiDB server 进行配置。

- TiKV: 默认配置
- PD : 默认配置

8.3.3 使用场景

本节描述以下几种使用场景：

- 将单表数据备份到网络盘（推荐生产环境使用）

- 从网络磁盘恢复备份数据（推荐生产环境使用）
- 将单表数据备份到本地磁盘（推荐测试环境试用）
- 从本地磁盘恢复备份数据（推荐测试环境试用）

推荐使用网络盘来进行备份和恢复操作，这样可以省去收集备份数据文件的繁琐步骤。尤其在 TiKV 集群规模较大的情况下，使用网络盘可以大幅提升操作效率。

在使用 BR 进行备份或恢复操作前，需要先进行如下准备工作。

8.3.3.1 备份前的准备工作

BR 工具已支持自适应 GC，会自动将 backupTS（默认是最新的 PD timestamp）注册到 PD 的 `safePoint`，保证 TiDB 的 GC Safe Point 在备份期间不会向前移动，即可避免手动设置 GC。

关于 `br backup` 命令的具体使用方法，参见使用备份与恢复工具 BR。

运行 `br backup` 命令前，请确保以下条件：

1. TiDB 集群中没有正在运行中的 DDL。
2. 用于创建备份的存储设备有足够的空间。

8.3.3.2 恢复前的准备工作

使用 BR 进行恢复前的准备工作如下：

运行`br restore` 命令前，需要检查新集群，确保集群内没有同名的表。

8.3.3.3 将单表数据备份到网络盘（推荐生产环境使用）

使用 `br backup` 命令，将单表数据 `--db batchmark --table order_line` 备份到指定的网络盘路径 `local:///br_data` 下。

8.3.3.3.1 前置要求

- 备份前的准备工作。
- 配置一台高性能 SSD 硬盘主机为 NFS server 存储数据。其他所有 BR 节点，TiKV 节点和 TiFlash 节点为 NFS client，挂载相同的路径（例如 `/br_data`）到 NFS server 上以访问 NFS server。
- NFS server 和 NFS client 间的数据传输速率至少要达到备份集群的 `TiKV 实例数 * 150MB/s`。否则网络 I/O 有可能成为性能瓶颈。

> 注意：
>
> - 因为备份时候只备份单副本 (leader) 数据，所以即使集群中存在 TiFlash 副本，无需挂载 TiFlash 节点 BR 也能完成备份。
> - BR 在恢复数据时，会恢复全部副本的数据。因此在恢复时，TiFlash 节点需要有备份数据的访问权限 BR 才能完成恢复，此时也必须将 TiFlash 节点挂载到 NFS server 上。

8.3.3.3.2 部署拓扑

部署拓扑如下图所示：

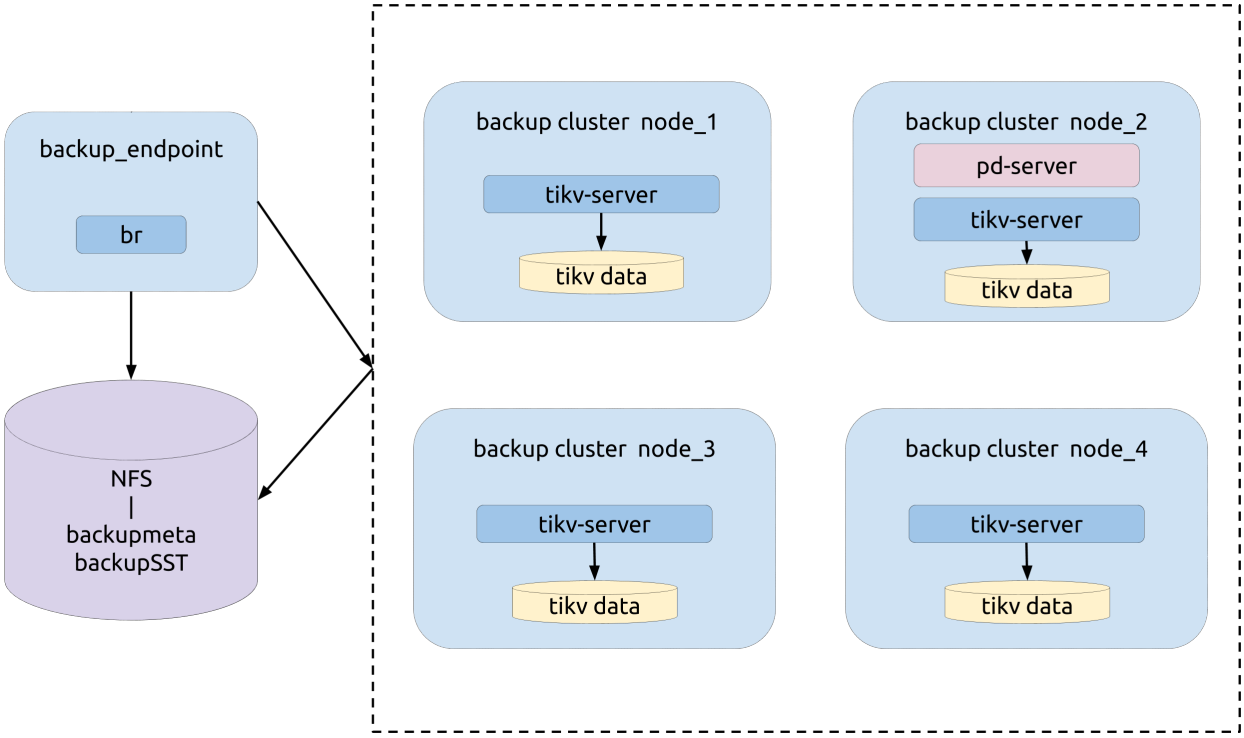


图 20: img

8.3.3.3.3 运行备份

备份操作前，在 TiDB 中使用 `admin checksum table order_line` 命令获得备份目标表 `--db batchmark --table order_line` 的统计信息。统计信息示例如下：

```
+-----------+------------+---------------------+------------+--------------+
| Db_name   | Table_name | Checksum_crc64_xor  | Total_kvs  | Total_bytes  |
+-----------+------------+---------------------+------------+--------------+
| batchmark | order_line | 10912722838344822475 | 5659888624 | 370385538778 |
+-----------+------------+---------------------+------------+--------------+
1 row in set (5 min 47.59 sec)
```

图 21: img

运行 BR 备份命令：

```
bin/br backup table \
    --db batchmark \
```

```
    --table order_line \
    -s local:///br_data \
    --pd ${PD_ADDR}:2379 \
    --log-file backup-nfs.log
```

8.3.3.3.4 备份过程中的运行指标

在 BR 备份过程中，需关注以下监控面版中的运行指标来了解备份的状态。

Backup CPU Utilization：参与备份的 TiKV 节点（例如 backup-worker 和 backup-endpoint）的 CPU 使用率。

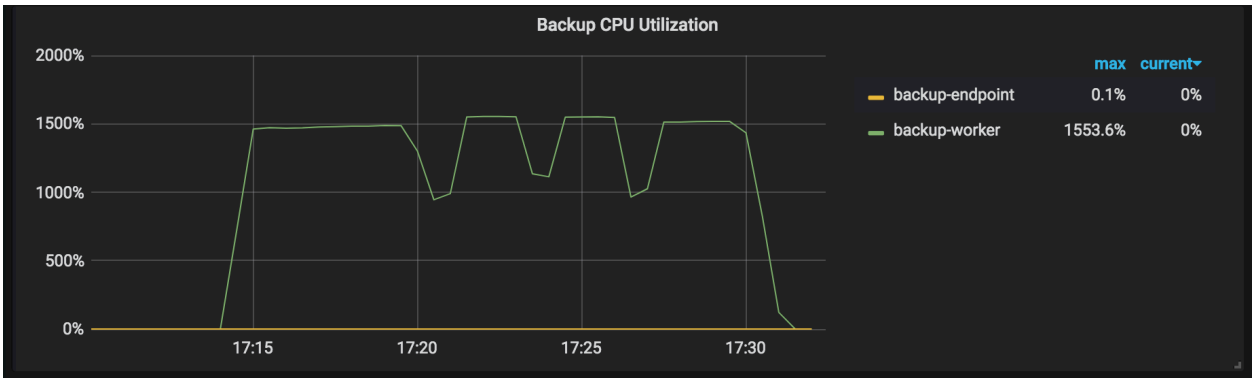


图 22: img

IO Utilization：参与备份的 TiKV 节点的 I/O 使用率。

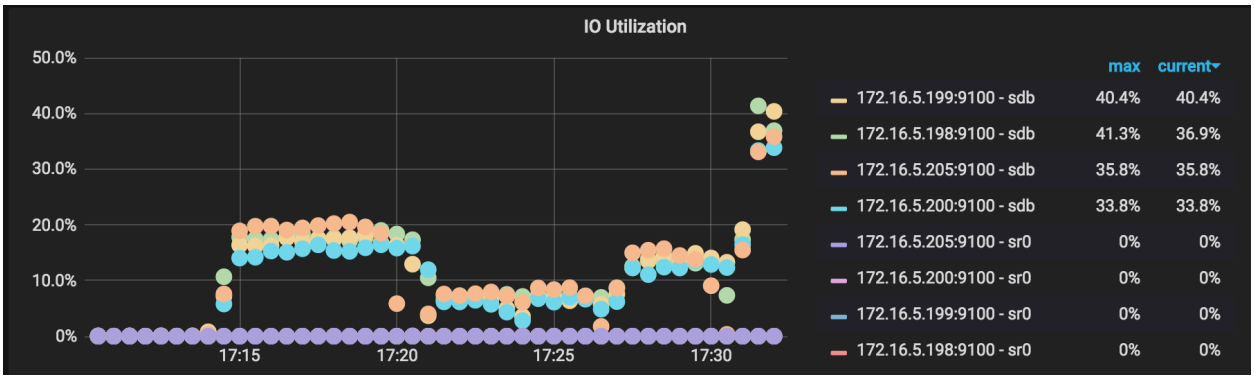


图 23: img

BackupSST Generation Throughput：参与备份的 TiKV 节点生成 backupSST 文件的吞吐。正常时单个 TiKV 节点的吞吐在 150MB/s 左右。

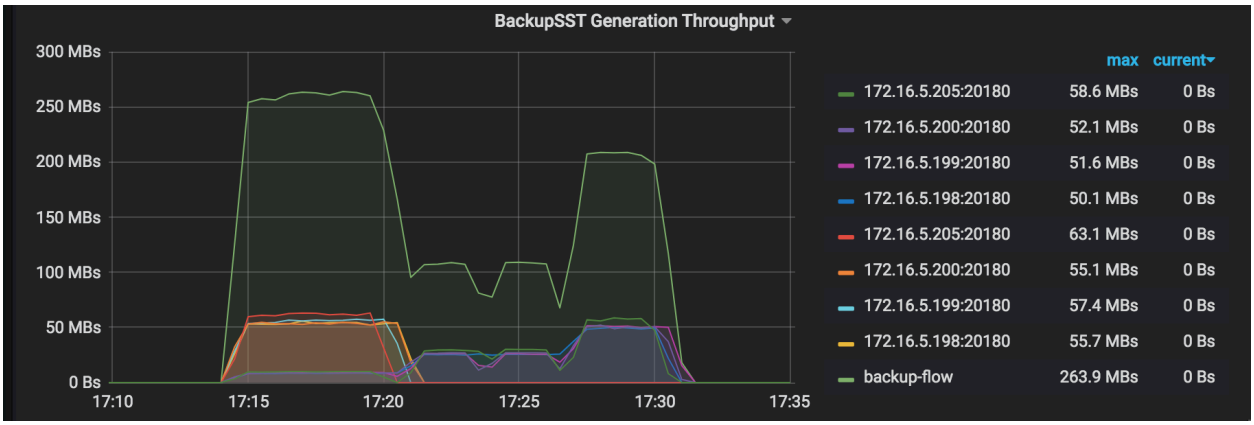


图 24: img

One Backup Range Duration：备份一个 range 的操作耗时，包括扫描耗时 (scan KV) 和保存耗时（保存为 backupSST 文件）。

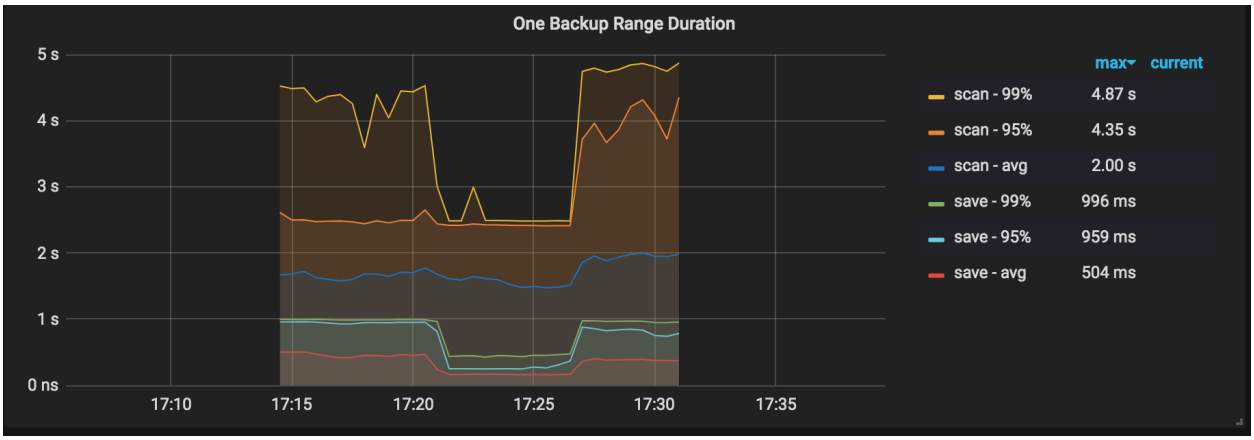


图 25: img

One Backup Subtask Duration：一次备份任务会被拆分成多个子任务。该监控项显示子任务的耗时。

注意：

- 虽然本次任务是备份单表，但因为表中有 3 个索引，所以正常会拆分成 4 个子任务。
- 下图中有 13 个点，说明有 9 次 (13-4) 重试。备份过程中可能发生 Region 调度行为，少量重试是正常的。

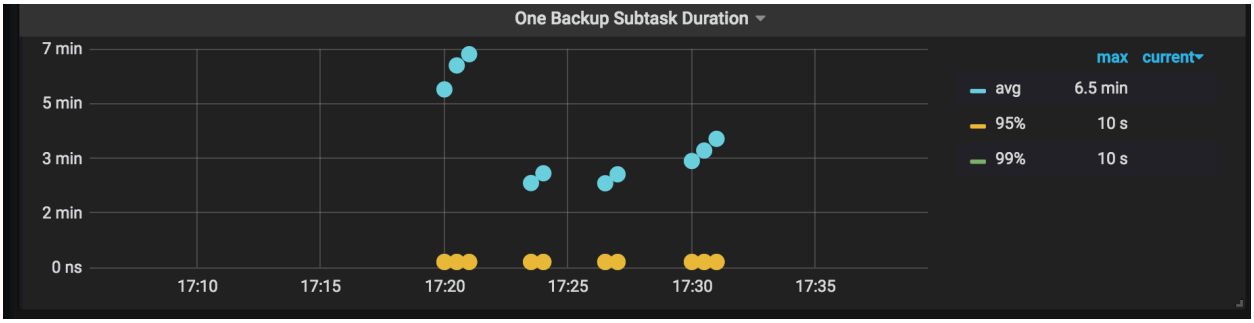


图 26: img

Backup Errors：备份过程中的错误。正常时无错误。即使出现少量错误，备份操作也有重试机制，可能会导致备份时间增加，但不会影响备份的正确性。

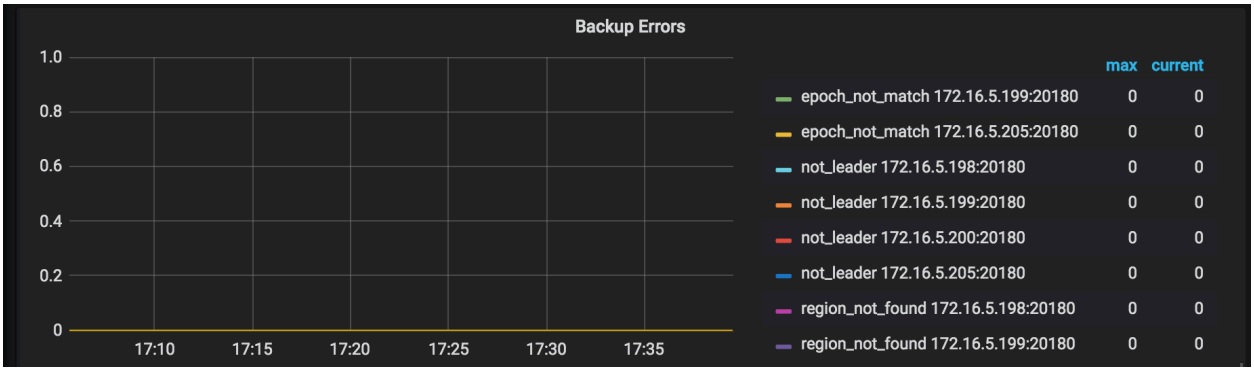


图 27: img

Checksum Request Duration：对备份集群执行 admin checksum 的耗时统计。

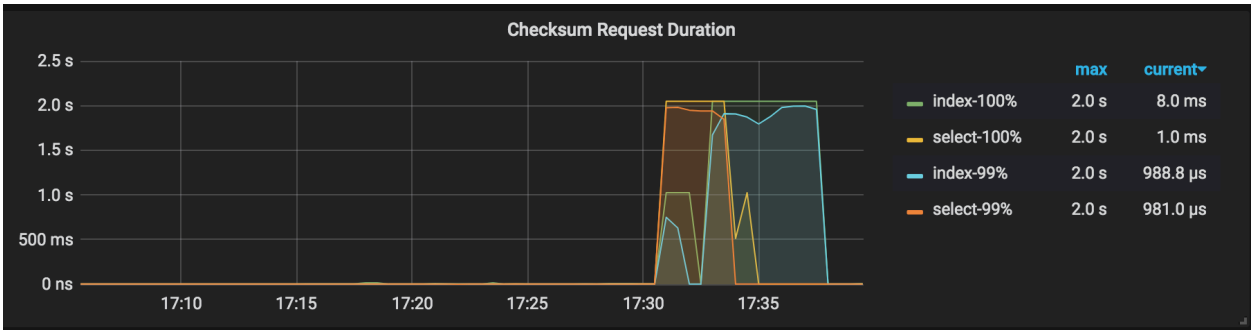


图 28: img

8.3.3.3.5 结果解读

BR 会在备份结束时输出备份总结到控制台。

同时使用 BR 前已设置日志的存放路径。从路径下存放的日志中可以获取此次备份的相关统计信息。在日志中搜关键字 “summary”，可以看到以下信息：

```
["Full backup Success summary:
    total backup ranges: 2,
    total success: 2,
    total failed: 0,
    total take(Full backup time): 31.802912166s,
    total take(real time): 49.799662427s,
    total size(MB): 5997.49,
    avg speed(MB/s): 188.58,
    total kv: 120000000"]
    ["backup checksum"=17.907153678s]
    ["backup fast checksum"=349.333µs]
    ["backup total regions"=43]
    [BackupTS=422618409346269185]
    [Size=826765915]
```

以上日志信息中包含以下内容：

- 备份耗时：`total take(Full backup time): 31.802912166s`
- 程序运行总耗时：`total take(real time): 49.799662427s`
- 备份数据大小：`total size(MB): 5997.49`
- 备份吞吐：`avg speed(MB/s): 188.58`
- 备份 KV 对数：`total kv: 120000000`
- 校验耗时：`["backup checksum"=17.907153678s]`
- 计算各表 checksum、KV 和 bytes 信息总和的耗时：`["backup fast checksum"=349.333µs]`
- 备份 Region 总数：`["backup total regions"=43]`
- 备份存档经压缩后在磁盘中的实际大小：`[Size=826765915]`
- 备份存档的快照时间戳：`[BackupTS=422618409346269185]`

通过以上数据可以计算得到单个 TiKV 实例的吞吐为：`avg speed(MB/s)/tikv_count = 62.86`。

8.3.3.3.6 性能调优

如果 TiKV 的资源使用没有出现明显的瓶颈（例如备份过程中的运行指标中的 Backup CPU Utilization 最高为 1500% 左右，IO Utilization 普遍低于 30%），可以尝试调大 `--concurrency`（默认是 4）参数以进行性能调优。该方法不适用于存在许多小表的场景。

示例如下：

```
bin/br backup table \
    --db batchmark \
    --table order_line \
    -s local:///br_data/ \
    --pd ${PD_ADDR}:2379 \
    --log-file backup-nfs.log \
    --concurrency 16
```

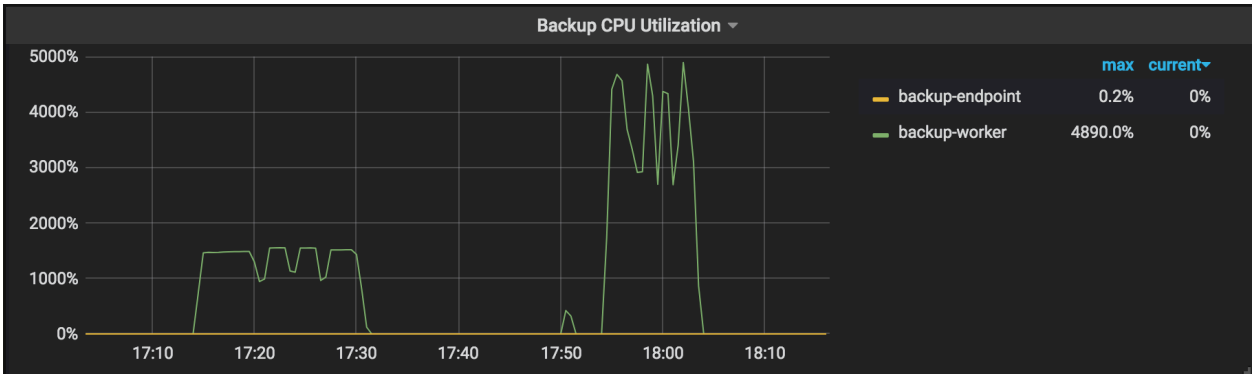


图 29: img

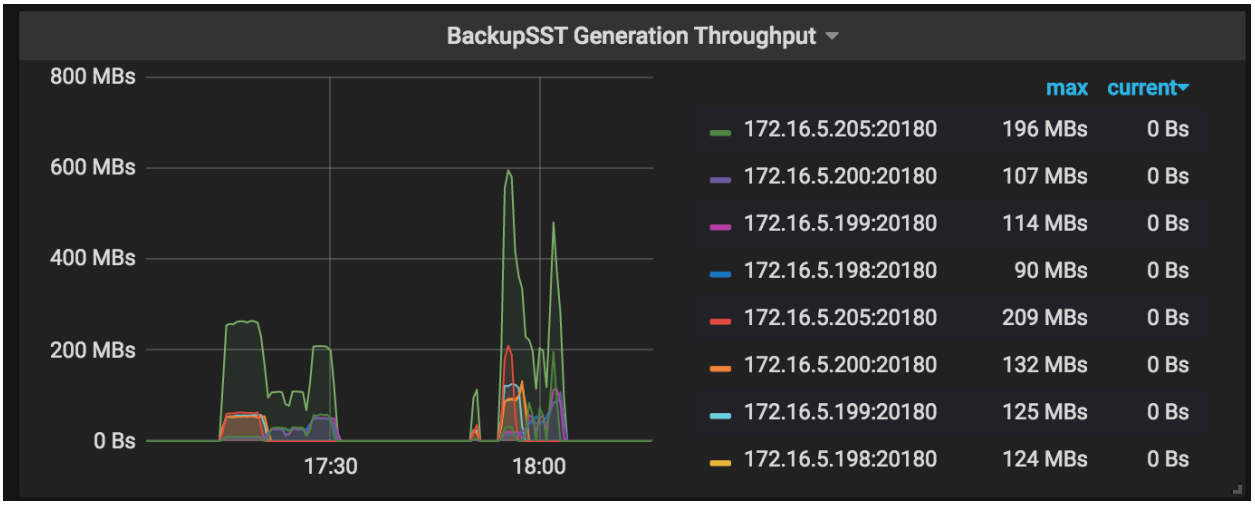


图 30: img

性能调优后的结果如下所示（保持数据大小不变）：

- 备份耗时：`total take(s)` 从 `986.43` 减少到 `535.53`
- 数据大小：`total size(MB): 353227.18`
- 备份吞吐：`avg speed(MB/s)` 从 `358.09` 提升到 `659.59`
- 单个 TiKV 实例的吞吐：`avg speed(MB/s)/tikv_count` 从 `89` 提升到 `164.89`

8.3.3.4 从网络磁盘恢复备份数据（推荐生产环境使用）

使用 `br restore` 命令，将一份完整的备份数据恢复到一个离线集群。暂不支持恢复到在线集群。

8.3.3.4.1 前置要求

- 恢复前的准备工作。

8.3.3.4.2 部署拓扑

部署拓扑如下图所示：

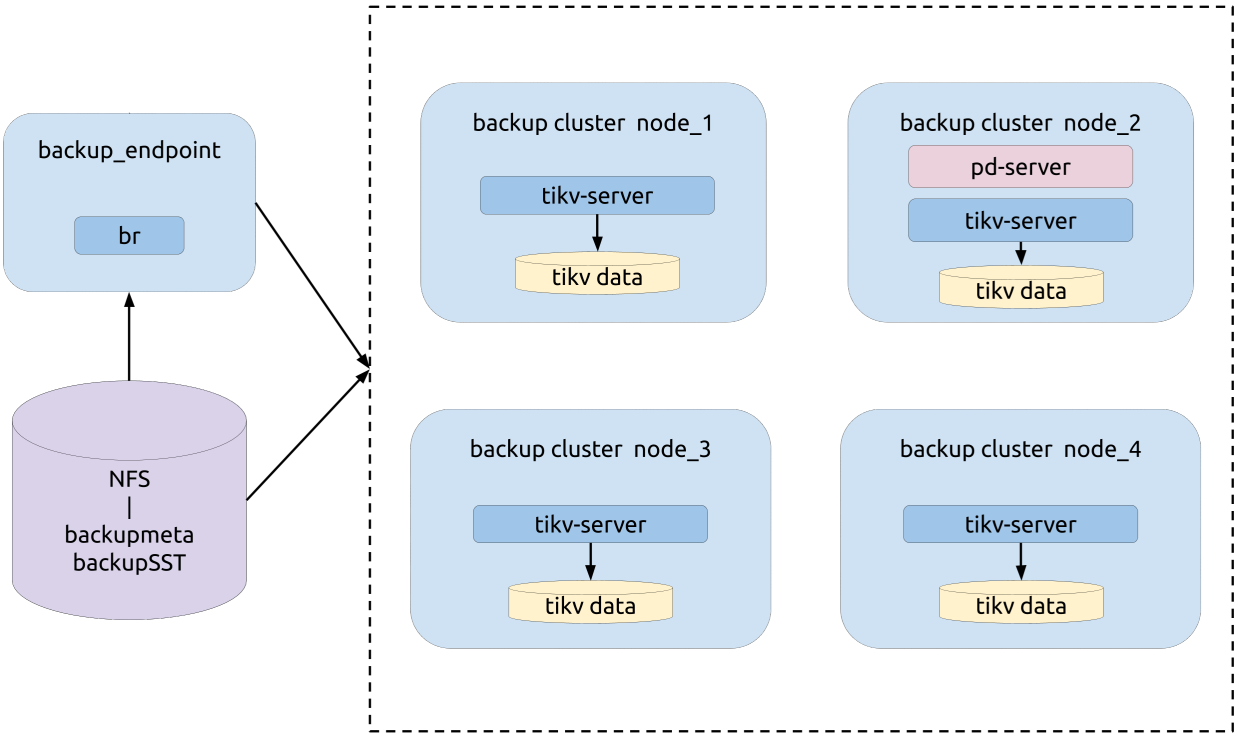


图 31: img

8.3.3.4.3 运行恢复

运行 `br restore` 命令：

```
bin/br restore table --db batchmark --table order_line -s local:///br_data --pd 172.16.5.198:2379
    ↪  --log-file restore-nfs.log
```

8.3.3.4.4 恢复过程中的运行指标

在 BR 恢复过程中，需关注以下监控面版中的运行指标来了解恢复的状态。

CPU Utilization：参与恢复的 TiKV 节点 CPU 使用率。

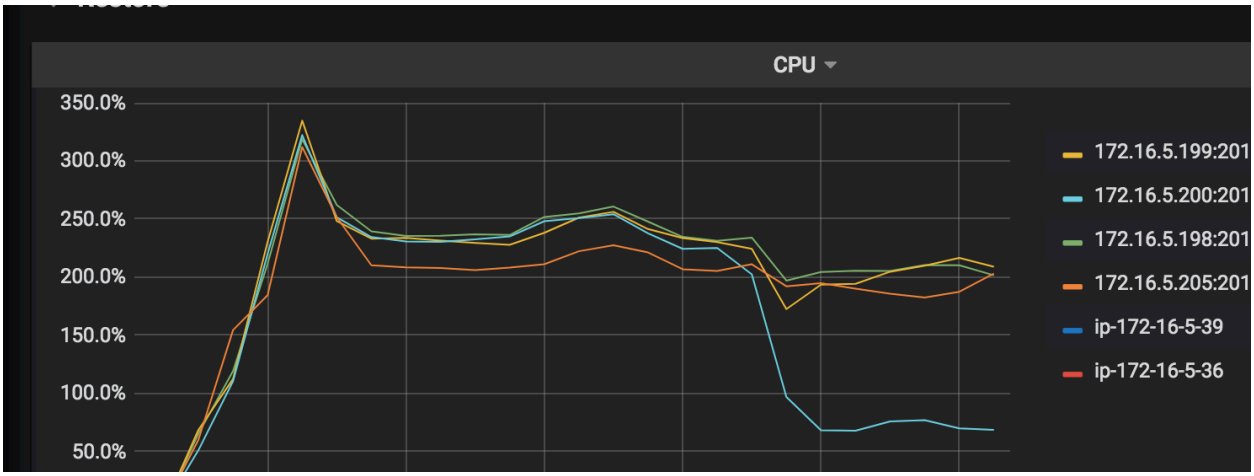


图 32: img

IO Utilization：参与恢复的 TiKV 节点的 I/O 使用率。

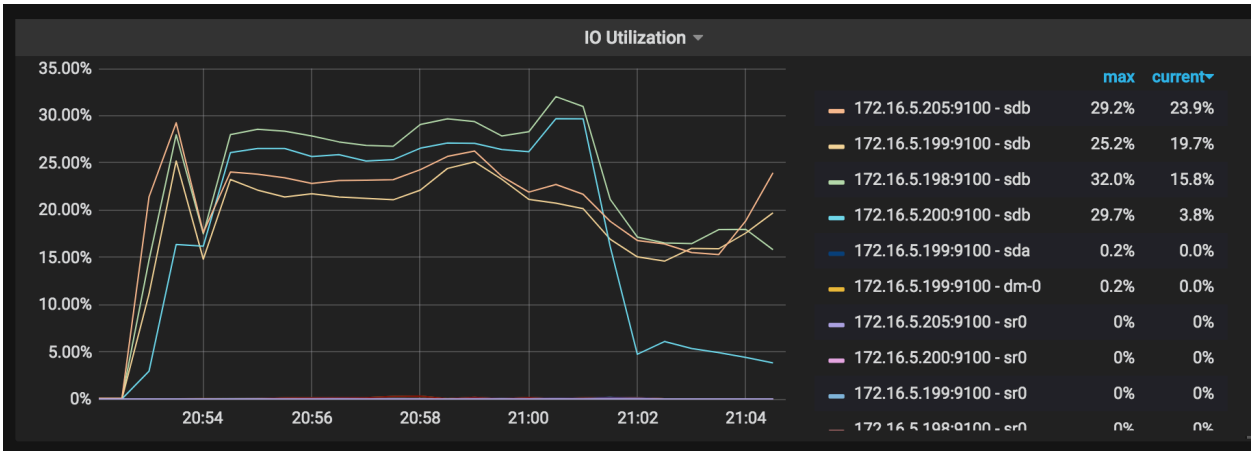


图 33: img

Region 分布：Region 分布越均匀，说明恢复资源利用越充分。

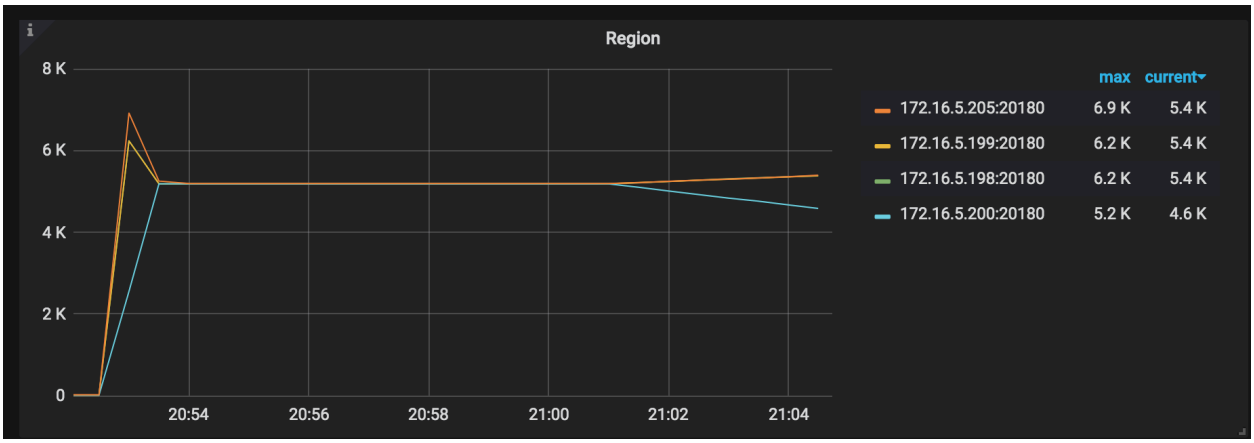


图 34: img

Process SST Duration：处理 SST 文件的延迟。恢复一张表时，如果 `tableID` 发生了变化，需要对 `tableID` 进行 `rewrite`，否则会进行 `rename`。通常 `rewrite` 延迟要高于 `rename` 的延迟。

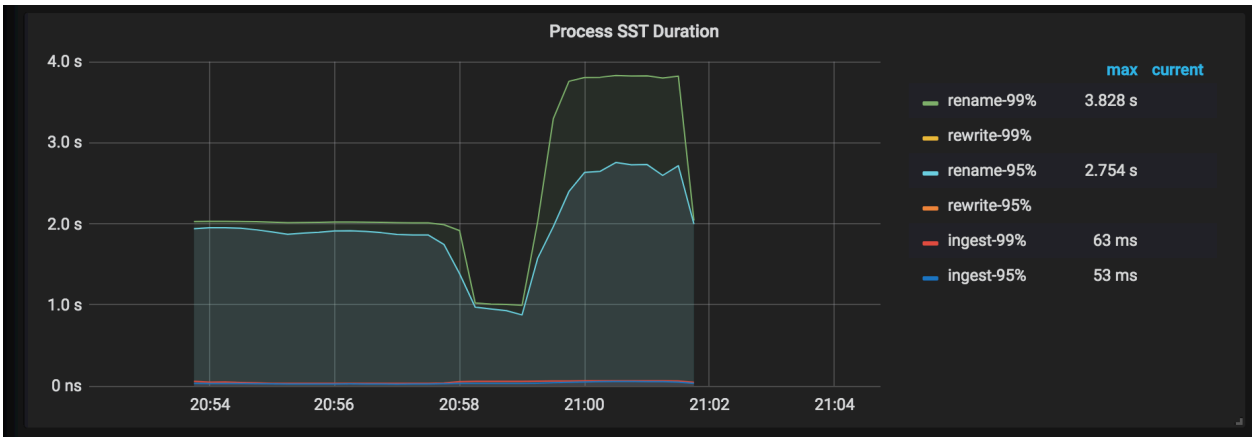


图 35: img

DownLoad SST Throughput：从 External Storage 下载 SST 文件的吞吐。

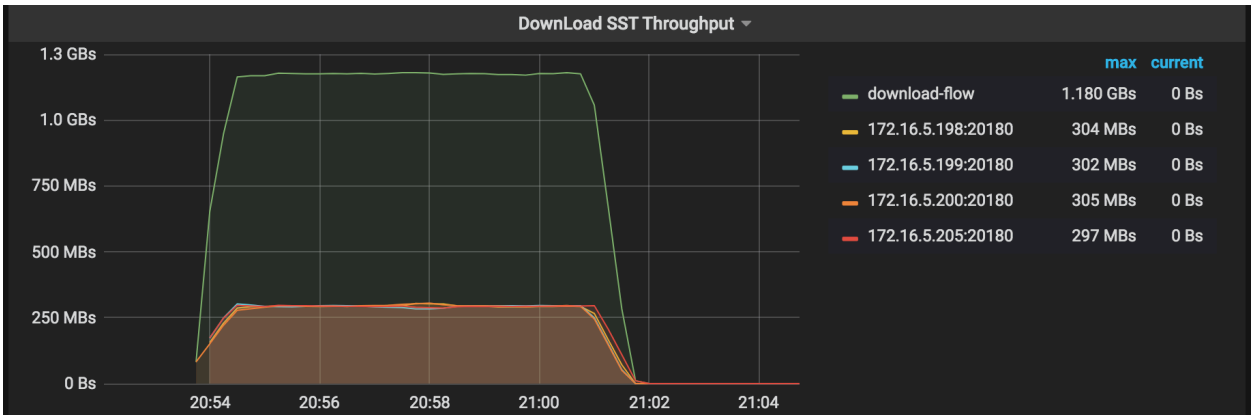


图 36: img

Restore Errors：恢复过程中的错误。

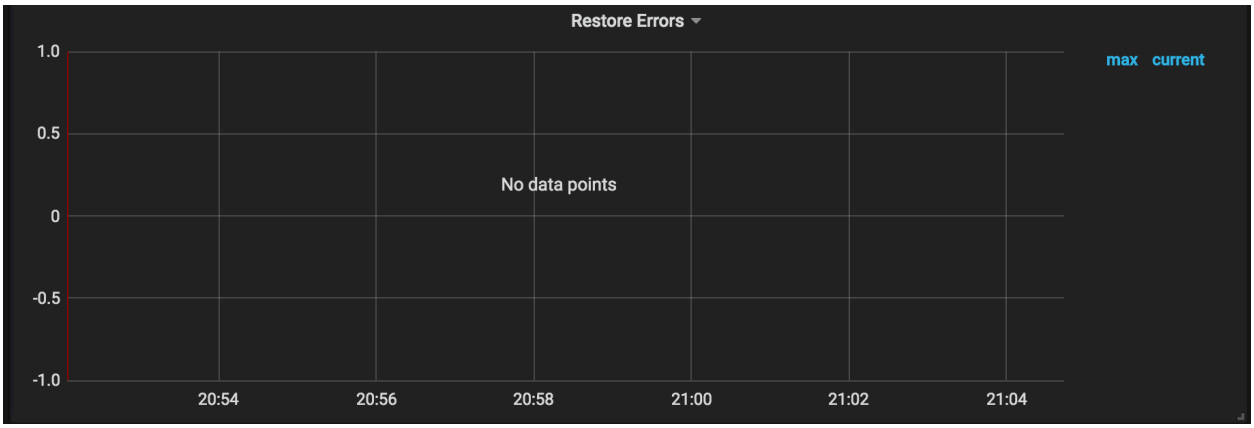


图 37: img

Checksum Request duration：对恢复集群执行 admin checksum 的耗时统计，会比备份时的 checksum 延迟高。

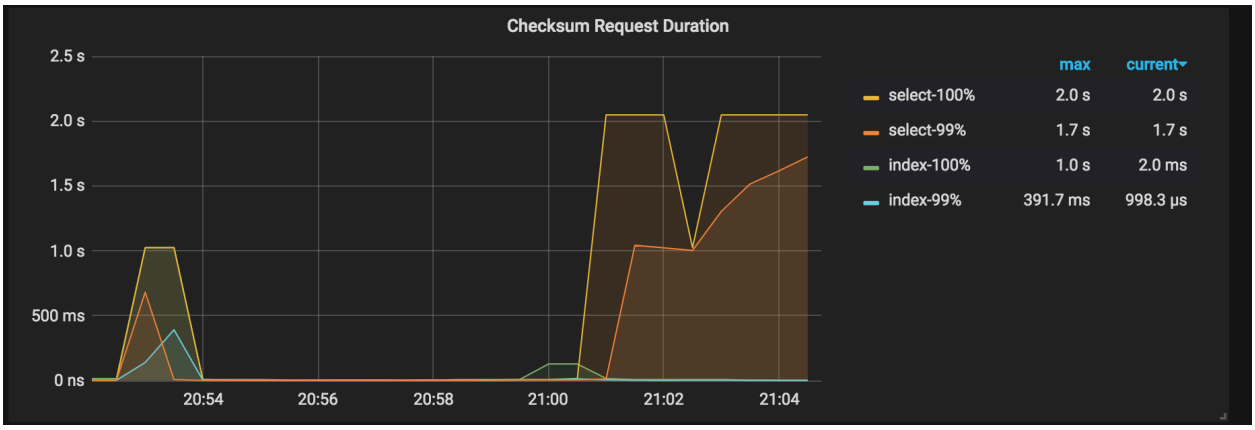


图 38: img

8.3.3.4.5 结果解读

使用 BR 前已设置日志的存放路径。从路径下存放的日志中可以获取此次恢复的相关统计信息。在日志中搜关键字“summary”，可以看到以下信息：

```
["Table Restore summary:
    total restore tables: 1,
    total success: 1,
    total failed: 0,
    total take(Full restore time): 17m1.001611365s,
    total take(real time): 16m1.371611365s,
    total kv: 5659888624,
    total size(MB): 353227.18,
    avg speed(MB/s): 367.42"]
    ["restore files"=9263]
    ["restore ranges"=6888]
    ["split region"=49.049182743s]
    ["restore checksum"=6m34.879439498s]
    [Size=48693068713]
```

以上日志信息中包含以下内容：

- 恢复耗时：`total take(Full restore time): 17m1.001611365s`
- 程序运行总耗时：`total take(real time): 16m1.371611365s`
- 恢复数据大小：`total size(MB): 353227.18`
- 恢复 KV 对数：`total kv: 5659888624`
- 恢复吞吐：`avg speed(MB/s): 367.42`
- Region Split 耗时：`take=49.049182743s`
- 校验耗时：`restore checksum=6m34.879439498s`
- 恢复存档在磁盘中的实际大小：`[Size=48693068713]`

根据上表数据可以计算得到：

- 单个 TiKV 吞吐：avg speed(MB/s)/tikv_count = 91.8
- 单个 TiKV 平均恢复速度：total size(MB)/(split time + restore time)/tikv_count = 87.4

8.3.3.4.6 性能调优

如果 TiKV 资源使用没有明显的瓶颈，可以尝试调大 --concurrency 参数（默认为 128），示例如下：

```
bin/br restore table --db batchmark --table order_line -s local:///br_data/ --pd
    ↪ 172.16.5.198:2379 --log-file restore-concurrency.log --concurrency 1024
```

性能调优后的结果如下表所示（保持数据大小不变）：

- 恢复耗时：total take(s) 从 961.37 减少到 443.49
- 恢复吞吐：avg speed(MB/s) 从 367.42 提升到 796.47
- 单个 TiKV 实例的吞吐：avg speed(MB/s)/tikv_count 从 91.8 提升到 199.1
- 单个 TiKV 实例的平均恢复速度：total size(MB)/(split time + restore time)/tikv_count 从 87.4 提升到 162.3

8.3.3.5 将单表数据备份到本地磁盘（推荐测试环境试用）

使用 br backup 命令，将单表数据 --db batchmark --table order_line 备份到指定的本地磁盘路径 local ↪ :///home/tidb/backup_local 下。

8.3.3.5.1 前置要求

- 备份前的准备工作。
- 各个 TiKV 节点有单独的磁盘用来存放 backupSST 数据。
- backup_endpoint 节点有单独的磁盘用来存放备份的 backupmeta 文件。
- TiKV 和 backup_endpoint 节点需要有相同的备份目录，例如 /home/tidb/backup_local。

8.3.3.5.2 部署拓扑

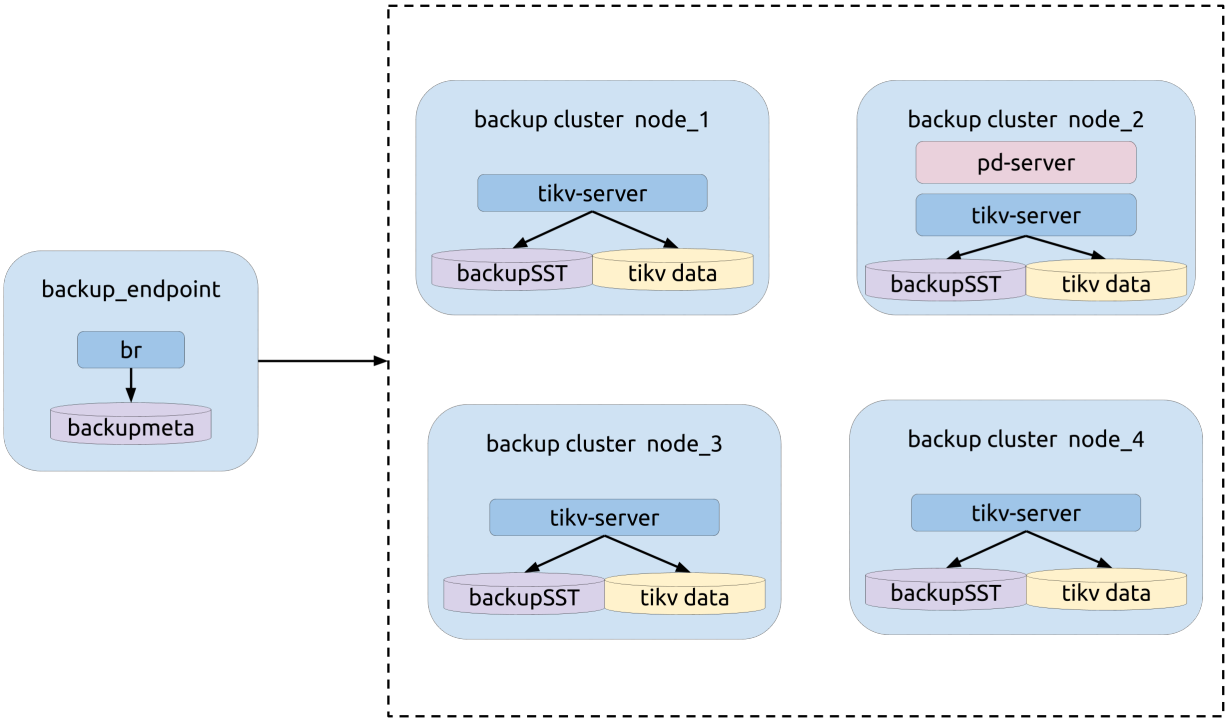


图 39: img

8.3.3.5.3 运行备份

备份前在 TiDB 里通过 `admin checksum table order_line` 获得备份的目标表 `--db batchmark --table order_line` 的统计信息。统计信息示例如下：

```
+----------+------------+---------------------+-----------+-------------+
| Db_name  | Table_name | Checksum_crc64_xor  | Total_kvs | Total_bytes |
+----------+------------+---------------------+-----------+-------------+
| batchmark | order_line | 10912722838344822475 | 5659888624 | 370385538778 |
+----------+------------+---------------------+-----------+-------------+
1 row in set (5 min 47.59 sec)
```

图 40: img

运行 `br backup` 命令：

```
bin/br backup table \
    --db batchmark \
    --table order_line \
    -s local:///home/tidb/backup_local/ \
    --pd ${PD_ADDR}:2379 \
    --log-file backup_local.log
```

运行备份时，参考备份过程中的运行指标对相关指标进行监控，以了解备份状态。

8.3.3.5.4 结果解读

使用 BR 前已设置日志的存放路径。从该路径下存放的日志获取此次备份的相关统计信息。在日志中搜关键字 “summary”，可以看到以下信息：

```
["Table backup summary: total backup ranges: 4, total success: 4, total failed: 0, total take(s):
    ↪  551.31, total kv: 5659888624, total size(MB): 353227.18, avg speed(MB/s): 640.71"] ["
    ↪ backup total regions"=6795] ["backup checksum"=6m33.962719217s] ["backup fast checksum
    ↪ "=22.995552ms]
```

以上日志信息中包含以下内容：

- 备份耗时：`total take(s): 551.31`
- 数据大小：`total size(MB): 353227.18`
- 备份吞吐：`avg speed(MB/s): 640.71`
- 校验耗时：`take=6m33.962719217s`

根据上表数据可以计算得到单个 TiKV 实例的吞吐：`avg speed(MB/s)/tikv_count` = 160。

8.3.3.6 从本地磁盘恢复备份数据（推荐测试环境试用）

使用 `br restore` 命令，将一份完整的备份数据恢复到一个离线集群。暂不支持恢复到在线集群。

8.3.3.6.1 前置要求

- 恢复前的准备工作。
- 集群中没有与备份数据相同的库表。目前 BR 不支持 table route。
- 集群中各个 TiKV 节点有单独的磁盘用来存放要恢复的 backupSST 数据。
- `restore_endpoint` 节点有单独的磁盘用来存放要恢复的 `backupmeta` 文件。
- 集群中 TiKV 和 `restore_endpoint` 节点需要有相同的备份目录，例如 `/home/tidb/backup_local/`。

如果备份数据存放在本地磁盘，那么需要执行以下的步骤：

1. 汇总所有 backupSST 文件到一个统一的目录下。
2. 将汇总后的 backupSST 文件到复制到集群的所有 TiKV 节点下。
3. 将 `backupmeta` 文件复制到到 `restore endpoint` 节点下。

8.3.3.6.2 部署拓扑

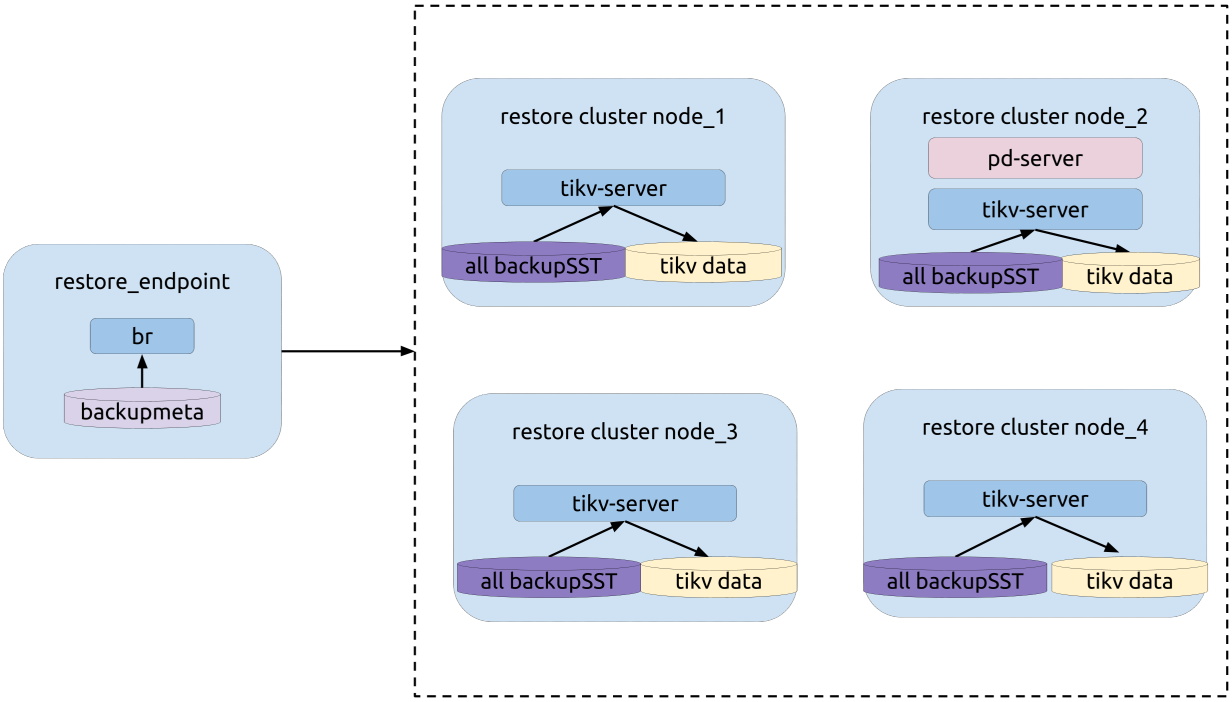


图 41: img

8.3.3.6.3 运行恢复

运行 `br restore` 命令：

```
bin/br restore table --db batchmark --table order_line -s local:///home/tidb/backup_local/ --pd
    ↪ 172.16.5.198:2379 --log-file restore_local.log
```

运行恢复时，参考恢复过程中的运行指标对相关指标进行监控，以了解恢复状态。

8.3.3.6.4 结果解读

使用 BR 前已设置日志的存放路径。从该日志中可以获取此次恢复的相关统计信息。在日志中搜关键字“summary”，可以看到以下信息：

```
["Table Restore summary: total restore tables: 1, total success: 1, total failed: 0, total take(s
    ↪ ): 908.42, total kv: 5659888624, total size(MB): 353227.18, avg speed(MB/s): 388.84"] ["
    ↪ restore files"=9263] ["restore ranges"=6888] ["split region"=58.7885518s] ["restore
    ↪ checksum"=6m19.349067937s]
```

以上日志信息中包含以下内容：

- 恢复耗时：`total take(s): 908.42`
- 数据大小：`total size(MB): 353227.18`
- 恢复吞吐：`avg speed(MB/s): 388.84`

- Region Split 耗时：take=58.7885518s
- 校验耗时：take=6m19.349067937s

根据上表数据可以计算得到：

- 单个 TiKV 实例的吞吐：avg speed(MB/s)/tikv_count = 97.2
- 单个 TiKV 实例的平均恢复速度：total size(MB)/(split time + restore time)/tikv_count = 92.4

8.3.4 备份过程中的异常处理

本节介绍如何处理备份过程中出现的常见错误。

8.3.4.1 备份日志中出现 key locked Error

日志中的错误消息：log - ["backup occur kv error"][error="{\"KvError\":{\"locked\":

如果在备份过程中遇到 key 被锁住，目前 BR 会尝试清锁。少量报错不会影响备份的正确性。

8.3.4.2 备份失败

日志中的错误消息：log - Error: msg:"Io(Custom { kind: AlreadyExists, error: \"[5_5359_42_123_default
↪ .sst] is already exists in /dir/backup_local/\" })"

若备份失败并出现以上错误消息，采取以下其中一种操作后再重新备份：

- 更换备份数据目录。例如将 /dir/backup-2020-01-01/ 改为 /dir/backup_local/。
- 删除所有 TiKV 和 BR 节点的备份目录。

8.4 外部存储

8.4.1 外部存储

Backup & Restore (BR)、TiDB Lightning 和 Dumpling 皆支持在本地文件系统和 Amazon S3 上读写数据；另外 BR 亦支持 Google Cloud Storage (GCS) 和Azure Blob Storage (Azblob)。通过传入不同 URL scheme 到 BR 的 --storage (-s) 参数、TiDB Lightning 的 -d 参数及 Dumpling 中的 --output (-o) 参数，可以区分不同的存储方式。

8.4.1.1 Scheme

TiDB 迁移工具支持以下存储服务：

服务	Scheme	示例
本地文件系统（分布在各节点上）	local	local:///path/to/dest/
Amazon S3 及其他兼容 S3 的服务	s3	s3://bucket-name/prefix/of/dest/
GCS	gcs, gs	gcs://bucket-name/prefix/of/dest/
Azure Blob Storage	azure, azblob	azure://container-name/prefix/of/dest/
不写入任何存储（仅作为基准测试）	noop	noop://

8.4.1.2 URL 参数

S3、GCS 和 Azblob 等云存储有时需要额外的连接配置，你可以为这类配置指定参数。例如：

- 用 Dumpling 导出数据到 S3：

```
./dumpling -u root -h 127.0.0.1 -P 3306 -B mydb -F 256MiB \
    -o 's3://my-bucket/sql-backup?region=us-west-2'
```

- 用 TiDB Lightning 从 S3 导入数据：

```
./tidb-lightning --tidb-port=4000 --pd-urls=127.0.0.1:2379 --backend=local --sorted-kv-dir=/
    ↪ tmp/sorted-kvs \
    -d 's3://my-bucket/sql-backup?region=us-west-2'
```

- 用 TiDB Lightning 从 S3 导入数据（使用路径类型的请求模式）：

```
./tidb-lightning --tidb-port=4000 --pd-urls=127.0.0.1:2379 --backend=local --sorted-kv-dir=/
    ↪ tmp/sorted-kvs \
    -d 's3://my-bucket/sql-backup?force-path-style=true&endpoint=http://10.154.10.132:8088'
```

- 用 BR 备份到 GCS：

```
./br backup full -u 127.0.0.1:2379 \
    -s 'gcs://bucket-name/prefix'
```

- 用 BR 备份到 Azblob：

```
./br backup full -u 127.0.0.1:2379 \
    -s 'azure://container-name/prefix'
```

8.4.1.2.1 S3 的 URL 参数

URL 参数	描述
access-key	访问密钥
secret-access-key	secret 访问密钥
region	Amazon S3 服务区域（默认为 us-east-1）
use-accelerate-endpoint	是否在 Amazon S3 上使用加速端点（默认为 false）
endpoint	S3 兼容服务自定义端点的 URL（例如 https://s3.example.com/）
force-path-style	使用 path-style，而不是 virtual-hosted style（默认为 false）
storage-class	上传对象的存储类别（例如 STANDARD、STANDARD_IA）
sse	用于加密上传的服务器端加密算法（可以设置为空，AES256 或 aws:kms）
sse-kms-key-id	如果 sse 设置为 aws:kms，则使用该参数指定 KMS ID
acl	上传对象的 canned ACL（例如，private、authenticated-read）

> **注意：**
>
> 不建议在存储 URL 中直接传递访问密钥和 secret 访问密钥，因为这些密钥是明文记录的。迁移工具尝试按照以下顺序从环境中推断这些密钥：

1. `$AWS_ACCESS_KEY_ID` 和 `$AWS_SECRET_ACCESS_KEY` 环境变量。
2. `$AWS_ACCESS_KEY` 和 `$AWS_SECRET_KEY` 环境变量。
3. 工具节点上的共享凭证文件，路径由 `$AWS_SHARED_CREDENTIALS_FILE` 环境变量指定。
4. 工具节点上的共享凭证文件，路径为 `~/.aws/credentials`。
5. 当前 Amazon EC2 容器的 IAM 角色。
6. 当前 Amazon ECS 任务的 IAM 角色。

8.4.1.2.2 GCS 的 URL 参数

URL 参数	描述
`credentials-file`	迁移工具节点上的凭证 JSON 文件的路径
`storage-class`	上传对象的存储类别（例如 `STANDARD`、`COLDLINE`）
`predefined-acl`	上传对象的预定义 ACL（例如 `private`、`project-private`）

如果没有指定 `credentials-file`，迁移工具尝试按照以下顺序从环境中推断出凭证：

1. 工具节点上位于 `$GOOGLE_APPLICATION_CREDENTIALS` 环境变量所指定路径的文件内容。
2. 工具节点上位于 `~/.config/gcloud/application_default_credentials.json` 的文件内容。
3. 在 GCE 或 GAE 中运行时，从元数据服务器中获取的凭证。

8.4.1.2.3 Azblob 的 URL 参数

URL 参数	描述
`account-name`	存储账户名
`account-key`	访问密钥

URL 参数	描述
access-tier	上传对象的存储类别（例如 Hot、Cool、Archive）。如果没有设置 access-tier 的值（该值为空），此值会默认设置为 Hot。

为了保证 TiKV 和迁移工具使用了同一个存储账户，account-name 会由迁移工具决定（即默认 send-credentials-to-tikv = true）。迁移工具按照以下顺序推断密钥：

1. 如果已指定 account-name 和 account-key，则使用该参数指定的密钥。
2. 如果没有指定 account-key，则尝试从工具节点上的环境变量读取相关凭证。
    - 迁移工具会优先读取 $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET。与此同时，工具会允许 TiKV 从各自节点上读取上述三个环境变量，采用 Azure AD (Azure Active Directory) 访问。
        - $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET 分别代表 Azure 应用程序的应用程序 ID client_id，租户 ID tenant_id 和客户端密码 client_secret。
        - 如需了解如何确认运行环境中存在环境变量 $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET，或需要将环境变量配置为参数，请参考配置环境变量作为参数
3. 如果上述的三个环境变量不存在于工具节点中，则尝试读取 $AZURE_STORAGE_KEY，采用密钥访问。
    - 如需确认 BR 运行环境和 TiKV 运行环境中是否存在这三个环境变量，请参考配置环境变量作为参数。

> **注意：**
>
> 将 Azure Blob Storage 作为外部存储时，必须设置 send-credentials-to-tikv = true（即默认情况），否则会导致备份失败。

8.4.1.3 命令行参数

除了使用 URL 参数，BR 和 Dumpling 工具亦支持从命令行指定这些配置，例如：

```
./dumpling -u root -h 127.0.0.1 -P 3306 -B mydb -F 256MiB \
    -o 's3://my-bucket/sql-backup' \
    --s3.region 'us-west-2'
```

如果同时指定了 URL 参数和命令行参数，命令行参数会覆盖 URL 参数。

8.4.1.3.1 S3 的命令行参数

命令行参数	描述
`--s3.region`	Amazon S3 服务区域（默认为 `us-east-1`）
`--s3.endpoint`	S3 兼容服务自定义端点的 URL（例如 `https://s3.example.com/`）
`--s3.storage-class`	上传对象的存储类别（例如 `STANDARD`、`STANDARD_IA`）
`--s3.sse`	用于加密上传的服务器端加密算法（可以设置为空，`AES256` 或 `aws:kms`）
`--s3.sse-kms-key-id`	如果 `--s3.sse` 设置为 `aws:kms`，则使用该参数指定 KMS ID
`--s3.acl`	上传对象的 canned ACL（例如，`private`、`authenticated-read`）
`--s3.provider`	S3 兼容服务类型（支持 `aws`、`alibaba`、`ceph`、`netease` 或 `other`）

8.4.1.3.2 GCS 的命令行参数

命令行参数	描述
`--gcs.credentials-file`	迁移工具节点上的凭证 JSON 文件的路径
`--gcs.storage-class`	上传对象的存储类别（例如 `STANDARD`、`COLDLINE`）
`--gcs.predefined-acl`	上传对象的预定义 ACL（例如 `private`、`project-private`）

8.4.1.3.3 Azblob 的命令行参数

命令行参数	描述
`--azblob.account-name`	存储账户名
`--azblob.account-key`	访问密钥

命令行参数	描述
--azblob. ↪ access- ↪ tier	上传对象的存储类别（例如 Hot、Cool、Archive）。如果没有设置 access- ↪ tier 的值（该值为空），此值会默认设置为 Hot。

8.4.1.4 BR 向 TiKV 发送凭证

在默认情况下，使用 S3、GCS 或 Azblob 存储时，BR 会将凭证发送到每个 TiKV 节点，以减少设置的复杂性。

但是，这个操作不适合云端环境，因为每个节点都有自己的角色和权限。在这种情况下，你需要用 --send-credentials-to-tikv=false（或简写为 -c=0）来禁止发送凭证：

```
./br backup full -c=0 -u pd-service:2379 -s 's3://bucket-name/prefix'
```

使用 SQL 进行备份恢复时，可加上 SEND_CREDENTIALS_TO_TIKV = FALSE 选项：

```
BACKUP DATABASE * TO 's3://bucket-name/prefix' SEND_CREDENTIALS_TO_TIKV = FALSE;
```

此参数不适用于 TiDB Lightning 和 Dumpling，因为目前它们都是单机程序。

8.4.2 在 Azure Blob Storage 备份恢复

Backup & Restore (BR) 工具支持将 Azure Blob Storage 作为外部存储来进行数据备份与恢复。

如需了解 BR 支持的其他外部存储，请参阅外部存储。

8.4.2.1 使用场景

Azure 虚拟机可以将大规模数据快速地存放到 Azure Blob Storage 上。如果你在使用 Azure 虚拟机来部署集群，可以考虑将数据备份到 Azure Blob Storage 中。

8.4.2.2 使用方法

使用 BR，你可以通过以下两种方法在 Azure Blob Storage 上进行备份与恢复：

- 使用 Azure AD 备份恢复
- 使用访问密钥备份恢复

在通常情况下，为了避免 account-key 等密钥信息记录在命令行中被泄漏，推荐使用第一种方法，即使用 Azure AD 备份恢复。

以下为使用上述两种方式在 Azure Blob Storage 上进行备份与恢复的操作示例，其中，具体操作目标如下：

- 备份：将数据库的 test 库备份到 Azure Blob Storage 的容器名为 container=test 且路径前缀为 t1 的空间中；
- 恢复：将 Azure Blob Storage 的容器名为 container=test 且路径前缀为 t1 的空间恢复到数据库的 test 库中。

8.4.2.2.1 方法一：使用 Azure AD 备份恢复（推荐）

在 BR 运行环境和 TiKV 运行环境中，需要存在环境变量 $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET。当存在这三项变量时，BR 可以使用 Azure AD 访问 Azure Blob Storage 且不用配置 account-key。这种方式更安全，因此也推荐被使用。在这里，$AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET 分别代表 Azure 应用程序的应用程序 ID client_id，租户 ID tenant_id 和客户端密码 client_secret。

如需了解如何确认运行环境中存在环境变量 $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET，或需要将环境变量配置为参数，请参考配置环境变量作为参数。

备份

使用 Azure AD 进行备份时，需要指定参数 account-name 和 access-tier。其中，如果没有设置 access-tier（即该值为空），该值会默认设置为 Hot。

> 注意：
>
> 将 Azure Blob Storage 作为外部存储时，必须设置 send-credentials-to-tikv = true（即默认情况），否则会导致备份失败。

本节中展示了备份到 cool tier，即上传对象的存储类别为 Cool 的案例。你可以通过以下两种方式指定 account-name 和 access-tier：

- 将参数信息放在 URL 参数中：

```
tiup br backup db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?account-name=
    ↪ devstoreaccount1&access-tier=Cool'
```

- 将参数信息放在命令行参数中：

```
tiup br backup db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?' --azblob.account-name=
    ↪ devstoreaccount1 --azblob.access-tier=Cool
```

恢复

使用 Azure AD 进行恢复时，需要指定参数 account-name。你可以通过以下两种方式指定 account-name：

- 将参数信息放在 URL 参数中：

```
tiup br restore db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?account-name=
    ↪ devstoreaccount1'
```

- 将参数信息放在命令行参数中：

```
tiup br restore db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?' --azblob.account-name=
    ↪ devstoreaccount1
```

8.4.2.2.2 方法二：使用访问密钥备份恢复（简易）

备份

使用访问密钥进行备份时，需要指定参数 `account-name`、`account-key` 和 `access-tier`。其中，如果没有设置 `access-tier`（即该值为空），该值会默认设置为 `Hot`。

> 注意：
>
> 将 Azure Blob Storage 作为外部存储时，必须设置 `send-credentials-to-tikv = true`（即默认情况），否则会导致备份失败。

本节中展示了备份到 `cool tier`，即上传对象的存储类别为 `Cool` 的案例。你可以通过以下两种方式指定 `account-name`、`account-key` 和 `access-tier`：

- 将参数信息放在 URL 参数中：

```
tiup br backup db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?account-name=
    ↪ devstoreaccount1&account-key=
    ↪ Eby8vdM02xNOcqFlqUwJPLlmEtlCDXJ1OUzFT50uSRZ6IFsuFq2UVErCz4I6tq/K1SZFPTOtr/
    ↪ KBHBeksoGMGw==&access-tier=Cool'
```

- 将参数信息放在命令行参数中：

```
tiup br backup db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?' --azblob.account-name=
    ↪ devstoreaccount1 --azblob.account-key=
    ↪ Eby8vdM02xNOcqFlqUwJPLlmEtlCDXJ1OUzFT50uSRZ6IFsuFq2UVErCz4I6tq/K1SZFPTOtr/
    ↪ KBHBeksoGMGw== --azblob.access-tier=Cool
```

恢复

使用访问密钥进行恢复时，需要指定参数 `account-name` 和 `account-key`。你可以通过以下两种方式指定这些参数：

- 将参数信息放在 URL 参数中：

```
tiup br restore db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?account-name=
    ↪ devstoreaccount1&account-key=
    ↪ Eby8vdM02xNOcqFlqUwJPLlmEtlCDXJ1OUzFT50uSRZ6IFsuFq2UVErCz4I6tq/K1SZFPTOtr/
    ↪ KBHBeksoGMGw=='
```

- 将参数信息放在命令行参数中：

```
tiup br restore db --db test -u 127.0.0.1:2379 -s 'azure://test/t1?' --azblob.account-name=
    ↪ devstoreaccount1 --azblob.account-key=
    ↪ Eby8vdM02xNOcqFlqUwJPLlmEtlCDXJ1OUzFT50uSRZ6IFsuFq2UVErCz4I6tq/K1SZFPTOtr/
    ↪ KBHBeksoGMGw==
```

8.4.2.3 参数说明

在进行备份和恢复过程中，你会用到 `account-name`、`account-key`、`access-tier` 三个参数。有关参数的详细介绍，请参阅以下内容：

- URL 参数
- 命令行参数

8.4.2.3.1 配置环境变量作为参数

当使用 Azure AD 备份恢复时，需要在 BR 运行环境和 TiKV 运行环境中配置环境变量 `$AZURE_CLIENT_ID`、`$AZURE_TENANT_ID` 和 `$AZURE_CLIENT_SECRET`。

- 当使用 TiUP 启动的集群时，TiKV 会使用 systemd 服务。以下示例提供如何为 TiKV 配置上述的三个环境变量作为参数：

> 注意：
> 该流程在第 3 步中需要重启 TiKV。如果你的集群不适合重启，请使用方法二进行备份恢复。

1. 假设该节点上 TiKV 端口为 24000（即 systemd 服务名为 tikv-24000）：

```
systemctl edit tikv-24000
```

2. 填入环境变量信息：

```
[Service]
Environment="AZURE_CLIENT_ID=aaaaaaaa-aaaa-aaaa-aaaa-aaaaaaaaaaaa"
Environment="AZURE_TENANT_ID=aaaaaaaa-aaaa-aaaa-aaaa-aaaaaaaaaaaa"
Environment="AZURE_CLIENT_SECRET=aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa"
```

3. 重新加载配置并重启 TiKV：

```
systemctl daemon-reload
systemctl restart tikv-24000
```

- 为命令行启动的 TiKV 和 BR 配置 Azure AD 的信息，只需要确定运行环境中存在 $AZURE_CLIENT_ID、$AZURE_TENANT_ID 和 $AZURE_CLIENT_SECRET。通过运行下列命令行，可以确认 BR 和 TiKV 运行环境中是否存在这三个环境变量：

```
echo $AZURE_CLIENT_ID
echo $AZURE_TENANT_ID
echo $AZURE_CLIENT_SECRET
```

8.4.2.4 兼容信息

该功能仅兼容 TiDB v5.4.0 及后续的版本。

8.5 BR 特性

8.5.1 BR 自动调节从 v5.4 版本开始引入

在 TiDB v5.4.0 之前，默认情况下，使用 BR 工具进行备份任务时使用的线程数量占总逻辑 CPU 数量的 75%。在没有限速的前提下，备份会消耗大量的集群资源，这会对在线集群的性能造成相当大的影响。虽然你可以通过调节线程池的大小的方式来减少备份对集群性能的影响，但观察负载、手动调节线程池大小也是一件繁琐的事情。

为了减少备份任务对在线集群的影响，从 TiDB v5.4.0 起，BR 引入了自动调节功能，此功能会默认开启。在集群资源占用率较高的情况下，BR 可以通过该功能自动限制备份使用的资源，从而减少对集群的影响。

8.5.1.1 使用场景

如果你希望减少 BR 备份对集群的影响，那么，你可以开启自动调节功能。开启该功能后，BR 会在不过度影响集群的前提下，以最快的速度进行备份。

或者，你也可以使用 TiKV 配置项`backup.num-threads` 或参数 `--ratelimit` 进行备份限速。

8.5.1.2 使用方法

自动调节功能默认打开，无需额外配置。

> **注意：**
>
> v5.3.x 版本的集群，在升级到 v5.4.0 及以上版本后，自动调节功能默认关闭，需手动开启。

如需开启 BR 自动调节功能，可以通过把 TiKV 配置项backup.enable-auto-tune 设置为 true 的方式来完成。

TiKV 支持动态配置自动调节功能，因此，在开启或关闭该功能时，无需重启集群。你可以运行以下命令动态启动或停止 BR 自动调节功能：

```
tikv-ctl modify-tikv-config -n backup.enable-auto-tune -v <true|false>
```

在离线备份场景中，你也可以使用 tikv-ctl 把 backup.num-threads 修改为更大的数字，从而提升备份速度。

8.5.1.3 使用限制

自动调节是一个粗粒度的限流方案，它的优势在无需手动调节。但是，由于调节的粒度不够精确，该功能有可能无法彻底移除备份对集群的影响。

该功能的已知问题及其解决方案如下：

- 问题 1：对于以写负载为主的集群，自动调节可能会让工作负载和备份进入一种“正反馈循环”：备份会占用较多资源，导致工作负载使用的资源变少。此时，自动调节会误以为资源使用率下降，从而让 BR 运行得更加激进。在这种情况下，自动调节实际上失效。
  - 解决方法：手动调节 backup.num-threads，限制处理备份的工作线程数量。具体原理如下：
    目前，备份过程会涉及大量的 SST 解码、编码、压缩、解压，而此过程会需要消耗大量的 CPU 资源。另外，以往的测试证明，备份过程中，用于备份的线程池的 CPU 利用率接近 100%。也就是说，备份任务会占用大量 CPU 资源。通过调整备份任务使用的线程数量，TiKV 可以控制备份任务使用的 CPU 核心数，从而减少其任务对集群性能的影响。
- 问题 2：对于存在热点的集群，产生热点的 TiKV 节点可能会被过度限流，从而拉慢备份的整体进度。
  - 解决方法：消除热点节点，或者在热点节点上关闭自动调节（关闭此功能可能会导致集群性能降低）。
- 问题 3：对于流量抖动非常大的场景，由于自动调节每隔auto-tune-refresh-interval（默认为一分钟）才会计算出新的限流，所以可能无法很好地应对流量抖动非常厉害的场景。
  - 解决方法：关闭自动调节。

8.5.1.4 实现原理

自动调节会通过调节 BR 备份时使用的工作线程池的大小，保证集群的 CPU 总体使用率不超过某个特定的值。

这个特性还有两个配置项未在 TiKV 文档中列出，仅在内部调试使用，正常备份时无需配置这两个参数。

- backup.auto-tune-remain-threads：
  - 通过控制备份任务占用的资源，自动调节会保证该节点中至少有该数量的核心会保持空闲的状态。
  - 默认值：round(0.2 * vCPU)
- backup.auto-tune-refresh-interval：
  - 每隔该值的时间段，自动调节会刷新统计信息并重新计算备份任务使用的 CPU 核心数的上限。
  - 默认值：1m

以下是一个使用自动调节功能的示例，其中 * 代表集群中被备份任务占用的 CPU，^ 代表其它任务占用的 CPU，- 代表空闲 CPU。

```
|--------| 系统总共有 8 颗逻辑 CPU。
|****----| 默认配置 `backup.num-threads` 为 `4`。请注意，在任何时候自动调节都不会让线程池大小大于
    ↪ `backup.num-threads`。
|^^****--| 默认配置 `auto-tune-remain-threads` = round(8 * 0.2) = 2。
    ↪ 自动调节会将备份任务的线程池大小调节至 `4`。
|^^^^**--| 由于集群的工作负载加重，自动调节将备份任务的线程池大小调节至 `2`。调节后，集群中仍有 2
    ↪ 个 CPU 核心数保持空闲。
```

在监控面板的 “Backup CPU Utilization” 中，可以看到自动限流目前选择的线程池的大小：

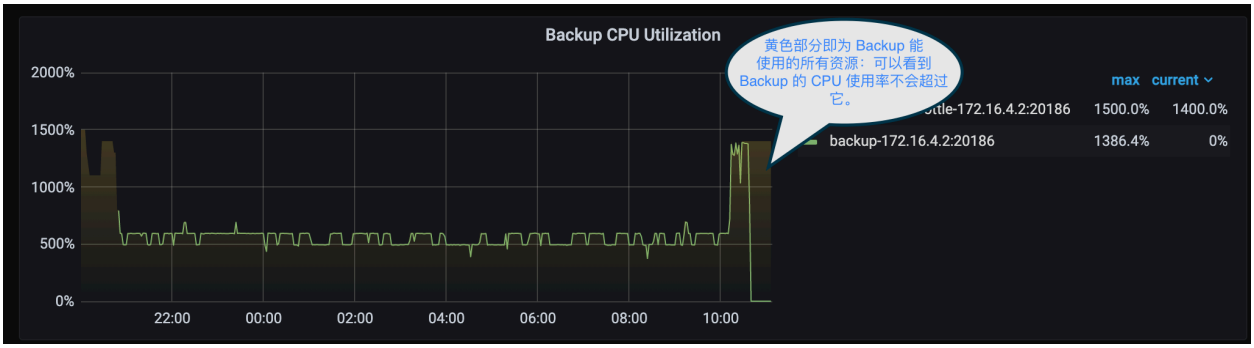


图 42: Grafana dashboard example of backup auto-tune metrics

图片中，黄色半透明的填充部分为开启自动调节后备份任务可用的线程，即备份任务能使用的所有资源。从中可以看到备份任务的 CPU 使用率不会超过黄色部分。

8.6 Backup & Restore 常见问题

本文列出了在使用 Backup & Restore (BR) 时，可能会遇到的问题及相应的解决方法。

如果遇到未包含在此文档且无法解决的问题，可以在 AskTUG 社区中提问。

8.6.1 在 TiDB v5.4.0 及后续版本中，当在有负载的集群进行备份时，备份速度为什么会变得很慢？

从 TiDB v5.4.0 起，TiKV 的备份新增了自动调节功能。对于 v5.4.0 及以上版本，该功能会默认开启。当集群负载较高时，该功能会自动限制备份任务使用的资源，从而减少备份对在线集群的性能造成的影响。如需了解关于自动调节功能的更多信息，请参见自动调节。

TiKV 支持动态配置自动调节功能。因此，在开启或关闭该功能时，你不需要重启集群。以下为该功能的具体使用方法：

- 关闭功能：把 TiKV 配置项backup.enable-auto-tune 设置为 false。
- 开启功能：把 backup.enable-auto-tune 设置为 true。对于 v5.3.x 版本的集群，当 TiDB 升级到 v5.4.0 及以上版本后，自动调节功能会默认关闭。这时，你可以通过该方式手动开启此功能。

如需了解通过 tikv-ctl 在线修改自动调节功能的命令行，请参阅自动调节功能的使用方法。

另外，该功能也降低了进行备份任务时默认使用的工作线程数量（详见backup.num-threads）。因此，你通过 Grafana 监控面板看到的备份速度、CPU 使用率、I/O 资源利用率都会小于 v5.4 之前的版本。在 v5.4 之前的版本中，backup.num-threads 的默认值为 CPU * 0.75，即处理备份任务的工作线程数量占了 75% 的逻辑 CPU，最大值为 32；在 v5.4 及之后的版本中，该配置项的默认值为 CPU * 0.5，最大值为 8。

在离线备份场景中，你也可以使用 tikv-ctl 把 backup.num-threads 修改为更大的数字，从而提升备份任务的速度。

8.6.2 恢复的时候，报错 could not read local://...:download sst failed，该如何处理？

在恢复的时候，每个节点都必须能够访问到所有的备份文件 (SST files)，默认情况下，假如使用 local storage，备份文件会分散在各个节点中，此时是无法直接恢复的，必须将每个 TiKV 节点的备份文件拷贝到其它所有 TiKV 节点才能恢复。

建议在备份的时候挂载一块 NFS 网盘作为备份盘，详见将单表数据备份到网络盘。

8.6.3 BR 备份时，对集群影响多大？

对于 TiDB v5.4.0 及以后的版本，BR 不仅调低了备份任务会使用的默认 CPU 使用率，还引入了自动调节功能。开启该功能后，当在集群高负载的场景下进行备份时，BR 能够自动限制备份任务使用的资源，从而限制 BR 备份的速度。因此，在 v5.4.0 高负载集群中使用默认配置进行备份时，备份任务对其集群性能的影响会显著小于 v5.4.0 之前的版本。有关自动调节功能的具体内容，请参阅BR 自动调节。

以下为针对内部的一个单节点进行的内部测试。通过测试结果，可以确定在全速备份场景下使用 v5.4.0 版本及其前期版本的默认配置时，BR 备份对集群性能的影响有较大区别。具体测试结果如下：

- 使用 v5.3.0 的默认配置时，纯写负载的 QPS 降低了 75%；
- 使用 v5.4.0 的默认配置时，相同负载的 QPS 降低了 25%。不过，使用该配置时，BR 备份任务的耗时也相应得拉慢，其所需时间为 v5.3.0 配置下耗时的 1.7 倍。

如果需要手动控制 BR 备份对集群性能带来的影响，可以通过以下方案实现。这两种方案在减少 BR 备份对集群的影响的同时，也会降低备份任务的速度。

- 使用 --ratelimit 参数对备份任务进行限速。请注意，这个参数限制的是把备份文件存储到外部存储的速度。计算备份文件的总大小的时候，请以备份日志中的 backup data size(after compressed) 为基准。
- 调节 TiKV 配置项backup.num-threads，限制备份任务使用的资源。该配置项决定处理备份任务的工作线程数量。过往的测试数据表明，当 BR 备份的线程数量不大于 8、集群总 CPU 利用率不超过 60% 时，BR 备份任务对集群（无论读写负载）几乎没有影响。

8.6.4 BR 会备份系统表吗？在数据恢复的时候，这些系统表会冲突吗？

在 v5.1.0 之前，BR 备份时会过滤掉系统库 mysql.* 的表数据。自 v5.1.0 起，BR 默认备份集群内的全部数据，包括系统库 mysql.* 中的数据。

恢复数据时，系统表不会冲突。由于恢复 `mysql.*` 中系统表数据的技术实现尚不完善，因此 BR 默认不恢复系统库 `mysql` 中的表数据，也就不会产生任何冲突。详情参阅备份和恢复 `mysql` 系统库下的表数据（实验特性）。

8.6.5 BR 遇到 Permission denied 或者 No such file or directory 错误，即使用 root 运行 BR 也无法解决，该如何处理？

需要确认 TiKV 是否有访问备份目录的权限。如果是备份，确认是否有写权限；如果是恢复，确认是否有读权限。

在进行备份操作时，如果使用本地磁盘或 NFS 作为存储介质，请确保执行 BR 的用户和启动 TiKV 的用户相同（如果 BR 和 TiKV 位于不同的机器，则需要用户的 UID 相同），否则备份可能会出现该问题。。

使用 root 运行 BR 仍旧有可能会因为磁盘权限而失败，因为备份文件 (SST) 的保存是由 TiKV 执行的。

> **注意：**
>
> 在恢复的时候也可能遇到同样的问题。
>
> 使用 BR 进行数据的恢复时，检验读权限的时机是在第一次读取 SST 文件时，考虑到执行 DDL 的耗时，这个时刻可能会离开始运行 BR 的时间很远。这样可能会出现等了很长时间之后遇到 Permission denied 错误失败的情况。
>
> 因此，最好在恢复前提前检查权限。

你可以按照如下步骤进行权限检查：

1. 执行 Linux 原生的进程查询命令

```
ps aux | grep tikv-server
```

命令输出示例如下:

```
tidb_ouo  9235 10.9  3.8 2019248 622776 ?      Ssl  08:28   1:12 bin/tikv-server --addr
    ↪ 0.0.0.0:20162 --advertise-addr 172.16.6.118:20162 --status-addr 0.0.0.0:20188 --
    ↪ advertise-status-addr 172.16.6.118:20188 --pd 172.16.6.118:2379 --data-dir /home/
    ↪ user1/tidb-data/tikv-20162 --config conf/tikv.toml --log-file /home/user1/tidb-
    ↪ deploy/tikv-20162/log/tikv.log
tidb_ouo  9236  9.8  3.8 2048940 631136 ?      Ssl  08:28   1:05 bin/tikv-server --addr
    ↪ 0.0.0.0:20161 --advertise-addr 172.16.6.118:20161 --status-addr 0.0.0.0:20189 --
    ↪ advertise-status-addr 172.16.6.118:20189 --pd 172.16.6.118:2379 --data-dir /home/
    ↪ user1/tidb-data/tikv-20161 --config conf/tikv.toml --log-file /home/user1/tidb-
    ↪ deploy/tikv-20161/log/tikv.log
```

或者执行以下命令：

```
ps aux | grep tikv-server | awk '{print $1}'
```

命令输出示例如下:

```
tidb_ouo
tidb_ouo
```

2. 使用 TiUP 命令查询集群的启动信息

```
tiup cluster list
```

命令输出示例如下：

```
[root@Copy-of-VM-EE-CentOS76-v1 br]# tiup cluster list
Starting component `cluster`: /root/.tiup/components/cluster/v1.5.2/tiup-cluster list
Name          User      Version  Path
    ↪ PrivateKey
----          ----      -------  ----
    ↪ ----------
tidb_cluster  tidb_ouo  v5.0.2   /root/.tiup/storage/cluster/clusters/tidb_cluster  /root/.
    ↪ tiup/storage/cluster/clusters/tidb_cluster/ssh/id_rsa
```

3. 检查备份目录的权限，例如 backup 目录是备份数据存储目录。命令示例如下：

```
ls -al backup
```

命令输出示例如下：

```
[root@Copy-of-VM-EE-CentOS76-v1 user1]# ls -al backup
total 0
drwxr-xr-x  2 root root   6 Jun 28 17:48 .
drwxr-xr-x 11 root root 310 Jul  4 10:35 ..
```

由以上命令输出结果可知，tikv-server 实例由用户 tidb_ouo 启动，但用户账号 tidb_ouo 没有 backup 目录的写入权限，所以备份失败。

8.6.6 BR 遇到错误信息 Io(Os...)，该如何处理？

这类问题几乎都是 TiKV 在写盘的时候遇到的系统调用错误。例如遇到 Io(Os { code: 13, kind: ↪ PermissionDenied...}) 或者 Io(Os { code: 2, kind: NotFound...}) 这类错误信息，首先检查备份目录的挂载方式和文件系统，试试看备份到其它文件夹或者其它硬盘。

目前已知备份到 samba 搭建的网盘时可能会遇到 Code: 22(invalid argument) 错误。

8.6.7 BR 遇到错误信息 rpc error: code = Unavailable desc =...，该如何处理？

该问题一般是因为使用 BR 恢复数据的时候，恢复集群的性能不足导致的。可以从恢复集群的监控或者 TiKV 日志来辅助确认。

要解决这类问题，可以尝试扩大集群资源，以及调小恢复时的并发度 (concurrency)，打开限速 (ratelimit) 设置。

8.6.8 使用 local storage 的时候，BR 备份的文件会存在哪里？

在使用 local storage 的时候，会在运行 BR 的节点生成 backupmeta，在各个 Region 的 Leader 节点生成备份文件。

8.6.9 备份数据有多大，备份会有副本吗？

备份的时候仅仅在每个 Region 的 Leader 处生成该 Region 的备份文件。因此备份的大小等于数据大小，不会有多余的副本数据。所以最终的总大小大约是 TiKV 数据总量除以副本数。

但是假如想要从本地恢复数据，因为每个 TiKV 都必须要能访问到所有备份文件，在最终恢复的时候会有等同于恢复时 TiKV 节点数量的副本。

8.6.10 BR 恢复到 TiCDC / Drainer 的上游集群时，要注意些什么？

- BR 恢复的数据无法被同步到下游，因为 BR 直接导入 SST 文件，而下游集群目前没有办法获得上游的 SST 文件。
- 在 4.0.3 版本之前，BR 恢复时产生的 DDL jobs 还可能会让 TiCDC / Drainer 执行异常的 DDL。所以，如果一定要在 TiCDC / Drainer 的上游集群执行恢复，请将 BR 恢复的所有表加入 TiCDC / Drainer 的阻止名单。

TiCDC 可以通过配置项中的 `filter.rules` 项完成，Drainer 则可以通过`syncer.ignore-table` 完成。

8.6.11 BR 会备份表的 SHARD_ROW_ID_BITS 和 PRE_SPLIT_REGIONS 信息吗？恢复出来的表会有多个 Region 吗？

会的，BR 会备份表的`SHARD_ROW_ID_BITS` 和 `PRE_SPLIT_REGIONS` 信息，并恢复成多个 Region。

8.6.12 使用 BR 恢复备份数据后，SQL 查询报错 region is unavailable

如果 BR 备份的集群有 TiFlash，恢复时会将 TiFlash 信息存进 TableInfo。此时如果恢复的集群没有 TiFlash，则会报该错误。计划在未来版本中修复该错误。

8.6.13 BR 是否支持就地 (in-place) 全量恢复某个历史备份？

不支持。

8.6.14 在 Kubernetes 环境中如何使用 BR 进行增量备份？

可以使用 kubectl 执行 `kubectl -n ${namespace} get bk ${name}` 以获得上次 BR 备份 `commitTs` 字段，该字段的内容可作为 `--lastbackupts` 使用。

8.6.15 BR backupTS 如何转化成 Unix 时间？

BR backupTS 默认是在备份开始前，从 PD 获取到的最新时间戳。可以使用 `pd-ctl tso timestamp` 来解析该时间戳，以获得精确值，也可以通过 `backupTS >> 18` 来快速获取估计值。

8.6.16 BR 恢复存档后是否需要对表执行 ANALYZE 以更新 TiDB 在表和索引上留下的统计信息？

BR 不会备份统计信息（v4.0.9 除外）。所以在恢复存档后需要手动执行 ANALYZE TABLE 或等待 TiDB 自动进行 ANALYZE。

BR v4.0.9 备份统计信息使 BR 消耗过多内存，为保证备份过程正常，从 v4.0.10 开始默认关闭备份统计信息的功能。

如果不对表执行 ANALYZE，TiDB 会因统计信息不准确而选不中最优化的执行计划。如果查询性能不是重点关注项，可以忽略 ANALYZE。

8.6.17 是否可以同时使用多个 BR 进程对单个集群进行恢复？

强烈不建议在单个集群中同时使用多个 BR 进程进行恢复，原因如下：

- BR 在恢复数据时，会修改 PD 的一些全局配置。如果同时使用多个 BR 进程进行恢复，这些配置可能会被错误地覆写，导致集群状态异常。
- 因为 BR 在恢复数据的时候会占用大量集群资源，事实上并行恢复能获得的速度提升也非常有限。
- 多个 BR 并行恢复的场景没有经过测试，无法保证成功。

9 监控与告警

9.1 TiDB 监控框架概述

TiDB 使用开源时序数据库 Prometheus 作为监控和性能指标信息存储方案，使用 Grafana 作为可视化组件进行展示。

9.1.1 Prometheus 在 TiDB 中的应用

Prometheus 是一个拥有多维度数据模型的、灵活的查询语句的时序数据库。Prometheus 作为热门的开源项目，拥有活跃的社区及众多的成功案例。

Prometheus 提供了多个组件供用户使用。目前，TiDB 使用了以下组件：

- Prometheus Server：用于收集和存储时间序列数据。
- Client 代码库：用于定制程序中需要的 Metric。
- Alertmanager：用于实现报警机制。

其结构如下图所示：

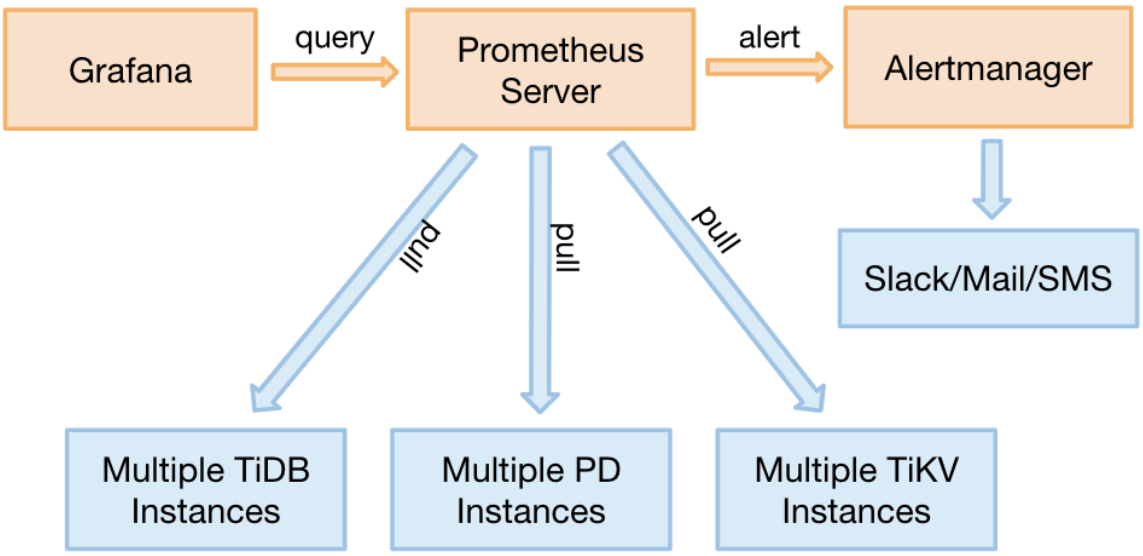


图 43: Prometheus in TiDB

9.1.2 Grafana 在 TiDB 中的应用

Grafana 是一个开源的 metric 分析及可视化系统。TiDB 使用 Grafana 来展示 TiDB 集群各组件的相关监控，监控项分组如下图所示：

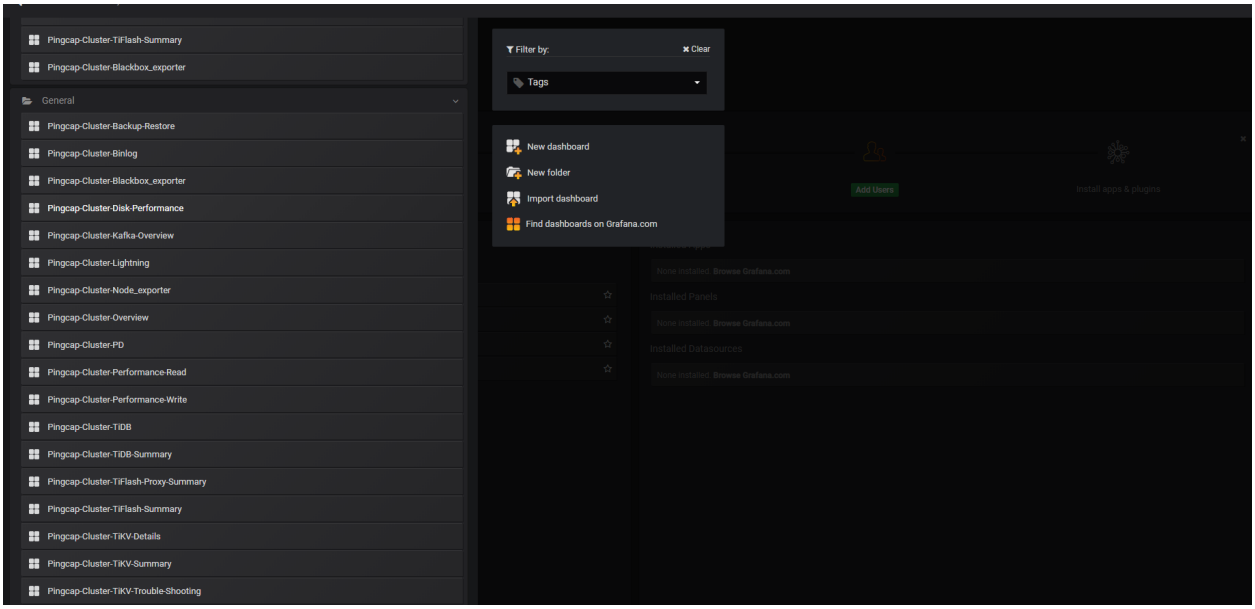


图 44: Grafana monitored_groups

- {TiDB_Cluster_name}-Backup-Restore：备份恢复相关的监控项。

- {TiDB_Cluster_name}-Binlog：TiDB Binlog 相关的监控项。
- {TiDB_Cluster_name}-Blackbox_exporter：网络探活相关监控项。
- {TiDB_Cluster_name}-Disk-Performance：磁盘性能相关监控项。
- {TiDB_Cluster_name}-Kafka-Overview：Kafka 相关监控项。
- {TiDB_Cluster_name}-Lightning：TiDB Lightning 组件相关监控项。
- {TiDB_Cluster_name}-Node_exporter：操作系统相关监控项。
- {TiDB_Cluster_name}-Overview：重要组件监控概览。
- {TiDB_Cluster_name}-PD：PD server 组件相关监控项。
- {TiDB_Cluster_name}-Performance-Read：读性能相关监控项。
- {TiDB_Cluster_name}-Performance-Write：写性能相关监控项。
- {TiDB_Cluster_name}-TiDB：TiDB server 组件详细监控项。
- {TiDB_Cluster_name}-TiDB-Summary：TiDB server 相关监控项概览。
- {TiDB_Cluster_name}-TiFlash-Proxy-Summary：数据同步到 TiFlash 的代理 server 监控项概览。
- {TiDB_Cluster_name}-TiFlash-Summary：TiFlash server 相关监控项概览。
- {TiDB_Cluster_name}-TiKV-Details：TiKV server 组件详细监控项。
- {TiDB_Cluster_name}-TiKV-Summary：TiKV server 监控项概览。
- {TiDB_Cluster_name}-TiKV-Trouble-Shooting：TiKV 错误诊断相关监控项。
- {TiDB_Cluster_name}-TiCDC：TiCDC 组件详细监控项。

每个分组包含多个监控项页签，页签中包含多个详细的监控项信息。以 Overview 监控组为例，其中包含 5 个页签，每个页签内有相应的监控指标看板，如下图所示：



图 45: Grafana Overview

要快速了解 TiDB 监控与报警系统的体系、该系统背后的数据流转方式、系统管理方法、系统使用方法和常用监控指标，建议观看下面的培训视频（时长 29 分钟）。注意本视频只作为学习参考，具体的监控指标与相关报警规则，请以文档内容为准。

9.2 TiDB 集群监控 API

TiDB 提供了以下几种接口来监控集群状态：

- 状态接口：通过 HTTP 接口对外汇报组件的信息。通过状态接口，你可获取当前 TiDB Server 的运行状态以及表的存储信息。
- Metrics 接口：使用 Prometheus 记录组件中各种操作的详细信息，使用 Grafana 进行可视化展示。

9.2.1 使用状态接口

状态接口用于监控组件的一些基本信息，并且可以作为 keepalive 的监测接口。另外，通过 PD 的状态接口可以看到整个 TiKV 集群的详细信息。

9.2.1.1 TiDB Server

- TiDB API 地址：`http://${host}:${port}`
- 默认端口：10080

9.2.1.1.1 运行状态

以下示例中，通过访问 `http://${host}:${port}/status` 获取当前 TiDB Server 的状态，并判断该 TiDB Server 是否存活。结果以 JSON 格式返回：

```
curl http://127.0.0.1:10080/status
```

```
{
    connections: 0,  # 当前 TiDB Server 上的客户端连接数
    version: "5.7.25-TiDB-v4.0.0-rc-141-g7267747ae",  # TiDB 版本号
    git_hash: "7267747ae0ec624dffc3fdedb00f1ed36e10284b"  # TiDB 当前代码的 Git Hash
}
```

9.2.1.1.2 存储信息

以下示例中，通过访问 `http://${host}:${port}/schema_storage/${db}/${table}` 获取指定数据表的存储信息。结果以 JSON 格式返回：

```
curl http://127.0.0.1:10080/schema_storage/mysql/stats_histograms
```

```
{
    "table_schema": "mysql",
    "table_name": "stats_histograms",
    "table_rows": 0,
    "avg_row_length": 0,
    "data_length": 0,
    "max_data_length": 0,
    "index_length": 0,
    "data_free": 0
}
```

```
curl http://127.0.0.1:10080/schema_storage/test
```

```
[
    {
        "table_schema": "test",
        "table_name": "test",
        "table_rows": 0,
        "avg_row_length": 0,
        "data_length": 0,
        "max_data_length": 0,
```

```
        "index_length": 0,
        "data_free": 0
    }
]
```

9.2.1.2 PD Server

- PD API 地址：`http://${host}:${port}/pd/api/v1/${api_name}`
- 默认端口：2379
- 各类 `api_name` 详细信息：参见 PD API Doc

通过该接口可以获取当前所有 TiKV 节点的状态以及负载均衡信息。下面以一个单节点的 TiKV 集群为例，说明用户需要了解的信息：

```
curl http://127.0.0.1:2379/pd/api/v1/stores
```

```
{
  "count": 1,  # TiKV 节点数量
  "stores": [  # TiKV 节点的列表
    # 集群中单个 TiKV 节点的信息
    {
      "store": {
        "id": 1,
        "address": "127.0.0.1:20160",
        "version": "4.0.0-rc.2",
        "status_address": "172.16.5.90:20382",
        "git_hash": "2fdb2804bf8ffaab4b18c4996970e19906296497",
        "start_timestamp": 1590029618,
        "deploy_path": "/data2/tidb_test/v4.0.rc.2/tikv-20372/bin",
        "last_heartbeat": 1590030038949235439,
        "state_name": "Up"
      },
      "status": {
        "capacity": "3.581TiB",  # 存储总容量
        "available": "3.552TiB",  # 存储剩余容量
        "used_size": "31.77MiB",
        "leader_count": 174,
        "leader_weight": 1,
        "leader_score": 174,
        "leader_size": 174,
        "region_count": 531,
        "region_weight": 1,
        "region_score": 531,
        "region_size": 531,
        "start_ts": "2020-05-21T10:53:38+08:00",
```

```
        "last_heartbeat_ts": "2020-05-21T11:00:38.949235439+08:00",
        "uptime": "7m0.949235439s"
      }
    }
]
```

9.2.2 使用 metrics 接口

Metrics 接口用于监控整个集群的状态和性能。

- 如果使用其他方式部署 TiDB 集群，在使用 metrics 接口前，需先部署 Prometheus 和 Grafana。

成功部署 Prometheus 和 Grafana 之后，配置 Grafana。

9.3 TiDB 集群监控部署

本文档适用于希望手动部署 TiDB 监控报警系统的用户。TiUP 部署方式，会同时自动部署监控报警系统，无需手动部署。

9.3.1 部署 Prometheus 和 Grafana

假设 TiDB 的拓扑结构如下：

节点	主机 IP	服务
Node1	192.168.199.113	PD1, TiDB, node_export, Prometheus, Grafana
Node2	192.168.199.114	PD2, node_export
Node3	192.168.199.115	PD3, node_export
Node4	192.168.199.116	TiKV1, node_export
Node5	192.168.199.117	TiKV2, node_export
Node6	192.168.199.118	TiKV3, node_export

9.3.1.1 第 1 步：下载二进制包

下载二进制包：

```
wget https://download.pingcap.org/prometheus-2.27.1.linux-amd64.tar.gz
wget https://download.pingcap.org/node_exporter-0.17.0.linux-amd64.tar.gz
wget https://download.pingcap.org/grafana-6.1.6.linux-amd64.tar.gz
```

解压二进制包：

```
tar -xzf prometheus-2.27.1.linux-amd64.tar.gz
tar -xzf node_exporter-0.17.0.linux-amd64.tar.gz
tar -xzf grafana-6.1.6.linux-amd64.tar.gz
```

9.3.1.2 第 2 步：在 Node1，Node2，Node3，Node4 上启动 node_exporter

```
cd node_exporter-0.17.0.linux-amd64
```

启动 node_exporter 服务：

```
./node_exporter --web.listen-address=":9100" \
    --log.level="info" &
```

9.3.1.3 第 3 步：在 Node1 上启动 Prometheus

编辑 Prometheus 的配置文件：

```
cd prometheus-2.27.1.linux-amd64 &&
vi prometheus.yml
```

```
...

global:
  scrape_interval:     15s
  evaluation_interval: 15s
  # scrape_timeout 设置为全局默认值 (10s)
  external_labels:
    cluster: 'test-cluster'
    monitor: "prometheus"

scrape_configs:
  - job_name: 'overwritten-nodes'
    honor_labels: true  # 不要覆盖 job 和实例的 label
    static_configs:
    - targets:
      - '192.168.199.113:9100'
      - '192.168.199.114:9100'
      - '192.168.199.115:9100'
      - '192.168.199.116:9100'
      - '192.168.199.117:9100'
      - '192.168.199.118:9100'

  - job_name: 'tidb'
    honor_labels: true  # 不要覆盖 job 和实例的 label
    static_configs:
    - targets:
      - '192.168.199.113:10080'

  - job_name: 'pd'
    honor_labels: true  # 不要覆盖 job 和实例的 label
```

```
    static_configs:
    - targets:
      - '192.168.199.113:2379'
      - '192.168.199.114:2379'
      - '192.168.199.115:2379'

  - job_name: 'tikv'
    honor_labels: true  # 不要覆盖 job 和实例的 label
    static_configs:
    - targets:
      - '192.168.199.116:20180'
      - '192.168.199.117:20180'
      - '192.168.199.118:20180'

...
```

启动 Prometheus 服务：

```
./prometheus \
    --config.file="./prometheus.yml" \
    --web.listen-address=":9090" \
    --web.external-url="http://192.168.199.113:9090/" \
    --web.enable-admin-api \
    --log.level="info" \
    --storage.tsdb.path="./data.metrics" \
    --storage.tsdb.retention="15d" &
```

9.3.1.4 第 4 步：在 Node1 上启动 Grafana

编辑 Grafana 的配置文件：

```
cd grafana-6.1.6 &&
vi conf/grafana.ini
```

```
...

[paths]
data = ./data
logs = ./data/log
plugins = ./data/plugins
[server]
http_port = 3000
domain = 192.168.199.113
[database]
[session]
[analytics]
```

```
check_for_updates = true
[security]
admin_user = admin
admin_password = admin
[snapshots]
[users]
[auth.anonymous]
[auth.basic]
[auth.ldap]
[smtp]
[emails]
[log]
mode = file
[log.console]
[log.file]
level = info
format = text
[log.syslog]
[event_publisher]
[dashboards.json]
enabled = false
path = ./data/dashboards
[metrics]
[grafana_net]
url = https://grafana.net

...
```

启动 Grafana 服务：

```
./bin/grafana-server \
    --config="./conf/grafana.ini" &
```

9.3.2 配置 Grafana

本小节介绍如何配置 Grafana。

9.3.2.1 第 1 步：添加 Prometheus 数据源

1. 登录 Grafana 界面。

    - 默认地址：`http://localhost:3000`
    - 默认账户：admin
    - 默认密码：admin

> **注意：**
>
> Change Password 步骤可以选择 Skip。

2. 点击 Grafana 侧边栏菜单 Configuration 中的 Data Source。
3. 点击 Add data source。
4. 指定数据源的相关信息：
    - 在 Name 处，为数据源指定一个名称。
    - 在 Type 处，选择 Prometheus。
    - 在 URL 处，指定 Prometheus 的 IP 地址。
    - 根据需求指定其它字段。
5. 点击 Add 保存新的数据源。

9.3.2.2 第 2 步：导入 Grafana 面板

执行以下步骤，为 PD Server、 TiKV Server 和 TiDB Server 分别导入 Grafana 面板：

1. 点击侧边栏的 Grafana 图标。
2. 在侧边栏菜单中，依次点击 Dashboards > Import 打开 Import Dashboard 窗口。
3. 点击 Upload .json File 上传对应的 JSON 文件（下载 TiDB Grafana 配置文件。

> **注意：**
>
> TiKV、 PD 和 TiDB 面板对应的 JSON 文件分别为 `tikv_summary.json`，`tikv_details.json`，`tikv_trouble_shooting.json`，`pd.json`，`tidb.json`，`tidb_summary.json`。

4. 点击 Load。
5. 选择一个 Prometheus 数据源。
6. 点击 Import，Prometheus 面板即导入成功。

9.3.3 查看组件 metrics

在顶部菜单中，点击 New dashboard，选择要查看的面板。

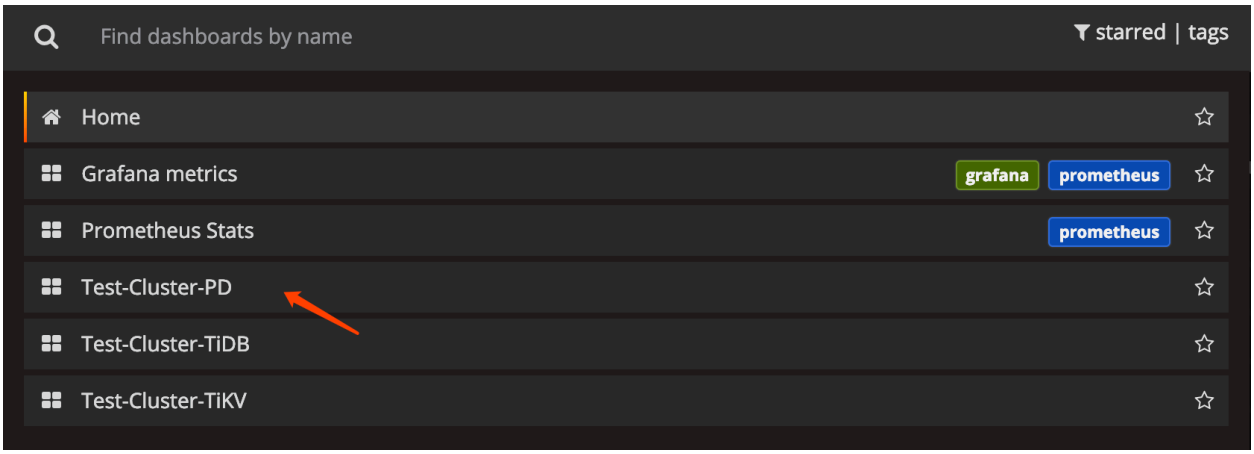


图 46: view dashboard

可查看以下集群组件信息：

- TiDB Server:
  - query 处理时间，可以看到延迟和吞吐
  - ddl 过程监控
  - TiKV client 相关的监控
  - PD client 相关的监控
- PD Server:
  - 命令执行的总次数
  - 某个命令执行失败的总次数
  - 某个命令执行成功的耗时统计
  - 某个命令执行失败的耗时统计
  - 某个命令执行完成并返回结果的耗时统计
- TiKV Server:
  - GC 监控
  - 执行 KV 命令的总次数
  - Scheduler 执行命令的耗时统计
  - Raft propose 命令的总次数
  - Raft 执行命令的耗时统计
  - Raft 执行命令失败的总次数
  - Raft 处理 ready 状态的总次数

9.4 将 Grafana 监控数据导出成快照

在故障诊断中，监控数据十分重要。当你请求远程协助时，技术支持人员有时需要查看 Grafana Dashboard 以确认问题所在。MetricsTool 用于将 Grafana Dashboard 的快照导出为本地文件，并将快照可视化。因此，你可以在不泄露 Grafana 服务器上其他敏感信息的前提下，将监控数据以快照形式分享给外部人员，同时也方便外部人员准确识读数据图表。

9.4.1 使用方法

可以通过访问 <https://metricstool.pingcap.com> 来使用 MetricsTool。它主要提供以下三种功能：

- 导出快照：提供一段在浏览器开发者工具上运行的用户脚本。你可以使用这个脚本在任意 Grafana v6.x.x 服务器上下载当前 Dashboard 中所有可见面板的快照。

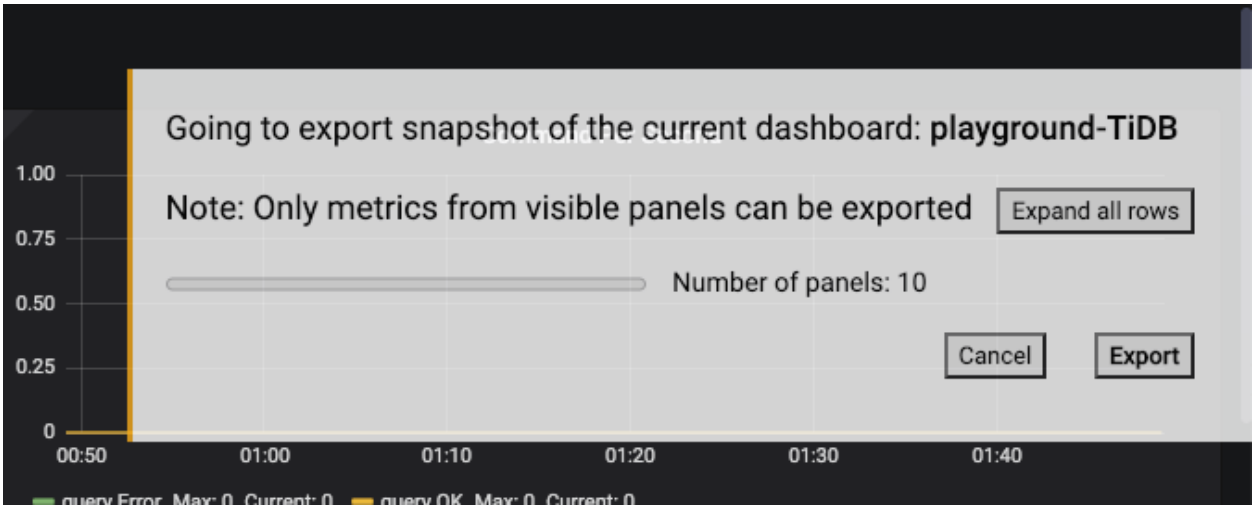


图 47: 运行用户脚本后的 MetricsTool Exporter 截图

- 快照可视化：通过网页端可视化工具 Visualizer 将快照导出文件可视化。快照经过可视化后，操作起来与实际的 Grafana Dashboard 无异。

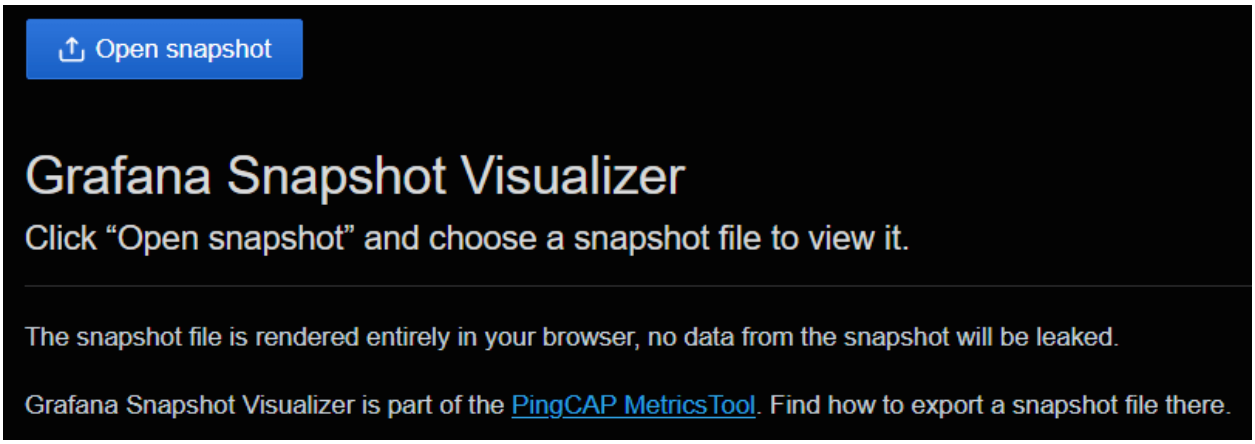


图 48: MetricsTool Visualizer 截图

- 导入快照：介绍如何将导出的快照重新导入到已有的 Grafana 实例中。

9.4.2 FAQs

9.4.2.1 与直接截图及导出 PDF 相比，MetricTool 有什么优势？

MetricsTool 导出的快照文件包含快照生成时的监控指标实际数值。你可以通过 Visualizer 与渲染的图表进行交互，比如切换序列、选择一个较小的时间范围以及检查特定时间点的监控数据值等，就像在操作一个实际的 Grafana Dashboard 一样，因此它比 PDF 文件和截图更强大。

9.4.2.2 快照文件里都包含什么？

快照文件包含所选时间范围内所有图表和面板数据的值，但不保存数据源的原始监控指标，所以无法在 Visualizer 中编辑查询表达式。

9.4.2.3 Visualizer 会将上传的快照文件保存到 PingCAP 的服务器上吗？

不会。快照文件解析全部在浏览器中完成，Visualizer 不会将任何信息发送给 PingCAP。你可以放心地使用 Visualizer 查看带有敏感信息的快照文件，不用担心信息会泄露给第三方。

9.4.2.4 MetricsTool 可以导出除 Grafana 外其他监控工具的数据吗？

不能。目前该工具仅支持在 Grafana v6.x.x 上使用。

9.4.2.5 可以在所有监控指标数据都加载完毕前就运行脚本吗？

可以。虽然脚本会弹出提示，让你等所有监控数据加载完毕后再运行，但可以手动跳过等待并导出快照，以免有些监控数据加载的时间过长。

9.4.2.6 快照文件可视化后，可以通过网页链接分享吗？

不能。但你可以分享快照文件，并说明如何使用 Visualizer 查看。如果确实需要通过网页链接分享，可以尝试使用 Grafana 内置的 `snapshot.raintank.io` 服务。但在这样做之前，要确保不会泄漏隐私信息。

9.5 TiDB 集群报警规则

本文介绍了 TiDB 集群中各组件的报警规则，包括 TiDB、TiKV、PD、TiFlash、TiDB Binlog、TiCDC、Node_exporter 和 Blackbox_exporter 的各报警项的规则描述及处理方法。

按照严重程度由高到低，报警项可分为紧急级别 > 严重级别 > 警告级别三类。该分级适用于以下各组件的报警项。

严重程度	说明
紧急级别	最高严重程度，服务不可用，通常由于服务停止或节点故障导致，此时需要马上进行人工干预
严重级别	服务可用性下降，需要用户密切关注异常指标
警告级别	对某一问题或错误的提醒

9.5.1 TiDB 报警规则

本节介绍了 TiDB 组件的报警项。

9.5.1.1 紧急级别报警项

9.5.1.1.1 TiDB_schema_error

- 报警规则：

  ```
  increase(tidb_session_schema_lease_error_total{type="outdated"}[15m])> 0
  ```

- 规则描述：

  TiDB 在一个 Lease 时间内没有重载到最新的 Schema 信息。如果 TiDB 无法继续对外提供服务，则报警。

- 处理方法：

  该问题通常由于 TiKV Region 不可用或超时导致，需要看 TiKV 的监控指标定位问题。

9.5.1.1.2 TiDB_tikvclient_region_err_total

- 报警规则：

  ```
  increase(tidb_tikvclient_region_err_total[10m])> 6000
  ```

- 规则描述：

  TiDB 访问 TiKV 时发生了 Region 错误。如果在 10 分钟之内该错误多于 6000 次，则报警。

- 处理方法：

  查看 TiKV 的监控状态。

9.5.1.1.3 TiDB_domain_load_schema_total

- 报警规则：

  ```
  increase(tidb_domain_load_schema_total{type="failed"}[10m])> 10
  ```

- 规则描述：

  TiDB 重载最新的 Schema 信息失败的总次数。如果在 10 分钟之内重载失败次数超过 10 次，则报警。

- 处理方法：

  参考TiDB_schema_error 的处理方法。

9.5.1.1.4 TiDB_monitor_keep_alive

- 报警规则：

  ```
  increase(tidb_monitor_keep_alive_total[10m])< 100
  ```

- 规则描述：

  表示 TiDB 的进程是否仍然存在。如果在 10 分钟之内 `tidb_monitor_keep_alive_total` 增加次数少于 100，则 TiDB 的进程可能已经退出，此时会报警。

- 处理方法：

  - 检查 TiDB 进程是否 OOM。
  - 检查机器是否发生了重启。

9.5.1.2 严重级别报警项

9.5.1.2.1 TiDB_server_panic_total

- 报警规则：

  ```
  increase(tidb_server_panic_total[10m])> 0
  ```

- 规则描述：

  发生崩溃的 TiDB 线程的数量。当出现崩溃的时候会报警。该线程通常会被恢复，否则 TiDB 会频繁重启。

- 处理方法：

  收集 panic 日志，定位原因。

9.5.1.3 警告级别报警项

9.5.1.3.1 TiDB_memory_abnormal

- 报警规则：

  ```
  go_memstats_heap_inuse_bytes{job="tidb"} > 1e+10
  ```

- 规则描述：

  对 TiDB 内存使用量的监控。如果内存使用大于 10 G，则报警。

- 处理方法：

  通过 HTTP API 来排查 goroutine 泄露的问题。

9.5.1.3.2 TiDB_query_duration

- 报警规则：

  ```
  histogram_quantile(0.99, sum(rate(tidb_server_handle_query_duration_seconds_bucket[1m]))BY (
  ↪ le, instance))> 1
  ```

- 规则描述：

  TiDB 处理请求的延时。如果.99 的延迟大于 1 秒，则报警。

- 处理方法：

  查看 TiDB 的日志，搜索 SLOW_QUERY 和 TIME_COP_PROCESS 关键字，查找慢 SQL。

9.5.1.3.3 TiDB_server_event_error

- 报警规则：

```
increase(tidb_server_event_total{type=~"server_start|server_hang"}[15m])> 0
```

- 规则描述：

  TiDB 服务中发生的事件数量。当出现以下事件的时候会报警：

  1. start：TiDB 服务启动。
  2. hang：当发生了 Critical 级别的事件时（目前只有 Binlog 写不进去一种情况），TiDB 进入 hang 模式，并等待人工 Kill。

- 处理方法：

  - 重启 TiDB 以恢复服务。
  - 检查 TiDB Binlog 服务是否正常。

9.5.1.3.4 TiDB_tikvclient_backoff_seconds_count

- 报警规则：

```
increase(tidb_tikvclient_backoff_seconds_count[10m])> 10
```

- 规则描述：

  TiDB 访问 TiKV 发生错误时发起重试的次数。如果在 10 分钟之内重试次数多于 10 次，则报警。

- 处理方法：

  查看 TiKV 的监控状态。

9.5.1.3.5 TiDB_monitor_time_jump_back_error

- 报警规则：

```
increase(tidb_monitor_time_jump_back_total[10m])> 0
```

- 规则描述：

  如果 TiDB 所在机器的时间发生了回退，则报警。

- 处理方法：

  排查 NTP 配置。

9.5.1.3.6 TiDB_ddl_waiting_jobs

- 报警规则：

```
sum(tidb_ddl_waiting_jobs)> 5
```

- 规则描述：

  如果 TiDB 中等待执行的 DDL 任务的数量大于 5，则报警。

- 处理方法：

  通过 `admin show ddl` 语句检查是否有耗时的 add index 操作正在执行。

9.5.2 PD 报警规则

本节介绍了 PD 组件的报警项。

9.5.2.1 紧急级别报警项

9.5.2.1.1 PD_cluster_down_store_nums

- 报警规则：

```
(sum(pd_cluster_status{type="store_down_count"})by (instance)> 0)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  PD 长时间（默认配置是 30 分钟）没有收到 TiKV/TiFlash 心跳。

- 处理方法：

  - 检查 TiKV/TiFlash 进程是否正常、网络是否隔离以及负载是否过高，并尽可能地恢复服务。
  - 如果确定 TiKV/TiFlash 无法恢复，可做下线处理。

9.5.2.2 严重级别报警项

9.5.2.2.1 PD_etcd_write_disk_latency

- 报警规则：

```
histogram_quantile(0.99, sum(rate(etcd_disk_wal_fsync_duration_seconds_bucket[1m]))by (
↪ instance, job, le))> 1
```

- 规则描述：

  fsync 操作延迟大于 1s，代表 etcd 写盘慢，这很容易引起 PD leader 超时或者 TSO 无法及时存盘等问题，从而导致整个集群停止服务。

- 处理方法：

  - 排查写入慢的原因。可能是由于其他服务导致系统负载过高。可以检查 PD 本身是否占用了大量 CPU 或 IO 资源。
  - 可尝试重启 PD 或手动 transfer leader 至其他的 PD 来恢复服务。
  - 如果由于环境原因无法恢复，可将有问题的 PD 下线替换。

9.5.2.2.2 PD_miss_peer_region_count

- 报警规则：

```
(sum(pd_regions_status{type="miss_peer_region_count"})by (instance)> 100)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  Region 的副本数小于 `max-replicas` 配置的值。这通常是由于 TiKV 宕机等问题导致一段时间内一些 Region 缺副本。

- 处理方法：

  - 查看是否有 TiKV 宕机或在做下线操作，尝试定位问题产生的原因。
  - 观察 region health 面板，查看 `miss_peer_region_count` 是否在不断减少。

9.5.2.3 警告级别报警项

9.5.2.3.1 PD_cluster_lost_connect_store_nums

- 报警规则：

```
(sum(pd_cluster_status{type="store_disconnected_count"})by (instance)> 0)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  PD 在 20 秒之内未收到 TiKV/TiFlash 上报心跳。正常情况下是每 10 秒收到 1 次心跳。

- 处理方法：

  - 排查是否在重启 TiKV/TiFlash。
  - 检查 TiKV/TiFlash 进程是否正常、网络是否隔离以及负载是否过高，并尽可能地恢复服务。
  - 如果确定 TiKV/TiFlash 无法恢复，可做下线处理。
  - 如果确定 TiKV/TiFlash 可以恢复，但在短时间内还无法恢复，可以考虑延长 `max-down-time` 配置，防止超时后 TiKV/TiFlash 被判定为无法恢复并开始搬移数据。

9.5.2.3.2 PD_cluster_low_space

- 报警规则：

```
(sum(pd_cluster_status{type="store_low_space_count"})by (instance)> 0)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  表示 TiKV/TiFlash 节点空间不足。

- 处理方法：

  - 检查集群中的空间是否普遍不足。如果是，则需要扩容。
  - 检查 Region balance 调度是否有问题。如果有问题，会导致数据分布不均衡。
  - 检查是否有文件占用了大量磁盘空间，比如日志、快照、core dump 等文件。
  - 降低该节点的 Region weight 来减少数据量。
  - 无法释放空间时，可以考虑主动下线该节点，防止由于磁盘空间不足而宕机。

9.5.2.3.3 PD_etcd_network_peer_latency

- 报警规则：

```
histogram_quantile(0.99, sum(rate(etcd_network_peer_round_trip_time_seconds_bucket[1m]))by (
↪ To, instance, job, le))> 1
```

- 规则描述：

  PD 节点之间网络延迟高，严重情况下会导致 leader 超时和 TSO 存盘超时，从而影响集群服务。

- 处理方法：

  - 检查网络状况和系统负载情况。
  - 如果由于环境原因无法恢复，可将有问题的 PD 下线替换。

9.5.2.3.4 PD_tidb_handle_requests_duration

- 报警规则：

```
histogram_quantile(0.99, sum(rate(pd_client_request_handle_requests_duration_seconds_bucket{
↪ type="tso"}[1m]))by (instance, job, le))> 0.1
```

- 规则描述：

  PD 处理 TSO 请求耗时过长，一般是由于负载过高。

- 处理方法：

  - 检查服务器负载状况。
  - 使用 pprof 抓取 PD 的 CPU profile 进行分析。
  - 手动切换 PD leader。
  - 如果是环境问题，则将有问题的 PD 下线替换。

9.5.2.3.5 PD_down_peer_region_nums

- 报警规则：

```
(sum(pd_regions_status{type="down-peer-region-count"})by (instance)> 0)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  Raft leader 上报有不响应 peer 的 Region 数量。

- 处理方法：

  - 检查是否有 TiKV 宕机，或刚发生重启，或者繁忙。
  - 观察 region health 面板，检查 down_peer_region_count 是否在不断减少。
  - 检查是否有 TiKV 之间网络不通。

9.5.2.3.6 PD_pending_peer_region_count

- 报警规则：

```
(sum(pd_regions_status{type="pending-peer-region-count"})by (instance)> 100)and (sum(etcd_server_is_leader
↪ )by (instance)> 0)
```

- 规则描述：

  Raft log 落后的 Region 过多。由于调度产生少量的 pending peer 是正常的，但是如果持续很高，就可能有问题。

- 处理方法：

  - 观察 region health 面板，检查 pending_peer_region_count 是否在不断减少。
  - 检查 TiKV 之间的网络状况，特别是带宽是否足够。

9.5.2.3.7 PD_leader_change

- 报警规则：

```
count(changes(pd_tso_events{type="save"}[10m])> 0)>= 2
```

- 规则描述：

  近期发生了 PD leader 切换。

- 处理方法：

  - 排除人为因素，比如重启 PD、手动 transfer leader 或调整 leader 优先级等。
  - 检查网络状况和系统负载情况。
  - 如果由于环境原因无法恢复，可将有问题的 PD 下线替换。

9.5.2.3.8 TiKV_space_used_more_than_80%

- 报警规则：

```
sum(pd_cluster_status{type="storage_size"})/ sum(pd_cluster_status{type="storage_capacity"})
↪ * 100 > 80
```

- 规则描述：

  集群空间占用超过 80%。

- 处理方法：

  - 确认是否需要扩容。
  - 排查是否有文件占用了大量磁盘空间，比如日志、快照或 core dump 等文件。

9.5.2.3.9 PD_system_time_slow

- 报警规则：

```
changes(pd_tso_events{type="system_time_slow"}[10m])>= 1
```

- 规则描述：

  系统时间可能发生回退。

- 处理方法：

  检查系统时间设置是否正确。

9.5.2.3.10 PD_no_store_for_making_replica

- 报警规则：

```
increase(pd_checker_event_count{type="replica_checker", name="no_target_store"}[1m])> 0
```

- 规则描述：

  没有合适的 store 用来补副本。

- 处理方法：

  - 检查 store 是否空间不足。
  - 根据 label 配置（如果有这个配置的话）来检查是否有可以补副本的 store。

9.5.3 TiKV 报警规则

本节介绍了 TiKV 组件的报警项。

9.5.3.1 紧急级别报警项

9.5.3.1.1 TiKV_memory_used_too_fast

- 报警规则：

```
process_resident_memory_bytes{job=~"tikv",instance=~".*"} - (process_resident_memory_bytes{
↪ job=~"tikv",instance=~".*"} offset 5m)> 5*1024*1024*1024
```

- 规则描述：

  目前没有和内存相关的 TiKV 的监控，你可以通过 Node_exporter 监控集群内机器的内存使用情况。如上规则表示，如果在 5 分钟之内内存使用超过 5GB（TiKV 内存占用的太快），则报警。

- 处理方法：

  调整 `rockdb.defaultcf` 和 `rocksdb.writecf` 的 `block-cache-size` 的大小。

9.5.3.1.2 TiKV_GC_can_not_work

- 报警规则：

```
sum(increase(tikv_gcworker_gc_tasks_vec{task="gc"}[1d]))< 1 and sum(increase(tikv_gc_compaction_filter_perf
↪ [1d]))< 1
```

- 规则描述：

  在 24 小时内一个 TiKV 实例上没有成功执行 GC，说明 GC 不能正常工作了。短期内 GC 不运行不会造成太大的影响，但如果 GC 一直不运行，版本会越来越多，从而导致查询变慢。

- 处理方法：

  1. 执行 `SELECT VARIABLE_VALUE FROM mysql.tidb WHERE VARIABLE_NAME="tikv_gc_leader_desc"` 来找到 gc leader 对应的 `tidb-server`；
  2. 查看该 `tidb-server` 的日志，grep gc_worker tidb.log；
  3. 如果发现这段时间一直在 resolve locks（最后一条日志是 `start resolve locks`）或者 delete ranges（最后一条日志是 `start delete {number} ranges`），说明 GC 进程是正常的。否则需要报备开发人员 support@pingcap.com 进行处理。

9.5.3.2 严重级别报警项

9.5.3.2.1 TiKV_server_report_failure_msg_total

- 报警规则：

```
sum(rate(tikv_server_report_failure_msg_total{type="unreachable"}[10m]))BY (store_id)> 10
```

- 规则描述：

  表明无法连接远端的 TiKV。

- 处理方法：

  1. 检查网络是否通畅。
  2. 检查远端 TiKV 是否挂掉。
  3. 如果远端 TiKV 没有挂掉，检查压力是否太大，参考TiKV_channel_full_total 处理方法。

9.5.3.2.2 TiKV_channel_full_total

- 报警规则：

```
sum(rate(tikv_channel_full_total[10m]))BY (type, instance)> 0
```

- 规则描述：

  该错误通常是因为 Raftstore 线程卡死，TiKV 的压力已经非常大了。

- 处理方法：

1. 观察 Raft Propose 监控，看这个报警的 TiKV 节点是否明显有比其他 TiKV 高很多。如果是，表明这个 TiKV 上有热点，需要检查热点调度是否能正常工作。
2. 观察 Raft IO 监控，看延迟是否升高。如果延迟很高，表明磁盘可能有瓶颈。一个能缓解但不怎么安全的办法是将 sync-log 改成 false。
3. 观察 Raft Process 监控，看 tick duration 是否很高。如果是，需要在 [raftstore] 配置下加上 raft-base-tick-interval = “2s”。

9.5.3.2.3 TiKV_write_stall

- 报警规则：

```
delta(tikv_engine_write_stall[10m])> 0
```

- 规则描述：

RocksDB 写入压力太大，出现了 stall。

- 处理方法：

1. 观察磁盘监控，排除磁盘问题。
2. 看 TiKV 是否有写入热点。
3. 在 [rocksdb] 和 [raftdb] 配置下调大 max-sub-compactions 的值。

9.5.3.2.4 TiKV_raft_log_lag

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_raftstore_log_lag_bucket[1m]))by (le, instance))>
↪ 5000
```

- 规则描述：

这个值偏大，表明 Follower 已经远远落后于 Leader，Raft 没法正常同步了。可能的原因是 Follower 所在的 TiKV 卡住或者挂掉了。

9.5.3.2.5 TiKV_async_request_snapshot_duration_seconds

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_storage_engine_async_request_duration_seconds_bucket{
↪ type="snapshot"}[1m]))by (le, instance, type))> 1
```

- 规则描述：

这个值偏大，表明 Raftstore 负载压力很大，可能已经卡住。

- 处理方法：

参考TiKV_channel_full_total 的处理方法。

9.5.3.2.6 TiKV_async_request_write_duration_seconds

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_storage_engine_async_request_duration_seconds_bucket{
↪ type="write"}[1m]))by (le, instance, type))> 1
```

- 规则描述：

  这个值偏大，表明 Raft write 耗时很长。

- 处理方法：

  1. 检查 Raftstore 上的压力，参考TiKV_channel_full_total 的处理方法。
  2. 检查 apply worker 线程的压力。

9.5.3.2.7 TiKV_coprocessor_request_wait_seconds

- 报警规则：

```
histogram_quantile(0.9999, sum(rate(tikv_coprocessor_request_wait_seconds_bucket[1m]))by (le
↪ , instance, req))> 10
```

- 规则描述：

  这个值偏大，表明 Coprocessor worker 压力很大。可能有比较慢的任务卡住了 Coprocessor 线程。

- 处理方法：

  1. 从 TiDB 日志中查看慢查询日志，看查询是否用到了索引或全表扫，或者看是否需要做 analyze。
  2. 排查是否有热点。
  3. 查看 Coprocessor 监控，看 `coporcessor table/index scan` 里 `total` 和 `process` 是否匹配。如果相差太大，表明做了太多的无效查询。看是否有 `over seek bound`，如果有，表明版本太多，GC 工作不及时，需要增大并行 GC 的线程数。

9.5.3.2.8 TiKV_raftstore_thread_cpu_seconds_total

- 报警规则：

```
sum(rate(tikv_thread_cpu_seconds_total{name=~"raftstore_.*"}[1m]))by (instance, name)> 1.6
```

- 规则描述：

  Raftstore 线程压力太大。

- 处理方法：

  参考TiKV_channel_full_total 的处理方法。

9.5.3.2.9 TiKV_raft_append_log_duration_secs

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_raftstore_append_log_duration_seconds_bucket[1m]))by
↪ (le, instance))> 1
```

- 规则描述：

表示 append Raft log 的耗时，如果高，通常是因为 IO 太忙了。

9.5.3.2.10 TiKV_raft_apply_log_duration_secs

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_raftstore_apply_log_duration_seconds_bucket[1m]))by (
↪ le, instance))> 1
```

- 规则描述：

表示 apply Raft log 耗时，如果高，通常是因为 IO 太忙了。

9.5.3.2.11 TiKV_scheduler_latch_wait_duration_seconds

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_scheduler_latch_wait_duration_seconds_bucket[1m]))by
↪ (le, instance, type))> 1
```

- 规则描述：

Scheduler 中写操作获取内存锁时的等待时间。如果这个值高，表明写操作冲突较多，也可能是某些引起冲突的操作耗时较长，阻塞了其它等待相同锁的操作。

- 处理方法：

1. 查看 Scheduler-All 监控中的 scheduler command duration，看哪一个命令耗时最大。
2. 查看 Scheduler-All 监控中的 scheduler scan details，看 `total` 和 `process` 是否匹配。如果相差太大，表明有很多无效的扫描，另外观察是否有 `over seek bound`，如果太多，表明 GC 不及时。
3. 查看 Storage 监控中的 storage async snapshot/write duration，看是否 Raft 操作不及时。

9.5.3.2.12 TiKV_thread_apply_worker_cpu_seconds

- 报警规则：

```
max(rate(tikv_thread_cpu_seconds_total{name=~"apply_.*"}[1m]))by (instance)> 0.9
```

- 规则描述：

Apply Raft log 线程压力太大，已经接近或超过 apply 线程的处理上限。通常是因为短期内写入的数据量太多造成的。

9.5.3.3 警告级别报警项

9.5.3.3.1 TiKV_leader_drops

- 报警规则：

```
delta(tikv_pd_heartbeat_tick_total{type="leader"}[30s])< -10
```

- 规则描述：

  该问题通常是因为 Raftstore 线程卡住了。

- 处理方法：

  1. 参考TiKV_channel_full_total 的处理方法。
  2. 如果 TiKV 压力很小，考虑 PD 的调度是否太频繁。可以查看 PD 页面的 Operator Create 面板，排查 PD 产生调度的类型和数量。

9.5.3.3.2 TiKV_raft_process_ready_duration_secs

- 报警规则：

```
histogram_quantile(0.999, sum(rate(tikv_raftstore_raft_process_duration_secs_bucket{type='
↪ ready'}[1m]))by (le, instance, type))> 2
```

- 规则描述：

  表示处理 Raft ready 的耗时。这个值大，通常是因为 append log 任务卡住了。

9.5.3.3.3 TiKV_raft_process_tick_duration_secs

- 报警规则：

```
histogram_quantile(0.999, sum(rate(tikv_raftstore_raft_process_duration_secs_bucket{type='
↪ tick' }[1m]))by (le, instance, type))> 2
```

- 规则描述：

  表示处理 Raft tick 的耗时，这个值大，通常是因为 Region 太多导致的。

- 处理方法：

  1. 考虑使用更高等级的日志，比如 warn 或者 error。
  2. 在 [raftstore] 配置下添加 raft-base-tick-interval = “2s”。

9.5.3.3.4 TiKV_scheduler_context_total

- 报警规则：

  ```
  abs(delta( tikv_scheduler_context_total[5m]))> 1000
  ```

- 规则描述：

  Scheduler 正在执行的写命令数量。这个值高，表示任务完成得不及时。

- 处理方法：

  参考TiKV_scheduler_latch_wait_duration_seconds 的处理方法。

9.5.3.3.5 TiKV_scheduler_command_duration_seconds

- 报警规则：

  ```
  histogram_quantile(0.99, sum(rate(tikv_scheduler_command_duration_seconds_bucket[1m]))by (le
  ↪ , instance, type)/ 1000)> 1
  ```

- 规则描述：

  表明 Scheduler 执行命令的耗时。

- 处理方法：

  参考TiKV_scheduler_latch_wait_duration_seconds 的处理方法。

9.5.3.3.6 TiKV_coprocessor_outdated_request_wait_seconds

- 报警规则：

  ```
  delta(tikv_coprocessor_outdated_request_wait_seconds_count[10m])> 0
  ```

- 规则描述：

  Coprocessor 已经过期的请求等待的时间。这个值高，表示 Coprocessor 压力已经非常大了。

- 处理方法：

  参考TiKV_coprocessor_request_wait_seconds 的处理方法。

9.5.3.3.7 TiKV_coprocessor_pending_request

- 报警规则：

  ```
  delta(tikv_coprocessor_pending_request[10m])> 5000
  ```

- 规则描述：

  Coprocessor 排队的请求。

- 处理方法：

  参考TiKV_coprocessor_request_wait_seconds 的处理方法。

9.5.3.3.8 TiKV_batch_request_snapshot_nums

- 报警规则：

```
sum(rate(tikv_thread_cpu_seconds_total{name=~"cop_.*"}[1m]))by (instance)/ (count(tikv_thread_cpu_seconds_t
↪ {name=~"cop_.*"})* 0.9)/ count(count(tikv_thread_cpu_seconds_total)by (instance))> 0
```

- 规则描述：

  某个 TiKV 的 Coprocessor CPU 使用率超过了 90%。

9.5.3.3.9 TiKV_pending_task

- 报警规则：

```
sum(tikv_worker_pending_task_total)BY (instance,name)> 1000
```

- 规则描述：

  TiKV 等待的任务数量。

- 处理方法：

  查看是哪一类任务的值偏高，通常 Coprocessor、apply worker 这类任务都可以在其他指标里找到解决办法。

9.5.3.3.10 TiKV_low_space

- 报警规则：

```
sum(tikv_store_size_bytes{type="available"})by (instance)/ sum(tikv_store_size_bytes{type="
↪ capacity"})by (instance)< 0.2
```

- 规则描述：

  TiKV 数据量超过节点配置容量或物理磁盘容量的 80%。

- 处理方法：

  确认节点空间均衡情况，做好扩容计划。

9.5.3.3.11 TiKV_approximate_region_size

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tikv_raftstore_region_size_bucket[1m]))by (le))> 1073741824
↪
```

- 规则描述：

  TiKV split checker 扫描到的最大的 Region approximate size 在 1 分钟内持续大于 1 GB。

- 处理方法：

  Region 分裂的速度不及写入的速度。为缓解这种情况，建议更新到支持 batch-split 的版本 (>= 2.1.0-rc1)。如暂时无法更新，可以使用 `pd-ctl operator add split-region <region_id> --policy=approximate` 手动分裂 Region。

9.5.4 TiFlash 报警规则

关于 TiFlash 报警规则的详细描述，参见TiFlash 报警规则。

9.5.5 TiDB Binlog 报警规则

关于 TiDB Binlog 报警规则的详细描述，参见TiDB Binlog 集群监控报警文档。

9.5.6 TiCDC 报警规则

关于 TiCDC 报警规则的详细描述，参见TiCDC 集群监控报警。

9.5.7 Node_exporter 主机报警规则

本节介绍了 Node_exporter 主机的报警项。

9.5.7.1 紧急级别报警项

9.5.7.1.1 NODE_disk_used_more_than_80%

- 报警规则：

```
node_filesystem_avail{fstype=~"(ext.|xfs)", mountpoint!~"/boot"} / node_filesystem_size{
↪ fstype=~"(ext.|xfs)", mountpoint!~"/boot"} * 100 <= 20
```

- 规则描述：

  机器磁盘空间使用率超过 80%。

- 处理方法：

  登录机器，执行 `df -h` 命令，查看磁盘空间使用率，做好扩容计划。

9.5.7.1.2 NODE_disk_inode_more_than_80%

- 报警规则：

```
node_filesystem_files_free{fstype=~"(ext.|xfs)"} / node_filesystem_files{fstype=~"(ext.|xfs)
↪ "} * 100 < 20
```

- 规则描述：

  机器磁盘挂载目录文件系统 inode 使用率超过 80%。

- 处理方法：

  登录机器，执行 `df -i` 命令，查看磁盘挂载目录文件系统 inode 使用率，做好扩容计划。

9.5.7.1.3 NODE_disk_readonly

- 报警规则：

```
node_filesystem_readonly{fstype=~"(ext.|xfs)"} == 1
```

- 规则描述：

  磁盘挂载目录文件系统只读，无法写入数据，一般是因为磁盘故障或文件系统损坏。

- 处理方法：

  - 登录机器创建文件测试是否正常。
  - 检查该服务器硬盘指示灯是否正常，如异常，需更换磁盘并修复该机器文件系统。

9.5.7.2 严重级别报警项

9.5.7.2.1 NODE_memory_used_more_than_80%

- 报警规则：

```
(((node_memory_MemTotal-node_memory_MemFree-node_memory_Cached)/(node_memory_MemTotal)*100))
↪ >= 80
```

- 规则描述：

  机器内存使用率超过 80%。

- 处理方法：

  - 在 Grafana Node Exporter 页面查看该主机的 Memory 面板，检查 Used 是否过高，Available 内存是否过低。
  - 登录机器，执行 `free -m` 命令查看内存使用情况，执行 top 看是否有异常进程的内存使用率过高。

9.5.7.3 警告级别报警项

9.5.7.3.1 NODE_node_overload

- 报警规则：

```
(node_load5 / count without (cpu, mode)(node_cpu_seconds_total{mode="system"}))> 1
```

- 规则描述：

  机器 CPU 负载较高。

- 处理方法：

  - 在 Grafana Node exporter 页面上查看该主机的 CPU Usage 及 Load Average，检查是否过高。
  - 登录机器，执行 top 查看 load average 及 CPU 使用率，看是否是异常进程的 CPU 使用率过高。

9.5.7.3.2 NODE_cpu_used_more_than_80%

- 报警规则：

```
avg(irate(node_cpu_seconds_total{mode="idle"}[5m]))by(instance)* 100 <= 20
```

- 规则描述：

  机器 CPU 使用率超过 80%。

- 处理方法：

  - 在 Grafana Node exporter 页面上查看该主机的 CPU Usage 及 Load Average，检查是否过高。
  - 登录机器，执行 top 查看 load average 及 CPU 使用率，看是否是异常进程的 CPU 使用率过高。

9.5.7.3.3 NODE_tcp_estab_num_more_than_50000

- 报警规则：

```
node_netstat_Tcp_CurrEstab > 50000
```

- 规则描述：

  机器 `establish` 状态的 TCP 链接超过 50,000。

- 处理方法：

  登录机器执行 `ss -s` 可查看当前系统 `estab` 状态的 TCP 链接数，执行 `netstat` 查看是否有异常链接。

9.5.7.3.4 NODE_disk_read_latency_more_than_32ms

- 报警规则：

```
((rate(node_disk_read_time_seconds_total{device=~".+"}[5m])/ rate(node_disk_reads_completed_total
↪ {device=~".+"}[5m]))or (irate(node_disk_read_time_seconds_total{device=~".+"}[5m])/ irate
↪ (node_disk_reads_completed_total{device=~".+"}[5m])))* 1000 > 32
```

- 规则描述：

  磁盘读延迟超过 32 毫秒。

- 处理方法：

  - 查看 Grafana Disk Performance Dashboard 观察磁盘使用情况。
  - 查看 Disk Latency 面板观察磁盘的读延迟。
  - 查看 Disk IO Utilization 面板观察 IO 使用率。

9.5.7.3.5 NODE_disk_write_latency_more_than_16ms

- 报警规则：

```
((rate(node_disk_write_time_seconds_total{device=~".+"}[5m])/ rate(node_disk_writes_completed_total
↪ {device=~".+"}[5m]))or (irate(node_disk_write_time_seconds_total{device=~".+"}[5m])/
↪ irate(node_disk_writes_completed_total{device=~".+"}[5m])))> 16
```

- 规则描述：

  机器磁盘写延迟超过 16 毫秒。

- 处理方法：

  - 查看 Grafana Disk Performance Dashboard 观察磁盘使用情况。
  - 查看 Disk Latency 面板可查看磁盘的写延迟。
  - 查看 Disk IO Utilization 面板可查看 IO 使用率。

9.5.8 Blackbox_exporter TCP、ICMP 和 HTTP 报警规则

本节介绍了 Blackbox_exporter TCP、ICMP 和 HTTP 的报警项。

9.5.8.1 紧急级别报警项

9.5.8.1.1 TiDB_server_is_down

- 报警规则：

```
probe_success{group="tidb"} == 0
```

- 规则描述：

  TiDB 服务端口探测失败。

- 处理方法：

  - 检查 TiDB 服务所在机器是否宕机。
  - 检查 TiDB 进程是否存在。
  - 检查监控机与 TiDB 服务所在机器之间网络是否正常。

9.5.8.1.2 TiFlash_server_is_down

- 报警规则：

```
probe_success{group="tiflash"} == 0
```

- 规则描述：

  TiFlash 服务端口探测失败。

- 处理方法：

  - 检查 TiFlash 服务所在机器是否宕机。
  - 检查 TiFlash 进程是否存在。
  - 检查监控机与 TiFlash 服务所在机器之间网络是否正常。

9.5.8.1.3 Pump_server_is_down

- 报警规则：

  ```
  probe_success{group="pump"} == 0
  ```

- 规则描述：

  Pump 服务端口探测失败。

- 处理方法：

  - 检查 Pump 服务所在机器是否宕机。
  - 检查 Pump 进程是否存在。
  - 检查监控机与 Pump 服务所在机器之间网络是否正常。

9.5.8.1.4 Drainer_server_is_down

- 报警规则：

  ```
  probe_success{group="drainer"} == 0
  ```

- 规则描述：

  Drainer 服务端口探测失败。

- 处理方法：

  - 检查 Drainer 服务所在机器是否宕机。
  - 检查 Drainer 进程是否存在。
  - 检查监控机与 Drainer 服务所在机器之间网络是否正常。

9.5.8.1.5 TiKV_server_is_down

- 报警规则：

  ```
  probe_success{group="tikv"} == 0
  ```

- 规则描述：

  TiKV 服务端口探测失败。

- 处理方法：

  - 检查 TiKV 服务所在机器是否宕机。
  - 检查 TiKV 进程是否存在。
  - 检查监控机与 TiKV 服务所在机器之间网络是否正常。

9.5.8.1.6 PD_server_is_down

- 报警规则：

  ```
  probe_success{group="pd"} == 0
  ```

- 规则描述：

  PD 服务端口探测失败。

- 处理方法：

  - 检查 PD 服务所在机器是否宕机。
  - 检查 PD 进程是否存在。
  - 检查监控机与 PD 服务所在机器之间网络是否正常。

9.5.8.1.7 Node_exporter_server_is_down

- 报警规则：

  ```
  probe_success{group="node_exporter"} == 0
  ```

- 规则描述：

  Node_exporter 服务端口探测失败。

- 处理方法：

  - 检查 Node_exporter 服务所在机器是否宕机。
  - 检查 Node_exporter 进程是否存在。
  - 检查监控机与 Node_exporter 服务所在机器之间网络是否正常。

9.5.8.1.8 Blackbox_exporter_server_is_down

- 报警规则：

  ```
  probe_success{group="blackbox_exporter"} == 0
  ```

- 规则描述：

  Blackbox_exporter 服务端口探测失败。

- 处理方法：

  - 检查 Blackbox_exporter 服务所在机器是否宕机。
  - 检查 Blackbox_exporter 进程是否存在。
  - 检查监控机与 Blackbox_exporter 服务所在机器之间网络是否正常。

9.5.8.1.9 Grafana_server_is_down

- 报警规则：

  ```
  probe_success{group="grafana"} == 0
  ```

- 规则描述：

  Grafana 服务端口探测失败。

- 处理方法：

  - 检查 Grafana 服务所在机器是否宕机。
  - 检查 Grafana 进程是否存在。
  - 检查监控机与 Grafana 服务所在机器之间网络是否正常。

9.5.8.1.10 Pushgateway_server_is_down

- 报警规则：

  ```
  probe_success{group="pushgateway"} == 0
  ```

- 规则描述：

  Pushgateway 服务端口探测失败。

- 处理方法：

  - 检查 Pushgateway 服务所在机器是否宕机。
  - 检查 Pushgateway 进程是否存在。
  - 检查监控机与 Pushgateway 服务所在机器之间网络是否正常。

9.5.8.1.11 Kafka_exporter_is_down

- 报警规则：

  ```
  probe_success{group="kafka_exporter"} == 0
  ```

- 规则描述：

  Kafka_exporter 服务端口探测失败。

- 处理方法：

  - 检查 Kafka_exporter 服务所在机器是否宕机。
  - 检查 Kafka_exporter 进程是否存在。
  - 检查监控机与 Kafka_exporter 服务所在机器之间网络是否正常。

9.5.8.1.12 Pushgateway_metrics_interface

- 报警规则：

  ```
  probe_success{job="blackbox_exporter_http"} == 0
  ```

- 规则描述：

  Pushgateway 服务 http 接口探测失败。

- 处理方法：

  - 检查 Pushgateway 服务所在机器是否宕机。
  - 检查 Pushgateway 进程是否存在。
  - 检查监控机与 Pushgateway 服务所在机器之间网络是否正常。

9.5.8.2 警告级别报警项

9.5.8.2.1 BLACKER_ping_latency_more_than_1s

- 报警规则：

  ```
  max_over_time(probe_duration_seconds{job=~"blackbox_exporter.*_icmp"}[1m])> 1
  ```

- 规则描述：

  Ping 延迟超过 1 秒。

- 处理方法：

  - 在 Grafana Blackbox Exporter 页面上检查两个节点间的 ping 延迟是否太高。
  - 在 Grafana Node Exporter 页面的 TCP 面板上检查是否有丢包。

9.6 TiFlash 报警规则

本文介绍了 TiFlash 集群的报警规则。

9.6.1 TiFlash_schema_error

- 报警规则：

  ```
  increase(tiflash_schema_apply_count{type="failed"}[15m])> 0
  ```

- 规则描述：

  出现 schema apply 错误时报警。

- 处理方法：

  可能是逻辑问题，联系 TiFlash 开发人员。

9.6.2 TiFlash_schema_apply_duration

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tiflash_schema_apply_duration_seconds_bucket[1m]))BY (le,
↪ instance))> 20
```

- 规则描述：

apply 时间超过 20 秒的概率超过 99% 时报警。

- 处理方法：

可能是 TiFlash 存储引擎内部问题，联系 TiFlash 开发人员。

9.6.3 TiFlash_raft_read_index_duration

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tiflash_raft_read_index_duration_seconds_bucket[1m]))BY (
↪ le, instance))> 3
```

- 规则描述：

read index 时间超过 3 秒的概率超过 99% 时报警。

> 注意：
>
> read index 请求是发送给 TiKV leader 的 kvproto 请求，TiKV region 的重试，或 Store 的繁忙/网络问题都可能导致 read index 请求时间过长。

- 处理方法：

可能 TiKV 集群分裂/迁移频繁，导致频繁重试，可以查看 TiKV 集群状态确认。

9.6.4 TiFlash_raft_wait_index_duration

- 报警规则：

```
histogram_quantile(0.99, sum(rate(tiflash_raft_wait_index_duration_seconds_bucket[1m]))BY (
↪ le, instance))> 2
```

- 规则描述：

TiFlash 等待 Region Raft Index 的时间超过 2 秒的概率超过 99% 时报警。

- 处理方法：

可能 TiKV 和 Proxy 的通信出现问题，联系 TiFlash 开发人员确认。

9.7 处理告警

本文档介绍 DM 中各主要告警信息的处理方法。

9.7.1 高可用告警

9.7.1.1 DM_master_all_down

当全部 DM-master 离线时触发该告警。发生该错误时，需要检查集群环境，并通过各节点日志排查错误。

9.7.1.2 DM_worker_offline

存在离线的 DM-worker 超过一小时会触发该告警。在高可用架构下，该告警可能不会直接中断任务，但是会提升任务中断的风险。处理告警可以查看对应 DM-worker 节点的工作状态，检查是否连通，并通过日志排查错误。

9.7.1.3 DM_DDL_error

处理 shard DDL 时出现错误，此时需要参考DM 故障诊断进行处理。

9.7.1.4 DM_pending_DDL

存在未完成的 shard DDL 并超过一小时会触发该告警。在某些应用场景下，存在未完成的 shard DDL 可能是用户所期望的。在用户预期以外的场景下，可以通过手动处理 Sharding DDL Lock解决。

9.7.2 任务状态告警

9.7.2.1 DM_task_state

当 DM-worker 内有子任务处于 Paused 状态超过 20 分钟时会触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.3 relay log 告警

9.7.3.1 DM_relay_process_exits_with_error

当 relay log 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.3.2 DM_remain_storage_of_relay_log

当 relay log 所在磁盘的剩余可用容量小于 10G 时会触发该告警，对应的处理方法包括：

- 手动清理该磁盘上其他无用数据以增加可用容量。
- 尝试调整 relay log 的自动清理策略或执行手动清理。
- 使用 pause-relay 命令暂停 relay log 的拉取，并在磁盘空间合适之后使用 resume-relay 命令恢复。需要注意上游数据源不要清理尚未拉取的 binlog。

9.7.3.3 DM_relay_log_data_corruption

当 relay log 处理单元在校验从上游读取到的 binlog event 且发现 checksum 信息异常时会转为 Paused 状态并立即触发告警，此时需要参考DM 故障诊断进行处理。

9.7.3.4 DM_fail_to_read_binlog_from_master

当 relay log 处理单元在尝试从上游读取 binlog event 发生错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.3.5 DM_fail_to_write_relay_log

当 relay log 处理单元在尝试将 binlog event 写入 relay log 文件发生错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.3.6 DM_binlog_file_gap_between_master_relay

当 relay log 处理单元已拉取到的最新的 binlog 文件个数落后于当前上游 MySQL/MariaDB 超过 1 个（不含 1 个）且持续 10 分钟时会触发该告警，此时需要参考性能问题及处理方法对 relay log 处理单元相关的性能问题进行排查与处理。

9.7.4 Dump/Load 告警

9.7.4.1 DM_dump_process_exists_with_error

当 Dump 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.4.2 DM_load_process_exists_with_error

当 Load 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.5 Binlog replication 告警

9.7.5.1 DM_sync_process_exists_with_error

当 Binlog replication 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

9.7.5.2 DM_binlog_file_gap_between_master_syncer

当 Binlog replication 处理单元已处理到的最新的 binlog 文件个数落后于当前上游 MySQL/MariaDB 超过 1 个（不含 1 个）且持续 10 分钟时 DM 会触发该告警，此时需要参考性能问题及处理方法对 Binlog replication 处理单元相关的性能问题进行排查与处理。

9.7.5.3 DM_binlog_file_gap_between_relay_syncer

当 Binlog replication 处理单元已处理到的最新的 binlog 文件个数落后于当前 relay log 处理单元超过 1 个（不含 1 个）且持续 10 分钟时 DM 会触发该告警，此时需要参考性能问题及处理方法对 Binlog replication 处理单元相关的性能问题进行排查与处理。

9.8 Grafana Dashboard 监控指标

9.8.1 Overview 面板重要监控指标详解

使用 TiUP 部署 TiDB 集群时，一键部署监控系统 (Prometheus & Grafana)，监控架构参见TiDB 监控框架概述。

目前 Grafana Dashboard 整体分为 PD、 TiDB、 TiKV、 Node_exporter、 Overview 等。

对于日常运维，我们单独挑选出重要的 Metrics 放在 Overview 页面，方便日常运维人员观察集群组件 (PD, TiDB, TiKV) 使用状态以及集群使用状态。

以下为 Overview Dashboard 监控说明：

9.8.1.1 Services Port Status

- Services Up：各服务在线节点数量

9.8.1.2 PD

- PD role：当前 PD 的角色
- Storage capacity：TiDB 集群总可用数据库空间大小
- Current storage size：TiDB 集群目前已用数据库空间大小，TiKV 多副本的空间占用也会包含在内
- Normal stores：处于正常状态的节点数目
- Abnormal stores：处于异常状态的节点数目，正常情况应当为 0
- Number of Regions：当前集群的 Region 总量，请注意 Region 数量与副本数无关
- 99% completed_cmds_duration_seconds：单位时间内，99% 的 pd-server 请求执行时间小于监控曲线的值，一般 <= 5ms
- Handle_requests_duration_seconds：PD 发送请求的网络耗时
- Region health：每个 Region 的状态，通常情况下，pending 的 peer 应该少于 100，miss 的 peer 不能一直大于 0
- Hot write Region’s leader distribution：每个 TiKV 实例上是写入热点的 leader 的数量
- Hot read Region’s leader distribution：每个 TiKV 实例上是读取热点的 leader 的数量
- Region heartbeat report：TiKV 向 PD 发送的心跳个数
- 99% Region heartbeat latency：99% 的情况下，心跳的延迟

9.8.1.3 TiDB

- Statement OPS：不同类型 SQL 语句每秒执行的数量。按 SELECT、 INSERT、 UPDATE 等来统计
- Duration：执行的时间
  - 客户端网络请求发送到 TiDB，到 TiDB 执行结束后返回给客户端的时间。一般情况下，客户端请求都是以 SQL 语句的形式发送，但也可以包含 COM_PING、COM_SLEEP、COM_STMT_FETCH、COM_SEND_LONG_DATA 之类的命令执行的时间
  - 由于 TiDB 支持 Multi-Query，因此，可以接受客户端一次性发送的多条 SQL 语句，如：`select 1; select 1;`。此时，统计的执行时间是所有 SQL 执行完之后的总时间
- CPS By Instance：每个 TiDB 实例上的命令统计。按照命令和执行结果成功或失败来统计

- Failed Query OPM：每个 TiDB 实例上，每秒钟执行 SQL 语句发生错误按照错误类型的统计（例如语法错误、主键冲突等）。包含了错误所属的模块和错误码
- Connection count：每个 TiDB 的连接数
- Memory Usage：每个 TiDB 实例的内存使用统计，分为进程占用内存和 Golang 在堆上申请的内存
- Transaction OPS：每秒事务执行数量统计
- Transaction Duration：事务执行的时间
- KV Cmd OPS：KV 命令执行数量统计
- KV Cmd Duration 99：KV 命令执行的时间
- PD TSO OPS：TiDB 每秒从 PD 获取 TSO 的数量
- PD TSO Wait Duration：TiDB 等待从 PD 获取 TS 的时间
- TiClient Region Error OPS：TiKV 返回 Region 相关错误信息的数量
- Lock Resolve OPS：TiDB 清理锁操作的数量。当 TiDB 的读写请求遇到锁时，会尝试进行锁清理
- Load Schema Duration：TiDB 从 TiKV 获取 Schema 的时间
- KV Backoff OPS：TiKV 返回错误信息的数量

9.8.1.4 TiKV

- leader：各个 TiKV 节点上 Leader 的数量分布
- region：各个 TiKV 节点上 Region 的数量分布
- CPU：各个 TiKV 节点的 CPU 使用率
- Memory：各个 TiKV 节点的内存使用量
- store size：每个 TiKV 实例的使用的存储空间的大小
- cf size：每个列族的大小
- channel full：每个 TiKV 实例上 channel full 错误的数量，正常情况下应当为 0
- server report failures：每个 TiKV 实例上报错的消息个数，正常情况下应当为 0
- scheduler pending commands：每个 TiKV 实例上 pending 命令的个数
- coprocessor executor count：TiKV 每秒收到的 coprocessor 操作数量，按照 coprocessor 类型统计
- coprocessor request duration：处理 coprocessor 读请求所花费的时间
- raft store CPU：raftstore 线程的 CPU 使用率，线程数量默认为 2（通过 `raftstore.store-pool-size` 配置）。如果单个线程使用率超过 80%，说明使用率很高
- Coprocessor CPU：coprocessor 线程的 CPU 使用率

9.8.1.5 System Info

- Vcores：CPU 核心数量
- Memory：内存总大小
- CPU Usage：CPU 使用率，最大为 100%
- Load [1m]：1 分钟的负载情况
- Memory Available：剩余内存大小
- Network Traffic：网卡流量统计
- TCP Retrans：TCP 重传数量统计
- IO Util：磁盘使用率，最高为 100%，一般到 80% - 90% 就需要考虑加节点

9.8.1.6 图例



图 49: overview

9.8.2 TiDB 重要监控指标详解

使用 TiUP 部署 TiDB 集群时，你可以一键部署监控系统 (Prometheus & Grafana)，参考监控架构TiDB 监控框架概述。

目前 Grafana Dashboard 整体分为 PD、TiDB、TiKV、Node_exporter、Overview 等。TiDB 分为 TiDB 和 TiDB Summary 面板，两个面板的区别如下：

- TiDB 面板：提供尽可能全面的信息，供排查集群异常。
- TiDB Summary 面板：将 TiDB 面板中用户最为关心的部分抽取出来，并做了些许修改。主要用于提供数据库日常运行中用户关心的数据，如 QPS、TPS、响应延迟等，以便作为外部展示、汇报用的监控信息。

以下为 TiDB Dashboard 关键监控指标的说明：

9.8.2.1 关键指标说明

- Query Summary

  - Duration：执行时间
    * 客户端网络请求发送到 TiDB，到 TiDB 执行结束后返回给客户端的时间。一般情况下，客户端请求都是以 SQL 语句的形式发送，但也可以包含 COM_PING、COM_SLEEP、COM_STMT_FETCH、COM_SEND_LONG_DATA 之类的命令执行时间。
    * 由于 TiDB 支持 Multi-Query，因此，客户端可以一次性发送多条 SQL 语句，如 `select 1; select 1; select 1;`。此时的执行时间是所有 SQL 语句执行完之后的总时间。
  - Command Per Second：TiDB 按照执行结果成功或失败来统计每秒处理的命令数。
  - QPS：按 SELECT、INSERT、UPDATE 类型统计所有 TiDB 实例上每秒执行的 SQL 语句数量。
  - CPS By Instance：按照命令和执行结果成功或失败来统计每个 TiDB 实例上的命令。
  - Failed Query OPM：每个 TiDB 实例上，对每秒钟执行 SQL 语句发生的错误按照错误类型进行统计（例如语法错误、主键冲突等）。包含了错误所属的模块和错误码。
  - Slow query：慢查询的处理时间（整个慢查询耗时、Coprocessor 耗时、Coprocessor 调度等待时间），慢查询分为 internal 和 general SQL 语句。
  - Connection Idle Duration：空闲连接的持续时间。
  - 999/99/95/80 Duration：不同类型的 SQL 语句执行耗时（不同百分位）。

- Query Detail
  - Duration 80/95/99/999 By Instance：每个 TiDB 实例执行 SQL 语句的耗时（不同百分位）。
  - Failed Query OPM Detail：每个 TiDB 实例执行 SQL 语句发生的错误按照错误类型统计（例如语法错误、主键冲突等）。
  - Internal SQL OPS：整个 TiDB 集群内部 SQL 语句执行的 QPS。内部 SQL 语句是指 TiDB 内部自动执行的 SQL 语句，一般由用户 SQL 语句来触发或者内部定时任务触发。

- Server
  - Uptime：每个 TiDB 实例的运行时间。
  - Memory Usage：每个 TiDB 实例的内存使用，分为进程占用内存和 Golang 在堆上申请的内存。
  - CPU Usage：每个 TiDB 实例的 CPU 使用。
  - Connection Count：每个 TiDB 的连接数。
  - Open FD Count：每个 TiDB 实例的打开的文件描述符数量。
  - Disconnection Count：每个 TiDB 实例断开连接的数量。
  - Event OPM：每个 TiDB 实例关键事件，例如 start，close，graceful-shutdown，kill，hang 等。
  - Goroutine Count：每个 TiDB 实例的 Goroutine 数量。
  - Prepare Statement Count：每个 TiDB 实例现存的 Prepare 语句数以及总数。
  - Keep Alive OPM：每个 TiDB 实例每分钟刷新监控的次数，通常不需要关注。
  - Panic And Critical Error：TiDB 中出现的 Panic、Critical Error 数量。
  - Time Jump Back OPS：每个 TiDB 实例上每秒操作系统时间回跳的次数。
  - Get Token Duration：每个连接获取 Token 的耗时。
  - Skip Binlog Count：TiDB 写入 Binlog 失败的数量。
  - Client Data Traffic：TiDB 和客户端的数据流量。

- Transaction
  - Transaction OPS：每秒事务的执行数量
  - Duration：事务执行时间
  - Transaction Statement Num：事务中的 SQL 语句数量
  - Transaction Retry Num：事务重试次数

- Session Retry Error OPS：事务重试时每秒遇到的错误数量，分为重试失败和超过最大重试次数两种类型
- Commit Token Wait Duration：事务提交时的流控队列等待时间。当出现较长等待时，代表提交事务过大，正在限流。如果系统还有资源可以使用，可以通过增大 TiDB 配置文件中 `committer-concurrency` 值来加速提交
- KV Transaction OPS：每个 TiDB 内部每秒执行的事务数量
  * 一个用户的事务，在 TiDB 内部可能会触发多次事务执行，其中包含，内部元数据的读取，用户事务原子性地多次重试执行等
  * TiDB 内部的定时任务也会通过事务来操作数据库，这部分也包含在这个面板里
- KV Transaction Duration：每个 TiDB 内部执行事务的耗时
- Transaction Regions Num：事务操作的 Region 数量

- Transaction Write KV Num Rate and Sum：事务写入 KV 的速率总和
- Transaction Write KV Num：事务操作的 KV 数量
- Statement Lock Keys：单个语句的加锁个数
- Send HeartBeat Duration：事务发送心跳的时间间隔
- Transaction Write Size Bytes Rate and sum：事务写入字节数的速率总和
- Transaction Write Size Bytes：事务写入的数据大小
- Acquire Pessimistic Locks Duration：加锁所消耗的时间
- TTL Lifetime Reach Counter：事务的 TTL 寿命上限。TTL 上限默认值 1 小时，它的含义是从悲观事务第一次加锁，或者乐观事务的第一个 prewrite 开始，超过了 1 小时。可以通过修改 TiDB 配置文件中 `max-txn-ttl` 来改变 TTL 寿命上限
- Load Safepoint OPS：加载 Safepoint 的次数。Safepoint 作用是在事务读数据时，保证不读到 Safepoint 之前的数据，保证数据安全。因为，Safepoint 之前的数据有可能被 GC 清理掉

- Pessimistic Statement Retry OPS：悲观语句重试次数。当语句尝试加锁时，可能遇到写入冲突，此时，语句会重新获取新的 snapshot 并再次加锁
- Transaction Types Per Seconds：每秒采用两阶段提交 (2PC)、异步提交（Async Commit) 和一阶段提交 (1PC) 机制的事务数量，提供成功和失败两种数量

- Executor

  - Parse Duration：SQL 语句解析耗时统计。
  - Compile Duration：将解析后的 SQL AST 编译成执行计划的耗时。
  - Execution Duration：执行 SQL 语句执行计划耗时。
  - Expensive Executor OPS：每秒消耗系统资源比较多的算子。包括 Merge Join、Hash Join、Index Look Up Join、Hash Agg、 Stream Agg、 Sort、 TopN 等。
  - Queries Using Plan Cache OPS：每秒使用 Plan Cache 的查询数量。

- Distsql

  - Distsql Duration： Distsql 处理的时长
  - Distsql QPS：每秒 Distsql 的数量
  - Distsql Partial QPS：每秒 Partial Results 的数量
  - Scan Keys Num：每个 Query 扫描的 Key 的数量
  - Scan Keys Partial Num：每一个 Partial Result 扫描的 Key 的数量
  - Partial Num：每个 SQL 语句 Partial Results 的数量

- KV Errors
  - KV Backoff Duration：KV 每个请求重试的总时间。TiDB 向 TiKV 的请求都有重试机制，这里统计的是向 TiKV 发送请求时遇到错误重试的总时间
  - TiClient Region Error OPS：TiKV 返回 Region 相关错误信息的数量
  - KV Backoff OPS：TiKV 返回错误信息的数量
  - Lock Resolve OPS：TiDB 清理锁操作的数量。当 TiDB 的读写请求遇到锁时，会尝试进行锁清理
  - Other Errors OPS：其他类型的错误数量，包括清锁和更新 SafePoint
- KV Request
  - KV Request OPS：KV Request 根据 TiKV 显示执行次数
  - KV Request Duration 99 by store：根据 TiKV 显示 KV Request 执行时间
  - KV Request Duration 99 by type：根据类型显示 KV Request 的执行时间
- PD Client
  - PD Client CMD OPS：PD Client 每秒执行命令的数量
  - PD Client CMD Duration：PD Client 执行命令耗时
  - PD Client CMD Fail OPS：PD Client 每秒执行命令失败的数量
  - PD TSO OPS：TiDB 每秒从 PD 获取 TSO 的数量
  - PD TSO Wait Duration：TiDB 等待从 PD 返回 TSO 的时间
  - PD TSO RPC Duration：TiDB 从向 PD 发送获取 TSO 的请求到接收到 TSO 的耗时
  - Start TSO Wait Duration：TiDB 从向 PD 发送获取 start tso 请求开始到开始等待 tso 返回的时间
- Schema Load
  - Load Schema Duration：TiDB 从 TiKV 获取 Schema 的时间
  - Load Schema OPS：TiDB 从 TiKV 每秒获取 Schema 的数量
  - Schema Lease Error OPM：Schema Lease 出错统计，包括 change 和 outdate 两种，change 代表 schema 发生了变化，outdate 代表无法更新 schema，属于较严重错误，出现 outdate 错误时会报警
  - Load Privilege OPS：TiDB 从 TiKV 每秒获取权限信息的数量
- DDL
  - DDL Duration 95：DDL 语句处理时间的 95% 分位
  - Batch Add Index Duration 100：创建索引时每个 Batch 所花费的最大时间
  - DDL Waiting Jobs Count：等待的 DDL 任务数量
  - DDL META OPM：DDL 每分钟获取 META 的次数
  - DDL Worker Duration 99：每个 DDL worker 执行时间的 99% 分位
  - Deploy Syncer Duration：Schema Version Syncer 初始化，重启，清空等操作耗时
  - Owner Handle Syncer Duration：DDL Owner 在执行更新，获取以及检查 Schema Version 的耗时
  - Update Self Version Duration：Schema Version Syncer 更新版本信息耗时
  - DDL OPM：DDL 语句的每秒执行次数
  - DDL add index progress in percentage：添加索引的进度展示
- Statistics
  - Auto Analyze Duration 95：自动 ANALYZE 耗时
  - Auto Analyze QPS：自动 ANALYZE 数量
  - Stats Inaccuracy Rate：统计信息不准确度
  - Pseudo Estimation OPS：使用假的统计信息优化 SQL 的数量

  - Dump Feedback OPS：存储统计信息 Feedback 的数量
  - Store Query Feedback QPS：存储合并查询的 Feedback 信息的每秒操作数量，该操作在 TiDB 内存中进行
  - Significant Feedback：重要的 Feedback 更新统计信息的数量
  - Update Stats OPS：利用 Feedback 更新统计信息的数量
  - Fast Analyze Status 100：快速收集统计信息的状态

- Owner
  - New ETCD Session Duration 95：创建一个新的 etcd 会话花费的时间。TiDB 通过 etcd client 连接 PD 中的 etcd 保存/读取部分元数据信息。这里记录了创建会话花费的时间
  - Owner Watcher OPS：DDL owner watch PD 的 etcd 的元数据的 goroutine 的每秒操作次数

- Meta
  - AutoID QPS：AutoID 相关操作的数量统计，包括全局 ID 分配、单个 Table AutoID 分配、单个 Table AutoID Rebase 三种操作
  - AutoID Duration：AutoID 相关操作的耗时
  - Region Cache Error OPS：TiDB 缓存的 region 信息每秒遇到的错误次数
  - Meta Operations Duration 99：元数据操作延迟

- GC
  - Worker Action OPM：GC 相关操作的数量，包括 run_job，resolve_lock，delete_range 等操作
  - Duration 99：GC 相关操作的耗时
  - Config：GC 的数据保存时长 (life time) 和 GC 运行间隔 (run interval) 配置
  - GC Failure OPM：GC 相关操作失败的数量
  - Delete Range Failure OPM：Delete range 失败的次数
  - Too Many Locks Error OPM：GC 清锁过多错误的数量
  - Action Result OPM：GC 相关操作结果数量
  - Delete Range Task Status：Delete range 的任务状态，包含完成和失败状态
  - Push Task Duration 95：将 GC 子任务推送给 GC worker 的耗时

- Batch Client
  - Pending Request Count by TiKV：TiKV 批量消息处理的等待数量
  - Wait Duration 95: 批量消息处理的等待时间。
  - Batch Client Unavailable Duration 95：批处理客户端的不可用时长。
  - No Available Connection Counter：批处理客户端不可用的连接数。

9.8.3 PD 重要监控指标详解

使用 TiUP 部署 TiDB 集群时，一键部署监控系统 (Prometheus & Grafana)，监控架构参见TiDB 监控框架概述。

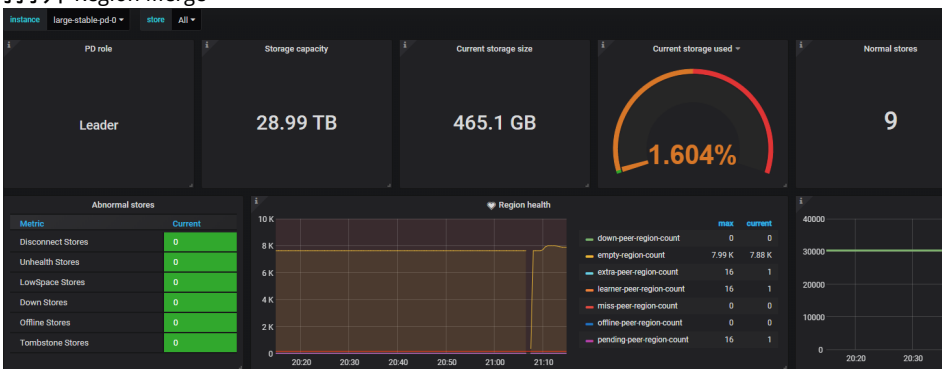
目前 Grafana Dashboard 整体分为 PD、 TiDB、 TiKV、 Node_exporter、 Overview 等。

对于日常运维，我们通过观察 PD 面板上的 Metrics，可以了解 PD 当前的状态。

以下为 PD Dashboard 监控说明：

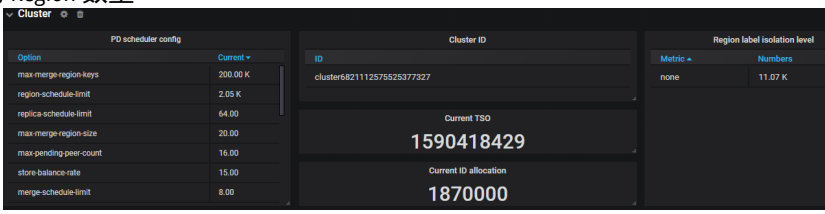
- PD role：当前 PD 的角色
- Storage capacity：TiDB 集群总可用数据库空间大小

- Current storage size：TiDB 集群目前已用数据库空间大小
- Current storage usage：TiDB 集群存储空间的使用率
- Normal stores：处于正常状态的节点数目
- Number of Regions：当前集群的 Region 总量
- Abnormal stores：处于异常状态的节点数目，正常情况应当为 0
- Region health：集群所有 Region 的状态。通常情况下，pending 或 down 的 peer 应该少于 100，miss 的 peer 不能一直大于 0，empty Region 过多需及时打开 Region Merge
- Current peer count：当前集群 peer 的总量



9.8.3.1 Cluster

- PD scheduler config：PD 调度配置列表
- Cluster ID：集群的 cluster id，唯一标识
- Current TSO：当前分配 TSO 的物理时间戳部分
- Current ID allocation：当前可分配 ID 的最大值
- Region label isolation level：不同 label 所在的 level 的 Region 数量
- Label distribution：集群中 TiKV 节点的 label 分布情况



9.8.3.2 Operator

- Schedule operator create：新创建的不同 operator 的数量，单位 opm 代表一分钟内创建的个数
- Schedule operator check：已检查的 operator 的次数，主要检查是否当前步骤已经执行完成，如果是，则执行下一个步骤
- Schedule operator finish：已完成调度的 operator 的数量
- Schedule operator timeout：已超时的 operator 的数量
- Schedule operator replaced or canceled：已取消或者被替换的 operator 的数量
- Schedule operators count by state：不同状态的 operator 的数量
- Operator finish duration：已完成的 operator 所花费的最长时间
- Operator step duration：已完成的 operator 的步骤所花费的最长时间

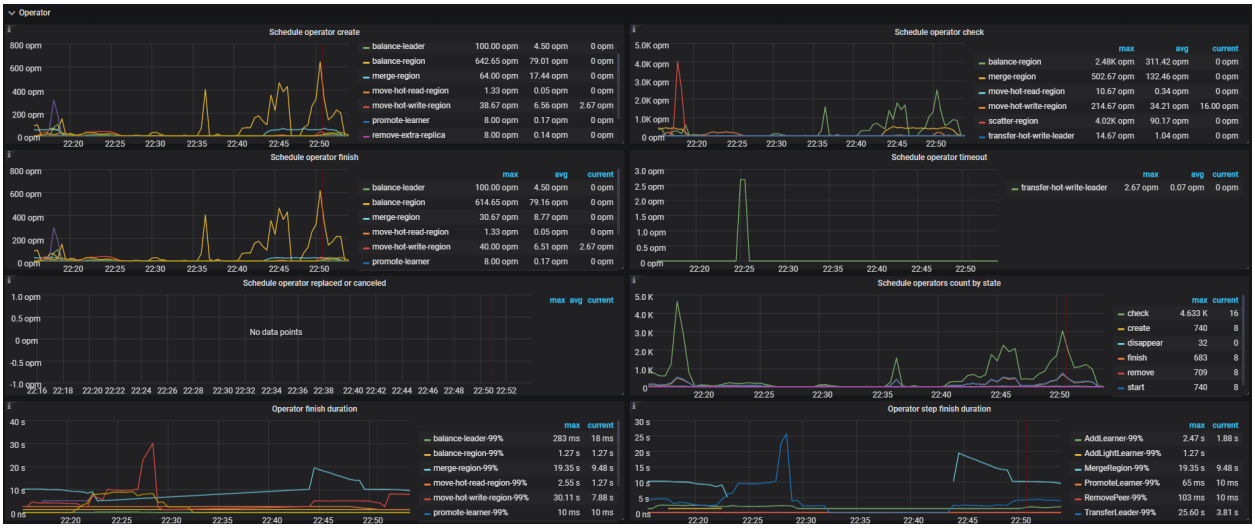


图 50: PD Dashboard - Operator metrics

9.8.3.3 Statistics - Balance

- Store capacity：每个 TiKV 实例的总的空间大小
- Store available：每个 TiKV 实例的可用空间大小
- Store used：每个 TiKV 实例的已使用空间大小
- Size amplification：每个 TiKV 实例的空间放大比率
- Size available ratio：每个 TiKV 实例的可用空间比率
- Store leader score：每个 TiKV 实例的 leader 分数
- Store Region score：每个 TiKV 实例的 Region 分数
- Store leader size：每个 TiKV 实例上所有 leader 的大小
- Store Region size：每个 TiKV 实例上所有 Region 的大小
- Store leader count：每个 TiKV 实例上所有 leader 的数量
- Store Region count：每个 TiKV 实例上所有 Region 的数量



图 51: PD Dashboard - Balance metrics

9.8.3.4 Statistics - hot write

- Hot Region’s leader distribution：每个 TiKV 实例上成为写入热点的 leader 的数量
- Total written bytes on hot leader Regions：每个 TiKV 实例上所有成为写入热点的 leader 的总的写入流量大小
- Hot write Region’s peer distribution：每个 TiKV 实例上成为写入热点的 peer 的数量
- Total written bytes on hot peer Regions：每个 TiKV 实例上所有成为写入热点的 peer 的写入流量大小
- Store Write rate bytes：每个 TiKV 实例总的写入的流量
- Store Write rate keys：每个 TiKV 实例总的写入 keys
- Hot cache write entry number：每个 TiKV 实例进入热点统计模块的 peer 的数量
- Selector events：热点调度中选择器的事件发生次数
- Direction of hotspot move leader：热点调度中 leader 的调度方向，正数代表调入，负数代表调出
- Direction of hotspot move peer：热点调度中 peer 的调度方向，正数代表调入，负数代表调出



图 52: PD Dashboard - Hot write metrics

9.8.3.5 Statistics - hot read

- Hot Region’s peer distribution：每个 TiKV 实例上成为读取热点的 peer 的数量
- Total read bytes on hot peer Regions：每个 TiKV 实例上所有成为读取热点的 peer 的总的读取流量大小
- Store read rate bytes：每个 TiKV 实例总的读取的流量
- Store read rate keys：每个 TiKV 实例总的读取 keys
- Hot cache read entry number：每个 TiKV 实例进入热点统计模块的 peer 的数量



图 53: PD Dashboard - Hot read metrics

9.8.3.6 Scheduler

- Scheduler is running：所有正在运行的 scheduler
- Balance leader movement：leader 移动的详细情况
- Balance Region movement：Region 移动的详细情况
- Balance leader event：balance leader 的事件数量
- Balance Region event：balance Region 的事件数量
- Balance leader scheduler：balance-leader scheduler 的状态
- Balance Region scheduler：balance-region scheduler 的状态
- Replica checker：replica checker 的状态
- Rule checker：rule checker 的状态
- Region merge checker：merge checker 的状态
- Filter target：尝试选择 Store 作为调度 taget 时没有通过 Filter 的计数
- Filter source：尝试选择 Store 作为调度 source 时没有通过 Filter 的计数
- Balance Direction：Store 被选作调度 target 或 source 的次数
- Store Limit：Store 的调度限流状态

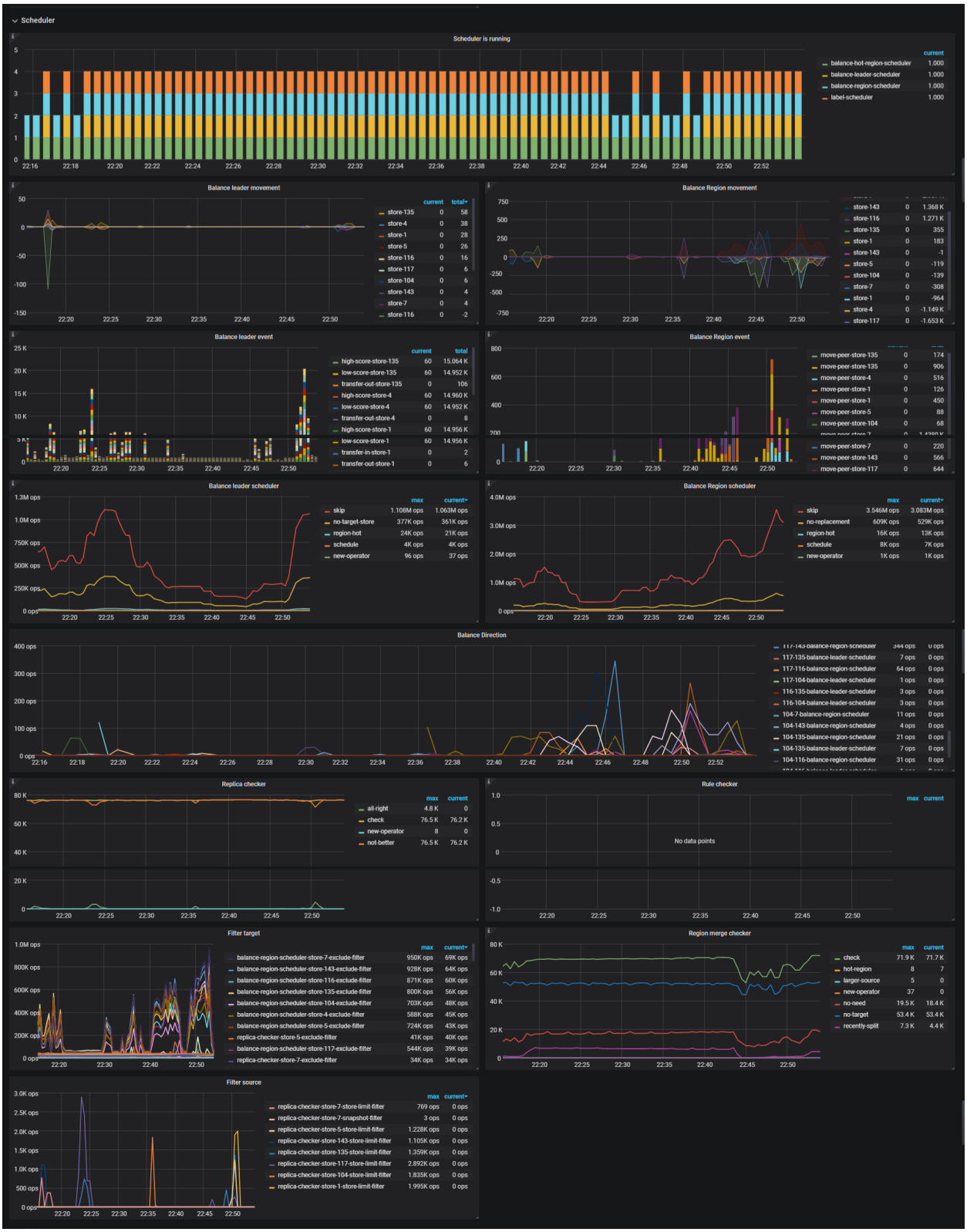


图 54: PD Dashboard - Scheduler metrics

9.8.3.7 gRPC

- Completed commands rate：gRPC 命令的完成速率
- 99% Completed commands duration：99% 命令的最长消耗时间



图 55: PD Dashboard - gRPC metrics

9.8.3.8 etcd

- Handle transactions count：etcd 的事务个数
- 99% Handle transactions duration：99% 的情况下，处理 etcd 事务所需花费的时间
- 99% WAL fsync duration：99% 的情况下，持久化 WAL 所需花费的时间，这个值通常应该小于 1s
- 99% Peer round trip time seconds：99% 的情况下，etcd 的网络延时，这个值通常应该小于 1s
- etcd disk WAL fsync rate：etcd 持久化 WAL 的速率
- Raft term：当前 Raft 的 term
- Raft committed index：最后一次 commit 的 Raft index
- Raft applied index：最后一次 apply 的 Raft index



图 56: PD Dashboard - etcd metrics

9.8.3.9 TiDB

- PD Server TSO handle time and Client recv time：从 PD 开始处理 TSO 请求到 client 端接收到 TSO 的总耗时
- Handle requests count：TiDB 的请求数量
- Handle requests duration：每个请求所花费的时间，99% 的情况下，应该小于 100ms

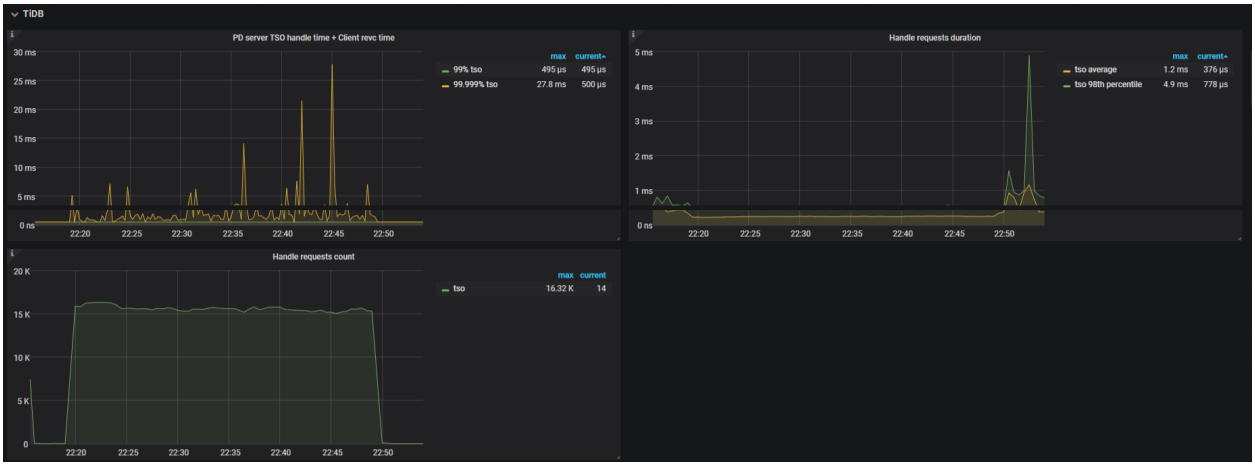


图 57: PD Dashboard - TiDB metrics

9.8.3.10 Heartbeat

- Heartbeat region event QPS：心跳处理 region 的 QPS，包括更新缓存和持久化
- Region heartbeat report：TiKV 向 PD 发送的心跳个数
- Region heartbeat report error：TiKV 向 PD 发送的异常的心跳个数
- Region heartbeat report active：TiKV 向 PD 发送的正常的心跳个数
- Region schedule push：PD 向 TiKV 发送的调度命令的个数
- 99% Region heartbeat latency：99% 的情况下，心跳的延迟

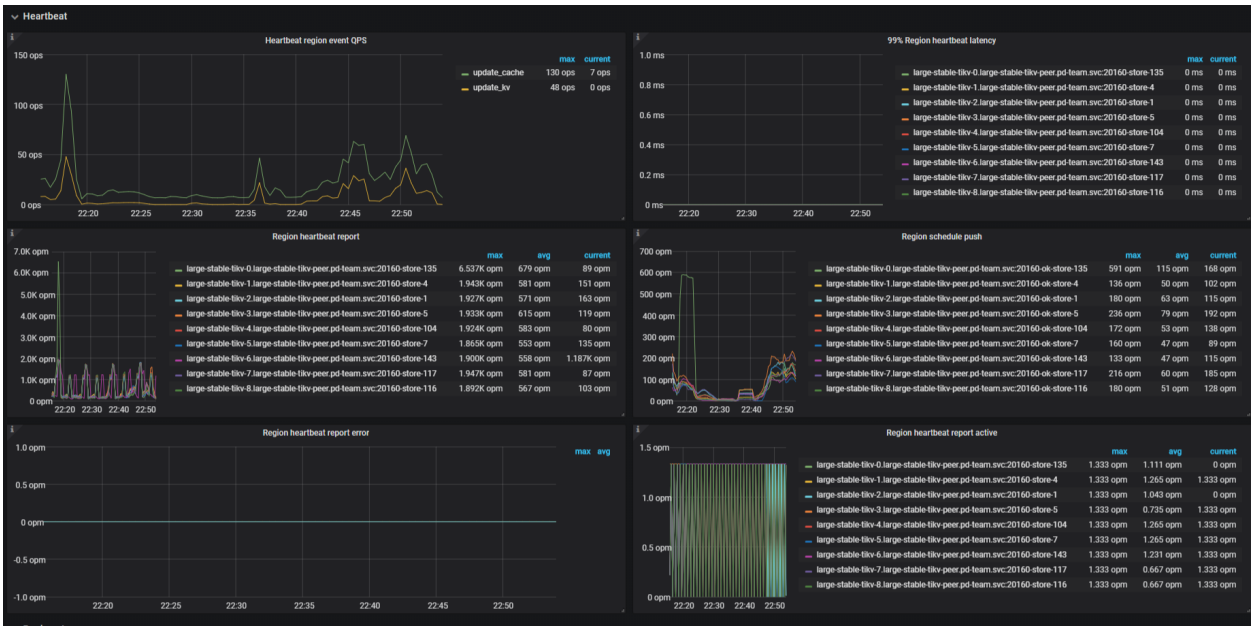


图 58: PD Dashboard - Heartbeat metrics

9.8.3.11 Region storage

- Syncer Index：Leader 记录 Region 变更历史的最大 index
- history last index：Follower 成功同步的 Region 变更历史的 index



图 59: PD Dashboard - Region storage

9.8.4 TiKV 监控指标详解

使用 TiUP 部署 TiDB 集群时，一键部署监控系统 (Prometheus & Grafana)，监控架构参见TiDB 监控框架概述。

目前 Grafana Dashboard 整体分为 PD、 TiDB、 TiKV、 Node_exporter、 Overview 等。

对于日常运维，我们通过观察 TiKV-Details 面板上的指标，可以了解 TiKV 当前的状态。根据性能地图可以检查集群的状态是否符合预期。

以下为 TiKV-Details 默认的监控信息：

9.8.4.1 Cluster

- Store size：每个 TiKV 实例的使用的存储空间的大小
- Available size：每个 TiKV 实例的可用的存储空间的大小
- Capacity size：每个 TiKV 实例的存储容量的大小
- CPU：每个 TiKV 实例 CPU 的使用率
- Memory：每个 TiKV 实例内存的使用情况
- IO utilization：每个 TiKV 实例 IO 的使用率
- MBps：每个 TiKV 实例写入和读取的数据量大小
- QPS：每个 TiKV 实例上各种命令的 QPS
- Errps：每个 TiKV 实例上 gRPC 消息失败的速率
- leader：每个 TiKV 实例 leader 的个数
- Region：每个 TiKV 实例 Region 的个数
- Uptime：自上次重启以来 TiKV 正常运行的时间

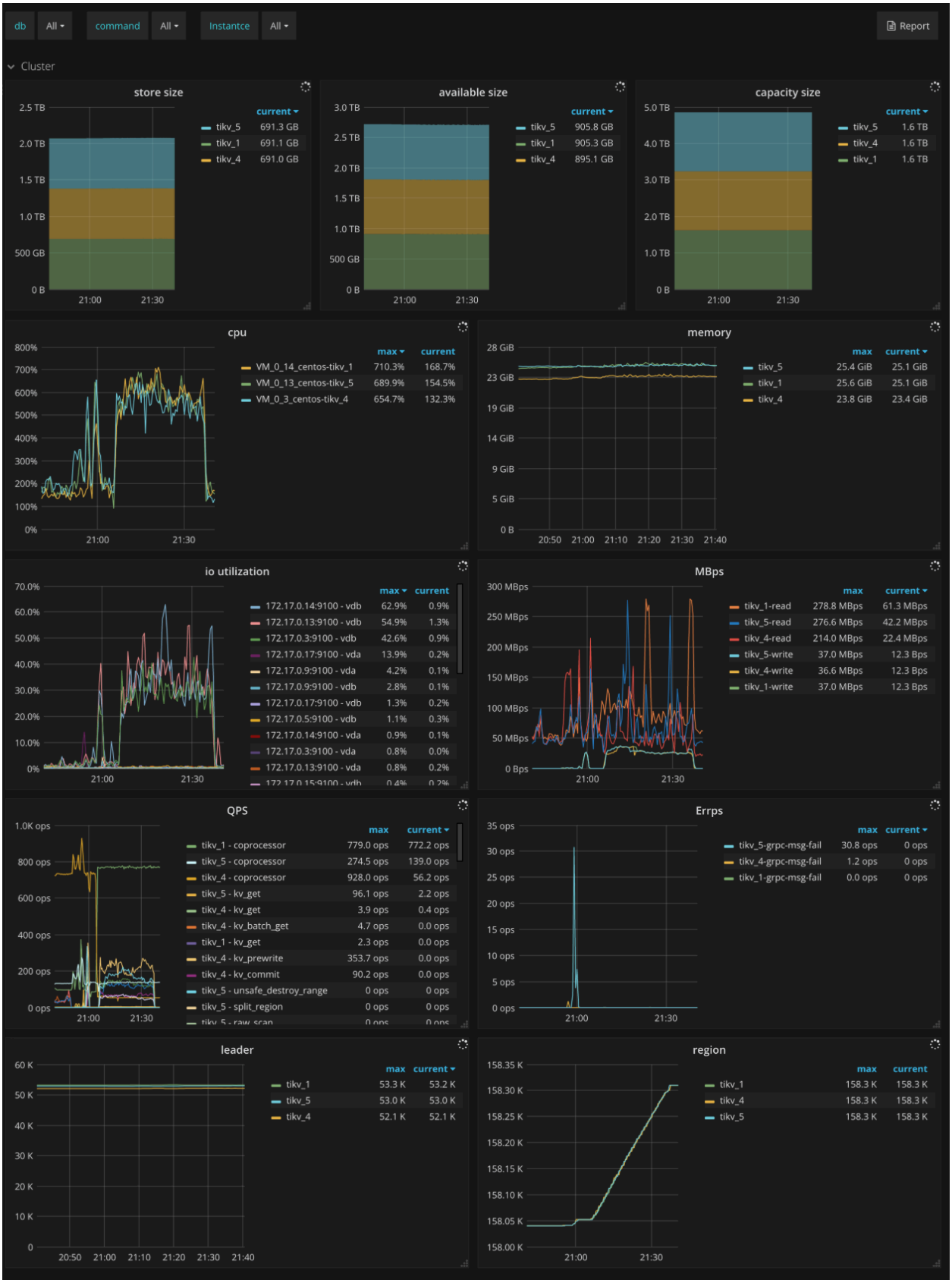


图 60: TiKV Dashboard - Cluster metrics

9.8.4.2 Errors

- Critical error：严重错误的数量
- Server is busy：各种会导致 TiKV 实例暂时不可用的事件个数，如 write stall，channel full 等，正常情况下应当为 0
- Server report failures：server 报错的消息个数，正常情况下应当为 0
- Raftstore error：每个 TiKV 实例上 raftstore 发生错误的个数
- Scheduler error：每个 TiKV 实例上 scheduler 发生错误的个数
- Coprocessor error：每个 TiKV 实例上 coprocessor 发生错误的个数
- gRPC message error：每个 TiKV 实例上 gRPC 消息发生错误的个数
- Leader drop：每个 TiKV 实例上 drop leader 的个数
- Leader missing：每个 TiKV 实例上 missing leader 的个数



图 61: TiKV Dashboard - Errors metrics

9.8.4.3 Server

- CF size：每个列族的大小
- Store size：每个 TiKV 实例的使用的存储空间的大小
- Channel full：每个 TiKV 实例上 channel full 错误的数量，正常情况下应当为 0
- Active written leaders：各个 TiKV 实例中正在被写入的 Leader 的数量
- Approximate Region size：每个 Region 近似的大小
- Approximate Region size Histogram：每个 Region 近似大小的直方图
- Region average written keys：每个 TiKV 实例上所有 Region 的平均 key 写入个数
- Region average written bytes：每个 TiKV 实例上所有 Region 的平均写入大小

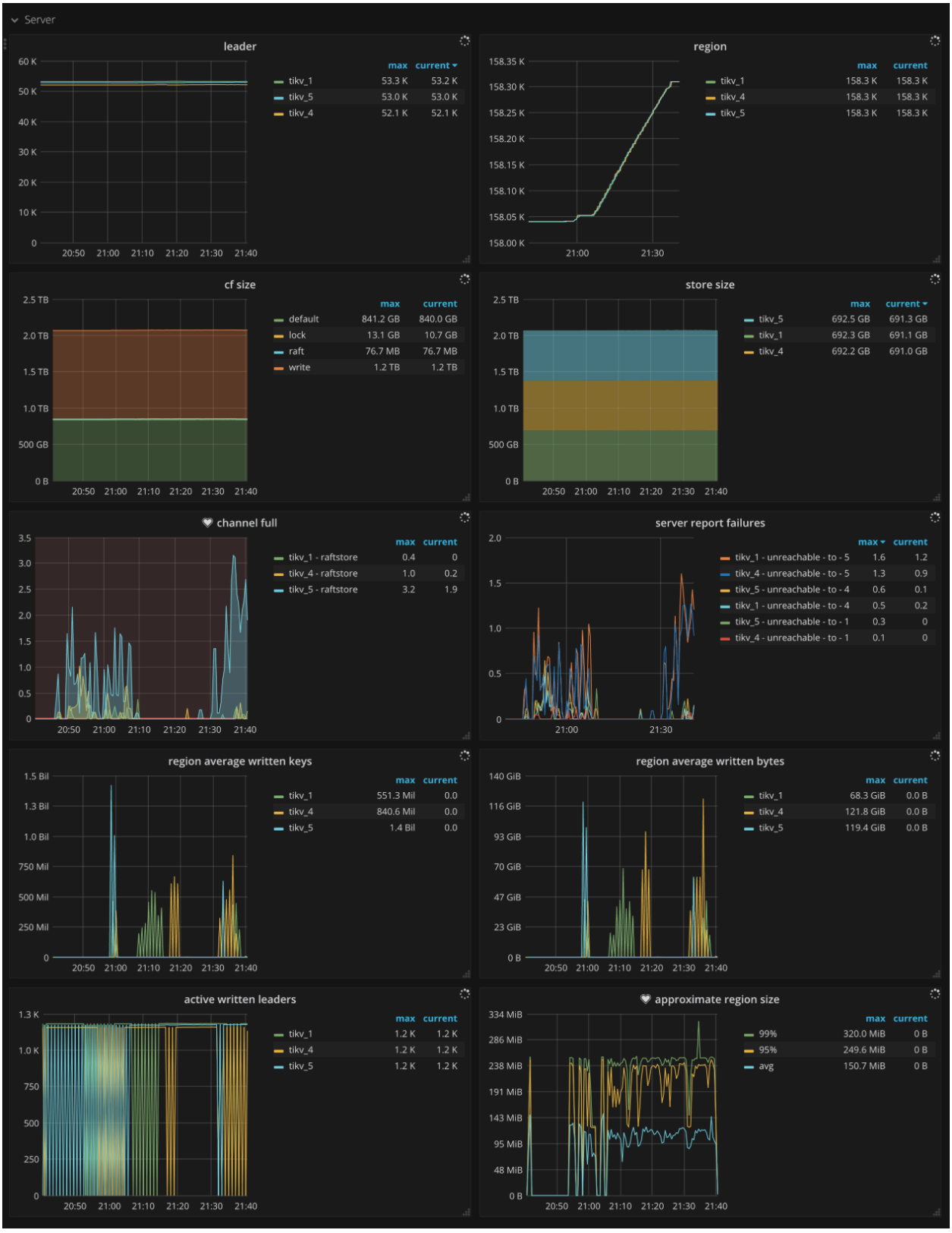


图 62: TiKV Dashboard - Server metrics

9.8.4.4 gRPC

- gRPC message count：每种 gRPC 请求的速度
- gRPC message failed：失败的 gRPC 请求的速度
- 99% gRPC message duration：99% gRPC 请求的执行时间小于该值
- Average gRPC message duration：gRPC 请求平均的执行时间
- gRPC batch size：TiDB 与 TiKV 之间 grpc 请求的 batch 大小
- raft message batch size：TiKV 与 TiKV 之间 raft 消息的 batch 大小

9.8.4.5 Thread CPU

- Raft store CPU：raftstore 线程的 CPU 使用率，通常应低于 80% * `raftstore.store-pool-size`
- Async apply CPU：async apply 线程的 CPU 使用率，通常应低于 90% * `raftstore.apply-pool-size`
- Scheduler worker CPU：scheduler worker 线程的 CPU 使用率，通常应低于 90% * `storage.scheduler-worker-pool-size`
- gRPC poll CPU：gRPC 线程的 CPU 使用率，通常应低于 80% * `server.grpc-concurrency`
- Unified read pool CPU：unified read pool 线程的 CPU 使用率
- Storage ReadPool CPU：storage read pool 线程的 CPU 使用率
- Coprocessor CPU：coprocessor 线程的 CPU 使用率
- RocksDB CPU：RocksDB 线程的 CPU 使用率
- GC worker CPU：GC worker 线程的 CPU 使用率
- BackGround worker CPU：background worker 线程的 CPU 使用率

9.8.4.6 PD

- PD requests：TiKV 发送给 PD 的请求速度
- PD request duration (average)：TiKV 发送给 PD 的请求处理的平均时间
- PD heartbeats：发送给 PD 的心跳的速度
- PD validate peers：TiKV 发送给 PD 用于验证 TiKV 的 peer 有效的消息的速度

9.8.4.7 Raft IO

- Apply log duration：Raft apply 日志所花费的时间
- Apply log duration per server：每个 TiKV 实例上 Raft apply 日志所花费的时间
- Append log duration：Raft append 日志所花费的时间
- Append log duration per server：每个 TiKV 实例上 Raft append 日志所花费的时间
- Commit log duration：Raft commit 日志所花费的时间
- Commit log duration per server：每个 TiKV 实例上 Raft commit 日志所花费的时间

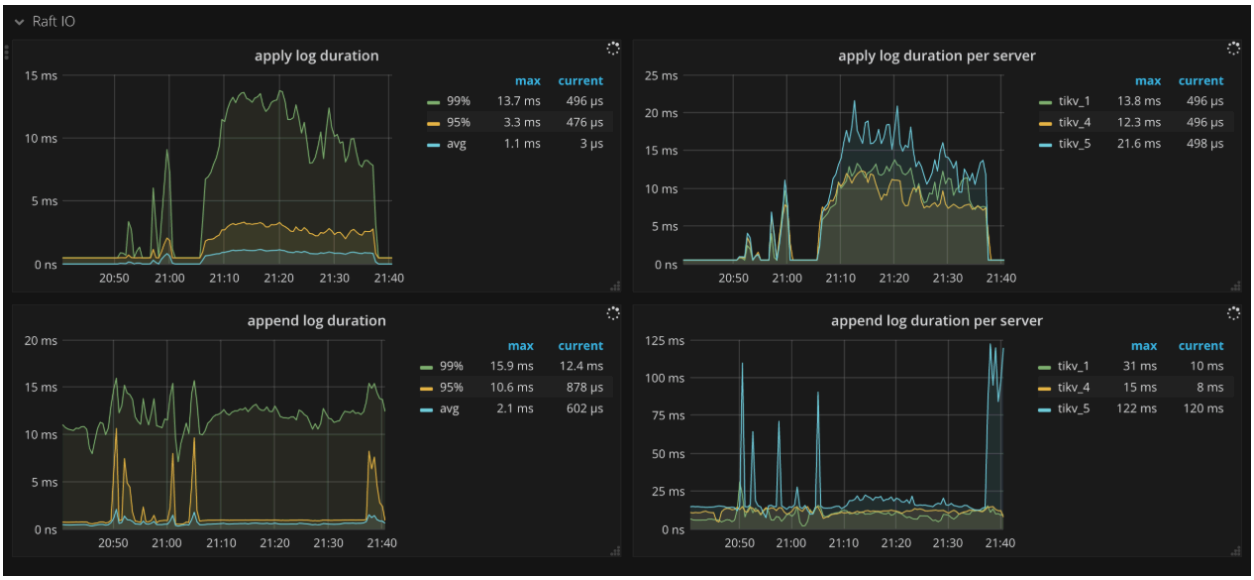


图 63: TiKV Dashboard - Raft IO metrics

9.8.4.8 Raft process

- Ready handled：Raft 中不同 ready 类型的 ops
  - count: 批量处理 ready 的 ops
  - has_ready_region: 获得 ready 的 Region 的 ops
  - pending_region: 被检查是否获得 ready 的 Region 的 ops，v3.0.0 后废弃
  - message: ready 内待发送 message 的 ops
  - append: ready 内 Raft log entry 的 ops
  - commit: ready 内 committed Raft log entry 的 ops
  - snapshot: 携带 snapshot 的 ready 的 ops
- 0.99 Duration of Raft store events：99% 的 raftstore 事件所花费的时间
- Process ready duration：处理 ready 所花费的时间
- Process ready duration per server：每个 TiKV 实例处理 ready 所花费的时间，99.99% 的情况下，应该小于 2s



图 64: TiKV Dashboard - Raft process metrics

9.8.4.9 Raft message

- Sent messages per server：每个 TiKV 实例发送 Raft 消息的 ops
- Flush messages per server：每个 TiKV 实例中 raft client 往外 flush Raft 消息的 ops
- Receive messages per server：每个 TiKV 实例接受 Raft 消息的 ops
- Messages：发送不同类型的 Raft 消息的 ops
- Vote：Raft 投票消息发送的 ops
- Raft dropped messages：每秒钟丢弃不同类型的 Raft 消息的个数

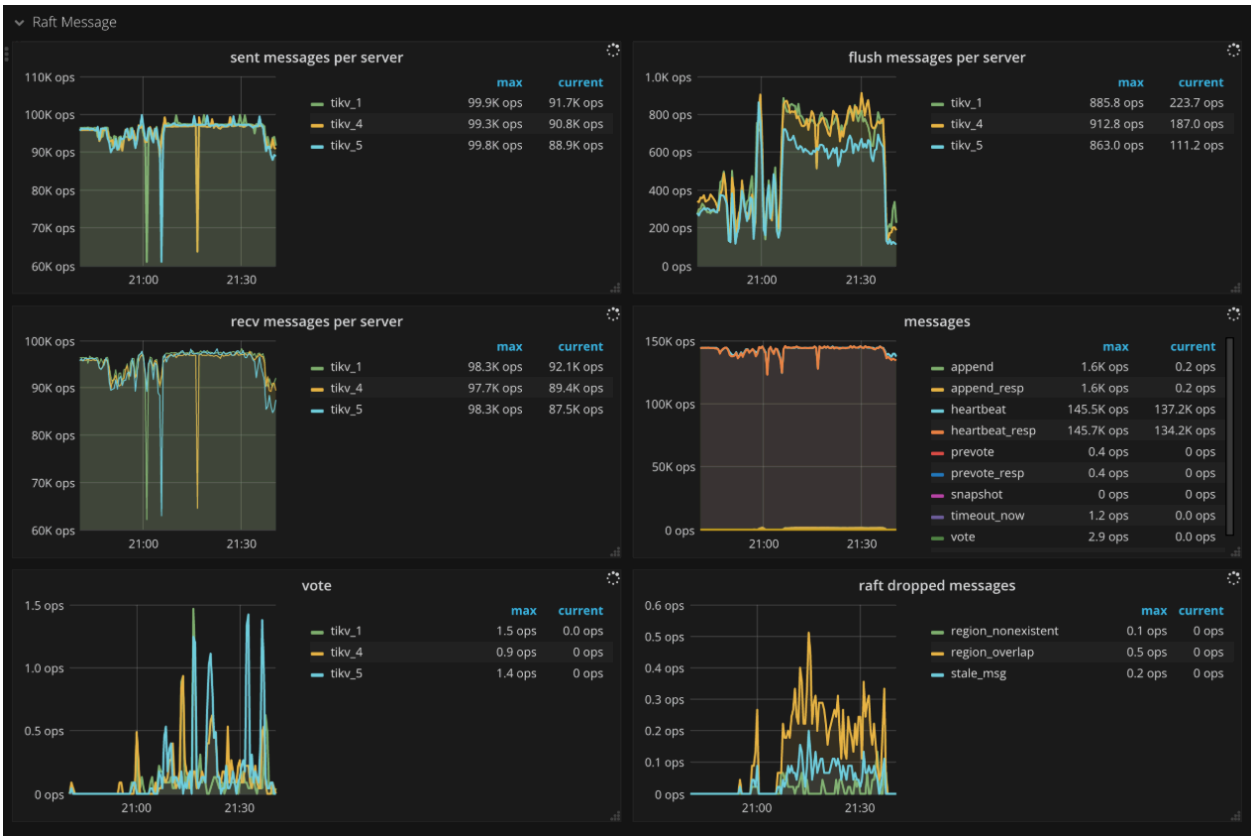


图 65: TiKV Dashboard - Raft message metrics

9.8.4.10 Raft propose

- Raft apply proposals per ready：在一个 batch 内，apply proposal 时每个 ready 中包含 proposal 的个数的直方图
- Raft read/write proposals：不同类型的 proposal 的 ops
- Raft read proposals per server：每个 TiKV 实例发起读 proposal 的 ops
- Raft write proposals per server：每个 TiKV 实例发起写 proposal 的 ops
- Propose wait duration：proposal 的等待时间的直方图
- Propose wait duration per server：每个 TiKV 实例上每个 proposal 的等待时间的直方图
- Apply wait duration：apply 的等待时间的直方图
- Apply wait duration per server：每个 TiKV 实例上每个 apply 的等待时间的直方图
- Raft log speed：peer propose 日志的平均速度

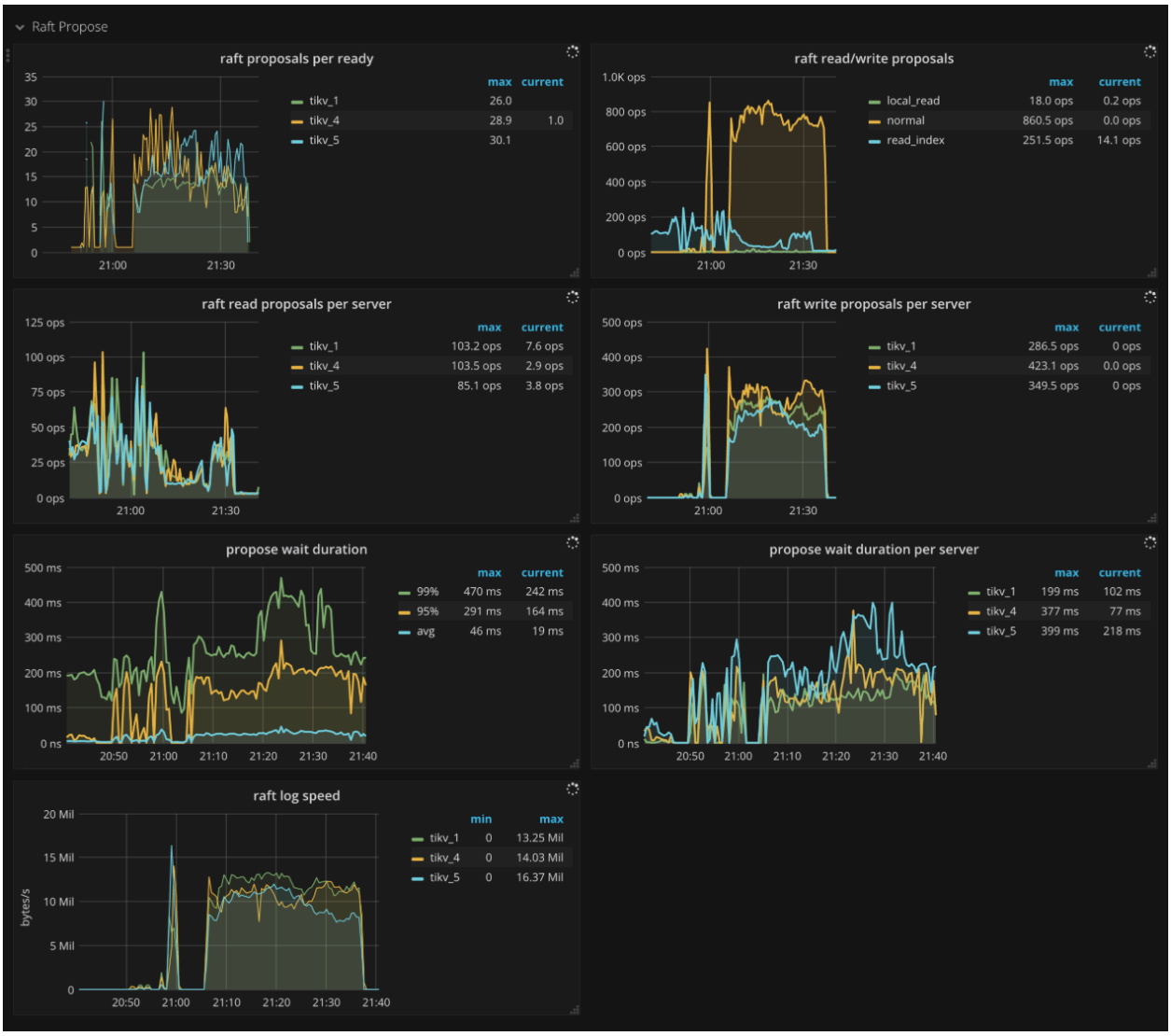


图 66: TiKV Dashboard - Raft propose metrics

9.8.4.11 Raft admin

- Admin proposals：admin proposal 的 ops
- Admin apply：apply 命令的 ops
- Check split：split check 命令的 ops
- 99.99% Check split duration：99.99% 的情况下，split check 所需花费的时间

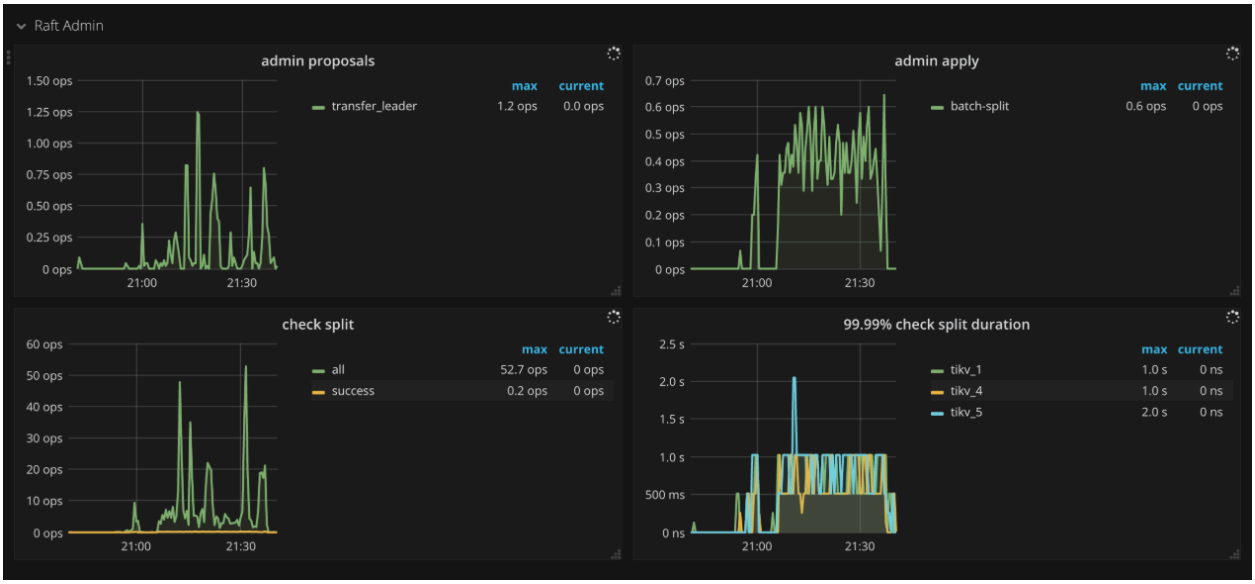


图 67: TiKV Dashboard - Raft admin metrics

9.8.4.12 Local reader

- Local reader requests：所有请求的总数以及 local read 线程拒绝的请求数量



图 68: TiKV Dashboard - Local reader metrics

9.8.4.13 Unified Read Pool

- Time used by level：在 unified read pool 中每个级别使用的时间，级别 0 指小查询

- Level 0 chance：在 unified read pool 中调度的 level 0 任务的比例
- Running tasks：在 unified read pool 中并发运行的任务数量

9.8.4.14 Storage

- Storage command total：收到不同命令的 ops
- Storage async request error：异步请求出错的 ops
- Storage async snapshot duration：异步处理 snapshot 所花费的时间，99% 的情况下，应该小于 1s
- Storage async write duration：异步写所花费的时间，99% 的情况下，应该小于 1s

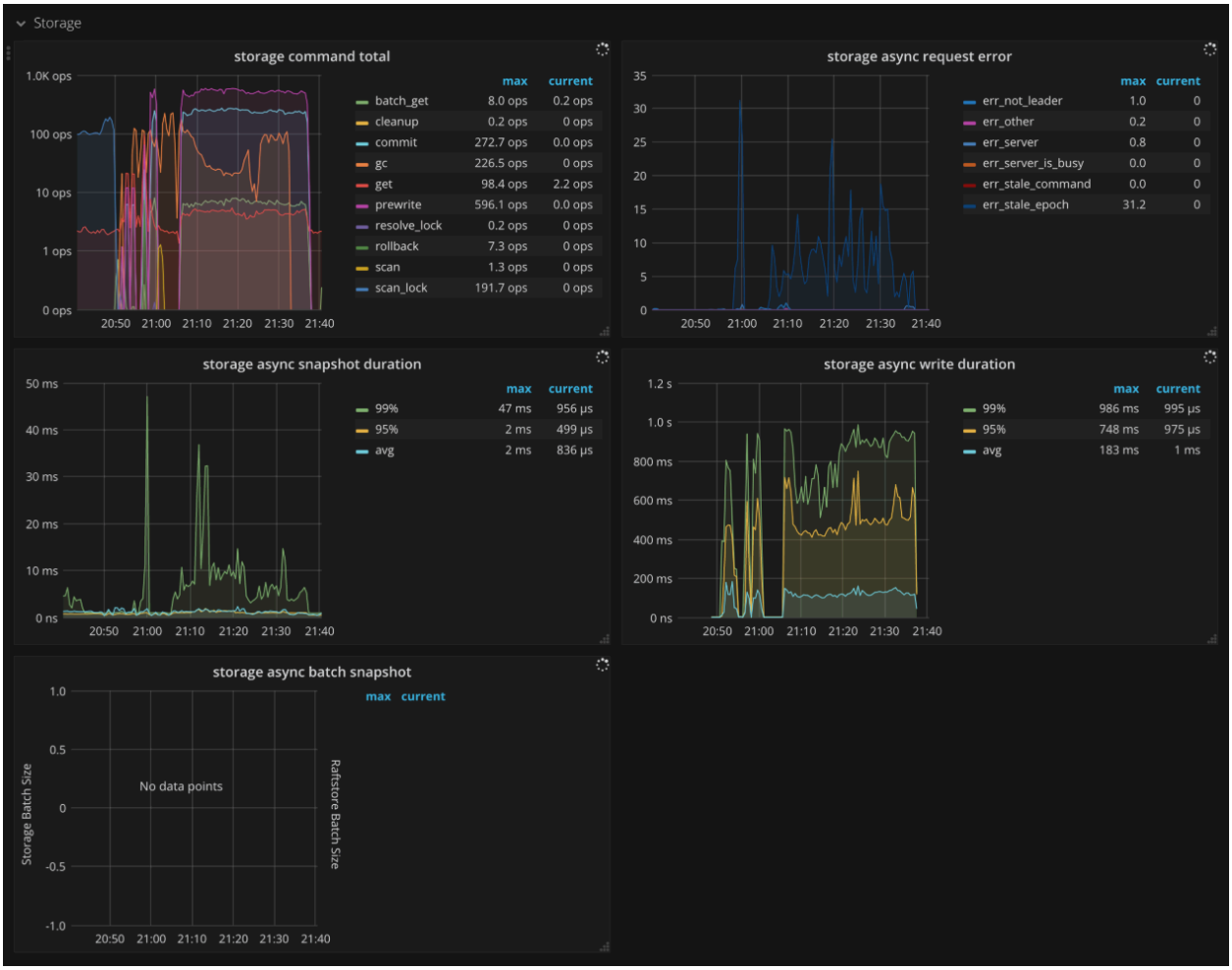


图 69: TiKV Dashboard - Storage metrics

9.8.4.15 Scheduler

- Scheduler stage total：每种命令不同阶段的 ops，正常情况下，不会在短时间内出现大量的错误
- Scheduler writing bytes：每个 TiKV 实例正在处理的命令的写入字节数量
- Scheduler priority commands：不同优先级命令的 ops

- Scheduler pending commands：每个 TiKV 实例上 pending 命令的 ops

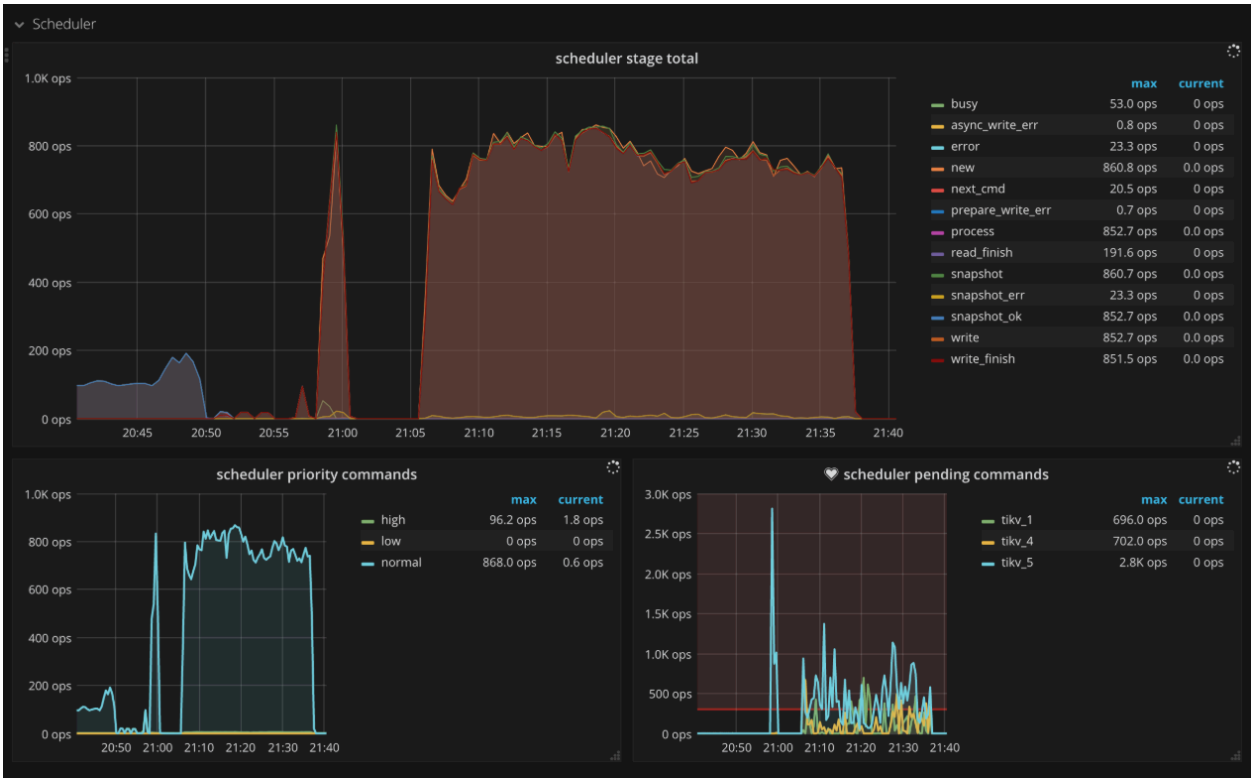


图 70: TiKV Dashboard - Scheduler metrics

9.8.4.16 Scheduler - commit

- Scheduler stage total：commit 中每个命令所处不同阶段的 ops，正常情况下，不会在短时间内出现大量的错误
- Scheduler command duration：执行 commit 命令所需花费的时间，正常情况下，应该小于 1s
- Scheduler latch wait duration：由于 latch wait 造成的时间开销，正常情况下，应该小于 1s
- Scheduler keys read：commit 命令读取 key 的个数
- Scheduler keys written：commit 命令写入 key 的个数
- Scheduler scan details：执行 commit 命令时，扫描每个 CF 中 key 的详细情况
- Scheduler scan details [lock]：执行 commit 命令时，扫描每个 lock CF 中 key 的详细情况
- Scheduler scan details [write]：执行 commit 命令时，扫描每个 write CF 中 key 的详细情况
- Scheduler scan details [default]：执行 commit 命令时，扫描每个 default CF 中 key 的详细情况



图 71: TiKV Dashboard - Scheduler commit metrics

9.8.4.17 Scheduler - pessimistic_rollback

- Scheduler stage total：pessimistic_rollback 中每个命令所处不同阶段的 ops，正常情况下，不会在短时间内出现大量的错误
- Scheduler command duration：执行 pessimistic_rollback 命令所需花费的时间，正常情况下，应该小于 1s
- Scheduler latch wait duration：由于 latch wait 造成的时间开销，正常情况下，应该小于 1s
- Scheduler keys read：pessimistic_rollback 命令读取 key 的个数
- Scheduler keys written：pessimistic_rollback 命令写入 key 的个数

- Scheduler scan details：执行 pessimistic_rollback 命令时，扫描每个 CF 中 key 的详细情况
- Scheduler scan details [lock]：执行 pessimistic_rollback 命令时，扫描每个 lock CF 中 key 的详细情况
- Scheduler scan details [write]：执行 pessimistic_rollback 命令时，扫描每个 write CF 中 key 的详细情况
- Scheduler scan details [default]：执行 pessimistic_rollback 命令时，扫描每个 default CF 中 key 的详细情况

9.8.4.18 Scheduler - prewrite

- Scheduler stage total：prewrite 中每个命令所处不同阶段的 ops，正常情况下，不会在短时间内出现大量的错误
- Scheduler command duration：执行 prewrite 命令所需花费的时间，正常情况下，应该小于 1s
- Scheduler latch wait duration：由于 latch wait 造成的时间开销，正常情况下，应该小于 1s
- Scheduler keys read：prewrite 命令读取 key 的个数
- Scheduler keys written：prewrite 命令写入 key 的个数
- Scheduler scan details：执行 prewrite 命令时，扫描每个 CF 中 key 的详细情况
- Scheduler scan details [lock]：执行 prewrite 命令时，扫描每个 lock CF 中 key 的详细情况
- Scheduler scan details [write]：执行 prewrite 命令时，扫描每个 write CF 中 key 的详细情况
- Scheduler scan details [default]：执行 prewrite 命令时，扫描每个 default CF 中 key 的详细情况

9.8.4.19 Scheduler - rollback

- Scheduler stage total：rollback 中每个命令所处不同阶段的 ops，正常情况下，不会在短时间内出现大量的错误
- Scheduler command duration：执行 rollback 命令所需花费的时间，正常情况下，应该小于 1s
- Scheduler latch wait duration：由于 latch wait 造成的时间开销，正常情况下，应该小于 1s
- Scheduler keys read：rollback 命令读取 key 的个数
- Scheduler keys written：rollback 命令写入 key 的个数
- Scheduler scan details：执行 rollback 命令时，扫描每个 CF 中 key 的详细情况
- Scheduler scan details [lock]：执行 rollback 命令时，扫描每个 lock CF 中 key 的详细情况
- Scheduler scan details [write]：执行 rollback 命令时，扫描每个 write CF 中 key 的详细情况
- Scheduler scan details [default]：执行 rollback 命令时，扫描每个 default CF 中 key 的详细情况

9.8.4.20 GC

- GC tasks：由 gc_worker 处理的 GC 任务的个数
- GC tasks Duration：执行 GC 任务时所花费的时间
- TiDB GC seconds：TiDB 执行 GC 花费的时间
- TiDB GC worker actions：TiDB GC worker 的不同 action 的个数
- TiKV AutoGC Working：Auto GC 管理器的工作状态
- ResolveLocks Progress：GC 第一阶段 (ResolveLocks) 的进度
- TiKV Auto GC Progress：GC 第二阶段的进度
- GC speed：GC 每秒删除的 key 的数量
- TiKV Auto GC SafePoint：TiKV GC 的 safe point 的数值，safe point 为当前 GC 的时间戳
- GC lifetime：TiDB 设置的 GC lifetime
- GC interval：TiDB 设置的 GC 间隔
- GC in Compaction Filter：write CF 的 Compaction Filter 中已过滤版本的数量

9.8.4.21 Snapshot

- Rate snapshot message：发送 Raft snapshot 消息的速率
- 99% Handle snapshot duration：99% 的情况下，处理 snapshot 所需花费的时间
- Snapshot state count：不同状态的 snapshot 的个数
- 99.99% Snapshot size：99.99% 的 snapshot 的大小
- 99.99% Snapshot KV count：99.99% 的 snapshot 包含的 key 的个数

9.8.4.22 Task

- Worker handled tasks：worker 每秒钟处理的任务的数量
- Worker pending tasks：当前 worker 中，每秒钟 pending 和 running 的任务的数量，正常情况下，应该小于 1000
- FuturePool handled tasks：future pool 每秒钟处理的任务的数量
- FuturePool pending tasks：当前 future pool 中，每秒钟 pending 和 running 的任务的数量

9.8.4.23 Coprocessor Overview

- Request duration：从收到 coprocessor 请求到处理结束所消耗的总时间
- Total Requests：每种类型的总请求的 ops
- Handle duration：每分钟实际处理 coprocessor 请求所消耗的时间的直方图
- Total Request Errors：Coprocessor 每秒请求错误的数量，正常情况下，短时间内不应该有大量的错误
- Total KV Cursor Operations：各种类型的 KV cursor 操作的总数量的 ops，例如 select、index、analyze_table、analyze_index、checksum_table、checksum_index 等
- KV Cursor Operations：每秒各种类型的 KV cursor 操作的数量，以直方图形式显示
- Total RocksDB Perf Statistics：RocksDB 性能统计数据
- Total Response Size：coprocessor 回应的数据大小

9.8.4.24 Coprocessor Detail

- Handle duration：每秒钟实际处理 coprocessor 请求所消耗的时间的直方图
- 95% Handle duration by store：每秒钟中 95% 的情况下，每个 TiKV 实例处理 coprocessor 请求所花费的时间
- Wait duration：coprocessor 每秒钟内请求的等待时间，99.99% 的情况下，应该小于 10s
- 95% Wait duration by store：每秒钟 95% 的情况下，每个 TiKV 实例上 coprocessor 请求的等待时间
- Total DAG Requests：DAG 请求的总数量的 ops
- Total DAG Executors：DAG executor 的总数量的 ops
- Total Ops Details (Table Scan)：coprocessor 中请求为 select 的 scan 过程中每秒钟各种事件发生的次数
- Total Ops Details (Index Scan)：coprocessor 中请求为 index 的 scan 过程中每秒钟各种事件发生的次数
- Total Ops Details by CF (Table Scan)：coprocessor 中对于每个 CF 请求为 select 的 scan 过程中每秒钟各种事件发生的次数
- Total Ops Details by CF (Index Scan)：coprocessor 中对于每个 CF 请求为 index 的 scan 过程中每秒钟各种事件发生的次数

9.8.4.25 Threads

- Threads state：TiKV 线程的状态
- Threads IO：TiKV 各个线程的 I/O 流量
- Thread Voluntary Context Switches：TiKV 线程自主切换的次数
- Thread Nonvoluntary Context Switches：TiKV 线程被动切换的次数

9.8.4.26 RocksDB - kv/raft

- Get operations：get 操作的 ops
- Get duration：get 操作的耗时
- Seek operations：seek 操作的 ops
- Seek duration：seek 操作的耗时
- Write operations：write 操作的 ops
- Write duration：write 操作的耗时
- WAL sync operations：sync WAL 操作的 ops
- Write WAL duration：write 操作中写 WAL 的耗时
- WAL sync duration：sync WAL 操作的耗时
- Compaction operations：compaction 和 flush 操作的 ops
- Compaction duration：compaction 和 flush 操作的耗时
- SST read duration：读取 SST 所需的时间
- Write stall duration：由于 write stall 造成的时间开销，正常情况下应为 0
- Memtable size：每个 CF 的 memtable 的大小
- Memtable hit：memtable 的命中率
- Block cache size：block cache 的大小。如果将 `shared block cache` 禁用，即为每个 CF 的 block cache 的大小
- Block cache hit：block cache 的命中率
- Block cache flow：不同 block cache 操作的流量
- Block cache operations 不同 block cache 操作的个数
- Keys flow：不同操作造成的 key 的流量
- Total keys：每个 CF 中 key 的个数
- Read flow：不同读操作的流量
- Bytes/Read：每次读的大小
- Write flow：不同写操作的流量
- Bytes/Write：每次写的大小
- Compaction flow：compaction 相关的流量
- Compaction pending bytes：等待 compaction 的大小
- Read amplification：每个 TiKV 实例的读放大
- Compression ratio：每一层的压缩比
- Number of snapshots：每个 TiKV 的 snapshot 的数量
- Oldest snapshots duration：最旧的 snapshot 保留的时间
- Number files at each level：每一层的文件个数
- Ingest SST duration seconds：ingest SST 所花费的时间
- Stall conditions changed of each CF：每个 CF stall 的原因

9.8.4.27 Titan - All

- Blob file count：Titan blob 文件的数量
- Blob file size：Titan blob 文件总大小
- Live blob size：有效 blob record 的总大小
- Blob cache hit：Titan 的 blob cache 命中率
- Iter touched blob file count：单个 Iterator 所涉及到 blob 文件的数量
- Blob file discardable ratio distribution：blob 文件的失效 blob record 比例的分布情况
- Blob key size：Titan 中 blob key 的大小
- Blob value size：Titan 中 blob value 的大小
- Blob get operations：blob 的 get 操作的数量
- Blob get duration：blob 的 get 操作的耗时
- Blob iter operations：blob 的 iter 操作的耗时
- Blob seek duration：blob 的 seek 操作的耗时
- Blob next duration：blob 的 next 操作的耗时
- Blob prev duration：blob 的 prev 操作的耗时
- Blob keys flow：Titan blob 读写的 key 数量
- Blob bytes flow：Titan blob 读写的 bytes 数量
- Blob file read duration：blob 文件的读取耗时
- Blob file write duration：blob 文件的写入耗时
- Blob file sync operations：blob 文件 sync 次数
- Blob file sync duration：blob 文件 sync 耗时
- Blob GC action：Titan GC 细分动作的次数
- Blob GC duration：Titan GC 的耗时
- Blob GC keys flow：Titan GC 读写的 key 数量
- Blob GC bytes flow：Titan GC 读写的 bytes 数量
- Blob GC input file size：Titan GC 输入文件的大小
- Blob GC output file size：Titan GC 输出文件的大小
- Blob GC file count：Titan GC 涉及的 blob 文件数量

9.8.4.28 Lock manager

- Thread CPU：lock manager 的线程 CPU 使用率
- Handled tasks：lock manager 处理的任务数量
- Waiter lifetime duration：事务等待锁释放的时间
- Wait table：wait table 的状态信息，包括锁的数量和等锁事务的数量
- Deadlock detect duration：处理死锁检测请求的耗时
- Detect error：死锁检测遇到的错误数量，包含死锁的数量
- Deadlock detector leader：死锁检测器 leader 所在节点的信息

9.8.4.29 Memory

- Allocator Stats：内存分配器的统计信息

9.8.4.30 Backup

- Backup CPU：backup 的线程 CPU 使用率
- Range Size：backup range 的大小直方图
- Backup Duration：backup 的耗时
- Backup Flow：backup 总的字节大小
- Disk Throughput：实例磁盘的吞吐量
- Backup Range Duration：backup range 的耗时
- Backup Errors：backup 中发生的错误数量

9.8.4.31 Encryption

- Encryption data keys：正在使用的加密 data key 的总数量
- Encrypted files：被加密的文件数量
- Encryption initialized：显示加密是否被启用，1 代表已经启用
- Encryption meta files size：加密相关的元数据文件的大小
- Encrypt/decrypt data nanos：每次加密/解密数据的耗时的直方图
- Read/write encryption meta duration：每秒钟读写加密文件所耗费的时间

9.8.4.32 面板常见参数的解释

9.8.4.32.1 gRPC 消息类型

1. 使用事务型接口的命令：

    - kv_get：事务型的 get 命令，获取指定 ts 能读到的最新版本数据
    - kv_scan：扫描连续的一段数据
    - kv_prewrite：2PC 的第一阶段，预写入事务要提交的数据
    - kv_pessimistic_lock：对 key 加悲观锁，防止其他事务修改
    - kv_pessimistic_rollback：删除 key 上的悲观锁
    - kv_txn_heart_beat：更新悲观事务或大事务的 `lock_ttl` 以防止其被回滚
    - kv_check_txn_status：检查事务的状态
    - kv_commit：2PC 的第二阶段，提交 prewrite 阶段写入的数据
    - kv_cleanup：回滚一个事务（此命令将会在 4.0 中废除）
    - kv_batch_get：与 `kv_get` 类似，一次性获取批量 key 的 value
    - kv_batch_rollback：批量回滚多个预写的事务
    - kv_scan_lock：扫描所有版本号在 `max_version` 之前的锁，用于清理过期的事务
    - kv_resolve_lock：根据事务状态，提交或回滚事务的锁
    - kv_gc：触发垃圾回收
    - kv_delete_range：从 TiKV 中删除连续的一段数据

2. 非事务型的裸命令：

    - raw_get：获取 key 所对应的 value
    - raw_batch_get：获取一批 key 所对应的 value
    - raw_scan：扫描一段连续的数据
    - raw_batch_scan：扫描多段连续的数据

- raw_put：写入一个 key/value 对
- raw_batch_put：直接写入一批 key/value 对
- raw_delete：删除一个 key/value 对
- raw_batch_delete：删除一批 key/value 对
- raw_delete_range：删除连续的一段区间

9.8.5 TiFlash 集群监控

使用 TiUP 部署 TiDB 集群时，一键部署监控系统 (Prometheus & Grafana)，监控架构参见TiDB 监控框架概述。

目前 Grafana Dashboard 整体分为 PD、 TiDB、 TiKV、 Node_exporter、 Overview 等。

TiFlash 面板一共包括 TiFlash-Summary、 TiFlash-Proxy-Summary、 TiFlash-Proxy-Details。通过面板上的指标，可以了解 TiFlash 当前的状态。其中 TiFlash-Proxy-Summary、 TiFlash-Proxy-Details 主要为 TiFlash 的 Raft 层信息，其监控指标信息可参考TiKV 监控指标详解。

> 注意：
>
> 低版本的 TiFlash 监控信息较不完善，如有需要推荐使用 v4.0.5 或更高版本的 TiDB 集群。

以下为 TiFlash-Summary 默认的监控信息：

9.8.5.1 Server

- Store size：每个 TiFlash 实例的使用的存储空间的大小。
- Available size：每个 TiFlash 实例的可用的存储空间的大小。
- Capacity size：每个 TiFlash 实例的存储容量的大小。
- Uptime：自上次重启以来 TiFlash 正常运行的时间。
- Memory：每个 TiFlash 实例内存的使用情况。
- CPU Usage：每个 TiFlash 实例 CPU 的使用率。
- FSync OPS：每个 TiFlash 实例每秒进行 fsync 操作的次数。
- File Open OPS：每个 TiFlash 实例每秒进行 open 操作的次数。
- Opened File Count：当前每个 TiFlash 实例打开的文件句柄数。

> 注意：
>
> Store size、FSync OPS、File Open OPS、Opened File Count 目前仅包含了 TiFlash 存储层的统计指标，未包括 TiFlash-Proxy 内的信息。

9.8.5.2 Coprocessor

- Request QPS：所有 TiFlash 实例收到的 coprocessor 请求数量。其中 batch 是 batch 请求数量，batch_cop 是 batch 请求中的 coprocessor 请求数量，cop 是直接通过 coprocessor 接口发送的 coprocessor 请求数量，cop_dag 是所有 coprocessor 请求中 dag 请求数量，super_batch 是开启 super batch 特性的请求数量。
- Executor QPS：所有 TiFlash 实例收到的请求中，每种 dag 算子的数量，其中 table_scan 是扫表算子，selection 是过滤算子，aggregation 是聚合算子，top_n 是 TopN 算子，limit 是 limit 算子。
- Request Duration：所有 TiFlash 实例处理 coprocessor request 总时间，总时间为接收到该 coprocessor 请求至请求应答完毕的时间。
- Error QPS：所有 TiFlash 实例处理 coprocessor 请求的错误数量。其中 meet_lock 为读取的数据有锁，region_not_found 为 Region 不存在，epoch_not_match 为读取的 Region epoch 与本地不一致，kv_client_error 为与 TiKV 通信产生的错误，internal_error 为 TiFlash 内部系统错误，other 为其他错误。
- Request Handle Duration：所有 TiFlash 实例处理 coprocessor 请求处理时间，处理时间为该 coprocessor 请求开始执行到执行结束的时间。
- Response Bytes/Seconds：所有 TiFlash 实例应答总字节数。
- Cop task memory usage：所有 TiFlash 实例处理 coprocessor 请求占用的总内存。
- Handling Request Number：所有 TiFlash 实例正在处理的 coprocessor 请求数量之和。请求的分类与 Request QPS 中的相同.

9.8.5.3 DDL

- Schema Version：每个 TiFlash 实例目前缓存的 schema 版本。
- Schema Apply OPM：所有 TiFlash 实例每分钟 apply 同步 TiDB schema diff 的次数。diff apply 是正常的单次 apply 过程，如果 diff apply 失败，则 failed apply +1，并回退到 full apply，拉取最新的 schema 信息以更新 TiFlash 的 schema 版本。
- Schema Internal DDL OPM：所有 TiFlash 实例每分钟执行的内部 DDL 次数。
- Schema Apply Duration：所有 TiFlash 实例 apply schema 消耗的时间。

9.8.5.4 Storage

- Write Command OPS：所有 TiFlash 实例存储层每秒收到的写请求数量。
- Write Amplification：每个 TiFlash 实例写放大倍数（实际磁盘写入量/逻辑数据写入量）。total 为自此次启动以来的写放大倍数，5min 为最近 5 分钟内的写放大倍数。
- Read Tasks OPS：每个 TiFlash 实例每秒存储层内部读取任务的数量。
- Rough Set Filter Rate：每个 TiFlash 实例最近 1 分钟内读取的 packet 数被存储层粗糙索引过滤的比例。
- Internal Tasks OPS：所有 TiFlash 实例每秒进行内部数据整理任务的次数。
- Internal Tasks Duration：所有 TiFlash 实例进行内部数据整理任务消耗的时间。
- Page GC Tasks OPM：所有 TiFlash 实例每分钟进行 Delta 部分数据整理任务的次数。
- Page GC Tasks Duration：所有 TiFlash 实例进行 Delta 部分数据整理任务消耗的时间分布。
- Disk Write OPS：所有 TiFlash 实例每秒进行磁盘写入的次数。
- Disk Read OPS：所有 TiFlash 实例每秒进行磁盘读取的次数。
- Write flow：所有 TiFlash 实例磁盘写操作的流量。
- Read flow：所有 TiFlash 实例磁盘读操作的流量。

> **注意：**
>
> 目前这部分监控指标仅包含了 TiFlash 存储层的统计指标，未包括 TiFlash-Proxy 内的信息。

9.8.5.5 Storage Write Stall

- Write & Delta Management Throughput：所有实例写入及数据整理的吞吐量。
  - `throughput_write` 表示通过 Raft 进行数据同步的吞吐量。
  - `throughput_delta-management` 表示数据整理的吞吐量。
  - `total_write` 表示自上次启动以来的总写入字节数。
  - `total_delta-management` 表示自上次启动以来数据整理的总字节数。
- Write Stall Duration：每个实例写入和移除 Region 数据产生的卡顿时长。
- Write Throughput By Instance：每个实例写入数据的吞吐量，包括 apply Raft 数据日志以及 Raft 快照的写入吞吐量。
- Write Command OPS By Instance：每个实例收到各种命令的总计数。
  - `write block` 表示通过 Raft 同步数据日志。
  - `delete_range` 表示从该实例中删除一些 Region 或移动一些 Region 到该实例中。
  - `ingest` 表示这些 Region 的快照被应用到这个实例中。

9.8.5.6 Raft

- Read Index OPS：每个 TiFlash 实例每秒触发 read_index 请求的次数，等于请求触发的 Region 总数。
- Read Index Duration：所有 TiFlash 实例在进行 read_index 消耗的时间，主要消耗在于和 Region leader 的交互和重试时间。
- Wait Index Duration：所有 TiFlash 实例在进行 wait_index 消耗的时间，即拿到 read_index 请求后，等待本地的 Region index >= read_index 所花费的时间。

9.8.6 TiCDC 重要监控指标详解

使用 TiUP 部署 TiDB 集群时，一键部署的监控系统面板包含 TiCDC 面板。本文档对 TiCDC 监控面板上的各项指标进行详细说明。在日常运维中，运维人员可通过观察 TiCDC 面板上的指标了解 TiCDC 当前的状态。

本文档的对指标的介绍基于以下同步任务，即使用默认配置同步数据到 MySQL。

```
cdc cli changefeed create --pd=http://10.0.10.25:2379 --sink-uri="mysql://root:123456@127
    ↪ .0.0.1:3306/" --changefeed-id="simple-replication-task"
```

下图显示了 TiCDC Dashboard 各监控面板：



图 72: TiCDC Dashboard - Overview

各监控面板说明如下：

- Server：TiDB 集群中 TiKV 节点和 TiCDC 节点的概要信息
- Changefeed：TiCDC 同步任务的详细信息
- Events：TiCDC 内部数据流转的详细信息
- TiKV：TiKV 中和 TiCDC 相关的详细信息

9.8.6.1 Server 面板

Server 面板示例如下：



图 73: TiCDC Dashboard - Server metrics

Server 面板的各指标说明如下：

- Uptime：TiKV 节点和 TiCDC 节点已经运行的时间
- Goroutine count：TiCDC 节点 Goroutine 的个数
- Open FD count：TiCDC 节点打开的文件句柄个数
- Ownership：TiCDC 集群中节点的当前状态
- Ownership history：TiCDC 集群中 Owner 节点的历史记录
- CPU usage：TiCDC 节点使用的 CPU
- Memory usage：TiCDC 节点使用的内存

9.8.6.2 Changefeed 面板

Changefeed 面板示例如下：



Changefeed 面板的各指标说明如下：

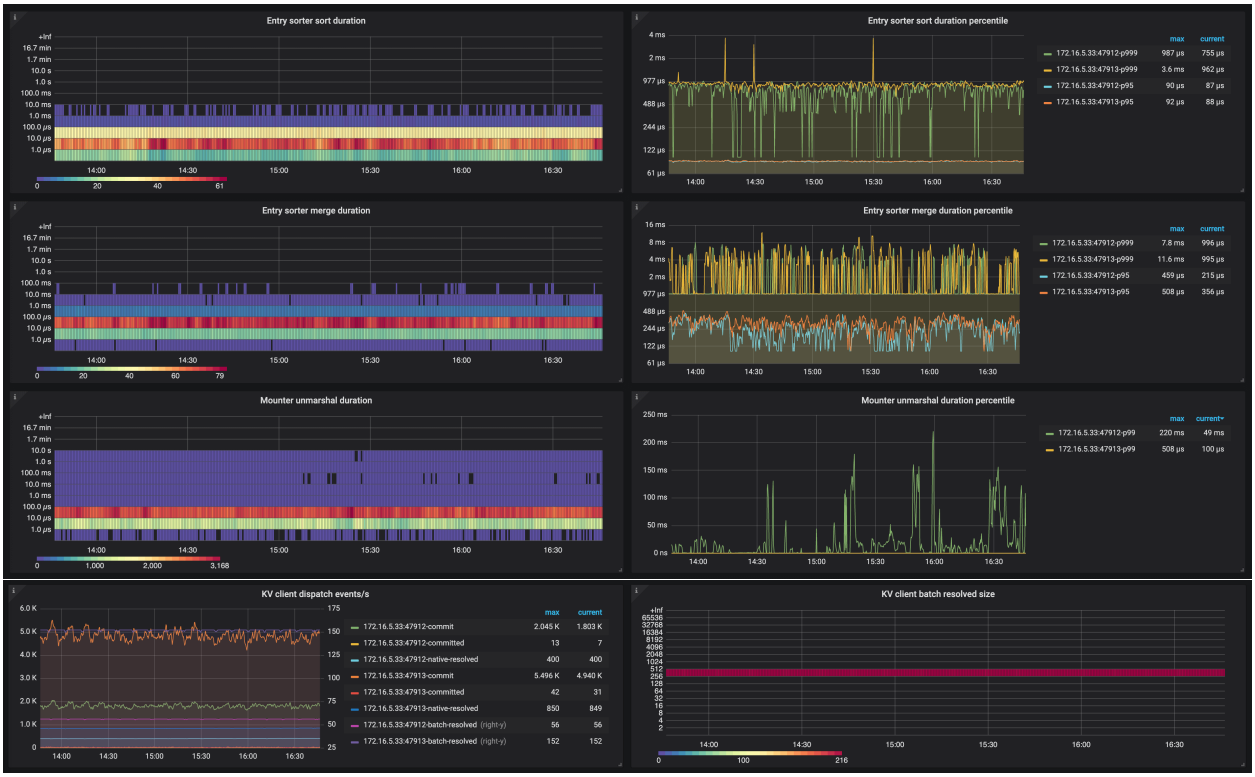
- Changefeed table count：一个同步任务中分配到各个 TiCDC 节点同步的数据表个数。
- Processor resolved ts：TiCDC 节点内部状态中已同步的时间点。
- Table resolved ts：同步任务中各数据表的同步进度。
- Changefeed checkpoint：同步任务同步到下游的进度，正常情况下绿柱应和黄线相接。
- PD etcd requests/s：TiCDC 节点每秒向 PD 读写数据的次数。
- Exit error count：每分钟内导致同步中断的错误发生次数。

- Changefeed checkpoint lag：同步任务上下游数据的进度差（以时间计算）。
- Changefeed resolved ts lag：TiCDC 节点内部同步状态与上游的进度差（以时间计算）。
- Flush sink duration：TiCDC 异步刷写数据入下游的耗时直方图。
- Flush sink duration percentile：每秒钟中 95%、99% 和 99.9% 的情况下，TiCDC 异步刷写数据入下游所花费的时间。
- Sink write duration：TiCDC 将一个事务的更改写到下游的耗时直方图。
- Sink write duration percentile：每秒钟中 95%、99% 和 99.9% 的情况下，TiCDC 将一个事务的更改写到下游所花费的时间。
- MySQL sink conflict detect duration：MySQL 写入冲突检测耗时直方图。
- MySQL sink conflict detect duration percentile：每秒钟中 95%、99% 和 99.9% 的情况下，MySQL 写入冲突检测耗时。
- MySQL sink worker load：TiCDC 节点中写 MySQL 线程的负载情况。

9.8.6.3 Events 面板

Events 面板示例如下：





Events 面板的各指标说明如下：

- Eventfeed count：TiCDC 节点中 Eventfeed RPC 的个数。
- Event size percentile：每秒钟中 95% 和 99.9% 的情况下，TiCDC 收到的来自 TiKV 的数据变更消息大小。
- Eventfeed error/m：TiCDC 节点中每分钟 Eventfeed RPC 遇到的错误个数。
- KV client receive events/s：TiCDC 节点中 KV client 模块每秒收到来自 TiKV 的数据变更个数。
- Puller receive events/s：TiCDC 节点中 Puller 模块每秒收到来自 KV client 模块的数据变更个数。
- Puller output events/s：TiCDC 节点中 Puller 模块每秒输出到 Sorter 模块的数据变更个数。
- Sink flush rows/s：TiCDC 节点每秒写到下游的数据变更的个数。
- Puller buffer size：TiCDC 节点中缓存在 Puller 模块中的数据变更个数。
- Entry sorter buffer size：TiCDC 节点中缓存在 Sorter 模块中的数据变更个数。
- Processor/Mounter buffer size：TiCDC 节点中缓存在 Processor 模块和 Mounter 模块中的数据变更个数。
- Sink row buffer size：TiCDC 节点中缓存在 Sink 模块中的数据变更个数。
- Entry sorter sort duration：TiCDC 节点对数据变更进行排序的耗时直方图。
- Entry sorter sort duration percentile：每秒钟中 95%，99% 和 99.9% 的情况下，TiCDC 排序数据变更所花费的时间。
- Entry sorter merge duration：TiCDC 节点合并排序后的数据变更的耗时直方图。
- Entry sorter merge duration percentile：每秒钟中 95%，99% 和 99.9% 的情况下，TiCDC 合并排序后的数据变更所花费的时间。
- Mounter unmarshal duration：TiCDC 节点解码数据变更的耗时直方图。
- Mounter unmarshal duration percentile：每秒钟中 95%，99% 和 99.9% 的情况下，TiCDC 解码数据变更所花费的时间。
- KV client dispatch events/s：TiCDC 节点内部 KV client 模块每秒分发数据变更的个数。
- KV client batch resolved size：TiKV 批量发给 TiCDC 的 resolved ts 消息的大小。

9.8.6.4 TiKV 面板

TiKV 面板示例如下：



TiKV 面板的各指标说明如下：

- CDC endpoint CPU：TiKV 节点上 CDC endpoint 线程使用的 CPU
- CDC worker CPU：TiKV 节点上 CDC worker 线程使用的 CPU
- Min resolved ts：TiKV 节点上最小的 resolved ts
- Min resolved region：TiKV 节点上最小的 resolved ts 的 Region ID
- Resolved ts lag duration percentile：TiKV 节点上最小的 resolved ts 与当前时间的差距
- Initial scan duration：TiKV 节点与 TiCDC 建立链接时增量扫的耗时直方图
- Initial scan duration percentile：每秒钟中 95%、99% 和 99.9% 的情况下，TiKV 节点增量扫的耗时
- Memory without block cache：TiKV 节点在减去 RocksDB block cache 后使用的内存
- CDC pending bytes in memory：TiKV 节点中 CDC 模块使用的内存
- Captured region count：TiKV 节点上捕获数据变更的 Region 个数

9.8.7 DM 监控指标

使用 TiUP 部署 DM 集群的时候，会默认部署一套监控系统。

9.8.7.1 Task

在 Grafana dashboard 中，DM 默认名称为 `DM-task`。

9.8.7.1.1 Overview

overview 下包含运行当前选定 task 的所有 DM-worker/master instance/source 的部分监控指标。当前默认告警规则只针对于单个 DM-worker/master instance/source。

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	N/A	N/A
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	N/A	N/A
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
binlog file gap between master and syncer	与上游 master 相比 binlog replication unit 落后的 binlog file 个数	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A

9.8.7.1.2 Operate error

metric 名称	说明	告警说明	告警级别
before any operate error	在进行操作之前出错的次数	N/A	N/A
source bound error	数据源绑定操作出错次数	N/A	N/A

metric 名称	说明	告警说明	告警级别
start error	子任务启动的出错次数	N/A	N/A
pause error	子任务暂停的出错次数	N/A	N/A
resume error	子任务恢复的出错次数	N/A	N/A
auto-resume error	子任务自动恢复的出错次数	N/A	N/A
update error	子任务更新的出错次数	N/A	N/A
stop error	子任务停止的出错次数	N/A	N/A

9.8.7.1.3 HA 高可用

metric 名称	说明	告警说明	告警级别
number of dm-masters start leader components per minute	每分钟内 DM-master 尝试启用 leader 相关组件次数	N/A	N/A
number of workers in different state	不同状态下有多少个 DM-worker	存在离线的 DM-worker 超过一小时	critical
workers' state	DM-worker 的状态	N/A	N/A
number of worker event error	不同类型的 DM-worker 错误出现次数	N/A	N/A
shard ddl error per minute	每分钟内不同类型的 shard DDL 错误次数	发生 shard DDL 错误	critical
number of pending shard ddl	未完成的 shard DDL 数目	存在未完成的 shard DDL 数目超过一小时	critical

metric 名称	说明	告警说明	告警级别

9.8.7.1.4 Task 状态

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	当子任务状态处于 Paused 超过 20 分钟时	critical

9.8.7.1.5 Dump/Load unit

下面 metrics 仅在 `task-mode` 为 `full` 或者 `all` 模式下会有值。

metric 名称	说明	告警说明	告警级别
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
data file size	load unit 导入的全量数据中数据文件（内含 `INSERT INTO` 语句）的总大小	N/A	N/A
dump process exits with error	dump unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
load process exits with error	load unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
table count	load unit 导入的全量数据中 table 的数量总和	N/A	N/A
data file count	load unit 导入的全量数据中数据文件（内含 `INSERT INTO` 语句）的数量总和	N/A	N/A
transaction execution latency	load unit 在执行事务的时延，单位：秒	N/A	N/A
statement execution latency	load unit 执行语句的耗时，单位：秒	N/A	N/A
remaining time	load unit 完成同步的剩余时间，单位：秒	N/A	N/A

9.8.7.1.6 Binlog replication

下面 metrics 仅在 `task-mode` 为 `incremental` 或者 `all` 模式下会有值。

metric 名称	说明	告警说明	告警级别
remaining time to sync	预计 Syncer 还需要多少分钟可以和上游 master 完全同步，单位：分钟	N/A	N/A
replicate lag gauge	上游 master 到下游的 binlog 复制延迟时间，单位：秒	N/A	N/A

metric 名称	说明	告警说明	告警级别
replicate lag histogram	上游 master 到下游的 binlog 复制延迟分布，单位：秒。注意由于统计机制不同，数据会有误差	N/A	N/A
process exist with error	binlog replication unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
binlog file gap between master and syncer	与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个 (不含 1 个) 且持续 10 分钟时	critical
binlog file gap between relay and syncer	与 relay 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个 (不含 1 个) 且持续 10 分钟时	critical
binlog event QPS	单位时间内接收到的 binlog event 数量 (不包含需要跳过的 event)	N/A	N/A
skipped binlog event QPS	单位时间内接收到的需要跳过的 binlog event 数量	N/A	N/A
read binlog event duration	binlog replication unit 从 relay log 或上游 MySQL 读取 binlog 的耗时，单位：秒	N/A	N/A

metric 名称	说明	告警说明	告警级别
transform binlog event duration	binlog replication unit 解析 binlog 并将 binlog 转换成 SQL 语句的耗时，单位：秒	N/A	N/A
dispatch binlog event duration	binlog replication unit 调度一条 binlog event 的耗时，单位：秒	N/A	N/A
transaction execution latency	binlog replication unit 执行事务到下游的耗时，单位：秒	N/A	N/A
binlog event size	binlog replication unit 从 relay log 或上游 MySQL 读取的单条 binlog event 的大小	N/A	N/A
DML queue remain length	剩余 DML job 队列的长度	N/A	N/A
total sqls jobs	单位时间内新增的 job 数量	N/A	N/A
finished sqls jobs	单位时间内完成的 job 数量	N/A	N/A
statement execution latency	binlog replication unit 执行语句到下游的耗时，单位：秒	N/A	N/A
add job duration	binlog replication unit 增加一条 job 到队列的耗时，单位：秒	N/A	N/A
DML conflict detect duration	binlog replication unit 检测 DML 间冲突的耗时，单位：秒	N/A	N/A
skipped event duration	binlog replication unit 跳过 binlog event 的耗时，单位：秒	N/A	N/A

metric 名称	说明	告警说明	告警级别
unsynced tables	当前子任务内还未收到 shard DDL 的分表数量	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A
ideal QPS	在 DM 运行耗时为 0 时可以达到的最高 QPS	N/A	N/A
binlog event row	一个 binlog 事件中的行数	N/A	N/A
finished transaction total	执行完毕的事务数量	N/A	N/A
replication transaction batch	执行到下游的事务里中 sql 行数	N/A	N/A
flush checkpoints time interval	检查点刷新时间间隔，单位：秒	N/A	N/A

9.8.7.1.7 Relay log

metric 名称	说明	告警说明	告警级别
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	小于 10G 的时候需要告警	critical
process exits with error	relay log 在 DM-worker 内部遇到错误并且退出了	立即告警	critical

metric 名称	说明	告警说明	告警级别
relay log data corruption	relay log 文件损坏的个数	立即告警	emergency
fail to read binlog from master	relay 从上游的 MySQL 读取 binlog 时遇到的错误数	立即告警	critical
fail to write relay log	relay 写 binlog 到磁盘时遇到的错误数	立即告警	critical
binlog file index	relay log 最大的文件序列号。如 value = 1 表示 relay-log.000001	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个（不含 1 个）且持续 10 分钟时	critical
binlog pos	relay log 最新文件的写入 offset	N/A	N/A
read binlog event duration	relay log 从上游的 MySQL 读取 binlog 的时延，单位：秒	N/A	N/A
write relay log duration	relay log 每次写 binlog 到磁盘的时延，单位：秒	N/A	N/A
binlog event size	relay log 写到磁盘的单条 binlog 的大小	N/A	N/A

9.8.7.2 Instance

在 Grafana dashboard 中，instance 的默认名称为 `DM-instance`。

9.8.7.2.1 Relay log

metric 名称	说明	告警说明	告警级别
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	小于 10G 的时候需要告警	critical
process exits with error	relay log 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
relay log data corruption	relay log 文件损坏的个数	立即告警	emergency
fail to read binlog from master	relay 从上游的 MySQL 读取 binlog 时遇到的错误数	立即告警	critical
fail to write relay log	relay 写 binlog 到磁盘时遇到的错误数	立即告警	critical
binlog file index	relay log 最大的文件序列号。如 value = 1 表示 relay-log.000001	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个（不含 1 个）且持续 10 分钟时	critical
binlog pos	relay log 最新文件的写入 offset	N/A	N/A
read binlog duration	relay log 从上游的 MySQL 读取 binlog 的时延，单位：秒	N/A	N/A

metric 名称	说明	告警说明	告警级别
write relay log duration	relay log 每次写 binlog 到磁盘的时延，单位：秒	N/A	N/A
binlog size	relay log 写到磁盘的单条 binlog 的大小	N/A	N/A

9.8.7.2.2 task

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	当子任务状态处于 paused 超过 10 分钟时	critical
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
binlog file gap between master and syncer	与上游 master 相比 binlog replication unit 落后的 binlog file 个数	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A

9.9 TiDB Dashboard 监控指标

9.9.1 TiDB Dashboard 介绍

TiDB Dashboard 是 TiDB 自 4.0 版本起提供的图形化界面，可用于监控及诊断 TiDB 集群。TiDB Dashboard 内置于 TiDB 的 PD 组件中，无需独立部署。

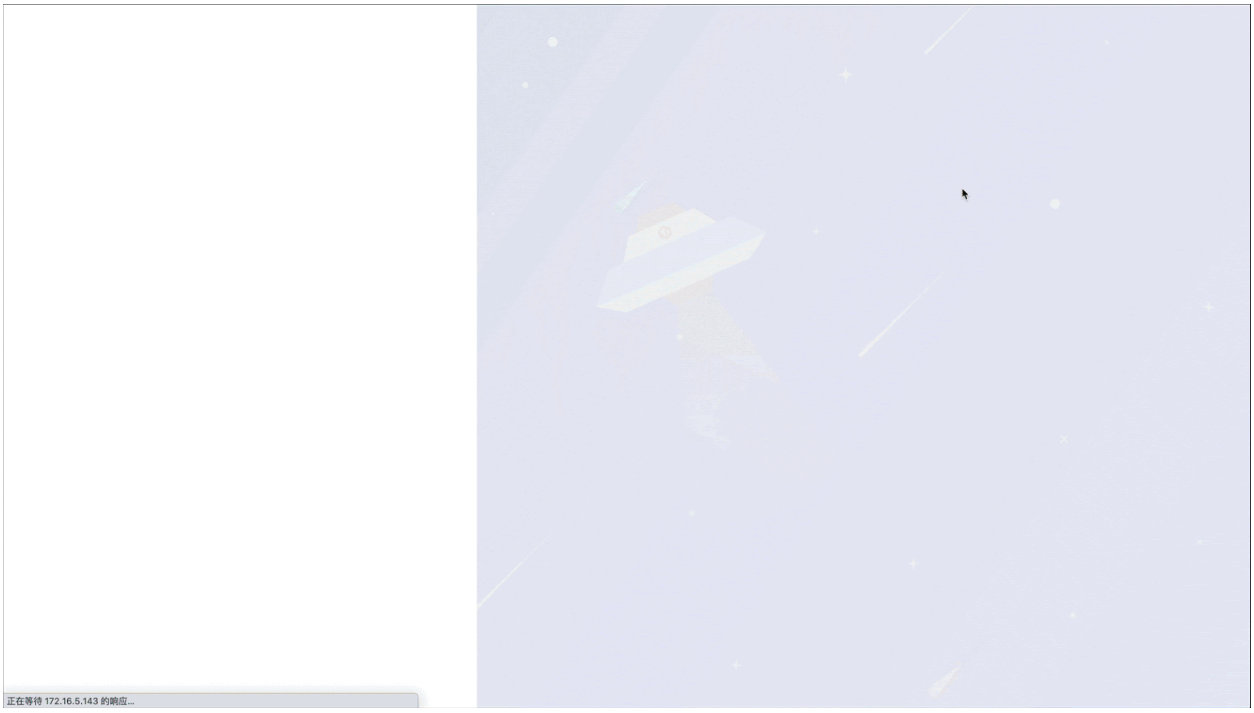


图 74: 界面

TiDB Dashboard 在 GitHub 上开源。

以下列出了 TiDB Dashboard 的主要功能，可分别点击小节内的链接进一步了解详情。

9.9.1.1 了解集群整体运行概况

查看集群整体 QPS 数值、执行耗时、消耗资源最多的几类 SQL 语句等概况信息。

参阅概况页面了解详情。

9.9.1.2 查看组件及主机运行状态

查看整个集群中 TiDB、TiKV、PD、TiFlash 组件的运行状态及其所在主机的运行状态。

参阅集群信息页面了解详情。

9.9.1.3 分析集群读写流量分布及趋势变化

通过热力图形式可视化地展示整个集群中读写流量随时间的变化情况，及时发现业务模式的变化，或定位性能不均衡的热点所在。

参阅流量可视化页面了解详情。

9.9.1.4 列出所有 SQL 查询的耗时等执行信息

列出所有 SQL 语句在集群上执行情况，了解各个阶段的执行时间、总运行次数等信息，帮助用户分析和定位集群中最消耗资源的查询，优化整体性能。

参阅SQL 语句分析页面了解详情。

9.9.1.5 详细了解耗时较长的 SQL 语句的执行信息

列出所有耗时较长的 SQL 语句文本及其执行信息，帮助用户定位 SQL 语句性能缓慢或发生性能抖动的原因。

参阅慢查询页面了解详情。

9.9.1.6 诊断常见集群问题并生成报告

自动判断集群中是否存在一些常见的风险（如配置不一致）或问题，生成报告并给出操作建议，或对比集群在不同时间段的各个指标状态，供用户分析可能存在问题的方向。

参阅集群诊断页面了解详情。

9.9.1.7 查询所有组件日志

按关键字、时间范围等条件快速搜索集群中所有运行实例的日志，并可打包下载到本地。

参阅日志搜索页面了解详情。

9.9.1.8 收集分析各个组件的性能数据

高级调试功能：无需第三方工具，在线地对各个组件进行性能分析，剖析组件实例在分析时间段内执行的各种内部操作及比例。

参阅实例性能分析页面了解详情。

> **注意：**
>
> TiDB Dashboard 默认会收集使用情况信息，并将这些信息分享给 PingCAP 用于改善产品。若要了解所收集的信息详情及如何禁用该行为，请参见遥测。

9.9.2 运维

9.9.2.1 部署 TiDB Dashboard

TiDB Dashboard 界面内置于 TiDB 4.0 或更高版本的 PD 组件中，无需额外部署。只需部署标准 TiDB 集群，TiDB Dashboard 就会原生集成。

请参阅下列文档了解如何部署标准 TiDB 集群：

- 快速试用 TiDB 集群
- 生产环境部署
- Kubernetes 环境部署

> **注意：**
>
> TiDB Dashboard 目前不能在低于 4.0 版本的集群中部署或使用。

9.9.2.1.1 多 PD 实例部署

当集群中部署了多个 PD 实例时，其中仅有一个 PD 实例会固定地提供 TiDB Dashboard 服务。

各个 PD 首次运行时会自动协商出其中某一个实例提供 TiDB Dashboard 服务。协商完毕后，无论重启或扩容，都会固定在这个实例上运行 TiDB Dashboard 服务，除非该实例被手动缩容。其他 PD 实例不会运行 TiDB Dashboard 服务。这个协商过程无需用户介入，会自动完成。

当用户访问不提供 TiDB Dashboard 服务的 PD 实例时，浏览器将会收到重定向指令，自动引导用户重新访问提供了 TiDB Dashboard 服务的 PD 实例，从而能正常使用。流程如下图所示。

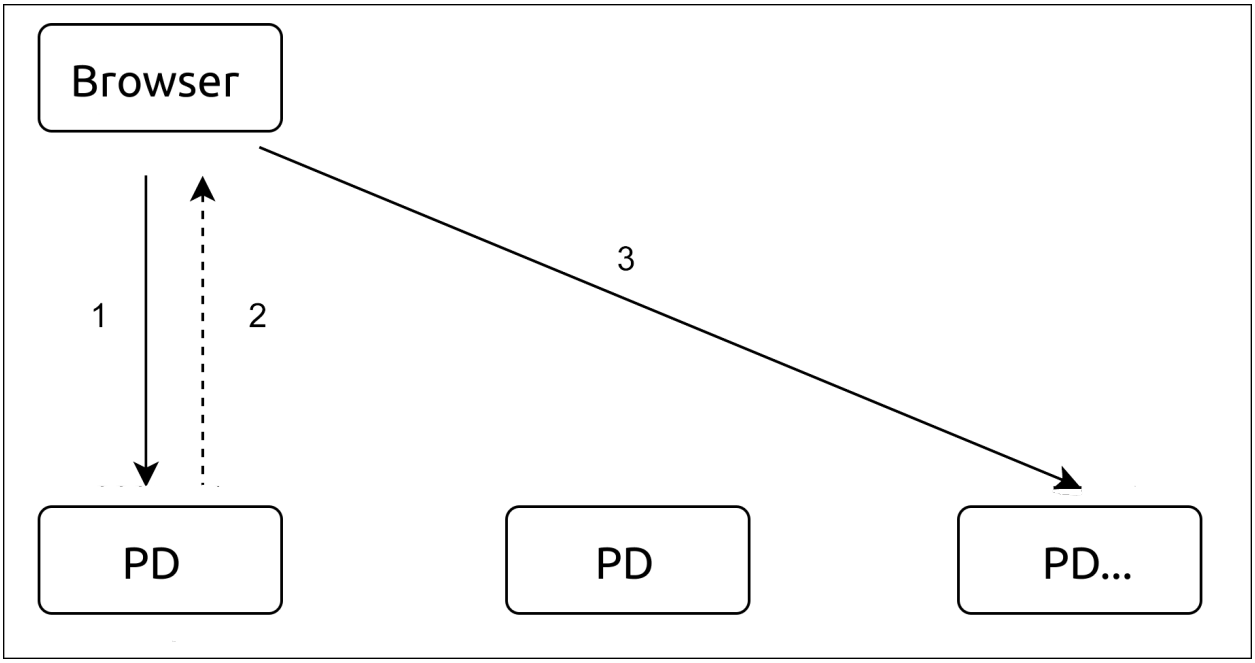


图 75: 流程示意

> **注意：**
>
> 提供 TiDB Dashboard 服务的 PD 实例不一定与 PD leader 一致。

查询实际运行 TiDB Dashboard 服务的 PD 实例

使用 TiUP 部署时，对于已启动的集群，可通过 `tiup cluster display` 命令查看哪个 PD 节点提供了 TiDB Dashboard 服务（将 `CLUSTER_NAME` 替换为集群名称）：

```
tiup cluster display CLUSTER_NAME --dashboard
```

输出样例如下：

```
http://192.168.0.123:2379/dashboard/
```

> **注意：**
>
> 该功能在 TiUP Cluster v1.0.3 或更高版本部署工具中提供。
>
> 升级 TiUP Cluster 步骤
>
> ```
> tiup update --self
> tiup update cluster --force
> ```

切换其他 PD 实例提供 TiDB Dashboard 服务

使用 TiUP 部署时，对于已启动的集群，可使用 `tiup ctl pd` 命令切换其他 PD 实例运行 TiDB Dashboard，或在禁用 TiDB Dashboard 的情况下重新指定一个 PD 实例运行 TiDB Dashboard：

```
tiup ctl pd -u http://127.0.0.1:2379 config set dashboard-address http://9.9.9.9:2379
```

其中：

- 将 `127.0.0.1:2379` 替换为任意 PD 实例的 IP 和端口
- 将 `9.9.9.9:2379` 替换为想运行 TiDB Dashboard 服务的新 PD 实例的 IP 和端口

修改完毕后，可使用 `tiup cluster display` 命令确认修改是否生效（将 CLUSTER_NAME 替换为集群名称）：

```
tiup cluster display CLUSTER_NAME --dashboard
```

> **警告：**
>
> 切换 TiDB Dashboard 将会丢失之前 TiDB Dashboard 实例所存储的本地数据，包括流量可视化历史、历史搜索记录等。

9.9.2.1.2 禁用 TiDB Dashboard

使用 TiUP 部署时，对于已启动的集群，可使用 `tiup ctl pd` 命令在所有 PD 实例上禁用 TiDB Dashboard（将 `127.0.0.1:2379` 替换为任意 PD 实例的 IP 和端口）：

```
tiup ctl pd -u http://127.0.0.1:2379 config set dashboard-address none
```

禁用 TiDB Dashboard 后，查询哪个 PD 实例提供 TiDB Dashboard 服务将会失败：

```
Error: TiDB Dashboard is disabled
```

浏览器访问任意 PD 实例的 TiDB Dashboard 地址也将提示失败：

```
Dashboard is not started.
```

9.9.2.1.3 重新启用 TiDB Dashboard

使用 TiUP 部署时，对于已启动的集群，可使用 `tiup ctl pd` 命令，要求 PD 重新协商出某一个实例运行 TiDB Dashboard（将 `127.0.0.1:2379` 替换为任意 PD 实例的 IP 和端口）：

```
tiup ctl pd -u http://127.0.0.1:2379 config set dashboard-address auto
```

修改完毕后，使用 `tiup cluster display` 命令查看 PD 自动协商出的 TiDB Dashboard 实例地址（将 CLUSTER_NAME 替换为集群名称）：

```
tiup cluster display CLUSTER_NAME --dashboard
```

还可以通过手动指定哪个 PD 实例运行 TiDB Dashboard 服务的方式重新启用 TiDB Dashboard，具体操作参见上文切换其他 PD 实例提供 TiDB Dashboard 服务。

> 警告：
>
> 若新启用的 TiDB Dashboard 实例与禁用前的实例不一致，将会丢失之前 TiDB Dashboard 实例所存储的本地数据，包括流量可视化历史、历史搜索记录等。

9.9.2.1.4 下一步

- 参阅访问 TiDB Dashboard 章节了解如何访问及登录集群上的 TiDB Dashboard 界面。
- 参阅提高 TiDB Dashboard 安全性章节了解如何增强 TiDB Dashboard 的安全性，如配置防火墙等。

9.9.2.2 通过反向代理使用 TiDB Dashboard

你可以使用反向代理将 TiDB Dashboard 服务安全从内部网络提供给外部网络。

9.9.2.2.1 操作步骤

第 1 步：获取实际 TiDB Dashboard 地址

当集群中部署有多个 PD 实例时，其中仅有一个 PD 实例会真正运行 TiDB Dashboard，因此需要确保反向代理的上游 (Upstream) 指向了正确的地址。关于该机制的详情，可参阅TiDB Dashboard 多 PD 实例部署章节。

使用 TiUP 部署工具时，操作命令如下（将 CLUSTER_NAME 替换为集群名称）：

```
tiup cluster display CLUSTER_NAME --dashboard
```

输出即为实际 TiDB Dashboard 地址。样例如下：

```
http://192.168.0.123:2379/dashboard/
```

> 注意：
>
> 该功能在 TiUP Cluster v1.0.3 或更高版本部署工具中提供。
>
> 升级 TiUP Cluster 步骤
>
> ```
> tiup update --self
> tiup update cluster --force
> ```

第 2 步：配置反向代理

使用 HAProxy 反向代理

HAProxy 作为反向代理时，方法如下：

1. 以在 8033 端口反向代理 TiDB Dashboard 为例，在 HAProxy 配置文件中，新增如下配置：

“‘haproxy frontend tidb_dashboard_front bind *:8033 use_backend tidb_dashboard_back if { path /dashboard } or { path_beg /dashboard/ }

backend tidb_dashboard_back mode http server tidb_dashboard 192.168.0.123:2379 “‘

其中 192.168.0.123:2379 需替换为第 1 步：获取实际 TiDB Dashboard 地址中取得的 TiDB Dashboard 实际地址中的 IP 及端口部分。

> 警告：
>
> 请务必保留 use_backend 指令中的 if 部分，确保只有该路径下的服务会被反向代理，否则将引入安全风险。参见提高 TiDB Dashboard 安全性。

2. 重启 HAProxy，以使配置生效。
3. 测试反向代理是否生效：访问 HAProxy 所在机器的 8033 端口下 /dashboard/ 地址，如 http://example.com:8033/dashboard/ ，即可访问 TiDB Dashboard。

使用 NGINX 反向代理

NGINX 作为反向代理时，方法如下：

1. 以在 8033 端口反向代理 TiDB Dashboard 为例，在 NGINX 配置文件中，新增如下配置：

```
nginx  server {  listen 8033; location /dashboard/ { proxy_pass http://192.168.0.123:2379/dashboard
↪ /; } }
```

其中 `http://192.168.0.123:2379/dashboard/` 需替换为第 1 步：获取实际 TiDB Dashboard 地址中取得的 TiDB Dashboard 实际地址。

> **警告：**
>
> 请务必保留 `proxy_pass` 指令中的 `/dashboard/` 路径，确保只有该路径下的服务会被反向代理，否则将引入安全风险。参见提高 TiDB Dashboard 安全性。

2. 重新载入 NGINX 以使配置生效：

```
shell  sudo nginx -s reload
```

3. 测试反向代理是否生效：访问 NGINX 所在机器的 8033 端口下 `/dashboard/` 地址，如 `http://example.com
↪ :8033/dashboard/` ，即可访问 TiDB Dashboard。

9.9.2.2.2 自定义路径前缀

TiDB Dashboard 默认在 `/dashboard/` 路径下提供服务，即使是反向代理也是如此，例如 `http://example.com
↪ :8033/dashboard/`。若要配置反向代理以非默认的路径提供 TiDB Dashboard 服务，例如 `http://example.com
↪ :8033/foo/` 或 `http://example.com:8033/`，可参考以下步骤。

第 1 步：修改 PD 配置指定 TiDB Dashboard 服务路径前缀

修改 PD 配置中 `[dashboard]` 类别的 `public-path-prefix` 配置项，可指定服务路径前缀。该配置修改后需要重启 PD 实例生效。

以 TiUP 部署且希望运行在 `http://example.com:8033/foo/` 为例，可指定以下配置：

```
server_configs:
  pd:
    dashboard.public-path-prefix: /foo
```

使用 TiUP 部署全新集群时修改配置

若要全新部署集群，可在 TiUP 拓扑文件 `topology.yaml` 中加入上述配置项后进行部署，具体步骤参阅TiUP 部署文档。

使用 TiUP 修改已部署集群的配置

1. 以编辑模式打开该集群的配置文件（将 `CLUSTER_NAME` 替换为集群名称）

```
shell  tiup cluster edit-config CLUSTER_NAME
```

2. 在 server_configs 的 pd 配置下修改或新增配置项，若没有 server_configs 请在最顶层新增：

```
yaml  server_configs: pd:     dashboard.public-path-prefix: /foo
```

修改完成后的配置文件类似于：

```
yaml  server_configs: pd:     dashboard.public-path-prefix: /foo global: user: tidb ...
```

或

```
yaml  monitored:  ...  server_configs: tidb: ...  tikv: ...  pd:     dashboard.public-path-
↪ prefix: /foo ...
```

3. 滚动重启所有 PD 实例生效配置（将 CLUSTER_NAME 替换为集群名称）

```
shell  tiup cluster reload CLUSTER_NAME -R pd
```

详情请参阅TiUP 常见运维操作 - 修改配置参数。

若希望运行在根路径（如 http://example.com:8033/）下，相应的配置为：

```
server_configs:
  pd:
    dashboard.public-path-prefix: /
```

> **警告：**
>
> 修改自定义路径前缀生效后，直接访问将不能正常使用 TiDB Dashboard，您只能通过和路径前缀匹配的反向代理访问。

第 2 步：修改反向代理配置

使用 HAProxy 反向代理

以 http://example.com:8033/foo/ 为例，HAProxy 配置如下：

```
frontend tidb_dashboard_front
  bind *:8033
  use_backend tidb_dashboard_back if { path /foo } or { path_beg /foo/ }

backend tidb_dashboard_back
  mode http
  http-request set-path %[path,regsub(^/foo/?,/dashboard/)]
  server tidb_dashboard 192.168.0.123:2379
```

其中 192.168.0.123:2379 需替换为第 1 步：获取实际 TiDB Dashboard 地址中取得的 TiDB Dashboard 实际地址中的 IP 及端口部分。

> **警告：**
>
> 请务必保留 `use_backend` 指令中的 `if` 部分，确保只有该路径下的服务会被反向代理，否则将引入安全风险。参见提高 TiDB Dashboard 安全性。

若希望运行在根路径（如 `http://example.com:8033/`），HAProxy 配置如下：

```
frontend tidb_dashboard_front
  bind *:8033
  use_backend tidb_dashboard_back

backend tidb_dashboard_back
  mode http
  http-request set-path /dashboard%[path]
  server tidb_dashboard 192.168.0.123:2379
```

修改配置并重启 HAProxy 后即可生效。

使用 NGINX 反向代理

以 `http://example.com:8033/foo/` 为例，相应的 NGINX 配置为：

```
server {
  listen 8033;
  location /foo/ {
    proxy_pass http://192.168.0.123:2379/dashboard/;
  }
}
```

其中 `http://192.168.0.123:2379/dashboard/` 需替换为第 1 步：获取实际 TiDB Dashboard 地址中取得的 TiDB Dashboard 实际地址。

> **警告：**
>
> 请务必保留 `proxy_pass` 指令中的 `/dashboard/` 路径，确保只有该路径下的服务会被反向代理，否则将引入安全风险。参见提高 TiDB Dashboard 安全性。

若希望运行在根路径（如 `http://example.com:8033/`），NGINX 配置为：

```
server {
  listen 8033;
  location / {
    proxy_pass http://192.168.0.123:2379/dashboard/;
  }
```

```
}
```

修改配置并重启 NGINX 后即可生效：

```
sudo nginx -s reload
```

9.9.2.2.3 下一步

参阅提高 TiDB Dashboard 安全性文档了解如何增强 TiDB Dashboard 的安全性，如配置防火墙等。

9.9.2.3 TiDB Dashboard 用户管理

TiDB Dashboard 与 TiDB 使用相同的用户权限体系和登录验证方式。你可以通过控制和管理 TiDB SQL 用户，从而限制和约束这些用户对 TiDB Dashboard 的访问。本文描述了 TiDB SQL 用户访问 TiDB Dashboard 所需的最小权限，并提供了如何创建最小权限 SQL 用户的示例。

要了解如何控制和管理 TiDB SQL 用户，请参见 TiDB 用户账户管理。

9.9.2.3.1 所需权限说明

- 当所连接的 TiDB 服务器未启用安全增强模式 (SEM) 时，要访问 TiDB Dashboard，SQL 用户应当拥有以下所有权限：
  - PROCESS
  - SHOW DATABASES
  - CONFIG
  - DASHBOARD_CLIENT
- 当所连接的 TiDB 服务器启用了安全增强模式 (SEM) 时，要访问 TiDB Dashboard，SQL 用户应当拥有以下所有权限：
  - PROCESS
  - SHOW DATABASES
  - CONFIG
  - DASHBOARD_CLIENT
  - RESTRICTED_TABLES_ADMIN
  - RESTRICTED_STATUS_ADMIN
  - RESTRICTED_VARIABLES_ADMIN
- 若希望 SQL 用户在登录 TiDB Dashboard 后允许修改界面上的各项配置，SQL 用户还应当拥有以下权限：
  - SYSTEM_VARIABLES_ADMIN

> **注意：**
>
> 拥有 ALL PRIVILEGES 或 SUPER 等粗粒度高权限的用户同样可以登录 TiDB Dashboard。出于最小权限原则，强烈建议创建用户时仅使用上述精细权限，从而防止用户执行非预期操作。请参阅权限管理了解这些权限的详细信息。

如果登录 TiDB Dashboard 时指定的 SQL 用户未满足上述权限需求，则登录将失败，如下图所示：

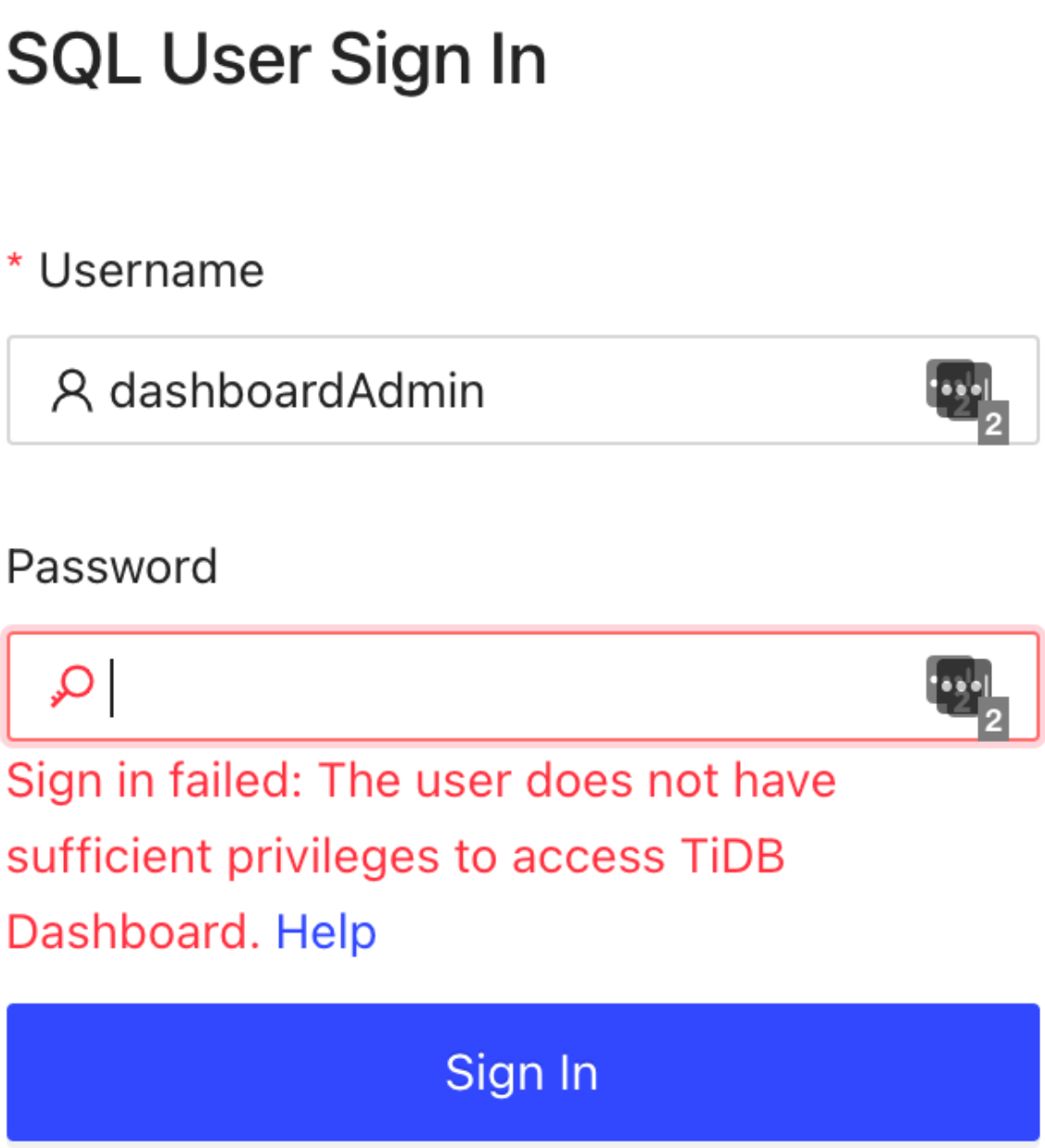


图 76: insufficient-privileges

9.9.2.3.2 示例：创建一个最小权限 SQL 用户用于登录 TiDB Dashboard

- 当所连接的 TiDB 服务器未启用安全增强模式 (SEM) 时，你可以通过执行以下示例 SQL 语句创建一个允许登录 TiDB Dashboard 的 SQL 用户 dashboardAdmin：

```
CREATE USER 'dashboardAdmin'@'%' IDENTIFIED BY '<YOUR_PASSWORD>';
```

```
GRANT PROCESS, CONFIG ON *.* TO 'dashboardAdmin'@'%';
GRANT SHOW DATABASES ON *.* TO 'dashboardAdmin'@'%';
GRANT DASHBOARD_CLIENT ON *.* TO 'dashboardAdmin'@'%';

-- 如果要使自定义的 SQL 用户能修改 TiDB Dashboard 界面上的各项配置，可以增加以下权限
GRANT SYSTEM_VARIABLES_ADMIN ON *.* TO 'dashboardAdmin'@'%';
```

- 当所连接的 TiDB 服务器启用了安全增强模式 (SEM) 时，先关闭 SEM，然后执行以下示例 SQL 语句创建一个允许登录 TiDB Dashboard 的 SQL 用户 dashboardAdmin，创建完成后，再重新开启 SEM：

```
CREATE USER 'dashboardAdmin'@'%' IDENTIFIED BY '<YOUR_PASSWORD>';
GRANT PROCESS, CONFIG ON *.* TO 'dashboardAdmin'@'%';
GRANT SHOW DATABASES ON *.* TO 'dashboardAdmin'@'%';
GRANT DASHBOARD_CLIENT ON *.* TO 'dashboardAdmin'@'%';
GRANT RESTRICTED_STATUS_ADMIN ON *.* TO 'dashboardAdmin'@'%';
GRANT RESTRICTED_TABLES_ADMIN ON *.* TO 'dashboardAdmin'@'%';
GRANT RESTRICTED_VARIABLES_ADMIN ON *.* TO 'dashboardAdmin'@'%';

-- 如果要使自定义的 SQL 用户能修改 TiDB Dashboard 界面上的各项配置，可以增加以下权限
GRANT SYSTEM_VARIABLES_ADMIN ON *.* TO 'dashboardAdmin'@'%';
```

9.9.2.3.3 示例：通过 RBAC 授权 SQL 用户登录 TiDB Dashboard

以下示例演示了如何在基于角色的访问控制（RBAC）机制下创建角色及用户来登录 TiDB Dashboard。

1. 创建一个包含登录 TiDB Dashboard 所需权限的角色 dashboard_access：

```
CREATE ROLE 'dashboard_access';
GRANT PROCESS, CONFIG ON *.* TO 'dashboard_access'@'%';
GRANT SHOW DATABASES ON *.* TO 'dashboard_access'@'%';
GRANT DASHBOARD_CLIENT ON *.* TO 'dashboard_access'@'%';
GRANT SYSTEM_VARIABLES_ADMIN ON *.* TO 'dashboard_access'@'%';
```

2. 为其他用户授权 dashboard_access 角色并设置为默认启用：

```
CREATE USER 'dashboardAdmin'@'%' IDENTIFIED BY '<YOUR_PASSWORD>';
GRANT 'dashboard_access' TO 'dashboardAdmin'@'%';
-- 需要默认启用 dashboard_access 角色
SET DEFAULT ROLE dashboard_access to 'dashboardAdmin'@'%';
```

完成以上步骤后，可以用 dashboardAdmin 用户登录 TiDB Dashboard。

9.9.2.3.4 登录 TiDB Dashboard

创建满足 TiDB Dashboard 权限要求的 SQL 用户后，你可以使用该用户登录 TiDB Dashboard。

9.9.2.4 提高 TiDB Dashboard 安全性

尽管访问 TiDB Dashboard 需要登录，但它被设计为默认供可信的用户实体访问。当你希望将 TiDB Dashboard 提供给外部网络用户或不可信用户访问时，需要注意采取以下措施以避免安全漏洞。

9.9.2.4.1 提高 TiDB 用户安全性

为 root 用户设置强密码

TiDB Dashboard 的账号体系与 TiDB SQL 用户一致。默认部署情况下，TiDB 的 root 用户没有密码，因而访问 TiDB Dashboard 也不需要密码验证。这将会给恶意访问者极大的集群权限，包括执行特权 SQL 等。

建议的措施：

- 为 TiDB 的 root 用户设置一个强密码（请参见 TiDB 用户账户管理了解详情），或禁用 root 账户。

为 TiDB Dashboard 创建最小权限用户

TiDB Dashboard 的账号体系与 TiDB SQL 用户一致，并基于 TiDB SQL 用户的权限进行 TiDB Dashboard 授权验证。TiDB Dashboard 所需的权限较少，甚至可以只有只读权限。可以基于最小权限原则配置合适的用户访问 TiDB Dashboard，减少高权限用户的使用场景。

建议的措施：

- 为访问 TiDB Dashboard 创建一个最小权限的 SQL 用户，并用该用户登录 TiDB Dashboard，避免使用高权限用户，提升安全性。请参见TiDB Dashboard 用户管理了解详情。

9.9.2.4.2 使用防火墙阻止不可信访问

TiDB Dashboard 通过 PD Client 端口提供服务，默认为 http://IP:2379/dashboard/。尽管 TiDB Dashboard 需要验证身份，但 PD Client 端口上承载的其他 PD 内部特权接口不需要验证身份，且能进行特权操作，例如 http://IP:2379/pd/api/v1/members。因此，将 PD Client 端口直接暴露给外部网络具有极大的风险。

建议的措施：

1. 使用防火墙禁止组件外部网络或不可信网络访问任何 PD 组件的 Client 端口。

注意，TiDB、TiKV 等组件需要通过 PD Client 端口与 PD 组件进行通信，因此请勿对组件内部网络阻止访问，这将导致集群不可用。

2. 参见通过反向代理使用 TiDB Dashboard 了解如何配置反向代理将 TiDB Dashboard 服务在另一个端口上安全地提供给外部网络。

如何在多 PD 实例部署时开放 TiDB Dashboard 端口访问

> 警告：
>
> 本章节描述了一个不安全的访问方案，仅供测试环境使用。在生产环境中，不要使用本方案。请参见本文的其余章节。

在测试环境中，您可能需要配置防火墙开放 TiDB Dashboard 端口供外部访问。

当部署了多个 PD 实例时，其中仅有一个 PD 实例会真正运行 TiDB Dashboard，访问其他 PD 实例时会发生浏览器重定向，因此需要确保防火墙配置了正确的 IP 地址。关于该机制的详情，可参阅TiDB Dashboard 多 PD 实例部署章节。

使用 TiUP 部署工具时，可使用以下命令查看实际运行 TiDB Dashboard 的 PD 实例地址（将 CLUSTER_NAME 替换为集群名称）：

```
tiup cluster display CLUSTER_NAME --dashboard
```

输出即为实际 TiDB Dashboard 地址。

> 注意：
>
> 该功能在 TiUP Cluster v1.0.3 或更高版本部署工具中提供。
>
> 升级 TiUP Cluster 步骤
>
> ```
> tiup update --self
> tiup update cluster --force
> ```

以下是一个样例输出：

```
http://192.168.0.123:2379/dashboard/
```

在这个样例中，需要为防火墙配置开放 IP 192.168.0.123 的 2379 端口入站访问，并通过 http://192.168.0.123:2379/dashboard/ 访问 TiDB Dashboard。

9.9.2.4.3 反向代理仅代理 TiDB Dashboard

如前文所述，PD Client 端口下提供的服务不仅有 TiDB Dashboard（位于 http://IP:2379/dashboard/），还有其他 PD 内部特权接口（如 http://IP:2379/pd/api/v1/members）。因此，使用反向代理将 TiDB Dashboard 提供给外部网络时，应当确保仅提供 /dashboard 前缀下的服务，而非该端口下所有服务，避免外部网络能通过反向代理访问到 PD 内部特权接口。

建议的措施：

- 参见通过反向代理使用 TiDB Dashboard 了解安全且推荐的反向代理配置。

9.9.2.4.4 为反向代理开启 TLS

为了进一步加强传输层安全性，可以为反向代理开启 TLS，甚至可以引入 mTLS 实现访问用户的证书验证。

请参阅 NGINX 文档或 HAProxy 文档了解如何为它们开启 TLS。

9.9.2.4.5 其他建议的安全措施

- 为 TiDB 组件间通信开启加密传输
- 为 TiDB 客户端服务端间通信开启加密传输

9.9.3 访问 TiDB Dashboard

通过浏览器访问 http://127.0.0.1:2379/dashboard/（将 127.0.0.1:2379 替换为实际 PD 实例的地址和端口）即可打开 TiDB Dashboard。

9.9.3.1 多 PD 实例访问

当集群中部署有多个 PD 实例、且您可以直接访问到每个 PD 实例地址和端口时，可以简单地将 http://127.0.0.1:2379/dashboard/ 地址中的 127.0.0.1:2379 替换为集群中任意一个 PD 实例的地址和端口进行访问。

> **注意：**
>
> 当处于防火墙或反向代理等环境下、无法直接访问每个 PD 实例时，可能会无法访问 TiDB Dashboard。这通常是防火墙或反向代理没有正确配置导致的。可参阅通过反向代理使用 TiDB Dashboard 或提高 TiDB Dashboard 安全性章节了解如何在多 PD 实例情况下正确配置防火墙或反向代理规则。

9.9.3.2 浏览器兼容性

TiDB Dashboard 可在常见的、更新及时的桌面浏览器中使用，具体版本号为：

- Chrome >= 77
- Firefox >= 68
- Edge >= 17

> **注意：**
>
> 若使用旧版本浏览器或其他浏览器访问 TiDB Dashboard，部分界面可能不能正常工作。

9.9.3.3 登录

访问 TiDB Dashboard 将会显示用户登录界面。

- 可使用 TiDB 的 root 用户登录。
- 如果创建了自定义 SQL 用户，也可以使用自定义的 SQL 用户和密码登录。

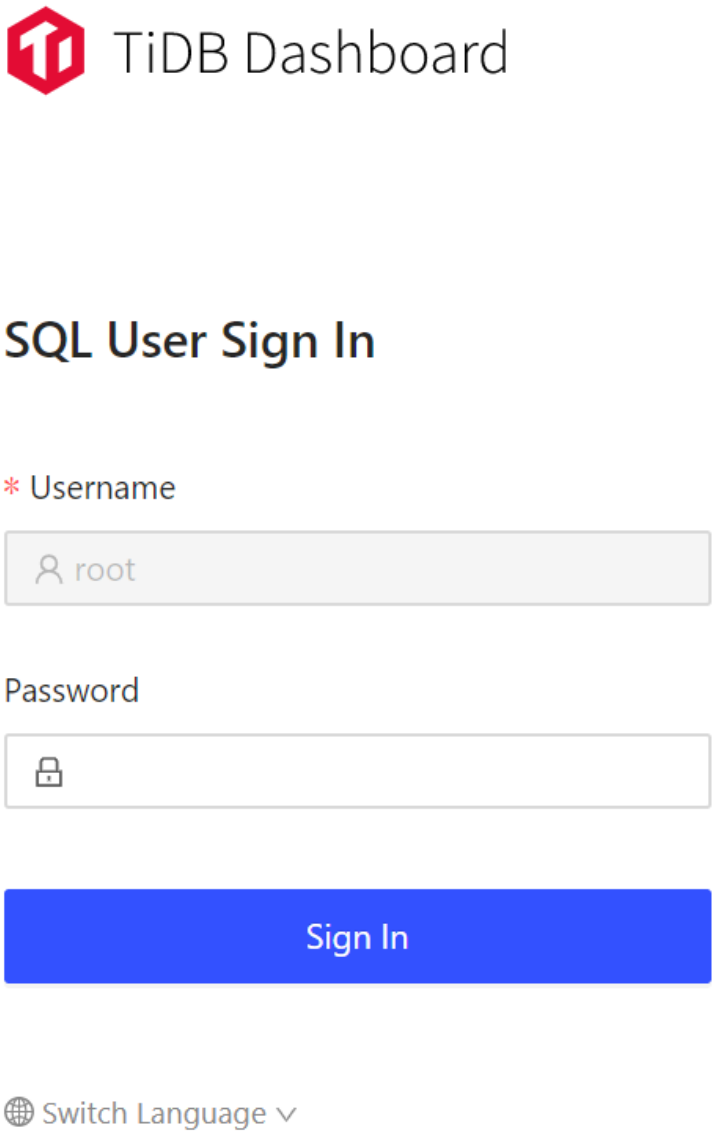


图 77: 登录界面

如果存在以下情况，则可能会登录失败：

- TiDB root 用户不存在
- PD 未启动或无法访问
- TiDB 未启动或无法访问
- root 密码错误

登录后，24 小时内将保持自动登录状态。参见登出章节了解如何登出用户。

9.9.3.4 切换语言

TiDB Dashboard 目前支持以下语言：

- 简体中文
- 英文

在登录界面中，可点击 Switch Language 下拉框切换界面显示语言：

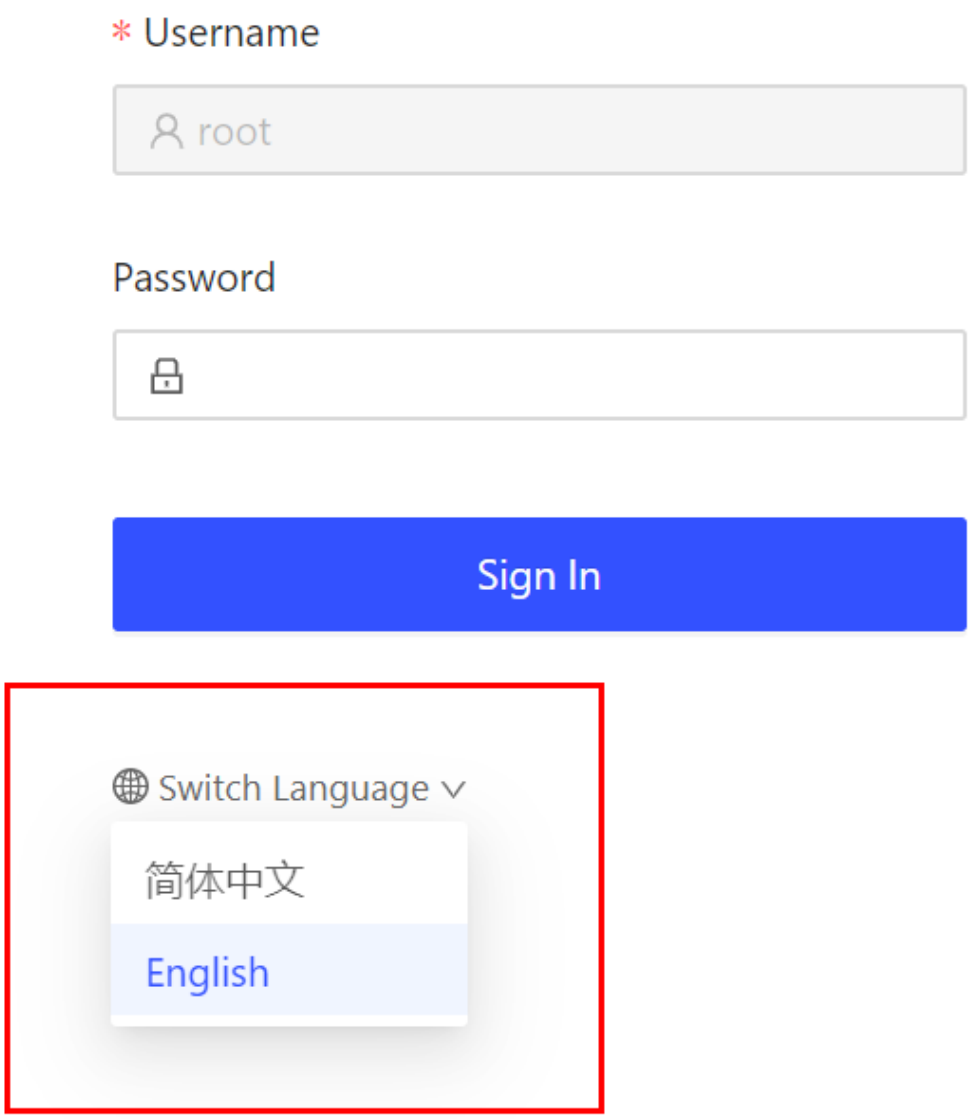


图 78: 切换语言

9.9.3.5 登出

登录后，在左侧导航处点击登录用户名，可切换到用户页面。在用户页面点击登出 (Logout) 按钮即可登出当前用户。登出后，需重新输入用户名密码。

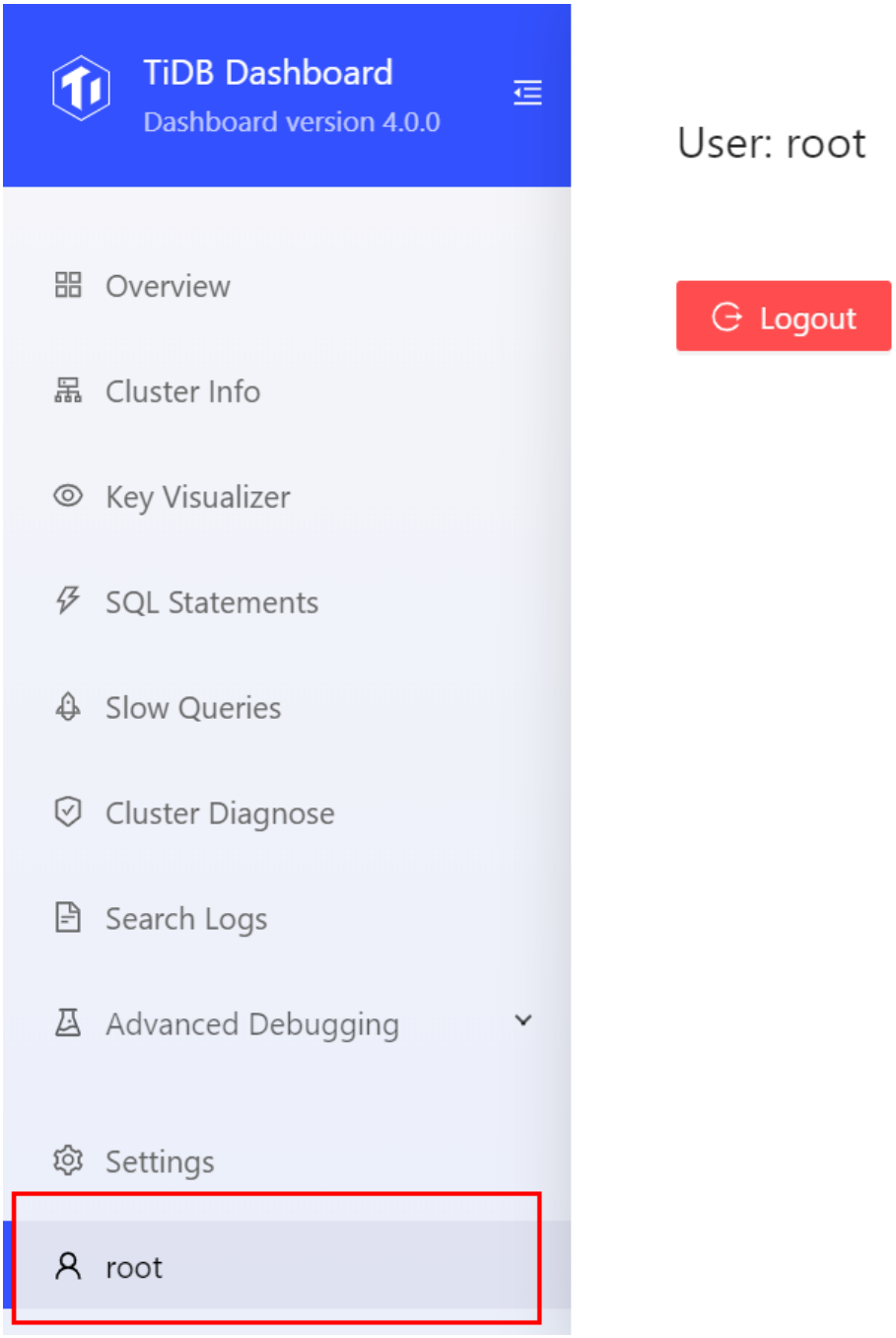


图 79: 登出

9.9.4 TiDB Dashboard 概况页面

该页面显示了整个集群的概况，包含以下信息：

- 整个集群的 QPS
- 整个集群的查询延迟

- 最近一段时间内累计耗时最多的若干 SQL 语句
- 最近一段时间内运行时间超过一定阈值的慢查询
- 各个实例的节点数和状态
- 监控及告警信息

9.9.4.1 访问

登录 TiDB Dashboard 后默认进入该页面，也可以左侧导航条点击概况 (Overview) 进入：



图 80: 访问

9.9.4.2 QPS

该区域显示最近一小时整个集群的每秒成功和失败查询数量：

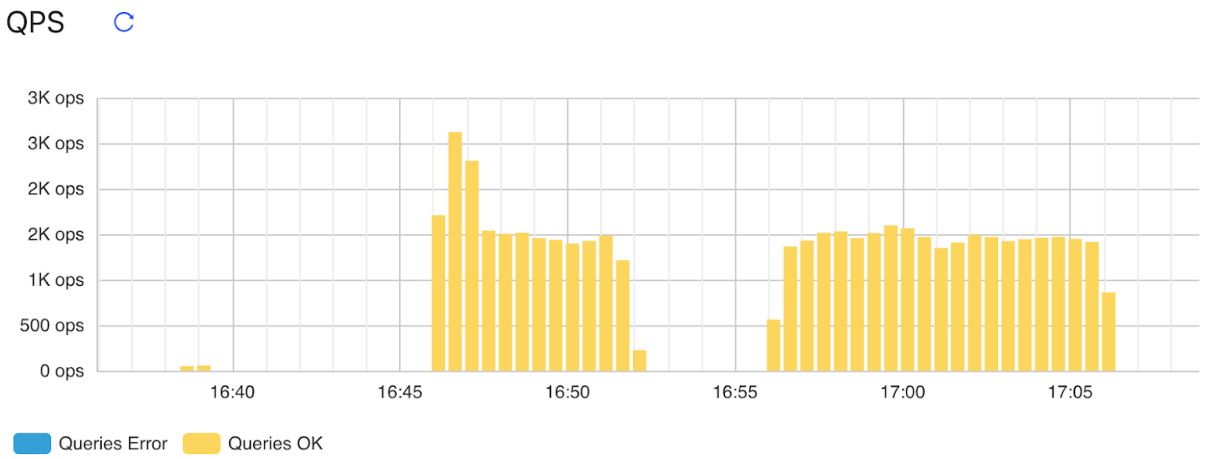


图 81: 界面

> **注意：**
>
> 该功能仅在部署了 Prometheus 监控组件的集群上可用，未部署监控组件的情况下会显示为失败。

9.9.4.3 延迟

该区域显示最近一小时整个集群中 99.9%、99% 和 90% 查询的延迟：

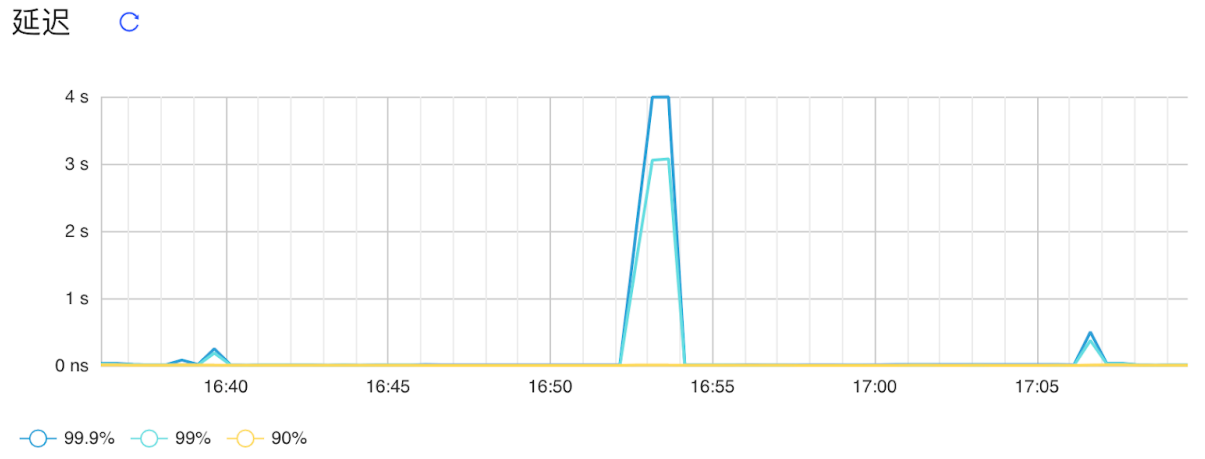


图 82: 界面

> 注意：
>
> 该功能仅在部署了 Prometheus 监控组件的集群上可用，未部署监控组件的情况下会显示为失败。

9.9.4.4 Top SQL 语句

该区域显示最近一段时间内整个群集中累计耗时最长的 10 类 SQL 语句。查询参数不一样但结构一样的 SQL 会归为同一类 SQL 语句，在同一行中显示：

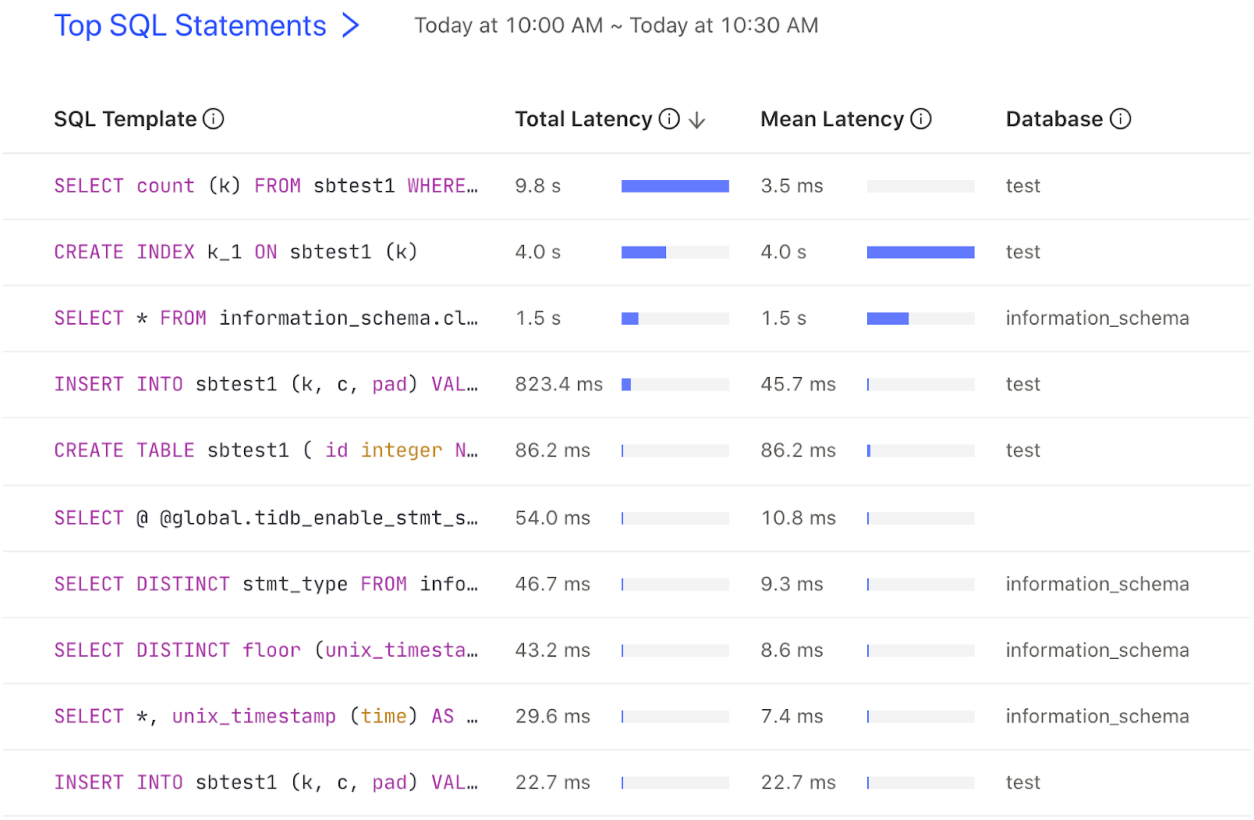


图 83: 界面

该区域显示的内容与SQL 语句分析页面一致，可点击 Top SQL 语句 (Top SQL Statements) 标题查看完整列表。关于该表格中各列详情，见SQL 语句分析页面。

> 注意：
>
> 该功能仅在开启了 SQL 语句分析功能的集群上可用。

9.9.4.5 最近的慢查询

该区域默认显示最近 30 分钟内整个集群中最新的 10 条慢查询：



图 84: 界面

默认情况下运行时间超过 300ms 的 SQL 查询即会被计为慢查询并显示在该表格中。可通过调整 tidb_slow_log_threshold 变量或 TiDB slow-threshold 参数调整阈值。

该区域显示的内容与慢查询页面一致，可点击最近的慢查询 (Recent Slow Queries) 标题查看完整列表。关于该表格中各列详情，见慢查询页面。

> **注意：**
>
> 该功能仅在配置开启了慢查询日志的集群中可用，使用 TiUP 部署的集群默认开启慢查询日志。

9.9.4.6 实例

该区域汇总显示了整个集群中 TiDB、TiKV、PD、TiFlash 的总实例数量及异常实例数量：

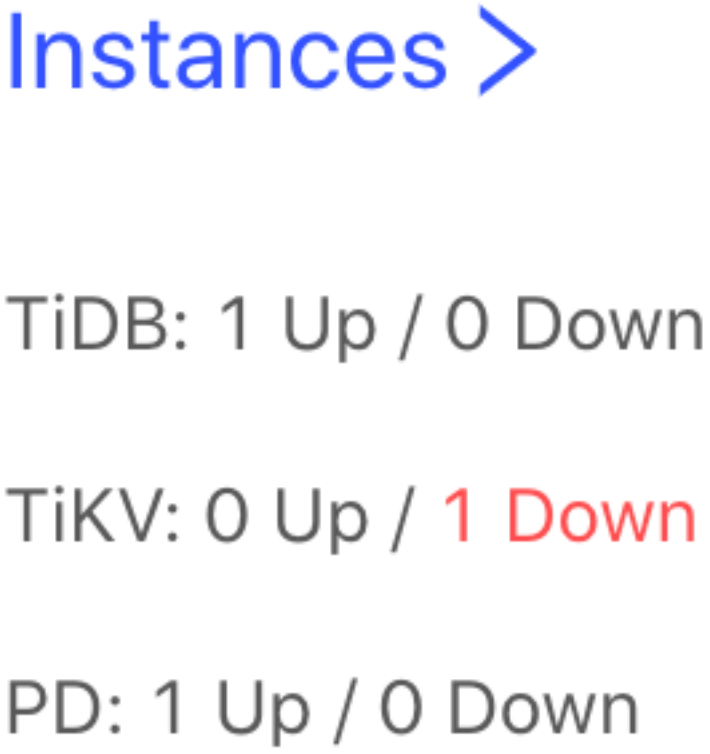


图 85: 界面

状态描述如下：

- Up：实例运行正常（含下线中的存储实例）。
- Down：实例运行异常，例如网络无法连接、进程已崩溃等。

点击实例标题可进入集群信息页面查看各个实例的详细运行状态。

9.9.4.7 监控和告警

该区域提供了便捷的链接方便用户查看详细监控或告警：

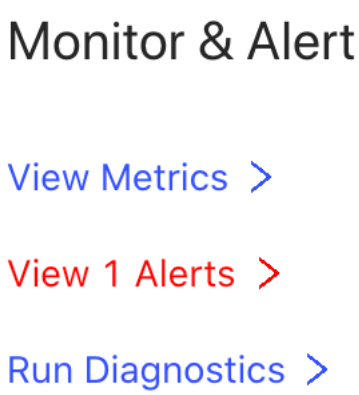


图 86: 界面

- 查看监控链接：点击后跳转至 Grafana 页面，可查看集群详细监控信息。关于 Grafana 监控面板中各个详细监控指标的解释，参见监控指标文档。
- 查看告警链接：点击后跳转至 AlertManager 页面，可查看集群详细告警信息。当集群中已有告警时，告警数量将会直接显示在链接文本上。
- 运行诊断链接：点击后跳转至集群诊断页面，参见集群诊断页面了解详情。

> 注意：
>
> 查看监控链接仅在集群中部署了 Grafana 节点时可用，查看告警链接仅在集群中部署了 AlertManager 节点时可用。

9.9.5 TiDB Dashboard 集群信息页面

该页面上允许用户查看整个集群中 TiDB、 TiKV、 PD、 TiFlash 组件的运行状态及其所在主机的运行状态。

9.9.5.1 访问

可以通过以下两种方法访问集群信息页面：

- 登录后，左侧导航条点击集群信息：

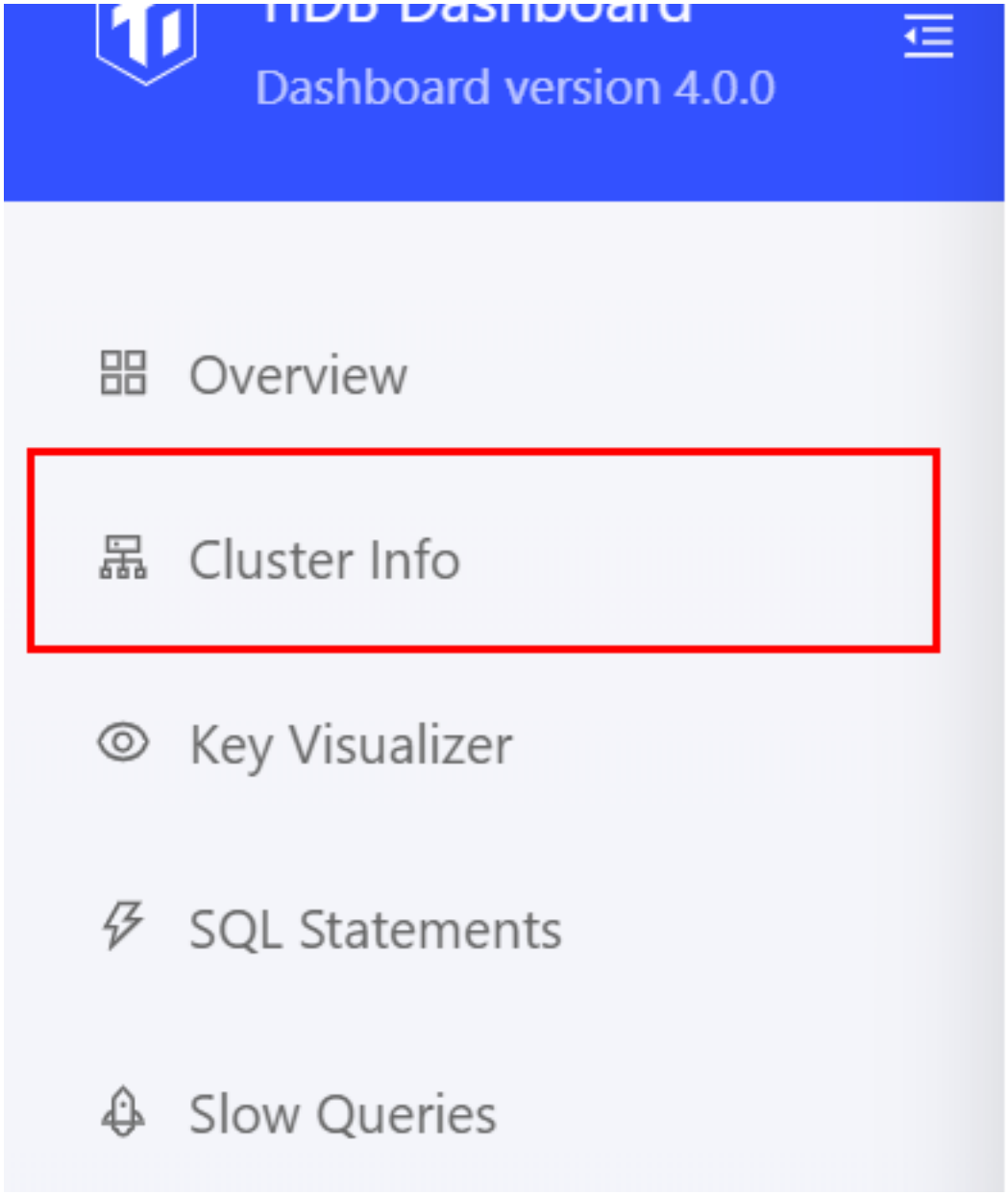


图 87: 访问

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/cluster_info/instance（将 127.0.0.1:2379 替换为实际 PD 实例地址和端口）。

9.9.5.2 实例列表

点击实例可查看实例列表：

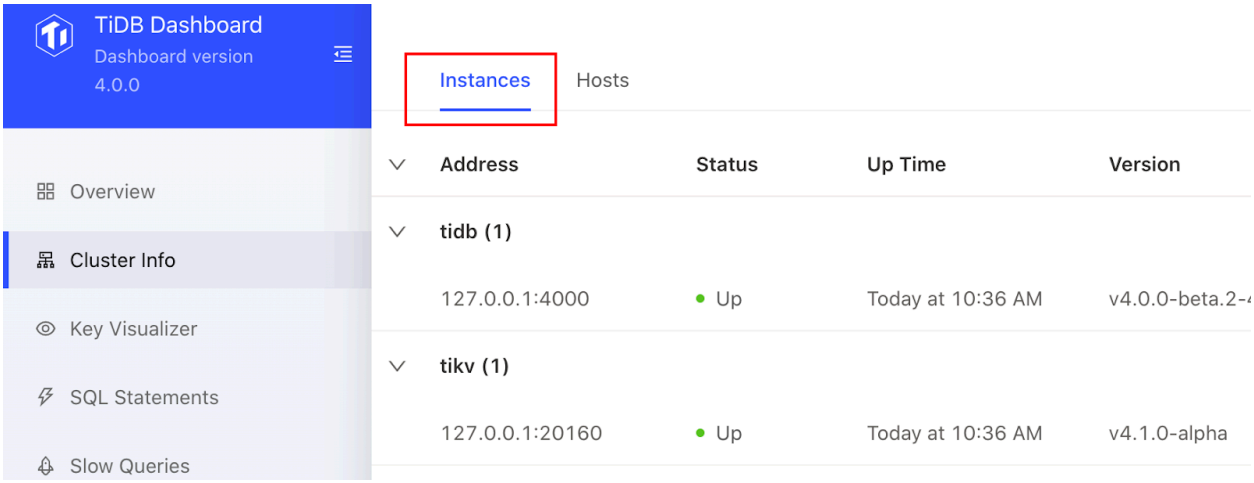


图 88: 实例

实例列表列出了该集群中 TiDB、 TiKV、 PD 和 TiFlash 组件所有实例的概况信息。

表格包含以下列：

- 地址：实例地址
- 状态：实例的运行状态
- 启动时间：实例的启动时间
- 版本：实例版本号
- 部署路径：实例二进制文件所在目录路径
- Git 哈希值：实例二进制对应的 Git 哈希值

实例的运行状态有：

- 在线 (Up)：实例正常运行。
- 离线 (Down) 或无法访问 (Unreachable)：实例未启动或对应主机存在网络问题。
- 已缩容下线 (Tombstone)：实例上的数据已被完整迁出并缩容完毕。仅 TiKV 或 TiFlash 实例存在该状态。
- 下线中 (Offline)：实例上的数据正在被迁出并缩容。仅 TiKV 或 TiFlash 实例存在该状态。
- 未知 (Unknown)：未知的实例运行状态。

> 注意：
>
> 表格中部分列仅在实例处于在线状态时能显示。

9.9.5.3 主机列表

点击主机可查看主机列表：

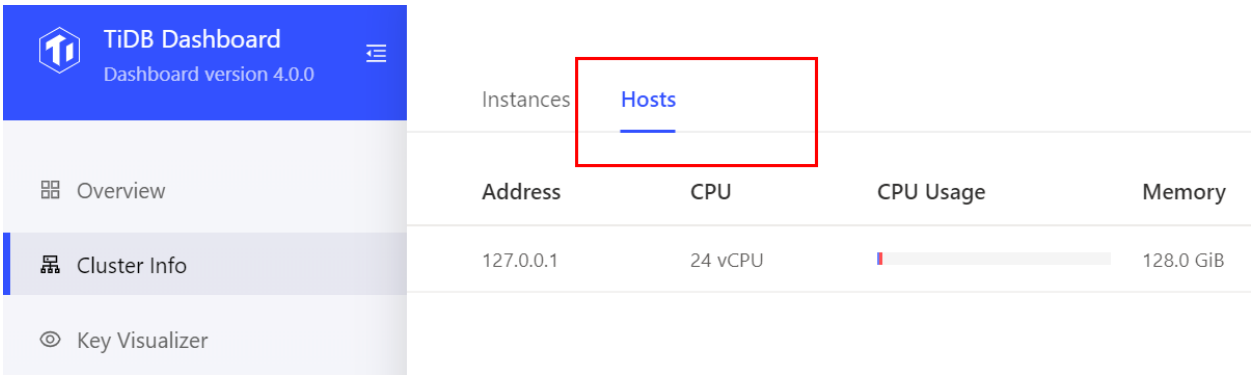


图 89: 主机

主机列表列出了该集群中 TiDB、 TiKV、 PD 和 TiFlash 组件所有实例对应主机的运行情况。

表格包含以下列：

- 地址：主机 IP 地址
- CPU：主机 CPU 逻辑核心数
- CPU 使用率：主机当前 1 秒的用户态和内核态 CPU 使用率
- 物理内存：主机总计的物理内存大小
- 内存使用率：主机当前内存使用率
- 部署磁盘：主机上运行实例所在磁盘的文件系统和磁盘挂载路径
- 磁盘使用率：主机上运行实例所在磁盘的空间使用率

> 注意：
>
> 主机列表信息由各个实例进程给出，因此当主机上所有实例都处于离线状态时，该主机信息将无法显示。

9.9.6 Top SQL

> 警告：
>
> Top SQL 目前为实验特性，不建议在生产环境中使用。

本文介绍如何使用 Top SQL 找到一段时间内对某个 TiDB 或 TiKV 节点消耗负载较大的 SQL 查询。例如，你可以通过 Top SQL 找出一个低负载的数据库上执行的一条消耗 99% 负载的分析查询。

针对指定的 TiDB 或 TiKV 节点，Top SQL 可以提供以下功能：

- 展示指定时间范围内消耗负载最多的 5 类 SQL 查询。
- 展示某一查询的 CPU 使用量、每秒请求数、平均延迟、查询计划等信息，作为改善业务的潜在性能优化点。

9.9.6.1 启用 Top SQL

Top SQL 功能默认关闭。你可以通过以下方法在整个集群范围内启用该功能：

- 方法一：登录 TiDB Dashboard，点击左侧面板中的 Top SQL，然后点击页面右上角的齿轮按钮，并打开 Top SQL 功能开关。
- 方法二：配置 TiDB 系统变量`tidb_enable_top_sql`的值为 `ON`。

> **注意：**
>
> 启用 Top SQL 会对集群的性能产生轻微的影响。

9.9.6.2 使用 Top SQL

开启 Top SQL 后，只需要登录 TiDB Dashboard，然后点击左侧面板中的 Top SQL 即可使用 Top SQL。

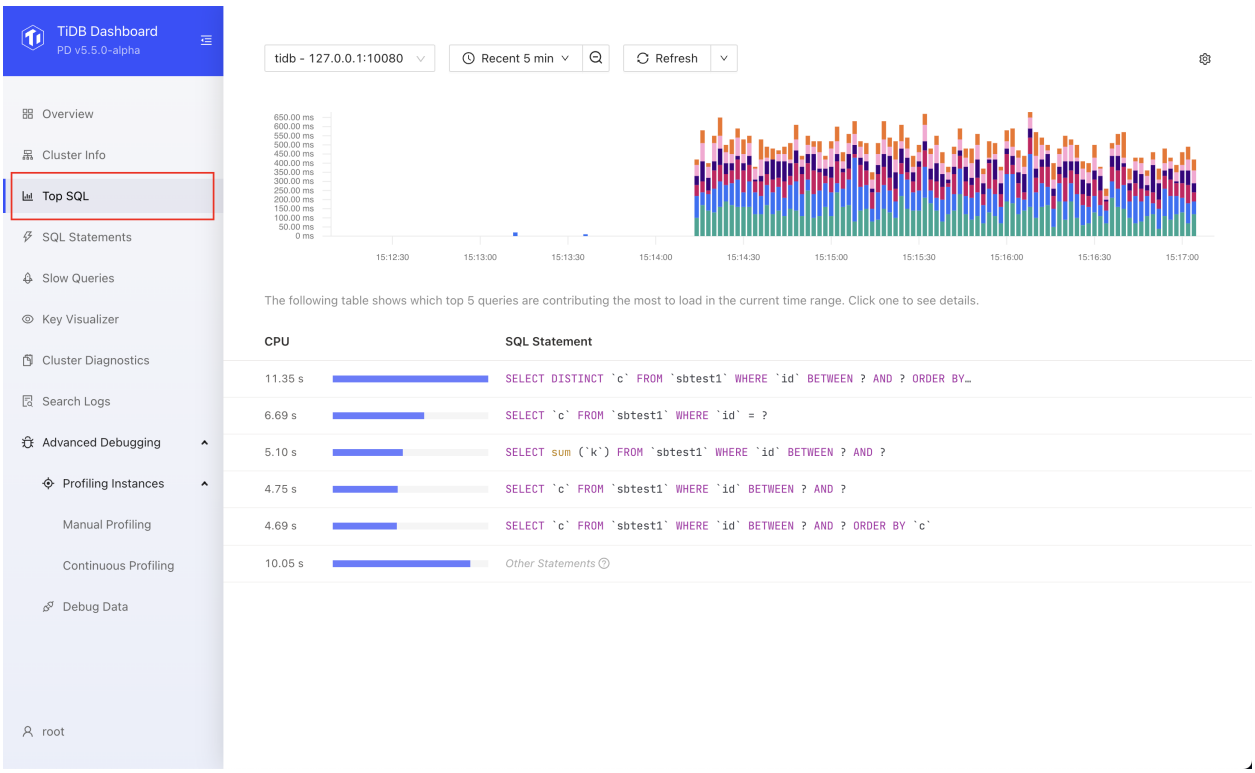


图 90: Top SQL

使用提示：

- 你可以在页面顶部的下拉列表里选择要展示的节点和时间范围，也可以在图表中选择时间范围。
- 如果图表中显示的数据已过时，你可以点击刷新 (Refresh) 按钮，也可以在刷新 (Refresh) 下拉列表中选择是否自动刷新，以及自动刷新间隔。
- 列表中展示了选中节点、选中时间范围内消耗负载最多的 5 类查询。
- 点击选中列表中的某个查询类型，可以查看这类查询在这个节点上的执行计划，以及详细执行信息，例如 Call/sec （平均每秒请求数）、Scan Rows/sec （平均每秒扫描行数）、Scan Indexes/sec （平均每秒扫描索引数）、Latency/call （平均延迟）。

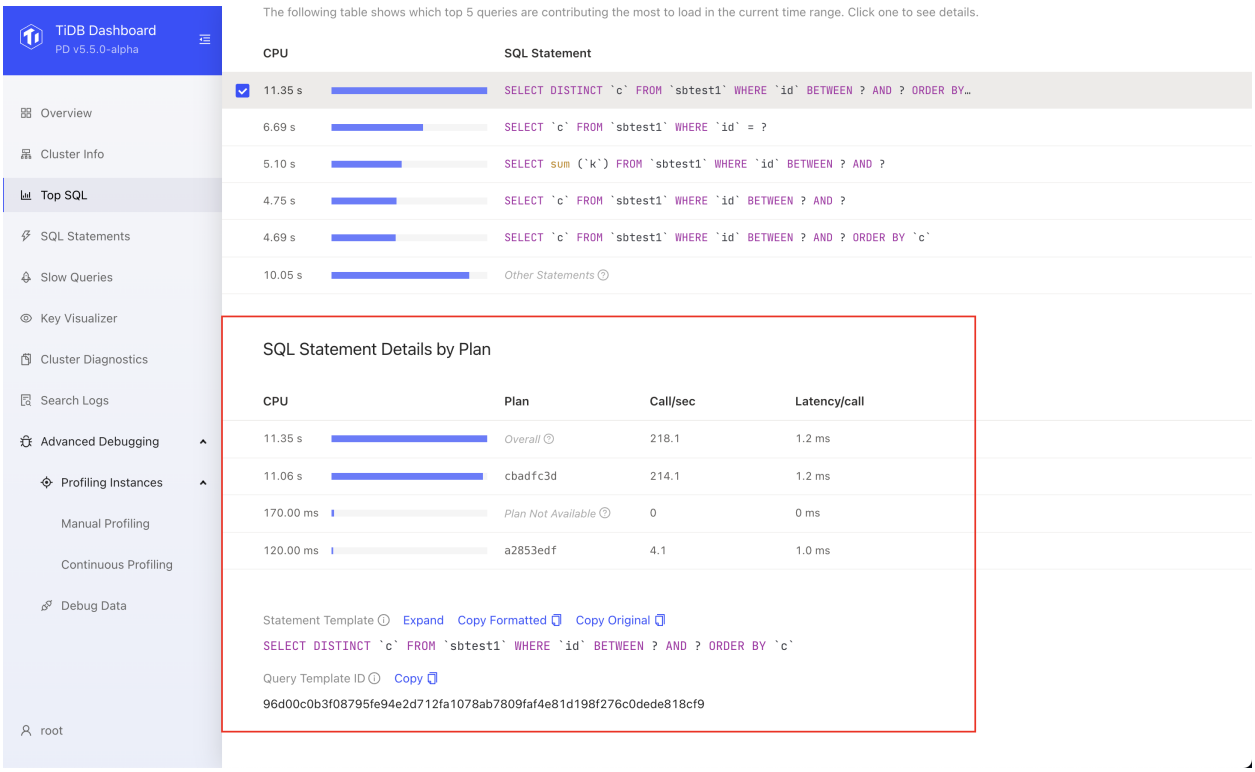


图 91: Top SQL Details

9.9.7 TiDB Dashboard 流量可视化页面

流量可视化页面 (Key Visualizer) 可用于分析 TiDB 集群的使用模式和排查流量热点。该页面可视化地呈现了 TiDB 集群一段时间的流量情况。

9.9.7.1 访问页面

可以通过以下两种方法访问 Key Visualizer 流量可视化页面：

- 登录 TiDB Dashboard 后，点击左侧导航条的 Key Visualizer （流量可视化）：

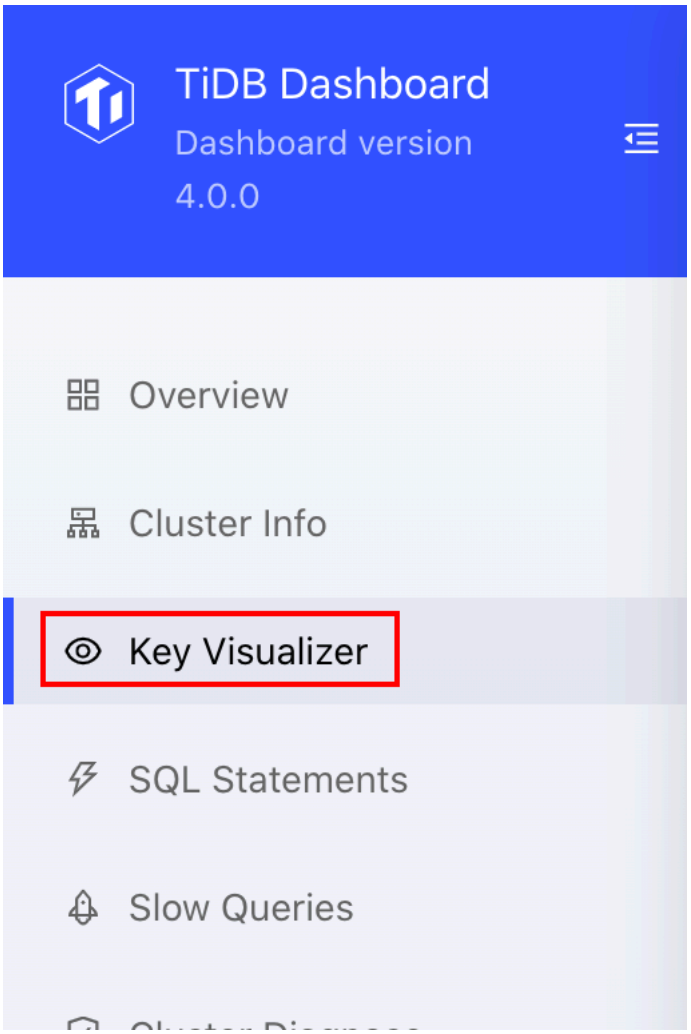


图 92: 访问

- 在浏览器中访问 `http://127.0.0.1:2379/dashboard/#/keyviz`（将 `127.0.0.1:2379` 替换为实际 PD 实例地址和端口）。

9.9.7.2 界面示例

流量可视化页面示例如下：

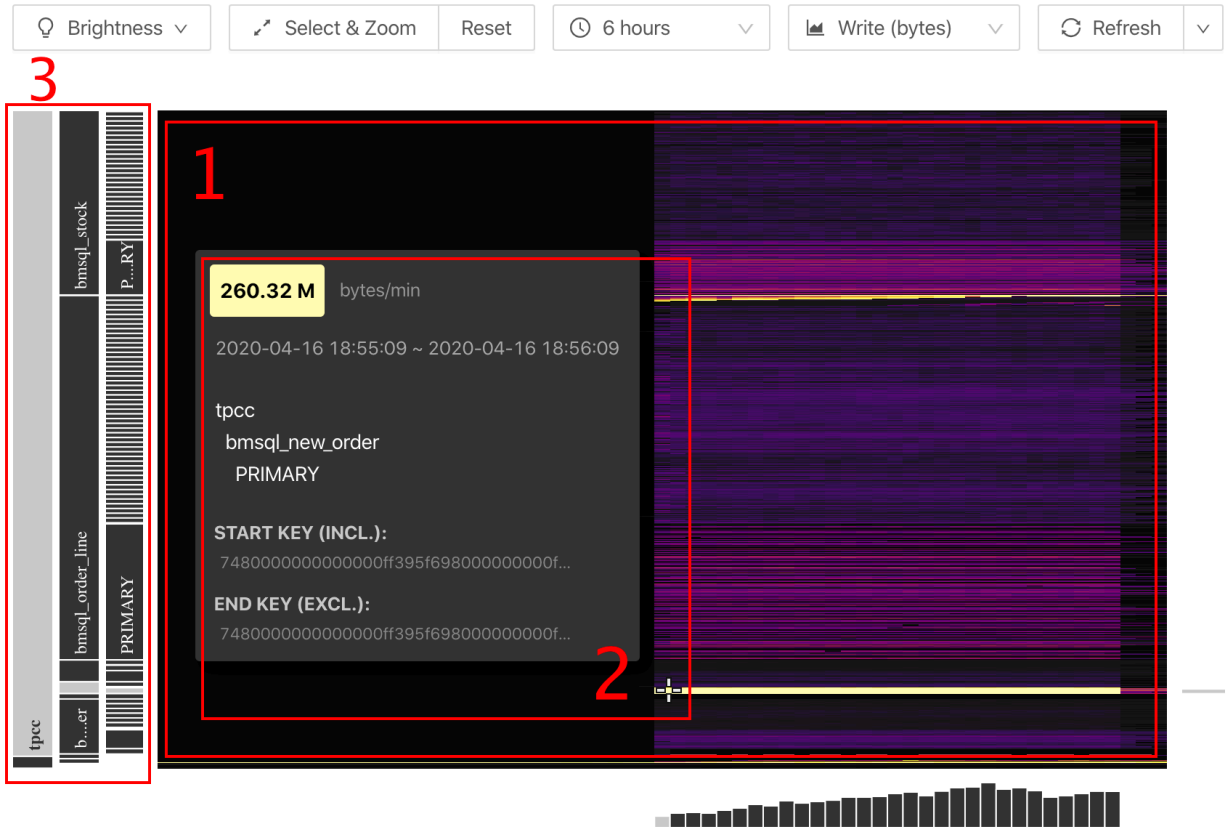


图 93: Key Visualizer 示例图

从以上流量可视化界面可以观察到以下信息：

- 一个大型热力图，显示整体访问流量随时间的变化情况。
- 热力图某个坐标的详细信息。
- 左侧为表、索引等标识信息。

9.9.7.3 基本概念

本节介绍流量可视化涉及的一些基本概念。

9.9.7.3.1 Region

在 TiDB 集群中，数据以分布式的方式存储在所有的 TiKV 实例中。TiKV 在逻辑上是一个巨大且有序的 KV Map。整个 Key-Value 空间分成很多 Region，每一个 Region 是一系列连续的 Key。

> 注意：
>
> 关于 Region 的详细介绍，请参考三篇文章了解 TiDB 技术内幕 - 说存储

9.9.7.3.2 热点

在使用 TiDB 的过程中，热点是一个典型的现象，它表现为大量的流量都在读写一小块数据。由于连续的数据往往由同一个 TiKV 实例处理，因此热点对应的 TiKV 实例的性能就成为了整个业务的性能瓶颈。常见的热点场景有使用自增主键连续写入相邻数据导致的写入表数据热点、时间索引下写入相邻时间数据导致的写入表索引热点等。

> 注意：
>
> 热点问题详情请参阅 TiDB 热点问题详解。

9.9.7.3.3 热力图

热力图是流量可视化页面的核心，它显示了一个指标随时间的变化。热力图的横轴 X 是时间，纵轴 Y 则是按 Key 排序的连续 Region，横跨 TiDB 集群上所有数据库和数据表。颜色越暗 (cold) 表示该区域的 Region 在这个时间段上读写流量较低，颜色越亮 (hot) 表示读写流量越高，即越热。

9.9.7.3.4 Region 压缩

一个 TiDB 集群中，Region 的数量可能多达数十万。在屏幕上是难以显示这么多 Region 信息的。因此，在一张热力图中，Region 会被压缩到约 1500 个连续范围，每个范围称为 Bucket。在一个热力图上，热的实例更需要关注，因此 Region 压缩总是倾向于将流量较小的大量 Region 压缩为一个 Bucket，而尽量让高流量的 Region 独立成 Bucket。

9.9.7.4 使用介绍

本节介绍如何使用流量可视化页面。

9.9.7.4.1 设置

首次使用流量可视化页面需要先通过设置页面手动开启此功能。参考页面指引，点击 Open Settings（打开设置）即可打开设置页面：

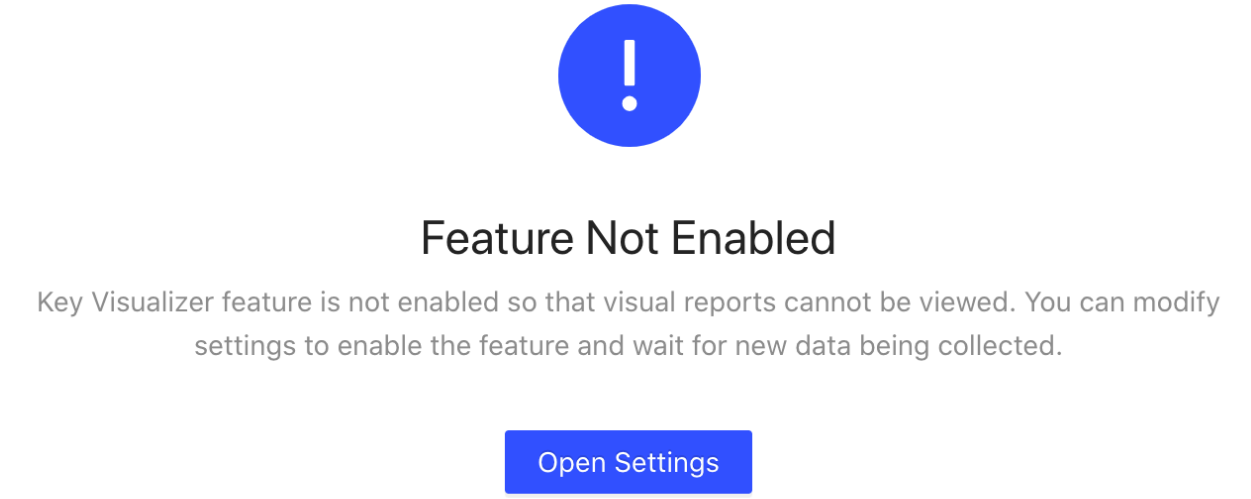


图 94: 功能未开启

在功能已经开启时，可通过右上角的设置图标打开设置页面：

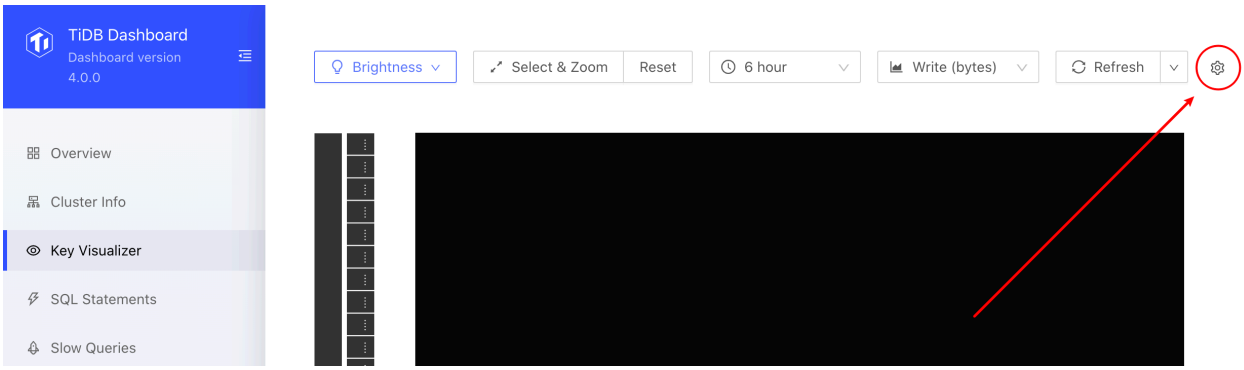


图 95: 设置按钮

设置页面如下图所示：

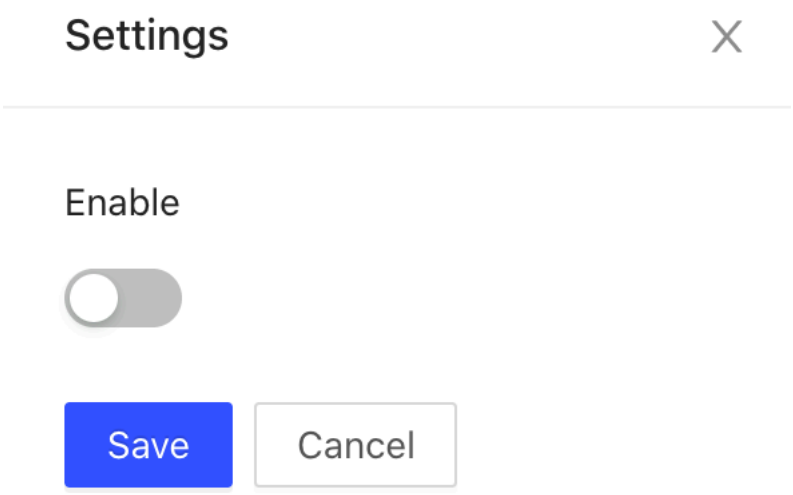


图 96: 设置页面

通过开关设定好是否开启收集，并点击 Save（保存）后生效。开启后，在界面上观察到工具栏已经可以使用：

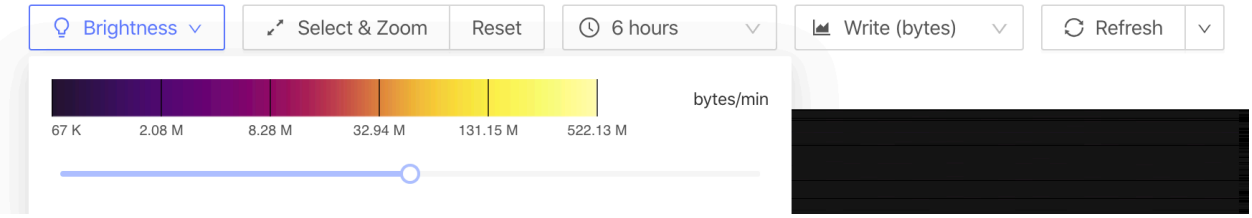


图 97: 工具栏

功能开启后，后台会持续收集数据，稍等一段时间即可看到热力图。

9.9.7.4.2 观察某一段时间或者 Region 范围

打开流量可视化页面时，默认会显示最近六小时整个数据库内的热力图。其中，越靠近右侧（当前时间）时，每列 Bucket 对应的时间间隔会越小。如果你想观察某个特定时间段或者特定的 Region 范围，则可以通过放大来获得更多细节。具体操作描述如下：

- 在热力图中向上或向下滚动。
- 点击 Select & Zoom（框选）按钮，然后点击并拖动以选择要放大的区域。

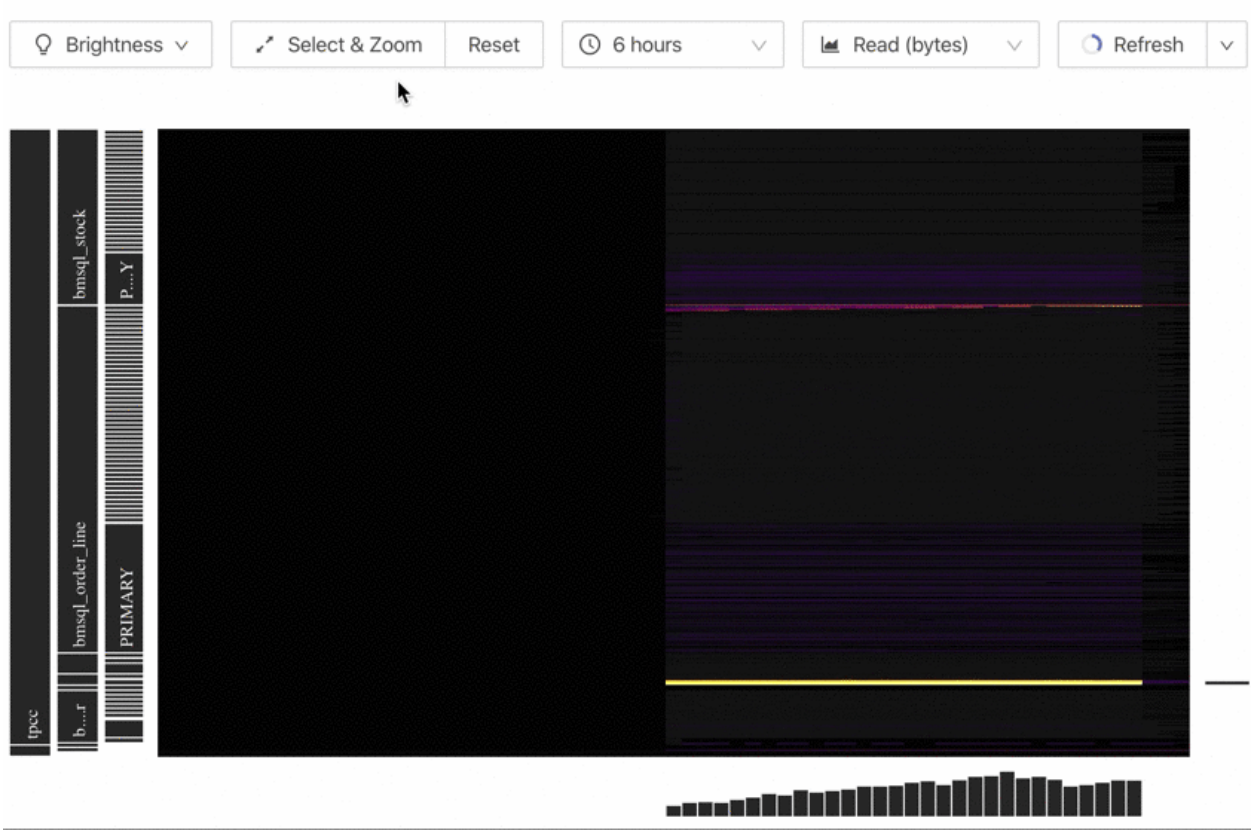


图 98: 框选

- 点击 Reset（重置）按钮，将 Region 范围重置为整个数据库。
- 点击「时间选择框」(界面的 6 hours 处)，重新选择观察时间段。

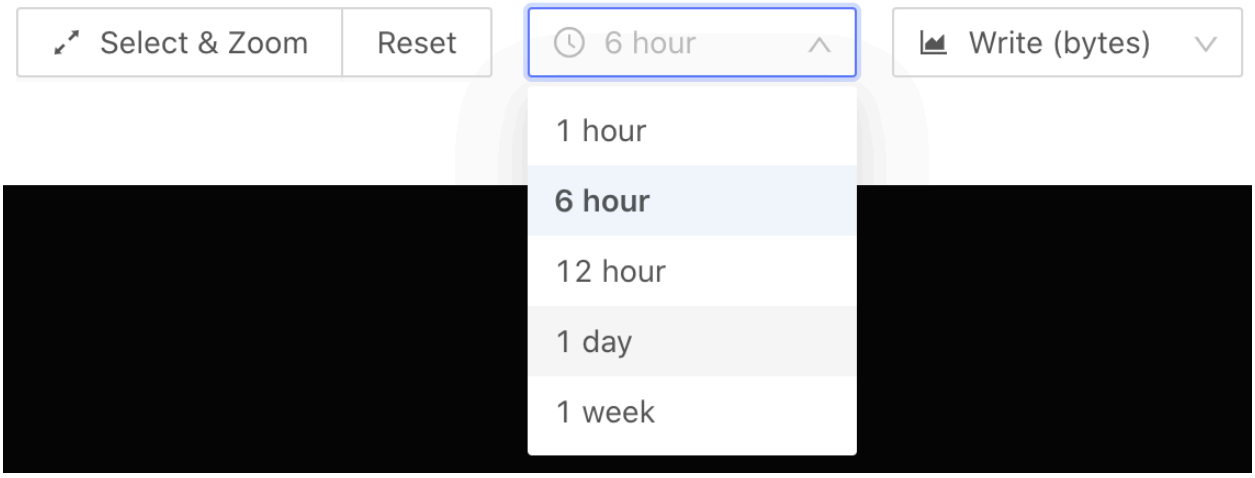


图 99: 时间选择

> 注意：
>
> 进行后三种操作，将引起热力图的重新绘制。你可能观察到热力图与放大前有较大差异。这是一个正常的现象，可能是由于在进行局部观察时，Region 压缩的粒度发生了变化，或者是局部范围内，“热”的基准发生了改变。

9.9.7.4.3 调整亮度

热力图使用颜色的明暗来表达一个 Bucket 的流量高低，颜色越暗 (cold) 表示该区域的 Region 在这个时间段上读写流量较低，颜色越亮 (hot) 表示读写流量越高，即越热。如果热力图中的颜色太亮或太暗，则可能很难观察到细节。此时，可以点击 Brightness（调整亮度）按钮，然后通过滑块来调节页面的亮度。

> 注意：
>
> 在显示一个区域内的热力图时，会根据区域内的流量情况来界定冷热。当整个区域流量较为平均时，即使整体流量在数值上很低，你依然有可能观察到较大的亮色区域。请注意一定要结合数值一起分析。

9.9.7.4.4 选择指标

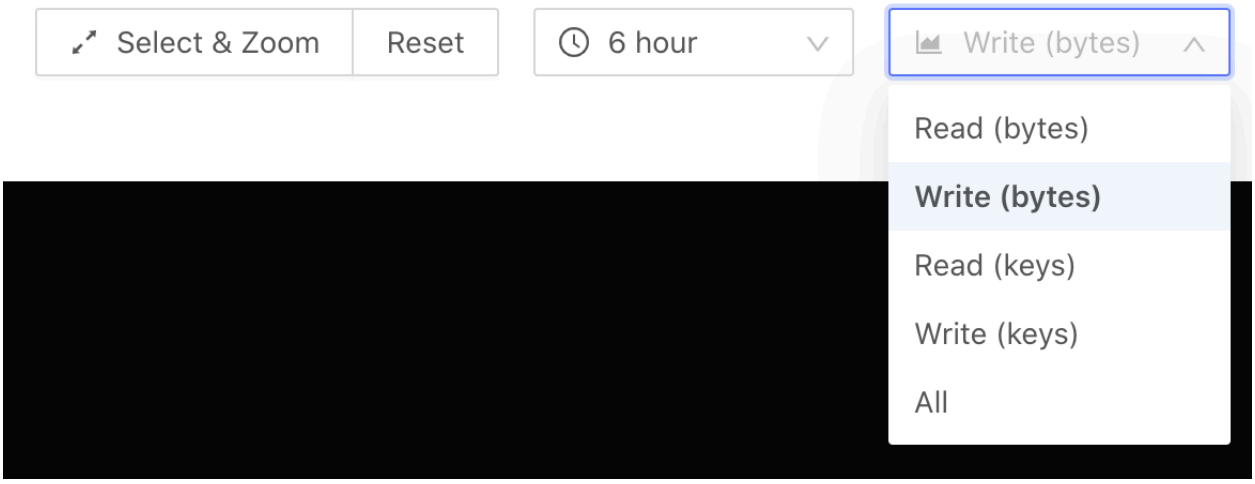


图 100: 指标选择

你可以通过「指标选择框」(以上界面中 Write (bytes) 处) 来查看你关心的指标：

- `Read (bytes)`：读取字节量
- `Write (bytes)`：读取字节量

- `Read (keys)`：读取次数
- `Write (keys)`：写入次数
- `All`：所有（读写流量的总和）

9.9.7.4.5 刷新与自动刷新

可以通过点击 Refresh（刷新）按钮来重新获得基于当前时间的热力图。当需要实时观察数据库的流量分布情况时，可以点击按钮右侧的向下箭头，选择一个固定的时间间隔让热力图按此间隔自动刷新。

> 注意：
>
> 如果进行了时间范围或者 Region 范围的调整，自动刷新会被关闭。

9.9.7.4.6 查看详情

可以将鼠标悬停在你所关注的 Bucket 上，来查看这个区域的详细信息：

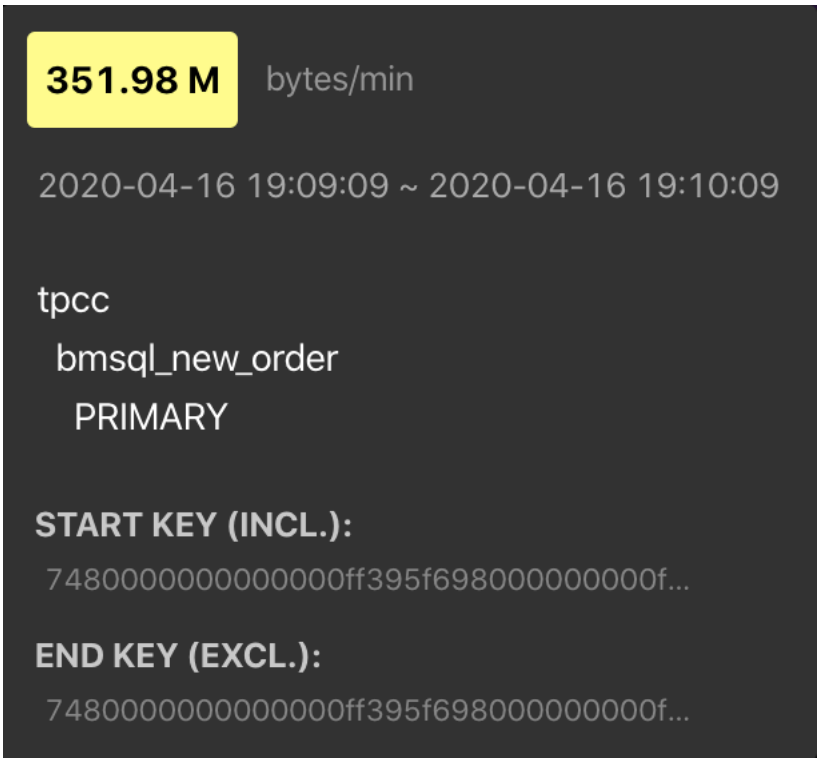


图 101: Bucket 详细信息

如果需要复制某个信息，可以进行点击 Bucket。此时相关详细信息的页面会被暂时钉住。点击你关注的信息，即可将其复制到剪切板：

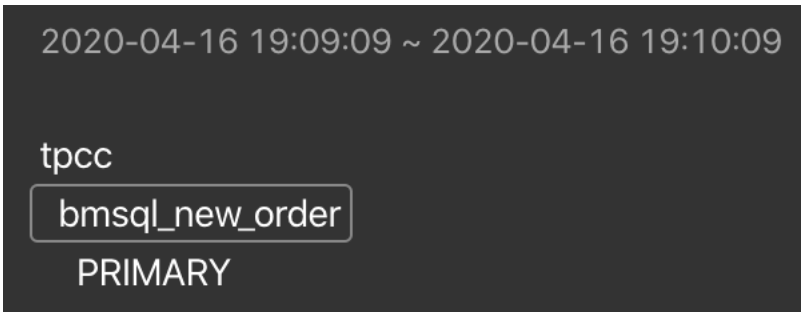


图 102: 复制 Bucket 详细信息

9.9.7.5 常见热力图解读

本章节选取了 Key Visualizer 中常见的四种热力图进行解读。

9.9.7.5.1 均衡：期望结果

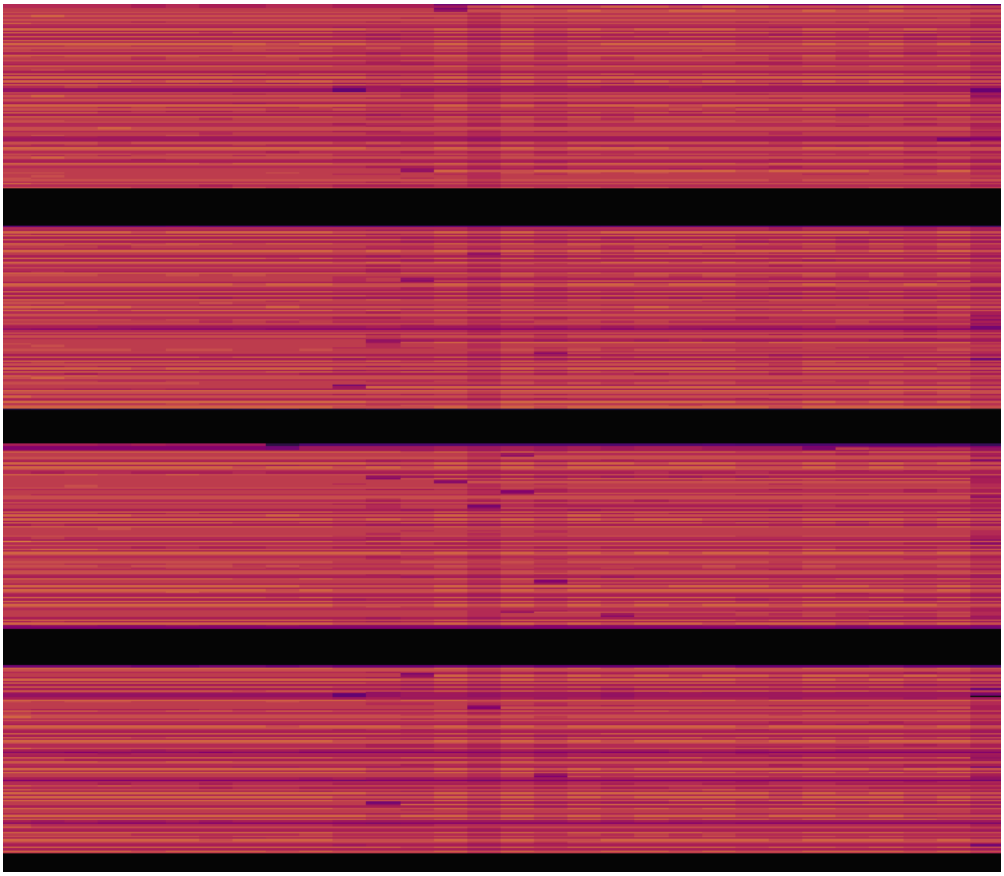


图 103: 均衡结果图

如上图所示，热力图颜色均匀或者深色和亮色混合良好，说明读取或写入在时间和 Region 空间范围上都分布得比较均衡，访问压力均匀地分摊在所有的机器上。这种负载是最适合分布式数据库的。

9.9.7.5.2 X 轴明暗交替：需要关注高峰期的资源情况

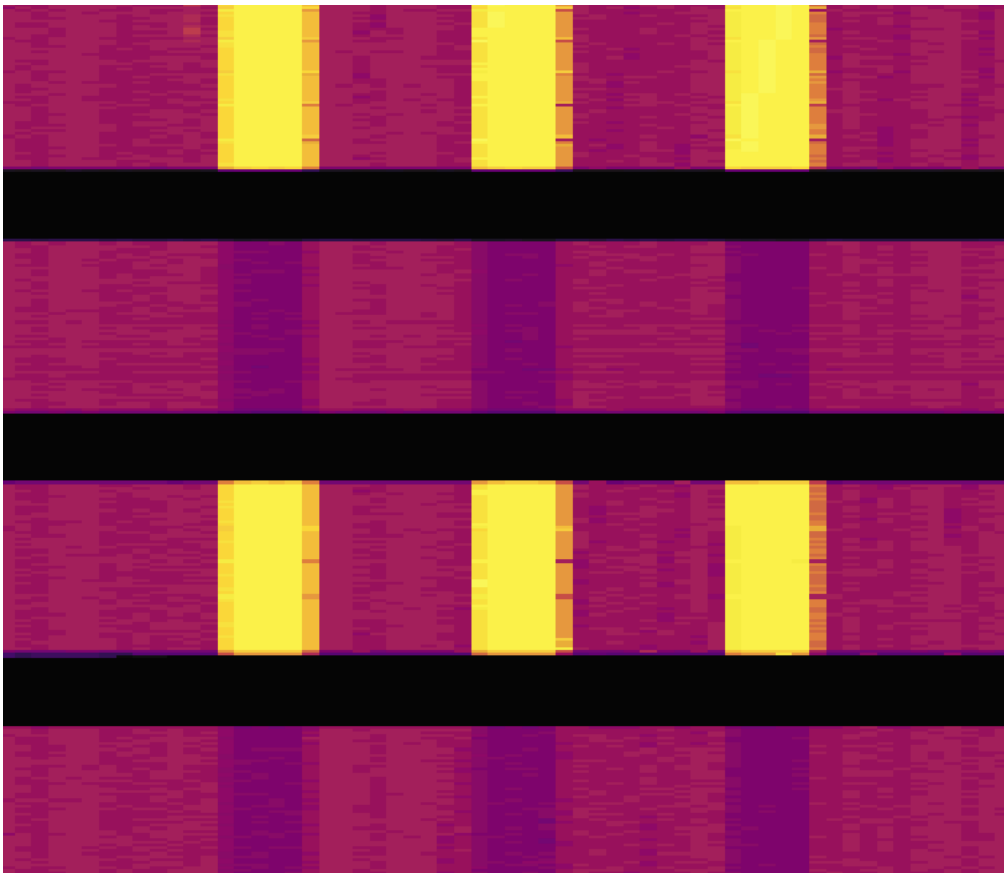


图 104: X 轴明暗交替

如上图所示，热力图在 X 轴（时间）上表现出明暗交替，但 Y 轴 (Region) 则比较均匀，说明读取或写入负载具有周期性的变化。这种情况可能出现在周期性的定时任务场景，如大数据平台每天定时从 TiDB 中抽取数据。一般来说可以关注一下使用高峰时期资源是否充裕。

9.9.7.5.3 Y 轴明暗交替：需要关注产生的热点聚集程度

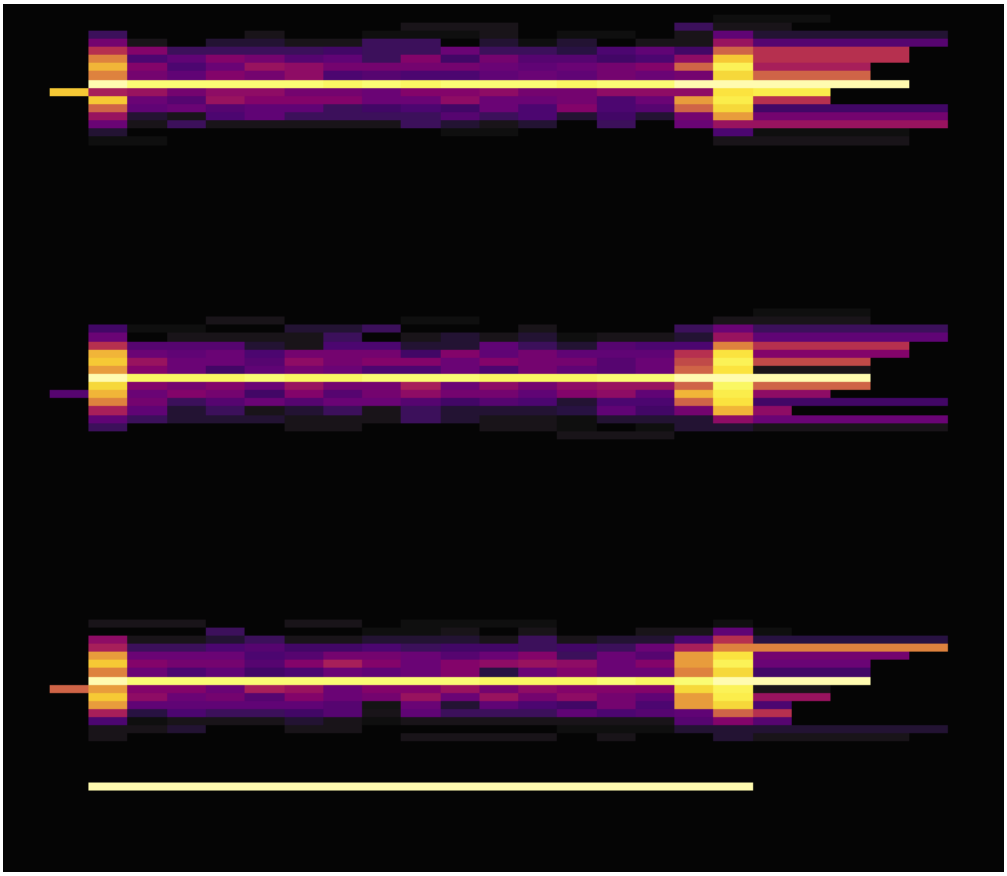


图 105: Y 轴明暗交替

如上图所示，热力图包含几个明亮的条纹，从 Y 轴来看条纹周围都是暗的，这表明明亮条纹区域的 Region 有很高的读写流量，可以从业务角度观察一下是否符合预期。例如，所有业务都关联用户表的情况下，用户表的整体流量就会很高，那么在热力图中表现为亮色区域就非常合理。

另外，明亮区域的高度（Y 轴方向的粗细）非常关键。由于 TiKV 自身拥有以 Region 为单位的热点平衡机制，因此涉及热点的 Region 越多其实越能有利于在所有 TiKV 实例上均衡流量。明亮条纹越粗、数量越多则意味着热点越分散、更多的 TiKV 能得到利用；明亮条纹越细、数量越少意味着热点越集中、热点 TiKV 越显著、越需要人工介入并关注。

9.9.7.5.4 明亮斜线：需要关注业务模式

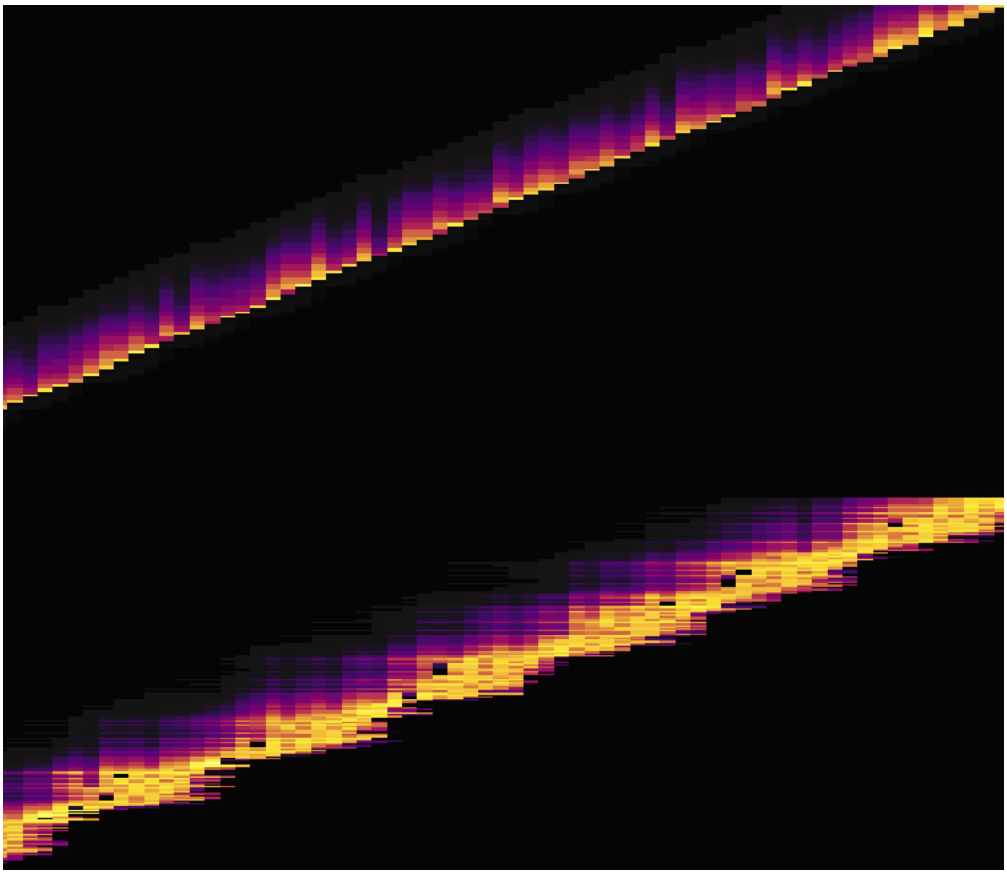


图 106: 明亮斜线

如上图所示，热力图显示了明亮的斜线，表明读写的 Region 是连续的。这种场景常常出现在带索引的数据导入或者扫描阶段。例如，向自增 ID 的表进行连续写入等等。图中明亮部分对应的 Region 是读写流量的热点，往往会成为整个集群的性能问题所在。这种时候，可能需要业务重新调整主键，尽可能打散以将压力分散在多个 Region 上，或者选择将业务任务安排在低峰期。

> 注意：
>
> 这里只是列出了几种常见的热力图模式。流量可视化页面中实际展示的是整个集群上所有数据库、数据表的热力图，因此非常有可能在不同的区域观察到不同的热力图模式，也可能观察到多种热力图模式的混合结果。使用的时候应当视实际情况灵活判断。

9.9.7.6 解决热点问题

TiDB 内置了不少帮助缓解常见热点问题的功能，深入了解请参考TiDB 高并发写入场景最佳实践。

9.9.8 TiDB Dashboard 监控关系图

TiDB Dashboard 监控关系图是 TiDB v4.0.7 起提供的新功能，可以将集群中各个内部流程的耗时监控数据绘制为关系图，帮助用户快速了解集群中各个环节的耗时及关系。

9.9.8.1 访问关系图

登录 TiDB Dashboard 后点击左侧导航的集群诊断可以进入此功能页面：



图 107: 生成监控关系图首页

设置区间起始时间和区间长度参数后，点击生成监控关系图按钮后，会进入监控关系图页面。

9.9.8.2 关系图解读

下面是一份监控耗时关系图示例，描述的是某个 TiDB 集群在 2020-07-29 16:36:00 开始往后 5 分钟内，TiDB 集群中各个监控的总耗时比例，以及各项监控之间的关系。

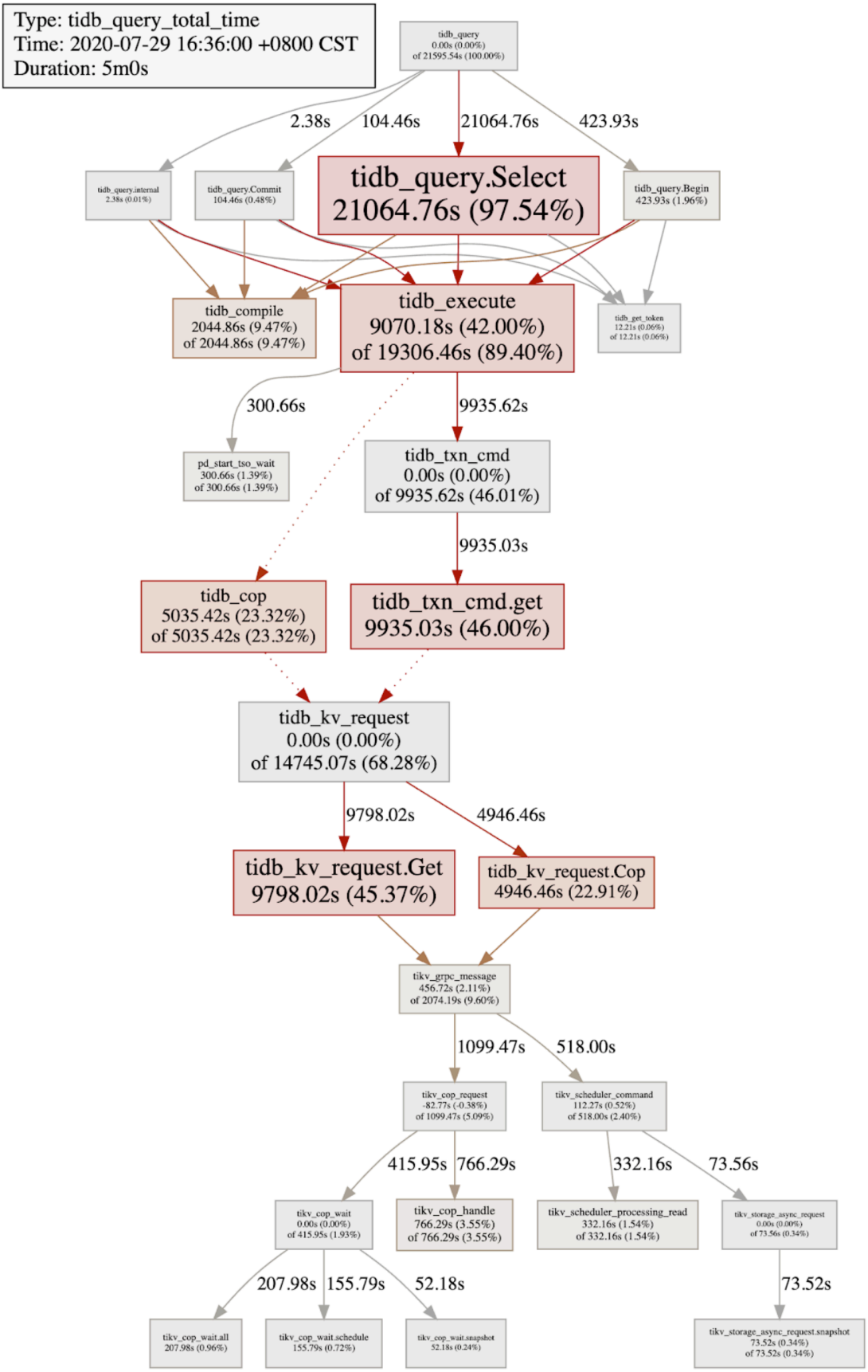


图 108: 监控关系图示例

例如以下 tidb_execute 节点监控图表示：tidb_execute 监控项的总耗时为 19306.46 秒，占总查询耗时的 89.4%，其中 tidb_execute 节点自身的耗时是 9070.18 秒，占总查询耗时的 42%。将鼠标悬停在该方框上，可以看到监控项的注释说明，总耗时、平均耗时、平均 P99 耗时等详细信息。

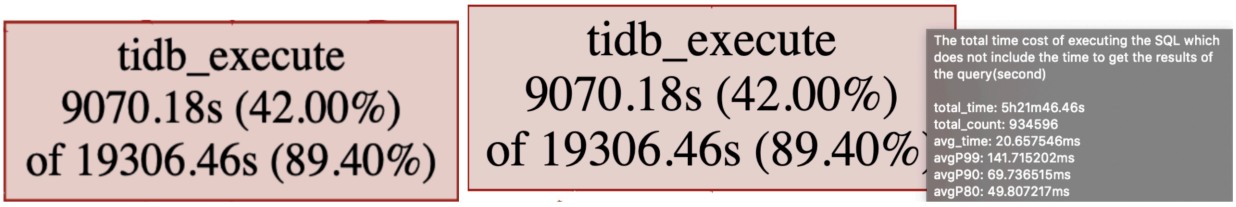


图 109: 监控关系图 tidb_execute 节点示例

9.9.8.2.1 节点的含义

每个方框节点代表一个监控项，包含了以下信息：

- 监控项的名称
- 监控项的总耗时
- 监控项总耗时和查询总耗时的比例

节点监控的总耗时 = 节点自身的耗时 + 子节点的耗时，所以某些节点监控图会显示节点自身的耗时和总耗时的比例。例如 tidb_execute 监控：

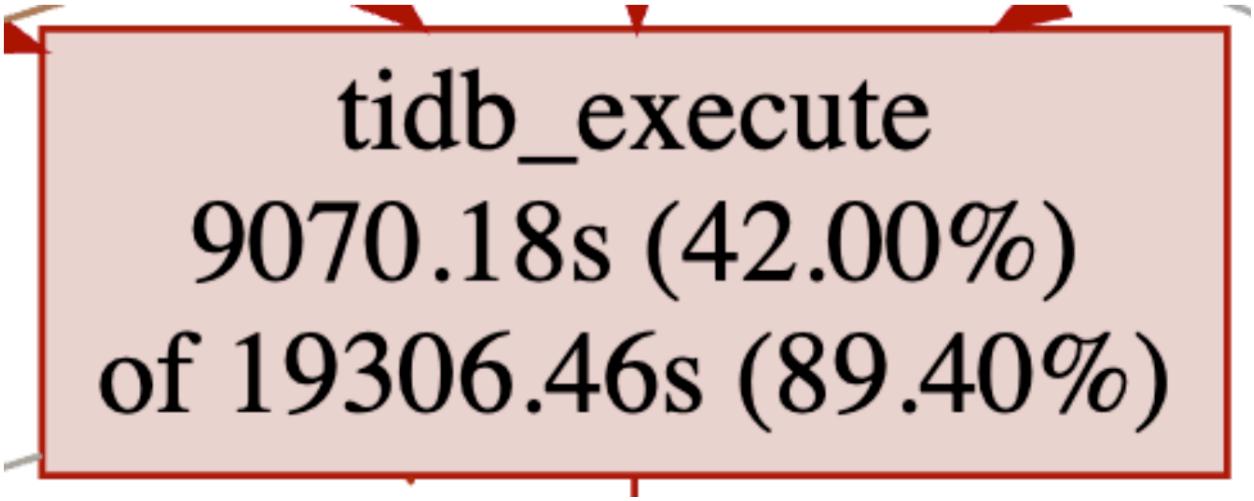


图 110: 监控关系图 tidb_execute 节点示例

- tidb_execute 是监控项的名字。该监控是指一条 SQL 请求在 TiDB 执行引擎中的执行耗时。
- 19306.46s 表示 tidb_execute 监控项消耗的总时间为 19306.46 秒。89.40% 表示 19306.46 秒占所有 SQL 查询总耗时（包括用户 SQL 和 TiDB 内部的 SQL）的比例为 89.40%。查询总耗时是 tidb_query 监控项的总耗时。

- 9070.18s 表示 tidb_execute 节点自身总执行耗时是 9070.18 秒，其余部分是被其子节点消耗的时间。42.00% 表示 9070.18 秒占所有查询总耗时的比例为 42.00%。

将鼠标悬停在该节点后，会显示监控项的更多详细信息：

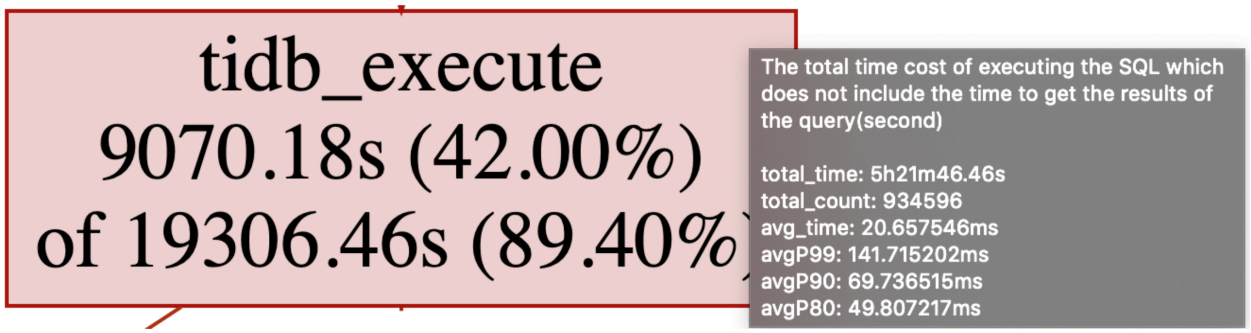


图 111: 监控关系图 tidb_execute 节点注释

上图信息为该项监控的注释说明，包括总耗时、总次数、平均耗时和平均 P99、P90、P80 耗时。

9.9.8.2.2 监控项之间的父子关系

下面以 tidb_execute 监控为例介绍该监控项相关的子节点：

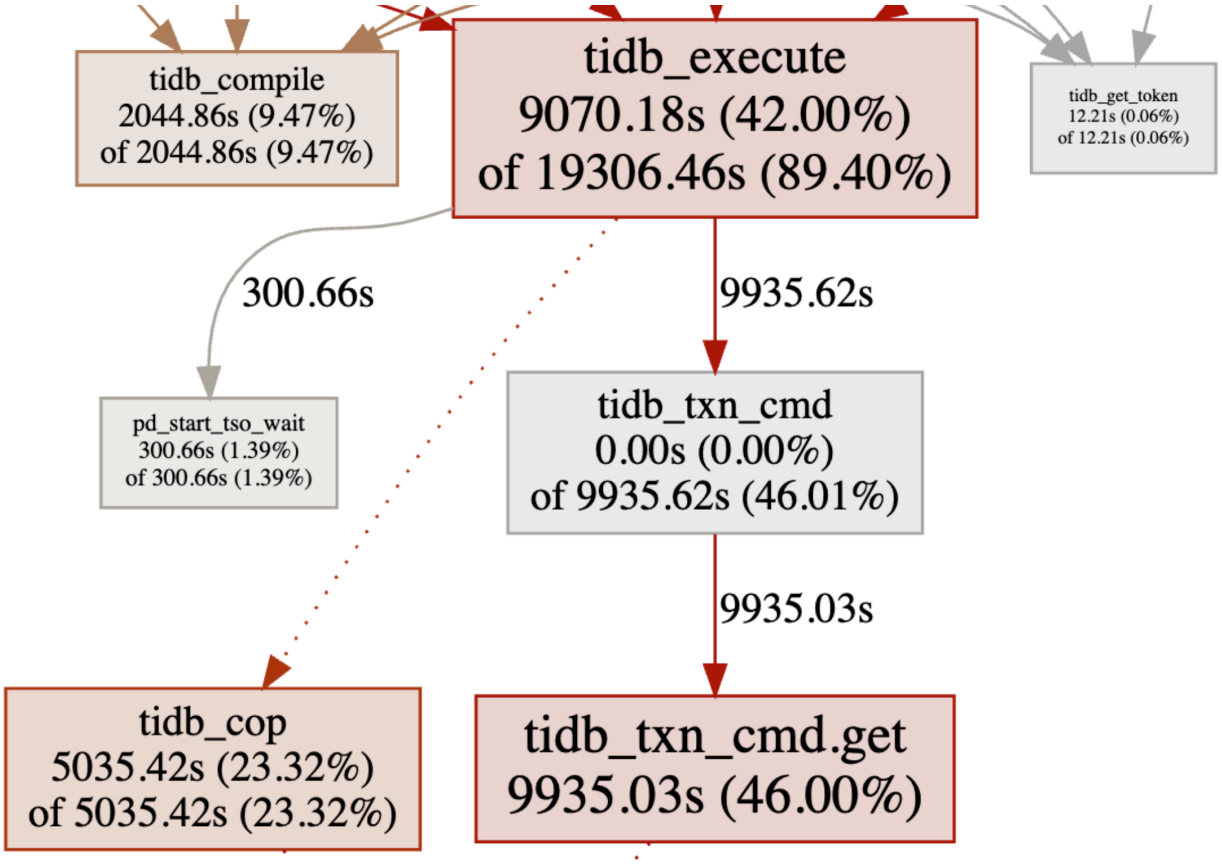


图 112: 监控关系图 tidb_execute 节点注释

可以看到，tidb_execute 包含两个子节点，分别是：

- pd_start_tso_wait：等待事务 start_tso 的耗时，其总耗时是 300.66 秒。
- tidb_txn_cmd：TiDB 执行事务相关命令的耗时，其总耗时是 9935.62 秒。

另外，tidb_execute 还有一条虚线箭头指向 tidb_cop 监控，这里虚线箭头的含义如下：

tidb_execute 包含 tidb_cop 监控的耗时，但是 cop 请求有可能并发执行。例如对两个表的进行 join 查询的 execute 耗时为 60 秒，其中 join 的两个表会并行地执行 cop 扫表请求。假如 cop 请求执行时间分别为 40 秒和 30 秒，那 cop 请求的总耗时是 70 秒，但是 execute 执行耗时只有 60 秒。所以如果父节点的耗时不完全包含子节点的耗时，就用虚线箭头来指向子节点。

> 注意：
>
> 当节点有虚线箭头指向的子节点时，该节点的本身的耗时是不准确的。例如 tidb_execute 监控中，tidb_execute 节点本身的耗时为 9070.18 = 19306.46 - 300.66 - 9935.62。这里 tidb_cop 节点的耗时并不会计入子节点耗时的计算，但实际上，tidb_execute 监控本身的耗时 9070.18 秒中包含了 tidb_cop 一部分监控节点的耗时，但无法确认具体包含了多少耗时。

9.9.8.2.3 tidb_kv_request 及其父节点

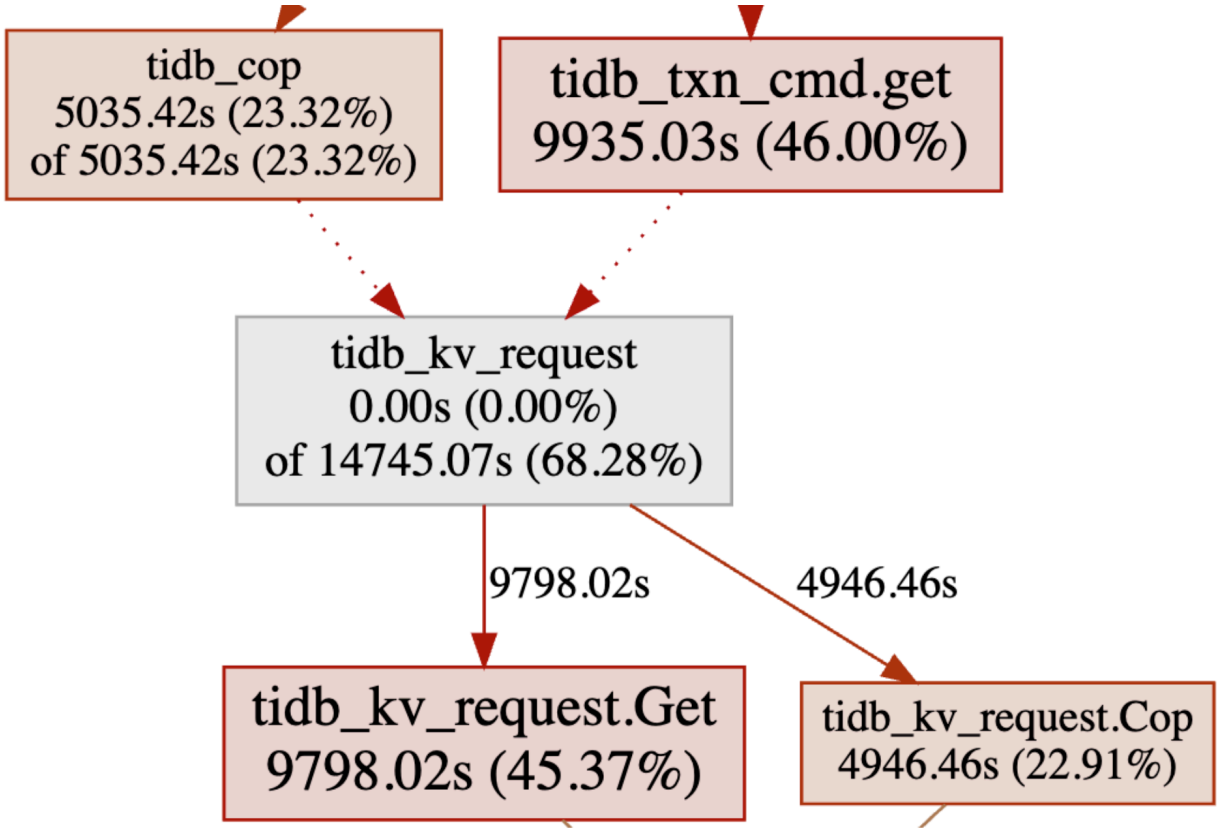


图 113: 监控关系图虚线节点关系

tidb_kv_request 的父节点 tidb_cop 和 tidb_txn_cmd.get 都用虚拟箭头指向 tidb_kv_request，这里表示：

- tidb_cop 的耗时包含部分 tidb_kv_request 的耗时
- tidb_txn_cmd.get 的耗时也包含部分的 tidb_kv_request 的耗时。

但是 tidb_cop 具体有多少耗时是 tidb_kv_request 消耗的，无法进行确认。

- tidb_kv_request.Get：TiDB 发送 Get 类型的 kv 请求的耗时。
- tidb_kv_request.Cop：TiDB 发送 Cop 类型的 kv 请求的耗时。

tidb_kv_request 与 tidb_kv_request.Get 和 tidb_kv_request.Cop 并不是父节点包含子节点的关系，而是组成关系。子节点的名称前缀是父节点的名称加上 .xxx，即为父节点的子类。这里可以理解为，TiDB 发送 kv 请求的总耗时为 14745.07 秒，其中 Get 和 Cop 类型的 kv 请求的总耗时分别为 9798.02 秒和 4946.46 秒。

9.9.9 SQL 语句分析

9.9.9.1 TiDB Dashboard SQL 语句分析执行详情页面

该页面可以查看所有 SQL 语句在集群上执行情况，常用于分析总耗时或单次耗时执行耗时较长的 SQL 语句。

在该页面中，结构一致的 SQL 查询（即使查询参数不一致）都会被归为同一个 SQL 语句，例如 `SELECT * FROM employee WHERE id IN (1, 2, 3)` 和 `select * from EMPLOYEE where ID in (4, 5)` 都属于同一 SQL 语句 `select * from employee where id in (...)`。

9.9.9.1.1 访问列表页面

可以通过以下两种方法访问 SQL 语句分析页面：

- 登录后，左侧导航条点击“SQL 语句分析” (SQL Statements)。

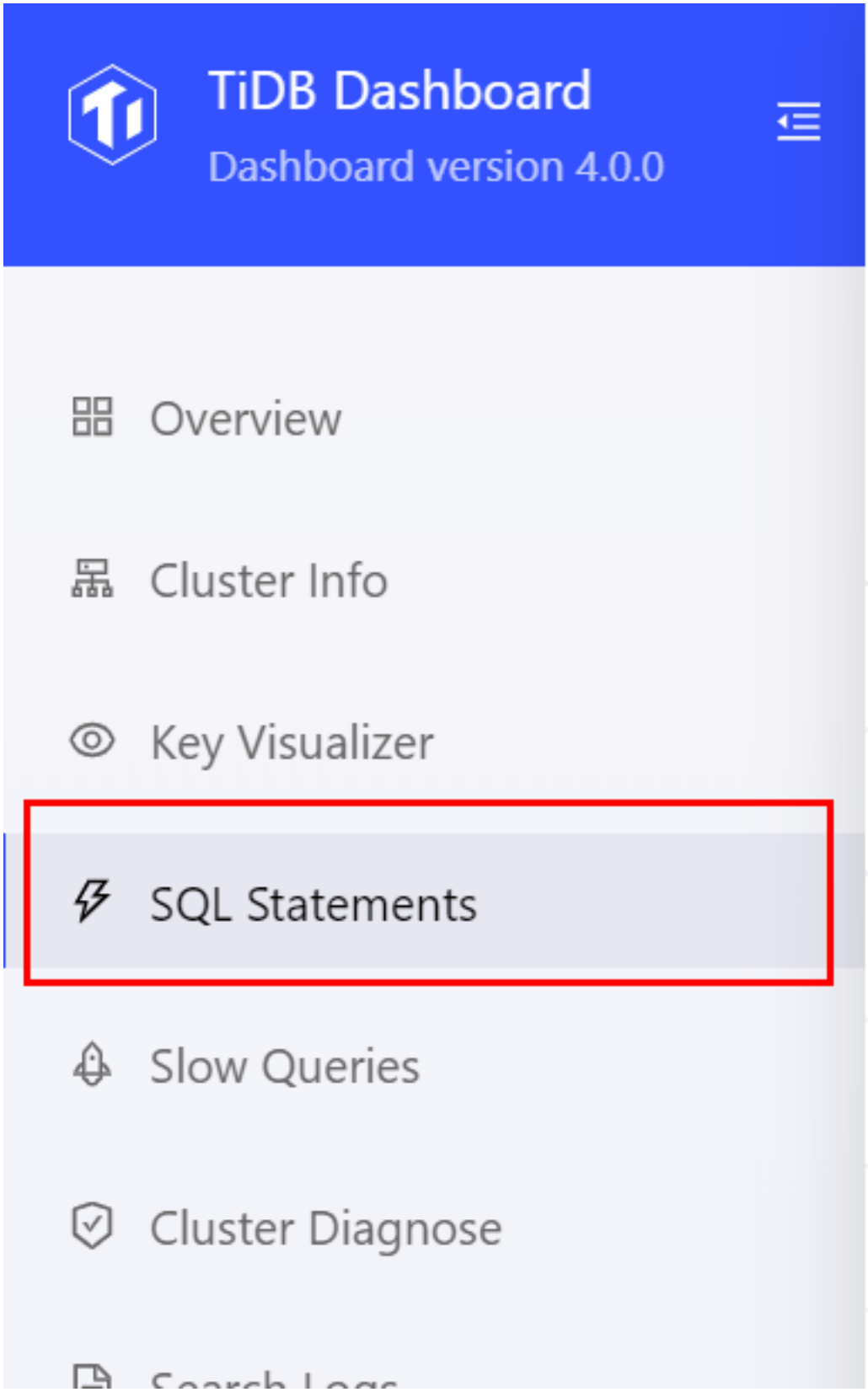

TiDB Dashboard
Dashboard version 4.0.0
Overview
Cluster Info
Key Visualizer
SQL Statements
Slow Queries
Cluster Diagnose

图 114: 访问

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/statement（将 127.0.0.1:2379 替换为实际 PD 实例地址和端口）。

SQL 语句分析页面所展示的所有数据都来自于 TiDB Statement 系统表，参见Statement Summary Tables 文档了解该系统表的详细情况。

> **注意：**
>
> 在 SQL 语句分析页面的 Mean Latency 列中，蓝色条表示平均耗时。如果某条 SQL 语句的蓝色条上有黄色的线，黄色线的左右两侧分别代表最近一个收集周期内该 SQL 语句耗时的最小值和最大值。

修改列表过滤条件

页面顶部可修改显示的时间范围、按执行 SQL 所在数据库进行过滤，或按 SQL 类型进行过滤，如下所示。默认显示最近一个收集周期（默认最近 30 分钟）内的所有 SQL 语句执行情况。

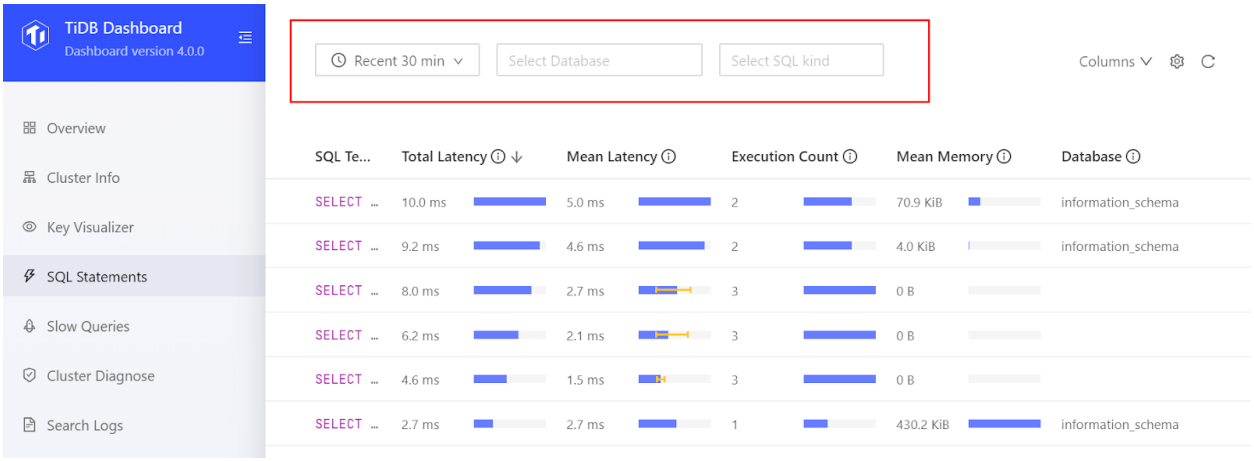


图 115: 修改过滤条件

显示其他字段

页面顶部选择列 (Columns) 选项中可选择显示更多列，可将鼠标移动到列名右侧的 (i) 图标处查看列的说明：

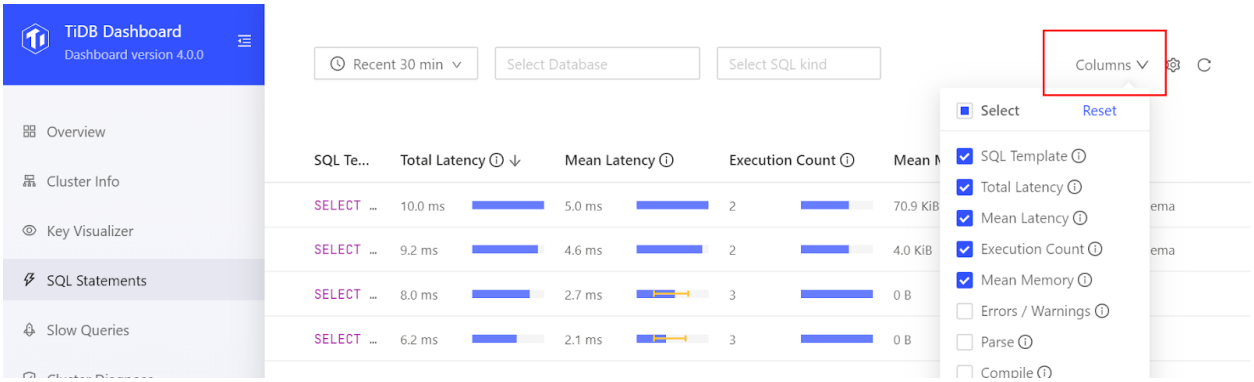


图 116: 选择列

修改列表排序依据

列表默认以累计耗时 (Total Latency) 从高到低进行排序，点击不同的列标题可以修改排序依据或切换排序顺序：

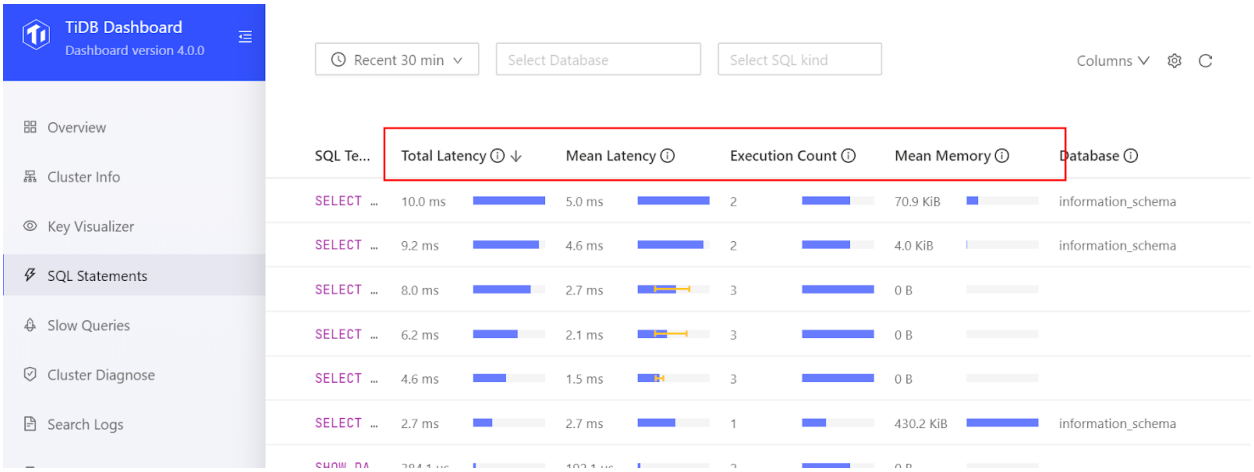


图 117: 修改列排序

修改数据收集设置

在列表页面，点击顶部右侧的设置 (Settings) 按钮，即可对 SQL 语句分析功能进行设置：

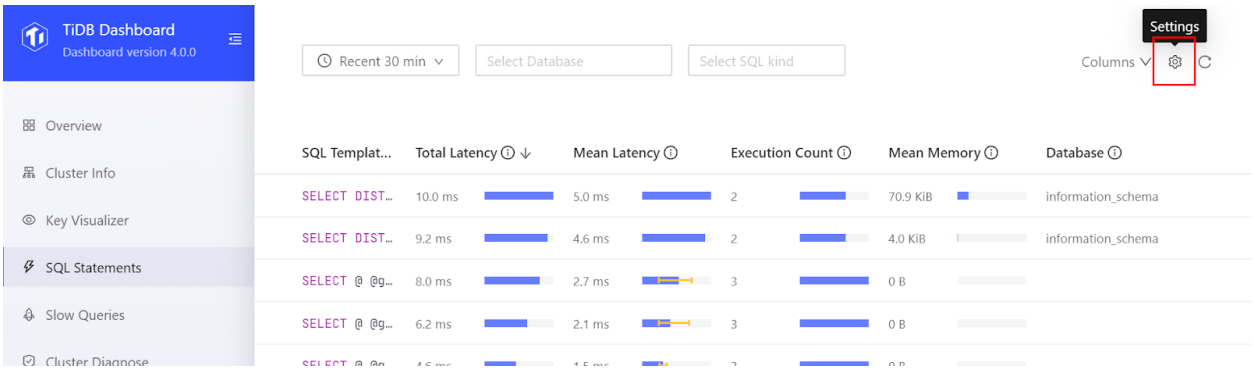


图 118: 设置入口

点击后设置界面如下图所示：

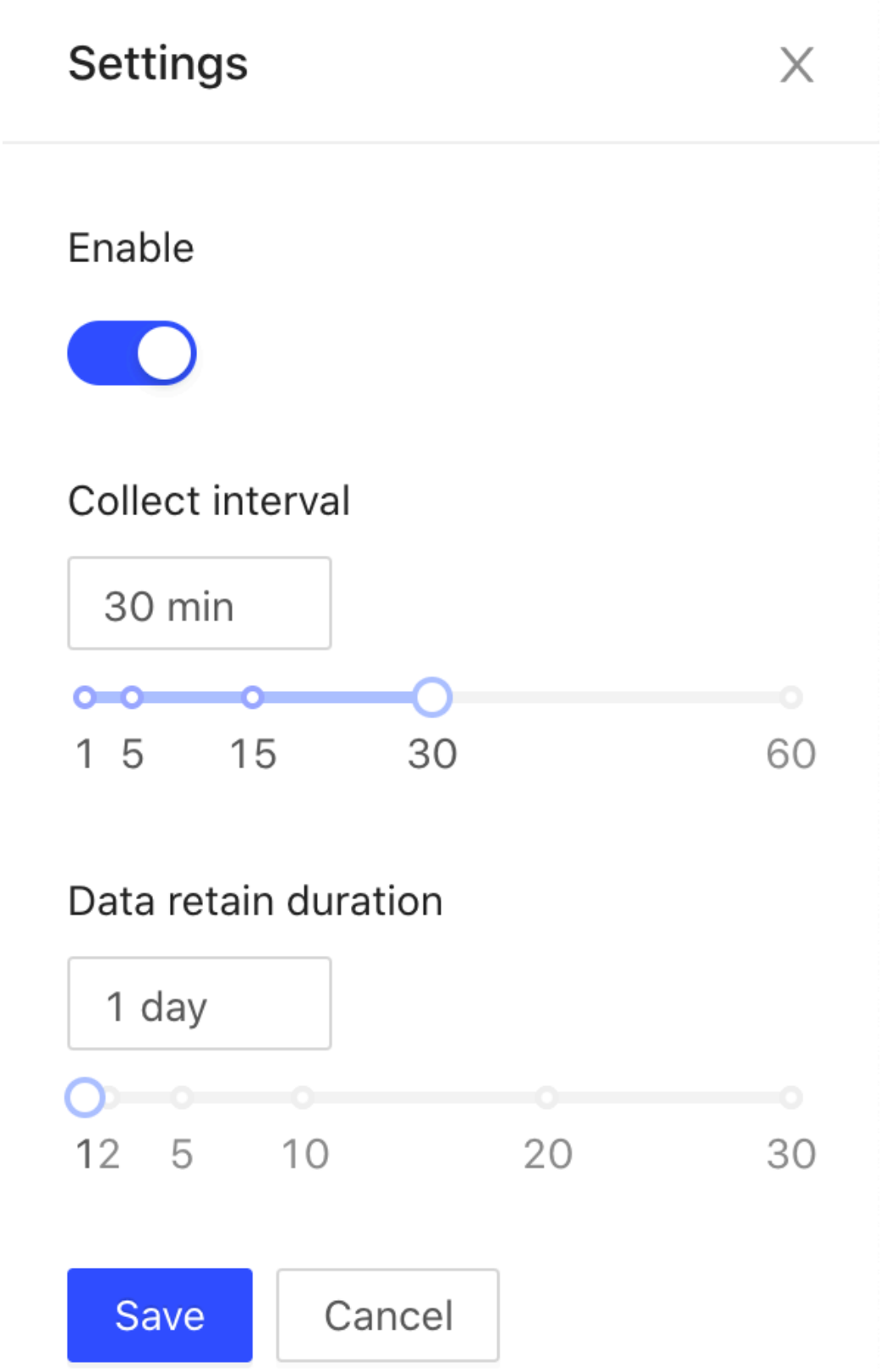

Settings
Enable
Collect interval
30 min
1 5 15 30 60
Data retain duration
1 day
12 5 10 20 30
Save
Cancel

图 119: 设置

在设置中可以选择关闭或开启 SQL 语句分析功能。在开启 SQL 语句分析功能时可以修改以下选项：

- 数据收集周期：默认 30 分钟，每次进行 SQL 语句分析的时间长度。SQL 语句分析功能每次对一段时间范围内的所有 SQL 语句进行汇总统计。

  如果这个时间范围过长，则统计的粒度粗，不利用定位问题；如果太短，则统计的粒度细，方便定位问题，但会导致在相同的数据保留时间内产生更多的记录，产生更多的内存占用。因此需要根据实际情况调整，在需要定位问题时适当地将值调低。

- 数据保留时间：默认 1 天，统计信息保留的时间，超过这个时间的数据会被从系统表中删除。

参见Statement Summary Tables 参数设置文档了解详细情况。

> 注意：
>
> 由于 Statement 系统表只存放在内存中，关闭此功能后，系统表中的数据会将清空。
>
> 数据收集周期和保留时间的值会影响内存占用，因此建议根据实际情况调整，保留时间不宜设置过大。

9.9.9.1.2 下一步

阅读查看执行详情章节了解如何进一步查看 SQL 语句的详细执行情况。

9.9.9.2 TiDB Dashboard SQL 语句分析执行详情页面

在列表中点击任意一行可以进入该 SQL 语句的详情页查看更详细的信息，这此信息包括三大部分：

- SQL 语句概况：包括 SQL 模板，SQL 模板 ID，当前查看的时间范围，执行计划个数以及执行所在的数据库（下图区域 1）
- 执行计划列表：如果该 SQL 语句有多个执行计划，则显示该列表，可以选择不同的执行计划，在列表下方会显示选中的执行计划详情；如果只有一个执行计划，则该列表不显示（下图区域 2）
- 执行计划详情：显示选中的执行计划的详细信息，具体见下一小节（下图区域 3）

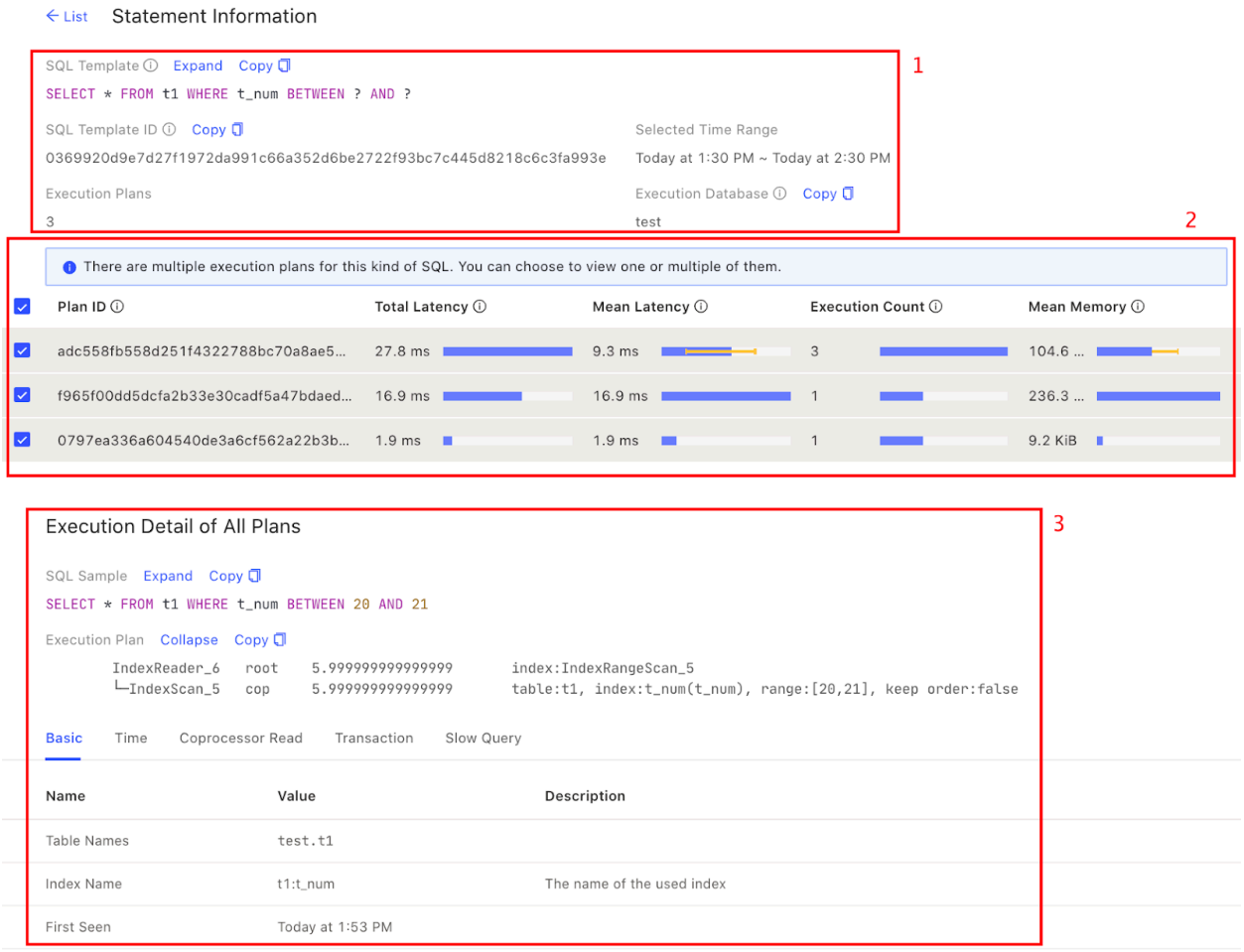


图 120: 详情

9.9.9.2.1 执行计划详情

执行计划详情包括以下内容：

- SQL 样本：该计划对应的实际执行的某一条 SQL 语句文本。时间范围内任何出现过的 SQL 都可能作为 SQL 样本。
- 执行计划：执行计划的完整内容，参阅理解 TiDB 执行计划文档了解如何解读执行计划。如果选择了多个执行计划，则显示的是其中任意一个。
- 其他关于该 SQL 的基本信息、执行时间、Coprocessor 读取、事务、慢查询等信息，可点击相应标签页标题切换。

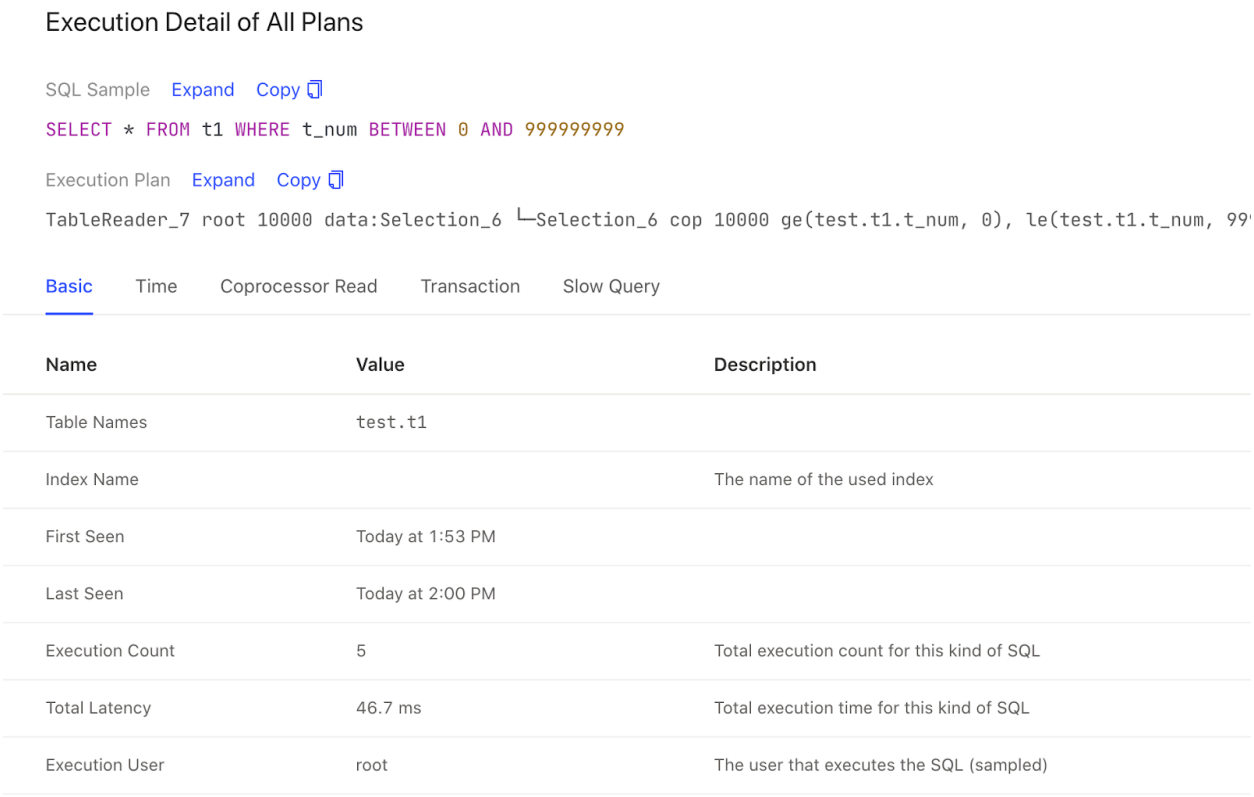


图 121: 执行计划详情

基本信息

包含关于表名、索引名、执行次数、累计耗时等信息。描述 (Description) 列对各个字段进行了具体描述。

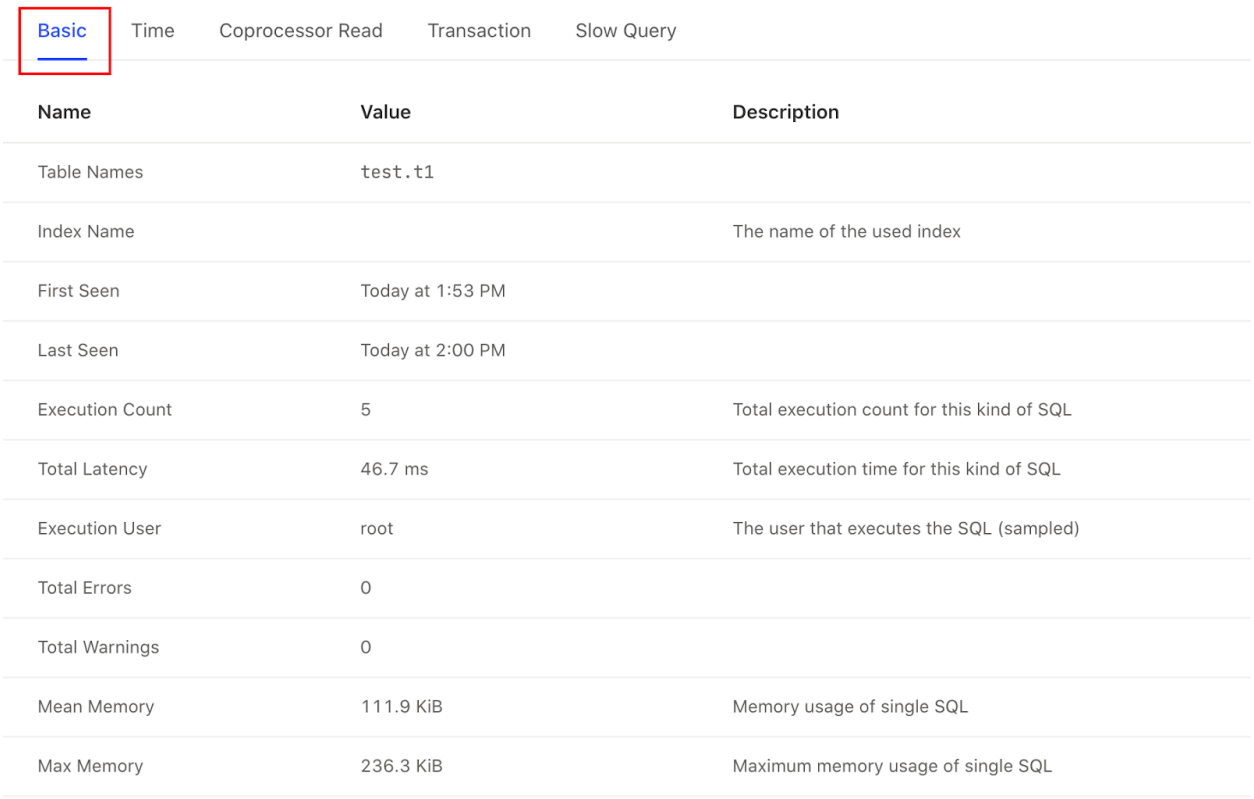


图 122: 基本信息

执行时间

显示执行计划执行的各阶段所耗费时间。

> **注意：**
>
> 由于单个 SQL 语句内部可能有并行执行的操作，因此各阶段累加时间可能超出该 SQL 语句的实际执行时间。

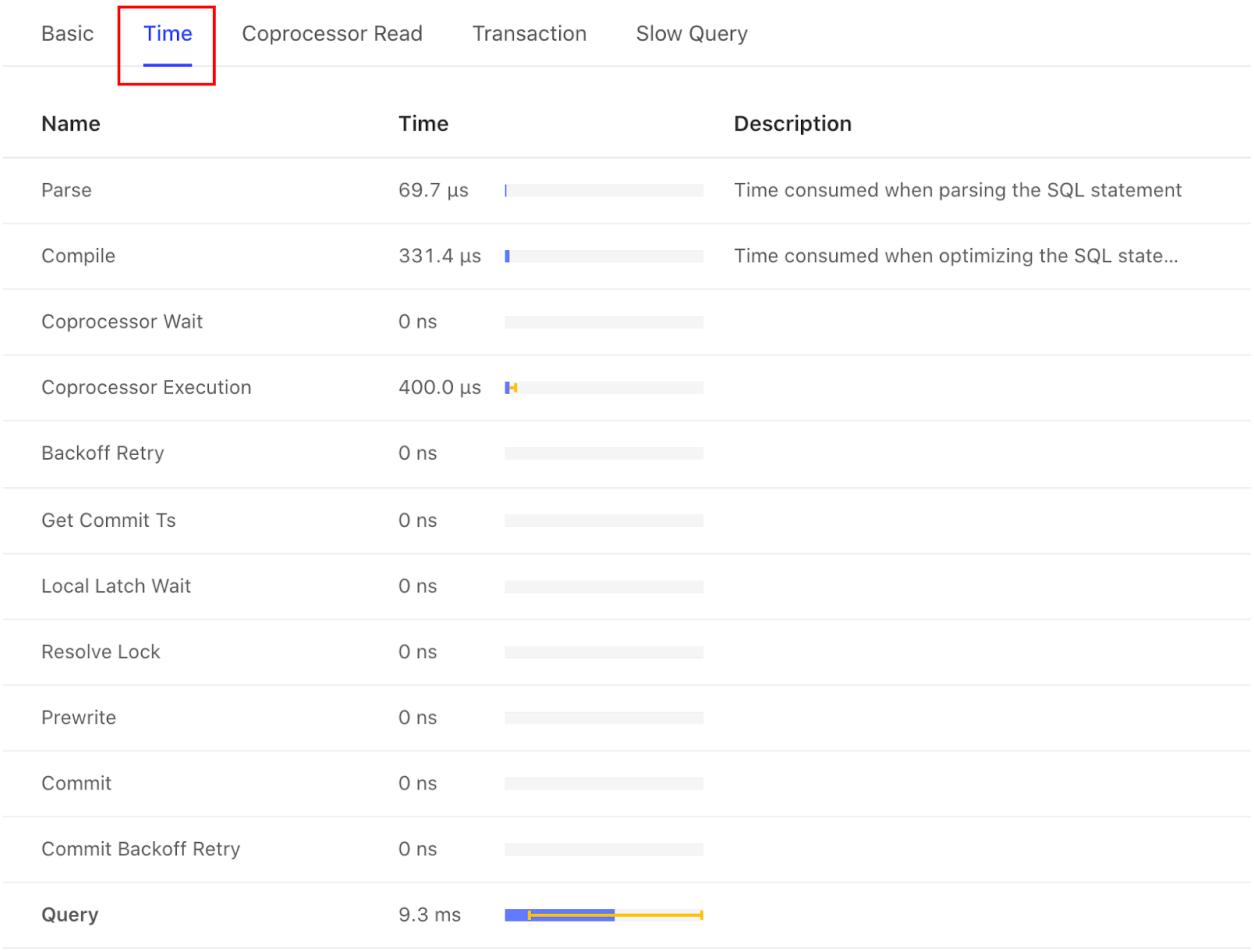


图 123: 执行时间

Coprocessor 读取

显示 Coprocessor 读取的相关信息。

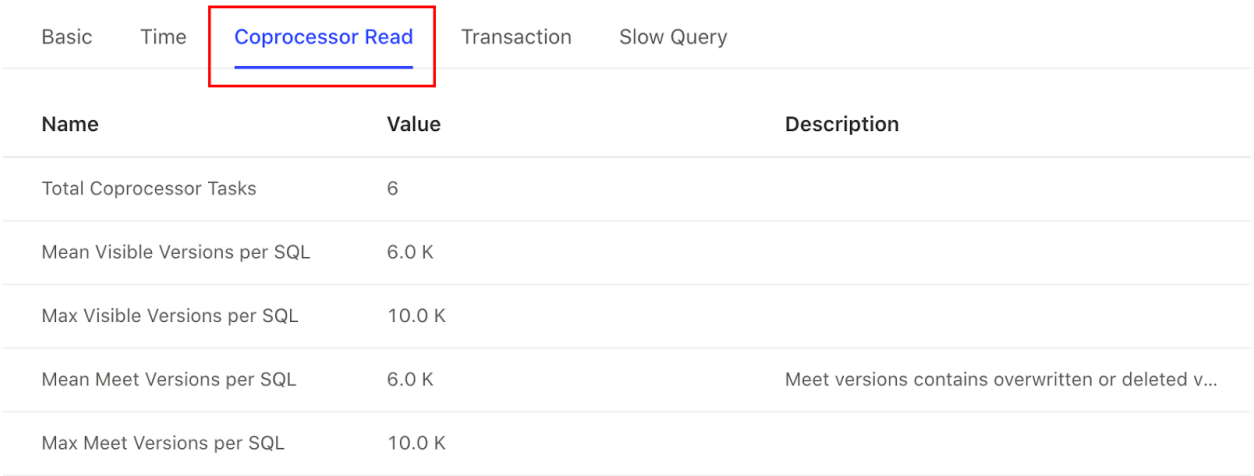


图 124: Coprocessor 读取

事务

显示执行计划与事务相关的信息，比如平均写入 key 个数，最大写入 key 个数等。

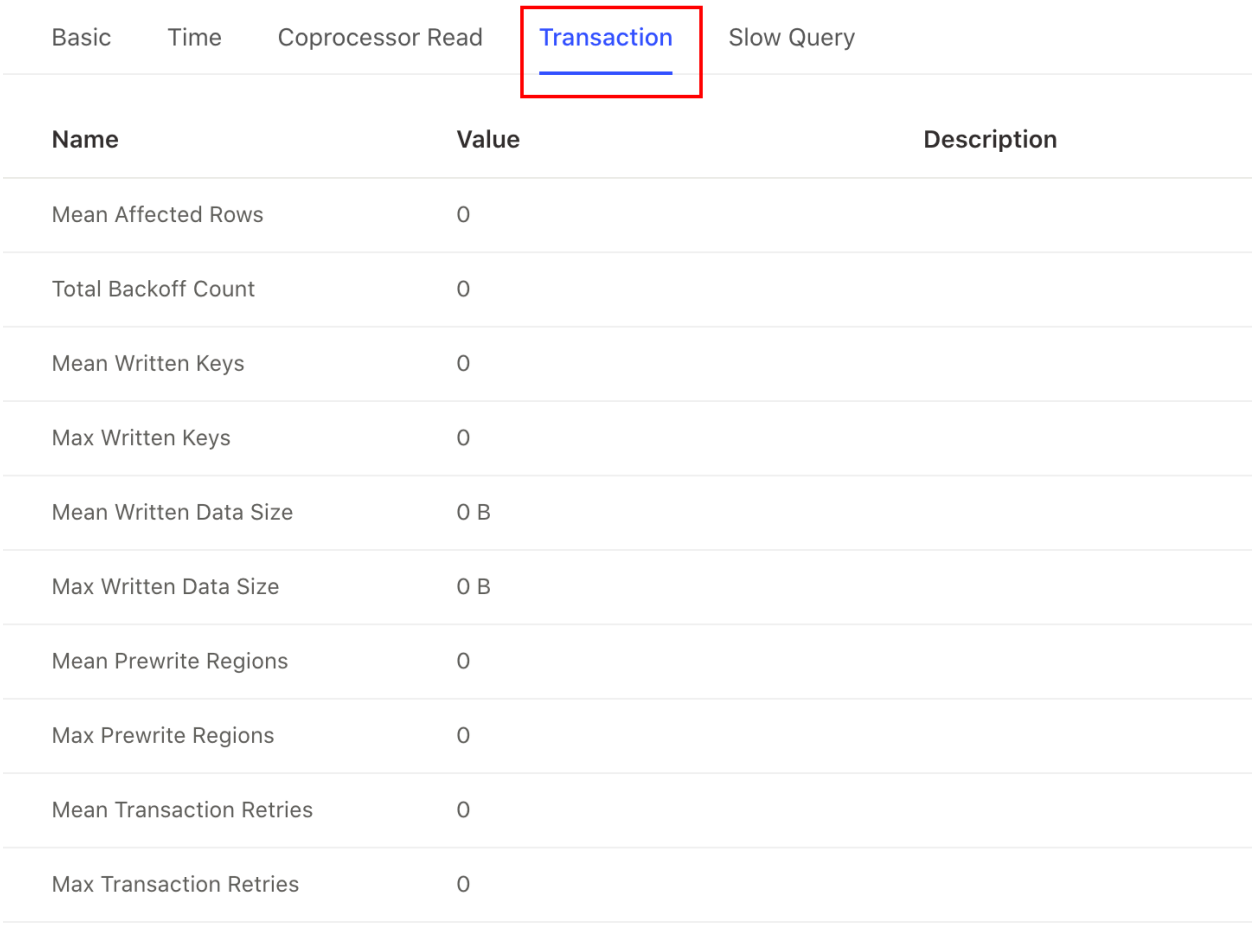


图 125: 事务

慢查询

如果该执行计划执行过慢，则在慢查询标签页下可以看到其关联的慢查询记录。

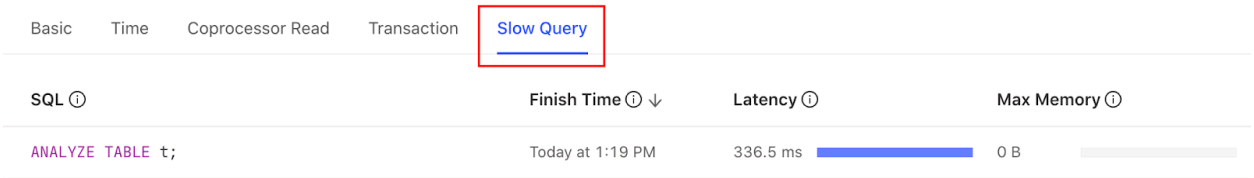


图 126: 慢查询

该区域显示的内容结构与慢查询页面一致，详见慢查询页面。

9.9.10 TiDB Dashboard 慢查询页面

该页面上能检索和查看集群中所有慢查询。

默认情况下，执行时间超过 300ms 的 SQL 查询就会被视为慢查询，被记录到慢查询日志中，并可通过本功能对记录到的慢查询进行查询。可调整`tidb_slow_log_threshold` SESSION 变量或 TiDB `slow-threshold` 参数调整慢查询阈值。

> **注意：**
>
> 若关闭了慢查询日志，则本功能不可用。慢查询日志默认开启，可通过修改 TiDB 配置`enable-slow-log` 开启或禁用。

9.9.10.1 访问列表页面

可以通过以下两种方法访问慢查询页面：

- 登录后，左侧导航条点击慢查询 (Slow Queries)：

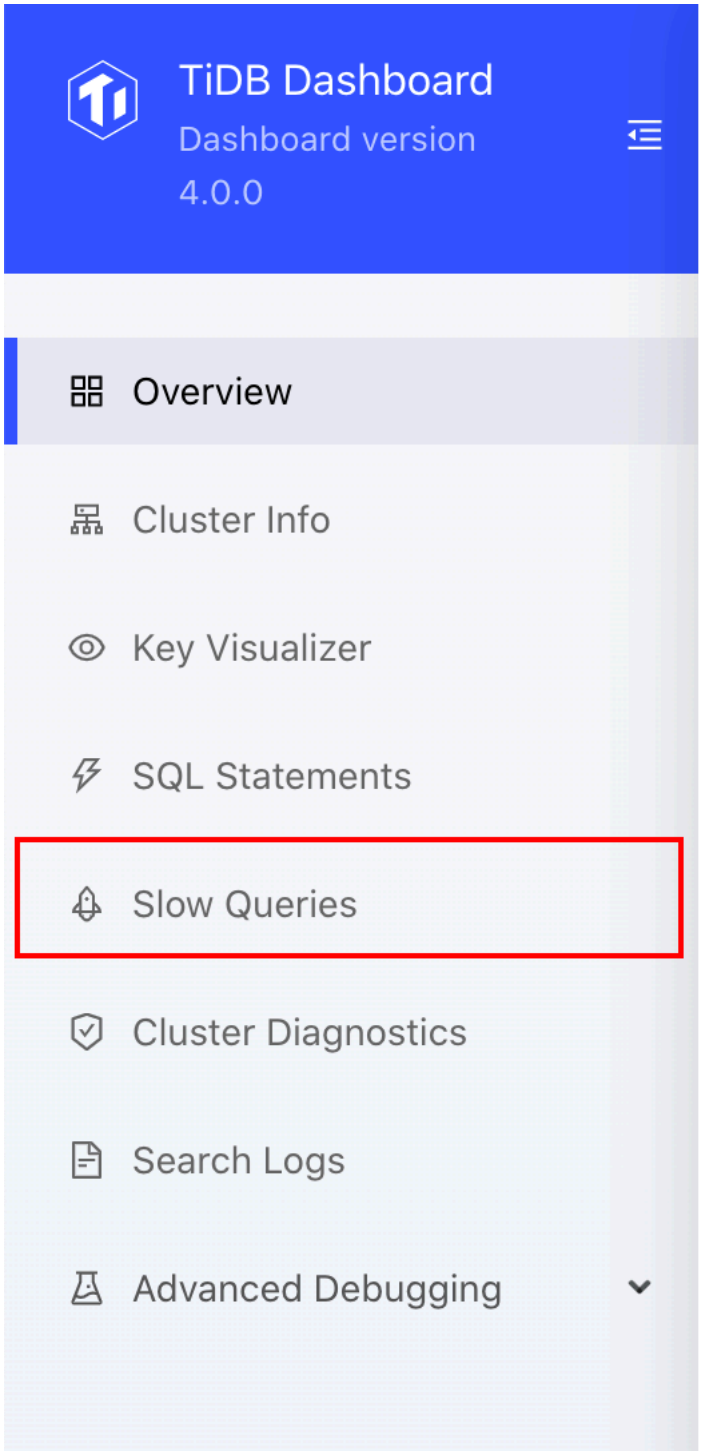


图 127: access 访问页面

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/slow_query（将 127.0.0.1:2379 替换为任意实际 PD 地址和端口）。

慢查询页面所展示的所有数据都来自于 TiDB 慢查询系统表及慢查询日志，参见慢查询日志文档了解详细情况。

9.9.10.1.1 修改列表过滤条件

可按时间范围、慢查询语句关联的数据库、SQL 关键字、SQL 类型、显示的慢查询语句数量等条件过滤，筛选慢查询句。如下所示，默认显示 30 分钟内最近 100 条慢查询。

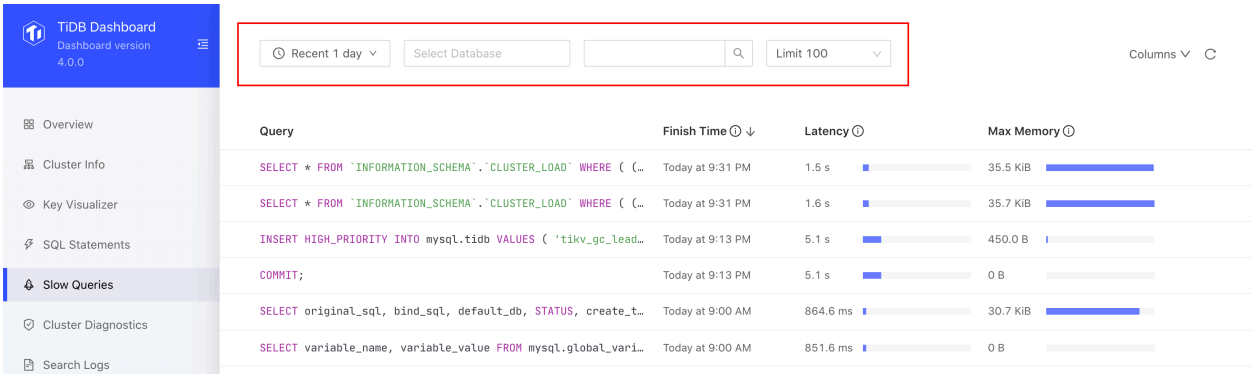


图 128: 修改列表过滤条件

9.9.10.1.2 显示更多列信息

页面顶部选择列 (Columns) 选项中可选择显示更多列，可将鼠标移动到列名右侧的 (i) 图标处查看列的说明：

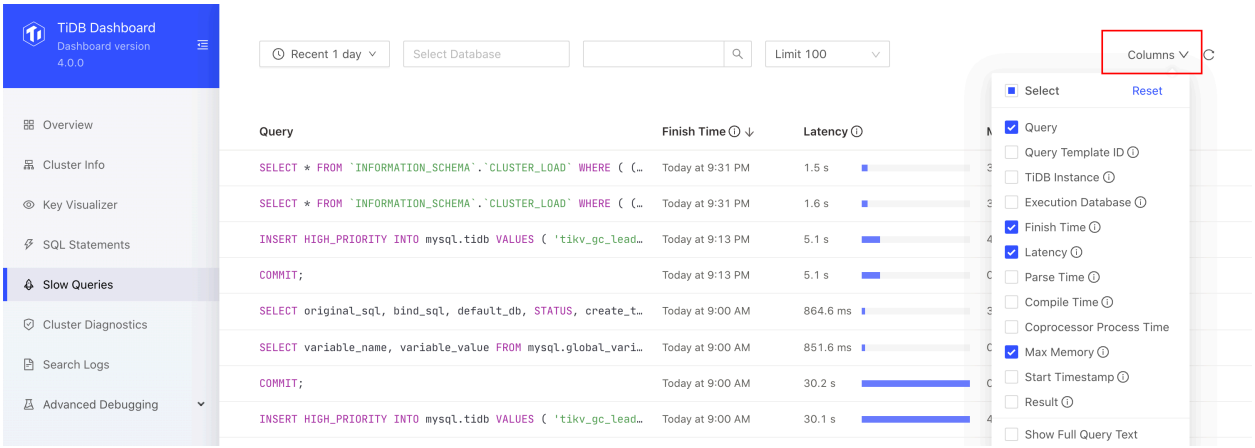


图 129: 显示更多列信息

9.9.10.1.3 修改列表排序依据

列表默认以结束运行时间 (Finish Time) 逆序排序，点击不同的列标题可以修改排序依据或切换排序顺序：

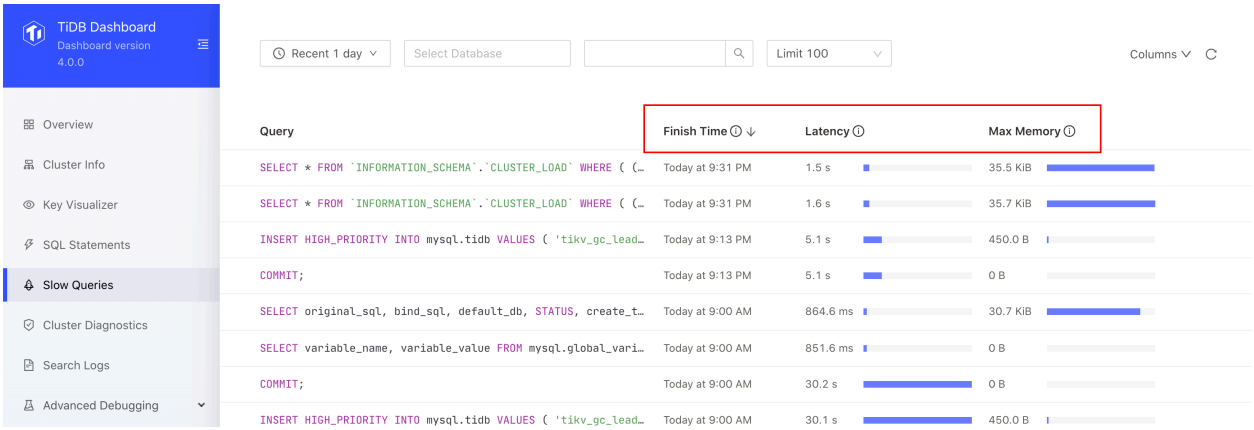


图 130: 修改列表排序依据

9.9.10.2 查看执行详情

在列表中点击任意一行可以显示该慢查询的详细执行信息，包含：

- SQL：慢查询 SQL 文本（下图中区域 1）
- 执行计划：慢查询的执行计划，参阅理解 TiDB 执行计划文档了解如何解读执行计划（下图中区域 2）
- 其他分类好的 SQL 执行信息（下图中区域 3）

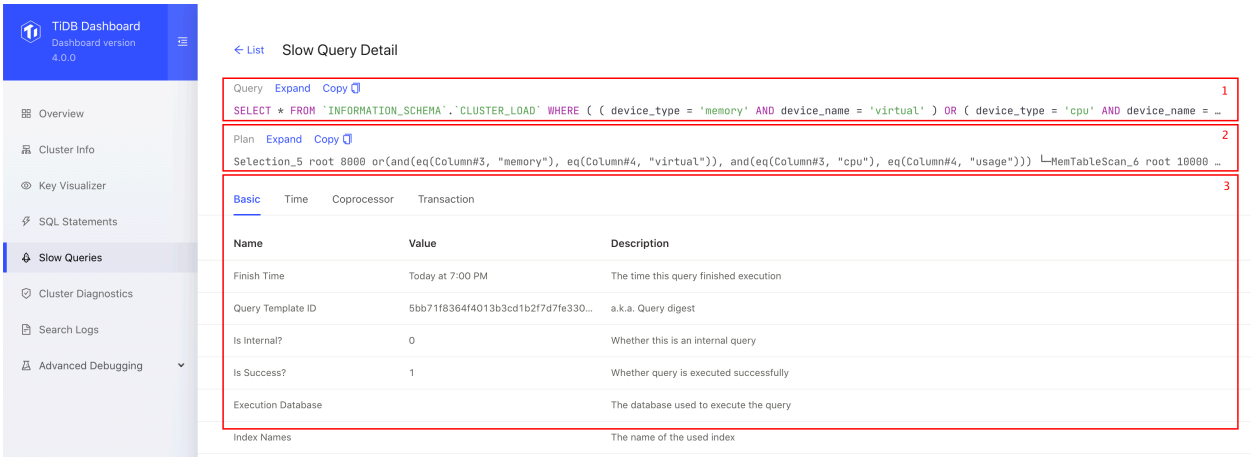


图 131: 查看执行详情

点击展开 (Expand) 链接可以展开相应项的完整内容，点击复制 (Copy) 链接可以复制完整内容到剪贴板。

点击标签页标题可切换显示不同分类的 SQL 执行信息：

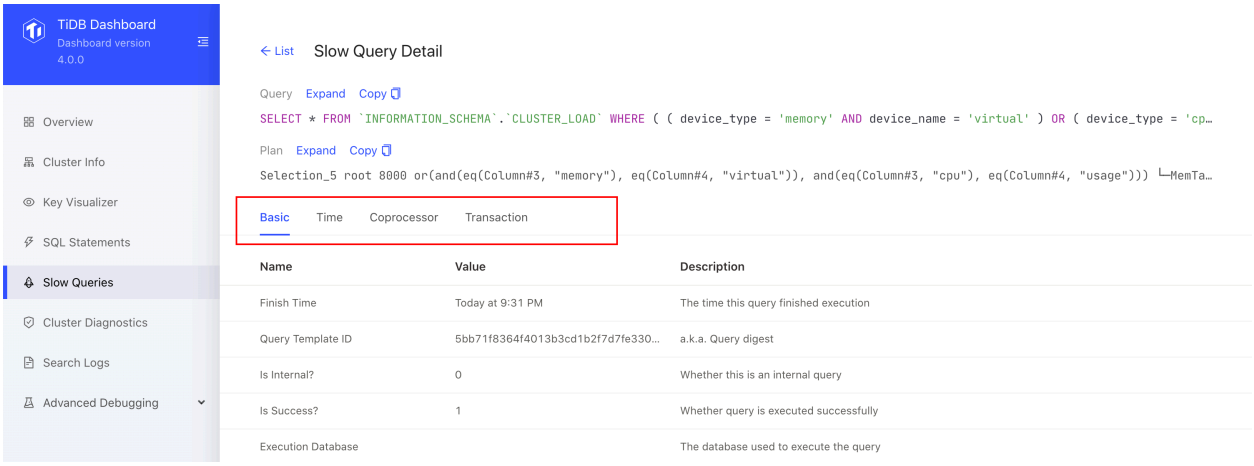


图 132: 显示不同分类执行信息

9.9.11 集群诊断页面

9.9.11.1 TiDB Dashboard 集群诊断页面

> **警告：**
>
> 该功能目前为实验特性，不建议在生产环境中使用。

集群诊断是在指定的时间范围内，对集群可能存在的问题进行诊断，并将诊断结果和一些集群相关的负载监控信息汇总成一个诊断报告。诊断报告是网页形式，通过浏览器保存后可离线浏览和传阅。

> **注意：**
>
> 集群诊断功能依赖于集群中部署有 Prometheus 监控组件，参见TiUP 部署文档了解如何部署监控组件。若集群中没有部署监控组件，生成的诊断报告中将提示生成失败。

9.9.11.1.1 访问

可以通过以下两种方法访问集群诊断页面：

- 登录后，左侧导航条点击集群诊断 (Cluster Diagnose)：

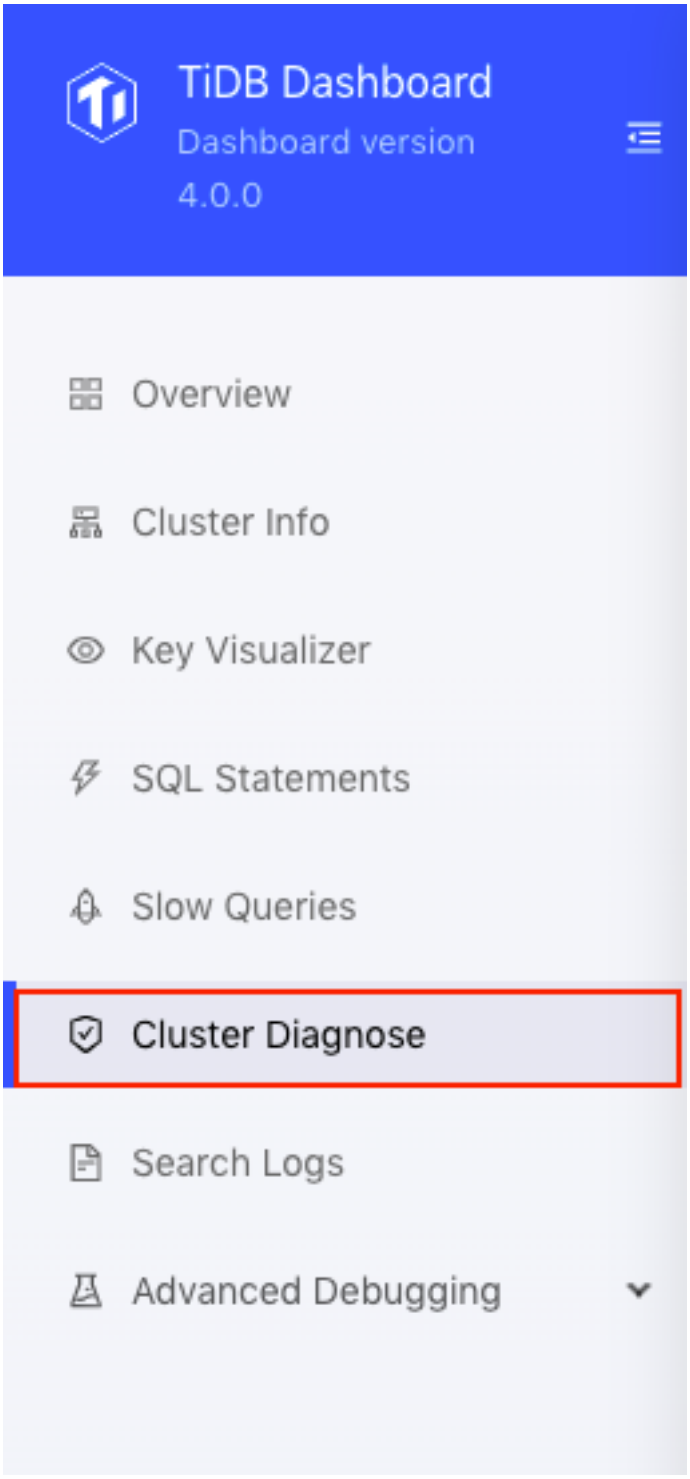


图 133: 访问

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/diagnose（将 127.0.0.1:2379 替换为任意实际 PD 地址和端口）。

9.9.11.1.2 生成诊断报告

如果想对一个时间范围内的集群进行诊断，查看集群的负载等情况，可以使用以下步骤来生成一段时间范围的诊断报告：

1. 设置区间的开始时间，例如 2020-05-21 14:40:00。
2. 设置区间长度。例如 10 min 。
3. 点击开始。

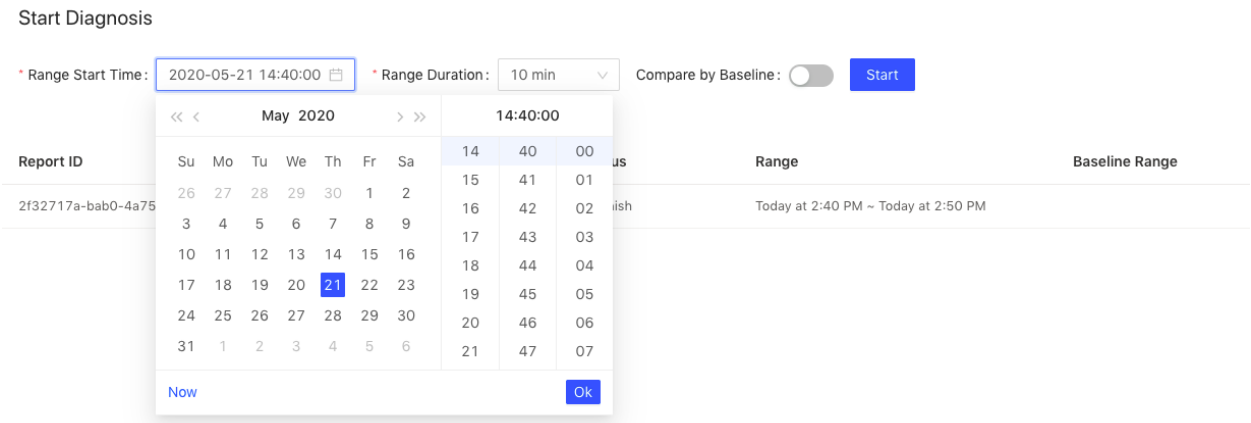


图 134: 生成单个时间段的诊断报告

> **注意：**
>
> 建议生成报告的时间范围在 1 min ~ 60 min 内，目前不建议生成超过 1 小时范围的报告。

以上操作会生成 2020-05-21 14:40:00 至 2020-05-21 14:50:00 时间范围的诊断报告。点击开始 (start) 后，会看到以下界面，生成进度 (progress) 是生成报告的进度条，生成报告完成后，点击查看报告 (View Full Report) 即可。

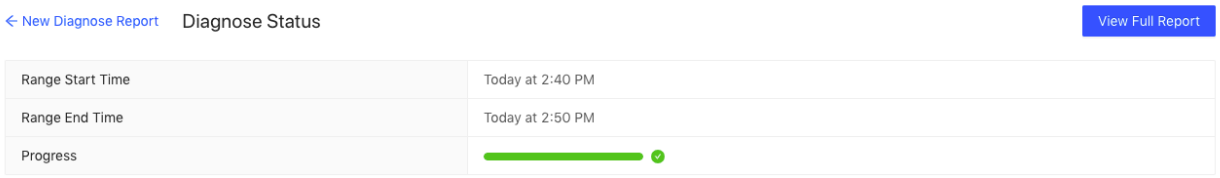


图 135: 生成报告的进度

9.9.11.1.3 生成对比诊断报告

如果系统在某个时间点发生异常，如 QPS 抖动或者延迟变高，可以生成一份异常时间范围和正常时间范围的对比报告，例如：

- 系统异常时间段：2020-05-21 14:40:00 ~ 2020-05-21 14:45:00，系统异常时间。
- 系统正常时间段：2020-05-21 14:30:00 ~ 2020-05-21 14:35:00，系统正常时间。

生成以上两个时间范围的对比报告的步骤如下：

1. 设置区间的开始时间，即异常时间段的开始时间，如 2020-05-21 14:40:00。
2. 设置区间长度。一般只系统异常的持续时间，例如 5 min。
3. 开启与基线区间对比开关。
4. 设置基线开始时间，即想要对比的系统正常时段的开始时间，如 2020-05-21 14:30:00。
5. 点击开始。

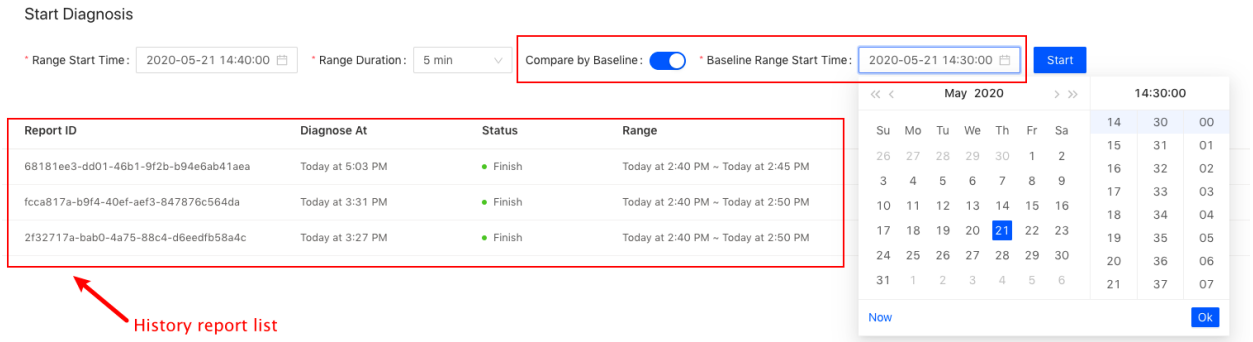


图 136: 生成对比报告

然后同样等报告生成完成后点击查看报告 (View Full Report) 即可。

另外，已生成的诊断报告会显式在诊断报告主页的列表里面，可以点击查看之前生成的报告，不用重复生成。

9.9.11.2 TiDB Dashboard 诊断报告

本文档主要介绍诊断报告的内容以及查看技巧，访问集群诊断和生成报告请参考诊断报告访问文档。

9.9.11.2.1 查看报告

诊断报告由以下几部分组成：

- 基本信息：包括生成报告的时间范围，集群的硬件信息，集群的拓扑版本信息。
- 诊断信息：显示自动诊断的结果。
- 负载信息：包括服务器，TIDB/PD/TiKV 相关的 CPU、内存等负载信息。
- 概览信息：包括 TiDB/PD/TiKV 的各个模块的耗时信息和错误信息。
- TiDB/PD/TiKV 监控信息：包括各个组件的监控信息。
- 配置信息：包括各个组件的配置信息。

报告中报表示例如下：

Total Time Consume

The table contain the event time consume in TiDB/TiKV/PD. METRIC_NAME is the event name; LABEL is the event label, such as instance, event type ...; TIME_RATIO is the TOTAL_TIME of this event devide by the TOTAL_TIME of upper event which TIME_RATIO is 1; TOTAL_TIME is the total time cost of this event; TOTAL_COUNT is the total count of this event; P999 is the max time of 0.999 quantile; P99 is the max time of 0.99 quantile; P90 is the max time of 0.90 quantile; P80 is the max time of 0.80 quantile;

METRIC_NAME	LABEL	TIME_RATIO	TOTAL_TIME	TOTAL_COUNT	P999	P99	P90	P80
tidb_query expand		1	23223.86	459625	63.54	45.55	0.1	0.06
tidb_get_token fold		0.000001	0.03	458888	0.00004	0.000001	0.000001	0.000001
|-- tidb_get_token	10.0.1.13:10080	0.000001	0.03	458888	0.00004	0.000001	0.000001	0.000001

图 137: 示例报表

上图中，最上面蓝框内的 Total Time Consume 是报表名。下方红框内的内容是对该报表意义的解释，以及报表中各个字段的含义。

报表中小按钮图标的解释如下：

- i 图标：鼠标移动到 i 图标处会显示该行的说明注释。
- expand：点击 expand 会看到这项监控更加详细的信息。如是上图中 `tidb_get_token` 的详细信息包括各个 TiDB 实例的延迟监控信息。
- fold：和 expand 相反，用于把监控的详细信息折叠起来。

所有监控基本上和 TiDB Grafna 监控面板上的监控内容相对应，发现某个模块异常后，可以在 TiDB Grafna 监控面板上查看更多详细的监控信息。

另外，报表中统计的 `TOTAL_TIME` 和 `TOTAL_COUNT` 由于是从 Prometheus 读取的监控数据，其统计会有一些计算上的精度误差。

以下介绍诊断报告的各部分内容。

基本信息

Report Time Range

Report Time Range 表显示生成报告的时间范围，包括开始时间和结束时间。

Report Time Range

START_TIME	END_TIME
2020-05-21 06:40:00	2020-05-21 06:50:00

图 138: report time range 报表

Cluster Hardware

Cluster Hardware 表显示集群中各服务器的硬件信息，包括 CPU、Memory、磁盘等信息。

Cluster Hardware

HOST	INSTANCE	CPU_CORES	MEMORY (GB)	DISK (GB)	UPTIME (DAY)
10.0.1.14	tiflash*1	20/40	125.64	nvme01: 498.638	2.9132028513567314
10.0.1.11	pd*1	20/40	125.64	sda1: 965.834	2.921851028772416
10.0.1.12	tikv*1	20/40	125.64	nvme0n1: 122.78 sda1: 965.834	2.9197623066052247
10.0.1.13	tidb*1	20/40	125.64	sda1: 498.638	2.917983893669314

图 139: Cluster Hardware 报表

上表中各个字段含义如下：

- HOST：服务器的 IP 地址。
- INSTANCE：该服务器部署的实例数量，如 pd * 1 代表这台服务器部署了 1 个 PD 实例。如 tidb * 2 pd * 1 表示这台服务器部署了 2 个 TiDB 实例和 1 个 PD 实例。
- CPU_CORES：表示服务器 CPU 的核心数，物理核心/逻辑核心。
- MEMORY：表示服务器的内存大小，单位是 GB。
- DISK：表示服务器磁盘大小，单位是 GB。
- UPTIME：服务器的启动时间，单位是 DAY。

Cluster Info

Cluster Info 为集群拓扑信息。表中信息来自 TiDB 的 information_schema.cluster_info 系统表。

Cluster Info

TYPE	INSTANCE	STATUS_ADDRESS	VERSION	GIT_HASH	START_TIME	UPTIME
pd	10.0.1.11:2379	10.0.1.11:2379	4.1.0-alpha	a80b99ef0d70807d106bdc1b7c6e97a118679c7e	2020-05-18T13:47:01Z	65h40m54.520385474s
tidb	10.0.1.13:4000	10.0.1.13:10080	4.0.0-beta.2	ac30f5322e253125002663ad7c789807acfc9305	2020-05-18T13:47:10Z	65h40m45.520383242s
tiflash	10.0.1.14:3930	10.0.1.14:20292	v4.1.0-alpha-37-gacf3f673d	acf3f673dfbc3e6ffff0dda575760670993fefa6	2020-05-18T13:47:18Z	65h40m37.520390633s
tikv	10.0.1.12:20160	10.0.1.12:20180	4.1.0-alpha	6b09cadbf55311ac07fc51e1b43e3b57b54e71f2	2020-05-18T13:47:04Z	65h40m51.520389948s

图 140: Cluster Info 报表

上表中各个字段含义如下：

- TYPE：节点类型。
- INSTANCE：实例地址，为 IP:PORT 格式的字符串。
- STATUS_ADDRESS：HTTP API 的服务地址。
- VERSION：对应节点的语义版本号。
- GIT_HASH：编译节点版本时的 Git Commit Hash，用于识别两个节点是否是绝对一致的版本。
- START_TIME：对应节点的启动时间。
- UPTIME：对应节点已经运行的时间。

诊断信息

TiDB 内置自动诊断的结果，具体各字段含义以及介绍可以参考 information_schema.inspection_result 系统表的内容。

负载信息

Node Load Info

Node Load Info 表显示服务器节点的负载信息，包括时间范围内，服务器以下指标的平均值 (AVG)、最大值 (MAX)、最小值 (MIN)：

- CPU 使用率，最大值是 100%
- 内存使用率
- 磁盘 I/O 使用率
- 磁盘写延迟
- 磁盘读延迟
- 磁盘每秒的读取字节数
- 磁盘每秒的写入字节数
- 节点网络每分钟收到的字节数
- 节点网络每分钟发送的字节数
- 节点正在使用的 TCP 连接数
- 节点所有的 TCP 连接数

Node Load Info

METRIC_NAME	instance	AVG	MAX	MIN
node_cpu_usage expand		40.42%	99.6%	6.98%
node_mem_usage expand		66.9%	96.28%	33.65%
node_disk_io_utilization expand		16%	97%	0%
Disk write latency expand		4000 us	50 ms	0 us
Disk read latency expand		2000 us	6000 us	0 us
tikv_disk_read_bytes expand		605.980 KB	15.842 MB	0.000 KB
tikv_disk_write_bytes expand		1.129 MB	12.235 MB	0.000 KB
node_network_in_traffic expand		225.706 KB	1.427 MB	0.000 KB
node_network_out_traffic expand		165.807 KB	1.659 MB	0.000 KB
node_tcp_in_use expand		19	47	5
node_tcp_connections expand		30	61	6

图 141: Node Load Info 报表

Instance CPU Usage

Instance CPU Usage 表显示各个 TiDB/PD/TiKV 进程的 CPU 使用率的平均值 (AVG)，最大值 (MAX)，最小值 (MIN)，这里进程 CPU 使用率最大值是 100% * CPU 逻辑核心数。

INSTANCE	JOB	AVG	MAX	MIN
172.16.5.40:10089	tidb	1891%	1906%	1878%
172.16.5.40:22151	tikv	517%	571%	459%
172.16.5.40:20151	tikv	498%	562%	444%
172.16.5.40:23151	tikv	352%	399%	311%
172.16.5.40:24799	pd	39%	45%	36%

图 142: Instance CPU Usage 报表

Instance Memory Usage

Instance Memory Usage 表显示各个 TiDB/PD/TiKV 进程占用内存字节数的平均值 (AVG)，最大值 (MAX)，最小值 (MIN)。

INSTANCE	JOB	AVG	MAX	MIN
172.16.5.40:20151	tikv	9.562 GB	9.605 GB	9.523 GB
172.16.5.40:23151	tikv	9.528 GB	9.574 GB	9.496 GB
172.16.5.40:22151	tikv	9.433 GB	9.498 GB	9.345 GB
172.16.5.40:10089	tidb	904.500 MB	911.660 MB	894.336 MB
172.16.5.40:24799	pd	61.089 MB	61.270 MB	60.980 MB

图 143: Instance Memory Usage 报表

TiKV Thread CPU Usage

TiKV Thread CPU Usage 表显示 TiKV 内部各个模块线程的 CPU 使用率的平均值 (AVG)、最大值 (MAX)、和最小值 (MIN)。这里进程 CPU 使用率最大值为 100% * 对应配置的线程数量。

TiKV Thread CPU Usage

METRIC_NAME	INSTANCE	AVG	MAX	MIN	CONFIG_KEY	CURRENT_CONFIG_VALUE
grpc expand		14.83%	16%	12.96%		
schedule worker fold		8.76%	9.47%	7.56%		
|-- sched_worker	10.0.1.12:20180	8.76%	9.47%	7.56%	storage.scheduler-worker-pool-size	4
storage read pool expand		3.24%	3.64%	2.67%		
storage read pool normal expand		3.22%	3.64%	2.67%		
Async apply fold		3.14%	3.4%	2.71%		
|-- Async apply	10.0.1.12:20180	3.14%	3.4%	2.71%	raftstore.apply-pool-size	2
raftstore expand		2.7%	2.93%	2.38%		
rocksdb expand		0.54%	2.13%	0%		
unified read pool expand		0.45%	4.24%	0%		
split_check expand		0.26%	0.84%	0%		
gc expand		0.04%	0.4%	0%		
storage read pool high expand		0.02%	0.04%	0%		
snapshot expand		0.004%	0.02%	0%		
cop normal						
cop high						
cop low						
storage read pool low expand		0%	0%	0%		
cop						

图 144: TiKV Thread CPU Usage 报表

上表中的字段解释如下：

- CONFIG_KEY：表示对应模块的相关线程数配置。
- CURRENT_CONFIG_VALUE：表示配置在生成报表时刻的当前值。

> **注意：**
>
> `CURRENT_CONFIG_VALUE` 是生成报告时的值，并不是报告时间范围内的值。目前不能获取历史时间某些配置的值。

TiDB/PD Goroutines Count

TiDB/PD Goroutines Count 表显示 TiDB/PD goroutines 数量的平均值 (AVG)，最大值 (MAX)，和最小值 (MIN)。如果 goroutines 数量超过 2000，说明该进程并发太高，会对整体请求的延迟有影响。

INSTANCE	JOB	AVG	MAX	MIN
172.16.5.40:10089	tidb	1434.33	1473	1394
172.16.5.40:24799	pd	157	157	157

图 145: TiDB/PD goroutines count 报表

概览信息

Time Consumed by Each Component

Time Consumed by Each Component 显示包括集群中 TiDB、PD、TiKV 各个模块的监控耗时以及各项耗时的占比。默认时间单位是秒。用户可以用该表快速定位哪些模块的耗时较多。

METRIC_NAME	LABEL	TIME_RATIO	TOTAL_TIME	TOTAL_COUNT	P999	P99	P90	P80
tidb_query expand		1	23223.86	459625	63.54	45.55	0.1	0.06
tidb_get_token expand		0.000001	0.03	458888	0.00004	0.000001	0.000001	0.000001
tidb_parse expand		0.00001	0.29	4603	0.0006	0.0004	0.0002	0.0001
tidb_compile expand		0.0008	17.82	463488	0.001	0.0009	0.0003	0.0002
tidb_execute expand		1	23237.39	463491	90.81	0.35	0.1	0.08
tidb_distsql_execution expand		0.0002	4.53	1064	0.13	0.1	0.02	0.01
tidb_cop expand		0.0005	12.39	1075	2.03	1.84	0.02	0.01
Transaction expand		1	23158.73	460023	8.16	7.91	5.38	0.06
tidb_transaction_local_latch_wait		0	0	0				
tidb_kv_backoff expand		0.00003	0.58	291	0.002	0.002	0.002	0.002
KV request expand		1.65	38325.58	2300949	2.03	1.84	0.04	0.03
Slow query expand		0.01	323.65	280	65.5	65.21	62.26	58.98
tidb_slow_query_cop_process		0	0	280	0.001	0.001	0.0009	0.0008
tidb_slow_query_cop_wait		0	0	280	0.001	0.001	0.0009	0.0008
DDL job		0	0	0				
tidb_ddl_worker		0	0	0				
tidb_ddl_update_self_version		0	0	0				
tidb_owner_handle_syncer		0	0	0				
Batch add index								
tidb_ddl_deploy_syncer		0	0	0				
Schema load expand		0.000008	0.19	27	0.06	0.06	0.06	0.05
TiDB meta operation expand		0.0001	3.21	1200	0.13	0.1	0.01	0.005
TiDB auto id request		0	0	0				
TiDB auto analyze expand		0.003	78.35	1	81.88	81.51	77.82	73.73
TiDB GC expand		0.000000001	0.00002	3	1	0.99	0.9	0.8
tidb_gc_push_task		0	0	0				
tidb_batch_client_unavailable		0	0	0				
tidb_batch_client_wait		0	0	0				
pd_tso_rpc expand		0.005	123.45	108540	0.02	0.008	0.003	0.002
pd_tso_wait expand		0.07	1662.29	920545	0.03	0.01	0.004	0.003
PD client cmd expand		0.15	3370.87	2761712	0.03	0.01	0.004	0.003
pd_client_request_rpc expand		0.005	123.45	108540	0.02	0.008	0.003	0.002
pd_grpc_completed_commands expand		0.00005	1.24	5255	0.01	0.01	0.006	0.005
pd_operator_finish								
pd_operator_step_finish								
Etcd transactions expand		0.000008	0.17	207	0.008	0.008	0.006	0.002
pd_region_heartbeat expand		0.00006	1.33	195	0	0	0	0
etcd_wal_fsync expand		0.00001	0.33	288	0.02	0.01	0.006	0.004
pd_peer_round_trip								

图 146: Total Time Consume 报表

上表各列的字段含义如下：

- METRIC_NAME：监控项的名称。
- Label：监控的 label 信息，点击 expand 后可以查看该项监控更加详细的各项 label 的监控信息。
- TIME_RATIO：该项为 TIME_RATIO 为 1 的监控行总时间与监控消耗的总时间的比例。如 kv_request 的总耗时占 tidb_query 总耗时的 1.65 = 38325.58/23223.86。因为 KV 请求会并行执行，所以所有 KV 请求的总时间有可能超过总查询 (tidb_query) 的执行时间。
- TOTAL_TIME：该项监控的总耗时。

- TOTAL_COUNT：该项监控执行的总次数。
- P999：该项监控的 P999 最大时间。
- P99：该项监控的 P99 最大时间。
- P90：该项监控的 P90 最大时间。
- P80：该项监控的 P80 最大时间。

以上监控中相关模块的耗时关系如下所示：

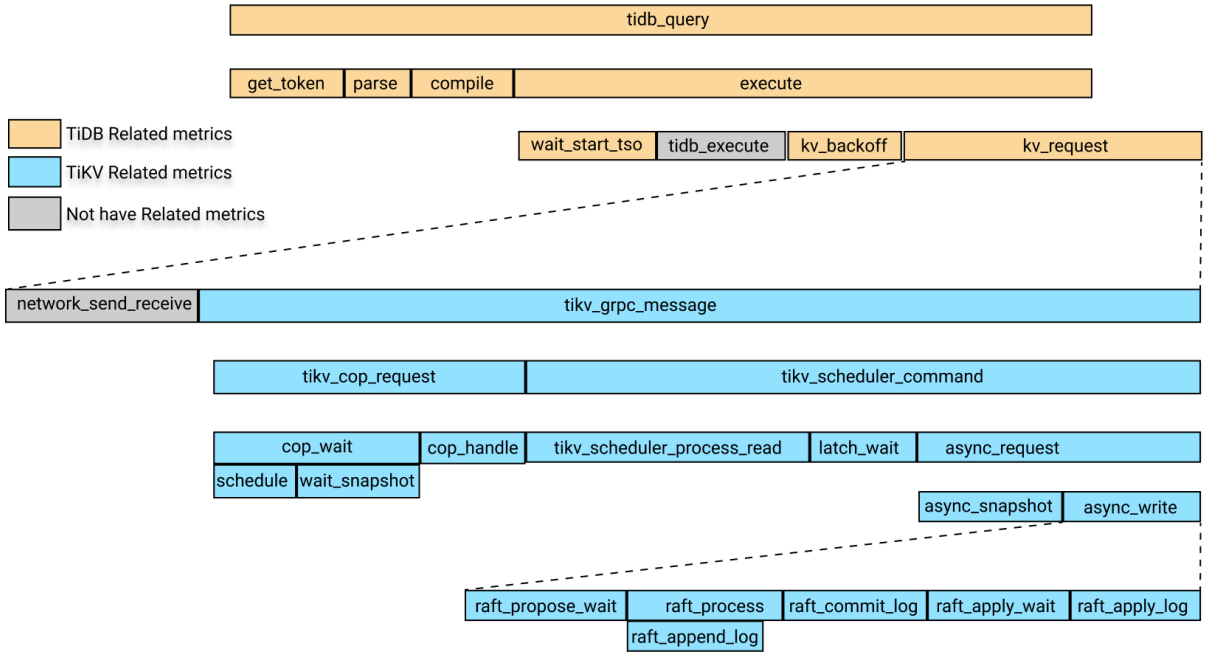


图 147: 各个模块耗时关系图

上图中，黄色部分是 TiDB 相关的监控，蓝色部分是 TiKV 相关的监控，灰色部分暂时没有具体对应的监控项。

上图中，tidb_query 的耗时包括以下部分的耗时：

- get_token
- parse
- compile
- execute

其中 execute 的耗时包括以下部分：

- wait_start_tso
- TiDB 层的执行时间，目前暂无监控
- KV 请求的时间
- tidb_kv_backoff 的时间，这是 KV 请求失败后进行 backoff 的时间

其中，KV 请求时间包含以下部分：

- 请求的网络发送以及接收耗时，目前该项暂无监控，可以大致用 KV 请求时间减去 `tikv_grpc_message` 的时间
- `tikv_grpc_message` 的耗时

其中，`tikv_grpc_message` 耗时包含以下部分：

- Coprocessor request 耗时，指用于处理 COP 类型的请求，该耗时包括以下部分：
  - `tikv_cop_wait`，请求排队等待的耗时
  - `Coprocessor handling request`，处理 COP 请求的耗时
- `tikv_scheduler_command` 耗时，该耗时包含以下部分：
  - `tikv_scheduler_processing_read`，处理读请求的耗时
  - `tikv_storage_async_request` 中获取 snapshot 的耗时（snapshot 是该项监控的 label）
  - 处理写请求的耗时，该耗时包括以下部分：
    * `tikv_scheduler_latch_wait`，等待 latch 的耗时
    * `tikv_storage_async_request` 中 write 的耗时（write 是该监控的 label）

其中，`tikv_storage_async_request` 中的 write 耗时是指 raft kv 写入的耗时，包括以下部分：

- `tikv_raft_propose_wait`
- `tikv_raft_process`，该耗时主要时间包括：
  - `tikv_raft_append_log`
  - `tikv_raft_commit_log`
  - `tikv_raft_apply_wait`
  - `tikv_raft_apply_log`

用户可以根据上述耗时之间的关系，利用 `TOTAL_TIME` 以及 P999，P99 的时间大致定位哪些模块耗时比较长，然后再看相关的监控。

> 注意：
>
> 由于 Raft KV 可能会将多个请求作为一个 batch 来写入，所以 `TOTAL_TIME` 不适用于来衡量 Raft KV 的写入相关监控项的耗时，这些监控项具体是 `tikv_raft_process`、`tikv_raft_append_log`、`tikv_raft_commit_log`、`tikv_raft_apply_wait`、`tikv_raft_apply_log`。此时用 P999 和 P99 的时间来对比各个模块的耗时更加合理。
>
> 原因是，假如有 10 个 async write 请求，Raft KV 内部将 10 个请求打包成一个 batch 执行，执行时间为 1 秒，所以每个请求的执行时间为 1 秒，10 个请求的总时间是 10 秒。但是 Raft KV 处理的总时间是 1 秒。如果用 `TOTAL_TIME` 来衡量，用户可能不明白剩余的 9 秒耗时在哪些模块下。这里从总请求数 (`TOTAL_COUNT`) 也能看出 Raft KV 的监控和其他监控的差异。

Errors Occurred in Each Component

Errors Occurred in Each Component 表显示包括 TiDB 和 TiKV 出现错误的总数。例如写 binlog 失败、`tikv server is busy`、`TiKV channel full`、`tikv write stall` 等错误，具体各项错误含义可以看行注释。

METRIC_NAME	LABEL	TOTAL_COUNT
tidb_binlog_error_total_count		0
tidb_handshake_error_total_count		0
tidb_transaction_retry_error_total_count		
tidb_kv_region_error_total_count expand		903
tidb_schema_lease_error_total_count		0
tikv_grpc_error_total_count		0
tikv_critical_error_total_count		
tikv_scheduler_is_busy_total_count		
tikv_channel_full_total_count		
tikv_coprocessor_request_error_total_count expand		1
tikv_engine_write_stall		0
tikv_server_report_failures_total_count expand		1396
tikv_storage_async_request_error expand		1176
tikv_lock_manager_detect_error_total_count		
tikv_backup_errors_total_count		
node_network_in_errors_total_count		0
node_network_out_errors_total_count		0

图 148: Errors Occurred in Each Component 报表

TiDB/PD/TiKV 的具体监控信息

这部分包括了 TiDB/PD/TiKV 更多的具体的监控信息。

TiDB 相关监控信息

Time Consumed by TiDB Component

Time Consumed by TiDB Component 表显示 TiDB 的各项监控耗时以及各项耗时的占比。和Time Consumed by Each Component 表类似，但是这个表的 label 信息会更丰富，细节更多。

TiDB Server Connections

TiDB Server Connections 表显示 TiDB 各个实例的客户端连接数。

TiDB Transaction

TiDB Transaction 表显示 TiDB 事务相关的监控。

METRIC_NAME	LABEL	TOTAL_VALUE	TOTAL_COUNT	P999	P99	P90	P80
tidb_transaction_retry_num		0	0	0	0	0	0
tidb_transaction_statement_num expand		6854588	6852371	4	4	3	2
tidb_txn_region_num expand		1070770	706887	3	2	2	2
tidb_txn_kv_write_num expand		513386	181312	234	2	2	2
tidb_txn_kv_write_size expand		266.772 MB	181296	116.913 KB	1.996 KB	1.905 KB	1.805 KB
tidb_load_safepoint_total_num expand		16					
tidb_lock_resolver_total_num expand		420514					

图 149: Transaction 报表

- TOTAL_VALUE：该项监控在报告时间段内所有值的和 (SUM)。
- TOTAL_COUNT：该项监控出现的总次数。
- P999: 该项监控的 P999 最大值。
- P99: 该项监控的 P99 最大值。
- P90: 该项监控的 P90 最大值。
- P80: 该项监控的 P80 最大值。

示例：

上表中，在报告时间范围的 tidb_txn_kv_write_size 表示一共约有 181296 次事务的 KV 写入，总 kv 写入大小是 266.772 MB，其中单次事务的 KV 写入的 P999、 P99、 P90、 P80 的最大值分别为 116.913 KB、 1.996 KB、 1.905 KB、 1.805 KB。

DDL Owner

MIN_TIME	DDL OWNER
2020-05-21 14:40:00	10.0.1.13:10080

图 150: TiDB DDL Owner 报表

上表表示从 2020-05-21 14:40:00 开始，集群的 DDL owner 在 10.0.1.13:10080 节点。如果 owner 发生变更，上表会有多行数据，其中 MinTime 列表示已知对应 Owner 的最小时间。

> 注意：
>
> 如果 owner 信息为空，不代表这个时间段内一定没有 owner，因为这里是依靠 ddl_worker 的监控信息来判断 DDL owner 的。也可能是这个时间段内 ddl_worker 没有做任何 DDL job 导致 owner 信息为空。

TiDB 中其他部分监控表如下：

- Statistics Info：TiDB 统计信息的相关监控。
- Top 10 Slow Query：报表时间范围内 Top 10 的慢查询信息。
- Top 10 Slow Query Group By Digest：报表时间范围内 Top 10 的慢查询信息，并按照 SQL 指纹聚合。
- Slow Query With Diff Plan：报表时间范围内执行计划发生变更的 SQL 语句。

PD 相关监控信息

PD 模块相关监控的报表如下：

- Time Consumed by PD Component：PD 中相关模块的耗时监控
- Scheduled Leader/Region：报表时间范围内集群发生的 `balance-region` 和 `balance leader` 监控，比如从 `tikv_note_1` 上调度走了多少个 leader，调度进了多少个 leader。
- Cluster Status：集群的状态信息，包括总 TiKV 数量、总集群存储容量、Region 数量、离线 TiKV 的数量等信息。
- Store Status：记录各个 TiKV 节点的状态信息，包括 Region score、 leader score、 Region/leader 的数量。
- etcd Status：PD 内部的 etcd 相关信息。

TiKV 相关监控信息

TIKV 模块的相关监控报表如下：

- Time Consumed by TiKV Component：TiKV 中相关模块的耗时监控。
- Time Consumed by RocksDB：TiKV 中 RocksDB 的耗时监控。
- TiKV Error：TiKV 中各个模块相关的 error 信息。
- TiKV Engine Size：TiKV 中各个节点 column family 的存储数据大小。
- Coprocessor Info：TiKV 中 Coprocessor 模块相关的监控。
- Raft Info：TiKV 中 Raft 模块的相关监控信息。
- Snapshot Info：TiKV 中 snapshot 相关监控信息。
- GC Info：TiKV 中 GC 相关的监控信息。
- Cache Hit：TiKV 中 Rocksdb 的各个缓存的命中率监控信息。

配置信息

配置信息中，部分模块的配置信息可以显示报告时间范围内的配置值，有部分配置则因为无法获取到历史的配置值，所以是生成报告时刻的当前配置值。

在报告时间范围内，以下表包括部分配置的在报告时间范围的开始时间的值：

- Scheduler Initial Config：PD 调度相关配置在报告开始时间的初始值。
- TiDB GC Initial Config：TiDB GC 相关配置在报告开始时间的初始值。
- TiKV RocksDB Initial Config：TiKV RocksDB 相关配置在报告开始时间的初始值。
- TiKV RaftStore Initial Config：TiKV RaftStore 相关配置在报告开始时间的初始值。

在报表时间范围内，如若有些配置被修改过，以下表包括部分配置被修改的记录：

- Scheduler Config Change History
- TiDB GC Config Change History

- TiKV RocksDB Config Change History
- TiKV RaftStore Config Change History

示例：

APPROXIMATE_CHANGE_TIME	CONFIG_ITEM	VALUE
2020-05-22T20:00:00+08:00	leader-schedule-limit	4
2020-05-22T20:07:00+08:00	leader-schedule-limit	8

图 151: Scheduler Config Change History 报表

上面报表显示，`leader-schedule-limit` 配置参数在报告时间范围内有被修改过：

- `2020-05-22T20:00:00+08:00`，即报告的开始时间 `leader-schedule-limit` 的配置值为 `4`，这里并不是指该配置被修改了，只是说明在报告时间范围的开始时间其配置值是 4。
- `2020-05-22T20:07:00+08:00`，`leader-schedule-limit` 的配置值为 `8`，说明在 `2020-05-22T20:07:00+08:00` 左右，该配置的值被修改了。

下面的报表是生成报告时，TiDB、PD、TiKV 的在生成报告时刻的当前配置：

- TiDB’s Current Config
- PD’s Current Config
- TiKV’s Current Config

9.9.11.2.2 对比报告

生成两个时间段的对比报告，其内容和单个时间段的报告是一样的，只是加入了对比列显示两个时间段的差别。下面主要介绍对比报告中的一些特有表以及如何查看对比报表。

首先在基本信息中的 Compare Report Time Range 报表会显示出对比的两个时间段：

Compare Report Time Range

T1.START_TIME	T1.END_TIME	T2.START_TIME	T2.END_TIME
2020-05-21 06:30:00	2020-05-21 06:35:00	2020-05-21 06:40:00	2020-05-21 06:45:00

图 152: Compare Report Time Range 报表

其中 `t1` 是正常时间段，或者叫参考时间段，`t2` 是异常时间段。

下面是一些慢查询相关的报表：

- Slow Queries In Time Range t2：仅出现在 `t2` 时间段但没有出现在 `t1` 时间段的慢查询。
- Top 10 slow query in time range t1： `t1` 时间段的 Top10 慢查询。
- Top 10 slow query in time range t2： `t2` 时间段的 Top10 慢查询。

DIFF_RATIO 介绍

本部分以 Instance CPU Usage 为例介绍 DIFF_RATIO。

Instance CPU Usage

INSTANCE	JOB	t1.AVG	t1.MAX	t1.MIN	t2.AVG	t2.MAX	t2.MIN	AVG_DIFF_RATIO	MAX_DIFF_RATIO	MIN_DIFF_RATIO
172.16.5.40:10089	tidb	410%	410%	410%	1240%	1240%	1240%	2.02	2.02	2.02
172.16.5.40:20151	tikv	221%	221%	221%	617%	617%	617%	1.79	1.79	1.79
172.16.5.40:20379	pd	82%	82%	82%	78%	78%	78%	-0.05	-0.05	-0.05

图 153: Compare Instance CPU Usage 报表

- t1.AVG、t1.MAX、t1.Min 分别是 t1 时间段内 CPU 使用率的平均值、最大值、最小值。
- t2.AVG、t2.MAX、t2.Min 分别是 t2 时间段内 CPU 使用率的平均值、最大值、最小值。
- AVG_DIFF_RATIO 表示 t1 和 t2 时间段平均值的 DIFF_RATIO。
- MAX_DIFF_RATIO 表示 t1 和 t2 时间段最大值的 DIFF_RATIO。
- MIN_DIFF_RATIO 表示 t1 和 t2 时间段最小值的 DIFF_RATIO。

DIFF_RATIO 表示两个时间段的差异大小，有以下几个取值方式：

- 如果该监控仅在 t2 时间内才有值，t1 时间段没有，则 DIFF_RATIO 取值为 1。
- 如果监控项仅在 t1 时间内才有值，t1 时间段没有，则 DIFF_RATIO 取值为 -1。
- 如果 t2 时间段的值比 t1 时间段的值大，则 DIFF_RATIO = (t2.value / t1.value)- 1。
- 如果 t2 时间段的值比 t1 时间段的值小，则 DIFF_RATIO = 1 - (t1.value / t2.value)。

例如上表中，tidb 节点的平均 CPU 使用率在 t2 时间段比 t1 时间段高 2.02 倍，2.02 = 1240/410 - 1。

Maximum Different Item 报表介绍

Maximum Different Item 的报表是对比两个时间段的监控项后，按照监控项的差异大小排序，通过这个表可以很快发现两个时间段哪些监控的差异最大。示例如下：

Maximum Different Item

The maximum different metrics between two time ranges.

TABLE	METRIC_NAME	LABEL	MAX_DIFF	t1.VALUE	t2.VALUE	VALUE_TYPE
TiKV, coprocessor_info Expand	TiKV Coprocessor scan	172.16.5.40:20151,select,get,lock	28916.85	61.33	1773532	TOTAL_VALUE
TiDB, statistics_info	store_query_feedback_total_count	172.16.5.40:10089,ok	7219.00		7219	TOTAL_COUNT
overview, total_time_consume	tidb_parse	general	1201.00		1201	TOTAL_COUNT
TiDB, tidb_time_consume	tidb_slow_query_cop_wait	172.16.5.40:10089	800.00		800	TOTAL_COUNT
overview, total_time_consume	tidb_slow_query_cop_process	172.16.5.40:10089	800.00		800	TOTAL_COUNT
overview, total_time_consume	Slow query	172.16.5.40:10089	799.00	1	800	TOTAL_COUNT
TiKV, coprocessor_info	TiKV Coprocessor response	172.16.5.40:20151	534.43	15.822 MB	8.273 GB	TOTAL_VALUE
overview, total_time_consume	tidb_distsql_execution	dag	470.06	0.68	320.32	TOTAL_TIME
load, tikv_thread_cpu_usage	unified read pool	172.16.5.40:20151	412.29	0.07%	28.93%	MIN
TiKV, coprocessor_info Expand	TiKV Coprocessor scan keys	172.16.5.40:20151,select	198.39	1067055.38	212755381.33	TOTAL_VALUE
overview, total_time_consume Expand	tidb_ddl_worker	drop view	139.00		139	TOTAL_COUNT
overview, total_error Expand	tikv_storage_async_request_error	snapshot	89.00		89	TOTAL_COUNT
overview, total_time_consume Expand	Coprocessor handling request	index	78.99	0.67	53.59	TOTAL_TIME
overview, total_time_consume	tidb_cop	172.16.5.40:10089	75.11	5.93	451.31	TOTAL_TIME
overview, total_time_consume	KV request	Cop	75.10	5.93	451.26	TOTAL_TIME
overview, total_time_consume Expand	Coprocessor request	select	73.12	4.73	350.6	TOTAL_TIME
PD, pd_time_consume Expand	pd_region_heartbeat	172.16.5.40:20150,1	-73.00	73		TOTAL_COUNT
overview, total_time_consume Expand	tikv_grpc_messge	coprocessor	71.75	5.56	404.49	TOTAL_TIME
TiDB, tidb_time_consume	tidb_ddl_update_self_version	172.16.5.40:10089,ok	71.00		71	TOTAL_COUNT
overview, total_time_consume Expand	TiDB meta operation	update_ddl_job	71.00		71	TOTAL_COUNT
overview, total_time_consume Expand	tidb_owner_handle_syncer	update_global_version	71.00		71	TOTAL_COUNT
TiDB, tidb_time_consume Expand	tidb_query	172.16.5.40:10089,internal	-62.84	15.96	0.25	P999
TiKV, scheduler_info Expand	Scheduler stage	172.16.5.40:20151,scan_lock,snapshot_ok	55.45		55.45	TOTAL_VALUE
overview, total_time_consume Expand	tikv_grpc_messge	kv_scan_lock	55.00		55	TOTAL_COUNT
TiKV, tikv_time_consume Expand	tikv_scheduler_command	172.16.5.40:20151,scan_lock	55.00		55	TOTAL_COUNT
overview, total_time_consume Expand	PD client cmd	get_region	44.00		44	TOTAL_COUNT
overview, total_time_consume Expand	tikv_cop_wait	select	39.40	171	6908	TOTAL_COUNT
TiDB, tidb_time_consume	Schema load	172.16.5.40:10089	37.50	0.02	0.77	TOTAL_TIME
overview, total_time_consume	Schema load	172.16.5.40:10089	30.25	0.008	0.25	P999

图 154: Maximum Different Item 报表

- `Table`：表示这个监控项来自于对比报告中报表，如 `TiKV, coprocessor_info` 表示是 TiKV 组件下的 `coprocessor_info` 报表。
- `METRIC_NAME`：监控项名，点击 expand 可以查看该监控的不同 label 的差异对比。
- `LABEL`：监控项对应的 label。比如 `TiKV Coprocessor scan` 监控项有两个 label，分别是 instance、req、tag、sql_type，分别表示为 TiKV 地址、请求类型、操作类型和操作的 column family。
- `MAX_DIFF`：差异大小，取值为 `t1.VALUE` 和 `t2.VALUE` 的 `DIFF_RATIO` 计算。

可以从上表中发现，`t2` 时间段比 `t1` 时间段多出了大量的 Coprocessor 请求，TiDB 的解析 SQL (parse) 时间也多了很多。

9.9.11.3 使用 TiDB Dashboard 诊断报告定位问题

本文介绍如何使用 TiDB Dashboard 诊断报告定位问题。

9.9.11.3.1 对比诊断功能示例

对比报告中对比诊断的功能，通过对比两个时间段的监控项差异来尝试帮助 DBA 定位问题。先看以下示例。

大查询/写入导致 QPS 抖动或延迟上升诊断

示例 1

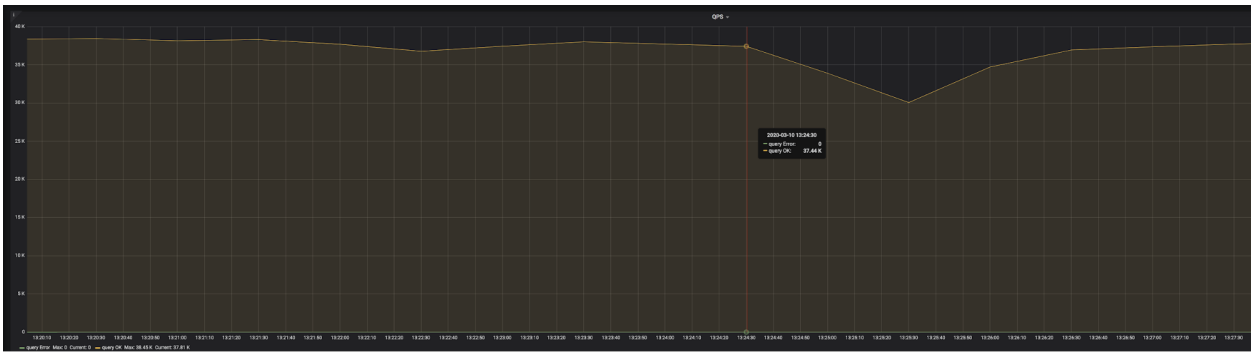


图 155: QPS 图

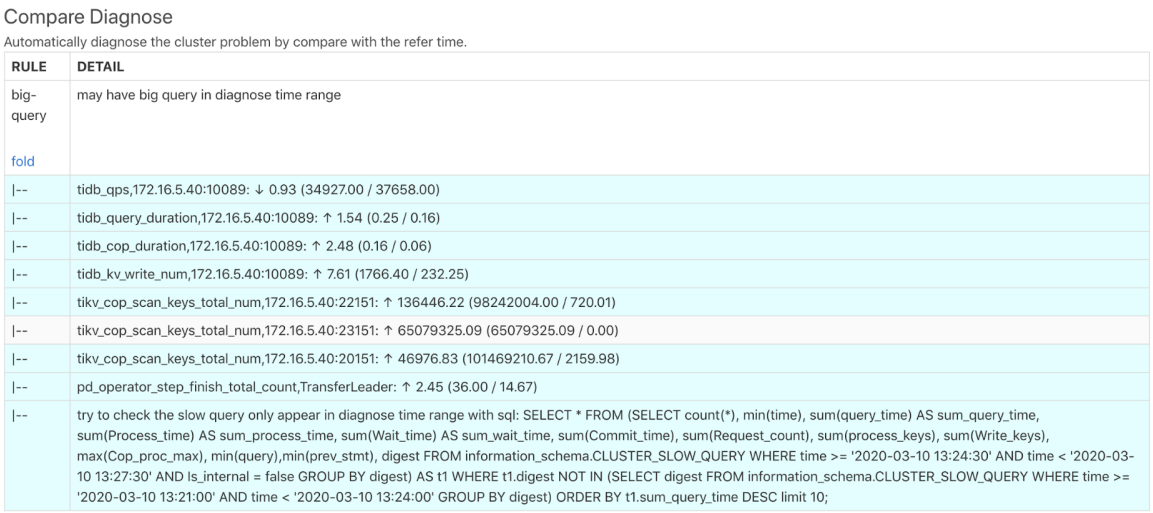
上图是 go-ycsb 压测的监控。可以发现，在 2020-03-10 13:24:30 时，QPS 突然开始下降，3 分钟后，QPS 开始恢复正常。

生成以下两个时间范围的对比报告：

- t1: 2020-03-10 13:21:00 - 2020-03-10 13:23:00，正常时间段，又叫参考时间段。
- t2: 2020-03-10 13:24:30 - 2020-03-10 13:27:30，QPS 开始下降的异常时间段。

这里两个时间间隔都是 3 分钟，因为抖动的影响范围为 3 分钟。因为诊断时会用一些监控的平均值做对比，所有间隔时间太长会导致平均值差异不明显，无法准确定位问题。

生成报告后查看 Compare Diagnose 报表内容如下：



Compare Diagnose

Automatically diagnose the cluster problem by compare with the refer time.

RULE	DETAIL
big-query fold	may have big query in diagnose time range
|--	tidb_qps,172.16.5.40:10089: ↓ 0.93 (34927.00 / 37658.00)
|--	tidb_query_duration,172.16.5.40:10089: ↑ 1.54 (0.25 / 0.16)
|--	tidb_cop_duration,172.16.5.40:10089: ↑ 2.48 (0.16 / 0.06)
|--	tidb_kv_write_num,172.16.5.40:10089: ↑ 7.61 (1766.40 / 232.25)
|--	tikv_cop_scan_keys_total_num,172.16.5.40:22151: ↑ 136446.22 (98242004.00 / 720.01)
|--	tikv_cop_scan_keys_total_num,172.16.5.40:23151: ↑ 65079325.09 (65079325.09 / 0.00)
|--	tikv_cop_scan_keys_total_num,172.16.5.40:20151: ↑ 46976.83 (101469210.67 / 2159.98)
|--	pd_operator_step_finish_total_count,TransferLeader: ↑ 2.45 (36.00 / 14.67)
|--	try to check the slow query only appear in diagnose time range with sql: SELECT * FROM (SELECT count(*), min(time), sum(query_time) AS sum_query_time, sum(Process_time) AS sum_process_time, sum(Wait_time) AS sum_wait_time, sum(Commit_time), sum(Request_count), sum(process_keys), sum(Write_keys), max(Cop_proc_max), min(query),min(prev_stmt), digest FROM information_schema.CLUSTER_SLOW_QUERY WHERE time >= '2020-03-10 13:24:30' AND time < '2020-03-10 13:27:30' AND Is_internal = false GROUP BY digest) AS t1 WHERE t1.digest NOT IN (SELECT digest FROM information_schema.CLUSTER_SLOW_QUERY WHERE time >= '2020-03-10 13:21:00' AND time < '2020-03-10 13:24:00' GROUP BY digest) ORDER BY t1.sum_query_time DESC limit 10;

图 156: 对比诊断结果

上面诊断结果显示，在诊断的时间内可能有大查询，下面的每一行的含义是：

- `tidb_qps`：QPS 下降 0.93 倍。
- `tidb_query_duration`：P999 的查询延迟上升 1.54 倍。
- `tidb_cop_duration`：P999 的 COP 请求的处理延迟上升 2.48 倍。
- `tidb_kv_write_num`：P999 的 tidb 的事务写入 kv 数量上升 7.61 倍。
- `tikv_cop_scan_keys_total_nun`：TiKV 的 coprocessor 扫描 key/value 的数量分别在 3 台 TiKV 上有很大的提升。
- `pd_operator_step_finish_total_count` 中，transfer leader 的数量上升 2.45 倍，说明异常时间段的调度比正常时间段要高。
- 提示可能有慢查询，并提示用 SQL 查询 TiDB 的慢日志。在 TiDB 中执行结果如下：

```
SELECT * FROM (SELECT count(*), min(time), sum(query_time) AS sum_query_time, sum(Process_time)
    ↪ AS sum_process_time, sum(Wait_time) AS sum_wait_time, sum(Commit_time), sum(Request_count
    ↪ ), sum(process_keys), sum(Write_keys), max(Cop_proc_max), min(query),min(prev_stmt),
    ↪ digest FROM information_schema.CLUSTER_SLOW_QUERY WHERE time >= '2020-03-10 13:24:30' AND
    ↪  time < '2020-03-10 13:27:30' AND Is_internal = false GROUP BY digest) AS t1 WHERE t1.
    ↪ digest NOT IN (SELECT digest FROM information_schema.CLUSTER_SLOW_QUERY WHERE time >= '
    ↪ 2020-03-10 13:21:00' AND time < '2020-03-10 13:24:00' GROUP BY digest) ORDER BY t1.
    ↪ sum_query_time DESC limit 10\G
***************************[ 1. row ]***************************
count(*)            | 196
min(time)           | 2020-03-10 13:24:30.204326
sum_query_time      | 46.878509117
sum_process_time    | 265.924
sum_wait_time       | 8.308
sum(Commit_time)    | 0.926820886
sum(Request_count) | 6035
```

```
sum(process_keys)   | 201453000
sum(Write_keys)     | 274500
max(Cop_proc_max)   | 0.263
min(query)          | delete from test.tcs2 limit 5000;
min(prev_stmt)      |
digest              | 24bd6d8a9b238086c9b8c3d240ad4ef32f79ce94cf5a468c0b8fe1eb5f8d03df
```

可以发现，从 13:24:30 开始有一个批量删除的大写入，一共执行了 196 次，每次删除 5000 行数据，总共耗时 46.8 秒。

示例 2

如果大查询一直没执行完，就不会记录慢日志，但仍可以进行诊断，示例如下：



图 157: QPS 图

上图中，也是在跑 go-ycsb 的压测。可以发现，在 2020-03-08 01:46:30 时，QPS 突然开始下降，并且一直没有恢复。

生成以下两个时间范围的对比报告：

- t1: 2020-03-08 01:36:00 - 2020-03-08 01:41:00，正常时间段，又叫参考时间段。
- t2: 2020-03-08 01:46:30 - 2020-03-08 01:51:30，QPS 开始下降的异常时间段。

生成报告后看 Compare Diagnose 报表的内容如下：

Compare Diagnose

Automatically diagnose the cluster problem by compare with the refer time.

RULE	DETAIL
big-query fold	may have big query in diagnose time range
|--	tidb_qps,172.16.5.40:10089: ↓ 0.85 (31881.00 / 37674.00)
|--	tidb_query_duration,172.16.5.40:10089: ↑ 1.35 (0.06 / 0.04)
|--	tidb_cop_duration,172.16.5.40:10089: ↑ 3.78 (0.08 / 0.02)
|--	tidb_kv_write_num,172.16.5.40:10089: ↑ 511.74 (511.74 / 0.00)
|--	tikv_cop_scan_keys_total_num,172.16.5.40:20151: ↑ 1277.21 (6270285.33 / 4909.36)
|--	try to check the slow query only appear in diagnose time range with sql: SELECT * FROM information_schema.cluster_log WHERE type='tidb' AND time >= '2020-03-08 01:46:30' AND time < '2020-03-08 01:51:30' AND level = 'warn' AND message LIKE '%expensive_query%'

图 158: 对比诊断结果

上面诊断结果的最后一行显示可能有慢查询，并提示用 SQL 查询 TiDB 日志中的 expensive query。在 TiDB 中执行结果如下：

```
> SELECT * FROM information_schema.cluster_log WHERE type='tidb' AND time >= '2020-03-08 01:46:30
    ↪ ' AND time < '2020-03-08 01:51:30' AND level = 'warn' AND message LIKE '%expensive_query%
    ↪ '\G
TIME     | 2020/03/08 01:47:35.846
TYPE     | tidb
INSTANCE | 172.16.5.40:4009
LEVEL    | WARN
MESSAGE  | [expensivequery.go:167] [expensive_query] [cost_time=60.085949605s] [process_time=2.52
    ↪ s] [wait_time=2.52s] [request_count=9] [total_keys=996009] [process_keys=996000] [
    ↪ num_cop_tasks=9] [process_avg_time=0.28s] [process_p90_time=0.344s] [process_max_time
    ↪ =0.344s] [process_max_addr=172.16.5.40:20150] [wait_avg_time=0.000777777s] [wait_p90_time
    ↪ =0.003s] [wait_max_time=0.003s] [wait_max_addr=172.16.5.40:20150] [stats=t_wide:pseudo] [
    ↪ conn_id=19717] [user=root] [database=test] [table_ids="[80,80]"] [txn_start_ts
    ↪ =415132076148785201] [mem_max="23583169 Bytes (22.490662574768066 MB)"] [sql="select
    ↪ count(*) from t_wide as t1 join t_wide as t2 where t1.c0>t2.c1 and t1.c2>0"]
```

以上查询结果显示，在 172.16.5.40:4009 这台 TiDB 上，2020/03/08 01:47:35.846 有一个已经执行了 60s 但还没有执行完的 expensive_query。

9.9.11.3.2 用对比报告定位问题

诊断有可能是误诊，使用对比报告或许可以帮助 DBA 更快速的定位问题。参考以下示例。

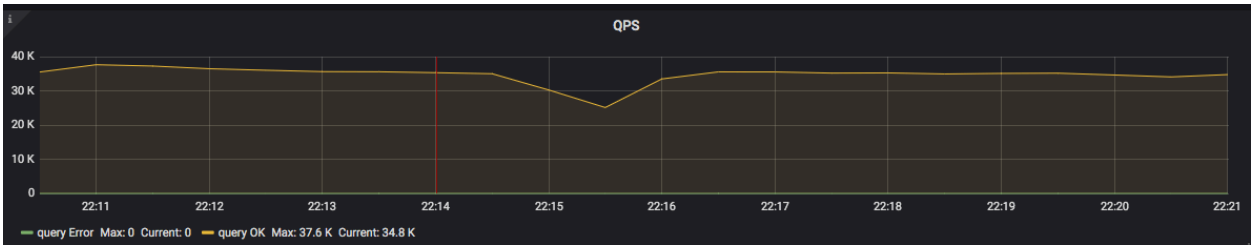


图 159: QPS 图

上图中，也是在跑 go-ycsb 的压测，可以发现，在 2020-05-22 22:14:00 时，QPS 突然开始下降，大概在持续 3 分钟后恢复。

生成以下 2 个时间范围的对比报告：

- t1: 2020-05-22 22:11:00 - 2020-05-22 22:14:00，正常时间段。
- t2: 2020-05-22 22:14:00 - 2020-05-22 22:17:00，QPS 开始下降的异常时间段。

生成对比报告后，查看 Max diff item 报表，该报表对比两个时间段的监控项后，按照监控项的差异大小排序，这个表的结果如下：

Maximum Different Item

The maximum different metrics between two time ranges.

TABLE	METRIC_NAME	LABEL	MAX_DIFF	t1.VALUE	t2.VALUE	VALUE_TYPE
TiDB, transaction	Transaction KV write count	172.16.5.40:10089	-21616.00	43234	2	P999
TiKV, coprocessor_info Expand	TiKV Coprocessor scan	172.16.5.40:20151,select,get,lock	10499.24	64	672015.48	TOTAL_VALUE
TiKV, coprocessor_info Expand	TiKV Coprocessor scan keys	172.16.5.40:20151,select	4264.45	15431.85	65823838.67	TOTAL_VALUE
TiDB, transaction	Transaction KV write size	172.16.5.40:10089	-2702.69	5.278 MB	1.999 KB	P999
overview, total_time_consume	Coprocessor handling request	select	2132.50	0.06	128.01	TOTAL_TIME
TiKV, coprocessor_info	TiKV Coprocessor response	172.16.5.40:20151	1535.00	1.792 MB	2.688 GB	TOTAL_VALUE
TiDB, statistics_info	store_query_feedback_total_count	172.16.5.40:10089,ok	1447.00		1447	TOTAL_COUNT
TiKV, scheduler_info Expand	tikv_scheduler_keys_written	172.16.5.40:20151,commit	-860.00	861	1	P99
TiKV, tikv_time_consume	Coprocessor request	172.16.5.40:20151,select	853.87	0.15	128.23	TOTAL_TIME
load, tikv_thread_cpu_usage	unified read pool	172.16.5.40:20151	667.27	0.11%	73.51%	AVG
TiKV, tikv_time_consume	Coprocessor handling request	172.16.5.40:20151,index	444.00	0.002	0.89	P999
TiKV, tikv_time_consume Expand	tikv_grpc_messge	172.16.5.40:20151,coprocessor	419.97	0.33	138.92	TOTAL_TIME
load, node_load_info	tikv_disk_read_bytes	172.16.5.40:19110,sda	-273.41	0.267 KB	0.000 KB	MAX
overview, total_time_consume Expand	KV request	Cop	244.29	0.66	161.89	TOTAL_TIME

图 160: 对比结果

从上面结果可以看出 t2 时间段新增了很多 Coprocessor 请求，可以猜测可能是 t2 时间段出现了一些大查询，或者是查询较多的负载。

实际上，在 t1 - t2 整个时间段内都在进行 go-ycsb 的压测，然后在 t2 时间段跑了 20 个 tpch 的查询，所以是因为 tpch 大查询导致了出现很多的 cop 请求。

这种大查询执行时间超过慢日志的阈值后也会记录在慢日志里面，可以继续查看 Slow Queries In Time ↪ Range t2 报表查看是否有一些慢查询。一些在 t1 时间段存在的查询，可能在 t2 时间段内就变成了慢查询，是因为 t2 时间段的其他负载影响导致该查询的执行变慢。

9.9.12 TiDB Dashboard 日志搜索页面

该页面上允许用户在集群中搜索所有节点上的日志，在页面上预览搜索结果和下载日志。

9.9.12.1 访问

登录 Dashboard 后点击左侧导航的日志搜索可以进入此功能页面：

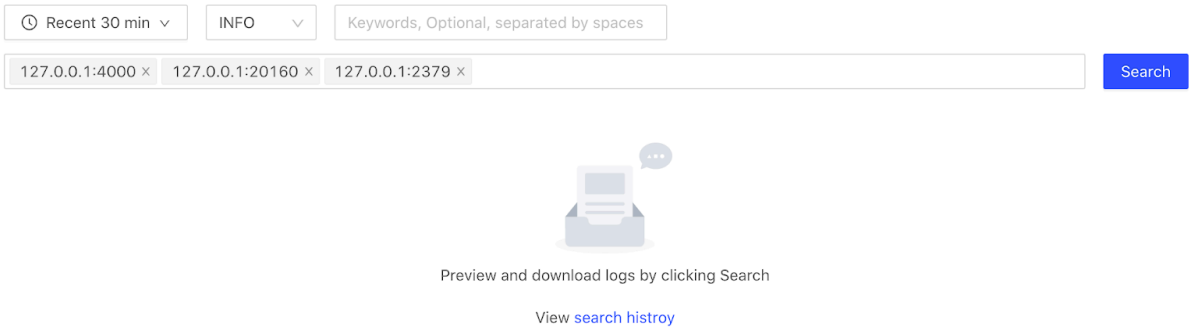


图 161: 日志搜索首页

该页面提供 4 个搜索参数，包括：

- 时间范围：限定搜索时间范围内的日志，默认值为最近 30 分钟
- 日志等级：限定最低日志等级，搜索该日志等级以上的所有日志，默认值为 INFO 等级
- 关键词：任何合法的任何字符串，可选，关键词之间以空格分割，支持正则表达式（大小写不敏感）
- 组件：选择要搜索的集群组件，多选，非空，默认选中所有组件

点击搜索按钮后，会进入搜索结果的详情页。

9.9.12.2 搜索结果详情

搜索结果详情页面如下图所示：



图 162: 搜索结果

整个页面分为三个部分：

- 搜索参数选项：与搜索页面的参数选项相同，可以在表单中重新选择参数并开始一次新的搜索（对应上图中框选的区域 1）
- 搜索进度：在页面右侧，展示当前搜索的进度，包括每个节点日志搜索状态和统计等（对应上图中框选的区域 2）
- 搜索结果（对应上图中框选的区域 3）：
  - 时间：日志产生的时间，时区与前端用户所处时区相同
  - 日志等级：日志等级
  - 组件类型：显示组件名及其地址
  - 日志：每条日志记录的正文部分，不包含日志的时间和日志等级。过长的日志会自动截断，鼠标单击该行可以展开查看完整内容，完整日志最长显示 512 个字符

> 注意：
>
> 在该页面上最多只显示 500 条搜索结果，完整的搜索结果可以通过下载得到。

9.9.12.2.1 搜索进度

在搜索进度区域中，对一个节点的搜索称为一次搜索任务，对搜索任务的状态分为：

- 运行中：开始搜索后，所有任务会进入运行中状态

- 成功：任务完成后自动转到成功，此时日志已缓存在 Dashboard 后端所在的本地磁盘中，可以提供给前端下载
- 失败：用户主动取消，或者某种原因报错退出的任务进入失败状态。任务失败时会自动清理本地临时文件

搜索进度区域包含三个控制按钮：

- 下载选中日志：下载被勾选组件的日志（只有已完成的才能被勾选），返回 tar 文件，解压得到一个或多个 zip 文件（每个组件对应一个 zip 文件），解压 zip 得到 log 文本文件
- 取消：取消所有正在运行的任务，只能在有运行中的任务时才能点击
- 重试：重试所有失败的任务，只有在有失败任务且无正在运行的任务时才能点击

9.9.12.3 搜索历史列表

在日志搜索首页点击查看搜索历史链接，进入搜索历史列表页面：

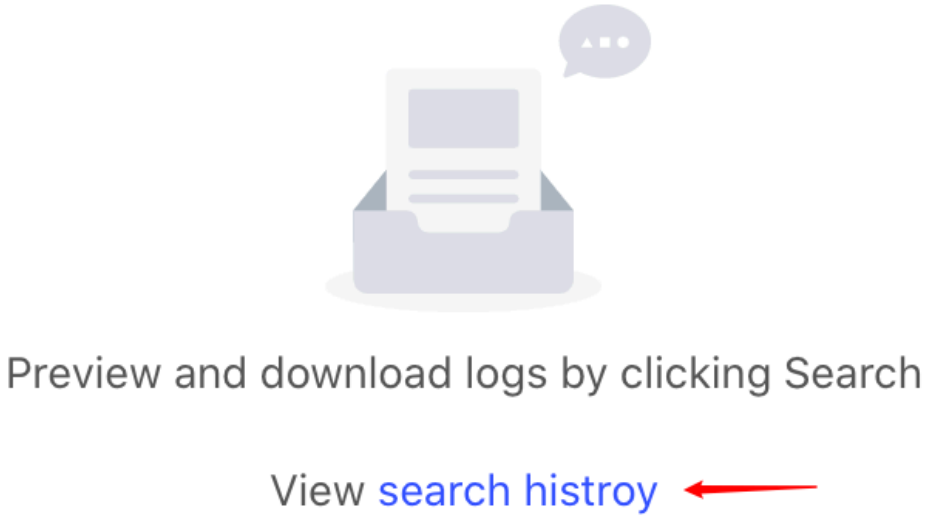


图 163: 搜索历史入口

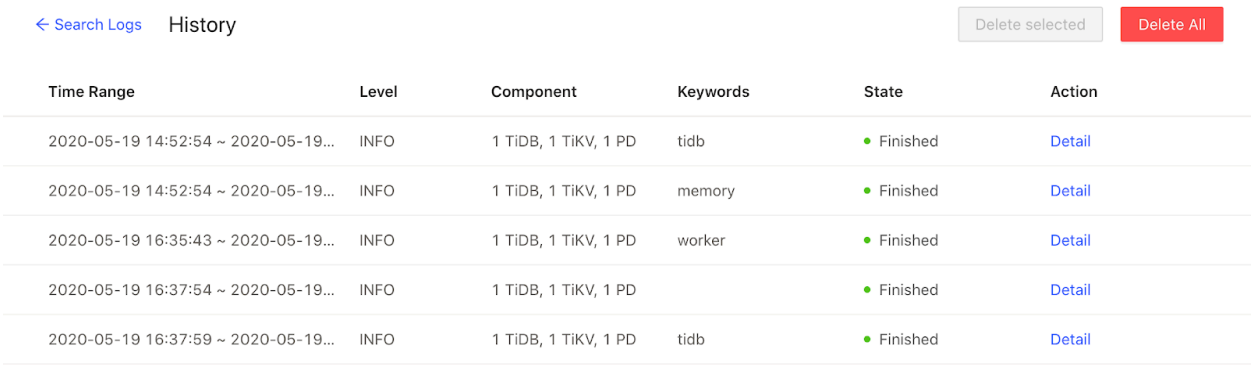


图 164: 搜索历史列表

该列表显示每次搜索的日志的时间范围、日志等级、组件、关键字以及搜索状态等信息。点击操作列的查看详情链接将跳转到此次搜索的结果详情页面。

可以对不再需要的搜索历史执行删除操作，点击右上角的删除全部任务，或者先选中需要删除的某些行后再点击删除选中的任务进行删除：

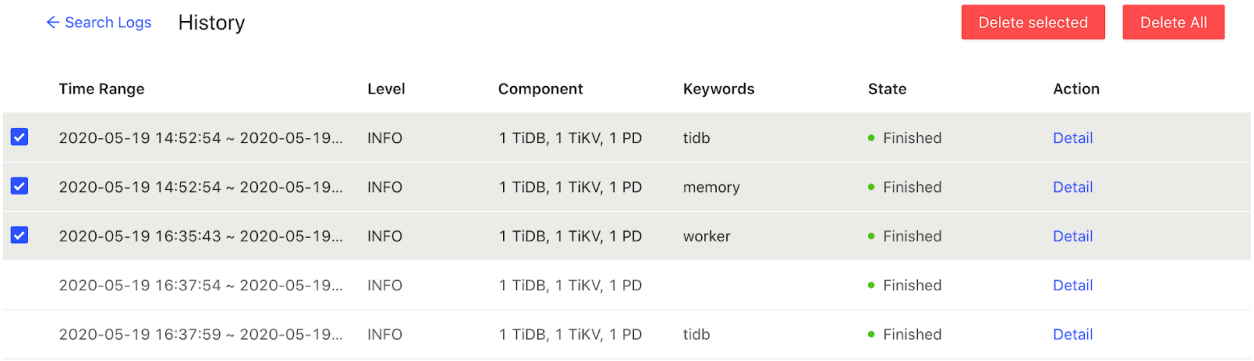


图 165: 删除搜索历史

9.9.13 实例性能分析

9.9.13.1 TiDB Dashboard 实例性能手动分析页面

该功能允许用户收集 TiDB、TiKV、PD、TiFlash 各个实例的性能数据。收集到的性能数据可显示为火焰图或有向无环图，直观展现实例在性能收集的时间段内执行的各种内部操作及比例，方便用户快速了解该实例 CPU 资源消耗细节。

9.9.13.1.1 分析内容

性能分析允许用户收集 TiDB、PD、TiKV、TiFlash 各个实例的性能数据。收集到的性能数据可显示为火焰图、有向无环图等，直观展现实例在性能收集的时间段内执行的各种内部操作及其比例，方便用户快速了解该实例性能消耗情况。目前支持的性能信息：

- TiDB/PD: CPU profile、 Heap、 Mutex、 Goroutine（ debug=2 ）
- TiKV/TiFlash: CPU Profile

9.9.13.1.2 访问页面

可以通过以下两种方法访问实例性能分析页面：

- 登录后，左侧导航条点击高级调试 (Advanced Debugging) > 实例性能分析 (Profile Instances) > 手动分析 (Manually Profile)：

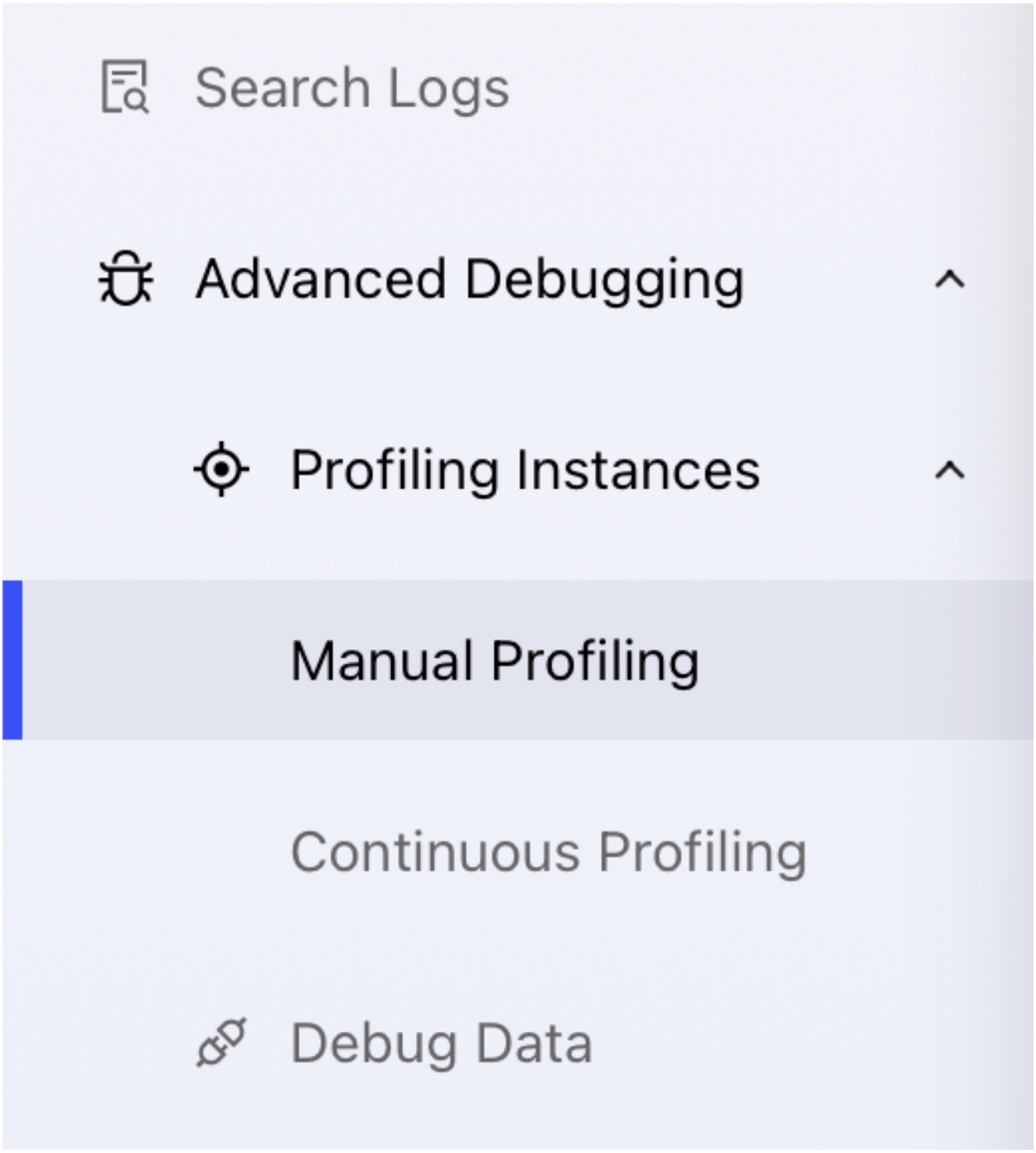


图 166: 访问页面

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/instance_profiling (将 `127.0.0.1:2379` 替换为实际 PD 实例地址和端口)。

9.9.13.1.3 开始性能分析

在实例性能分析页面，选择至少一个目标实例和分析类型，确定性能分析时长（默认为 30 秒）。点击开始分析 (Start Profiling)：

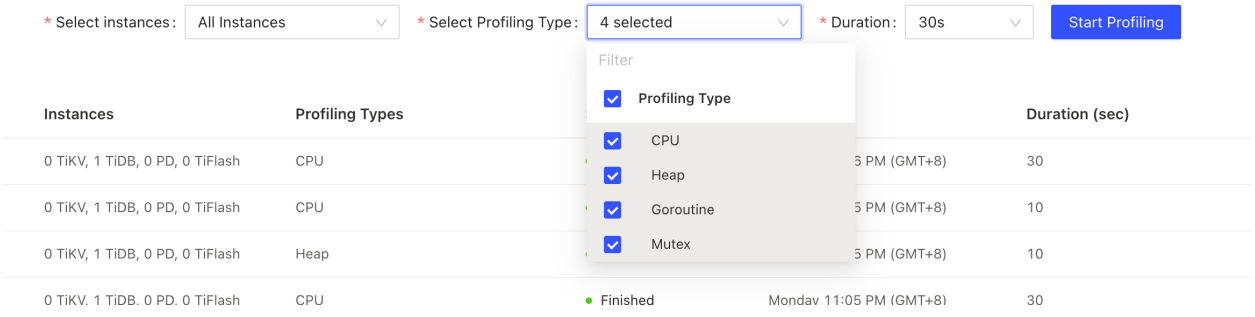


图 167: 开始分析

9.9.13.1.4 查看性能分析状态

开始性能分析后，可以看到实时性能分析状态和进度：

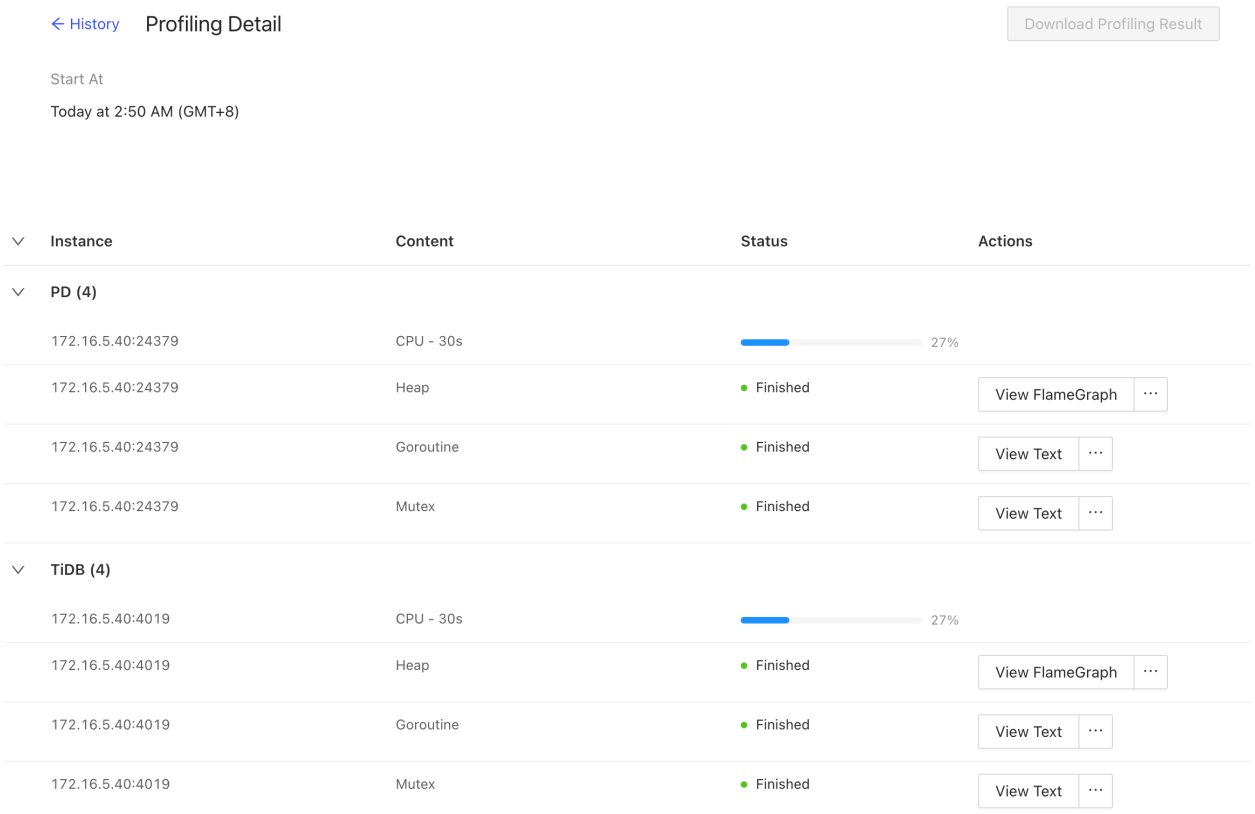


图 168: 实时状态

性能分析会在后台运行，刷新或退出当前页面不会终止已经运行的性能分析任务。

9.9.13.1.5 下载性能分析结果

所有实例的性能分析都完成后，可点击右上角下载按钮 (Download Profiling Result) 打包下载所有性能分析结果：

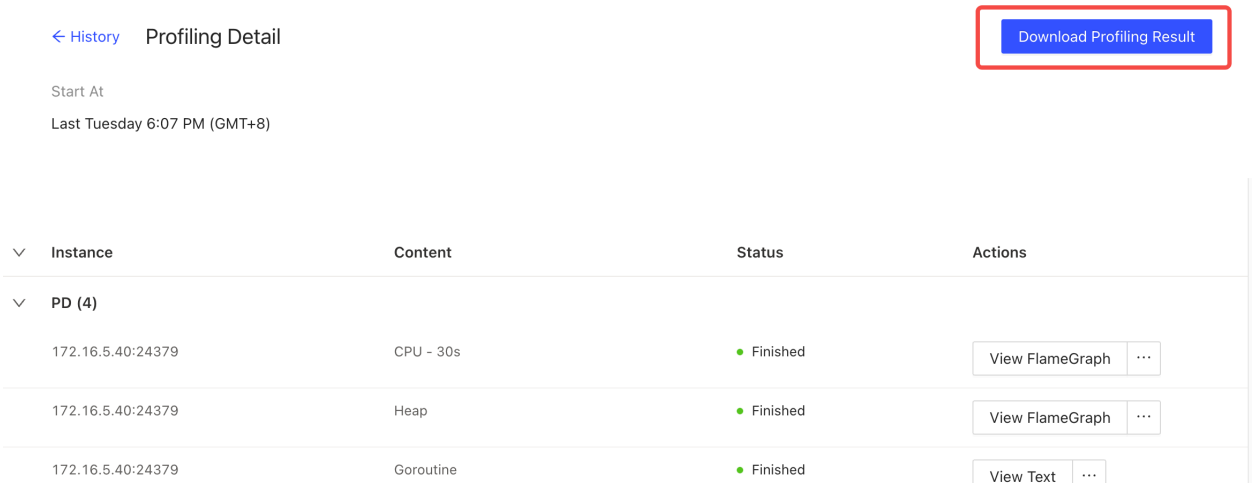


图 169: 下载分析结果

也可以点击列表中的单个实例查看其性能分析结果（包括火焰图、有向无环图、文本），或者悬浮到 ··· 按钮上下载原始数据：

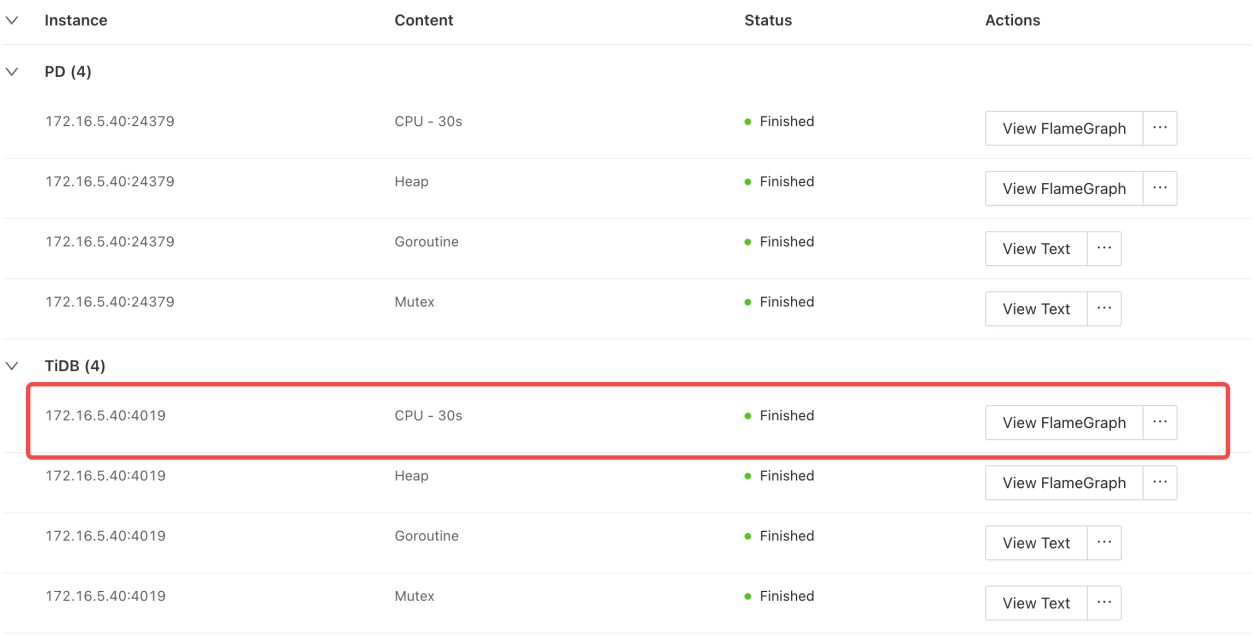


图 170: 在线查看分析结果

9.9.13.1.6 查看性能分析历史

在性能分析页面下方，列出了性能分析历史。点击任意一行，即可查看其状态详情：

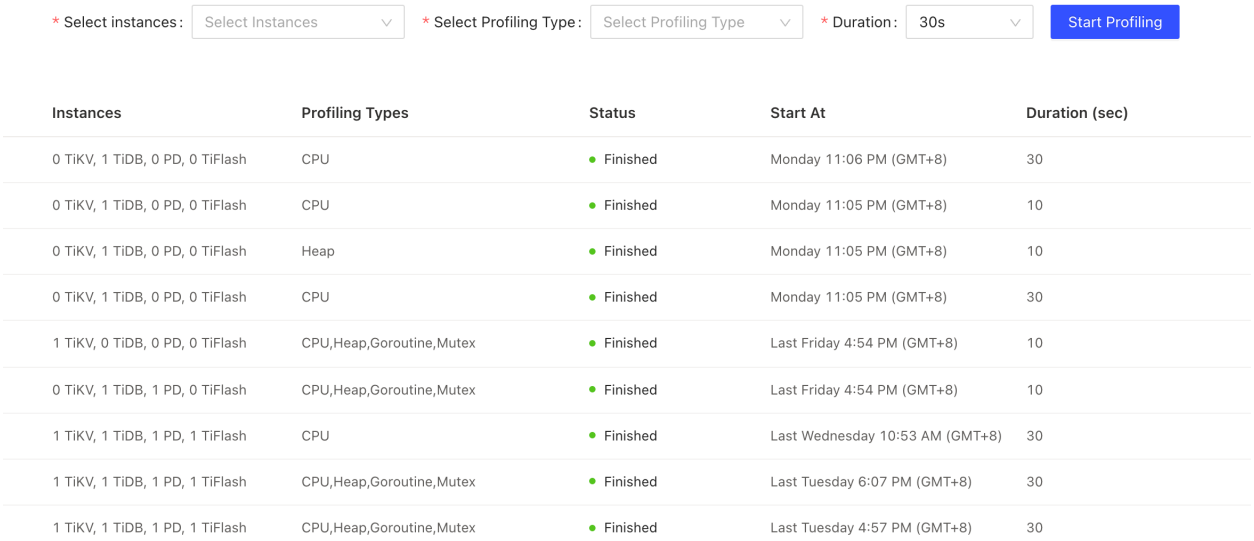


图 171: 历史列表

关于状态详情页的操作，参见查看性能分析状态章节。

9.9.13.2 TiDB Dashboard 实例性能分析 - 持续分析页面

> **警告：**
>
> 持续性能分析目前为实验特性，不建议在生产环境中使用。

持续性能分析是 TiDB v5.3.0 引入的一种从系统调用层面解读资源开销的方法。引入该方法后，TiDB 可提供数据库源码级性能观测，通过火焰图的形式帮助研发、运维人员定位性能问题的根因。

该功能以低于 0.5% 的性能损耗，对数据库内部运行状态持续打快照（类似 CT 扫描），让原本“黑盒”的数据库变成“白盒”，具备更高的可观测性。该功能一键开启后自动运行，存储结果提供了保留时长的设定，过期的结果将会被回收，确保存储空间的有效利用。

9.9.13.2.1 使用限制

使用持续性能分析时，应留意如下使用限制：

- 该功能可在 x86 架构下支持 TiDB、PD、TiKV、TiFlash；而在 ARM 框架下还未完全兼容，不可开启。
- 该功能适用于使用 v1.9.0 及以上版本 TiUP 或 v1.3.0 及以上版本 TiDB Operator 部署和升级的集群，不支持二进制包部署和升级的集群。

9.9.13.2.2 分析内容

持续性能分析允许用户在不重启的情况下持续收集 TiDB、TiKV、PD、TiFlash 各个实例的性能数据，并且持久监控节点。收集到的性能数据可显示为火焰图、有向无环图等，直观展现实例在性能收集的时间段内执行的各种内部操作及其比例，方便用户快速了解该实例 CPU 资源消耗细节。目前支持的性能信息：

- TiDB/PD: CPU profile、Heap、Mutex、Goroutine（debug=2）
- TiKV/TiFlash: CPU Profile

9.9.13.2.3 启用持续性能分析

本节介绍如何在使用 TiUP 或 TiDB Operator 部署的 TiDB 集群上启用持续性能分析。

使用 TiUP 部署的集群

对于使用 TiUP 部署的集群，你可以通过以下步骤启用持续性能分析。

1. 进入 TiDB Dashboard，选择高级调试 (Advanced Debugging) > 实例性能分析 (Profiling Instances) > 持续分析 (Continuous Profiling)。
2. 点击打开设置 (Open Settings)。在右侧设置 (Settings) 页面，将启用特性 (Enable Feature) 下方的开关打开。设置保留时间 (Retention Period) 或保留默认值。
3. 点击保存 (Save)。

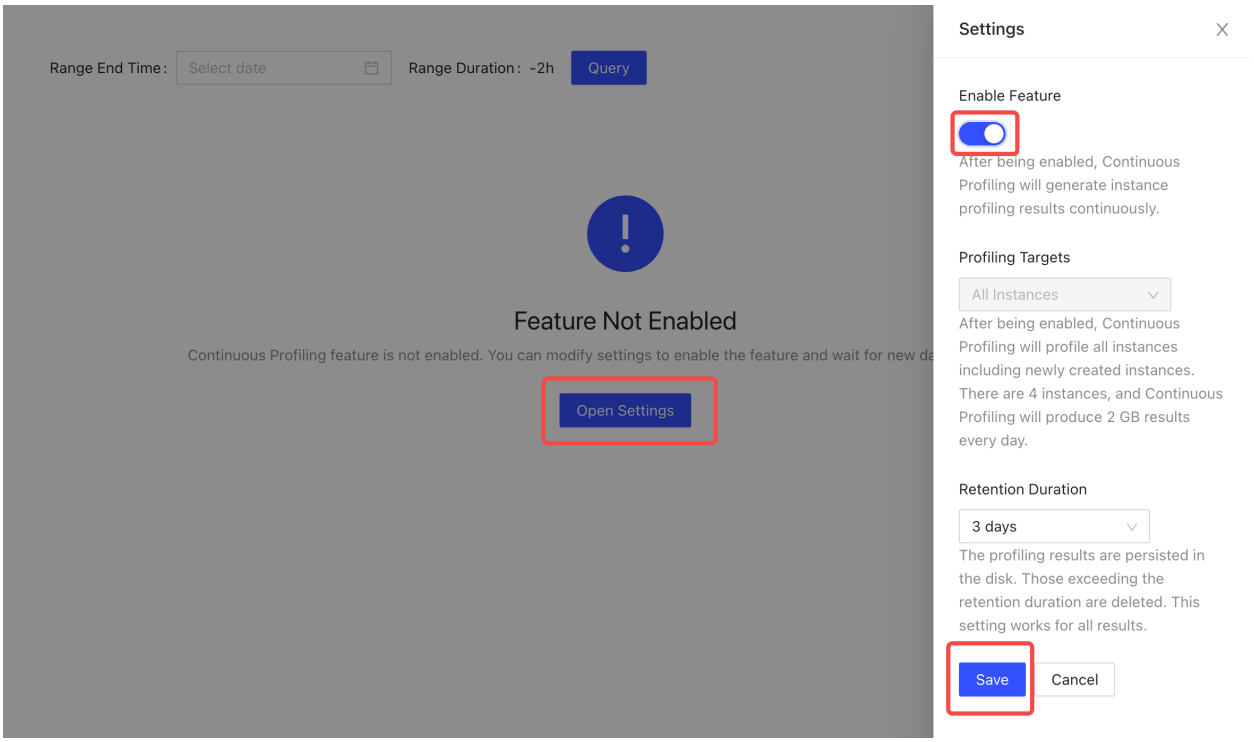


图 172: 启用功能

使用 TiDB Operator 部署的集群

详见 TiDB Operator 的 启用持续性能分析

9.9.13.2.4 访问页面

启用持续性能分析功能后，你可以通过以下任一方式访问实例性能分析页面：

- 登录后，左侧导航条点击高级调试 (Advanced Debugging) > 实例性能分析 (Profiling Instances) > 持续分析 (Continuous Profiling)：

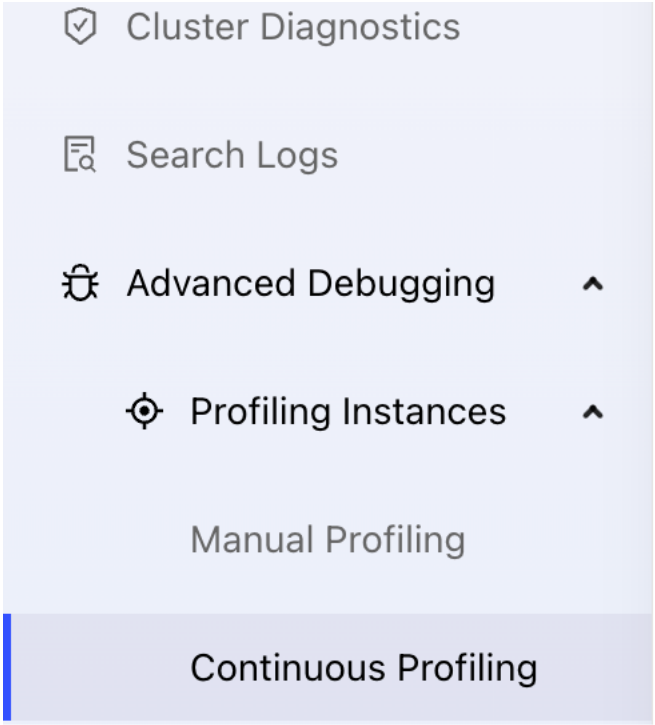


图 173: 访问页面

- 在浏览器中访问 http://127.0.0.1:2379/dashboard/#/continuous_profiling（将 `127.0.0.1:2379` 替换为实际 PD 实例地址和端口）。

9.9.13.2.5 查看性能分析历史结果

开始持续性能分析后，可以在列表中看到已经完成的性能分析结果：

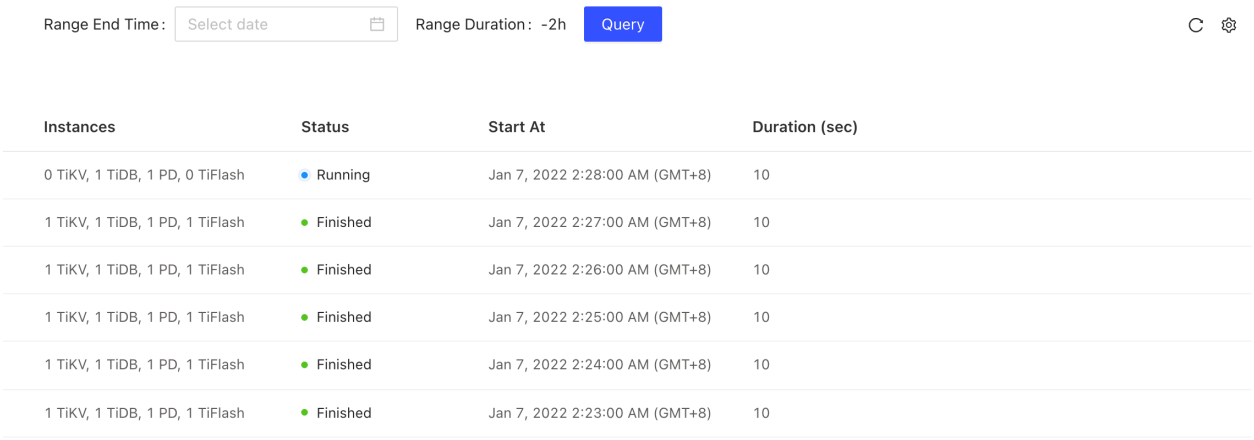


图 174: 历史结果

性能分析在后台运行，刷新或退出当前页面不会终止正在运行的性能分析任务。

9.9.13.2.6 下载性能分析结果

进入某次分析结果后，可点击右上角下载按钮 (Download Profiling Result) 打包下载所有性能分析结果：

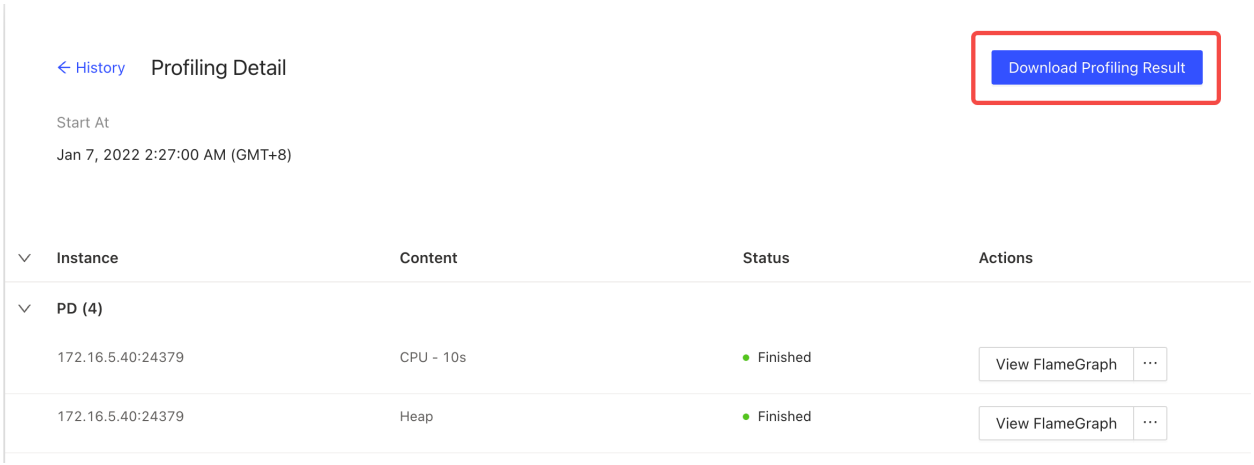


图 175: 下载某次分析结果

也可以点击列表中的单个实例查看其性能分析结果（包括火焰图、有向无环图、文本），或者悬浮在 ··· 按钮上下载原始数据：

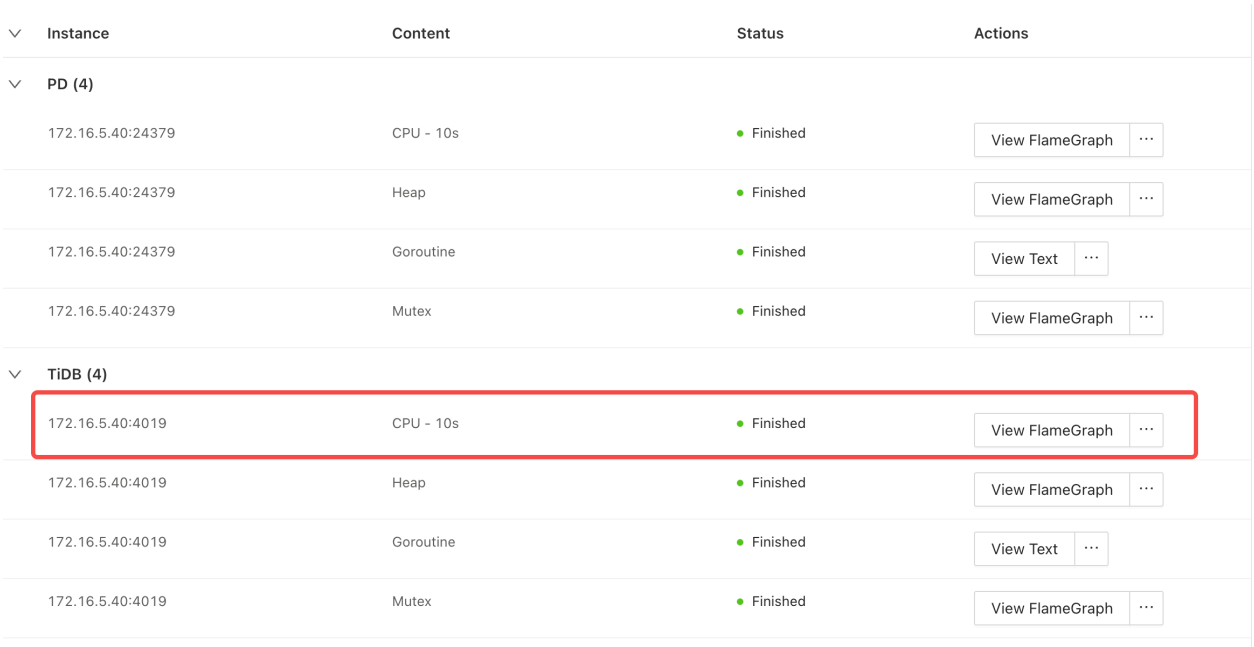


图 176: 查看单个实例分析结果

9.9.13.2.7 停用持续性能分析

1. 进入 TiDB Dashboard，选择高级调试 (Advanced Debugging) > 实例性能分析 (Profiling Instances) > 持续分析 (Continuous Profiling)。
2. 点击设置 (Settings)，将启用特性 (Enable Feature) 下方的开关关闭。
3. 在弹出的停用持续分析 (Disable Continuous Profiling Feature) 对话框中，选择停用 (Disable)。
4. 点击保存 (Save)。
5. 点击弹窗的确认（Disable）。

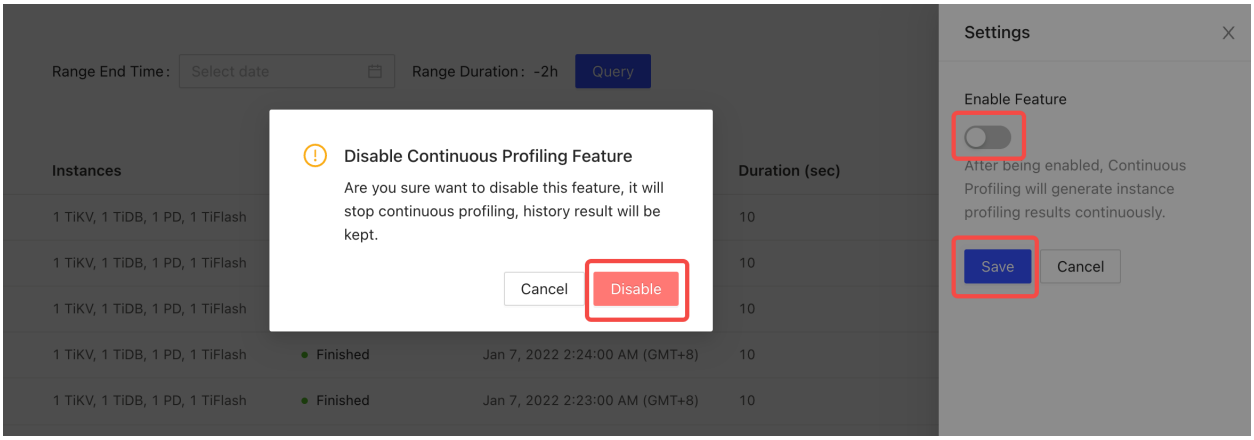


图 177: 停用功能

9.9.14 会话管理与配置

9.9.14.1 分享 TiDB Dashboard 会话

你可以将当前的 TiDB Dashboard 会话安全地分享给其他用户访问，这样其他用户无需要知道登录账号密码即可访问 TiDB Dashboard 并进行操作。

9.9.14.1.1 分享者操作步骤

1. 登录 TiDB Dashboard。
2. 点击边栏左下角的用户名访问配置界面。
3. 点击分享当前会话 (Share Current Session)。

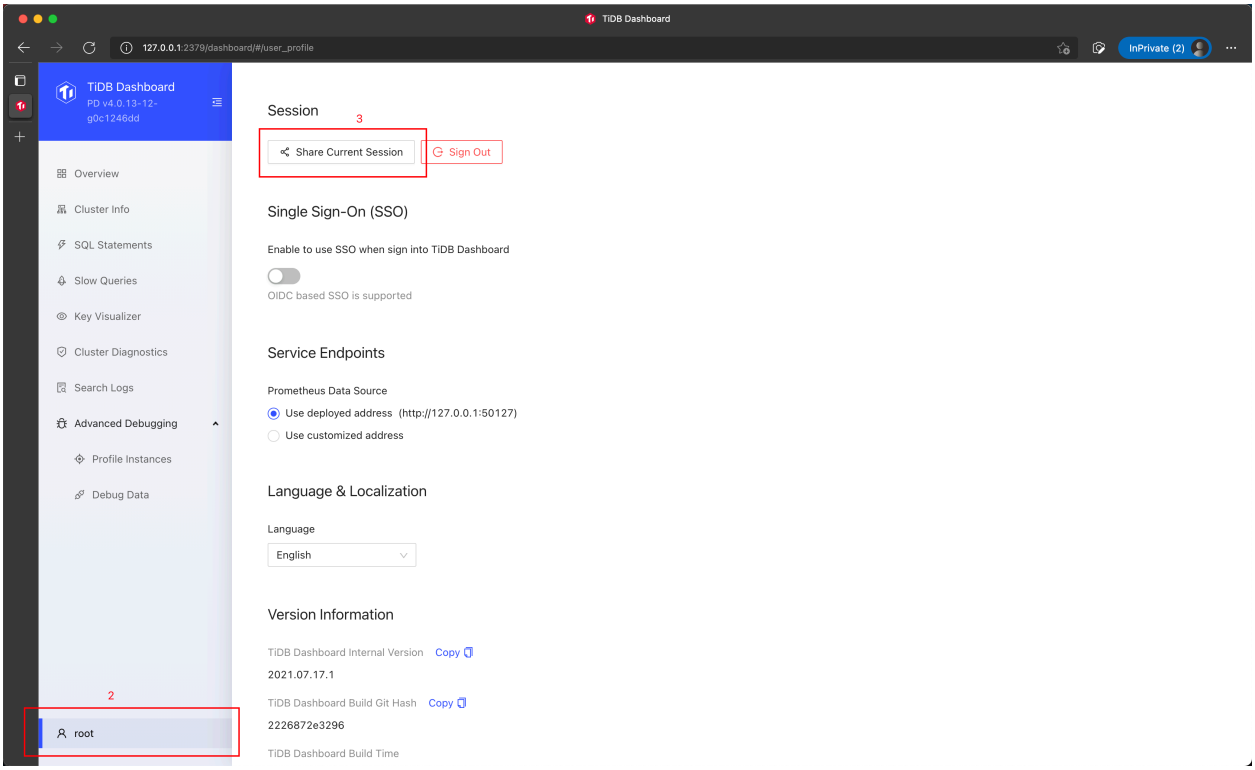


图 178: 操作示例

> **注意：**
>
> 出于安全考虑，已分享的会话中不能使用分享功能将该会话再次分享给其他人。

4. 在弹出的对话框中，对分享进行细节配置：

- 有效时间：分享的会话在多少时间内有效。登出当前会话不影响已分享会话的有效时间。
- 以只读权限分享：分享的会话为只读，例如不允许进行配置修改等操作。

5. 点击生成授权码 (Generate Authorization Code)。

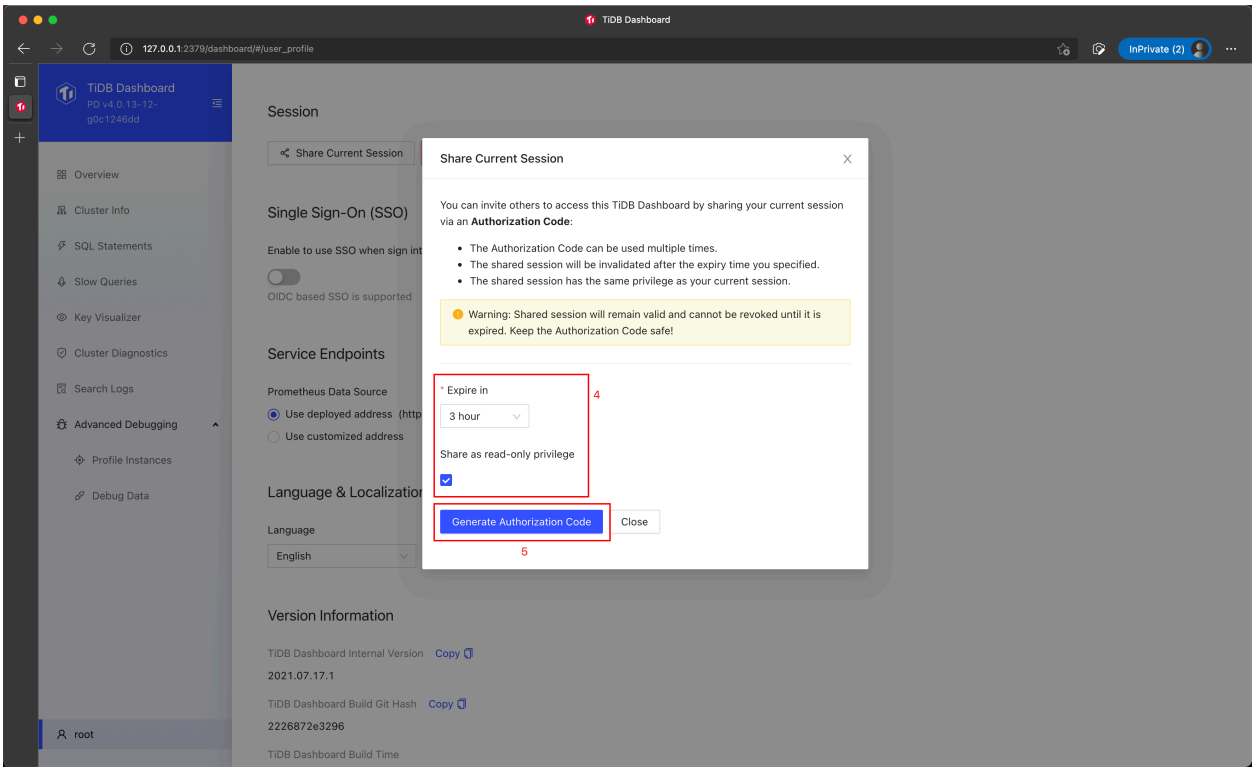


图 179: 操作示例

6. 将生成出来的授权码提供给要分享的用户。

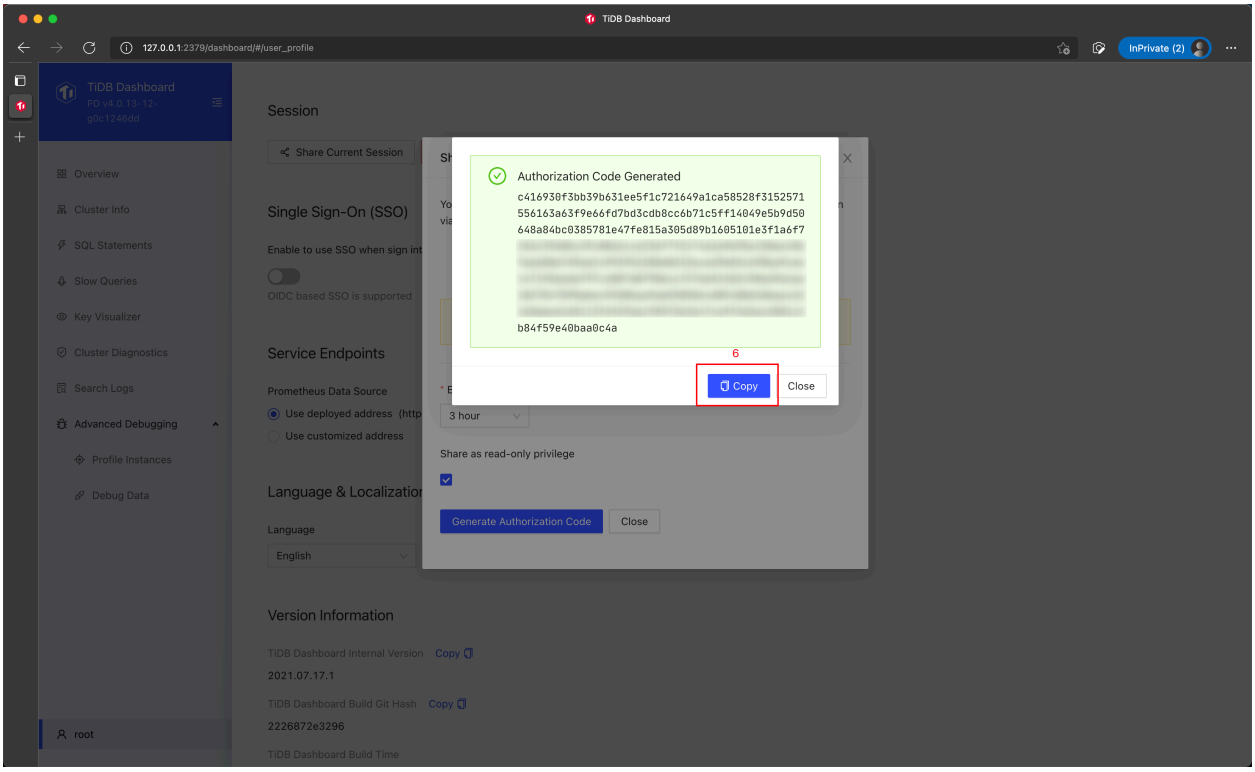


图 180: 操作示例

警告：

请妥善地保管授权码。不要将授权码分发给不受信任的用户，否则他们也将具备访问和操作 TiDB Dashboard 的能力。

9.9.14.1.2 受邀请者操作步骤

1. 在 TiDB Dashboard 登录界面上，点击使用其他登录方式 (Use Alternative Authentication)。

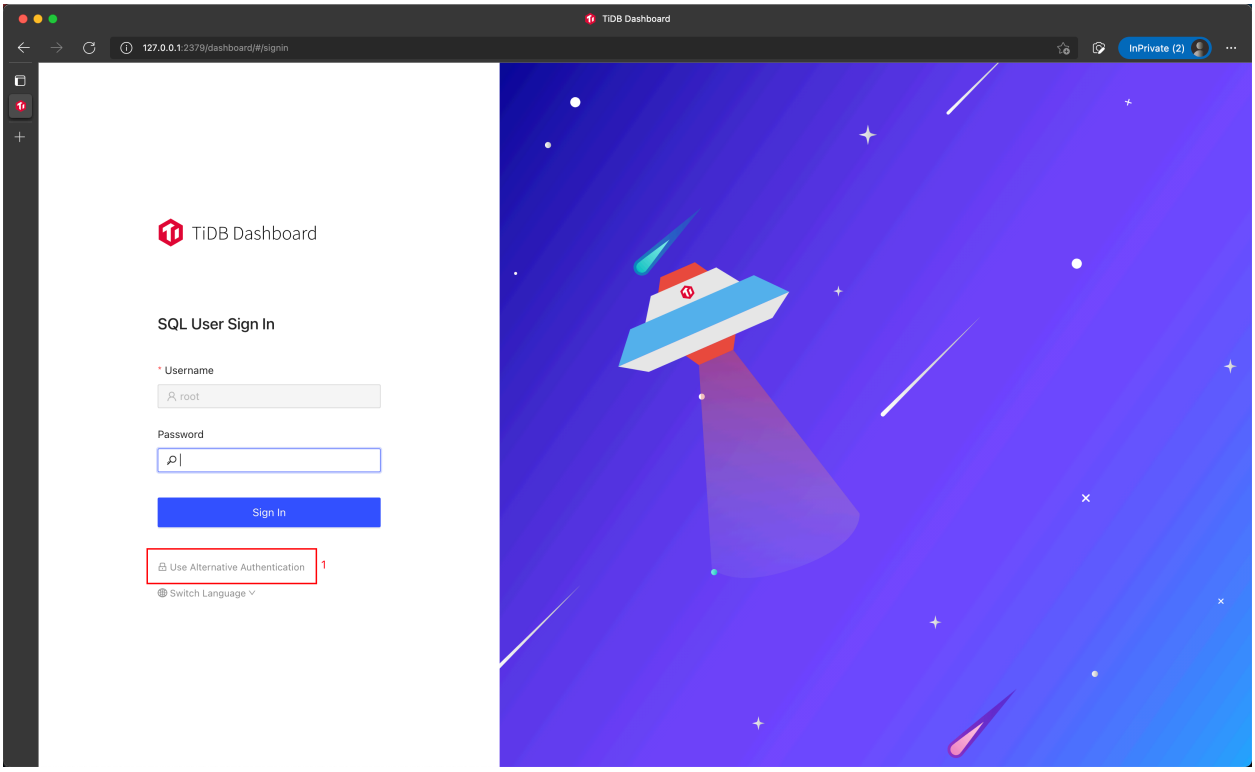


图 181: 操作示例

2. 选择使用授权码 (Authorization Code) 登录。

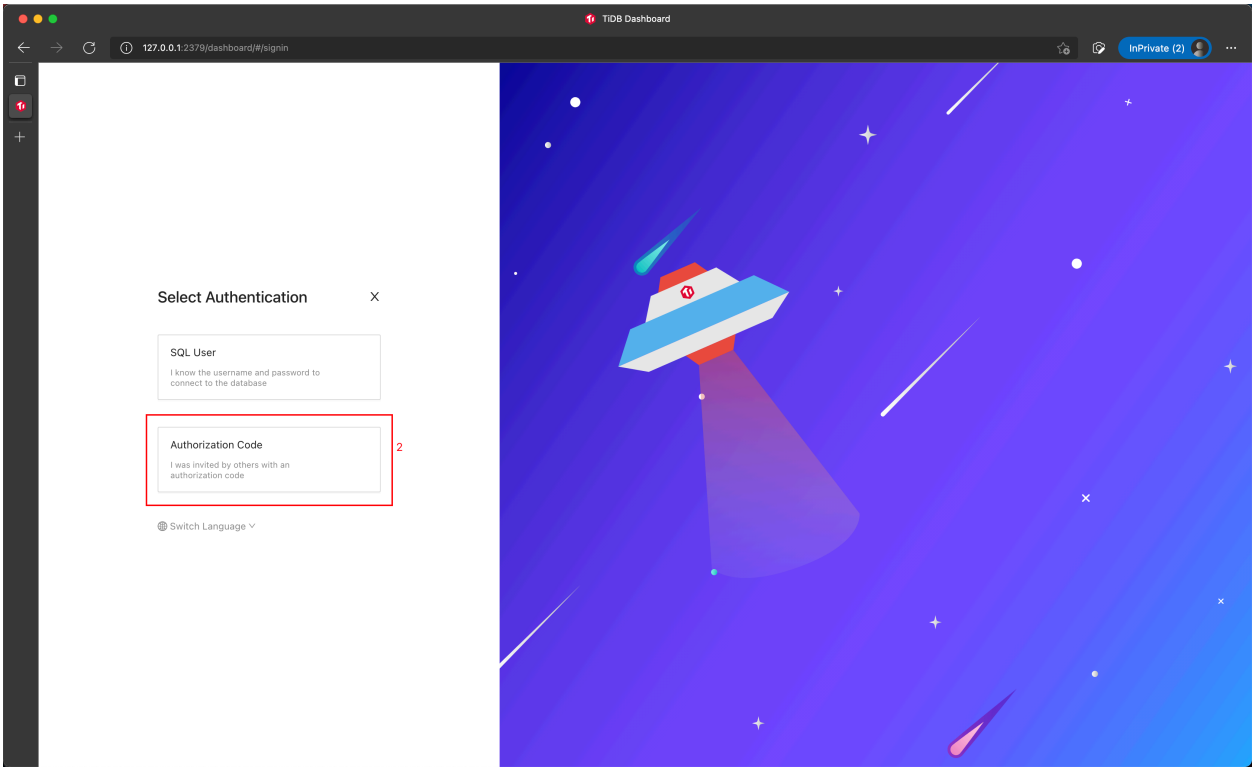


图 182: 操作示例

3. 输入从分享者取得的授权码。
4. 点击登录 (Sign In)。

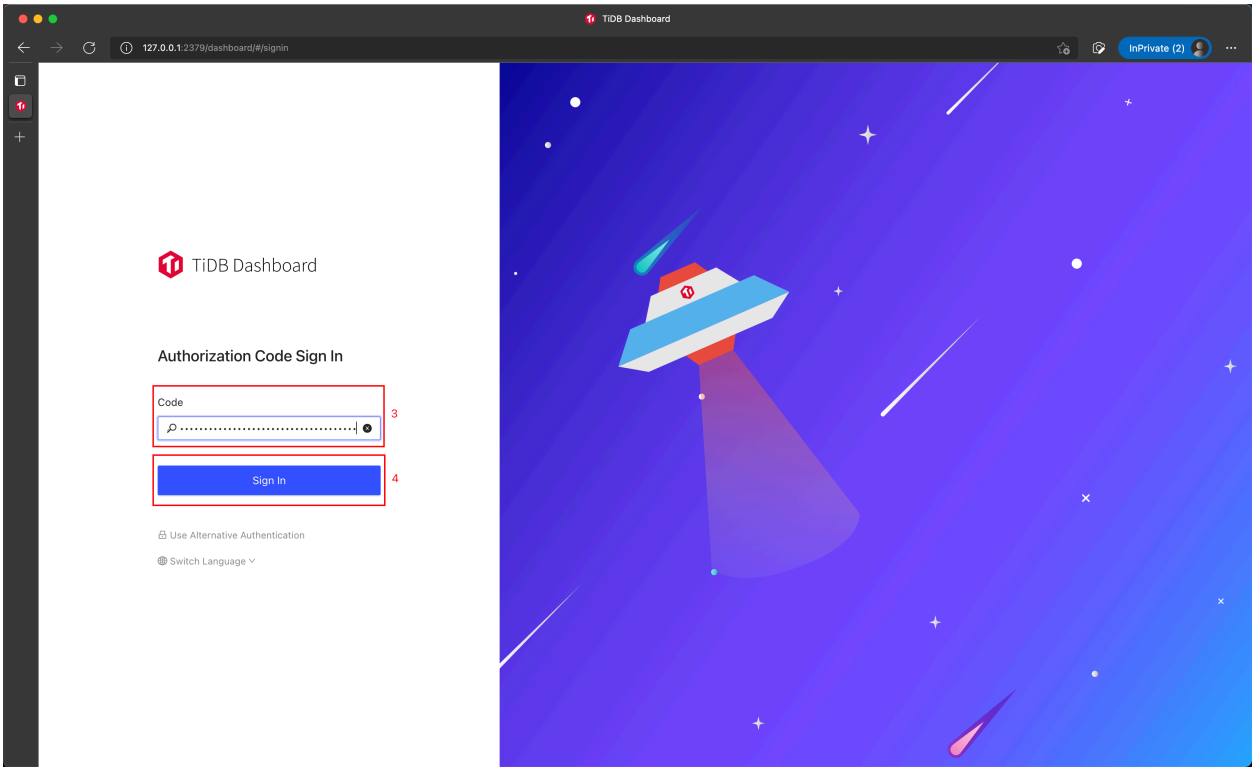


图 183: 操作示例

9.9.14.2 配置 TiDB Dashboard 使用 SSO 登录

TiDB Dashboard 支持基于 OIDC 协议的单点登录 (Single Sign-On)。配置 TiDB Dashboard 启用 SSO 登录后，你可以通过配置的 SSO 服务进行登录鉴权，无需输入 SQL 用户名和密码即可登录到 TiDB Dashboard。

9.9.14.2.1 配置 OIDC SSO

启用 SSO

1. 登录 TiDB Dashboard。
2. 点击边栏左下角的用户名访问配置界面。
3. 在单点登录 (Single Sign-On) 区域下，开启允许使用 SSO 登录到 TiDB Dashboard (Enable to use SSO when sign into TiDB Dashboard)。
4. 在表单中填写 OIDC Client ID 和 OIDC Discovery URL 字段。

   一般可以从 SSO 服务的提供商处获取到这两个字段信息：

   - OIDC Client ID 有时也被称为 OIDC Token Issuer
   - OIDC Discovery URL 有时也被称为 OIDC Token Audience。
5. 将 SQL 登录密码录入到 TiDB Dashboard 中，以便在 SSO 鉴权通过后完成登录。点击授权登录为该用户 (Authorize Impersonation) 录入密码。

这是因为 TiDB Dashboard SSO 的原理是在 SSO 成功鉴权后，采用 TiDB Dashboard 内加密存储的 SQL 登录密码进行替代登录。

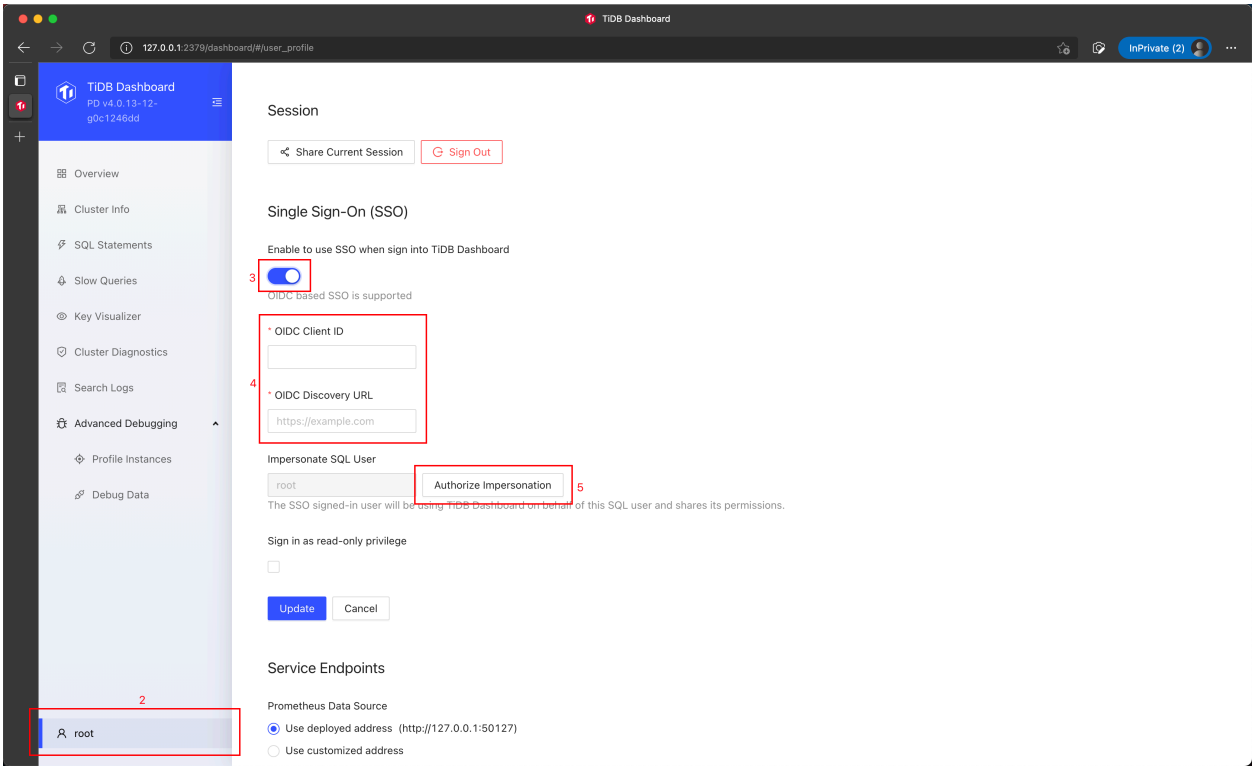


图 184: 操作示例

> **注意：**
>
> 你录入的密码将被加密存储。若 SQL 用户密码后续发生了变更，将导致 SSO 登录失败。这时可以重新录入密码使 SSO 登录恢复正常。

6. 在对话框中填写完密码后，点击授权并保存 (Authorize and Save)。

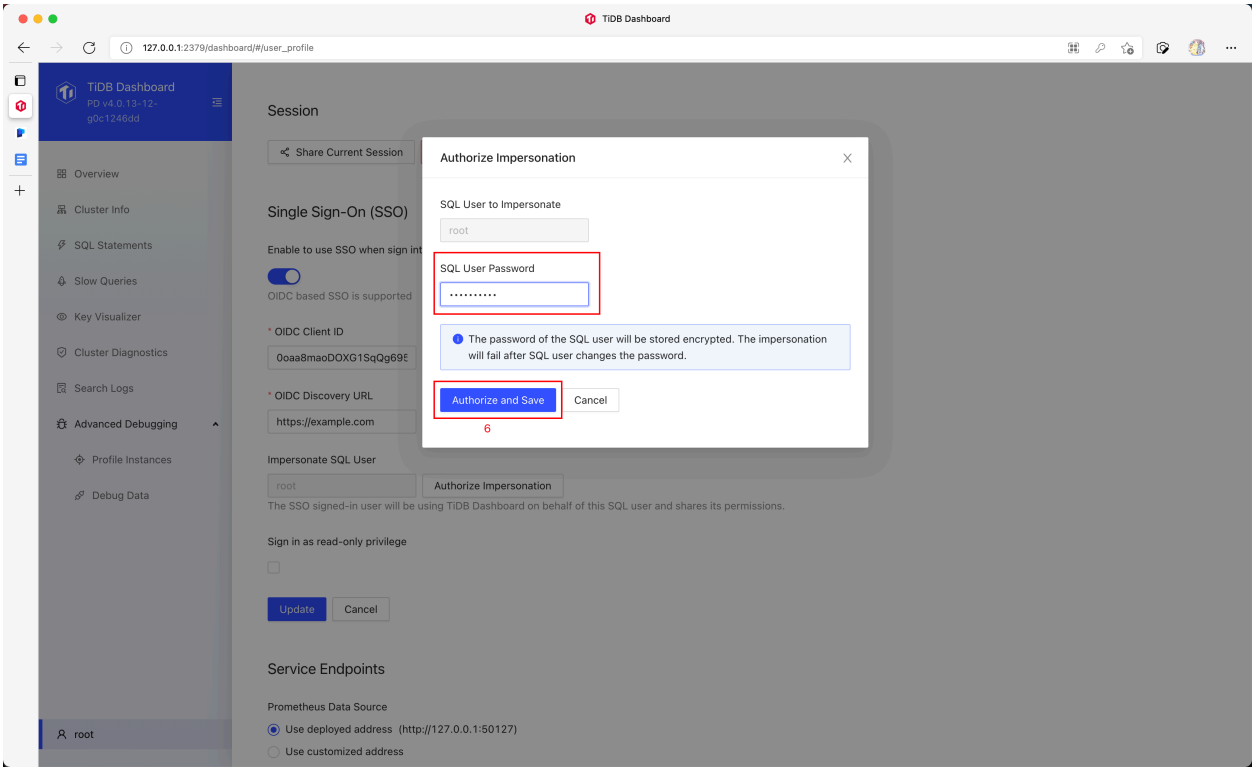


图 185: 操作示例

7. 点击更新 (Update) 保存配置。

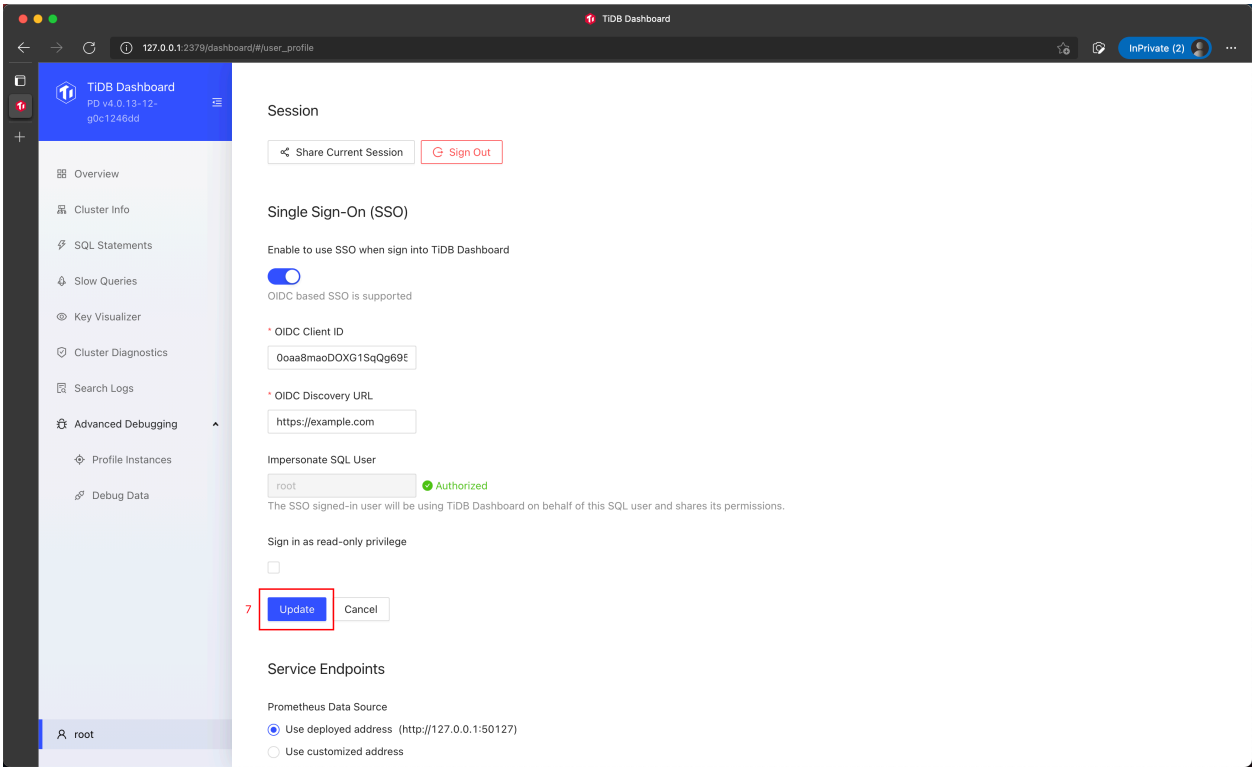


图 186: 操作示例

至此 TiDB Dashboard 中已经成功开启了 SSO 登录。

> **注意：**
>
> 出于安全原因，部分 SSO 服务还需要你进一步在 SSO 服务中配置受信任的登录和登出跳转地址，请参见 SSO 服务的具体帮助完成配置。

禁用 SSO

你可以随时禁用 SSO。禁用后，之前已录入并存储在本地的替代登录 SQL 密码将被彻底清除。禁用步骤如下：

1. 登录 TiDB Dashboard。
2. 点击边栏左下角用户名访问配置界面。
3. 在单点登录 (Single Sign-On) 区域下，关闭允许使用 SSO 登录到 TiDB Dashboard (Enable to use SSO when sign into TiDB Dashboard)。
4. 点击更新 (Update) 保存配置。

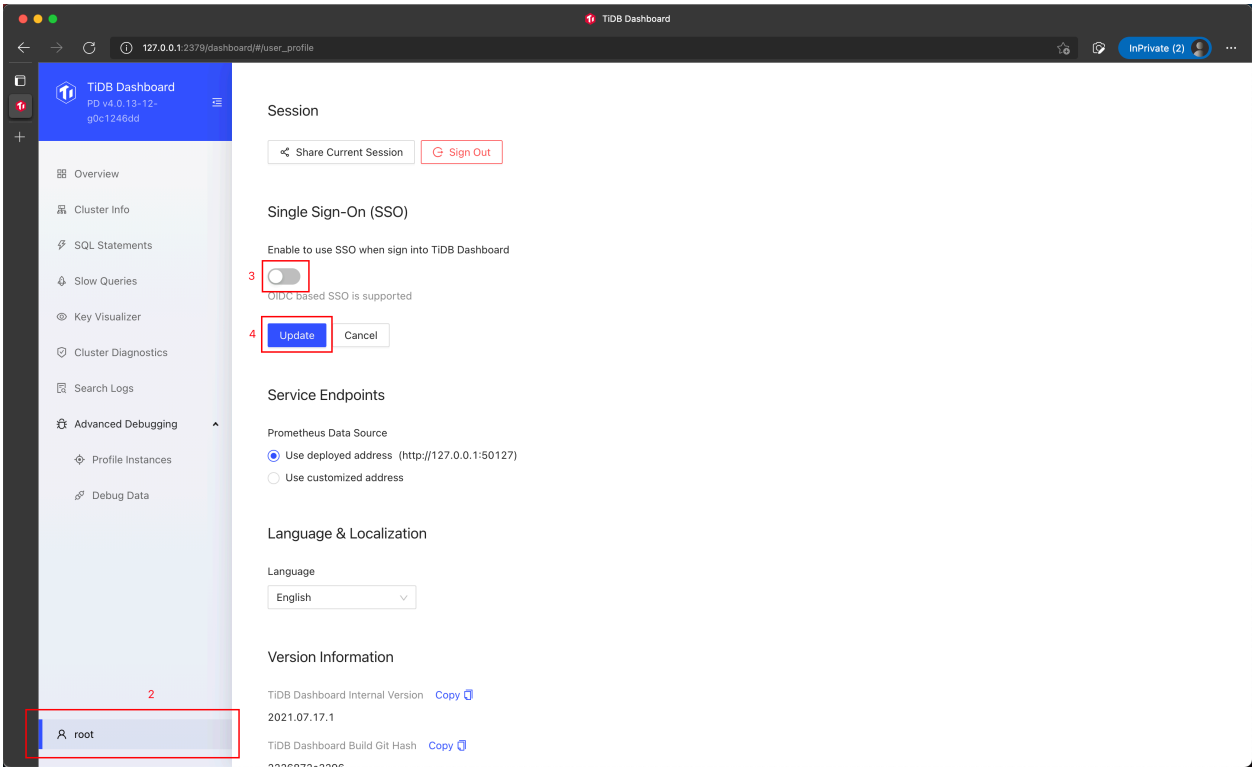


图 187: 操作示例

密码发生变更后重新录入密码

若替代登录的 SQL 用户密码发生了变更，则 SSO 登录将会失败。此时，你可以将新的登录密码录入到 TiDB Dashboard 中以恢复正常 SSO 登录功能，步骤如下：

1. 登录 TiDB Dashboard。
2. 点击边栏左下角用户名访问配置界面。
3. 在单点登录 (Single Sign-On) 区域下，点击授权登录为该用户 (Authorize Impersonation) 来录入新的密码。

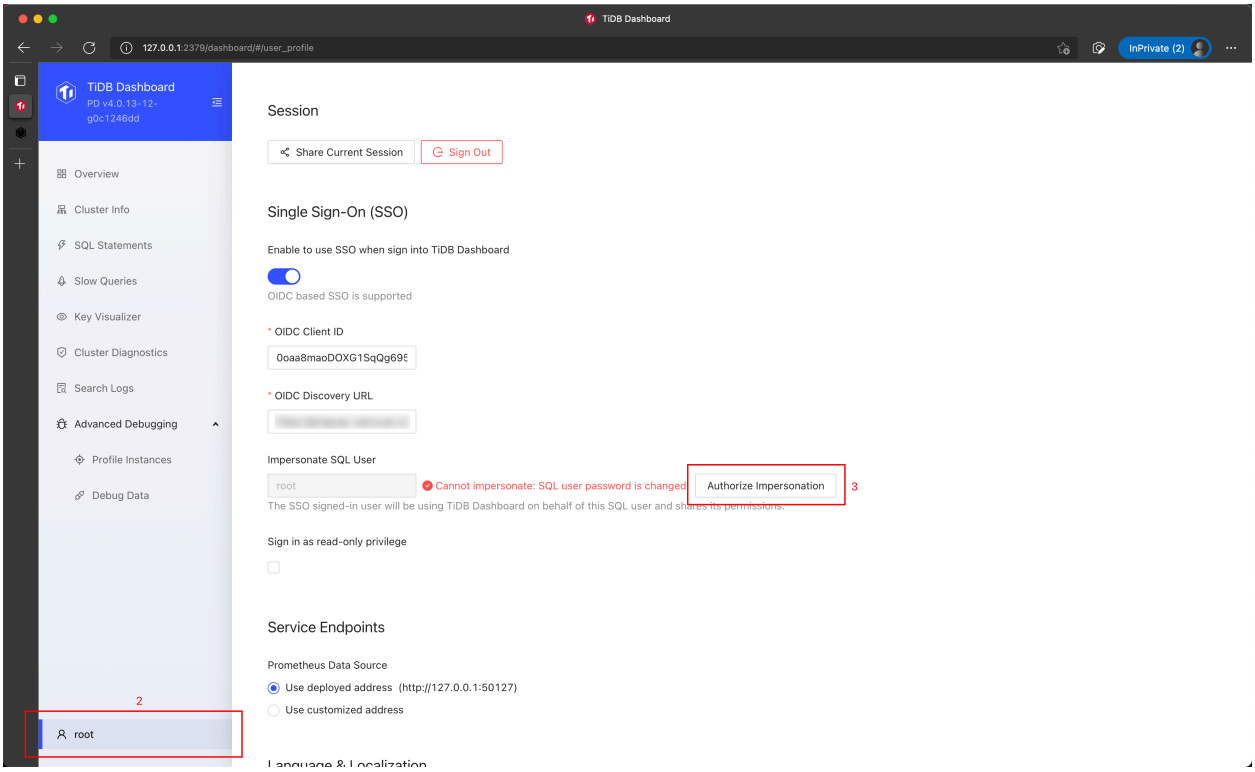


图 188: 操作示例

4. 在对话框中填写完毕密码后，点击授权并保存 (Authorize and Save)。

9.9.14.2.2 使用 SSO 登录

若 TiDB Dashboard 已经完成了 SSO 的配置，你可使用以下步骤完成登录：

1. 在 TiDB Dashboard 登录界面上，点击使用公司账号 SSO 登录 (Sign in via Company Account)。

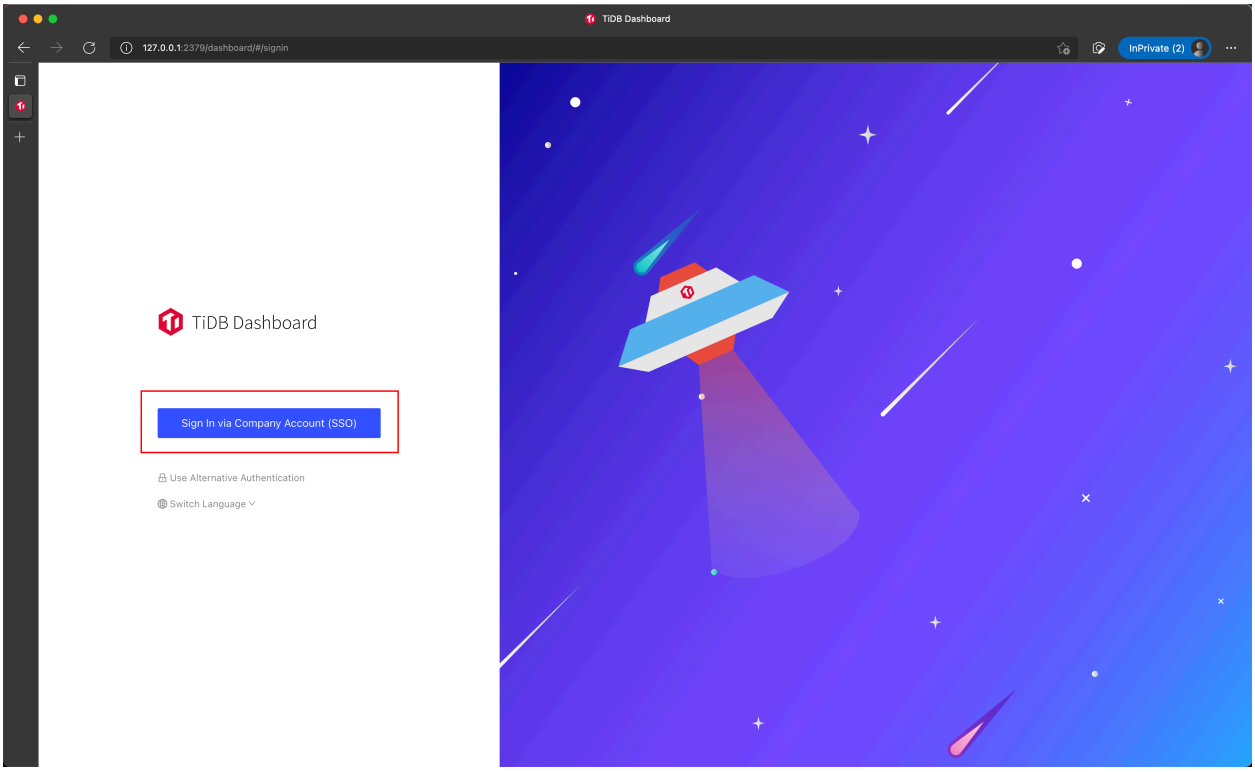


图 189: 操作示例

2. 在配置 SSO 的系统中进行登录。
3. 你将被重定向回 TiDB Dashboard 完成登录。

9.9.14.2.3 示例一：使用 Okta 进行 TiDB Dashboard SSO 登录认证

Okta 是一个提供 OIDC SSO 的身份认证服务。以下步骤展示了如何配置 Okta 及 TiDB Dashboard，使得 TiDB Dashboard 可以通过 Okta 进行 SSO 登录。

步骤一：配置 Okta

首先需要在 Okta 中创建一个用于集成 SSO 的 Application Integration。

1. 访问 Okta 管理后台。
2. 点击左侧边栏的 Applications > Applications。
3. 点击 Create App Integration。

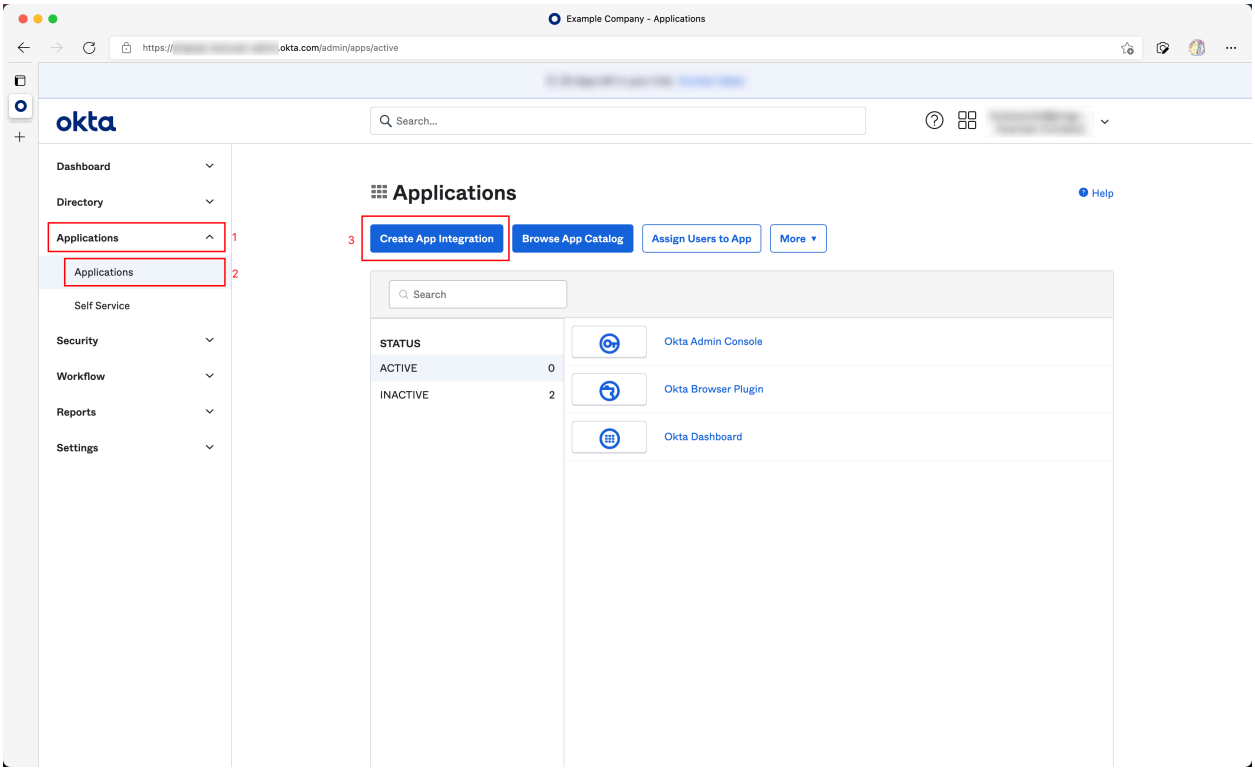


图 190: 操作示例

4. 在弹出的对话框中，Sign-in method 字段选择 OIDC - OpenID Connect。
5. Application Type 字段选择 Single-Page Application。
6. 对话框中点击 Next 按钮。

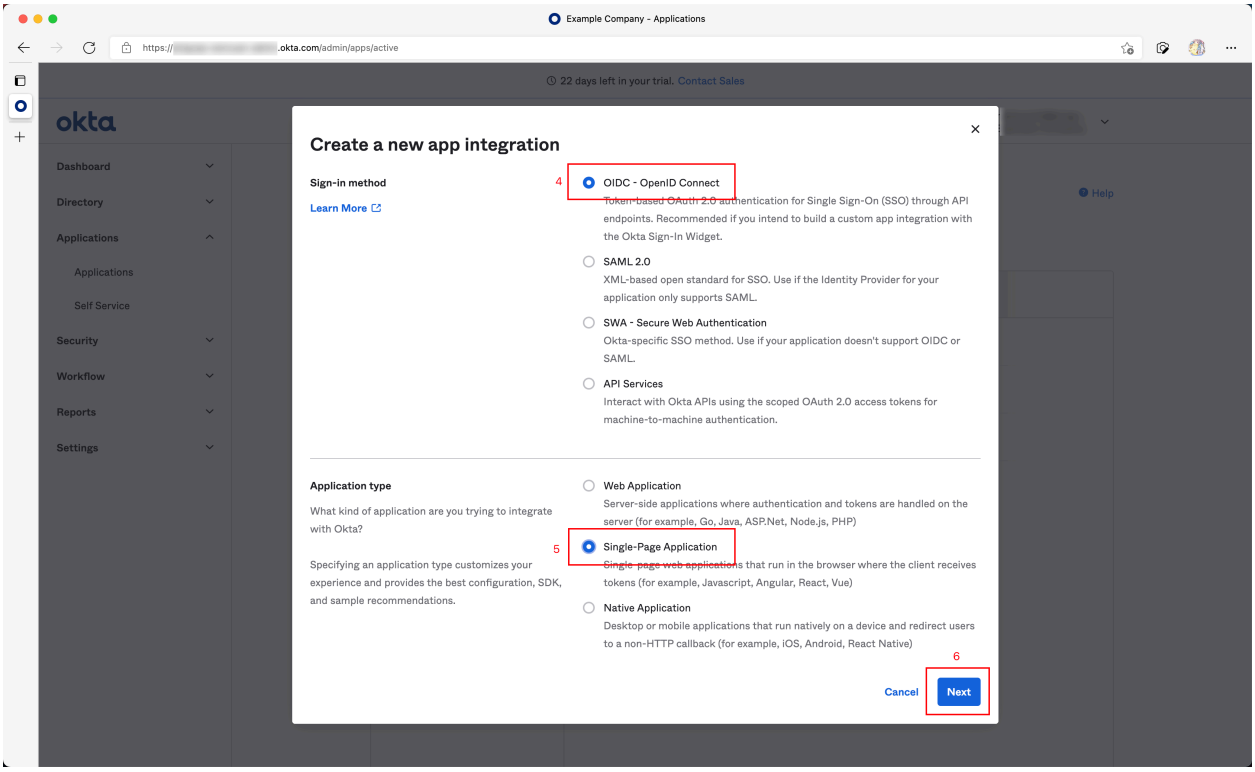


图 191: 操作示例

7. Sign-in redirect URIs 字段填写如下内容：

```
http://DASHBOARD_IP:PORT/dashboard/?sso_callback=1
```

以上内容中，将 DASHBOARD_IP:PORT 替换为你在浏览器中实际访问 TiDB Dashboard 所使用的域名（或 IP）及端口。

8. Sign-out redirect URIs 字段填写如下内容：

```
http://DASHBOARD_IP:PORT/dashboard/
```

类似地，将 DASHBOARD_IP:PORT 替换为实际的域名（或 IP）及端口。

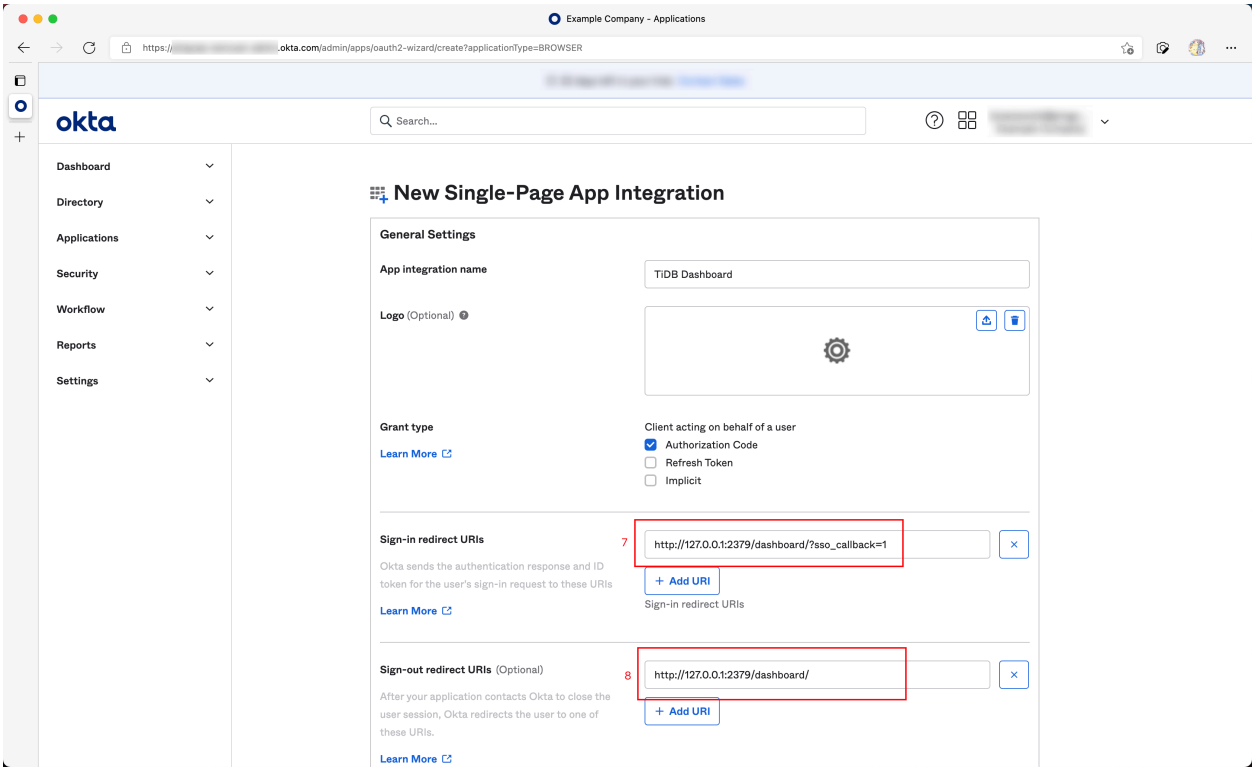


图 192: 操作示例

9. 在 Assignments 中按你的实际需求配置组织中哪些用户可以通过这个 SSO 登录 TiDB Dashboard，然后点击 Save 保存配置。

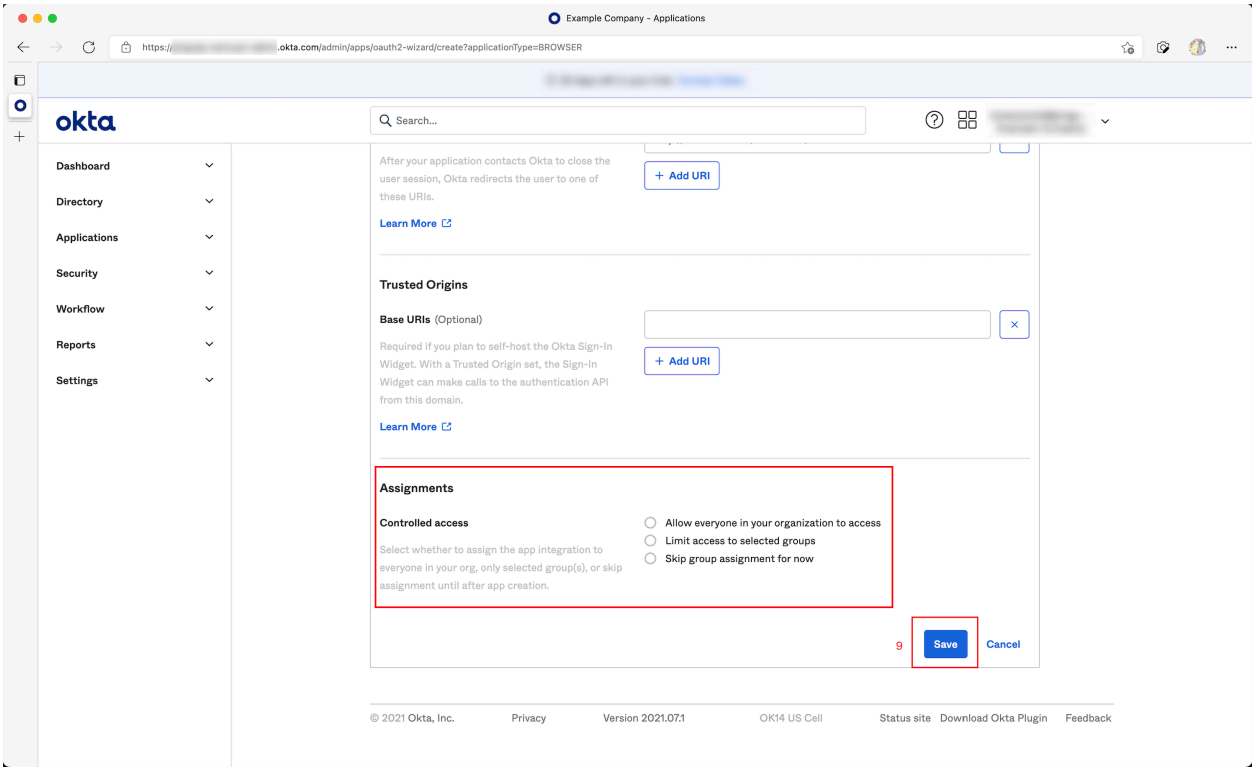


图 193: 操作示例

步骤二：获取 TiDB Dashboard 所需的配置参数并填入 TiDB Dashboard

1. 在 Okta 创建的 App Integration 中，点击 Sign On。

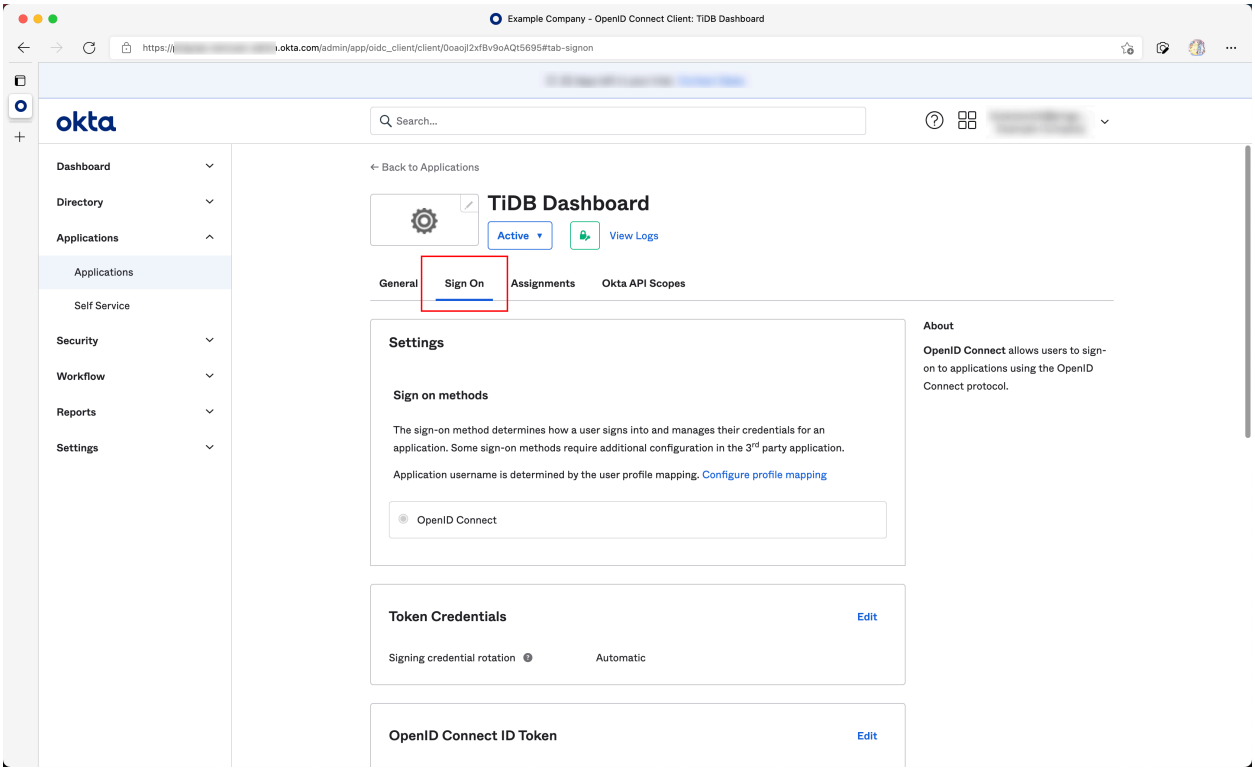


图 194: 操作示例 1

2. OpenID Connect ID Token 区域中有 Issuer 和 Audience 字段，复制这两个字段的值。

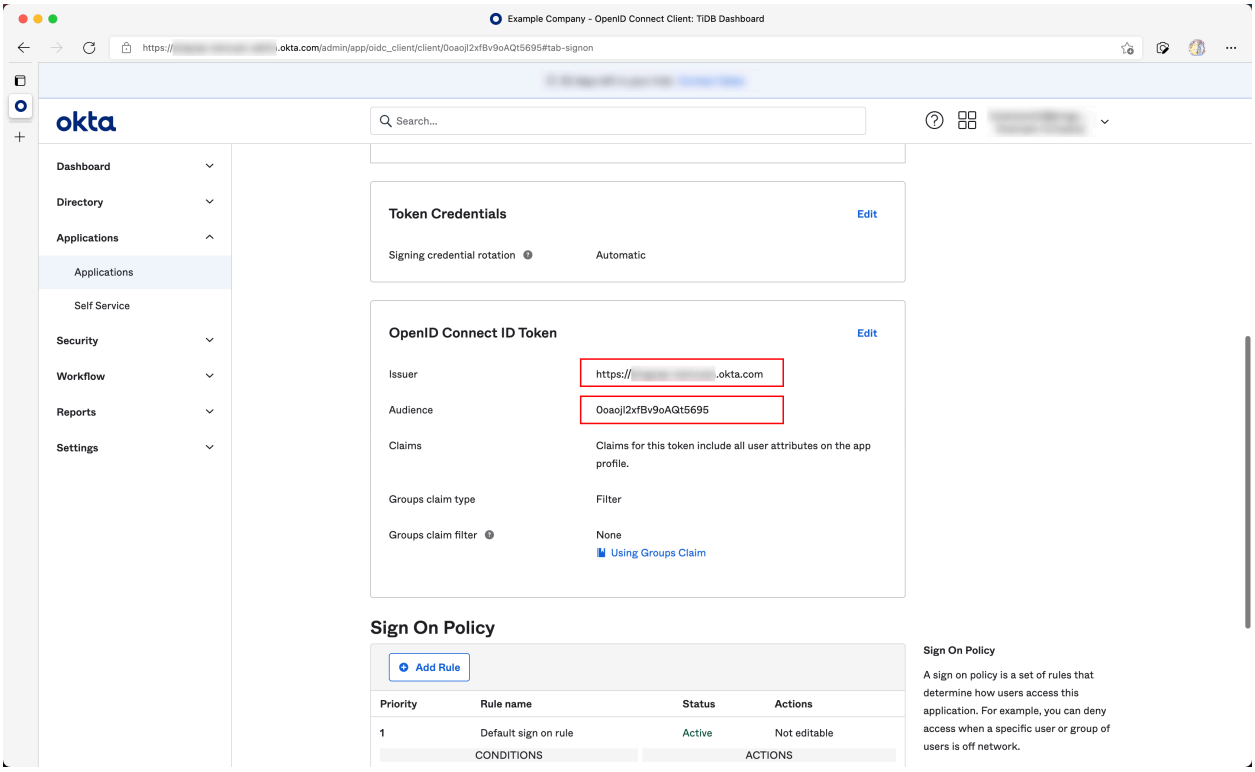


图 195: 操作示例 2

3. 打开 TiDB Dashboard 配置界面，将上一步获取到的 Issuer 填入 OIDC Client ID，将 Audience 填入 OIDC Discovery URL 后，完成授权并保存配置。示例如下：

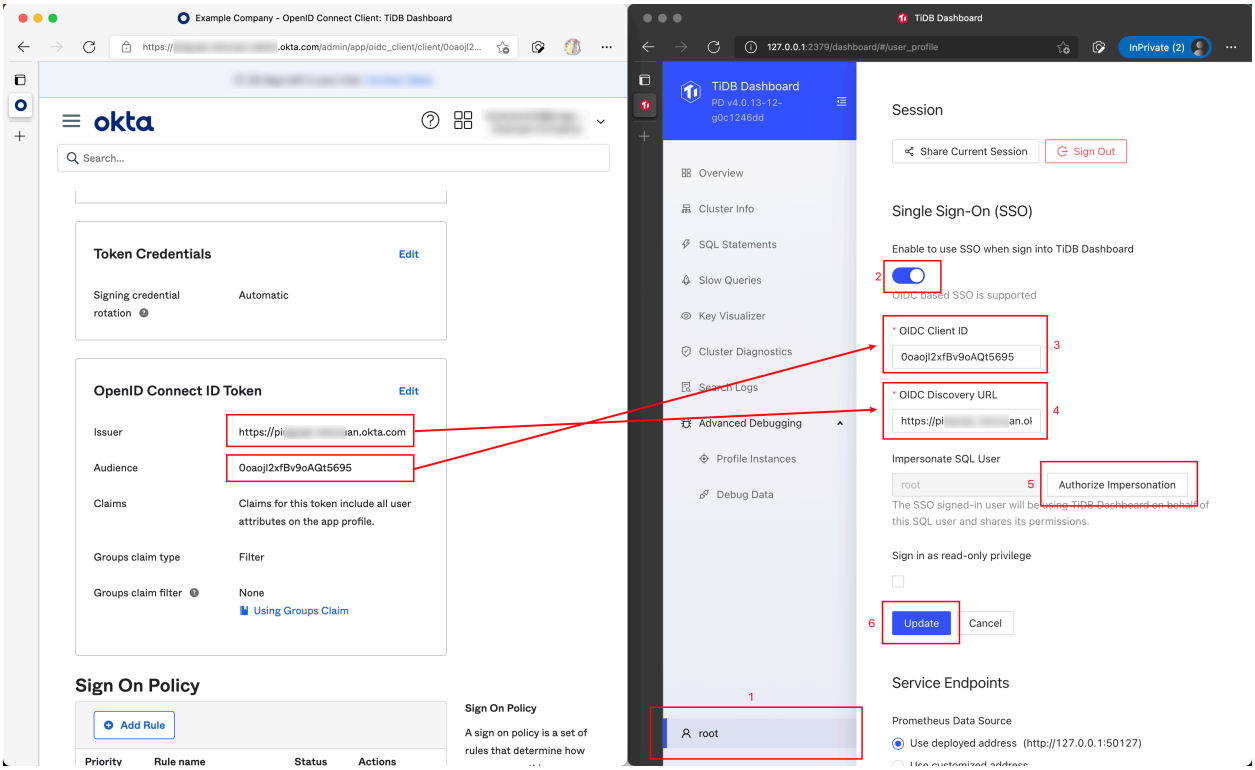


图 196: 操作示例 3

至此，TiDB Dashboard 已被配置为使用 Okta 进行 SSO 登录。

9.9.14.2.4 示例二：使用 Auth0 进行 TiDB Dashboard SSO 登录认证

和 Okta 类似，Auth0 也可以提供 OIDC SSO 的身份认证服务。

步骤一：配置 Auth0

1. 访问 Auth0 的管理后台。
2. 点击左侧边栏的 Applications > Applications。
3. 点击 Create Application，在弹出窗口中输入 Name，例如 “TiDB Dashboard”。在 Choose an application type 下选择 “Single Page Web Application”。

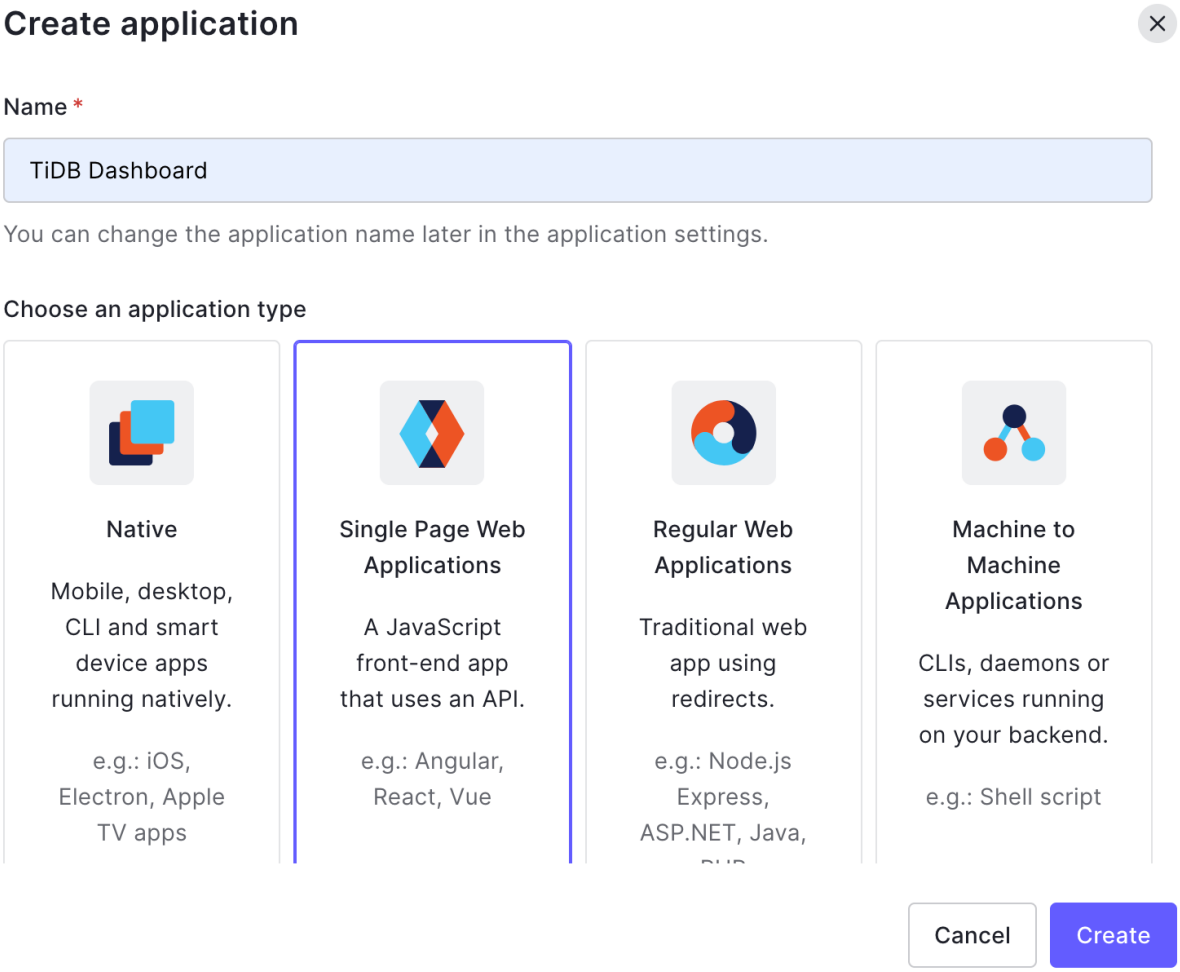


图 197: Create Application

4. 点击 Settings 栏。

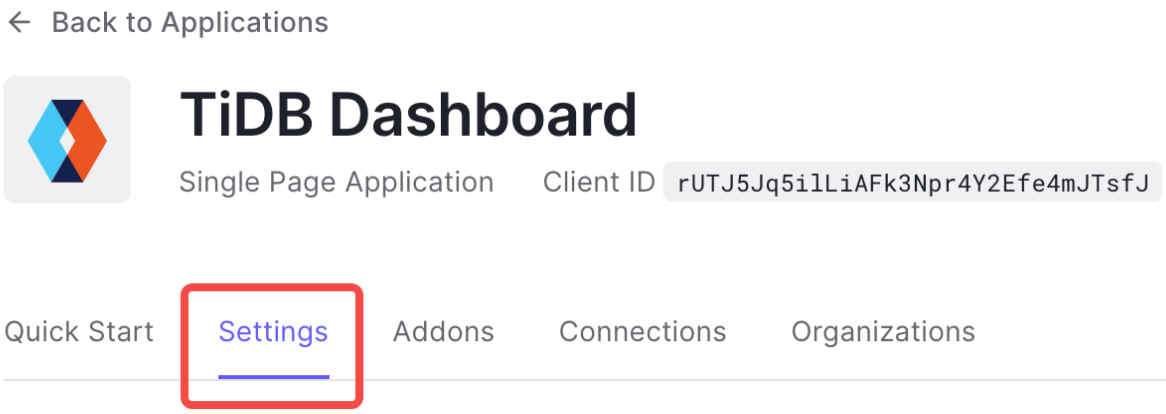


图 198: Settings

5. 在 Allowed Callback URLs 字段中填写如下内容：

```
http://DASHBOARD_IP:PORT/dashboard/?sso_callback=1
```

在以上内容中，将 DASHBOARD_IP:PORT 替换为你在浏览器中实际访问 TiDB Dashboard 所使用的域名（或 IP）及端口。

6. 在 Allowed Logout URLs 字段中填写如下内容：

```
http://DASHBOARD_IP:PORT/dashboard/
```

类似地，将 DASHBOARD_IP:PORT 替换为实际的域名（或 IP）及端口。

Allowed Callback URLs

http://localhost:3001/dashboard/?sso_callback=1,
http://127.0.0.1:2379/dashboard/?sso_callback

After the user authenticates we will only call back to any of these URLs. You can specify multiple valid URLs by comma-separating them (typically to handle different environments like QA or testing). Make sure to specify the protocol (`https://`) otherwise the callback may fail in some cases. With the exception of custom URI schemes for native clients, all callbacks should use protocol `https://`. You can use Organization URL parameters in these URLs.

Allowed Logout URLs

http://localhost:3001/dashboard/,
http://127.0.0.1:2379/dashboard/

A set of URLs that are valid to redirect to after logout from Auth0. After a user logs out from Auth0 you can redirect them with the `returnTo` query parameter. The URL that you use in `returnTo` must be listed here. You can specify multiple valid URLs by comma-separating them. You can use the star symbol as a wildcard for subdomains (`*.google.com`). Query strings and hash information are not taken into account when validating these URLs. Read more about this at https://auth0.com/docs/login/logout

图 199: Settings

7. 其它设置保持默认，点击 Save Changes 保存。

步骤二：获取 TiDB Dashboard 所需的配置参数并填入 TiDB Dashboard

1. 将 Auth0 Settings 栏 Basic Information 项的 Client ID 字段的值填入 TiDB Dashboard 的 OIDC Client ID，将 Domain 字段的值，加上 `https://` 前缀和 `/` 后缀后填入 OIDC Discovery URL 中，例如 `https://example.us.auth0.com/`。完成授权并保存配置即可。

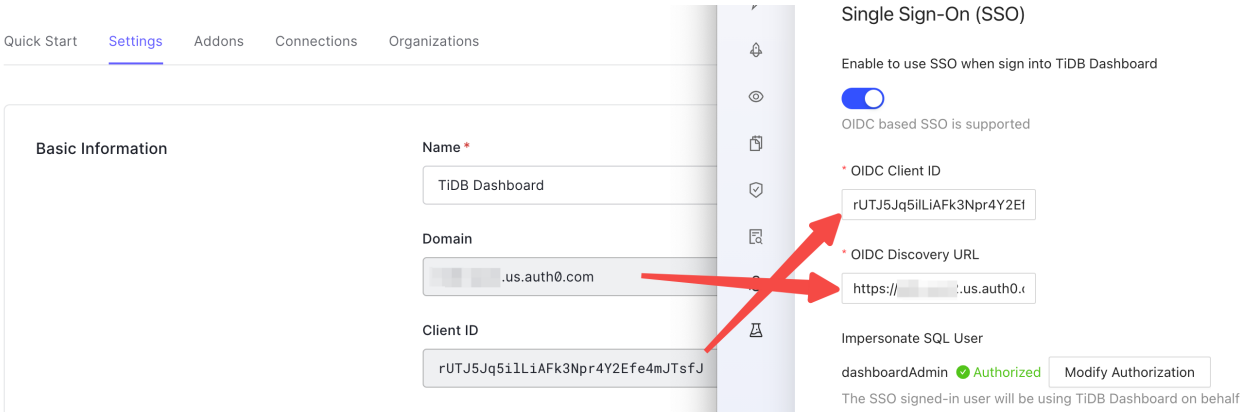


图 200: Settings

至此，TiDB Dashboard 已被配置为使用 Auth0 进行 SSO 登录。

9.9.14.2.5 示例三：使用 Casdoor 进行 TiDB Dashboard SSO 登录认证

Casdoor 是一个开源的、可以部署在私人服务器上的 SSO 平台。它与 TiDB Dashboard 的 SSO 功能兼容。以下步骤展示了如何配置 Casdoor 及 TiDB Dashboard，使得 TiDB Dashboard 可以通过 Casdoor 进行 SSO 登录。

步骤一：配置 Casdoor

1. 部署并访问 Casdoor 的管理后台。
2. 点击上方菜单栏的应用。
3. 点击添加按钮。

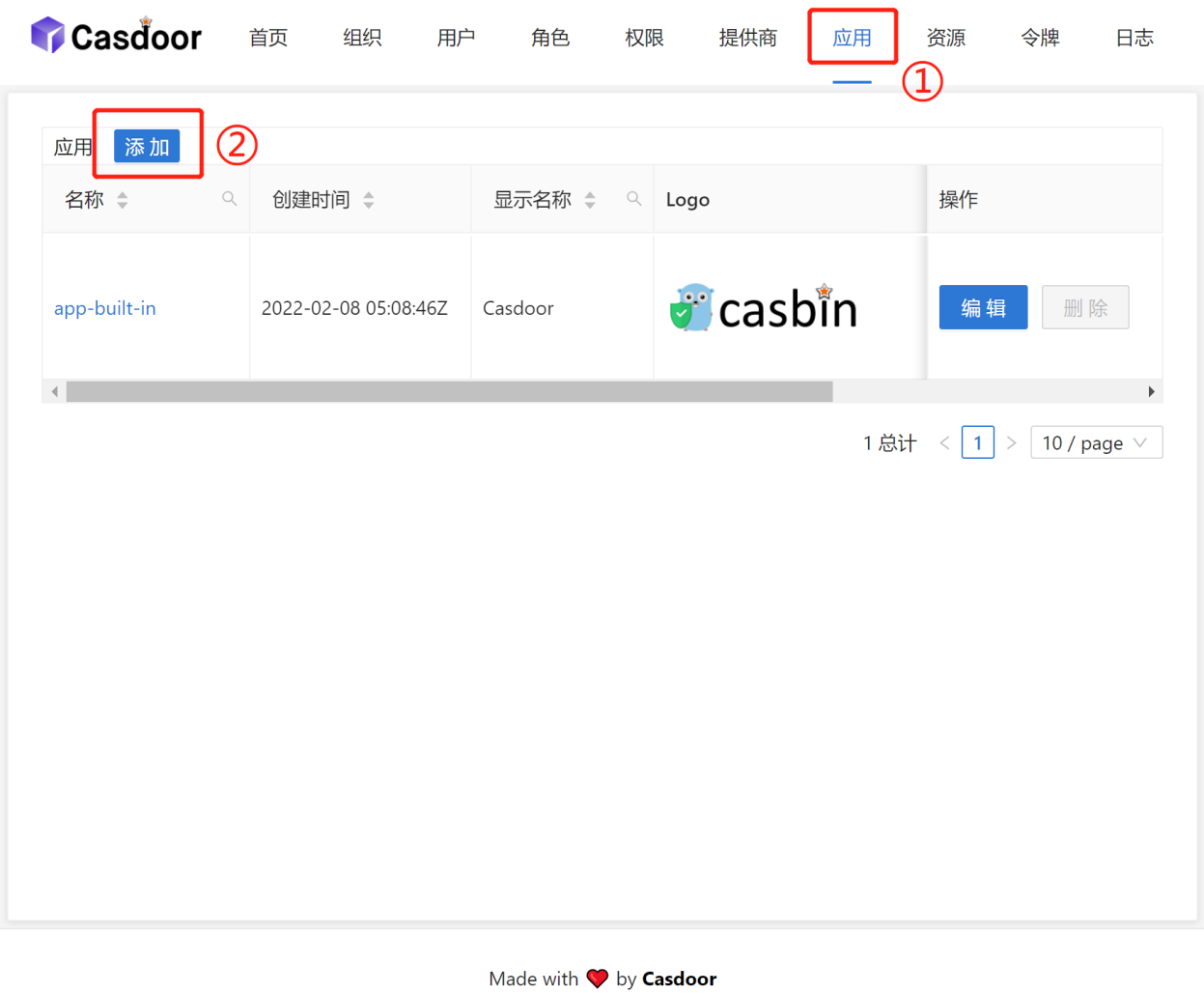


图 201: Settings

4. 填写名称和显示名称，比如：TiDB Dashboard 。
5. 在回调 URLs 中添加如下内容：

```
http://DASHBOARD_IP:PORT/dashboard/?sso_callback=1
```

在以上内容中，将 DASHBOARD_IP:PORT 替换为你在浏览器中实际访问 TiDB Dashboard 所使用的域名（或 IP）及端口。

回调URLs ⓘ：

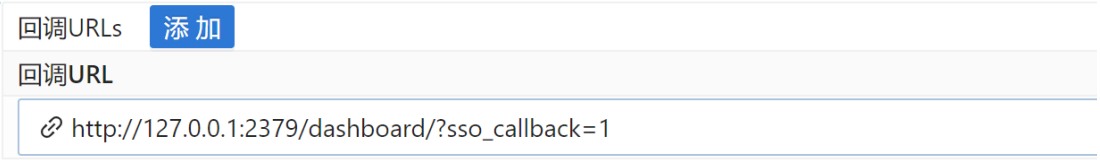


图 202: Settings

6. 其他设置保留默认值，点击保存 & 退出。
7. 保存页面上的客户端 ID。

步骤二：获取 TiDB Dashboard 所需的配置参数并填入 TiDB Dashboard

1. 将上一步获得的客户端 ID 字段的值填入 TiDB Dashboard 的 OIDC Client ID。
2. 将 Casdoor 部署地址加上 `https://` 前缀和 `/` 后缀后填入 OIDC Discovery URL 中，例如 `https://casdoor.example.com/`。完成授权并保存配置即可。

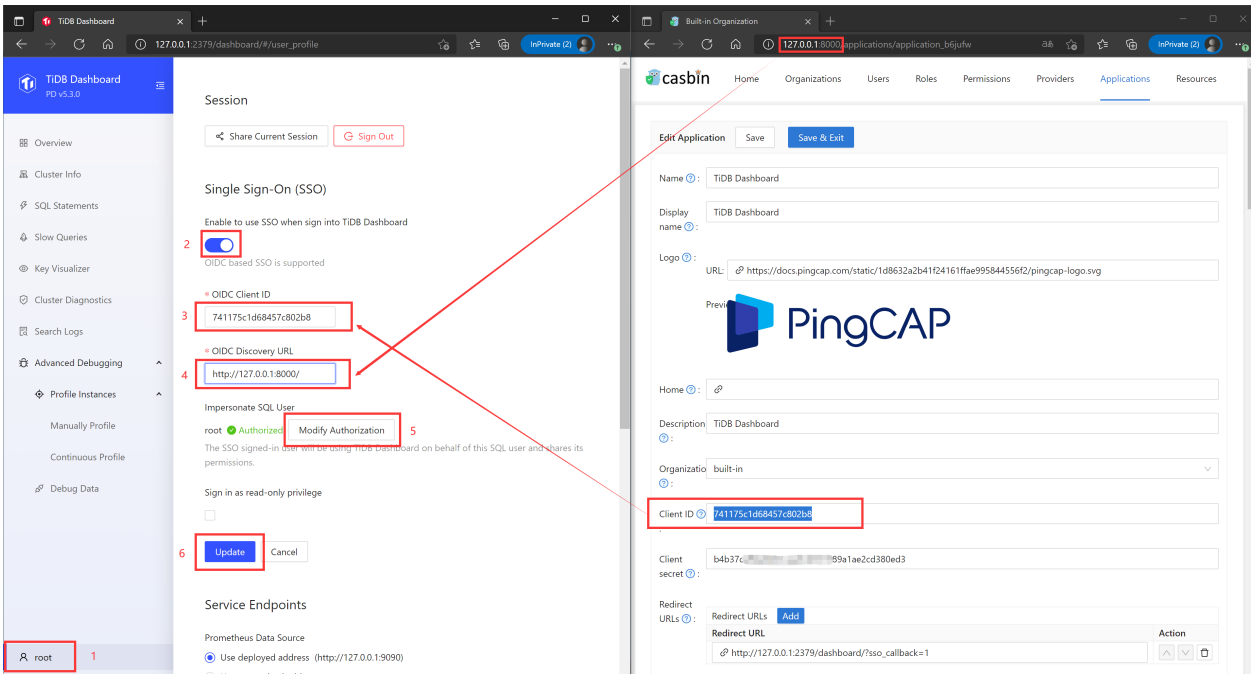


图 203: Settings

至此，TiDB Dashboard 已被配置为使用 Casdoor 进行 SSO 登录。

9.9.15 TiDB Dashboard 常见问题

本文汇总了使用 TiDB Dashboard 过程中的常见问题与解决办法。若无法找到对应问题，或者根据指引操作后问题仍然存在，请联系 PingCAP 技术支持获取帮助。

9.9.15.1 访问

9.9.15.1.1 已配置防火墙或反向代理，但访问后被跳转到一个内部地址无法访问 TiDB Dashboard

集群部署有多个 PD 实例的情况下，只有其中某一个 PD 实例会真正运行 TiDB Dashbaord 服务，访问其他 PD 实例时会发生浏览器端重定向。若防火墙或反向代理没有为此进行正确配置，就可能出现访问后被重定向到一个被防火墙或反向代理保护的内部地址的情况。

- 参阅TiDB Dashboard 多 PD 实例部署章节了解多 PD 实例下 TiDB Dashboard 的工作原理。
- 参阅通过反向代理使用 TiDB Dashboard 章节了解如何正确配置反向代理。
- 参阅提高 TiDB Dashboard 安全性章节了解如何正确配置防火墙。

9.9.15.1.2 双网卡部署时无法通过另一个网卡访问 TiDB Dashboard

PD 中的 TiDB Dashboard 出于安全考虑仅监听部署时所指定的 IP 地址（即只监听在一个网卡上），而非 `0.0.0.0`，因此当主机上安装了多个网卡时，通过另一个网卡将无法访问。

当你使用 `tiup cluster` 或 `tiup playground` 命令部署时，目前尚没有方法改变该行为。推荐使用反向代理将 TiDB Dashboard 安全地暴露给另一个网卡，具体参见通过反向代理使用 TiDB Dashboard 章节。

9.9.15.2 界面功能

9.9.15.2.1 概况页面中 QPS 及 Latency 显示 `prometheus_not_found` 错误

QPS 及 Latency 监控依赖于集群中已正常部署 Prometheus 监控实例，没有部署的情况下就会显示为错误。向集群中新部署 Prometheus 实例即可解决该问题。

若已经部署 Prometheus 监控实例但仍然显示为错误，可能的原因是您使用的部署工具（TiUP 或 TiDB Operator）版本比较旧，没有自动汇报监控地址，导致 TiDB Dashboard 无法感知并查询监控数据。可以升级到最新的部署工具并重试。

以下给出 TiUP 部署工具的操作方法，对于其他部署工具，请参阅工具对应文档。

1. 升级 TiUP、TiUP Cluster：

```
bash  tiup update --self tiup update cluster --force
```

2. 升级后，部署包含监控节点的新集群时，应当能正常显示监控。
3. 升级后，对于现有集群，可通过再次启动集群的方法汇报监控地址（将 CLUSTER_NAME 替换为实际集群名称）：

```
bash  tiup cluster start CLUSTER_NAME
```

即使集群已经启动，请仍然执行该命令。该命令不会影响集群上正常的业务，但会刷新并上报监控地址，从而能让监控在 TiDB Dashboard 中正常显示。

9.9.15.2.2 概况页面中 Top SQL 语句、最近慢查询显示 invalid connection 错误

可能的原因是你开启了 TiDB 的 prepared-plan-cache 功能。prepared-plan-cache 是实验性功能，在某些版本的 TiDB 中可能无法正常运行，开启后可能会导致 TiDB Dashboard（及其他应用）出现该问题。可以通过修改 TiDB 配置文件来关闭 prepared-plan-cache 功能。

9.9.15.2.3 慢查询页面显示 unknown field 错误

集群升级后，如果慢查询页面出现 unknown field 错误，是由于升级后新版本 TiDB Dashboard 字段与浏览器缓存内的用户偏好设置的字段不兼容导致的。该问题已修复。如果你的集群版本低于 v5.0.3 或 v4.0.14，可以根据以下步骤进行处理：

1. 打开 TiDB Dashboard 页面。
2. 打开浏览器的开发者工具。各浏览器的打开方式不同。

    - Firefox：菜单 > Web 开发者 > 切换工具箱（译者注：此处修改为最新的 Firefox Quantum），或者工具栏 > 切换工具箱。
    - Chrome：菜单 > 更多工具 > 开发者工具。
    - Safari：Develop > Show Web Inspector。如果你看不到 Develop 菜单，选择 Preferences > Advanced，然后点击 Show Develop menu in menu bar 复选框。

    以 Chrome 为例：

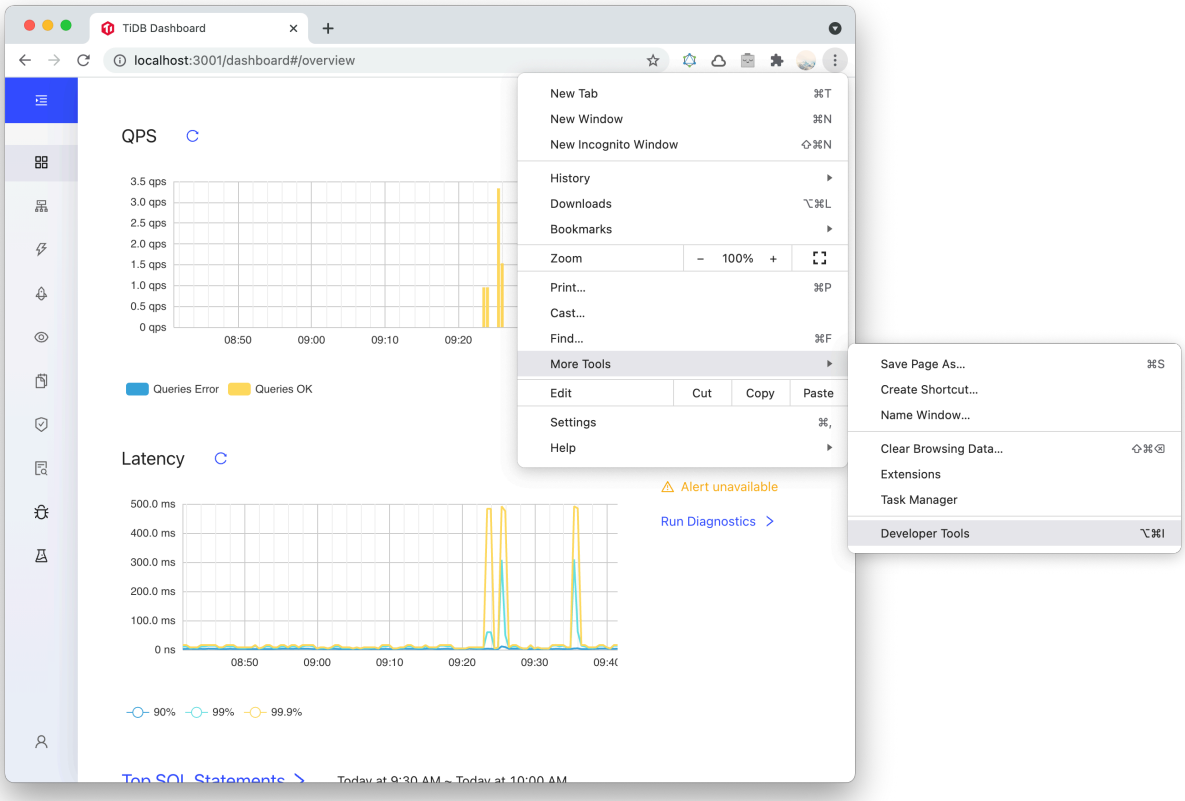


图 204: 打开开发者工具

3. 选中 Application 面板，展开 Local Storage 菜单并选中 TiDB Dashboard 页面的域名，点击 Clear All。

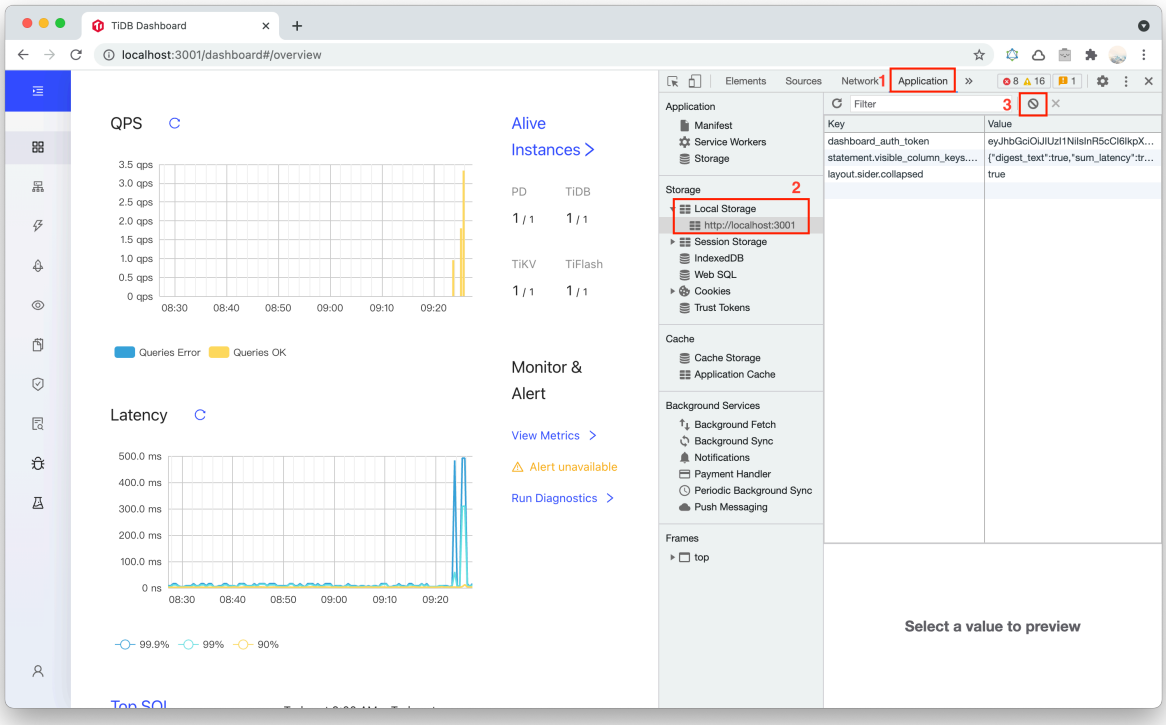


图 205: 清理 Local Storage

9.9.15.2.4 界面提示 集群中未启动必要组件 NgMonitoring

NgMonitoring 是 TiDB v5.4.0 及以上集群中内置的高级监控组件，用于支撑 TiDB Dashboard 的 持续分析 (Continuous Profiling) 和 Top SQL 等功能。使用 TiUP 部署集群时，NgMonitoring 会自动部署；使用 TiDB Operator 部署集群时，需要依据启用持续性能分析手动部署 NgMonitoring。

如果界面提示 集群中未启动必要组件 NgMonitoring，需要按 TiDB 集群部署方式排查问题。

使用 TiUP 部署的集群

第 1 步：检查 TiUP Cluster 版本

1. 检查 TiUP Cluster 版本，NgMonitoring 组件需要较高版本的部署工具支持（TiUP v1.9.0 及以上）：

```
```shell
tiup cluster --version
```

上述命令可查看 TiUP Cluster 的具体版本。例如：

```
tiup version 1.9.0 tiup
Go Version: go1.17.2
```
```

```
Git Ref: v1.9.0
```

2. 如果 TiUP 版本低于 v1.9.0，升级 TiUP 和 TiUP Cluster 版本至最新。

```shell
tiup update --all
```

第 2 步：在中控机上，通过 TiUP 添加 ng_port 配置项，然后重启 Prometheus 节点。

1. 以编辑模式打开集群的配置文件：

```shell
tiup cluster edit-config ${cluster-name}
```

2. 在 monitoring_servers 下面增加 ng_port:${port} 参数：

```
monitoring_servers:
- host: 172.16.6.6
  ng_port: ${port}
```

3. 重启 Prometheus 节点：

```shell
tiup cluster reload ${cluster-name} --role prometheus
```

执行完以上操作后，在 TiDB Dashboard 上启用持续性能分析特性。如果 NgMonitoring 依然无法加载，请联系 PingCAP 技术支持获取帮助。

使用 TiDB Operator 部署的集群

参考启用持续性能分析中的步骤部署 NgMonitoring。

10 性能调优

10.1 系统调优

10.1.1 操作系统性能参数调优

本文档仅用于描述如何优化 CentOS 7 的各个子系统。

注意：

- CentOS 7 操作系统的默认配置适用于中等负载下运行的大多数服务。调整特定子系统的性能可能会对其他子系统产生负面影响。因此在调整系统之前，请备份所有用户数据和配置信息；
- 请在测试环境下对所有修改做好充分测试后，再应用到生产环境中。

10.1.1.1 性能分析工具

系统调优需要根据系统性能分析的结果做指导，因此本文先列出常用的性能分析方法。

10.1.1.1.1 60 秒分析法

此分析法由《性能之巅》的作者 Brendan Gregg 及其所在的 Netflix 性能工程团队公布。所用到的工具均可从发行版的官方源获取，通过分析以下清单中的输出，可定位大部分常见的性能问题。

- uptime
- dmesg | tail
- vmstat 1
- mpstat -P ALL 1
- pidstat 1
- iostat -xz 1
- free -m
- sar -n DEV 1
- sar -n TCP,ETCP 1
- top

具体用法可查询相应 `man` 手册。

10.1.1.1.2 perf

Perf 是 Linux 内核提供的一个重要的性能分析工具，它涵盖硬件级别（CPU/PMU 和性能监视单元）功能和软件功能（软件计数器和跟踪点）。详细用法请参考 perf Examples。

10.1.1.1.3 BCC/bpftrace

CentOS 从 7.6 版本起，内核已实现对 bpf 的支持，因此可根据上述清单的结果，选取适当的工具进行深入分析。相比 perf/ftrace，bpf 提供了可编程能力和更小的性能开销。相比 kprobe，bpf 提供了更高的安全性，更适合在生产环境上使用。关于 BCC 工具集的使用请参考 BPF Compiler Collection (BCC)。

10.1.1.2 性能调优

性能调优将根据内核子系统进行分类描述。

10.1.1.2.1 处理器 ——动态节能技术

cpufreq 是一个动态调整 CPU 频率的模块，可支持五种模式。为保证服务性能应选用 performance 模式，将 CPU 频率固定工作在其支持的最高运行频率上，不进行动态调节，操作命令为 `cpupower frequency-set --governor performance`。

10.1.1.2.2 处理器 ——中断亲和性

- 自动平衡：可通过 `irqbalance` 服务实现。
- 手动平衡：
  - 确定需要平衡中断的设备，从 CentOS 7.5 开始，系统会自动为某些设备及其驱动程序配置最佳的中断关联性。不能再手动配置其亲和性。目前已知的有使用 `be2iscsi` 驱动的设备，以及 NVMe 设置；
  - 对于其他设备，可查询其芯片手册，是否支持分发中断，若不支持，则该设备的所有中断会路由到同一个 CPU 上，无法对其进行修改。若支持，则计算 `smp_affinity` 掩码并设置对应的配置文件，具体请参考内核文档。

10.1.1.2.3 NUMA 绑核

为尽可能的避免跨 NUMA 访问内存，可以通过设置线程的 CPU 亲和性来实现 NUMA 绑核。对于普通程序，可使用 `numactl` 命令来绑定，具体用法请查询 man 手册。对于网卡中断，请参考下文网络章节。

10.1.1.2.4 内存 ——透明大页

对于数据库应用，不推荐使用 THP，因为数据库往往具有稀疏而不是连续的内存访问模式，且当高阶内存碎片化比较严重时，分配 THP 页面会出现较大的延迟。若开启针对 THP 的直接内存规整功能，也会出现系统 CPU 使用率激增的现象，因此建议关闭 THP。

```
echo never > /sys/kernel/mm/transparent_hugepage/enabled
echo never > /sys/kernel/mm/transparent_hugepage/defrag
```

10.1.1.2.5 内存 ——虚拟内存参数

- `dirty_ratio` 百分比值。当脏的 page cache 总量达到系统内存总量的这一百分比后，系统将开始使用 pdflush 操作将脏的 page cache 写入磁盘。默认值为 20%，通常不需调整。对于高性能 SSD，比如 NVMe 设备来说，降低其值有利于提高内存回收时的效率。
- `dirty_background_ratio` 百分比值。当脏的 page cache 总量达到系统内存总量的这一百分比后，系统开始在后台将脏的 page cache 写入磁盘。默认值为 10%，通常不需调整。对于高性能 SSD，比如 NVMe 设备来说，设置较低的值有利于提高内存回收时的效率。

10.1.1.2.6 存储及文件系统

内核 I/O 栈链路较长，包含了文件系统层、块设备层和驱动层。

I/O 调度器

I/O 调度程序确定 I/O 操作何时在存储设备上运行以及持续多长时间。也称为 I/O 升降机。对于 SSD 设备，宜设置为 noop。

```
echo noop > /sys/block/${SSD_DEV_NAME}/queue/scheduler
```

格式化参数 ——块大小

块是文件系统的工作单元。块大小决定了单个块中可以存储多少数据，因此决定了一次写入或读取的最小数据量。

默认块大小适用于大多数使用情况。但是，如果块大小（或多个块的大小）与通常一次读取或写入的数据量相同或稍大，则文件系统将性能更好，数据存储效率更高。小文件仍将使用整个块。文件可以分布在多个块中，但这会增加运行时开销。

使用 mkfs 命令格式化设备时，将块大小指定为文件系统选项的一部分。指定块大小的参数随文件系统的不同而不同。有关详细信息，请查询对应文件系统的 `mkfs` 手册页，比如 `man mkfs.ext4`。

挂载参数

noatime 读取文件时，将禁用对元数据的更新。它还启用了 nodiratime 行为，该行为会在读取目录时禁用对元数据的更新。

10.1.1.2.7 网络

网络子系统由具有敏感连接的许多不同部分组成。因此，CentOS 7 网络子系统旨在为大多数工作负载提供最佳性能，并自动优化其性能。因此，通常无需手动调整网络性能。

网络问题通常由硬件或相关设施出现问题导致的，因此在调优协议栈前，请先排除硬件问题。

尽管网络堆栈在很大程度上是自我优化的。但是在网络数据包处理过程中，以下方面可能会成为瓶颈并降低性能：

- 网卡硬件缓存：正确观察硬件层面的丢包方法是使用 `ethtool -S ${NIC_DEV_NAME}` 命令观察 drops 字段。当出现丢包现象时，主要考虑是硬/软中断的处理速度跟不上网卡接收速度。若接收缓存小于最大限制时，也可尝试增加 RX 缓存来防止丢包。查询命令为：`ethtool -g ${NIC_DEV_NAME}`，修改命令为 `ethtool -G ${NIC_DEV_NAME}`。
- 硬中断：若网卡支持 Receive-Side Scaling（RSS 也称为多网卡接收）功能，则观察 `/proc/interrputs` 网卡中断，如果出现了中断不均衡的情况，请参考处理器调优章节。若不支持 RSS 或 RSS 数量远小于物理 CPU 核数，则可配置 Receive Packet Steering（RPS，可以看作 RSS 的软件实现），及 RPS 的扩展 Receive Flow Steering (RFS)。具体设置请参考内核文档。
- 软中断：观察 `/proc/net/softnet\_stat` 监控。如果除第三列的其他列的数值在增长，则应适度调大 `net.core.netdev\_budget` 或 `net.core.dev\_weight` 值，使 softirq 可以获得更多的 CPU 时间。除此之外，也需要检查 CPU 使用情况，确定哪些任务在频繁占用 CPU，能否优化。
- 应用的套接字接收队列：监控 `ss -nmp` 的 Recv-q 列，若队列已满，则应考虑增大应用程序套接字的缓存大小或使用自动调整缓存的方式。除此之外，也要考虑能否优化应用层的架构，降低读取套接字的间隔。
- 以太网流控：若网卡和交换机支持流控功能，可通过使能此功能，给内核一些时间来处理网卡队列中的数据，来规避网卡缓存溢出的问题。对于网卡测，可通过 `ethtool -a ${NIC_DEV_NAME}` 命令检查是否支持/使能，并通过 `ethtool -A ${NIC_DEV_NAME}` 命令开启。对于交换机，请查询其手册。
- 中断合并：过于频繁的硬件中断会降低系统性能，而过晚的硬件中断会导致丢包。对于较新的网卡支持中断合并功能，并允许驱动自动调节硬件中断数。可通过 `ethtool -c ${NIC_DEV_NAME}` 命令检查，`ethtool -C ${NIC_DEV_NAME}` 命令开启。自适应模式使网卡可以自动调节中断合并。在自适应模式下，

驱动程序将检查流量模式和内核接收模式，并实时评估合并设置，以防止数据包丢失。不同品牌的网卡具有不同的功能和默认配置，具体请参考网卡手册。

- 适配器队列：在协议栈处理之前，内核利用此队列缓存网卡接收的数据，每个 CPU 都有各自的 backlog 队列。此队列可缓存的最大 packets 数量为 `netdev\_max\_backlog`。观察 `/proc/net/softnet\_stat` 第二列，当某行的第二列持续增加，则意味着 CPU [行 -1] 队列已满，数据包被丢失，可通过持续加倍 `net.core.netdev\_max\_backlog` 值来解决。
- 发送队列：发送队列长度值确定在发送之前可以排队的数据包数量。默认值是 1000，对于 10 Gbps 足够。但若从 `ip -s link` 的输出中观察到 `TX errors` 值时，可尝试加倍该数据包数量：`ip link set dev ${NIC_DEV_NAME} txqueuelen 2000`。
- 驱动：网卡驱动通常也会提供调优参数，请查询设备硬件手册及其驱动文档。

10.2 软件调优

10.2.1 配置

10.2.1.1 TiDB 内存控制文档

目前 TiDB 已经能够做到追踪单条 SQL 查询过程中的内存使用情况，当内存使用超过一定阈值后也能采取一些操作来预防 OOM 或者排查 OOM 原因。在 TiDB 的配置文件中，我们可以使用如下配置来控制内存使用超阈值时 TiDB 的行为：

```
#### Valid options: ["log", "cancel"]
oom-action = "cancel"
```

- 如果上面的配置项使用的是 “log”，那么当一条 SQL 的内存使用超过一定阈值（由 session 变量 `tidb_mem_quota_query` 来控制）后，TiDB 会在 log 文件中打印一条 LOG，然后这条 SQL 继续执行，之后如果发生了 OOM 可以在 LOG 中找到对应的 SQL。
- 如果上面的配置项使用的是 “cancel”，那么当一条 SQL 的内存使用超过一定阈值后，TiDB 会立即中断这条 SQL 的执行并给客户端返回一个 error，error 信息中会详细写明这条 SQL 执行过程中各个占用内存比较多的物理执行算子的内存使用情况。

10.2.1.1.1 如何配置一条 SQL 执行过程中的内存使用阈值

可以在配置文件中设置每个 Query 默认的 Memory Quota，例如将其设置为 32GB：

```
mem-quota-query = 34359738368
```

此外还可通过 session 变量 `tidb_mem_quota_query` 来控制一条 Query 中的内存使用。几个使用例子：

配置整条 SQL 的内存使用阈值为 8GB：

```
set @@tidb_mem_quota_query = 8 << 30;
```

配置整条 SQL 的内存使用阈值为 8MB：

```
set @@tidb_mem_quota_query = 8 << 20;
```

配置整条 SQL 的内存使用阈值为 8KB：

```
set @@tidb_mem_quota_query = 8 << 10;
```

10.2.1.1.2 如何配置 tidb-server 实例使用内存的阈值

可以在配置文件中设置 tidb-server 实例的内存使用阈值。相关配置项为`server-memory-quota`。

例如，配置 tidb-server 实例的内存使用总量，将其设置成为 32 GB：

```
[performance]
server-memory-quota = 34359738368
```

在该配置下，当 tidb-server 实例内存使用到达 32 GB 时，正在执行的 SQL 语句会被随机强制终止，直至 tidb-server 实例内存使用下降到 32 GB 以下。被强制终止的 SQL 操作会向客户端返回 `Out Of Global Memory Limit!` 错误信息。

> 警告：
>
> - `server-memory-quota` 目前为实验性特性，不建议在生产环境中使用。
> - `server-memory-quota` 默认值为 0，表示无内存限制。

10.2.1.1.3 tidb-server 内存占用过高时的报警

默认配置下，tidb-server 实例会在机器内存使用达到总内存量的 80% 时打印报警日志，并记录相关状态文件。该内存使用率可以通过配置项`memory-usage-alarm-ratio`进行设置。具体报警规则请参考该配置项的说明部分。

注意，当触发一次报警后，只有在内存使用率连续低于阈值超过 10 秒并再次达到阈值时，才会再次触发报警。此外，为避免报警时产生的状态文件积累过多，目前只会保留最近 5 次报警时所生成的状态文件。

下例通过构造一个占用大量内存的 SQL 语句触发报警，对该报警功能进行演示：

1. 配置报警比例为 `0.8`：

   ```
   mem-quota-query = 34359738368  // 将单条 SQL 内存限制调高，以便于构造占用内存较大的 SQL
   [performance]
   memory-usage-alarm-ratio = 0.8
   ```

2. 创建单表 `CREATE TABLE t(a int);` 并插入 1000 行数据。

3. 执行 `select * from t t1 join t t2 join t t3 order by t1.a`。该 SQL 语句会输出 1000000000 条记录，占用巨大的内存，进而触发报警。

4. 检查 `tidb.log` 文件，其中会记录系统总内存、系统当前内存使用量、tidb-server 实例的内存使用量以及状态文件所在目录。

```
[2020/11/30 15:25:17.252 +08:00] [WARN] [memory_usage_alarm.go:141] ["tidb-server has the
    ↪ risk of OOM. Running SQLs and heap profile will be recorded in record path"] ["is
    ↪ server-memory-quota set"=false] ["system memory total"=33682427904] ["system memory
    ↪ usage"=27142864896] ["tidb-server memory usage"=22417922896] [memory-usage-alarm-
    ↪ ratio=0.8] ["record path"="/tmp/1000_tidb/MC4wLjAuMDo0MDAwLzAuMC4wLjA6MTAwODA=/tmp-
    ↪ storage/record"]
```

以上 Log 字段的含义如下：

- `is server-memory-quota set`：表示配置项`server-memory-quota` 是否被设置
- `system memory total`：表示当前系统的总内存
- `system memory usage`：表示当前系统的内存使用量
- `tidb-server memory usage`：表示 tidb-server 实例的内存使用量
- `memory-usage-alarm-ratio`：表示配置项`memory-usage-alarm-ratio` 的值
- `record path`：表示状态文件存放的目录

5. 通过访问状态文件所在目录 (该示例中的目录为 `/tmp/1000_tidb/MC4wLjAuMDo0MDAwLzAuMC4wLjA6MTAwODA=/tmp-storage/record`)，可以得到一组文件，其中包括 goroutinue、heap、running_sql 3 个文件，文件以记录状态文件的时间为后缀。这 3 个文件分别用来记录报警时的 goroutine 栈信息，堆内存使用状态，及正在运行的 SQL 信息。其中 `running_sql` 文件内的日志格式请参考`expensive-queries`。

10.2.1.1.4 tidb-server 其它内存控制策略

流量控制

- TiDB 支持对读数据算子的动态内存控制功能。读数据的算子默认启用`tidb_distsql_scan_concurrency` 所允许的最大线程数来读取数据。当单条 SQL 语句的内存使用每超过`tidb_mem_quota_query` 一次，读数据的算子就会停止一个线程。
- 流控行为由参数`tidb_enable_rate_limit_action` 控制。
- 当流控被触发时，会在日志中打印一条包含关键字 `memory exceeds quota, destroy one token now` 的日志。

数据落盘

TiDB 支持对执行算子的数据落盘功能。当 SQL 的内存使用超过 Memory Quota 时，tidb-server 可以通过落盘执行算子的中间数据，缓解内存压力。支持落盘的算子有：Sort、 MergeJoin、 HashJoin、 HashAgg。

- 落盘行为由参数`mem-quota-query`、`oom-use-tmp-storage`、`tmp-storage-path`、`tmp-storage-quota` 共同控制。
- 当落盘被触发时，TiDB 会在日志中打印一条包含关键字 `memory exceeds quota, spill to disk now` 或 `memory exceeds quota, set aggregate mode to spill-mode` 的日志。
- Sort、 MergeJoin、 HashJoin 落盘是从 v4.0.0 版本开始引入的，HashAgg 落盘是从 v5.2.0 版本开始引入的。
- 当包含 Sort、MergeJoin 或 HashJoin 的 SQL 语句引起内存 OOM 时，TiDB 默认会触发落盘。当包含 HashAgg 算子的 SQL 语句引起内存 OOM 时，TiDB 默认不触发落盘，请设置系统变量 `tidb_executor_concurrency = 1` 来触发 HashAgg 落盘功能。

> **注意：**
>
> - HashAgg 落盘功能目前不支持 distinct 聚合函数。使用 distinct 函数且内存占用过大时，无法进行落盘。

本示例通过构造一个占用大量内存的 SQL 语句，对 HashAgg 落盘功能进行演示：

1. 将 SQL 语句的 Memory Quota 配置为 1GB（默认 1GB）：

```
set tidb_mem_quota_query = 1 << 30;
```

2. 创建单表 `CREATE TABLE t(a int);` 并插入 256 行不同的数据。

3. 尝试执行以下 SQL 语句：

```
[tidb]> explain analyze select /*+ HASH_AGG() */ count(*) from t t1 join t t2 join t t3
    ↪ group by t1.a, t2.a, t3.a;
```

   该 SQL 语句占用大量内存，返回 Out of Memory Quota 错误。

```
ERROR 1105 (HY000): Out Of Memory Quota![conn_id=3]
```

4. 设置系统变量 `tidb_executor_concurrency` 将执行器的并发度调整为 1。在此配置下，内存不足时 HashAgg 会自动尝试触发落盘。

```
set tidb_executor_concurrency = 1;
```

5. 执行相同的 SQL 语句，不再返回错误，可以执行成功。从详细的执行计划可以看出，HashAgg 使用了 600MB 的硬盘空间。

```
[tidb]> explain analyze select /*+ HASH_AGG() */ count(*) from t t1 join t t2 join t t3
    ↪ group by t1.a, t2.a, t3.a;
```

```
+--
    ↪ -------------------------------+-------------+----------+-----------+---------------+---------------
    ↪
| id                              | estRows     | actRows  | task      | access object |
    ↪ execution info
    ↪
    ↪ | operator info                                          | memory    | disk
    ↪      |
+--
    ↪ -------------------------------+-------------+----------+-----------+---------------+---------------
    ↪
```

```
| HashAgg_11                      | 204.80      | 16777216 | root      |               |
    ↪ time:1m37.4s, loops:16385
    ↪
    ↪ | group by:test.t.a, test.t.a, test.t.a, funcs:count(1)->Column#7 | 1.13 GB   |
    ↪ 600.0 MB |
| └─HashJoin_12                   | 16777216.00 | 16777216 | root      |               |
    ↪ time:21.5s, loops:16385, build_hash_table:{total:267.2µs, fetch:228.9µs, build:38.2
    ↪ µs}, probe:{concurrency:1, total:35s, max:35s, probe:35s, fetch:962.2µs}         |
    ↪ CARTESIAN inner join                                          | 8.23 KB   | 4 KB
    ↪     |
|   ├─TableReader_21(Build)       | 256.00      | 256      | root      |               |
    ↪ time:87.2µs, loops:2, cop_task: {num: 1, max: 150µs, proc_keys: 0, rpc_num: 1,
    ↪ rpc_time: 145.1µs, copr_cache_hit_ratio: 0.00}
    ↪                                       | data:TableFullScan_20
    ↪                                           | 885 Bytes | N/A      |
|   │ └─TableFullScan_20          | 256.00      | 256      | cop[tikv] | table:t3      |
    ↪ tikv_task:{time:23.2µs, loops:256}
    ↪
    ↪ | keep order:false, stats:pseudo                                  | N/A       | N/A
    ↪     |
|   └─HashJoin_14(Probe)          | 65536.00    | 65536    | root      |               |
    ↪ time:728.1µs, loops:65, build_hash_table:{total:307.5µs, fetch:277.6µs, build:29.9µs
    ↪ }, probe:{concurrency:1, total:34.3s, max:34.3s, probe:34.3s, fetch:278µs}      |
    ↪ CARTESIAN inner join                                          | 8.23 KB   | 4 KB
    ↪     |
|     ├─TableReader_19(Build)     | 256.00      | 256      | root      |               |
    ↪ time:126.2µs, loops:2, cop_task: {num: 1, max: 308.4µs, proc_keys: 0, rpc_num: 1,
    ↪ rpc_time: 295.3µs, copr_cache_hit_ratio: 0.00}                                  |
    ↪ data:TableFullScan_18                                         | 885 Bytes | N/A
    ↪     |
|     │ └─TableFullScan_18        | 256.00      | 256      | cop[tikv] | table:t2      |
    ↪ tikv_task:{time:79.2µs, loops:256}
    ↪
    ↪ | keep order:false, stats:pseudo                                  | N/A       | N/A
    ↪     |
|     └─TableReader_17(Probe)     | 256.00      | 256      | root      |               |
    ↪ time:211.1µs, loops:2, cop_task: {num: 1, max: 295.5µs, proc_keys: 0, rpc_num: 1,
    ↪ rpc_time: 279.7µs, copr_cache_hit_ratio: 0.00}                                  |
    ↪ data:TableFullScan_16                                         | 885 Bytes | N/A
    ↪     |
|       └─TableFullScan_16        | 256.00      | 256      | cop[tikv] | table:t1      |
    ↪ tikv_task:{time:71.4µs, loops:256}
    ↪
    ↪ | keep order:false, stats:pseudo                                  | N/A       | N/A
    ↪     |
```

```
+--
    ↪ ------------------------------+-------------+----------+-----------+---------------+--------------
    ↪
9 rows in set (1 min 37.428 sec)
```

10.2.1.2 调度调优

10.2.1.2.1 Placement Rules 使用文档

> **注意：**
>
> TiDB 在 v5.3.0 中引入了实验特性 Placement Rules in SQL。使用该功能，你可以更方便地配置表和分区的位置。在未来版本中，Placement Rules in SQL 可能取代本文档介绍的 Placement Rules 功能。

Placement Rules 是 PD 在 4.0 版本引入的一套副本规则系统，用于指导 PD 针对不同类型的数据生成对应的调度。通过组合不同的调度规则，用户可以精细地控制任何一段连续数据的副本数量、存放位置、主机类型、是否参与 Raft 投票、是否可以担任 Raft leader 等属性。

Placement Rules 特性在 TiDB v5.0 及以上的版本中默认开启。如需关闭 Placement Rules 特性，请参考关闭 Placement Rules。

规则系统介绍

整个规则系统的配置由多条规则即 Rule 组成。每条 Rule 可以指定不同的副本数量、Raft 角色、放置位置等属性，以及这条规则生效的 key range。PD 在进行调度时，会先根据 Region 的 key range 在规则系统中查到该 Region 对应的规则，然后再生成对应的调度，来使得 Region 副本的分布情况符合 Rule。

多条规则的 key range 可以有重叠部分的，即一个 Region 能匹配到多条规则。这种情况下 PD 根据 Rule 的属性来决定规则是相互覆盖还是同时生效。如果有多条规则同时生效，PD 会按照规则的堆叠次序依次去生成调度进行规则匹配。

此外，为了满足不同来源的规则相互隔离的需求，支持更灵活的方式来组织规则，还引入了分组 (Group) 的概念。通常情况下，用户可根据规则的不同来源把规则放置在不同的 Group。

Placement Rules 示意图如下所示：

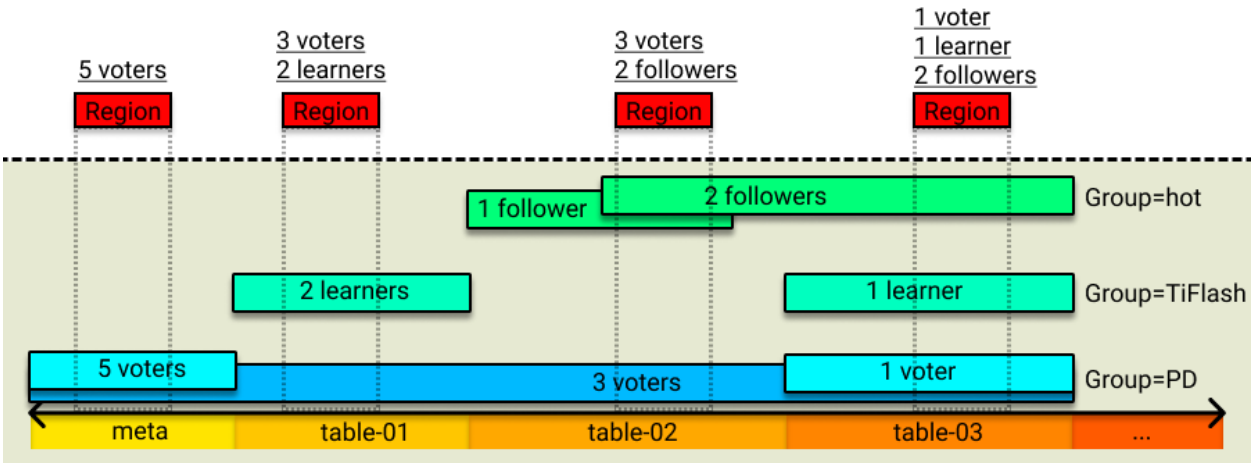


图 206: Placement rules overview

规则字段

以下是每条规则中各个字段的具体含义：

字段名	类型及约束	说明
GroupID	string	分组 ID，标识规则的来源
ID	string	分组内唯一 ID
Index	int	分组内堆叠次序
Override	true/false	是否覆盖 index 的更小 Rule（限分组内）
StartKey	string，十六进制编码	适用 Range 起始 key
EndKey	string，十六进制编码	适用 Range 终止 key
Role	string	副本角色，包括 leader/follower/learner
Count	int，正整数	副本数量
LabelConstraint	[]Constraint	用于按 label 筛选节点
LocationLabels	[]string	用于物理隔离
IsolationLevel	string	用于设置最小强制物理隔离级别

LabelConstraint 与 Kubernetes 中的功能类似，支持通过 in、notIn、exists 和 notExists 四种原语来筛选 label。这四种原语的意义如下：

- in：给定 key 的 label value 包含在给定列表中。
- notIn：给定 key 的 label value 不包含在给定列表中。
- exists：包含给定的 label key。
- notExists：不包含给定的 label key。

LocationLabels 的意义和作用与 PD v4.0 之前的版本相同。比如配置 [zone,rack,host] 定义了三层的拓扑结构：集群分为多个 zone（可用区），每个 zone 下有多个 rack（机架），每个 rack 下有多个 host（主机）。PD 在调度时首先会尝试将 Region 的 Peer 放置在不同的 zone，假如无法满足（比如配置 3 副本但总共只有 2 个 zone）

则保证放置在不同的 rack；假如 rack 的数量也不足以保证隔离，那么再尝试 host 级别的隔离，以此类推。

`IsolationLevel` 的意义和作用详细请参考配置集群拓扑。例如已配置 `LocationLabels` 为 `[zone,rack,host]` 的前提下，设置 `IsolationLevel` 为 `zone`，则 PD 在调度时会保证每个 Region 的所有 Peer 均被放置在不同的 zone。假如无法满足 `IsolationLevel` 的最小强制隔离级别限制（比如配置 3 副本但总共只有 2 个 zone），PD 也不会尝试补足，以满足该限制。`IsolationLevel` 默认值为空字符串，即禁用状态。

规则分组字段

以下是规则分组字段的含义：

字段名	类型及约束	说明
`ID`	`string`	分组 ID，用于标识规则来源
`Index`	`int`	不同分组的堆叠次序
`Override`	`true/false`	是否覆盖 index 更小的分组

如果不单独设置规则分组，默认 `Override=false`，对应的行为是不同分组之间相互不影响，不同分组内的规则是同时生效的。

配置规则操作步骤

本节的操作步骤以使用 pd-ctl 工具为例，涉及到的命令也支持通过 HTTP API 进行调用。

开启 Placement Rules 特性

Placement Rules 特性在 TiDB v5.0 及以上的版本中默认开启。如需关闭 Placement Rules 特性，请参考关闭 Placement Rules。如需在关闭后重新开启该特性，可以集群初始化以前设置 PD 配置文件：

```
[replication]
enable-placement-rules = true
```

这样，PD 在初始化成功后会开启这个特性，并根据 `max-replicas` 及 `location-labels` 配置生成对应的规则：

```
{
  "group_id": "pd",
  "id": "default",
  "start_key": "",
  "end_key": "",
  "role": "voter",
  "count": 3,
  "location_labels": ["zone", "rack", "host"],
  "isolation_level": ""
}
```

如果是已经初始化过的集群，也可以通过 pd-ctl 进行在线开启：

```
pd-ctl config placement-rules enable
```

PD 同样将根据系统的 `max-replicas` 及 `location-labels` 生成默认的规则。

> **注意：**
>
> 开启 Placement Rules 后，原先的 `max-replicas` 及 `location-labels` 配置项将不再生效。如果需要调整副本策略，应当使用 Placement Rules 相关接口。

关闭 Placement Rules 特性

使用 pd-ctl 可以关闭 Placement Rules 特性，切换为之前的调度策略。

```
pd-ctl config placement-rules disable
```

> **注意：**
>
> 关闭 Placement Rules 后，PD 将使用原先的 `max-replicas` 及 `location-labels` 配置。在 Placement Rules 开启期间对 Rule 的修改不会导致这两项配置的同步更新。此外，设置好的所有 Rule 都会保留在系统中，会在下次开启 Placement Rules 时被使用。

使用 pd-ctl 设置规则

> **注意：**
>
> 规则的变更将实时地影响 PD 调度，不恰当的规则设置可能导致副本数较少，影响系统的高可用。

pd-ctl 支持使用多种方式查看系统中的 Rule，输出是 json 格式的 Rule 或 Rule 列表：

- 查看所有规则列表

```
pd-ctl config placement-rules show
```

- 查看 PD Group 的所有规则列表

```
pd-ctl config placement-rules show --group=pd
```

- 查看对应 Group 和 ID 的某条规则

```
pd-ctl config placement-rules show --group=pd --id=default
```

- 查看 Region 所匹配的规则列表

```
pd-ctl config placement-rules show --region=2
```

上面的例子中 2 为 Region ID。

新增和编辑规则是类似的，需要把对应的规则写进文件，然后使用 save 命令保存至 PD：

```
cat > rules.json <<EOF
[
    {
        "group_id": "pd",
        "id": "rule1",
        "role": "voter",
        "count": 3,
        "location_labels": ["zone", "rack", "host"]
    },
    {
        "group_id": "pd",
        "id": "rule2",
        "role": "voter",
        "count": 2,
        "location_labels": ["zone", "rack", "host"]
    }
]
EOF
pd-ctl config placement save --in=rules.json
```

以上操作会将 rule1、rule2 两条规则写入 PD，如果系统中已经存在 GroupID+ID 相同的规则，则会覆盖该规则。

如果需要删除某条规则，只需要将规则的 count 置为 0 即可，对应 GroupID+ID 相同的规则会被删除。以下命令将删除 pd/rule2 这条规则：

```
cat > rules.json <<EOF
[
    {
        "group_id": "pd",
        "id": "rule2"
    }
]
EOF
pd-ctl config placement save --in=rules.json
```

使用 pd-ctl 设置规则分组

- 查看所有的规则分组列表

```
pd-ctl config placement-rules rule-group show
```

- 查看指定 ID 的规则分组

```
pd-ctl config placement-rules rule-group show pd
```

- 设置规则分组的 index 和 override 属性

```
pd-ctl config placement-rules rule-group set pd 100 true
```

- 删除规则分组配置（如组内还有规则，则使用默认分组配置）

```
pd-ctl config placement-rules rule-group delete pd
```

使用 pd-ctl 批量更新分组及组内规则

使用 `rule-bundle` 子命令，可以方便地同时查看和修改规则分组及组内的所有规则。

该子命令中 `get {group_id}` 用来查询一个分组，输出结果为嵌套形式的规则分组和组内规则：

```
pd-ctl config placement-rules rule-bundle get pd
```

输出示例：

```
{
  "group_id": "pd",
  "group_index": 0,
  "group_override": false,
  "rules": [
    {
      "group_id": "pd",
      "id": "default",
      "start_key": "",
      "end_key": "",
      "role": "voter",
      "count": 3
    }
  ]
}
```

`rule-bundle get` 子命令中可以添加 `-out` 参数来将输出写入文件，方便后续修改保存。

```
pd-ctl config placement-rules rule-bundle get pd -out="group.json"
```

修改完成后，使用 `rule-bundle set` 子命令将文件中的配置保存至 PD 服务器。与前面介绍的 `save` 不同，此命令会替换服务器端该分组内的所有规则。

```
pd-ctl config placement-rules rule-bundle set pd -in="group.json"
```

使用 pd-ctl 查看和修改所有配置

用户还可以使用 pd-ctl 查看和修改所有配置，即把全部配置保存至文件，修改后再覆盖保存。该操作同样使用 `rule-bundle` 子命令。

下面的命令将所有配置保存至 `rules.json` 文件：

```
pd-ctl config placement-rules rule-bundle load --out="rules.json"
```

编辑完文件后，使用下面的命令将配置保存至 PD 服务器：

```
pd-ctl config placement-rules rule-bundle save --in="rules.json"
```

使用 tidb-ctl 查询表相关的 key range

若需要针对元数据或某个特定的表进行特殊配置，可以通过 tidb-ctl 的 keyrange 命令来查询相关的 key。注意要添加 --encode 返回 PD 中的表示形式。

```
tidb-ctl keyrange --database test --table ttt --encode
```

```
global ranges:
  meta: (6d00000000000000f8, 6e00000000000000f8)
  table: (7400000000000000f8, 7500000000000000f8)
table ttt ranges: (NOTE: key range might be changed after DDL)
  table: (7480000000000000ff2d00000000000000f8, 7480000000000000ff2e00000000000000f8)
  table indexes: (7480000000000000ff2d5f690000000000fa, 7480000000000000ff2d5f720000000000fa)
    index c2: (7480000000000000ff2d5f698000000000ff0000010000000000fa, 7480000000000000
        ↪ ff2d5f698000000000ff0000020000000000fa)
    index c3: (7480000000000000ff2d5f698000000000ff0000020000000000fa, 7480000000000000
        ↪ ff2d5f698000000000ff0000030000000000fa)
    index c4: (7480000000000000ff2d5f698000000000ff0000030000000000fa, 7480000000000000
        ↪ ff2d5f698000000000ff0000040000000000fa)
  table rows: (7480000000000000ff2d5f720000000000fa, 7480000000000000ff2e00000000000000f8)
```

> **注意：**
>
> DDL 等操作会导致 table ID 发生变化，需要同步更新对应的规则。

典型场景示例

本部分介绍 Placement Rules 的使用场景示例。

场景一：普通的表使用 3 副本，元数据使用 5 副本提升集群容灾能力

只需要增加一条规则，将 key range 限定在 meta 数据的范围，并把 count 值设为 5。添加规则示例如下：

```
{
  "group_id": "pd",
  "id": "meta",
  "index": 1,
  "override": true,
  "start_key": "6d00000000000000f8",
  "end_key": "6e00000000000000f8",
```

```
  "role": "voter",
  "count": 5,
  "location_labels": ["zone", "rack", "host"]
}
```

场景二：5 副本按 2-2-1 的比例放置在 3 个数据中心，且第 3 个中心不产生 Leader

创建三条规则，分别设置副本数为 2、2、1，并且在每个规则内通过 `label_constraints` 将副本限定在对应的数据中心内。另外，不需要 leader 的数据中心将 `role` 改为 `follower`。

```
[
    {
        "group_id": "pd",
        "id": "zone1",
        "start_key": "",
        "end_key": "",
        "role": "voter",
        "count": 2,
        "label_constraints": [
            {"key": "zone", "op": "in", "values": ["zone1"]}
        ],
        "location_labels": ["rack", "host"]
    },
    {
        "group_id": "pd",
        "id": "zone2",
        "start_key": "",
        "end_key": "",
        "role": "voter",
        "count": 2,
        "label_constraints": [
            {"key": "zone", "op": "in", "values": ["zone2"]}
        ],
        "location_labels": ["rack", "host"]
    },
    {
        "group_id": "pd",
        "id": "zone3",
        "start_key": "",
        "end_key": "",
        "role": "follower",
        "count": 1,
        "label_constraints": [
            {"key": "zone", "op": "in", "values": ["zone3"]}
        ],
        "location_labels": ["rack", "host"]
```

```
    }
]
```

场景三：为某张表添加 2 个 TiFlash Learner 副本

为表的 row key 单独添加一条规则，限定数量为 2，并且通过 label_constraints 保证副本产生在 engine=tiflash 的节点。注意这里使用了单独的 group_id，保证这条规则不会与系统中其他来源的规则互相覆盖或产生冲突。

```
{
  "group_id": "tiflash",
  "id": "learner-replica-table-ttt",
  "start_key": "7480000000000000ff2d5f720000000000fa",
  "end_key": "7480000000000000ff2e00000000000000f8",
  "role": "learner",
  "count": 2,
  "label_constraints": [
    {"key": "engine", "op": "in", "values": ["tiflash"]}
  ],
  "location_labels": ["host"]
}
```

场景四：为某张表在有高性能磁盘的北京节点添加 2 个 Follower 副本

这个例子展示了比较复杂的 label_constraints 配置，下面的例子限定了副本放置在 bj1 或 bj2 机房，且磁盘类型不能为 hdd。

```
{
  "group_id": "follower-read",
  "id": "follower-read-table-ttt",
  "start_key": "7480000000000000ff2d00000000000000f8",
  "end_key": "7480000000000000ff2e00000000000000f8",
  "role": "follower",
  "count": 2,
  "label_constraints": [
    {"key": "zone", "op": "in", "values": ["bj1", "bj2"]},
    {"key": "disk", "op": "notIn", "values": ["hdd"]}
  ],
  "location_labels": ["host"]
}
```

场景五：将某张表迁移至 TiFlash 集群

与场景三不同，这个场景不是要在原有配置的基础上增加新副本，而是要强制覆盖一段数据的其它配置，因此需要通过配置规则分组来指定一个足够大的 index 以及设置 override 来覆盖原有规则。

规则：

```
{
```

```
  "group_id": "tiflash-override",
  "id": "learner-replica-table-ttt",
  "start_key": "748000000000000ff2d5f720000000000fa",
  "end_key": "748000000000000ff2e000000000000f8",
  "role": "voter",
  "count": 3,
  "label_constraints": [
    {"key": "engine", "op": "in", "values": ["tiflash"]}
  ],
  "location_labels": ["host"]
}
```

规则分组：

```
{
  "id": "tiflash-override",
  "index": 1024,
  "override": true,
}
```

10.2.1.2.2 Load Base Split

Load Base Split 是 TiKV 在 4.0 版本引入的特性，旨在解决 Region 访问分布不均匀造成的热点问题，比如小表的全表扫描。

场景描述

在 TiDB 中，当流量集中在某些节点时很容易形成热点。PD 会尝试通过调度 Hot Region，尽可能让这些 Hot Region 均匀分布在各个节点上，以求获得更好的性能。

但是 PD 的调度的最小粒度是 Region。如果集群的热点数目少于节点数目，或者说存在某几个热点流量远高于其他 Region，对 PD 的热点调度来说，能做到的也只是让热点从一个节点转移到另一个节点，而无法让整个集群承担负载。

这种场景在读请求居多的 workload 中尤为常见。例如对小表的全表扫描和索引查找，或者是对一些字段的频繁访问。

在此之前解决此类问题的办法是手动输入命令去拆分一个或几个热点 Region，但是这样的操作存在以下两个问题：

- 均匀拆分 Region 并不一定是最好的选择，请求可能集中在某几个 Key 上，即使均匀拆分后热点可能仍然集中在其中一个 Region 上，可能需要经过多次均匀拆分才能达到目标。
- 人工介入不够及时和易用。

实现原理

Load Base Split 会基于统计信息自动拆分 Region。通过统计去识别出那些读流量在 10s 内持续超过阈值的 Region，并在合适的位置将这些 Region 拆分。在选择拆分的位置时，会尽可能平衡拆分后两个 Region 的访问量，并尽量避免跨 Region 的访问。

Load Base Split 后的 Region 不会被迅速 Merge。一方面，PD 的 MergeChecker 会跳过 hot Region，另一方面 PD 也会针对心跳信息中的 QPS去进行判断，避免 Merge 两个 QPS 很高的 Region。

使用方法

目前的 Load Base Split 的控制参数为 split.qps-threshold（QPS 阈值）和 split.byte-threshold（流量阈值）。如果连续 10s 内，某个 Region 每秒的各类读请求之和都超过 QPS 阈值或流量阈值，那么就对此 Region 进行拆分。

目前默认开启 Load Base Split，但配置相对保守，split.qps-threshold 默认为 3000，split.byte-threshold 默认为 30MB/s。如果想要关闭这个功能，将这两个阈值全部都调到足够高即可。

目前有两种办法修改配置：

- 通过 SQL 语句修改，例如：

```
set config tikv split.qps-threshold=3000
```

- 通过 TiKV 修改，例如：

```
curl -X POST "http://ip:status_port/config" -H "accept: application/json" -d '{"split.qps-
    ↪ threshold":"3000"}'
```

同理，目前也有两种办法查看配置：

- 通过 SQL 查看，例如：

```
show config where type='tikv' and name like '%split.qps-threshold%'
```

- 通过 TiKV 查看，例如：

```
curl "http://ip:status_port/config"
```

> 注意：
>
> 从 v4.0.0-rc.2 起可以使用 SQL 语句来修改和查看配置。

10.2.1.2.3 Store Limit

Store Limit 是 PD 在 3.0 版本引入的特性，旨在能够更加细粒度地控制调度的速度，针对不同调度场景进行调优。

实现原理

PD 的调度是以 operator 为单位执行的。一个 operator 可能包含多个调度操作。示例如下；

```
"replace-down-replica {mv peer: store [2] to [3]} (kind:region,replica, region:10(4,5), createAt
    ↪ :2020-05-18 06:40:25.775636418 +0000 UTC m=+2168762.679540369, startAt:2020-05-18
    ↪ 06:40:25.775684648 +0000 UTC m=+2168762.679588599, currentStep:0, steps:[add learner peer
    ↪  20 on store 3, promote learner peer 20 on store 3 to voter, remove peer on store 2])"
```

以上示例中，`replace-down-replica` 这个 operator 具体包含以下操作：

1. 在 `store 3` 上添加一个 learner peer，ID 为 20。
2. 将 `store 3` 上 ID 为 20 的 learner peer 提升为 voter。
3. 删除 `store 2` 上的 peer。

Store Limit 是通过在内存中维护了一个 store ID 到令牌桶的映射，来实现 store 级别的限速。这里不同的操作对应不同的令牌桶，目前仅支持限制添加 learner/peer 和删除 peer 两种操作的速度，即对应于每个 store 存在两种类型的令牌桶。

每次 operator 产生后会检查所包含的操作对应的令牌桶中是否有足够的 token。如果 token 充足才会将该 operator 加入到调度的队列中，同时从令牌桶中拿走对应的 token，否则该 operator 被丢弃。令牌桶会按照固定的速率补充 token，从而实现限速的目的。

Store Limit 与 PD 其他 limit 相关的参数（如 `region-schedule-limit`，`leader-schedule-limit` 等）不同的是，Store Limit 限制的主要是 operator 的消费速度，而其他的 limit 主要是限制 operator 的产生速度。引入 Store Limit 特性之前，调度的限速主要是全局的，所以即使限制了全局的速度，但还是有可能存在调度都集中在部分 store 上面，因而影响集群的性能。而 Store Limit 通过将限速的粒度进一步细化，可以更好的控制调度的行为。

使用方法

Store Limit 相关的参数可以通过 `pd-ctl` 进行设置。

查看当前 store 的 limit 设置

查看当前 store 的 limit 示例如下：

```
store limit                          // 显示所有 store 添加和删除 peer 的速度上限。
store limit add-peer                 // 显示所有 store 添加 peer 的速度上限。
store limit remove-peer              // 显示所有 store 删除 peer 的速度上限。
```

设置全部 store 的 limit

设置全部 store 的 limit 示例如下：

```
store limit all 5                    // 设置所有 store 添加和删除 peer 的速度上限为每分钟 5 个。
store limit all 5 add-peer           // 设置所有 store 添加 peer 的速度上限为每分钟 5 个。
store limit all 5 remove-peer        // 设置所有 store 删除 peer 的速度上限为每分钟 5 个。
```

设置单个 store 的 limit

设置单个 store 的 limit 示例如下：

```
store limit 1 5                      // 设置 store 1 添加和删除 peer 的速度上限为每分钟 5 个。
store limit 1 5 add-peer             // 设置 store 1 添加 peer 的速度上限为每分钟 5 个。
store limit 1 5 remove-peer          // 设置 store 1 删除 peer 的速度上限为每分钟 5 个。
```

10.2.1.3 TiKV 线程池性能调优

本文主要介绍 TiKV 线程池性能调优的主要手段，以及 TiKV 内部线程池的主要用途。

10.2.1.3.1 线程池介绍

在 TiKV 中，线程池主要由 gRPC、Scheduler、UnifyReadPool、Raftstore、StoreWriter、Apply、RocksDB 以及其它一些占用 CPU 不多的定时任务与检测组件组成，这里主要介绍几个占用 CPU 比较多且会对用户读写请求的性能产生影响的线程池。

- gRPC 线程池：负责处理所有网络请求，它会把不同任务类型的请求转发给不同的线程池。
- Scheduler 线程池：负责检测写事务冲突，把事务的两阶段提交、悲观锁上锁、事务回滚等请求转化为 key-value 对数组，然后交给 Raftstore 线程进行 Raft 日志复制。
- Raftstore 线程池：
  - 处理所有的 Raft 消息以及添加新日志的提议 (Propose)。
  - 处理 Raft 日志。如果`store-io-pool-size`配置项的值为 0，Raftstore 线程将日志写入到磁盘；如果该值不为 0，Raftstore 线程将日志发送给 StoreWriter 线程处理。
  - 当日志在多数副本中达成一致后，Raftstore 线程把该日志发送给 Apply 线程处理。
- StoreWriter 线程池：负责将所有 Raft 日志写入到磁盘，再把结果返回到 Raftstore 线程。
- Apply 线程池：当收到从 Raftstore 线程池发来的已提交日志后，负责将其解析为 key-value 请求，然后写入 RocksDB 并且调用回调函数通知 gRPC 线程池中的写请求完成，返回结果给客户端。
- RocksDB 线程池：RocksDB 进行 Compact 和 Flush 任务的线程池，关于 RocksDB 的架构与 Compact 操作请参考 RocksDB: A Persistent Key-Value Store for Flash and RAM Storage。
- UnifyReadPool 线程池：由 Coprocessor 线程池与 Storage Read Pool 合并而来，所有的读取请求包括 kv get、 kv batch get、 raw kv get、 coprocessor 等都会在这个线程池中执行。

10.2.1.3.2 TiKV 的只读请求

TiKV 的读取请求分为两类：

- 一类是指定查询某一行或者某几行的简单查询，这类查询会运行在 Storage Read Pool 中。
- 另一类是复杂的聚合计算、范围查询，这类请求会运行在 Coprocessor Read Pool 中。

从 TiKV 5.0 版本起，默认所有的读取请求都通过统一的线程池进行查询。如果是从 TiKV 4.0 升级上来的 TiKV 集群且升级前未打开 `readpool.storage` 的 `use-unified-pool` 配置，则升级后所有的读取请求仍然继续使用独立的线程池进行查询，可以将 `readpool.storage.use-unified-pool` 设置为 `true` 使所有的读取请求通过统一的线程池进行查询。

10.2.1.3.3 TiKV 线程池调优

- gRPC 线程池的大小默认配置 (`server.grpc-concurrency`) 是 5。由于 gRPC 线程池几乎不会有多少计算开销，它主要负责网络 IO、反序列化请求，因此该配置通常不需要调整。
  - 如果部署的机器 CPU 核数特别少（小于等于 8 ），可以考虑将该配置 (`server.grpc-concurrency`) 设置为 2。
  - 如果机器配置很高，并且 TiKV 承担了非常大量的读写请求，观察到 Grafana 上的监控 Thread CPU 的 gRPC poll CPU 的数值超过了 server.grpc-concurrency 大小的 80%，那么可以考虑适当调大 `server.grpc-concurrency` 以控制该线程池使用率在 80% 以下（即 Grafana 上的指标低于 `80% * server.grpc-concurrency` 的值 ）。

- Scheduler 线程池的大小配置 (`storage.scheduler-worker-pool-size`) 在 TiKV 检测到机器 CPU 核数大于等于 16 时默认为 8，小于 16 时默认为 4。它主要用于将复杂的事务请求转化为简单的 key-value 读写。但是 scheduler 线程池本身不进行任何写操作。

  - 如果检测到有事务冲突，那么它会提前返回冲突结果给客户端。
  - 如果未检测到事务冲突，那么它会把需要写入的 key-value 合并成一条 Raft 日志交给 Raftstore 线程进行 Raft 日志复制。

  通常来说为了避免过多的线程切换，最好确保 scheduler 线程池的利用率保持在 50% ~ 75% 之间。(如果线程池大小为 8 的话，那么 Grafana 上的 TiKV-Details.Thread CPU.scheduler worker CPU 应当在 400% ~ 600% 之间较为合理)

- Raftstore 线程池是 TiKV 中最复杂的一个线程池，默认大小 (由 `raftstore.store-pool-size` 控制) 为 2。StoreWriter 线程池的默认大小 (由 `raftstore.store-io-pool-size` 控制) 为 0。

  - 当 StoreWriter 线程池大小为 0 时，所有的写请求都会由 Raftstore 线程以 fsync 的方式写入 RocksDB。此时建议采取如下调优操作：
    * 将 Raftstore 线程的整体 CPU 使用率控制在 60% 以下。当把 Raftstore 线程数设为默认值 2 时，将 Grafana 监控上 TiKV-Details、Thread CPU、Raft store CPU 面版上的数值控制在 120% 以内。由于存在 I/O 请求，理论上 Raftstore 线程的 CPU 使用率总是低于 100%。
    * 不建议为了提升写性能而盲目增大 Raftstore 线程池大小，这样可能会适得其反，增加磁盘负担，导致性能变差。
  - 当 StoreWriter 线程池大小不为 0 时，所有写请求都由 StoreWriter 线程以 fsync 的方式写入 RocksDB。此时建议采取如下调优操作：
    * 仅在整体 CPU 资源比较充裕的情况下启用 StoreWriter 线程池，并将 StoreWriter 线程和 Raftstore 线程的 CPU 使用率控制在 80% 以下。
      与写请求在 Raftstore 线程完成的情况相比，理论上 StoreWriter 线程处理写请求能够显著地降低写延迟和读的尾延迟。然而，写入速度变得更快意味着 Raft 日志也变得更多，从而导致 Raftstore 线程、Apply 线程和 gRPC 线程的 CPU 开销增多。在这种情况下，CPU 资源不足可能会抵消优化效果，反而还可能比原来的写速度更慢，因此若是 CPU 资源不充裕则不建议开启 StoreWriter 线程。由于 Raftstore 线程把绝大部分的 I/O 请求交给 StoreWriter，因此 Raftstore 线程的 CPU 使用率控制在 80% 以下即可。
    * 大多数情况下将 StoreWriter 线程池的大小设为 1 或 2 即可。这是因为 StoreWriter 线程池的大小会影响 Raft 日志数量，所以该值不宜过大。如果 CPU 使用率高于 80%，可以考虑再增加其大小。
    * 注意 Raft 日志增多对其他线程池 CPU 开销的影响，必要的时候需要相应地增加 Raftstore 线程、Apply 线程和 gRPC 线程的数量。

- UnifyReadPool 负责处理所有的读取请求。默认配置 (`readpool.unified.max-thread-count`) 大小为机器 CPU 数的 80% (如机器为 16 核，则默认线程池大小为 12)。

  通常建议根据业务负载特性调整其 CPU 使用率在线程池大小的 60% ~ 90% 之间 (如果用户 Grafana 上 TiKV-Details.Thread CPU.Unified read pool CPU 的峰值不超过 800%，那么建议用户将 `readpool.unified.max-thread-count` 设置为 10，过多的线程数会造成更频繁的线程切换，并且抢占其他线程池的资源)。

- RocksDB 线程池是 RocksDB 进行 Compact 和 Flush 任务的线程池，通常不需要配置。

- 如果机器 CPU 核数较少，可将 `rocksdb.max-background-jobs` 与 `raftdb.max-background-jobs` 同时设置为 4。
- 如果遇到了 Write Stall，可查看 Grafana 监控上 RocksDB-kv 中的 Write Stall Reason 有哪些指标不为 0。
  * 如果是由 pending compaction bytes 相关原因引起的，可将 `rocksdb.max-sub-compactions` 设置为 2 或者 3（该配置表示单次 compaction job 允许使用的子线程数量，TiKV 4.0 版本默认值为 3，3.0 版本默认值为 1）。
  * 如果原因是 memtable count 相关，建议调大所有列的 `max-write-buffer-number`（默认为 5）。
  * 如果原因是 level0 file limit 相关，建议调大如下参数为 64 或者更高：

```
rocksdb.defaultcf.level0-slowdown-writes-trigger
rocksdb.writecf.level0-slowdown-writes-trigger
rocksdb.lockcf.level0-slowdown-writes-trigger
rocksdb.defaultcf.level0-stop-writes-trigger
rocksdb.writecf.level0-stop-writes-trigger
rocksdb.lockcf.level0-stop-writes-trigger
```

10.2.1.4 TiKV 内存参数性能调优

本文档用于描述如何根据机器配置情况来调整 TiKV 的参数，使 TiKV 的性能达到最优。

TiKV 最底层使用的是 RocksDB 做为持久化存储，所以 TiKV 的很多性能相关的参数都是与 RocksDB 相关的。TiKV 使用了两个 RocksDB 实例，默认 RocksDB 实例存储 KV 数据，Raft RocksDB 实例（简称 RaftDB）存储 Raft 数据。

TiKV 使用了 RocksDB 的 `Column Families` (CF) 特性。

- 默认 RocksDB 实例将 KV 数据存储在内部的 `default`、`write` 和 `lock` 3 个 CF 内。
  - `default` CF 存储的是真正的数据，与其对应的参数位于 `[rocksdb.defaultcf]` 项中；
  - `write` CF 存储的是数据的版本信息 (MVCC) 以及索引相关的数据，相关的参数位于 `[rocksdb.writecf]` 项中；
  - `lock` CF 存储的是锁信息，系统使用默认参数。
- Raft RocksDB 实例存储 Raft log。
  - `default` CF 主要存储的是 Raft log，与其对应的参数位于 `[raftdb.defaultcf]` 项中。

所有的 CF 默认共同使用一个 block cache 实例。通过在 `[storage.block-cache]` 下设置 `capacity` 参数，你可以配置该 block cache 的大小。block cache 越大，能够缓存的热点数据越多，读取数据越容易，同时占用的系统内存也越多。如果要为每个 CF 使用单独的 block cache 实例，需要在 `[storage.block-cache]` 下设置 `shared=false`，并为每个 CF 配置单独的 block cache 大小。例如，可以在 `[rocksdb.writecf]` 下设置 `block-cache-size` 参数来配置 `write` CF 的大小。

> **注意：**
>
> 在 TiKV 3.0 之前的版本中，不支持使用 `shared block cache`，需要为每个 CF 单独配置 block cache。

每个 CF 有各自的 write-buffer，大小通过 write-buffer-size 控制。

10.2.1.4.1 参数说明

```
#### 日志级别，可选值为：trace，debug，warn，error，info，off
log-level = "info"

[server]
#### 监听地址
#### addr = "127.0.0.1:20160"

#### gRPC 线程池大小
#### grpc-concurrency = 4
#### TiKV 每个实例之间的 gRPC 连接数
#### grpc-raft-conn-num = 10

#### TiDB 过来的大部分读请求都会发送到 TiKV 的 Coprocessor 进行处理，该参数用于设置
#### coprocessor 线程的个数，如果业务是读请求比较多，增加 coprocessor 的线程数，但应比系统的
#### CPU 核数小。例如：TiKV 所在的机器有 32 core，在重读的场景下甚至可以将该参数设置为 30。在没有
#### 设置该参数的情况下，TiKV 会自动将该值设置为 CPU 总核数乘以 0.8。
#### end-point-concurrency = 8

#### 可以给 TiKV 实例打标签，用于副本的调度
#### labels = {zone = "cn-east-1", host = "118", disk = "ssd"}

[storage]
#### 数据目录
#### data-dir = "/tmp/tikv/store"

#### 通常情况下使用默认值就可以了。在导数据的情况下建议将该参数设置为 1024000。
#### scheduler-concurrency = 102400
#### 该参数控制写入线程的个数，当写入操作比较频繁的时候，需要把该参数调大。使用 top -H -p tikv-
    ↪ pid
#### 发现名称为 sched-worker-pool 的线程都特别忙，这个时候就需要将 scheduler-worker-pool-size
#### 参数调大，增加写线程的个数。
#### scheduler-worker-pool-size = 4

[storage.block-cache]
##### 是否为 RocksDB 的所有 CF 都创建一个 `shared block cache`。
##### ## RocksDB 使用 block cache 来缓存未压缩的数据块。较大的 block cache 可以加快读取速度。
##### 推荐开启 `shared block cache` 参数。这样只需要设置全部缓存大小，使配置过程更加方便。
##### 在大多数情况下，可以通过 LRU 算法在各 CF 间自动平衡缓存用量。
##### ## `storage.block-cache` 会话中的其余配置仅在开启 `shared block cache` 时起作用。
#### shared = true
##### `shared block cache` 的大小。正常情况下应设置为系统全部内存的 30%-50%。
```

```
##### 如果未设置该参数，则由以下字段或其默认值的总和决定。
##### ##   * rocksdb.defaultcf.block-cache-size 或系统全部内存的 25%
#####   * rocksdb.writecf.block-cache-size 或系统全部内存的 15%
#####   * rocksdb.lockcf.block-cache-size 或系统全部内存的 2%
#####   * raftdb.defaultcf.block-cache-size 或系统全部内存的 2%
##### ## 要在单个物理机上部署多个 TiKV 节点，需要显式配置该参数。
##### 否则，TiKV 中可能会出现 OOM 错误。
#### capacity = "1GB"

[pd]
#### pd 的地址
#### endpoints = ["127.0.0.1:2379","127.0.0.2:2379","127.0.0.3:2379"]

[metric]
#### 将 metrics 推送给 Prometheus pushgateway 的时间间隔
interval = "15s"
#### Prometheus pushgateway 的地址
address = ""
job = "tikv"

[raftstore]
#### Raft RocksDB 目录。默认值是 [storage.data-dir] 的 raft 子目录。
#### 如果机器上有多块磁盘，可以将 Raft RocksDB 的数据放在不同的盘上，提高 TiKV 的性能。
#### raftdb-path = "/tmp/tikv/store/raft"

#### 当 Region 写入的数据量超过该阈值的时候，TiKV 会检查该 Region 是否需要分裂。为了减少检查过程
#### 中扫描数据的成本，导入数据过程中可以将该值设置为 32 MB，正常运行状态下使用默认值即可。
region-split-check-diff = "32MB"

[coprocessor]

##### 当区间为 [a,e) 的 Region 的大小超过 `region_max_size`，TiKV 会尝试分裂该 Region，例如分裂成
    ↪  [a,b)、[b,c)、[c,d)、[d,e) 等区间的 Region 后
##### 这些 Region [a,b), [b,c), [c,d) 的大小为 `region_split_size`（或者稍大于 `region_split_size
    ↪ `）
#### region-max-size = "144MB"
#### region-split-size = "96MB"

[rocksdb]
#### RocksDB 进行后台任务的最大线程数，后台任务包括 compaction 和 flush。具体 RocksDB
    ↪ 为什么需要进行 compaction，
#### 请参考 RocksDB 的相关资料。在写流量比较大的时候（例如导数据），建议开启更多的线程，
#### 但应小于 CPU 的核数。例如在导数据的时候，32 核 CPU 的机器，可以设置成 28。
#### max-background-jobs = 8
```

```
#### RocksDB 能够打开的最大文件句柄数。
#### max-open-files = 40960

#### RocksDB MANIFEST 文件的大小限制.# 更详细的信息请参考：https://github.com/facebook/rocksdb/
    ↪ wiki/MANIFEST
max-manifest-file-size = "20MB"

#### RocksDB write-ahead logs 目录。如果机器上有两块盘，可以将 RocksDB 的数据和 WAL 日志放在
#### 不同的盘上，提高 TiKV 的性能。
#### wal-dir = "/tmp/tikv/store"

#### 下面两个参数用于怎样处理 RocksDB 归档 WAL。
#### 更多详细信息请参考：https://github.com/facebook/rocksdb/wiki/How-to-persist-in-memory-
    ↪ RocksDB-database%3F
#### wal-ttl-seconds = 0
#### wal-size-limit = 0

#### RocksDB WAL 日志的最大总大小，通常情况下使用默认值就可以了。
#### max-total-wal-size = "4GB"

#### 可以通过该参数打开或者关闭 RocksDB 的统计信息。
#### enable-statistics = true

#### 开启 RocksDB compaction 过程中的预读功能，如果使用的是机械磁盘，建议该值至少为2MB。
#### compaction-readahead-size = "2MB"

[rocksdb.defaultcf]
#### 数据块大小。RocksDB 是按照 block 为单元对数据进行压缩的，同时 block 也是缓存在 block-cache
#### 中的最小单元（类似其他数据库的 page 概念）。
block-size = "64KB"

#### RocksDB 每一层数据的压缩方式，可选的值为：no,snappy,zlib,bzip2,lz4,lz4hc,zstd。
#### no:no:lz4:lz4:lz4:zstd:zstd 表示 level0 和 level1 不压缩，level2 到 level4 采用 lz4 压缩算法
    ↪ ，
#### level5 和 level6 采用 zstd 压缩算法,。
#### no 表示没有压缩，lz4 是速度和压缩比较为中庸的压缩算法，zlib 的压缩比很高，对存储空间比较友
#### 好，但是压缩速度比较慢，压缩的时候需要占用较多的 CPU 资源。不同的机器需要根据 CPU 以及 I/O
    ↪ 资
#### 源情况来配置怎样的压缩方式。例如：如果采用的压缩方式为"no:no:lz4:lz4:lz4:zstd:zstd"，在大量
#### 写入数据的情况下（导数据），发现系统的 I/O 压力很大（使用 iostat 发现 %util 持续 100% 或者使
#### 用 top 命令发现 iowait 特别多），而 CPU 的资源还比较充裕，这个时候可以考虑将 level0 和
#### level1 开启压缩，用 CPU 资源换取 I/O 资源。如果采用的压缩方式
#### 为"no:no:lz4:lz4:lz4:zstd:zstd"，在大量写入数据的情况下，发现系统的 I/O 压力不大，但是 CPU
#### 资源已经吃光了，top -H 发现有大量的 bg 开头的线程（RocksDB 的 compaction 线程）在运行，这
#### 个时候可以考虑用 I/O 资源换取 CPU 资源，将压缩方式改成"no:no:no:lz4:lz4:zstd:zstd"。总之，目
```

```
#### 的是为了最大限度地利用系统的现有资源，使 TiKV 的性能在现有的资源情况下充分发挥。
compression-per-level = ["no", "no", "lz4", "lz4", "lz4", "zstd", "zstd"]

#### RocksDB memtable 的大小。
write-buffer-size = "128MB"

#### 最多允许几个 memtable 存在。写入到 RocksDB 的数据首先会记录到 WAL 日志里面，然后会插入到
#### memtable 里面，当 memtable 的大小到达了 write-buffer-size 限定的大小的时候，当前的
#### memtable 会变成只读的，然后生成一个新的 memtable 接收新的写入。只读的 memtable 会被
#### RocksDB 的 flush 线程（max-background-flushes 参数能够控制 flush 线程的最大个数）
#### flush 到磁盘，成为 level0 的一个 sst 文件。当 flush 线程忙不过来，导致等待 flush 到磁盘的
#### memtable 的数量到达 max-write-buffer-number 限定的个数的时候，RocksDB 会将新的写入
#### stall 住，stall 是 RocksDB 的一种流控机制。在导数据的时候可以将 max-write-buffer-number
#### 的值设置的更大一点，例如 10。
max-write-buffer-number = 5

#### 当 level0 的 sst 文件个数到达 level0-slowdown-writes-trigger 指定的限度的时候，
#### RocksDB 会尝试减慢写入的速度。因为 level0 的 sst 太多会导致 RocksDB 的读放大上升。
#### level0-slowdown-writes-trigger 和 level0-stop-writes-trigger 是 RocksDB 进行流控的
#### 另一个表现。当 level0 的 sst 的文件个数到达 4（默认值），level0 的 sst 文件会和 level1 中
#### 有 overlap 的 sst 文件进行 compaction，缓解读放大的问题。
level0-slowdown-writes-trigger = 20

#### 当 level0 的 sst 文件个数到达 level0-stop-writes-trigger 指定的限度的时候，RocksDB 会
#### stall 住新的写入。
level0-stop-writes-trigger = 36

#### 当 level1 的数据量大小达到 max-bytes-for-level-base 限定的值的时候，会触发 level1 的
#### sst 和 level2 种有 overlap 的 sst 进行 compaction。
#### 黄金定律：max-bytes-for-level-base 的设置的第一参考原则就是保证和 level0 的数据量大致相
#### 等，这样能够减少不必要的 compaction。例如压缩方式为"no:no:lz4:lz4:lz4:lz4:lz4"，那么
#### max-bytes-for-level-base 的值应该是 write-buffer-size 的大小乘以 4，因为 level0 和
#### level1 都没有压缩，而且 level0 触发 compaction 的条件是 sst 的个数到达 4（默认值）。在
#### level0 和 level1 都采取了压缩的情况下，就需要分析下 RocksDB 的日志，看一个 memtable 的压
#### 缩成一个 sst 文件的大小大概是多少，例如 32MB，那么 max-bytes-for-level-base 的建议值就应
#### 该是 32MB * 4 = 128MB。
max-bytes-for-level-base = "512MB"

#### sst 文件的大小。level0 的 sst 文件的大小受 write-buffer-size 和 level0 采用的压缩算法的
#### 影响，target-file-size-base 参数用于控制 level1-level6 单个 sst 文件的大小。
target-file-size-base = "32MB"

[rocksdb.writecf]
#### 保持和 rocksdb.defaultcf.compression-per-level 一致。
compression-per-level = ["no", "no", "lz4", "lz4", "lz4", "zstd", "zstd"]
```

```
#### 保持和 rocksdb.defaultcf.write-buffer-size 一致。
write-buffer-size = "128MB"
max-write-buffer-number = 5
min-write-buffer-number-to-merge = 1

#### 保持和 rocksdb.defaultcf.max-bytes-for-level-base 一致。
max-bytes-for-level-base = "512MB"
target-file-size-base = "32MB"

[raftdb]
#### RaftDB 能够打开的最大文件句柄数。
#### max-open-files = 40960

#### 可以通过该参数打开或者关闭 RaftDB 的统计信息。
#### enable-statistics = true

#### 开启 RaftDB compaction 过程中的预读功能，如果使用的是机械磁盘，建议该值至少为2MB。
#### compaction-readahead-size = "2MB"

[raftdb.defaultcf]
#### 保持和 rocksdb.defaultcf.compression-per-level 一致。
compression-per-level = ["no", "no", "lz4", "lz4", "lz4", "zstd", "zstd"]

#### 保持和 rocksdb.defaultcf.write-buffer-size 一致。
write-buffer-size = "128MB"
max-write-buffer-number = 5
min-write-buffer-number-to-merge = 1

#### 保持和 rocksdb.defaultcf.max-bytes-for-level-base 一致。
max-bytes-for-level-base = "512MB"
target-file-size-base = "32MB"
```

10.2.1.4.2 TiKV 内存使用情况

除了以上列出的 block-cache 以及 write-buffer 会占用系统内存外：

1. 需预留一些内存作为系统的 page cache
2. TiKV 在处理大的查询的时候（例如 select * from ...）会读取数据然后在内存中生成对应的数据结构返回给 TiDB，这个过程中 TiKV 会占用一部分内存

10.2.1.4.3 TiKV 机器配置推荐

1. 生产环境中，不建议将 TiKV 部署在 CPU 核数小于 8 或内存低于 32GB 的机器上

2. 如果对写入吞吐要求比较高，建议使用吞吐能力比较好的磁盘
3. 如果对读写的延迟要求非常高，建议使用 IOPS 比较高的 SSD 盘

10.2.1.5 Follower Read

当系统中存在读取热点 Region 导致 leader 资源紧张成为整个系统读取瓶颈时，启用 Follower Read 功能可明显降低 leader 的负担，并且通过在多个 follower 之间均衡负载，显著地提升整体系统的吞吐能力。本文主要介绍 Follower Read 的使用方法与实现机制。

10.2.1.5.1 概述

Follower Read 功能是指在强一致性读的前提下使用 Region 的 follower 副本来承载数据读取的任务，从而提升 TiDB 集群的吞吐能力并降低 leader 负载。Follower Read 包含一系列将 TiKV 读取负载从 Region 的 leader 副本上 offload 到 follower 副本的负载均衡机制。TiKV 的 Follower Read 可以保证数据读取的一致性，可以为用户提供强一致的数据读取能力。

> **注意：**
>
> 为了获得强一致读取的能力，在当前的实现中，follower 节点需要向 leader 节点询问当前的执行进度（即 `ReadIndex`），这会产生一次额外的网络请求开销，因此目前 Follower Read 的主要优势是处理隔离集群的读写请求以及提升整体读取吞吐。

10.2.1.5.2 使用方式

要开启 TiDB 的 Follower Read 功能，将变量 `tidb_replica_read` 的值设置为 `follower` 或 `leader-and-follower` 即可：

```
set [session | global] tidb_replica_read = '<目标值>';
```

作用域：SESSION | GLOBAL

默认值：leader

该变量用于设置期待的数据读取方式。

- 当设定为默认值 `leader` 或者空字符串时，TiDB 会维持原有行为方式，将所有的读取操作都发送给 leader 副本处理。
- 当设置为 `follower` 时，TiDB 会选择 Region 的 follower 副本完成所有的数据读取操作。
- 当 `tidb_replica_read` 的值设为 `leader-and-follower` 时，TiDB 可以选择任意副本来执行读取操作。
- 当设置为 `leader-and-follower` 时，读请求会在 leader 和 follower 之间负载均衡。

10.2.1.5.3 实现机制

在 Follower Read 功能出现之前，TiDB 采用 strong leader 策略将所有的读写操作全部提交到 Region 的 leader 节点上完成。虽然 TiKV 能够很均匀地将 Region 分散到多个物理节点上，但是对于每一个 Region 来说，只有 leader 副

本能够对外提供服务，另外的 follower 除了时刻同步数据准备着 failover 时投票切换成为 leader 外，没有办法对 TiDB 的请求提供任何帮助。

为了允许在 TiKV 的 follower 节点进行数据读取，同时又不破坏线性一致性和 Snapshot Isolation 的事务隔离，Region 的 follower 节点需要使用 Raft ReadIndex 协议确保当前读请求可以读到当前 leader 上已经 commit 的最新数据。在 TiDB 层面，Follower Read 只需根据负载均衡策略将某个 Region 的读取请求发送到 follower 节点。

Follower 强一致读

TiKV follower 节点处理读取请求时，首先使用 Raft ReadIndex 协议与 Region 当前的 leader 进行一次交互，来获取当前 Raft group 最新的 commit index。本地 apply 到所获取的 leader 最新 commit index 后，便可以开始正常的读取请求处理流程。

Follower 副本选择策略

由于 TiKV 的 Follower Read 不会破坏 TiDB 的 Snapshot Isolation 事务隔离级别，因此 TiDB 选择 follower 的策略可以采用 round robin 的方式。目前，对于 Coprocessor 请求，Follower Read 负载均衡策略粒度是连接级别的，对于一个 TiDB 的客户端连接在某个具体的 Region 上会固定使用同一个 follower，只有在选中的 follower 发生故障或者因调度策略发生调整的情况下才会进行切换。而对于非 Coprocessor 请求（点查等），Follower Read 负载均衡策略粒度是事务级别的，对于一个 TiDB 的事务在某个具体的 Region 上会固定使用同一个 follower，同样在 follower 发生故障或者因调度策略发生调整的情况下才会进行切换。

10.2.1.6 TiFlash 性能调优

本文介绍了使 TiFlash 性能达到最优的几种方式，包括规划机器资源、TiDB 参数调优、配置 TiKV Region 大小等。

10.2.1.6.1 资源规划

对于希望节省机器资源，并且完全没有隔离要求的场景，可以使用 TiKV 和 TiFlash 联合部署。建议为 TiKV 与 TiFlash 分别留够资源，并且不要共享磁盘。

10.2.1.6.2 TiDB 相关参数调优

1. 对于 OLAP/TiFlash 专属的 TiDB 节点，建议调大读取并发数tidb_distsql_scan_concurrency 到 80：

```
set @@tidb_distsql_scan_concurrency = 80;
```

2. 开启 Super batch 功能：

   tidb_allow_batch_cop 变量用来设置从 TiFlash 读取时，是否把 Region 的请求进行合并。当查询中涉及的 Region 数量比较大，可以尝试设置该变量为 1（对带 aggregation 下推到 TiFlash Coprocessor 的请求生效），或设置该变量为 2（对全部下推到 TiFlash Coprocessor 请求生效）。

```
set @@tidb_allow_batch_cop = 1;
```

3. 尝试开启聚合推过 Join / Union 等 TiDB 算子的优化：

   tidb_opt_agg_push_down 变量用来设置优化器是否执行聚合函数下推到 Join 之前的优化操作。当查询中聚合操作执行很慢时，可以尝试设置该变量为 1。

```
set @@tidb_opt_agg_push_down = 1;
```

4. 尝试开启 Distinct 推过 Join / Union 等 TiDB 算子的优化：

   tidb_opt_distinct_agg_push_down 变量用来设置优化器是否执行带有 Distinct 的聚合函数（比如 select count(distinct a)from t）下推到 Coprocessor 的优化操作。当查询中带有 Distinct 的聚合操作执行很慢时，可以尝试设置该变量为 1。

```
set @@tidb_opt_distinct_agg_push_down = 1;
```

10.2.1.7 DM 配置优化

本文档介绍如何对迁移任务的配置进行优化，从而提高 DM 的数据迁移性能。

10.2.1.7.1 全量导出

全量导出相关的配置项为 mydumpers，下面介绍和性能相关的参数如何配置。

rows

设置 rows 选项可以开启单表多线程并发导出，值为导出的每个 chunk 包含的最大行数。开启后，DM 会在 MySQL 的单表并发导出时，优先选出一列做拆分基准，选择的优先级为主键 > 唯一索引 > 普通索引，选出目标列后需保证该列为整数类型（如 INT、MEDIUMINT、BIGINT 等）。

rows 的值可以设置为 10000，具体设置的值可以根据表中包含数据的总行数以及数据库的性能做调整。另外也需要设置 threads 来控制并发线程数量，默认值为 4，可以适当做些调整。

chunk-filesize

DM 全量备份时会根据 chunk-filesize 参数的值把每个表的数据划分成多个 chunk，每个 chunk 保存到一个文件中，大小约为 chunk-filesize。根据这个参数把数据切分到多个文件中，这样就可以利用 DM Load 处理单元的并行处理逻辑提高导入速度。该参数默认值为 64（单位为 MB），正常情况下不需要设置，也可以根据全量数据的大小做适当的调整。

> 注意：
>
> - mydumpers 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
> - mydumpers.threads 可以使用配置项 mydumper-thread 替代来简化配置。
> - 如果设置了 rows，DM 会忽略 chunk-filesize 的值。

10.2.1.7.2 全量导入

全量导入相关的配置项为 loaders，下面介绍和性能相关的参数如何配置。

pool-size

pool-size 为 DM Load 阶段线程数量的设置，默认值为 16，正常情况下不需要设置，也可以根据全量数据的大小以及数据库的性能做适当的调整。

注意：

- loaders 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
- loaders.pool-size 可以使用配置项 loader-thread 替代来简化配置。

10.2.1.7.3 增量复制

增量复制相关的配置为 syncers，下面介绍和性能相关的参数如何配置。

worker-count

worker-count 为 DM Sync 阶段并发迁移 DML 的线程数量设置，默认值为 16，如果对迁移速度有较高的要求，可以适当调高改参数的值。

batch

batch 为 DM Sync 阶段迁移数据到下游数据库时，每个事务包含的 DML 的数量，默认值为 100，正常情况下不需要调整。

注意：

- syncers 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
- syncers.worker-count 可以使用配置项 syncer-thread 替代来简化配置。
- worker-count 和 batch 的设置需要根据实际的场景进行调整，例如：DM 到下游数据库的网络延迟较高，可以适当调高 worker-count，调低 batch。

10.2.2 下推计算结果缓存

TiDB 从 4.0 起支持下推计算结果缓存（即 Coprocessor Cache 功能）。开启该功能后，将在 TiDB 实例侧缓存下推给 TiKV 计算的结果，在部分场景下起到加速效果。

10.2.2.1 配置

Coprocessor Cache 的配置均位于 TiDB 的 tikv-client.copr-cache 配置项中。Coprocessor 的具体开启和配置方法，见 TiDB 配置文件描述。

10.2.2.2 特性说明

- 所有 SQL 在单个 TiDB 实例上的首次执行都不会被缓存。
- 缓存仅存储在 TiDB 内存中，TiDB 重启后缓存会失效。
- 不同 TiDB 实例之间不共享缓存。
- 缓存的是下推计算结果，即使缓存命中，后续仍有 TiDB 计算。
- 缓存以 Region 为单位。对 Region 的写入会导致涉及该 Region 的缓存失效。基于此原因，该功能主要会对很少变更的数据有效果。
- 下推计算请求相同时，缓存会被命中。通常在以下场景下，下推计算的请求是相同或部分相同的：
  - SQL 语句完全一致，例如重复执行相同的 SQL 语句。
    该场景下所有下推计算的请求都是一致的，所有请求都能利用上下推计算缓存。
  - SQL 语句包含一个变化的条件，其他部分一致，变化的条件是表主键或分区主键。
    该场景下一部分下推计算的请求会与之前出现过的一致，部分请求能利用上下推计算结果缓存。
  - SQL 语句包含多个变化的条件，其他部分一致，变化的条件完全匹配一个复合索引列。
    该场景下一部分下推计算的请求会与之前出现过的一致，部分请求能利用上下推计算结果缓存。
- 该功能对用户透明，开启和关闭都不影响计算结果，仅影响 SQL 执行时间。

10.2.2.3 检验缓存效果

可以通过执行 EXPLAIN ANALYZE 或查看 Grafana 监控面板来检查 Coprocessor 的缓存效果。

10.2.2.3.1 使用 EXPLAIN ANALYZE

用户可以通过 EXPLAIN ANALYZE 语句查看读表算子中的缓存命中率，示例如下：

```
EXPLAIN ANALYZE SELECT * FROM t USE INDEX(a);
+--
    ↪ ----------------------------+-----------+---------+-----------+-------------------------+---------------
    ↪
| id                            | estRows   | actRows | task      | access object           |
    ↪ execution info
    ↪
    ↪ | operator info                  | memory               | disk |
+--
    ↪ ----------------------------+-----------+---------+-----------+-------------------------+---------------
    ↪
| IndexLookUp_6                 | 262400.00 | 262400  | root      |                         | time
    ↪ :620.513742ms, loops:258, cop_task: {num: 4, max: 5.530817ms, min: 1.51829ms, avg:
    ↪ 2.70883ms, p95: 5.530817ms, max_proc_keys: 2480, p95_proc_keys: 2480, tot_proc: 1ms,
    ↪ tot_wait: 1ms, rpc_num: 4, rpc_time: 10.816328ms, copr_cache_hit_rate: 0.75} |
    ↪                                  | 6.685169219970703 MB  | N/A  |
| ├─IndexFullScan_4(Build)      | 262400.00 | 262400  | cop[tikv] | table:t, index:a(a, c) |
    ↪ proc max:93ms, min:1ms, p80:93ms, p95:93ms, iters:275, tasks:4
    ↪
    ↪ | keep order:false, stats:pseudo | 1.7549400329589844 MB | N/A  |
```

```
| └─TableRowIDScan_5(Probe)     | 262400.00 | 0       | cop[tikv] | table:t               |
    ↪ time:0ns, loops:0
    ↪
    ↪ | keep order:false, stats:pseudo | N/A                   | N/A  |
+--
    ↪ ----------------------------+----------+---------+----------+-----------------------+---------------
    ↪
3 rows in set (0.62 sec)
```

执行结果的 execution info 列有 copr_cache_hit_ratio 信息，表示下推计算结果缓存的命中率。以上示例的 0.75 表示命中率大约是 75%。

10.2.2.3.2 查看 Grafana 监控面板

在 Grafana 监控中，tidb 命名空间下 distsql 子系统中可见 copr-cache 面板，该面板监控整个集群中下推计算结果缓存的命中次数、未命中次数和缓存丢弃次数。

10.3 SQL 性能调优

10.3.1 SQL 性能调优

SQL 是一种声明性语言。一条 SQL 语句描述的是最终结果应该如何，而非按顺序执行的步骤。TiDB 会优化 SQL 语句的执行，语义上允许以任何顺序执行查询的各部分，前提是能正确返回语句所描述的最终结果。

SQL 性能优化的过程，可以理解为 GPS 导航的过程。你提供地址后，GPS 软件利用各种统计信息（例如以前的行程、速度限制等元数据，以及实时交通信息）规划出一条最省时的路线。这与 TiDB 中的 SQL 性能优化过程相对应。

本章节包括以下文档，可帮助你更好地理解查询执行计划：

- 理解 TiDB 执行计划介绍如何使用 EXPLAIN 语句来理解 TiDB 是如何执行某个查询的。
- SQL 优化流程概览介绍 TiDB 可以使用的几种优化，以提高查询性能。
- 控制执行计划介绍如何控制执行计划的生成。TiDB 的执行计划非最优时，建议控制执行计划。

10.3.2 理解 TiDB 执行计划

10.3.2.1 EXPLAIN 概览

> **注意：**
>
> 使用 MySQL 客户端连接到 TiDB 时，为避免输出结果在终端中换行，可先执行 pager less -S 命令。执行命令后，新的 EXPLAIN 的输出结果不再换行，可按右箭头 → 键水平滚动阅读输出结果。

使用 EXPLAIN 可查看 TiDB 执行某条语句时选用的执行计划。也就是说，TiDB 在考虑上数百或数千种可能的执行计划后，最终认定该执行计划消耗的资源最少、执行的速度最快。

EXPLAIN 示例如下：

```
CREATE TABLE t (id INT NOT NULL PRIMARY KEY auto_increment, a INT NOT NULL, pad1 VARCHAR(255),
    ↪ INDEX(a));
INSERT INTO t VALUES (1, 1, 'aaa'),(2,2, 'bbb');
EXPLAIN SELECT * FROM t WHERE a = 1;
```

返回的结果如下：

```
Query OK, 0 rows affected (0.96 sec)

Query OK, 2 rows affected (0.02 sec)
Records: 2  Duplicates: 0  Warnings: 0

+--
    ↪ ----------------------------+---------+-----------+---------------------+-------------------------------
    ↪
| id                            | estRows | task      | access object       | operator info
    ↪                                 |
+--
    ↪ ----------------------------+---------+-----------+---------------------+-------------------------------
    ↪
| IndexLookUp_10                | 10.00   | root      |                     |
    ↪                                                 |
| ├─IndexRangeScan_8(Build)     | 10.00   | cop[tikv] | table:t, index:a(a) | range:[1,1], keep
    ↪ order:false, stats:pseudo |
| └─TableRowIDScan_9(Probe)     | 10.00   | cop[tikv] | table:t             | keep order:false,
    ↪ stats:pseudo              |
+--
    ↪ ----------------------------+---------+-----------+---------------------+-------------------------------
    ↪
3 rows in set (0.00 sec)
```

EXPLAIN 实际不会执行查询。EXPLAIN ANALYZE 可用于实际执行查询并显示执行计划。如果 TiDB 所选的执行计划非最优，可用 EXPLAIN 或 EXPLAIN ANALYZE 来进行诊断。有关 EXPLAIN 用法的详细内容，参阅以下文档：

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.1.1 解读 EXPLAIN 的返回结果

EXPLAIN 的返回结果包含以下字段：

- id 为算子名，或执行 SQL 语句需要执行的子任务。详见算子简介。
- estRows 为显示 TiDB 预计会处理的行数。该预估数可能基于字典信息（例如访问方法基于主键或唯一键），或基于 CMSketch 或直方图等统计信息估算而来。
- task 显示算子在执行语句时的所在位置。详见Task 简介。
- access-object 显示被访问的表、分区和索引。显示的索引为部分索引。以上示例中 TiDB 使用了 a 列的索引。尤其是在有组合索引的情况下，该字段显示的信息很有参考意义。
- operator info 显示访问表、分区和索引的其他信息。详见operator info 结果。

算子简介

算子是为返回查询结果而执行的特定步骤。真正执行扫表（读盘或者读 TiKV Block Cache）操作的算子有如下几类：

- TableFullScan：全表扫描。
- TableRangeScan：带有范围的表数据扫描。
- TableRowIDScan：根据上层传递下来的 RowID 扫描表数据。时常在索引读操作后检索符合条件的行。
- IndexFullScan：另一种“全表扫描”，扫的是索引数据，不是表数据。
- IndexRangeScan：带有范围的索引数据扫描操作。

TiDB 会汇聚 TiKV/TiFlash 上扫描的数据或者计算结果，这种“数据汇聚”算子目前有如下几类：

- TableReader：将 TiKV 上底层扫表算子 TableFullScan 或 TableRangeScan 得到的数据进行汇总。
- IndexReader：将 TiKV 上底层扫表算子 IndexFullScan 或 IndexRangeScan 得到的数据进行汇总。
- IndexLookUp：先汇总 Build 端 TiKV 扫描上来的 RowID，再去 Probe 端上根据这些 RowID 精确地读取 TiKV 上的数据。Build 端是 IndexFullScan 或 IndexRangeScan 类型的算子，Probe 端是 TableRowIDScan 类型的算子。
- IndexMerge：和 IndexLookupReader 类似，可以看做是它的扩展，可以同时读取多个索引的数据，有多个 Build 端，一个 Probe 端。执行过程也很类似，先汇总所有 Build 端 TiKV 扫描上来的 RowID，再去 Probe 端上根据这些 RowID 精确地读取 TiKV 上的数据。Build 端是 IndexFullScan 或 IndexRangeScan 类型的算子，Probe 端是 TableRowIDScan 类型的算子。

算子的执行顺序

算子的结构是树状的，但在查询执行过程中，并不严格要求子节点任务在父节点之前完成。TiDB 支持同一查询内的并行处理，即子节点“流入”父节点。父节点、子节点和同级节点可能并行执行查询的一部分。

在以上示例中，├─IndexRangeScan_8(Build) 算子为 a(a) 索引所匹配的行查找内部 RowID。└─TableRowIDScan_9 ↪ (Probe) 算子随后从表中检索这些行。

Build 总是先于 Probe 执行，并且 Build 总是出现在 Probe 前面。即如果一个算子有多个子节点，子节点 ID 后面有 Build 关键字的算子总是先于有 Probe 关键字的算子执行。TiDB 在展现执行计划的时候，Build 端总是第一个出现，接着才是 Probe 端。

范围查询

在 WHERE/HAVING/ON 条件中，TiDB 优化器会分析主键或索引键的查询返回。如数字、日期类型的比较符，如大于、小于、等于以及大于等于、小于等于，字符类型的 LIKE 符号等。

若要使用索引，条件必须是 “Sargable” (Search ARGument ABLE) 的。例如条件 YEAR(date_column)< 1992 不能使用索引，但 date_column < '1992-01-01 就可以使用索引。

推荐使用同一类型的数据以及同一类型的字符串和排序规则进行比较，以避免引入额外的 cast 操作而导致不能利用索引。

可以在范围查询条件中使用 AND（求交集）和 OR（求并集）进行组合。对于多维组合索引，可以对多个列使用条件。例如对组合索引 (a, b, c)：

- 当 a 为等值查询时，可以继续求 b 的查询范围。
- 当 b 也为等值查询时，可以继续求 c 的查询范围。
- 反之，如果 a 为非等值查询，则只能求 a 的范围。

Task 简介

目前 TiDB 的计算任务分为两种不同的 task：cop task 和 root task。Cop task 是指使用 TiKV 中的 Coprocessor 执行的计算任务，root task 是指在 TiDB 中执行的计算任务。

SQL 优化的目标之一是将计算尽可能地下推到 TiKV 中执行。TiKV 中的 Coprocessor 能支持大部分 SQL 内建函数（包括聚合函数和标量函数）、SQL LIMIT 操作、索引扫描和表扫描。但是，所有的 Join 操作都只能作为 root task 在 TiDB 上执行。

operator info 结果

EXPLAIN 返回结果中 operator info 列可显示诸如条件下推等信息。本文以上示例中，operator info 结果各字段解释如下：

- range: [1,1] 表示查询的 WHERE 字句 (a = 1) 被下推到了 TiKV，对应的 task 为 cop[tikv]。
- keep order:false 表示该查询的语义不需要 TiKV 按顺序返回结果。如果查询指定了排序（例如 SELECT * FROM t WHERE a = 1 ORDER BY id），该字段的返回结果为 keep order:true。
- stats:pseudo 表示 estRows 显示的预估数可能不准确。TiDB 定期在后台更新统计信息。也可以通过执行 ANALYZE TABLE t 来手动更新统计信息。

EXPLAIN 执行后，不同算子返回不同的信息。你可以使用 Optimizer Hints 来控制优化器的行为，以此控制物理算子的选择。例如 /*+ HASH_JOIN(t1, t2)*/ 表示优化器将使用 Hash Join 算法。详细内容见Optimizer Hints。

10.3.2.1.2 算子相关的系统变量

TiDB 在 MySQL 的基础上，定义了一些专用的系统变量和语法用来优化性能。其中一些系统变量和具体的算子相关，比如算子的并发度，算子的内存使用上限，是否允许使用分区表等。这些都可以通过系统变量进行控制，从而影响各个算子执行的效率。

如果读者想要详细了解所有的系统变量及其使用规则，可以参见系统变量和语法。

10.3.2.1.3 另请参阅

- EXPLAIN
- EXPLAIN ANALYZE
- ANALYZE TABLE

- TRACE
- TiDB in Action
- 系统变量

10.3.2.2 使用 EXPLAIN 解读执行计划

SQL 是一种声明性语言，因此用户无法根据 SQL 语句直接判断一条查询的执行是否有效率。用户首先要使用 EXPLAIN 语句查看当前的执行计划。

以 bikeshare 数据库示例（英文）中的一个 SQL 语句为例，该语句统计了 2017 年 7 月 1 日的行程次数：

```
EXPLAIN SELECT count(*) FROM trips WHERE start_date BETWEEN '2017-07-01 00:00:00' AND '2017-07-01
    ↪  23:59:59';
```

```
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
| id                           | estRows  | task      | access object | operator info
    ↪
    ↪ |
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
| StreamAgg_20                 | 1.00     | root      |               | funcs:count(Column#13)->
    ↪ Column#11
    ↪                                                                                                 |
| └─TableReader_21             | 1.00     | root      |               | data:StreamAgg_9
    ↪
    ↪ |
|   └─StreamAgg_9              | 1.00     | cop[tikv] |               | funcs:count(1)->Column
    ↪ #13
    ↪
    ↪ |
|     └─Selection_19           | 250.00   | cop[tikv] |               | ge(bikeshare.trips.
    ↪ start_date, 2017-07-01 00:00:00.000000), le(bikeshare.trips.start_date, 2017-07-01
    ↪ 23:59:59.000000) |
|       └─TableFullScan_18     | 10000.00 | cop[tikv] | table:trips   | keep order:false, stats:
    ↪ pseudo
    ↪                                                                                                 |
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
5 rows in set (0.00 sec)
```

以上是该查询的执行计划结果。从 └─TableFullScan_18 算子开始向上看，查询的执行过程如下（非最佳执行计划）：

1. Coprocessor (TiKV) 读取整张 `trips` 表的数据，作为一次 TableFullScan 操作，再将读取到的数据传递给 Selection_19 算子。Selection_19 算子仍在 TiKV 内。
2. Selection_19 算子根据谓词 `WHERE start_date BETWEEN ..` 进行数据过滤。预计大约有 250 行数据满足该过滤条件（基于统计信息以及算子的执行逻辑估算而来）。└─TableFullScan_18 算子显示 `stats:pseudo`，表示该表没有实际统计信息，执行 `ANALYZE TABLE trips` 收集统计信息后，预计的估算的数字会更加准确。
3. COUNT 函数随后应用于满足过滤条件的行，这一过程也是在 TiKV (`cop[tikv]`) 中的 StreamAgg_9 算子内完成的。TiKV coprocessor 能执行一些 MySQL 内置函数，COUNT 是其中之一。
4. StreamAgg_9 算子执行的结果会被传递给 TableReader_21 算子（位于 TiDB 进程中，即 `root` 任务）。执行计划中，TableReader_21 算子的 `estRows` 为 1，表示该算子将从每个访问的 TiKV Region 接收一行数据。这一请求过程的详情，可参阅 `EXPLAIN ANALYZE`。
5. StreamAgg_20 算子随后对 └─TableReader_21 算子传来的每行数据计算 COUNT 函数的结果。StreamAgg_20 是根算子，会将结果返回给客户端。

> **注意：**
>
> 要查看 TiDB 中某张表的 Region 信息，可执行 `SHOW TABLE REGIONS` 语句。

10.3.2.2.1 评估当前的性能

`EXPLAIN` 语句只返回查询的执行计划，并不执行该查询。若要获取实际的执行时间，可执行该查询，或使用 `EXPLAIN ANALYZE` 语句：

```
EXPLAIN ANALYZE SELECT count(*) FROM trips WHERE start_date BETWEEN '2017-07-01 00:00:00' AND '
    ↪ 2017-07-01 23:59:59';
```

```
+--
    ↪ ---------------------------+----------+----------+-----------+---------------+-----------------------------
    ↪
| id                          | estRows  | actRows  | task      | access object | execution info
    ↪
    ↪
    ↪ | operator info
    ↪
    ↪ | memory    | disk |
+--
    ↪ ---------------------------+----------+----------+-----------+---------------+-----------------------------
    ↪
| StreamAgg_20                | 1.00     | 1        | root      |               | time
    ↪ :1.031417203s, loops:2
    ↪
```

```
    ↪ | funcs:count(Column#13)->Column#11
    ↪                                                                                |
    ↪ 632 Bytes | N/A  |
| └─TableReader_21              | 1.00     | 56       | root      |               | time
    ↪ :1.031408123s, loops:2, cop_task: {num: 56, max: 782.147269ms, min: 5.759953ms, avg:
    ↪ 252.005927ms, p95: 609.294603ms, max_proc_keys: 910371, p95_proc_keys: 704775, tot_proc:
    ↪ 11.524s, tot_wait: 580ms, rpc_num: 56, rpc_time: 14.111932641s} | data:StreamAgg_9
    ↪
    ↪ | 328 Bytes | N/A  |
|   └─StreamAgg_9               | 1.00     | 56       | cop[tikv] |               | proc max:640
    ↪ ms, min:8ms, p80:276ms, p95:480ms, iters:18695, tasks:56
    ↪
    ↪ | funcs:count(1)->Column#13
    ↪
    ↪ | N/A       | N/A  |
|     └─Selection_19            | 250.00   | 11409    | cop[tikv] |               | proc max:640
    ↪ ms, min:8ms, p80:276ms, p95:476ms, iters:18695, tasks:56
    ↪
    ↪ | ge(bikeshare.trips.start_date, 2017-07-01 00:00:00.000000), le(bikeshare.trips.
    ↪ start_date, 2017-07-01 23:59:59.000000) | N/A       | N/A  |
|       └─TableFullScan_18      | 10000.00 | 19117643 | cop[tikv] | table:trips   | proc max:612
    ↪ ms, min:8ms, p80:248ms, p95:460ms, iters:18695, tasks:56
    ↪
    ↪ | keep order:false, stats:pseudo
    ↪                                                                                |
    ↪  N/A       | N/A  |
+--
    ↪ ----------------------------+----------+----------+-----------+---------------+-------------------------
    ↪
5 rows in set (1.03 sec)
```

执行以上示例查询耗时 1.03 秒，说明执行性能较为理想。

以上 EXPLAIN ANALYZE 的结果中，actRows 表明一些 estRows 预估数不准确（预估返回 10000 行数据但实际返回 19117643 行）。└─TableFullScan_18 算子的 operator info 列 (stats:pseudo) 信息也表明该算子的预估数不准确。

如果先执行 ANALYZE TABLE 再执行 EXPLAIN ANALYZE，预估数与实际数会更接近：

```
ANALYZE TABLE trips;
EXPLAIN ANALYZE SELECT count(*) FROM trips WHERE start_date BETWEEN '2017-07-01 00:00:00' AND '
    ↪ 2017-07-01 23:59:59';
```

```
Query OK, 0 rows affected (10.22 sec)

+--
    ↪ ----------------------------+-------------+----------+-----------+---------------+-------------------------
```

```
    ↪
| id                           | estRows     | actRows  | task      | access object | execution
    ↪ info
    ↪
    ↪ | operator info
    ↪
    ↪ | memory    | disk |
+--
    ↪ ----------------------------+-------------+----------+-----------+---------------+----------------------
    ↪
| StreamAgg_20                 | 1.00        | 1        | root      |               | time
    ↪ :926.393612ms, loops:2
    ↪
    ↪ | funcs:count(Column#13)->Column#11
    ↪                                                                                  |
    ↪ 632 Bytes | N/A  |
| └─TableReader_21             | 1.00        | 56       | root      |               | time
    ↪ :926.384792ms, loops:2, cop_task: {num: 56, max: 850.94424ms, min: 6.042079ms, avg:
    ↪ 234.987725ms, p95: 495.474806ms, max_proc_keys: 910371, p95_proc_keys: 704775, tot_proc:
    ↪ 10.656s, tot_wait: 904ms, rpc_num: 56, rpc_time: 13.158911952s} | data:StreamAgg_9
    ↪
    ↪ | 328 Bytes | N/A  |
|   └─StreamAgg_9              | 1.00        | 56       | cop[tikv] |               | proc max
    ↪ :592ms, min:4ms, p80:244ms, p95:480ms, iters:18695, tasks:56
    ↪
    ↪ | funcs:count(1)->Column#13
    ↪
    ↪ | N/A       | N/A  |
|     └─Selection_19           | 432.89      | 11409    | cop[tikv] |               | proc max
    ↪ :592ms, min:4ms, p80:244ms, p95:480ms, iters:18695, tasks:56
    ↪
    ↪ | ge(bikeshare.trips.start_date, 2017-07-01 00:00:00.000000), le(bikeshare.trips.
    ↪ start_date, 2017-07-01 23:59:59.000000) | N/A       | N/A  |
|       └─TableFullScan_18     | 19117643.00 | 19117643 | cop[tikv] | table:trips   | proc max
    ↪ :564ms, min:4ms, p80:228ms, p95:456ms, iters:18695, tasks:56
    ↪
    ↪ | keep order:false
    ↪
    ↪ | N/A       | N/A  |
+--
    ↪ ----------------------------+-------------+----------+-----------+---------------+----------------------
    ↪
5 rows in set (0.93 sec)
```

执行 ANALYZE TABLE 后，可以看到 └─TableFullScan_18 算子的预估行数是准确的，└─Selection_19 算子的预

估行数也更接近实际行数。以上两个示例中的执行计划（即 TiDB 执行查询所使用的一组算子）未改变，但过时的统计信息常常导致 TiDB 选择到非最优的执行计划。

除 ANALYZE TABLE 外，达到tidb_auto_analyze_ratio 阈值后，TiDB 会自动在后台重新生成统计数据。若要查看 TiDB 有多接近该阈值（即 TiDB 判断统计数据有多健康），可执行 SHOW STATS_HEALTHY 语句。

```
SHOW STATS_HEALTHY;
```

```
+-----------+------------+----------------+---------+
| Db_name   | Table_name | Partition_name | Healthy |
+-----------+------------+----------------+---------+
| bikeshare | trips      |                |     100 |
+-----------+------------+----------------+---------+
1 row in set (0.00 sec)
```

10.3.2.2.2 确定优化方案

当前执行计划是有效率的：

- 大部分任务是在 TiKV 内处理的，需要通过网络传输给 TiDB 处理的仅有 56 行数据，每行都满足过滤条件，而且都很短。
- 在 TiDB (StreamAgg_20) 中和在 TiKV (└─StreamAgg_9) 中汇总行数都使用了 Stream Aggregate，该算法在内存使用方面很有效率。

当前执行计划存在的最大问题在于谓词 start_date BETWEEN '2017-07-01 00:00:00' AND '2017-07-01 ↪ 23:59:59' 并未立即生效，先是 TableFullScan 算子读取所有行数据，然后才进行过滤选择。可以在 SHOW CREATE TABLE trips 的返回结果中找出问题原因：

```
SHOW CREATE TABLE trips\G
```

```
*************************** 1. row ***************************
       Table: trips
Create Table: CREATE TABLE `trips` (
  `trip_id` bigint(20) NOT NULL AUTO_INCREMENT,
  `duration` int(11) NOT NULL,
  `start_date` datetime DEFAULT NULL,
  `end_date` datetime DEFAULT NULL,
  `start_station_number` int(11) DEFAULT NULL,
  `start_station` varchar(255) DEFAULT NULL,
  `end_station_number` int(11) DEFAULT NULL,
  `end_station` varchar(255) DEFAULT NULL,
  `bike_number` varchar(255) DEFAULT NULL,
  `member_type` varchar(255) DEFAULT NULL,
  PRIMARY KEY (`trip_id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_bin AUTO_INCREMENT=20477318
1 row in set (0.00 sec)
```

以上返回结果显示，start_date 列没有索引。要将该谓词下推到 index reader 算子，还需要一个索引。添加索引如下：

```
ALTER TABLE trips ADD INDEX (start_date);
```

```
Query OK, 0 rows affected (2 min 10.23 sec)
```

> **注意：**
>
> 你可通过执行 ADMIN SHOW DDL JOBS 语句来查看 DDL 任务的进度。TiDB 中的默认值的设置较为保守，因此添加索引不会对生产环境下的负载造成太大影响。测试环境下，可以考虑调大tidb_ddl_reorg_batch_size 和tidb_ddl_reorg_worker_cnt 的值。在参照系统上，将批处理大小设为 10240，将 worker count 并发度设置为 32，该系统可获得 10 倍的性能提升（较之使用默认值）。

添加索引后，可以使用 EXPLAIN 重复该查询。在以下返回结果中，可见 TiDB 选择了新的执行计划，而且不再使用 TableFullScan 和 Selection 算子。

```
EXPLAIN SELECT count(*) FROM trips WHERE start_date BETWEEN '2017-07-01 00:00:00' AND '2017-07-01
    ↪  23:59:59';
```

```
+--
    ↪ --------------------------+---------+-----------+-----------------------------------------+----------
    ↪
| id                          | estRows | task      | access object                           |
    ↪  operator info                                                  |
+--
    ↪ --------------------------+---------+-----------+-----------------------------------------+----------
    ↪
| StreamAgg_17                | 1.00    | root      |                                         |
    ↪  funcs:count(Column#13)->Column#11                              |
| └─IndexReader_18            | 1.00    | root      |
    ↪ | index:StreamAgg_9                                             |
|   └─StreamAgg_9             | 1.00    | cop[tikv] |
    ↪ | funcs:count(1)->Column#13                                      |
|     └─IndexRangeScan_16     | 8471.88 | cop[tikv] | table:trips, index:start_date(start_date)
    ↪ | range:[2017-07-01 00:00:00,2017-07-01 23:59:59], keep order:false |
+--
    ↪ --------------------------+---------+-----------+-----------------------------------------+----------
    ↪
4 rows in set (0.00 sec)
```

若要比较实际的执行时间，可再次使用 EXPLAIN ANALYZE 语句：

```
EXPLAIN ANALYZE SELECT count(*) FROM trips WHERE start_date BETWEEN '2017-07-01 00:00:00' AND '
    ↪ 2017-07-01 23:59:59';
```

```
+--
    ↪ -------------------------+---------+---------+-----------+-------------------------------------+-
    ↪
| id                          | estRows | actRows | task      | access object
    ↪                              | execution info
    ↪
    ↪ | operator info                                                 | memory    | disk |
+--
    ↪ -------------------------+---------+---------+-----------+-------------------------------------+-
    ↪
| StreamAgg_17                | 1.00    | 1       | root      |
    ↪                                          | time:4.516728ms, loops:2
    ↪                                                                                                |
    ↪  funcs:count(Column#13)->Column#11                                | 372 Bytes | N/A  |
| └─IndexReader_18            | 1.00    | 1       | root      |
    ↪                                          | time:4.514278ms, loops:2, cop_task: {num: 1,
    ↪ max:4.462288ms, proc_keys: 11409, rpc_num: 1, rpc_time: 4.457148ms} | index:StreamAgg_9
    ↪                                               | 238 Bytes | N/A  |
|   └─StreamAgg_9             | 1.00    | 1       | cop[tikv] |
    ↪                                          | time:4ms, loops:12
    ↪
    ↪ | funcs:count(1)->Column#13                                       | N/A       | N/A  |
|     └─IndexRangeScan_16     | 8471.88 | 11409   | cop[tikv] | table:trips, index:start_date(
    ↪ start_date) | time:4ms, loops:12
    ↪
    ↪ | range:[2017-07-01 00:00:00,2017-07-01 23:59:59], keep order:false | N/A       | N/A  |
+--
    ↪ -------------------------+---------+---------+-----------+-------------------------------------+-
    ↪
4 rows in set (0.00 sec)
```

从以上结果可看出，查询时间已从 1.03 秒减少到 0.0 秒。

> **注意：**
>
> 以上示例另一个可用的优化方案是coprocessor cache。如果你无法添加索引，可考虑开启 coprocessor cache 功能。开启后，只要算子上次执行以来 Region 未作更改，TiKV 将从缓存中返回值。这也有助于减少 TableFullScan 和 Selection 算子的大部分运算成本。

10.3.2.3 用 EXPLAIN 查看 MPP 模式查询的执行计划

TiDB 支持使用 MPP 模式来执行查询。在 MPP 执行模式下，SQL 优化器会生成 MPP 的执行计划。注意 MPP 模式仅对有 TiFlash 副本的表生效。

本文档使用的示例数据如下：

```
CREATE TABLE t1 (id int, value int);
INSERT INTO t1 values(1,2),(2,3),(1,3);
ALTER TABLE t1 set tiflash replica 1;
ANALYZE TABLE t1;
SET tidb_allow_mpp = 1;
```

10.3.2.3.1 MPP 查询片段和 MPP 任务

在 MPP 模式下，一个查询在逻辑上会被切分为多个 MPP 查询片段 (query fragment)。示例如下：

```
EXPLAIN SELECT COUNT(*) FROM t1 GROUP BY id;
```

这个查询在 MPP 模式下会包含两个查询片段，一个为一阶段聚合，一个为二阶段聚合（最终聚合）。在查询执行的时候每个查询片段都会被实例化为一个或者多个 MPP 任务。

10.3.2.3.2 Exchange 算子

MPP 查询的执行计划中有两个 MPP 特有的 Exchange 算子，分别为 ExchangeReceiver 和 ExchangeSender。ExchangeReceiver 表示从下游查询片段读取数据，ExchangeSender 表示下游查询片段向上游查询片段发送数据。在 MPP 执行模式下，每个 MPP 查询片段的根算子均为 ExchangeSender 算子，即每个查询片段以 ExchangeSender 为界进行划分。一个简单的 MPP 计划如下：

```
EXPLAIN SELECT COUNT(*) FROM t1 GROUP BY id;
```

```
+--
    ↪ ---------------------------------+---------+-------------------+--------------+-------------------------
    ↪
| id                                | estRows | task              | access object | operator
    ↪ info                                    |
+--
    ↪ ---------------------------------+---------+-------------------+--------------+-------------------------
    ↪
| TableReader_31                    | 2.00    | root              |               | data:
    ↪ ExchangeSender_30                              |
| └─ExchangeSender_30               | 2.00    | batchCop[tiflash] |               |
    ↪ ExchangeType: PassThrough                      |
|   └─Projection_26                 | 2.00    | batchCop[tiflash] |               | Column#4
    ↪                                          |
|     └─HashAgg_27                  | 2.00    | batchCop[tiflash] |               | group by:
    ↪ test.t1.id, funcs:sum(Column#7)->Column#4 |
```

```
|       └─ExchangeReceiver_29        | 2.00    | batchCop[tiflash] |               |
    ↪                                                          |
|         └─ExchangeSender_28        | 2.00    | batchCop[tiflash] |               |
    ↪ ExchangeType: HashPartition, Hash Cols: test.t1.id |
|           └─HashAgg_9              | 2.00    | batchCop[tiflash] |               | group by:
    ↪ test.t1.id, funcs:count(1)->Column#7      |
|             └─TableFullScan_25     | 3.00    | batchCop[tiflash] | table:t1      | keep order:
    ↪ false                                    |
+--
    ↪ ---------------------------------+---------+-------------------+---------------+-----------------------
    ↪
```

以上执行计划中有两个查询片段：

- `[TableFullScan_25, HashAgg_9, ExchangeSender_28]` 为第一个查询片段，其主要完成一阶段聚合的计算。
- `[ExchangeReceiver_29, HashAgg_27, Projection_26, ExchangeSender_30]` 为第二个查询片段，其主要完成二阶段聚合的计算。

ExchangeSender 算子的 `operator info` 列输出了 ExchangeType 信息。目前有以下三种 ExchangeType：

- HashPartition：ExchangeSender 把数据按 Hash 值进行分区之后分发给上游的 MPP 任务的 ExchangeReceiver 算子，通常在 Hash Aggregation 以及 Shuffle Hash Join 算法中使用。
- Broadcast：ExchangeSender 通过广播的方式把数据分发给上游的 MPP 任务，通常在 Broadcast Join 中使用。
- PassThrough：ExchangeSender 把数据分发给上游的 MPP Task，与 Broadcast 的区别是此时上游有且仅有一个 MPP 任务，通常用于向 TiDB 返回数据。

上述例子中 ExchangeSender 的 ExchangeType 为 HashPartition 以及 PassThrough，分别对应于 Hash Aggregation 运算以及向 TiDB 返回数据。

另外一个典型的 MPP 应用为 join 运算。TiDB MPP 支持两种类型的 join，分别为：

- Shuffle Hash Join：join 的 input 通过 HashPartition 的方式 shuffle 数据，上游的 MPP 任务进行分区内的 join。
- Broadcast Join：join 中的小表以 Broadcast 的方式把数据广播到各个节点，各个节点各自进行 join。

典型的 Shuffle Hash Join 执行计划如下：

```
SET tidb_opt_broadcast_join=0; SET tidb_broadcast_join_threshold_count=0; SET
    ↪ tidb_broadcast_join_threshold_size=0; EXPLAIN SELECT COUNT(*) FROM t1 a JOIN t1 b ON a.id
    ↪  = b.id;
```

```
+--
    ↪ ------------------------------------+---------+-------------+--------------+-----------------------
    ↪
| id                                  | estRows | task        | access object | operator info
    ↪                                      |
```

```
+--
    ↪ ------------------------------------+---------+--------------+---------------+-------------------------
    ↪
| StreamAgg_14                          | 1.00    | root         |               | funcs:count
    ↪ (1)->Column#7                             |
| └─TableReader_48                      | 9.00    | root         |               | data:
    ↪ ExchangeSender_47                           |
|   └─ExchangeSender_47                 | 9.00    | cop[tiflash] |               | ExchangeType
    ↪ : PassThrough                           |
|     └─HashJoin_44                     | 9.00    | cop[tiflash] |               | inner join,
    ↪ equal:[eq(test.t1.id, test.t1.id)]      |
|       ├─ExchangeReceiver_19(Build)    | 6.00    | cop[tiflash] |               |
    ↪                                                 |
|       │ └─ExchangeSender_18           | 6.00    | cop[tiflash] |               | ExchangeType
    ↪ : HashPartition, Hash Cols: test.t1.id |
|       │   └─Selection_17              | 6.00    | cop[tiflash] |               | not(isnull(
    ↪ test.t1.id))                            |
|       │     └─TableFullScan_16        | 6.00    | cop[tiflash] | table:a       | keep order:
    ↪ false                                   |
|       └─ExchangeReceiver_23(Probe)    | 6.00    | cop[tiflash] |               |
    ↪                                                 |
|         └─ExchangeSender_22           | 6.00    | cop[tiflash] |               | ExchangeType
    ↪ : HashPartition, Hash Cols: test.t1.id |
|           └─Selection_21              | 6.00    | cop[tiflash] |               | not(isnull(
    ↪ test.t1.id))                            |
|             └─TableFullScan_20        | 6.00    | cop[tiflash] | table:b       | keep order:
    ↪ false                                   |
+--
    ↪ ------------------------------------+---------+--------------+---------------+-------------------------
    ↪
12 rows in set (0.00 sec)
```

以上执行计划中，

- [TableFullScan_20, Selection_21, ExchangeSender_22] 完成表 b 的数据读取并通过 HashPartition 的方式把数据 shuffle 给上游 MPP 任务。
- [TableFullScan_16, Selection_17, ExchangeSender_18] 完成表 a 的数据读取并通过 HashPartition 的方式把数据 shuffle 给上游 MPP 任务。
- [ExchangeReceiver_19, ExchangeReceiver_23, HashJoin_44, ExchangeSender_47] 完成 join 并把数据返回给 TiDB。

典型的 Broadcast Join 执行计划如下：

```
EXPLAIN SELECT COUNT(*) FROM t1 a JOIN t1 b ON a.id = b.id;
```

```
+--
    ↪ ------------------------------------+---------+-------------+---------------+-------------------------
    ↪
| id                                  | estRows | task         | access object | operator info
    ↪                                  |
+--
    ↪ ------------------------------------+---------+-------------+---------------+-------------------------
    ↪
| StreamAgg_15                        | 1.00    | root         |               | funcs:count
    ↪ (1)->Column#7                      |
| └─TableReader_47                     | 9.00    | root         |               | data:
    ↪ ExchangeSender_46                         |
|   └─ExchangeSender_46                | 9.00    | cop[tiflash] |               | ExchangeType
    ↪ : PassThrough                      |
|     └─HashJoin_43                    | 9.00    | cop[tiflash] |               | inner join,
    ↪ equal:[eq(test.t1.id, test.t1.id)] |
|       ├─ExchangeReceiver_20(Build)   | 6.00    | cop[tiflash] |               |
    ↪                                           |
|       | └─ExchangeSender_19          | 6.00    | cop[tiflash] |               | ExchangeType
    ↪ : Broadcast                        |
|       |   └─Selection_18             | 6.00    | cop[tiflash] |               | not(isnull(
    ↪ test.t1.id))                       |
|       |     └─TableFullScan_17       | 6.00    | cop[tiflash] | table:a       | keep order:
    ↪ false                              |
|       └─Selection_22(Probe)          | 6.00    | cop[tiflash] |               | not(isnull(
    ↪ test.t1.id))                       |
|         └─TableFullScan_21           | 6.00    | cop[tiflash] | table:b       | keep order:
    ↪ false                              |
+--
    ↪ ------------------------------------+---------+-------------+---------------+-------------------------
    ↪
```

以上执行计划中，

- `[TableFullScan_17, Selection_18, ExchangeSender_19]` 从小表（表 a）读数据并广播给大表（表 b）数据所在的各个节点。
- `[TableFullScan_21, Selection_22, ExchangeReceiver_20, HashJoin_43, ExchangeSender_46]` 完成 join 并将数据返回给 TiDB。

10.3.2.3.3 对 MPP 模式的查询使用 EXPLAIN ANALYZE

`EXPLAIN ANALYZE` 语句与 `EXPLAIN` 类似，但还会输出一些运行时的信息。一个简单的 `EXPLAIN ANALYZE` 输出信息如下：

```
EXPLAIN ANALYZE SELECT COUNT(*) FROM t1 GROUP BY id;
```

```
+--
    ↪ ---------------------------------+---------+---------+------------------+---------------+--------------
    ↪
| id                                | estRows | actRows | task              | access object |
    ↪ execution info
    ↪                                                                                          | operator
    ↪ info                                                  | memory | disk |
+--
    ↪ ---------------------------------+---------+---------+------------------+---------------+--------------
    ↪
| TableReader_31                      | 4.00    | 2       | root              |               |
    ↪ time:44.5ms, loops:2, cop_task: {num: 1, max: 0s, proc_keys: 0, copr_cache_hit_ratio:
    ↪ 0.00} | data:ExchangeSender_30                                          | N/A    | N/A  |
| └─ExchangeSender_30                | 4.00    | 2       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:16.5ms, loops:1, threads:1}
    ↪                                                 | ExchangeType: PassThrough, tasks: [2, 3,
    ↪ 4]                     | N/A    | N/A  |
|   └─Projection_26                  | 4.00    | 2       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:16.5ms, loops:1, threads:1}
    ↪                                                 | Column#4
    ↪                                                         | N/A    | N/A  |
|     └─HashAgg_27                   | 4.00    | 2       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:16.5ms, loops:1, threads:1}
    ↪                                                 | group by:test.t1.id, funcs:sum(Column#7)->
    ↪ Column#4              | N/A    | N/A  |
|       └─ExchangeReceiver_29        | 4.00    | 2       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:14.5ms, loops:1, threads:20}
    ↪                                                |
    ↪                                                                     | N/A    | N/A  |
|         └─ExchangeSender_28        | 4.00    | 0       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:9.49ms, loops:0, threads:0}
    ↪                                                 | ExchangeType: HashPartition, Hash Cols:
    ↪ test.t1.id, tasks: [1] | N/A    | N/A  |
|           └─HashAgg_9              | 4.00    | 0       | batchCop[tiflash] |               |
    ↪ tiflash_task:{time:9.49ms, loops:0, threads:0}
    ↪                                                 | group by:test.t1.id, funcs:count(1)->
    ↪ Column#7                  | N/A    | N/A  |
|             └─TableFullScan_25     | 6.00    | 0       | batchCop[tiflash] | table:t1      |
    ↪ tiflash_task:{time:9.49ms, loops:0, threads:0}
    ↪                                                 | keep order:false
    ↪                                                  | N/A    | N/A  |
+--
    ↪ ---------------------------------+---------+---------+------------------+---------------+--------------
    ↪
```

与 EXPLAIN 相比，ExchangeSender 的 operator info 中多了 task id 的输出，其记录了该查询片段实例化成的 MPP 任务的任务 ID。此外 MPP 算子中都会有 threads 这一列，这列记录了 MPP 在执行该算子时使用的并发数 (如果集群由多个节点组成，该并发数是所有节点并发数相加的结果)。

10.3.2.3.4 其他类型查询的执行计划

- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.4 用 EXPLAIN 查看索引查询的执行计划

SQL 查询可能会使用索引，可以通过 EXPLAIN 语句来查看索引查询的执行计划。本文提供多个示例，以帮助用户理解索引查询是如何执行的。

TiDB 支持以下使用索引的算子来提升查询速度：

- IndexLookup
- IndexReader
- Point_Get 和 Batch_Point_Get
- IndexFullScan

本文档中的示例都基于以下数据：

```
CREATE TABLE t1 (
 id INT NOT NULL PRIMARY KEY auto_increment,
 intkey INT NOT NULL,
 pad1 VARBINARY(1024),
 INDEX (intkey)
);

INSERT INTO t1 SELECT NULL, FLOOR(RAND()*1024), RANDOM_BYTES(1024) FROM dual;
INSERT INTO t1 SELECT NULL, FLOOR(RAND()*1024), RANDOM_BYTES(1024) FROM t1 a JOIN t1 b JOIN t1 c
    ↪ LIMIT 10000;
INSERT INTO t1 SELECT NULL, FLOOR(RAND()*1024), RANDOM_BYTES(1024) FROM t1 a JOIN t1 b JOIN t1 c
    ↪ LIMIT 10000;
INSERT INTO t1 SELECT NULL, FLOOR(RAND()*1024), RANDOM_BYTES(1024) FROM t1 a JOIN t1 b JOIN t1 c
    ↪ LIMIT 10000;
```

10.3.2.4.1 IndexLookup

TiDB 从二级索引检索数据时会使用 IndexLookup 算子。例如，以下所有查询均会在 intkey 列的索引上使用 IndexLookup 算子：

```
EXPLAIN SELECT * FROM t1 WHERE intkey = 123;
EXPLAIN SELECT * FROM t1 WHERE intkey < 10;
EXPLAIN SELECT * FROM t1 WHERE intkey BETWEEN 300 AND 310;
EXPLAIN SELECT * FROM t1 WHERE intkey IN (123,29,98);
EXPLAIN SELECT * FROM t1 WHERE intkey >= 99 AND intkey <= 103;
```

```
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
    ↪  info                       |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 1.00    | root      |                               |
    ↪                                       |
| ├─IndexRangeScan_8(Build)     | 1.00    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[123,123], keep order:false |
| └─TableRowIDScan_9(Probe)     | 1.00    | cop[tikv] | table:t1                       | keep
    ↪ order:false                   |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
    ↪  info                       |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 3.60    | root      |                               |
    ↪                                       |
| ├─IndexRangeScan_8(Build)     | 3.60    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[-inf,10), keep order:false |
| └─TableRowIDScan_9(Probe)     | 3.60    | cop[tikv] | table:t1                       | keep
    ↪ order:false                   |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
```

```
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
    ↪  info                       |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 5.67    | root      |                               |
    ↪                                      |
| ├─IndexRangeScan_8(Build)     | 5.67    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[300,310], keep order:false |
| └─TableRowIDScan_9(Probe)     | 5.67    | cop[tikv] | table:t1                      | keep
    ↪ order:false                  |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
    ↪  info                       |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 5.67    | root      |                               |
    ↪                                      |
| ├─IndexRangeScan_8(Build)     | 5.67    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[300,310], keep order:false |
| └─TableRowIDScan_9(Probe)     | 5.67    | cop[tikv] | table:t1                      | keep
    ↪ order:false                  |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
```

```
    ↪  info                                        |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 4.00    | root      |                               |
    ↪                                                 |
| ├─IndexRangeScan_8(Build)     | 4.00    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[29,29], [98,98], [123,123], keep order:false |
| └─TableRowIDScan_9(Probe)     | 4.00    | cop[tikv] | table:t1                       | keep
    ↪ order:false                                    |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                 | operator
    ↪  info                     |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| IndexLookUp_10                | 6.00    | root      |                               |
    ↪                                       |
| ├─IndexRangeScan_8(Build)     | 6.00    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[99,103], keep order:false |
| └─TableRowIDScan_9(Probe)     | 6.00    | cop[tikv] | table:t1                       | keep
    ↪ order:false                   |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)
```

IndexLookup 算子有以下两个子节点：

- ├─IndexRangeScan_8(Build) 算子节点对 intkey 列的索引执行范围扫描，并检索内部的 RowID 值（对此表而言，即为主键）。
- └─TableRowIDScan_9(Probe) 算子节点随后从表数据中检索整行。

IndexLookup 任务分以上两步执行。如果满足条件的行较多，SQL 优化器可能会根据统计信息选择使用 TableFullScan 算子。在以下示例中，很多行都满足 intkey > 100 这一条件，因此优化器选择了 TableFullScan：

```
EXPLAIN SELECT * FROM t1 WHERE intkey > 100;
```

```
+-------------------------+---------+-----------+---------------+-------------------------+
| id                      | estRows | task      | access object | operator info           |
+-------------------------+---------+-----------+---------------+-------------------------+
| TableReader_7           | 898.50  | root      |               | data:Selection_6        |
| └─Selection_6           | 898.50  | cop[tikv] |               | gt(test.t1.intkey, 100) |
|   └─TableFullScan_5     | 1010.00 | cop[tikv] | table:t1      | keep order:false        |
+-------------------------+---------+-----------+---------------+-------------------------+
3 rows in set (0.00 sec)
```

IndexLookup 算子能在带索引的列上有效优化 LIMIT：

```
EXPLAIN SELECT * FROM t1 ORDER BY intkey DESC LIMIT 10;
```

```
+--
    ↪ ------------------------------+---------+-----------+--------------------------------+------------------
    ↪
| id                              | estRows | task      | access object                  |
    ↪ operator info                       |
+--
    ↪ ------------------------------+---------+-----------+--------------------------------+------------------
    ↪
| IndexLookUp_21                  | 10.00   | root      |                                | limit
    ↪ embedded(offset:0, count:10) |
| ├─Limit_20(Build)               | 10.00   | cop[tikv] |                                | offset
    ↪ :0, count:10                  |
| │ └─IndexFullScan_18            | 10.00   | cop[tikv] | table:t1, index:intkey(intkey) | keep
    ↪ order:true, desc              |
| └─TableRowIDScan_19(Probe)      | 10.00   | cop[tikv] | table:t1                       | keep
    ↪ order:false, stats:pseudo     |
+--
    ↪ ------------------------------+---------+-----------+--------------------------------+------------------
    ↪
4 rows in set (0.00 sec)
```

以上示例中，TiDB 从 intkey 索引读取最后 10 行，然后从表数据中检索这些行的 RowID 值。

10.3.2.4.2 IndexReader

TiDB 支持覆盖索引优化 (covering index optimization)。如果 TiDB 能从索引中检索出所有行，就会跳过 IndexLookup 任务中通常所需的第二步（即从表数据中检索整行）。示例如下：

```
EXPLAIN SELECT * FROM t1 WHERE intkey = 123;
EXPLAIN SELECT id FROM t1 WHERE intkey = 123;
```

```
+--
    ↪ ----------------------------+---------+-----------+--------------------------------+-------------------
    ↪
| id                            | estRows | task      | access object                  | operator
    ↪  info                      |
+--
    ↪ ----------------------------+---------+-----------+--------------------------------+-------------------
    ↪
| IndexLookUp_10                | 1.00    | root      |                                |
    ↪                                      |
| ├─IndexRangeScan_8(Build)     | 1.00    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[123,123], keep order:false |
| └─TableRowIDScan_9(Probe)     | 1.00    | cop[tikv] | table:t1                       | keep
    ↪ order:false                  |
+--
    ↪ ----------------------------+---------+-----------+--------------------------------+-------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ------------------------+---------+-----------+--------------------------------+----------------------
    ↪
| id                        | estRows | task      | access object                  | operator info
    ↪                       |
+--
    ↪ ------------------------+---------+-----------+--------------------------------+----------------------
    ↪
| Projection_4              | 1.00    | root      |                                | test.t1.id
    ↪                          |
| └─IndexReader_6           | 1.00    | root      |                                | index:
    ↪ IndexRangeScan_5             |
|   └─IndexRangeScan_5      | 1.00    | cop[tikv] | table:t1, index:intkey(intkey) | range
    ↪ :[123,123], keep order:false |
+--
    ↪ ------------------------+---------+-----------+--------------------------------+----------------------
    ↪
3 rows in set (0.00 sec)
```

以上结果中，id 也是内部的 RowID 值，因此 id 也存储在 intkey 索引中。部分 └─IndexRangeScan_5 任务使用 intkey 索引后，可直接返回 RowID 值。

10.3.2.4.3 Point_Get 和 Batch_Point_Get

TiDB 直接从主键或唯一键检索数据时会使用 Point_Get 或 Batch_Point_Get 算子。这两个算子比 IndexLookup 更有效率。示例如下：

```
EXPLAIN SELECT * FROM t1 WHERE id = 1234;
EXPLAIN SELECT * FROM t1 WHERE id IN (1234,123);

ALTER TABLE t1 ADD unique_key INT;
UPDATE t1 SET unique_key = id;
ALTER TABLE t1 ADD UNIQUE KEY (unique_key);

EXPLAIN SELECT * FROM t1 WHERE unique_key = 1234;
EXPLAIN SELECT * FROM t1 WHERE unique_key IN (1234, 123);
```

```
+-------------+---------+------+---------------+---------------+
| id          | estRows | task | access object | operator info |
+-------------+---------+------+---------------+---------------+
| Point_Get_1 | 1.00    | root | table:t1      | handle:1234   |
+-------------+---------+------+---------------+---------------+
1 row in set (0.00 sec)

+--
    ↪ -----------------+---------+------+---------------+-----------------------------------------------+
    ↪
| id                | estRows | task | access object | operator info
    ↪                                    |
+--
    ↪ -----------------+---------+------+---------------+-----------------------------------------------+
    ↪
| Batch_Point_Get_1 | 2.00    | root | table:t1      | handle:[1234 123], keep order:false, desc:
    ↪ false |
+--
    ↪ -----------------+---------+------+---------------+-----------------------------------------------+
    ↪
1 row in set (0.00 sec)

Query OK, 0 rows affected (0.27 sec)

Query OK, 1010 rows affected (0.06 sec)
Rows matched: 1010  Changed: 1010  Warnings: 0

Query OK, 0 rows affected (0.37 sec)

+-------------+---------+------+-----------------------------------------+---------------+
| id          | estRows | task | access object                           | operator info |
+-------------+---------+------+-----------------------------------------+---------------+
| Point_Get_1 | 1.00    | root | table:t1, index:unique_key(unique_key) |               |
+-------------+---------+------+-----------------------------------------+---------------+
```

```
1 row in set (0.00 sec)

+--
    ↪ -----------------+---------+------+----------------------------------------+-----------------------------
    ↪
| id                | estRows | task | access object                          | operator info
    ↪                 |
+--
    ↪ -----------------+---------+------+----------------------------------------+-----------------------------
    ↪
| Batch_Point_Get_1 | 2.00    | root | table:t1, index:unique_key(unique_key) | keep order:false,
    ↪  desc:false |
+--
    ↪ -----------------+---------+------+----------------------------------------+-----------------------------
    ↪
1 row in set (0.00 sec)
```

10.3.2.4.4 IndexFullScan

索引是有序的，所以优化器可以使用 `IndexFullScan` 算子来优化常见的查询，例如在索引值上使用 `MIN` 或 `Max` 函数：

```
EXPLAIN SELECT MIN(intkey) FROM t1;
EXPLAIN SELECT MAX(intkey) FROM t1;
```

```
+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-----------------------
    ↪
| id                           | estRows | task      | access object                  | operator
    ↪ info                         |
+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-----------------------
    ↪
| StreamAgg_12                 | 1.00    | root      |                                | funcs:min
    ↪ (test.t1.intkey)->Column#4 |
| └─Limit_16                   | 1.00    | root      |                                 | offset
    ↪ :0, count:1                    |
|   └─IndexReader_29           | 1.00    | root      |                                 | index:
    ↪ Limit_28                       |
|     └─Limit_28               | 1.00    | cop[tikv] |                                 | offset
    ↪ :0, count:1                    |
|       └─IndexFullScan_27     | 1.00    | cop[tikv] | table:t1, index:intkey(intkey) | keep
    ↪ order:true                      |
+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-----------------------
    ↪
```

```
5 rows in set (0.00 sec)

+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-------------------
    ↪
| id                           | estRows | task      | access object                  | operator
    ↪ info                         |
+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-------------------
    ↪
| StreamAgg_12                 | 1.00    | root      |                                | funcs:max
    ↪ (test.t1.intkey)->Column#4 |
| └─Limit_16                   | 1.00    | root      |                                | offset
    ↪ :0, count:1                  |
|   └─IndexReader_29           | 1.00    | root      |                                | index:
    ↪ Limit_28                     |
|     └─Limit_28               | 1.00    | cop[tikv] |                                | offset
    ↪ :0, count:1                  |
|       └─IndexFullScan_27     | 1.00    | cop[tikv] | table:t1, index:intkey(intkey) | keep
    ↪ order:true, desc             |
+--
    ↪ ---------------------------+---------+-----------+--------------------------------+-------------------
    ↪
5 rows in set (0.00 sec)
```

以上语句的执行过程中，TiDB 在每一个 TiKV Region 上执行 IndexFullScan 操作。虽然算子名为 FullScan 即全扫描，TiDB 只读取第一行 (└─Limit_28)。每个 TiKV Region 返回各自的 MIN 或 MAX 值给 TiDB，TiDB 再执行流聚合运算来过滤出一行数据。即使表为空，带 MAX 或 MIN 函数的流聚合运算也能保证返回 NULL 值。

相反，在没有索引的值上执行 MIN 函数会在每一个 TiKV Region 上执行 TableFullScan 操作。该查询会要求在 TiKV 中扫描所有行，但 TopN 计算可保证每个 TiKV Region 只返回一行数据给 TiDB。尽管 TopN 能减少 TiDB 和 TiKV 之间的多余数据传输，但该查询的效率仍远不及以上示例 (MIN 能够使用索引)。

```
EXPLAIN SELECT MIN(pad1) FROM t1;
```

```
+--
    ↪ ----------------------------+---------+-----------+--------------+--------------------------------+
    ↪
| id                           | estRows | task      | access object | operator info
    ↪                      |
+--
    ↪ ----------------------------+---------+-----------+--------------+--------------------------------+
    ↪
| StreamAgg_13                 | 1.00    | root      |               | funcs:min(test.t1.pad1)
    ↪ ->Column#4 |
| └─TopN_14                    | 1.00    | root      |               | test.t1.pad1, offset:0,
    ↪  count:1   |
```

```
|   └─TableReader_23             | 1.00    | root      |               | data:TopN_22
    ↪                       |
|     └─TopN_22                  | 1.00    | cop[tikv] |               | test.t1.pad1, offset:0,
    ↪  count:1   |
|       └─Selection_21           | 1008.99 | cop[tikv] |               | not(isnull(test.t1.pad1
    ↪ ))         |
|         └─TableFullScan_20     | 1010.00 | cop[tikv] | table:t1      | keep order:false
    ↪                   |
+--
    ↪ ------------------------------+---------+-----------+---------------+--------------------------------+
    ↪
6 rows in set (0.00 sec)
```

执行以下语句时，TiDB 将使用 IndexFullScan 算子扫描索引中的每一行：

```
EXPLAIN SELECT SUM(intkey) FROM t1;
EXPLAIN SELECT AVG(intkey) FROM t1;
```

```
+--
    ↪ -------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                         | estRows | task      | access object                 | operator
    ↪ info                        |
+--
    ↪ -------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| StreamAgg_20               | 1.00    | root      |                               | funcs:sum(
    ↪ Column#6)->Column#4        |
| └─IndexReader_21           | 1.00    | root      |                               | index:
    ↪ StreamAgg_8                    |
|   └─StreamAgg_8            | 1.00    | cop[tikv] |                               | funcs:sum(
    ↪ test.t1.intkey)->Column#6 |
|     └─IndexFullScan_19     | 1010.00 | cop[tikv] | table:t1, index:intkey(intkey) | keep order
    ↪ :false                     |
+--
    ↪ -------------------------+---------+-----------+-------------------------------+-------------------
    ↪
4 rows in set (0.00 sec)

+--
    ↪ -------------------------+---------+-----------+-------------------------------+-------------------
    ↪
| id                         | estRows | task      | access object                 | operator
    ↪ info                                                          |
+--
    ↪ -------------------------+---------+-----------+-------------------------------+-------------------
```

```
    ↪
| StreamAgg_20               | 1.00    | root      |                                | funcs:avg(
    ↪ Column#7, Column#8)->Column#4                                                    |
| └─IndexReader_21           | 1.00    | root      |                                 | index:
    ↪ StreamAgg_8                                                                      |
|   └─StreamAgg_8            | 1.00    | cop[tikv] |                                 | funcs:
    ↪ count(test.t1.intkey)->Column#7, funcs:sum(test.t1.intkey)->Column#8 |
|     └─IndexFullScan_19     | 1010.00 | cop[tikv] | table:t1, index:intkey(intkey) | keep order
    ↪ :false                                                                           |
+--
    ↪ -------------------------+---------+-----------+--------------------------------+-----------------------
    ↪
4 rows in set (0.00 sec)
```

以上示例中，IndexFullScan 比 TableFullScan 更有效率，因为 (intkey + RowID) 索引中值的长度小于整行的长度。

以下语句不支持使用 IndexFullScan 算子，因为涉及该表中的其他列：

```
EXPLAIN SELECT AVG(intkey), ANY_VALUE(pad1) FROM t1;
```

```
+--
    ↪ ---------------------------+---------+-----------+--------------+--------------------------------
    ↪
| id                           | estRows | task      | access object | operator info
    ↪
    ↪ |
+--
    ↪ ---------------------------+---------+-----------+--------------+--------------------------------
    ↪
| Projection_4                 | 1.00    | root      |               | Column#4, any_value(test.
    ↪ t1.pad1)->Column#5
    ↪                                                                                |
| └─StreamAgg_16               | 1.00    | root      |               | funcs:avg(Column#10,
    ↪ Column#11)->Column#4, funcs:firstrow(Column#12)->test.t1.pad1
    ↪                                      |
|   └─TableReader_17           | 1.00    | root      |               | data:StreamAgg_8
    ↪
    ↪ |
|     └─StreamAgg_8            | 1.00    | cop[tikv] |               | funcs:count(test.t1.
    ↪ intkey)->Column#10, funcs:sum(test.t1.intkey)->Column#11, funcs:firstrow(test.t1.pad1)->
    ↪ Column#12 |
|       └─TableFullScan_15     | 1010.00 | cop[tikv] | table:t1      | keep order:false
    ↪
    ↪ |
+--
    ↪ ---------------------------+---------+-----------+--------------+--------------------------------
```

```
    ↪
5 rows in set (0.00 sec)
```

10.3.2.4.5 其他类型查询的执行计划

- MPP 模式查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.5 用 EXPLAIN 查看 JOIN 查询的执行计划

SQL 查询中可能会使用 JOIN 进行表连接，可以通过 EXPLAIN 语句来查看 JOIN 查询的执行计划。本文提供多个示例，以帮助用户理解表连接查询是如何执行的。

在 TiDB 中，SQL 优化器需要确定数据表的连接顺序，且要判断对于某条特定的 SQL 语句，哪一种 Join 算法最为高效。

本文档使用的示例数据如下：

```
CREATE TABLE t1 (id BIGINT NOT NULL PRIMARY KEY auto_increment, pad1 BLOB, pad2 BLOB, pad3 BLOB,
    ↪ int_col INT NOT NULL DEFAULT 0);
CREATE TABLE t2 (id BIGINT NOT NULL PRIMARY KEY auto_increment, t1_id BIGINT NOT NULL, pad1 BLOB,
    ↪  pad2 BLOB, pad3 BLOB, INDEX(t1_id));
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM
    ↪ dual;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
```

```
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
UPDATE t1 SET int_col = 1 WHERE pad1 = (SELECT pad1 FROM t1 ORDER BY RAND() LIMIT 1);
SELECT SLEEP(1);
ANALYZE TABLE t1, t2;
```

10.3.2.5.1 Index Join

如果预计需要连接的行数较少（一般小于 1 万行），推荐使用 Index Join 算法。这个算法与 MySQL 主要使用的 Join 算法类似。在下表的示例中，TableReader_28(Build) 算子首先读取表 t1，然后根据在 t1 中匹配到的每行数据，依次探查表 t2 中的数据：

```
EXPLAIN SELECT /*+ INL_JOIN(t1, t2) */ * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id;
```

```
+--
    ↪ ------------------------------+-----------+-----------+-----------------------------+-----------------
    ↪
| id                            | estRows   | task      | access object               |
    ↪ operator info                                                                   |
+--
    ↪ ------------------------------+-----------+-----------+-----------------------------+-----------------
    ↪
| IndexJoin_10                  | 180000.00 | root      |                             | inner
    ↪ join, inner:IndexLookUp_9, outer key:test.t1.id, inner key:test.t2.t1_id |
| ├─TableReader_28(Build)       | 142020.00 | root      |                             | data:
    ↪ TableFullScan_27                                                          |
```

```
| | └─TableFullScan_27            | 142020.00 | cop[tikv] | table:t1                      | keep
    ↪ order:false                                                                          |
| └─IndexLookUp_9(Probe)          | 1.27      | root      |                               |
    ↪                                                                                      |
|   ├─IndexRangeScan_7(Build)     | 1.27      | cop[tikv] | table:t2, index:t1_id(t1_id) | range
    ↪ : decided by [eq(test.t2.t1_id, test.t1.id)], keep order:false                       |
|   └─TableRowIDScan_8(Probe)     | 1.27      | cop[tikv] | table:t2                      | keep
    ↪ order:false                                                                          |
+--
    ↪ -------------------------------+-----------+-----------+-------------------------------+-----------------
    ↪
6 rows in set (0.00 sec)
```

Index Join 算法对内存消耗较小，但如果需要执行大量探查操作，运行速度可能会慢于其他 Join 算法。以下面这条查询语句为例：

```
SELECT * FROM t1 INNER JOIN t2 ON t1.id=t2.t1_id WHERE t1.pad1 = 'value' and t2.pad1='value';
```

在 Inner Join 操作中，TiDB 会先执行 Join Reorder 算法，所以不能确定会先读取 t1 还是 t2。假设 TiDB 先读取了 t1 来构建 Build 端，那么 TiDB 会在探查 t2 前先根据谓词 t1.col = 'value' 筛选数据，但接下来每次探查 t2 时都要应用谓词 t2.col='value'。所以对于这条语句，Index Join 算法可能不如其他 Join 算法高效。

但如果 Build 端的数据量比 Probe 端小，且 Probe 端的数据已预先建立了索引，那么这种情况下 Index Join 算法效率更高。在下面这段查询语句中，因为 Index Join 比 Hash Join 效率低，所以 SQL 优化器选择了 Hash Join 算法：

```
-- 删除已有索引
ALTER TABLE t2 DROP INDEX t1_id;

EXPLAIN ANALYZE SELECT /*+ INL_JOIN(t1, t2) */  * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id WHERE
    ↪  t1.int_col = 1;
EXPLAIN ANALYZE SELECT /*+ HASH_JOIN(t1, t2) */  * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id
    ↪ WHERE t1.int_col = 1;
EXPLAIN ANALYZE SELECT * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id WHERE t1.int_col = 1;
```

```
Query OK, 0 rows affected (0.29 sec)

+--
    ↪ --------------------------+----------+---------+-----------+--------------+-------------------------
    ↪
| id                          | estRows  | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                                                  | memory
    ↪                 | disk |
+--
    ↪ --------------------------+----------+---------+-----------+--------------+-------------------------
    ↪
```

```
| IndexJoin_13                | 90000.00 | 20000   | root      |               | time:613.19955ms
    ↪ , loops:21, inner:{total:42.494047ms, concurrency:5, task:12, construct:33.149671ms,
    ↪ fetch:9.322956ms, build:8.66µs}, probe:32.435355ms
    ↪
    ↪ | inner join, inner:TableReader_9, outer key:test.t2.t1_id, inner key:test.t1.id |
    ↪ 269.63341903686523 MB | N/A  |
| ├─TableReader_19(Build)     | 90000.00 | 90000   | root      |               | time:586.613252
    ↪ ms, loops:95, cop_task: {num: 3, max: 205.893949ms, min: 185.051354ms, avg: 194.878702ms,
    ↪  p95: 205.893949ms, max_proc_keys: 31715, p95_proc_keys: 31715, tot_proc: 332ms, tot_wait
    ↪ : 4ms, rpc_num: 4, rpc_time: 584.907774ms, copr_cache_hit_ratio: 0.00}, backoff{
    ↪ regionMiss: 2ms} | data:TableFullScan_18
    ↪                                                          | 182.624906539917 MB   | N/A  |
| │ └─TableFullScan_18        | 90000.00 | 90000   | cop[tikv] | table:t2      | time:0ns, loops
    ↪ :0, tikv_task:{proc max:72ms, min:64ms, p80:72ms, p95:72ms, iters:102, tasks:3}
    ↪
    ↪ | keep order:false                                                          | N/A
    ↪                   | N/A  |
| └─TableReader_9(Probe)      | 0.00     | 5       | root      |               | time:8.432051ms
    ↪ , loops:14, cop_task: {num: 14, max: 629.805µs, min: 226.129µs, avg: 420.979µs, p95:
    ↪ 629.805µs, max_proc_keys: 4, p95_proc_keys: 4, rpc_num: 15, rpc_time: 5.953229ms,
    ↪ copr_cache_hit_ratio: 0.00}
    ↪                                                                        | data:
    ↪ Selection_8                                                  | N/A
    ↪                   | N/A  |
|   └─Selection_8             | 0.00     | 5       | cop[tikv] |               | time:0ns, loops
    ↪ :0, tikv_task:{proc max:0s, min:0s, p80:0s, p95:0s, iters:14, tasks:14}
    ↪
    ↪ | eq(test.t1.int_col, 1)                                                    | N/A
    ↪                   | N/A  |
|     └─TableRangeScan_7      | 1.00     | 25      | cop[tikv] | table:t1      | time:0ns, loops
    ↪ :0, tikv_task:{proc max:0s, min:0s, p80:0s, p95:0s, iters:14, tasks:14}
    ↪
    ↪ | range: decided by [test.t2.t1_id], keep order:false                       | N/A
    ↪                   | N/A  |
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------------
    ↪
6 rows in set (0.61 sec)

+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------------
    ↪
| id                          | estRows  | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                | memory                | disk    |
```

```
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------
    ↪
| HashJoin_19                  | 90000.00 | 20000   | root      |               | time:406.098528
    ↪ ms, loops:22, build_hash_table:{total:148.574644ms, fetch:146.843636ms, build:1.731008ms
    ↪ }, probe:{concurrency:5, total:2.026547436s, max:406.039309ms, probe:205.337813ms, fetch
    ↪ :1.821209623s}                                          | inner join, equal:[eq(test.t1.
    ↪ id, test.t2.t1_id)] | 30.00731658935547 MB  | 0 Bytes |
| ├─TableReader_22(Build)      | 71.01    | 10000   | root      |               | time
    ↪ :147.072725ms, loops:12, cop_task: {num: 3, max: 145.847743ms, min: 50.932527ms, avg:
    ↪ 113.009029ms, p95: 145.847743ms, max_proc_keys: 31724, p95_proc_keys: 31724, tot_proc:
    ↪ 284ms, rpc_num: 3, rpc_time: 338.950488ms, copr_cache_hit_ratio: 0.00} | data:
    ↪ Selection_21                                  | 29.679713249206543 MB | N/A     |
| │ └─Selection_21              | 71.01    | 10000   | cop[tikv] |               | time:0ns,
    ↪ loops:0, tikv_task:{proc max:132ms, min:48ms, p80:132ms, p95:132ms, iters:83, tasks:3}
    ↪
    ↪ | eq(test.t1.int_col, 1)                            | N/A                   | N/A     |
| │   └─TableFullScan_20       | 71010.00 | 71010   | cop[tikv] | table:t1      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:128ms, min:48ms, p80:128ms, p95:128ms, iters:83, tasks:3}
    ↪
    ↪ | keep order:false                                  | N/A                   | N/A     |
| └─TableReader_24(Probe)      | 90000.00 | 90000   | root      |               | time
    ↪ :365.918504ms, loops:91, cop_task: {num: 3, max: 398.62145ms, min: 338.460345ms, avg:
    ↪ 358.732721ms, p95: 398.62145ms, max_proc_keys: 31715, p95_proc_keys: 31715, tot_proc: 536
    ↪ ms, rpc_num: 3, rpc_time: 1.076128895s, copr_cache_hit_ratio: 0.00}  | data:
    ↪ TableFullScan_23                              | 182.62489891052246 MB | N/A     |
|   └─TableFullScan_23         | 90000.00 | 90000   | cop[tikv] | table:t2      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:100ms, min:40ms, p80:100ms, p95:100ms, iters:102, tasks:3}
    ↪
    ↪ | keep order:false                                  | N/A                   | N/A     |
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------
    ↪
6 rows in set (0.41 sec)

+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------
    ↪
| id                           | estRows  | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                     | memory                | disk    |
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------
    ↪
| HashJoin_20                  | 90000.00 | 20000   | root      |               | time:441.897092
```

```
    ↪ ms, loops:21, build_hash_table:{total:138.600864ms, fetch:136.353899ms, build:2.246965ms
    ↪ }, probe:{concurrency:5, total:2.207403854s, max:441.850032ms, probe:148.01937ms, fetch
    ↪ :2.059384484s}                                                | inner join, equal:[eq(test.
    ↪ t1.id, test.t2.t1_id)] | 30.00731658935547 MB  | 0 Bytes |
| ├─TableReader_25(Build)      | 71.01    | 10000   | root      |               | time
    ↪ :138.081807ms, loops:12, cop_task: {num: 3, max: 134.702901ms, min: 53.356202ms, avg:
    ↪ 93.372186ms, p95: 134.702901ms, max_proc_keys: 31724, p95_proc_keys: 31724, tot_proc: 236
    ↪ ms, rpc_num: 3, rpc_time: 280.017658ms, copr_cache_hit_ratio: 0.00}   | data:Selection_24
    ↪                                  | 29.680171966552734 MB | N/A     |
| │ └─Selection_24             | 71.01    | 10000   | cop[tikv] |               | time:0ns,
    ↪ loops:0, tikv_task:{proc max:80ms, min:52ms, p80:80ms, p95:80ms, iters:83, tasks:3}
    ↪
    ↪ | eq(test.t1.int_col, 1)                            | N/A                   | N/A     |
| │   └─TableFullScan_23       | 71010.00 | 71010   | cop[tikv] | table:t1      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:80ms, min:52ms, p80:80ms, p95:80ms, iters:83, tasks:3}
    ↪
    ↪ | keep order:false                                  | N/A                   | N/A     |
| └─TableReader_22(Probe)      | 90000.00 | 90000   | root      |               | time
    ↪ :413.560548ms, loops:91, cop_task: {num: 3, max: 432.938474ms, min: 231.263355ms, avg:
    ↪ 365.710741ms, p95: 432.938474ms, max_proc_keys: 31715, p95_proc_keys: 31715, tot_proc:
    ↪ 488ms, rpc_num: 3, rpc_time: 1.097021983s, copr_cache_hit_ratio: 0.00} | data:
    ↪ TableFullScan_21                              | 182.62489891052246 MB | N/A     |
|   └─TableFullScan_21         | 90000.00 | 90000   | cop[tikv] | table:t2      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:80ms, min:80ms, p80:80ms, p95:80ms, iters:102, tasks:3}
    ↪
    ↪ | keep order:false                                  | N/A                   | N/A     |
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------
    ↪
6 rows in set (0.44 sec)
```

在上面所示的 Index Join 操作中，`t1.int_col` 一列的索引被删除了。如果加上这个索引，操作执行速度可以从 `0.61` 秒 提高到 `0.14` 秒，如下表所示：

```
-- 重新添加索引
ALTER TABLE t2 ADD INDEX (t1_id);

EXPLAIN ANALYZE SELECT /*+ INL_JOIN(t1, t2) */  * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id WHERE
    ↪  t1.int_col = 1;
EXPLAIN ANALYZE SELECT /*+ HASH_JOIN(t1, t2) */  * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id
    ↪ WHERE t1.int_col = 1;
EXPLAIN ANALYZE SELECT * FROM t1 INNER JOIN t2 ON t1.id = t2.t1_id WHERE t1.int_col = 1;
```

```
Query OK, 0 rows affected (3.65 sec)
```

```
+--
    ↪ ------------------------------+----------+---------+-----------+---------------------------+---------
    ↪
| id                            | estRows  | actRows | task      | access object
    ↪                | execution info
    ↪
    ↪ | operator info                                                                 |
    ↪ memory               | disk |
+--
    ↪ ------------------------------+----------+---------+-----------+---------------------------+---------
    ↪
| IndexJoin_11                  | 90000.00 | 0       | root      |
    ↪                              | time:136.876686ms, loops:1, inner:{total:114.948158ms,
    ↪ concurrency:5, task:7, construct:5.329114ms, fetch:109.610054ms, build:2.38µs}, probe
    ↪ :1.699799ms
    ↪
    ↪ | inner join, inner:IndexLookUp_10, outer key:test.t1.id, inner key:test.t2.t1_id |
    ↪ 29.864535331726074 MB | N/A  |
| ├─TableReader_32(Build)       | 10000.00 | 10000   | root      |
    ↪                              | time:95.755212ms, loops:12, cop_task: {num: 3, max:
    ↪ 95.652443ms, min: 30.758712ms, avg: 57.545129ms, p95: 95.652443ms, max_proc_keys: 31724,
    ↪ p95_proc_keys: 31724, tot_proc: 124ms, rpc_num: 3, rpc_time: 172.528417ms,
    ↪ copr_cache_hit_ratio: 0.00}                     | data:Selection_31
    ↪                                                                  | 29.679298400878906 MB | N
    ↪ /A  |
| │ └─Selection_31             | 10000.00 | 10000   | cop[tikv] |
    ↪                              | time:0ns, loops:0, tikv_task:{proc max:44ms, min:28ms, p80
    ↪ :44ms, p95:44ms, iters:84, tasks:3}
    ↪
    ↪ | eq(test.t1.int_col, 1)                                                        | N/A
    ↪                   | N/A  |
| │   └─TableFullScan_30        | 71010.00 | 71010   | cop[tikv] | table:t1
    ↪                     | time:0ns, loops:0, tikv_task:{proc max:44ms, min:28ms, p80:44ms,
    ↪ p95:44ms, iters:84, tasks:3}
    ↪
    ↪ | keep order:false                                                              | N/A
    ↪                   | N/A  |
| └─IndexLookUp_10(Probe)       | 9.00     | 0       | root      |
    ↪                              | time:103.93801ms, loops:7
    ↪
    ↪ |                                                                             |
    ↪ 2.1787109375 KB       | N/A  |
|   ├─IndexRangeScan_8(Build)   | 9.00     | 0       | cop[tikv] | table:t2, index:t1_id(t1_id
    ↪ ) | time:0s, loops:0, cop_task: {num: 7, max: 23.969244ms, min: 12.003682ms, avg:
    ↪ 14.659066ms, p95: 23.969244ms, tot_proc: 100ms, rpc_num: 7, rpc_time: 102.435966ms,
```

```
    ↪ copr_cache_hit_ratio: 0.00}, tikv_task:{proc max:24ms, min:12ms, p80:16ms, p95:24ms,
    ↪ iters:7, tasks:7} | range: decided by [eq(test.t2.t1_id, test.t1.id)], keep order:false
    ↪             | N/A                   | N/A  |
|   └─TableRowIDScan_9(Probe)     | 9.00     | 0       | cop[tikv] | table:t2
    ↪                     | time:0ns, loops:0
    ↪
    ↪ | keep order:false                                                          | N/A
    ↪                   | N/A  |
+--
    ↪ ------------------------------+----------+---------+-----------+---------------------------+---------
    ↪
7 rows in set (0.14 sec)

+--
    ↪ ---------------------------+----------+---------+-----------+--------------+-------------------------
    ↪
| id                          | estRows  | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                  | memory                | disk    |
+--
    ↪ ---------------------------+----------+---------+-----------+--------------+-------------------------
    ↪
| HashJoin_31                 | 90000.00 | 0       | root      |               | time:402.263795
    ↪ ms, loops:1, build_hash_table:{total:128.467151ms, fetch:126.871282ms, build:1.595869ms},
    ↪  probe:{concurrency:5, total:2.010969815s, max:402.212295ms, probe:8.924769ms, fetch
    ↪ :2.002045046s}                                          | inner join, equal:[eq(test.
    ↪ t1.id, test.t2.t1_id)] | 29.689788818359375 MB | 0 Bytes |
| ├─TableReader_34(Build)     | 10000.00 | 10000   | root      |               | time
    ↪ :126.765972ms, loops:11, cop_task: {num: 3, max: 126.721293ms, min: 54.375481ms, avg:
    ↪ 84.518849ms, p95: 126.721293ms, max_proc_keys: 31724, p95_proc_keys: 31724, tot_proc: 208
    ↪ ms, rpc_num: 3, rpc_time: 253.478218ms, copr_cache_hit_ratio: 0.00} | data:Selection_33
    ↪                                 | 29.679292678833008 MB | N/A     |
| │ └─Selection_33            | 10000.00 | 10000   | cop[tikv] |               | time:0ns,
    ↪ loops:0, tikv_task:{proc max:72ms, min:56ms, p80:72ms, p95:72ms, iters:84, tasks:3}
    ↪
    ↪ | eq(test.t1.int_col, 1)                          | N/A                   | N/A     |
| │   └─TableFullScan_32      | 71010.00 | 71010   | cop[tikv] | table:t1      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:72ms, min:56ms, p80:72ms, p95:72ms, iters:84, tasks:3}
    ↪
    ↪ | keep order:false                                | N/A                   | N/A     |
| └─TableReader_36(Probe)     | 90000.00 | 90000   | root      |               | time
    ↪ :400.447175ms, loops:90, cop_task: {num: 3, max: 400.838264ms, min: 309.474053ms, avg:
    ↪ 341.01943ms, p95: 400.838264ms, max_proc_keys: 31719, p95_proc_keys: 31719, tot_proc: 528
    ↪ ms, rpc_num: 3, rpc_time: 1.02298055s, copr_cache_hit_ratio: 0.00} | data:
    ↪ TableFullScan_35                          | 182.62786674499512 MB | N/A     |
```

```
|   └─TableFullScan_35          | 90000.00 | 90000    | cop[tikv] | table:t2      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:108ms, min:72ms, p80:108ms, p95:108ms, iters:102, tasks:3}
    ↪
    ↪ | keep order:false                                | N/A                  | N/A     |
+--
    ↪ ---------------------------+----------+---------+-----------+--------------+------------------------------
    ↪
6 rows in set (0.40 sec)

+--
    ↪ ---------------------------+----------+---------+-----------+--------------+------------------------------
    ↪
| id                            | estRows  | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                   | memory               | disk    |
+--
    ↪ ---------------------------+----------+---------+-----------+--------------+------------------------------
    ↪
| HashJoin_32                   | 90000.00 | 0       | root      |               | time:356.282882
    ↪ ms, loops:1, build_hash_table:{total:154.187155ms, fetch:151.259305ms, build:2.92785ms},
    ↪ probe:{concurrency:5, total:1.781087041s, max:356.238312ms, probe:7.406146ms, fetch
    ↪ :1.773680895s}                                                | inner join, equal:[eq(test
    ↪ .t1.id, test.t2.t1_id)] | 29.689788818359375 MB | 0 Bytes |
| ├─TableReader_41(Build)       | 10000.00 | 10000    | root      |               | time
    ↪ :151.190175ms, loops:11, cop_task: {num: 3, max: 151.055697ms, min: 56.214348ms, avg:
    ↪ 96.70463ms, p95: 151.055697ms, max_proc_keys: 31724, p95_proc_keys: 31724, tot_proc: 240
    ↪ ms, rpc_num: 3, rpc_time: 290.019942ms, copr_cache_hit_ratio: 0.00}   | data:Selection_40
    ↪                                    | 29.679292678833008 MB | N/A     |
| │ └─Selection_40              | 10000.00 | 10000    | cop[tikv] |               | time:0ns,
    ↪ loops:0, tikv_task:{proc max:80ms, min:56ms, p80:80ms, p95:80ms, iters:84, tasks:3}
    ↪
    ↪ | eq(test.t1.int_col, 1)                          | N/A                  | N/A     |
| │   └─TableFullScan_39        | 71010.00 | 71010    | cop[tikv] | table:t1      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:80ms, min:56ms, p80:80ms, p95:80ms, iters:84, tasks:3}
    ↪
    ↪ | keep order:false                                | N/A                  | N/A     |
| └─TableReader_43(Probe)       | 90000.00 | 90000    | root      |               | time:354.68523
    ↪ ms, loops:90, cop_task: {num: 3, max: 354.313475ms, min: 328.460762ms, avg: 345.530558ms,
    ↪  p95: 354.313475ms, max_proc_keys: 31719, p95_proc_keys: 31719, tot_proc: 508ms, rpc_num:
    ↪  3, rpc_time: 1.036492374s, copr_cache_hit_ratio: 0.00} | data:TableFullScan_42
    ↪                              | 182.62786102294922 MB | N/A     |
|   └─TableFullScan_42          | 90000.00 | 90000    | cop[tikv] | table:t2      | time:0ns,
    ↪ loops:0, tikv_task:{proc max:84ms, min:64ms, p80:84ms, p95:84ms, iters:102, tasks:3}
    ↪
    ↪ | keep order:false                                | N/A                  | N/A     |
```

```
+--
    ↪ ---------------------------+----------+---------+-----------+--------------+-------------------------
    ↪
6 rows in set (0.36 sec)
```

> 注意：
>
> 在上方示例中，SQL 优化器之所以选择了性能较差的 Hash Join 算法，而不是 Index Join 算法，原因在于查询优化是一个 NP 完全问题，可能会选择不太理想的计划。如果需要频繁调用这个查询，建议通过执行计划管理的方式将 Hint 与 SQL 语句绑定，这样要比在发送给 TiDB 的 SQL 语句中插入 Hint 更容易管理。

Index Join 相关算法

如果使用 Hint INL_JOIN 进行 Index Join 操作，TiDB 会在连接外表之前创建一个中间结果的 Hash Table。TiDB 同样也支持使用 Hint INL_HASH_JOIN 在外表上建 Hash Table。以上所述的 Index Join 相关算法都由 SQL 优化器自动选择。

配置

Index Join 算法的性能受以下系统变量影响：

- tidb_index_join_batch_size（默认值：25000）- index lookup join 操作的 batch 大小。
- tidb_index_lookup_join_concurrency（默认值：4）- 可以并发执行的 index lookup 任务数。

10.3.2.5.2 Hash Join

在 Hash Join 操作中，TiDB 首先读取 Build 端的数据并将其缓存在 Hash Table 中，然后再读取 Probe 端的数据，使用 Probe 端的数据来探查 Hash Table 以获得所需行。与 Index Join 算法相比，Hash Join 要消耗更多内存，但如果需要连接的行数很多，运行速度会比 Index Join 快。TiDB 中的 Hash Join 算子是多线程的，并且可以并发执行。

下面是一个 Hash Join 示例：

```
EXPLAIN SELECT /*+ HASH_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

```
+--
    ↪ ---------------------------+-----------+-----------+---------------+--------------------------------
    ↪
| id                          | estRows   | task      | access object | operator info
    ↪                                 |
+--
    ↪ ---------------------------+-----------+-----------+---------------+--------------------------------
    ↪
| HashJoin_27                 | 142020.00 | root      |               | inner join, equal:[eq(
    ↪ test.t1.id, test.t2.id)] |
```

```
| ├─TableReader_29(Build)     | 142020.00 | root      |               | data:TableFullScan_28
    ↪                          |
| │ └─TableFullScan_28        | 142020.00 | cop[tikv] | table:t1      | keep order:false
    ↪                               |
| └─TableReader_31(Probe)     | 180000.00 | root      |               | data:TableFullScan_30
    ↪                          |
|   └─TableFullScan_30        | 180000.00 | cop[tikv] | table:t2      | keep order:false
    ↪                               |
+--
    ↪ --------------------------+----------+----------+---------------+--------------------------------
    ↪
5 rows in set (0.00 sec)
```

TiDB 会按照以下顺序执行 HashJoin_27 算子：

1. 将 Build 端数据缓存在内存中。
2. 根据缓存数据在 Build 端构造一个 Hash Table。
3. 读取 Probe 端的数据。
4. 使用 Probe 端的数据来探查 Hash Table。
5. 将符合条件的结果返回给用户。

EXPLAIN 返回结果中的 operator info 一列记录了 HashJoin_27 的其他信息，包括该查询是 Inner Join 还是 Outer Join 以及 Join 的条件是什么等。在上面给出的示例中，该查询为 Inner Join，Join 条件是 equal:[eq(test.t1.id, ↪ test.t2.id)]，与查询语句中的 WHERE t1.id = t2.id 部分对应。下面例子中其他几个 Join 算子的 operator info 和此处类似。

运行数据

如果在执行操作时，内存使用超过了tidb_mem_quota_query 规定的值（默认为 1GB），且 oom-use-tmp-storage 的值为 true （默认为 true），那么 TiDB 会尝试使用临时存储，在磁盘上创建 Hash Join 的 Build 端。EXPLAIN ↪ ANALYZE 返回结果中的 execution info 一栏记录了有关内存使用情况等运行数据。下面的例子展示了 tidb_mem_quota_query 的值分别设为 1GB（默认）及 500MB 时，EXPLAIN ANALYZE 的返回结果（当内存配额设为 500MB 时，磁盘用作临时存储区）：

```
EXPLAIN ANALYZE SELECT /*+ HASH_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
SET tidb_mem_quota_query=500 * 1024 * 1024;
EXPLAIN ANALYZE SELECT /*+ HASH_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

```
+--
    ↪ --------------------------+----------+---------+----------+--------------+--------------------------
    ↪
| id                          | estRows   | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                  | memory                | disk    |
+--
    ↪ --------------------------+----------+---------+----------+--------------+--------------------------
    ↪
```

```
| HashJoin_27                 | 142020.00 | 71010   | root      |               | time:647.508572
    ↪ ms, loops:72, build_hash_table:{total:579.254415ms, fetch:566.91012ms, build:12.344295ms
    ↪ }, probe:{concurrency:5, total:3.23315006s, max:647.520113ms, probe:330.884716ms, fetch
    ↪ :2.902265344s}                                          | inner join, equal:[eq(test.
    ↪ t1.id, test.t2.id)] | 209.61642456054688 MB | 0 Bytes |
| ├─TableReader_29(Build)     | 142020.00 | 71010   | root      |               | time
    ↪ :567.088247ms, loops:72, cop_task: {num: 2, max: 569.809411ms, min: 369.67451ms, avg:
    ↪ 469.74196ms, p95: 569.809411ms, max_proc_keys: 39245, p95_proc_keys: 39245, tot_proc: 400
    ↪ ms, rpc_num: 2, rpc_time: 939.447231ms, copr_cache_hit_ratio: 0.00}   | data:
    ↪ TableFullScan_28                          | 210.2100534439087 MB  | N/A     |
| │ └─TableFullScan_28        | 142020.00 | 71010   | cop[tikv] | table:t1      | proc max:64ms,
    ↪  min:48ms, p80:64ms, p95:64ms, iters:79, tasks:2
    ↪
    ↪ | keep order:false                                | N/A                   | N/A     |
| └─TableReader_31(Probe)     | 180000.00 | 90000   | root      |               | time
    ↪ :337.233636ms, loops:91, cop_task: {num: 3, max: 569.790741ms, min: 332.758911ms, avg:
    ↪ 421.543165ms, p95: 569.790741ms, max_proc_keys: 31719, p95_proc_keys: 31719, tot_proc:
    ↪ 500ms, rpc_num: 3, rpc_time: 1.264570696s, copr_cache_hit_ratio: 0.00} | data:
    ↪ TableFullScan_30                          | 267.1126985549927 MB  | N/A     |
|   └─TableFullScan_30        | 180000.00 | 90000   | cop[tikv] | table:t2      | proc max:84ms,
    ↪  min:72ms, p80:84ms, p95:84ms, iters:102, tasks:3
    ↪
    ↪ | keep order:false                                | N/A                   | N/A     |
+--
    ↪ ---------------------------+----------+---------+----------+---------------+-----------------------------
    ↪
5 rows in set (0.65 sec)

Query OK, 0 rows affected (0.00 sec)

+--
    ↪ ---------------------------+----------+---------+----------+---------------+-----------------------------
    ↪
| id                          | estRows   | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                                   | memory                | disk
    ↪                 |
+--
    ↪ ---------------------------+----------+---------+----------+---------------+-----------------------------
    ↪
| HashJoin_27                 | 142020.00 | 71010   | root      |               | time:963.983353
    ↪ ms, loops:72, build_hash_table:{total:775.961447ms, fetch:503.789677ms, build:272.17177ms
    ↪ }, probe:{concurrency:5, total:4.805454793s, max:963.973133ms, probe:922.156835ms, fetch
    ↪ :3.883297958s}                                          | inner join, equal:[eq(test.t1.
    ↪ id, test.t2.id)] | 93.53974533081055 MB  | 210.7459259033203 MB |
```

```
| ├─TableReader_29(Build)     | 142020.00 | 71010   | root      |               | time
    ↪ :504.062018ms, loops:72, cop_task: {num: 2, max: 509.276857ms, min: 402.66386ms, avg:
    ↪ 455.970358ms, p95: 509.276857ms, max_proc_keys: 39245, p95_proc_keys: 39245, tot_proc:
    ↪ 384ms, rpc_num: 2, rpc_time: 911.893237ms, copr_cache_hit_ratio: 0.00}  | data:
    ↪ TableFullScan_28                          | 210.20934200286865 MB | N/A
    ↪ |
| | └─TableFullScan_28        | 142020.00 | 71010   | cop[tikv] | table:t1      | proc max:88ms,
    ↪  min:72ms, p80:88ms, p95:88ms, iters:79, tasks:2
    ↪
    ↪ | keep order:false                                | N/A                   | N/A
    ↪                   |
| └─TableReader_31(Probe)     | 180000.00 | 90000   | root      |               | time
    ↪ :363.058382ms, loops:91, cop_task: {num: 3, max: 412.659191ms, min: 358.489688ms, avg:
    ↪ 391.463008ms, p95: 412.659191ms, max_proc_keys: 31719, p95_proc_keys: 31719, tot_proc:
    ↪ 484ms, rpc_num: 3, rpc_time: 1.174326746s, copr_cache_hit_ratio: 0.00} | data:
    ↪ TableFullScan_30                          | 267.11340618133545 MB | N/A
    ↪ |
|   └─TableFullScan_30        | 180000.00 | 90000   | cop[tikv] | table:t2      | proc max:92ms,
    ↪  min:64ms, p80:92ms, p95:92ms, iters:102, tasks:3
    ↪
    ↪ | keep order:false                                | N/A                   | N/A
    ↪                   |
+--
    ↪ --------------------------+----------+---------+----------+---------------+-------------------------
    ↪
5 rows in set (0.98 sec)
```

配置

Hash Join 算法的性能受以下系统变量影响：

- tidb_mem_quota_query（默认值：1GB）- 如果某条查询的内存消耗超出了配额，TiDB 会尝试将 Hash Join 的 Build 端移到磁盘上以节省内存。
- tidb_hash_join_concurrency（默认值：5）- 可以并发执行的 Hash Join 任务数量。

10.3.2.5.3 Merge Join

Merge Join 是一种特殊的 Join 算法。当两个关联表要 Join 的字段需要按排好的顺序读取时，适用 Merge Join 算法。由于 Build 端和 Probe 端的数据都会读取，这种算法的 Join 操作是流式的，类似“拉链式合并”的高效版。Merge Join 占用的内存要远低于 Hash Join，但 Merge Join 不能并发执行。

下面是一个使用 Merge Join 的例子：

```
EXPLAIN SELECT /*+ MERGE_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

```
+--
    ↪ --------------------------+----------+----------+--------------+--------------------------------
    ↪
```

```
| id                          | estRows    | task      | access object | operator info
    ↪                                      |
+--
    ↪ ---------------------------+-----------+-----------+---------------+-------------------------------
    ↪
| MergeJoin_7                 | 142020.00 | root      |               | inner join, left key:test
    ↪ .t1.id, right key:test.t2.id |
| ├─TableReader_12(Build)     | 180000.00 | root      |               | data:TableFullScan_11
    ↪                                     |
| │ └─TableFullScan_11        | 180000.00 | cop[tikv] | table:t2      | keep order:true
    ↪                                       |
| └─TableReader_10(Probe)     | 142020.00 | root      |               | data:TableFullScan_9
    ↪                                      |
|   └─TableFullScan_9         | 142020.00 | cop[tikv] | table:t1      | keep order:true
    ↪                                       |
+--
    ↪ ---------------------------+-----------+-----------+---------------+-------------------------------
    ↪
5 rows in set (0.00 sec)
```

TiDB 会按照以下顺序执行 Merge Join 算子：

1. 从 Build 端把一个 Join Group 的数据全部读取到内存中。
2. 读取 Probe 端的数据。
3. 将 Probe 端的每行数据与 Build 端的一个完整 Join Group 比较，依次查看是否匹配（除了满足等值条件以外，还有其他非等值条件，这里的“匹配”主要是指查看是否满足非等值条件）。Join Group 指的是所有 Join Key 上值相同的数据。

10.3.2.5.4 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.6 用 EXPLAIN 查看子查询的执行计划

TiDB 会执行多种子查询相关的优化，以提升子查询的执行性能。本文档介绍一些常见子查询的优化方式，以及如何解读 EXPLAIN 语句返回的执行计划信息。

本文档所使用的示例表数据如下：

```
CREATE TABLE t1 (id BIGINT NOT NULL PRIMARY KEY auto_increment, pad1 BLOB, pad2 BLOB, pad3 BLOB,
    ↪ int_col INT NOT NULL DEFAULT 0);
CREATE TABLE t2 (id BIGINT NOT NULL PRIMARY KEY auto_increment, t1_id BIGINT NOT NULL, pad1 BLOB,
    ↪  pad2 BLOB, pad3 BLOB, INDEX(t1_id));
CREATE TABLE t3 (
 id INT NOT NULL PRIMARY KEY auto_increment,
 t1_id INT NOT NULL,
 UNIQUE (t1_id)
);

INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM
    ↪ dual;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024), 0 FROM t1
    ↪  a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
```

```
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t2 SELECT NULL, a.id, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪  t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
UPDATE t1 SET int_col = 1 WHERE pad1 = (SELECT pad1 FROM t1 ORDER BY RAND() LIMIT 1);
INSERT INTO t3 SELECT NULL, id FROM t1 WHERE id < 1000;

SELECT SLEEP(1);
ANALYZE TABLE t1, t2, t3;
```

10.3.2.6.1 Inner join（无 UNIQUE 约束的子查询）

以下示例中，IN 子查询会从表 t2 中搜索一列 ID。为保证语义正确性，TiDB 需要保证 t1_id 列的值具有唯一性。使用 EXPLAIN 可查看到该查询的执行计划去掉重复项并执行 Inner Join 内连接操作：

```
EXPLAIN SELECT * FROM t1 WHERE id IN (SELECT t1_id FROM t2);
```

```
+--
    ↪ --------------------------------+----------+-----------+------------------------------+----------------
    ↪
| id                                | estRows  | task      | access object                |
    ↪ operator info
    ↪
    ↪ |
+--
    ↪ --------------------------------+----------+-----------+------------------------------+----------------
    ↪
| IndexJoin_14                      | 5.00     | root      |                              | inner
    ↪ join, inner:IndexLookUp_13, outer key:test.t2.t1_id, inner key:test.t1.id, equal cond:eq(
    ↪ test.t2.t1_id, test.t1.id) |
| ├─StreamAgg_49(Build)             | 5.00     | root      |                              | group
    ↪  by:test.t2.t1_id, funcs:firstrow(test.t2.t1_id)->test.t2.t1_id
    ↪                                                        |
| │ └─IndexReader_50                | 5.00     | root      |                              | index
    ↪ :StreamAgg_39
    ↪
    ↪ |
| │   └─StreamAgg_39                | 5.00     | cop[tikv] |                              | group
    ↪  by:test.t2.t1_id,
    ↪
    ↪ |
| │     └─IndexFullScan_31          | 50000.00 | cop[tikv] | table:t2, index:t1_id(t1_id) | keep
    ↪ order:true
```

```
    ↪
    ↪ |
| └─IndexLookUp_13(Probe)          | 1.00     | root      |                                |
    ↪
    ↪ |
|   ├─IndexRangeScan_11(Build)     | 1.00     | cop[tikv] | table:t1, index:PRIMARY(id)  | range
    ↪ : decided by [eq(test.t1.id, test.t2.t1_id)], keep order:false
    ↪                                                      |
|   └─TableRowIDScan_12(Probe)     | 1.00     | cop[tikv] | table:t1                     | keep
    ↪ order:false
    ↪
    ↪ |
+--
    ↪ --------------------------------+----------+-----------+------------------------------+----------------
    ↪
8 rows in set (0.00 sec)
```

由上述查询结果可知，TiDB 通过索引连接操作 | IndexJoin_14 将子查询做了连接转化。该执行计划首先在 TiKV 侧通过索引扫描算子 └─IndexFullScan_31 读取 t2.t1_id 列的值，然后由 └─StreamAgg_39 算子的部分任务在 TiKV 中对 t1_id 值进行去重，然后采用 ├─StreamAgg_49(Build) 算子的部分任务在 TiDB 中对 t1_id 值再次进行去重，去重操作由聚合函数 firstrow(test.t2.t1_id) 执行；之后将操作结果与 t1 表的主键相连接，连接条件是 eq(test.t1.id, test.t2.t1_id)。

10.3.2.6.2 Inner join（有 UNIQUE 约束的子查询）

在上述示例中，为了确保 t1_id 值在与表 t1 连接前具有唯一性，需要执行聚合运算。在以下示例中，由于 UNIQUE 约束已能确保 t3.t1_id 列值的唯一：

```
EXPLAIN SELECT * FROM t1 WHERE id IN (SELECT t1_id FROM t3);
```

```
+--
    ↪ --------------------------------+---------+-----------+------------------------------+----------------
    ↪
| id                               | estRows | task      | access object                | operator
    ↪  info
    ↪
    ↪ |
+--
    ↪ --------------------------------+---------+-----------+------------------------------+----------------
    ↪
| IndexJoin_17                     | 1978.13 | root      |                              | inner
    ↪ join, inner:IndexLookUp_16, outer key:test.t3.t1_id, inner key:test.t1.id, equal cond:eq(
    ↪ test.t3.t1_id, test.t1.id) |
| ├─TableReader_44(Build)          | 1978.00 | root      |                               | data:
    ↪ TableFullScan_43
```

```
    ↪
    ↪ |
| | └─TableFullScan_43             | 1978.00 | cop[tikv] | table:t3                    | keep
    ↪ order:false
    ↪
    ↪ |
| └─IndexLookUp_16(Probe)          | 1.00    | root      |                             |
    ↪
    ↪ |
|   ├─IndexRangeScan_14(Build)     | 1.00    | cop[tikv] | table:t1, index:PRIMARY(id) | range:
    ↪ decided by [eq(test.t1.id, test.t3.t1_id)], keep order:false
    ↪                                                     |
|   └─TableRowIDScan_15(Probe)     | 1.00    | cop[tikv] | table:t1                    | keep
    ↪ order:false
    ↪
    ↪ |
+--
    ↪ --------------------------------+---------+-----------+-----------------------------+-----------------
    ↪
6 rows in set (0.01 sec)
```

从语义上看，因为约束保证了 `t3.t1_id` 列值的唯一性，TiDB 可以直接执行 `INNER JOIN` 查询。

10.3.2.6.3 Semi Join（关联查询）

在前两个示例中，通过 HashAgg 聚合操作或通过 `UNIQUE` 约束保证子查询数据的唯一性之后，TiDB 才能够执行 `Inner Join` 操作。这两种连接均使用了 `Index Join`。

下面的例子中，TiDB 优化器则选择了一种不同的执行计划：

```
EXPLAIN SELECT * FROM t1 WHERE id IN (SELECT t1_id FROM t2 WHERE t1_id != t1.int_col);
```

```
+--
    ↪ --------------------------+----------+----------+-----------------------------+-----------------
    ↪
| id                          | estRows   | task      | access object               | operator
    ↪ info
    ↪
    ↪ |
+--
    ↪ --------------------------+----------+----------+-----------------------------+-----------------
    ↪
| MergeJoin_9                 | 45446.40  | root      |                             | semi join,
    ↪  left key:test.t1.id, right key:test.t2.t1_id, other cond:ne(test.t2.t1_id, test.t1.
    ↪ int_col) |
```

```
| ├─IndexReader_24(Build)     | 180000.00 | root      |                                 | index:
    ↪ IndexFullScan_23
    ↪                                                                                          |
| │ └─IndexFullScan_23        | 180000.00 | cop[tikv] | table:t2, index:t1_id(t1_id) | keep
    ↪ order:true
    ↪                                                                                                |
| └─TableReader_22(Probe)     | 56808.00  | root      |                                 | data:
    ↪ Selection_21
    ↪                                                                                                     |
|   └─Selection_21            | 56808.00  | cop[tikv] |                                 | ne(test.
    ↪ t1.id, test.t1.int_col)
    ↪                                                                                  |
|     └─TableFullScan_20      | 71010.00  | cop[tikv] | table:t1                        | keep
    ↪ order:true
    ↪                                                                                                |
+--
    ↪ --------------------------+----------+----------+-----------------------------+-------------------
    ↪
6 rows in set (0.00 sec)
```

由上述查询结果可知，TiDB 执行了 Semi Join。不同于 Inner Join，Semi Join 仅允许右键 (t2.t1_id) 上的第一个值，也就是该操作将去除 Join 算子任务中的重复数据。Join 算法也包含 Merge Join，会按照排序顺序同时从左侧和右侧读取数据，这是一种高效的 Zipper Merge。

可以将原语句视为关联子查询，因为它引入了子查询外的 t1.int_col 列。然而，EXPLAIN 语句的返回结果显示的是关联子查询去关联后的执行计划。条件 t1_id != t1.int_col 会被重写为 t1.id != t1.int_col。TiDB 可以从表 t1 中读取数据并且在 └─Selection_21 中执行此操作，因此这种去关联和重写操作会极大提高执行效率。

10.3.2.6.4 Anti Semi Join（NOT IN 子查询）

在以下示例中，除非子查询中存在 t3.t1_id，否则该查询将（从语义上）返回表 t3 中的所有行：

```
EXPLAIN SELECT * FROM t3 WHERE t1_id NOT IN (SELECT id FROM t1 WHERE int_col < 100);
```

```
+--
    ↪ -------------------------------+---------+-----------+-----------------------------+-------------------
    ↪
| id                              | estRows | task      | access object               | operator
    ↪  info
    ↪
    ↪ |
+--
    ↪ -------------------------------+---------+-----------+-----------------------------+-------------------
    ↪
```

```
| IndexJoin_14                    | 1582.40 | root      |                              | anti
    ↪ semi join, inner:IndexLookUp_13, outer key:test.t3.t1_id, inner key:test.t1.id, equal
    ↪ cond:eq(test.t3.t1_id, test.t1.id) |
| ├─TableReader_35(Build)         | 1978.00 | root      |                              | data:
    ↪ TableFullScan_34
    ↪
    ↪ |
| │ └─TableFullScan_34            | 1978.00 | cop[tikv] | table:t3                     | keep
    ↪ order:false
    ↪
    ↪ |
| └─IndexLookUp_13(Probe)         | 1.00    | root      |                              |
    ↪
    ↪ |
|   ├─IndexRangeScan_10(Build)    | 1.00    | cop[tikv] | table:t1, index:PRIMARY(id) | range:
    ↪ decided by [eq(test.t1.id, test.t3.t1_id)], keep order:false
    ↪                                                          |
|   └─Selection_12(Probe)         | 1.00    | cop[tikv] |                              | lt(test
    ↪ .t1.int_col, 100)
    ↪
    ↪ |
|     └─TableRowIDScan_11         | 1.00    | cop[tikv] | table:t1                     | keep
    ↪ order:false
    ↪
    ↪ |
+--
    ↪ --------------------------------+---------+-----------+------------------------------+-------------------
    ↪
7 rows in set (0.00 sec)
```

上述查询首先读取了表 t3，然后根据主键开始探测 (probe) 表 t1。连接类型是 anti semi join，即反半连接：之所以使用 anti，是因为上述示例有不存在匹配值（即 NOT IN）的情况；使用 Semi Join 则是因为仅需要匹配第一行后就可以停止查询。

10.3.2.6.5 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.7 用 EXPLAIN 查看聚合查询执行计划

SQL 查询中可能会使用聚合计算，可以通过 EXPLAIN 语句来查看聚合查询的执行计划。本文提供多个示例，以帮助用户理解聚合查询是如何执行的。

SQL 优化器会选择以下任一算子实现数据聚合：

- Hash Aggregation
- Stream Aggregation

为了提高查询效率，数据聚合在 Coprocessor 层和 TiDB 层均会执行。现有示例如下：

```
CREATE TABLE t1 (id INT NOT NULL PRIMARY KEY auto_increment, pad1 BLOB, pad2 BLOB, pad3 BLOB);
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM dual;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM t1 a
    ↪ JOIN t1 b JOIN t1 c LIMIT 10000;
SELECT SLEEP(1);
ANALYZE TABLE t1;
```

以上示例创建表格 t1 并插入数据后，再执行 SHOW TABLE REGIONS 语句。从以下 SHOW TABLE REGIONS 的执行结果可知，表 t1 被切分为多个 Region：

```
SHOW TABLE t1 REGIONS;
```

```
+--
    ↪ ---------+--------------+-------------+----------+----------------+-------+------------+--------------
    ↪
| REGION_ID | START_KEY     | END_KEY       | LEADER_ID | LEADER_STORE_ID | PEERS | SCATTERING |
    ↪ WRITTEN_BYTES | READ_BYTES | APPROXIMATE_SIZE(MB) | APPROXIMATE_KEYS |
+--
    ↪ ---------+--------------+-------------+----------+----------------+-------+------------+--------------
    ↪
|        64 | t_64_         | t_64_r_31766 |        65 |               1 | 65    |          0 |
    ↪          1325 |  102033520 |                   98 |            52797 |
|        66 | t_64_r_31766 | t_64_r_63531 |        67 |               1 | 67    |          0 |
    ↪          1325 |   72522521 |                  104 |            78495 |
|        68 | t_64_r_63531 | t_64_r_95296 |        69 |               1 | 69    |          0 |
    ↪          1325 |          0 |                  104 |            95433 |
|         2 | t_64_r_95296 |              |         3 |               1 | 3     |          0 |
    ↪          1501 |          0 |                   81 |            63211 |
+--
    ↪ ---------+--------------+-------------+----------+----------------+-------+------------+--------------
    ↪
4 rows in set (0.00 sec)
```

使用 EXPLAIN 查看以下聚合语句的执行计划。可以看到 └─StreamAgg_8 算子先执行在 TiKV 内每个 Region 上，然后 TiKV 的每个 Region 会返回一行数据给 TiDB，TiDB 在 StreamAgg_16 算子上对每个 Region 返回的数据进行聚合：

```
EXPLAIN SELECT COUNT(*) FROM t1;
```

```
+--
    ↪ -------------------------+----------+----------+--------------+---------------------------------+
    ↪
| id                          | estRows   | task      | access object | operator info
    ↪                   |
+--
    ↪ -------------------------+----------+----------+--------------+---------------------------------+
    ↪
| StreamAgg_16                | 1.00      | root      |               | funcs:count(Column#7)->
    ↪ Column#5 |
| └─TableReader_17            | 1.00      | root      |               | data:StreamAgg_8
    ↪                  |
|   └─StreamAgg_8             | 1.00      | cop[tikv] |               | funcs:count(1)->Column#7
    ↪         |
|     └─TableFullScan_15      | 242020.00 | cop[tikv] | table:t1      | keep order:false
    ↪                  |
```

```
+--
    ↪ -------------------------+----------+----------+--------------+-----------------------------+
    ↪
4 rows in set (0.00 sec)
```

同样，通过执行 EXPLAIN ANALYZE 语句可知，actRows 与 SHOW TABLE REGIONS 返回结果中的 Region 数匹配，这是因为执行使用了 TableFullScan 全表扫并且没有二级索引：

```
EXPLAIN ANALYZE SELECT COUNT(*) FROM t1;
```

```
+--
    ↪ -------------------------+----------+---------+----------+--------------+-------------------------
    ↪
| id                          | estRows    | actRows | task      | access object | execution info
    ↪
    ↪ | operator info                    | memory    | disk |
+--
    ↪ -------------------------+----------+---------+----------+--------------+-------------------------
    ↪
| StreamAgg_16                | 1.00       | 1       | root      |               | time:12.609575ms
    ↪ , loops:2
    ↪
    ↪ | funcs:count(Column#7)->Column#5 | 372 Bytes | N/A  |
| └─TableReader_17            | 1.00       | 4       | root      |               | time:12.605155
    ↪ ms, loops:2, cop_task: {num: 4, max: 12.538245ms, min: 9.256838ms, avg: 10.895114ms, p95:
    ↪  12.538245ms, max_proc_keys: 31765, p95_proc_keys: 31765, tot_proc: 48ms, rpc_num: 4,
    ↪ rpc_time: 43.530707ms, copr_cache_hit_ratio: 0.00} | data:StreamAgg_8                |
    ↪ 293 Bytes | N/A  |
|   └─StreamAgg_8             | 1.00       | 4       | cop[tikv] |               | proc max:12ms,
    ↪ min:12ms, p80:12ms, p95:12ms, iters:122, tasks:4
    ↪
    ↪ | funcs:count(1)->Column#7        | N/A       | N/A  |
|     └─TableFullScan_15      | 242020.00 | 121010  | cop[tikv] | table:t1      | proc max:12ms,
    ↪ min:12ms, p80:12ms, p95:12ms, iters:122, tasks:4
    ↪
    ↪ | keep order:false                | N/A       | N/A  |
+--
    ↪ -------------------------+----------+---------+----------+--------------+-------------------------
    ↪
4 rows in set (0.01 sec)
```

10.3.2.7.1 Hash Aggregation

Hash Aggregation 算法在执行聚合时使用 Hash 表存储中间结果。此算法采用多线程并发优化，执行速度快，但与 Stream Aggregation 算法相比会消耗较多内存。

下面是一个使用 Hash Aggregation（即 HashAgg 算子）的例子：

```
EXPLAIN SELECT /*+ HASH_AGG() */ count(*) FROM t1;
```

```
+--
    ↪ ------------------------+----------+----------+--------------+------------------------------+
    ↪
| id                        | estRows    | task      | access object | operator info
    ↪                    |
+--
    ↪ ------------------------+----------+----------+--------------+------------------------------+
    ↪
| HashAgg_9                 | 1.00       | root      |               | funcs:count(Column#6)->
    ↪ Column#5 |
| └─TableReader_10          | 1.00       | root      |               | data:HashAgg_5
    ↪                    |
|   └─HashAgg_5             | 1.00       | cop[tikv] |               | funcs:count(1)->Column#6
    ↪         |
|     └─TableFullScan_8     | 242020.00 | cop[tikv] | table:t1      | keep order:false
    ↪                  |
+--
    ↪ ------------------------+----------+----------+--------------+------------------------------+
    ↪
4 rows in set (0.00 sec)
```

`operator info` 列显示，用于聚合数据的 Hash 函数为 `funcs:count(1)->Column#6`。

10.3.2.7.2 Stream Aggregation

Stream Aggregation 算法通常会比 Hash Aggregation 算法占用更少的内存。但是此算法要求数据按顺序发送，以便对依次到达的值实现流式数据聚合。

下面是一个使用 Stream Aggregation 的例子：

```
CREATE TABLE t2 (id INT NOT NULL PRIMARY KEY, col1 INT NOT NULL);
INSERT INTO t2 VALUES (1, 9),(2, 3),(3,1),(4,8),(6,3);
EXPLAIN SELECT /*+ STREAM_AGG() */ col1, count(*) FROM t2 GROUP BY col1;
```

```
Query OK, 0 rows affected (0.11 sec)

Query OK, 5 rows affected (0.01 sec)
Records: 5  Duplicates: 0  Warnings: 0

+--
    ↪ --------------------------+----------+----------+--------------+--------------------------------
    ↪
```

```
| id                          | estRows  | task      | access object | operator info
    ↪                                                                          |
+--
    ↪ ---------------------------+----------+-----------+---------------+-------------------------------
    ↪
| Projection_4                | 8000.00  | root      |               | test.t2.col1, Column#3
    ↪                                                                        |
| └─StreamAgg_8               | 8000.00  | root      |               | group by:test.t2.col1,
    ↪ funcs:count(1)->Column#3, funcs:firstrow(test.t2.col1)->test.t2.col1 |
|   └─Sort_13                 | 10000.00 | root      |               | test.t2.col1
    ↪                                                                              |
|     └─TableReader_12        | 10000.00 | root      |               | data:TableFullScan_11
    ↪                                                                                |
|       └─TableFullScan_11    | 10000.00 | cop[tikv] | table:t2      | keep order:false, stats:
    ↪ pseudo                                                                       |
+--
    ↪ ---------------------------+----------+-----------+---------------+-------------------------------
    ↪
5 rows in set (0.00 sec)
```

以上示例中，可以在 col1 上添加索引来消除 └─Sort_13 算子。添加索引后，TiDB 就可以按顺序读取数据并消除 └─Sort_13 算子。

```
ALTER TABLE t2 ADD INDEX (col1);
EXPLAIN SELECT /*+ STREAM_AGG() */ col1, count(*) FROM t2 GROUP BY col1;
```

```
Query OK, 0 rows affected (0.28 sec)

+--
    ↪ ---------------------------+---------+-----------+---------------------------+-----------------------
    ↪
| id                          | estRows | task      | access object             | operator info
    ↪                                                                                 |
+--
    ↪ ---------------------------+---------+-----------+---------------------------+-----------------------
    ↪
| Projection_4                | 4.00    | root      |                           | test.t2.col1,
    ↪  Column#3                                                                       |
| └─StreamAgg_14              | 4.00    | root      |                           | group by:
    ↪ test.t2.col1, funcs:count(Column#4)->Column#3, funcs:firstrow(test.t2.col1)->test.t2.col1
    ↪  |
|   └─IndexReader_15          | 4.00    | root      |                           | index:
    ↪ StreamAgg_8
    ↪                                                                                    |
|     └─StreamAgg_8           | 4.00    | cop[tikv] |                           | group by:
    ↪ test.t2.col1, funcs:count(1)->Column#4
```

```
    ↪                                                                    |
|       └─IndexFullScan_13     | 5.00    | cop[tikv] | table:t2, index:col1(col1) | keep order:
    ↪ true, stats:pseudo                                                                    |
+--
    ↪ ---------------------------+---------+-----------+---------------------------+----------------------
    ↪
5 rows in set (0.00 sec)
```

10.3.2.7.3 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.8 用 EXPLAIN 查看带视图的 SQL 执行计划

EXPLAIN 语句返回的结果会显示视图引用的表和索引，而不是视图本身的名称。这是因为视图是一张虚拟表，本身并不存储任何数据。视图的定义会和查询语句的其余部分在 SQL 优化过程中进行合并。

参考 bikeshare 数据库示例（英文），以下两个示例查询的执行方式类似：

```
ALTER TABLE trips ADD INDEX (duration);
CREATE OR REPLACE VIEW long_trips AS SELECT * FROM trips WHERE duration > 3600;
EXPLAIN SELECT * FROM long_trips;
EXPLAIN SELECT * FROM trips WHERE duration > 3600;
```

```
Query OK, 0 rows affected (2 min 10.11 sec)

Query OK, 0 rows affected (0.13 sec)

+--
    ↪ ------------------------------+------------+-----------+-----------------------------------+--------
    ↪
| id                            | estRows    | task      | access object
    ↪                          | operator info                         |
+--
    ↪ ------------------------------+------------+-----------+-----------------------------------+--------
    ↪
| IndexLookUp_12                | 6372547.67 | root      |
    ↪                                        |                                       |
| ├─IndexRangeScan_10(Build)    | 6372547.67 | cop[tikv] | table:trips, index:duration(duration
    ↪ ) | range:(3600,+inf], keep order:false |
```

```
| └─TableRowIDScan_11(Probe)     | 6372547.67 | cop[tikv] | table:trips
    ↪                            | keep order:false                    |
+--
    ↪ ----------------------------+------------+-----------+-------------------------------------+--------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ----------------------------+-----------+-----------+-------------------------------------+----------
    ↪
| id                            | estRows    | task      | access object                       |
    ↪  operator info                          |
+--
    ↪ ----------------------------+-----------+-----------+-------------------------------------+----------
    ↪
| IndexLookUp_10                | 833219.37 | root      |                                     |
    ↪                                         |
| ├─IndexRangeScan_8(Build)     | 833219.37 | cop[tikv] | table:trips, index:duration(duration)
    ↪ | range:(3600,+inf], keep order:false |
| └─TableRowIDScan_9(Probe)     | 833219.37 | cop[tikv] | table:trips
    ↪ | keep order:false                    |
+--
    ↪ ----------------------------+-----------+-----------+-------------------------------------+----------
    ↪
3 rows in set (0.00 sec)
```

同样，该视图中的谓词被下推至基表：

```
EXPLAIN SELECT * FROM long_trips WHERE bike_number = 'W00950';
EXPLAIN SELECT * FROM trips WHERE bike_number = 'W00950';
```

```
+--
    ↪ ----------------------------+---------+-----------+-------------------------------------+------------
    ↪
| id                            | estRows | task      | access object                       |
    ↪ operator info                           |
+--
    ↪ ----------------------------+---------+-----------+-------------------------------------+------------
    ↪
| IndexLookUp_14                | 3.33    | root      |                                     |
    ↪                                         |
| ├─IndexRangeScan_11(Build)    | 3333.33 | cop[tikv] | table:trips, index:duration(duration) |
    ↪  range:(3600,+inf], keep order:false, stats:pseudo |
| └─Selection_13(Probe)         | 3.33    | cop[tikv] |                                     |
    ↪  eq(bikeshare.trips.bike_number, "W00950")        |
```

```
|   └─TableRowIDScan_12          | 3333.33 | cop[tikv] | table:trips                          |
    ↪  keep order:false, stats:pseudo                    |
+--
    ↪ ------------------------------+---------+-----------+--------------------------------------+-----------
    ↪
4 rows in set (0.00 sec)

+--
    ↪ ----------------------+-------------+-----------+--------------+-------------------------------------
    ↪
| id                      | estRows     | task      | access object | operator info
    ↪                            |
+--
    ↪ ----------------------+-------------+-----------+--------------+-------------------------------------
    ↪
| TableReader_7           | 43.00       | root      |               | data:Selection_6
    ↪                           |
| └─Selection_6           | 43.00       | cop[tikv] |               | eq(bikeshare.trips.
    ↪ bike_number, "W00950") |
|   └─TableFullScan_5     | 19117643.00 | cop[tikv] | table:trips   | keep order:false
    ↪                           |
+--
    ↪ ----------------------+-------------+-----------+--------------+-------------------------------------
    ↪
3 rows in set (0.00 sec)
```

执行以上第一条语句时使用了索引，满足视图定义，接着在 TiDB 读取行时应用了 `bike_number = 'W00950'` 条件。执行以上第二条语句时，不存在满足该语句的索引，因此使用了 `TableFullScan`。

TiDB 使用的索引可以同时满足视图定义和语句本身，如以下组合索引所示：

```
ALTER TABLE trips ADD INDEX (bike_number, duration);
EXPLAIN SELECT * FROM long_trips WHERE bike_number = 'W00950';
EXPLAIN SELECT * FROM trips WHERE bike_number = 'W00950';
```

```
Query OK, 0 rows affected (2 min 31.20 sec)

+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------
    ↪
| id                            | estRows  | task      | access object
    ↪                                       | operator info
    ↪                                       |
+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------
    ↪
```

```
| IndexLookUp_13                 | 63725.48 | root      |
    ↪                                                    |
    ↪                                                    |
| ├─IndexRangeScan_11(Build)     | 63725.48 | cop[tikv] | table:trips, index:bike_number(
    ↪ bike_number, duration) | range:("W00950" 3600,"W00950" +inf], keep order:false |
| └─TableRowIDScan_12(Probe)     | 63725.48 | cop[tikv] | table:trips
    ↪                                          | keep order:false
    ↪                                      |
+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------------------
    ↪
3 rows in set (0.00 sec)

+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------------------
    ↪
| id                             | estRows  | task      | access object
    ↪                                          | operator info                                 |
+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------------------
    ↪
| IndexLookUp_10                 | 19117.64 | root      |
    ↪                                                    |
    ↪                                               |
| ├─IndexRangeScan_8(Build)      | 19117.64 | cop[tikv] | table:trips, index:bike_number(
    ↪ bike_number, duration) | range:["W00950","W00950"], keep order:false |
| └─TableRowIDScan_9(Probe)      | 19117.64 | cop[tikv] | table:trips
    ↪                                          | keep order:false                              |
+--
    ↪ ------------------------------+----------+-----------+-------------------------------------------------------
    ↪
3 rows in set (0.00 sec)
```

在第一条语句中，TiDB 能够使用组合索引的两个部分 (bike_number, duration)。在第二条语句，TiDB 仅使用了索引 (bike_number, duration) 的第一部分 bike_number。

10.3.2.8.1 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 分区查询的执行计划
- 索引合并查询的执行计划

10.3.2.9 用 EXPLAIN 查看分区查询的执行计划

使用 EXPLAIN 语句可以查看 TiDB 在执行查询时需要访问的分区。由于存在分区裁剪，所显示的分区通常只是所有分区的一个子集。本文档介绍了常见分区表的一些优化方式，以及如何解读 EXPLAIN 语句返回的执行计划信息。

本文档所使用的示例数据如下：

```
CREATE TABLE t1 (
 id BIGINT NOT NULL auto_increment,
 d date NOT NULL,
 pad1 BLOB,
 pad2 BLOB,
 pad3 BLOB,
 PRIMARY KEY (id,d)
) PARTITION BY RANGE (YEAR(d)) (
 PARTITION p2016 VALUES LESS THAN (2017),
 PARTITION p2017 VALUES LESS THAN (2018),
 PARTITION p2018 VALUES LESS THAN (2019),
 PARTITION p2019 VALUES LESS THAN (2020),
 PARTITION pmax VALUES LESS THAN MAXVALUE
);

INSERT INTO t1 (d, pad1, pad2, pad3) VALUES
 ('2016-01-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2016-06-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2016-09-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2017-01-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2017-06-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2017-09-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2018-01-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2018-06-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2018-09-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2019-01-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2019-06-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2019-09-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2020-01-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2020-06-01', RANDOM_BYTES(102), RANDOM_BYTES(1024), RANDOM_BYTES(1024)),
 ('2020-09-01', RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024));

INSERT INTO t1 SELECT NULL, a.d, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪ t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, a.d, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪ t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, a.d, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
    ↪ t1 a JOIN t1 b JOIN t1 c LIMIT 10000;
INSERT INTO t1 SELECT NULL, a.d, RANDOM_BYTES(1024), RANDOM_BYTES(1024), RANDOM_BYTES(1024) FROM
```

```
    ↪ t1 a JOIN t1 b JOIN t1 c LIMIT 10000;

SELECT SLEEP(1);
ANALYZE TABLE t1;
```

以下示例解释了基于新建分区表 t1 的一条语句：

```
EXPLAIN SELECT COUNT(*) FROM t1 WHERE d = '2017-06-01';
```

```
+--
    ↪ ---------------------------+---------+-----------+--------------------------+----------------------------
    ↪
| id                           | estRows | task      | access object            | operator info
    ↪                              |
+--
    ↪ ---------------------------+---------+-----------+--------------------------+----------------------------
    ↪
| StreamAgg_21                 | 1.00    | root      |                          | funcs:count(
    ↪ Column#8)->Column#6           |
| └─TableReader_22             | 1.00    | root      |                          | data:
    ↪ StreamAgg_10                    |
|   └─StreamAgg_10             | 1.00    | cop[tikv] |                          | funcs:count
    ↪ (1)->Column#8                 |
|     └─Selection_20           | 8.87    | cop[tikv] |                          | eq(test.t1.d,
    ↪  2017-06-01 00:00:00.000000) |
|       └─TableFullScan_19     | 8870.00 | cop[tikv] | table:t1, partition:p2017 | keep order:
    ↪ false                         |
+--
    ↪ ---------------------------+---------+-----------+--------------------------+----------------------------
    ↪
5 rows in set (0.01 sec)
```

由上述 EXPLAIN 结果可知，从最末尾的 └─TableFullScan_19 算子开始，再返回到根部的 StreamAgg_21 算子的执行过程如下：

- TiDB 成功地识别出只需要访问一个分区 (p2017)，并将该信息在 access object 列中注明。
- └─TableFullScan_19 算子先对整个分区进行扫描，然后执行 └─Selection_20 算子筛选起始日期为 2017-06-01 00:00:00.000000 的行。
- 之后，└─Selection_20 算子匹配的行在 Coprocessor 中进行流式聚合，Coprocessor 本身就可以理解聚合函数 count。
- 每个 Coprocessor 请求会发送一行数据给 TiDB 的 └─TableReader_22 算子，然后将数据在 StreamAgg_21 算子下进行流式聚合，再将一行数据返回给客户端。

以下示例中，分区裁剪不会消除任何分区：

```
EXPLAIN SELECT COUNT(*) FROM t1 WHERE YEAR(d) = 2017;
```

```
+--
    ↪ --------------------------------+----------+----------+--------------------------+-----------------
    ↪
| id                                | estRows  | task      | access object             |
    ↪ operator info                     |
+--
    ↪ --------------------------------+----------+----------+--------------------------+-----------------
    ↪
| HashAgg_20                          | 1.00     | root      |                           | funcs:
    ↪ count(Column#7)->Column#6  |
| └─PartitionUnion_21                 | 5.00     | root      |                           |
    ↪                                   |
|   ├─StreamAgg_36                    | 1.00     | root      |                           | funcs:
    ↪ count(Column#9)->Column#7  |
|   │ └─TableReader_37                | 1.00     | root      |                           | data:
    ↪ StreamAgg_25                |
|   │   └─StreamAgg_25                | 1.00     | cop[tikv] |                           | funcs:
    ↪ count(1)->Column#9         |
|   │     └─Selection_35              | 6000.00  | cop[tikv] |                           | eq(
    ↪ year(test.t1.d), 2017)        |
|   │       └─TableFullScan_34        | 7500.00  | cop[tikv] | table:t1, partition:p2016 | keep
    ↪ order:false                 |
|   ├─StreamAgg_55                    | 1.00     | root      |                           | funcs:
    ↪ count(Column#11)->Column#7 |
|   │ └─TableReader_56                | 1.00     | root      |                           | data:
    ↪ StreamAgg_44                |
|   │   └─StreamAgg_44                | 1.00     | cop[tikv] |                           | funcs:
    ↪ count(1)->Column#11        |
|   │     └─Selection_54              | 14192.00 | cop[tikv] |                           | eq(
    ↪ year(test.t1.d), 2017)        |
|   │       └─TableFullScan_53        | 17740.00 | cop[tikv] | table:t1, partition:p2017 | keep
    ↪ order:false                 |
|   ├─StreamAgg_74                    | 1.00     | root      |                           | funcs:
    ↪ count(Column#13)->Column#7 |
|   │ └─TableReader_75                | 1.00     | root      |                           | data:
    ↪ StreamAgg_63                |
|   │   └─StreamAgg_63                | 1.00     | cop[tikv] |                           | funcs:
    ↪ count(1)->Column#13        |
|   │     └─Selection_73              | 3977.60  | cop[tikv] |                           | eq(
    ↪ year(test.t1.d), 2017)        |
|   │       └─TableFullScan_72        | 4972.00  | cop[tikv] | table:t1, partition:p2018 | keep
    ↪ order:false                 |
|   ├─StreamAgg_93                    | 1.00     | root      |                           | funcs:
    ↪ count(Column#15)->Column#7 |
```

```
|   |   └─TableReader_94               | 1.00     | root      |                              | data:
    ↪ StreamAgg_82                |
|   |     └─StreamAgg_82               | 1.00     | cop[tikv] |                              | funcs:
    ↪ count(1)->Column#15         |
|   |       └─Selection_92             | 20361.60 | cop[tikv] |                              | eq(
    ↪ year(test.t1.d), 2017)         |
|   |         └─TableFullScan_91       | 25452.00 | cop[tikv] | table:t1, partition:p2019    | keep
    ↪ order:false                 |
|   └─StreamAgg_112                    | 1.00     | root      |                              | funcs:
    ↪ count(Column#17)->Column#7  |
|     └─TableReader_113                | 1.00     | root      |                              | data:
    ↪ StreamAgg_101               |
|       └─StreamAgg_101                | 1.00     | cop[tikv] |                              | funcs:
    ↪ count(1)->Column#17         |
|         └─Selection_111              | 8892.80  | cop[tikv] |                              | eq(
    ↪ year(test.t1.d), 2017)         |
|           └─TableFullScan_110        | 11116.00 | cop[tikv] | table:t1, partition:pmax     | keep
    ↪ order:false                 |
+--
    ↪ ---------------------------------+----------+-----------+---------------------------+----------------
    ↪
27 rows in set (0.00 sec)
```

由上述 EXPLAIN 结果可知：

- TiDB 认为需要访问所有分区 (p2016..pMax)。这是因为 TiDB 将谓词 YEAR (d) = 2017 视为 non-sargable。这个问题并非是 TiDB 特有的。
- 在扫描每个分区时，Selection 算子将筛选出年份不为 2017 的行。
- 在每个分区上会执行流式聚合，以计算匹配的行数。
- └─PartitionUnion_21 算子会合并访问每个分区后的结果。

10.3.2.9.1 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 索引合并查询的执行计划

10.3.2.10 用 EXPLAIN 查看索引合并的 SQL 执行计划

索引合并是从 TiDB v4.0 起引入的一种新的表访问方式。在这种访问方式下，TiDB 优化器可以选择对一张表使用多个索引，并将每个索引的返回结果进行合并。在某些场景下，这种访问方式能够减少大量不必要的数据扫描，提升查询的执行效率。

```
EXPLAIN SELECT * FROM t WHERE a = 1 OR b = 1;
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| id                     | estRows  | task      | access object | operator info
    ↪                        |
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| TableReader_7          | 8000.00  | root      |               | data:Selection_6
    ↪                       |
| └─Selection_6           | 8000.00  | cop[tikv] |               | or(eq(test.t.a, 1), eq(test.t
    ↪ .b, 1)) |
|   └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo       |
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
EXPLAIN SELECT /*+ USE_INDEX_MERGE(t) */ * FROM t WHERE a > 1 OR b > 1;
+--
    ↪ ----------------------------+---------+-----------+-------------------------+-----------------------
    ↪
| id                           | estRows | task      | access object           | operator info
    ↪                                |
+--
    ↪ ----------------------------+---------+-----------+-------------------------+-----------------------
    ↪
| IndexMerge_16                | 6666.67 | root      |                         |
    ↪                                                |
| ├─IndexRangeScan_13(Build)     | 3333.33 | cop[tikv] | table:t, index:idx_a(a) | range:(1,+inf
    ↪ ], keep order:false, stats:pseudo |
| ├─IndexRangeScan_14(Build)     | 3333.33 | cop[tikv] | table:t, index:idx_b(b) | range:(1,+inf
    ↪ ], keep order:false, stats:pseudo |
| └─TableRowIDScan_15(Probe)     | 6666.67 | cop[tikv] | table:t                 | keep order:
    ↪ false, stats:pseudo                |
+--
    ↪ ----------------------------+---------+-----------+-------------------------+-----------------------
    ↪
```

例如，在上述示例中，过滤条件是使用 OR 条件连接的 WHERE 子句。在启用索引合并前，每个表只能使用一个索引，不能将 a = 1 下推到索引 a，也不能将 b = 1 下推到索引 b。当 t 中存在大量数据时，全表扫描的效率会很低。针对这类场景，TiDB 引入了对表的新访问方式：索引合并。

在索引合并访问方式下，优化器可以选择对一张表使用多个索引，并将每个索引的返回结果进行合并，生成以上示例中后一个执行计划。此时的 IndexMerge_16 算子有三个子节点，其中 IndexRangeScan_13 和

IndexRangeScan_14 根据范围扫描得到符合条件的所有 RowID，再由 TableRowIDScan_15 算子根据这些 RowID 精确地读取所有满足条件的数据。

其中对于 IndexRangeScan/TableRangeScan 一类按范围进行的扫表操作，EXPLAIN 表中 operator info 列相比于其他扫表操作，多了被扫描数据的范围这一信息。比如上面的例子中，IndexRangeScan_13 算子中的 range:(1,+inf] 这一信息表示该算子扫描了从 1 到正无穷这个范围的数据。

> 注意：
>
> - TiDB 的索引合并特性在 v5.4.0 及之后的版本默认开启，即tidb_enable_index_merge 为 ON。
> - 如果查询中使用了 SQL 优化器 Hint USE_INDEX_MERGE ，无论 tidb_enable_index_merge 开关是否开启，都会强制使用索引合并特性。当过滤条件中有无法下推的表达式时，必须使用 Hint USE_INDEX_MERGE 才能开启索引合并。
> - 索引合并目前仅支持析取范式（or 连接的表达式），不支持合取范式（and 连接的表达式）。
> - 索引合并目前无法在临时表上使用。

10.3.2.10.1 其他类型查询的执行计划

- MPP 模式查询的执行计划
- 索引查询的执行计划
- Join 查询的执行计划
- 子查询的执行计划
- 聚合查询的执行计划
- 视图查询的执行计划
- 分区查询的执行计划

10.3.3 SQL 优化流程

10.3.3.1 SQL 优化流程简介

在 TiDB 中，从输入的查询文本到最终的执行计划执行结果的过程可以见下图。

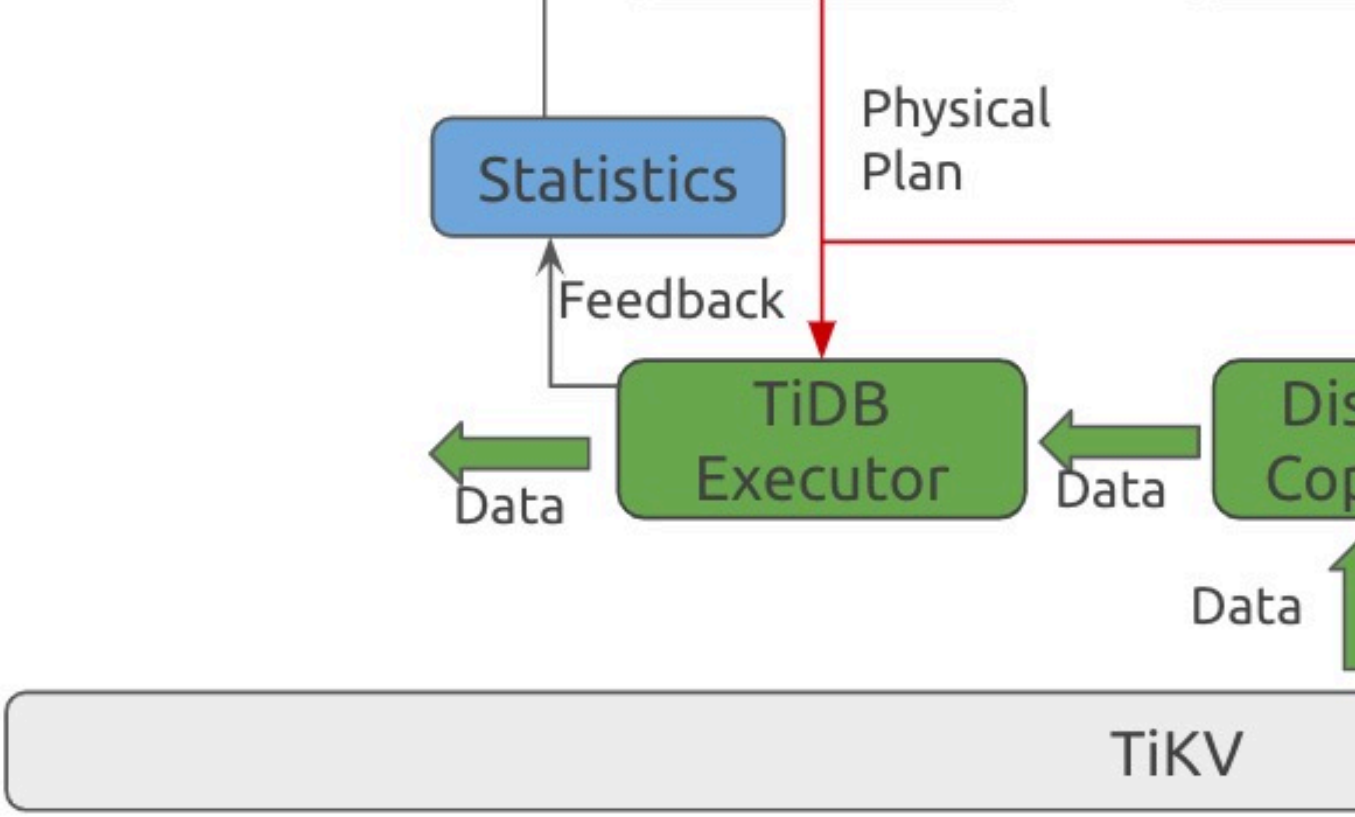


图 207: SQL Optimization

在经过了 parser 对原始查询文本的解析以及一些简单的合法性验证后，TiDB 首先会对查询做一些逻辑上的等价变化，详细的变化可以查询逻辑优化章节。

通过这些等价变化，使得这个查询在逻辑执行计划上可以变得更易于处理。在等价变化结束之后，TiDB 会得到一个与原始查询等价的查询计划结构，之后根据数据分布、以及一个算子具体的执行开销，来获得一个最终的执行计划，这部分内容可以查询物理优化章节。

同时，TiDB 在执行 PREPARE 语句时，可以选择开启缓存来降低 TiDB 生成执行计划的开销，这部分内容会在执行计划缓存一节中介绍。

10.3.3.2 逻辑优化

10.3.3.2.1 逻辑优化

本章节将对一些比较关键的逻辑改写进行说明，帮助大家理解 TiDB 如何生成最终的查询计划。比如在 TiDB 输入 `select * from t where t.a in (select t1.a from t1 where t1.b=t.b)` 这个查询时，在最终的执行计划中将看不到这个 `t.a in (select t1.a from t1 where t1.b=t.b)` 这个 IN 子查询的存在，这便是因为 TiDB 对这里进行了一些改写。

本章节会介绍如下几个关键改写：

- 子查询相关的优化
- 列裁剪
- 关联子查询去关联
- Max/Min 消除

- 谓词下推
- 分区裁剪
- TopN 和 Limit 下推
- Join Reorder

10.3.3.2.2 子查询相关的优化

本文主要介绍子查询相关的优化。

通常会遇到如下情况的子查询：

- `NOT IN (SELECT ... FROM ...)`
- `NOT EXISTS (SELECT ... FROM ...)`
- `IN (SELECT ... FROM ..)`
- `EXISTS (SELECT ... FROM ...)`
- `... >/>=/</<=/=/!= (SELECT ... FROM ...)`

有时，子查询中包含了非子查询中的列，如 `select * from t where t.a in (select * from t2 where t.b =t2.b)` 中，子查询中的 `t.b` 不是子查询中的列，而是从子查询外面引入的列。这种子查询通常会被称为关联子查询，外部引入的列会被称为关联列，关联子查询相关的优化参见关联子查询去关联。本文主要关注不涉及关联列的子查询。

子查询默认会以理解 TiDB 执行计划中提到的 `semi join` 作为默认的执行方式，同时对于一些特殊的子查询，TiDB 会做一些逻辑上的替换使得查询可以获得更好的执行性能。

`... < ALL (SELECT ... FROM ...)` 或者 `... > ANY (SELECT ... FROM ...)`

对于这种情况，可以将 `ALL` 或者 `ANY` 用 `MAX` 以及 `MIN` 来代替。不过由于在表为空时，`MAX(EXPR)` 以及 `MIN(EXPR)` 的结果会为 `NULL`，其表现形式和 `EXPR` 是有 `NULL` 值的结果一样。以及外部表达式结果为 `NULL` 时也会影响表达式的最终结果，因此这里完整的改写会是如下的形式：

- `t.id < all(select s.id from s)` 会被改写为 `t.id < min(s.id)and if(sum(s.id is null)!= 0, null, true)`。
- `t.id < any (select s.id from s)` 会被改写为 `t.id < max(s.id)or if(sum(s.id is null)!= 0, null, false)`。

`... != ANY (SELECT ... FROM ...)`

对于这种情况，当子查询中不同值的个数只有一种的话，那只要和这个值对比就即可。如果子查询中不同值的个数多于 1 个，那么必然会有不相等的情况出现。因此这样的子查询可以采取如下的改写手段：

- `select * from t where t.id != any (select s.id from s)` 会被改写为 `select t.* from t, (select s.id, count(distinct s.id)as cnt_distinct from s)where (t.id != s.id or cnt_distinct > 1)`

`... = ALL (SELECT ... FROM ...)`

对于这种情况，当子查询中不同值的个数多于一种的话，那么这个表达式的结果必然为假。因此这样的子查询在 TiDB 中会改写为如下的形式：

- select * from t where t.id = all (select s.id from s) 会被改写为 select t.* from t, (select s .id, count(distinct s.id)as cnt_distinct from s)where (t.id = s.id and cnt_distinct <= 1)

... IN (SELECT ... FROM ...)

对于这种情况，会将其 IN 的子查询改写为 SELECT ... FROM ... GROUP ... 的形式，然后将 IN 改写为普通的 JOIN 的形式。如 select * from t1 where t1.a in (select t2.a from t2) 会被改写为 select t1.* from t1 , (select distinct(a)a from t2)t2 where t1.a = t2.a 的形式。同时这里的 DISTINCT 可以在 t2.a 具有 UNIQUE 属性时被自动消去。

```
explain select * from t1 where t1.a in (select t2.a from t2);
```

```
+--
    ↪ ---------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
| id                           | estRows | task      | access object         | operator info
    ↪                                                                  |
+--
    ↪ ---------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
| IndexJoin_12                 | 9990.00 | root      |                       | inner join, inner
    ↪ :TableReader_11, outer key:test.t2.a, inner key:test.t1.a |
| ├─HashAgg_21(Build)          | 7992.00 | root      |                       | group by:test.t2
    ↪ .a, funcs:firstrow(test.t2.a)->test.t2.a                  |
| │ └─IndexReader_28           | 9990.00 | root      |                       | index:
    ↪ IndexFullScan_27                                                 |
| │   └─IndexFullScan_27       | 9990.00 | cop[tikv] | table:t2, index:idx(a) | keep order:false
    ↪ , stats:pseudo                                            |
| └─TableReader_11(Probe)      | 1.00    | root      |                       | data:
    ↪ TableRangeScan_10                                                |
|   └─TableRangeScan_10        | 1.00    | cop[tikv] | table:t1              | range: decided
    ↪ by [test.t2.a], keep order:false, stats:pseudo             |
+--
    ↪ ---------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
```

这个改写会在 IN 子查询相对较小，而外部查询相对较大时产生更好的执行性能。因为不经过改写的情况下，我们无法使用以 t2 为驱动表的 index join。同时这里的弊端便是，当改写删成的聚合无法被自动消去且 t2 表比较大时，反而会影响查询的性能。目前 TiDB 中使用tidb_opt_insubq_to_join_and_agg 变量来控制这个优化的打开与否。当遇到不合适这个优化的情况可以手动关闭。

EXISTS 子查询以及 ... >/>=/</<=/=/!= (SELECT ... FROM ...)

当前对于这种场景的子查询，当它不是关联子查询时，TiDB 会在优化阶段提前展开它，将其直接替换为一个结果集直接判断结果。如下图中，EXISTS 会提前在优化阶段被执行为 TRUE，从而不会在最终的执行结果中看到它。

```
create table t1(a int);
```

```
create table t2(a int);
insert into t2 values(1);
explain select * from t where exists (select * from t2);
```

```
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| id                    | estRows  | task      | access object | operator info
    ↪ |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| TableReader_12          | 10000.00 | root      |               | data:TableFullScan_11
    ↪ |
| └─TableFullScan_11      | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:pseudo
    ↪  |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
```

10.3.3.2.3 列裁剪

列裁剪的基本思想在于：对于算子中实际用不上的列，优化器在优化的过程中没有必要保留它们。对这些列的删除会减少 I/O 资源占用，并为后续的优化带来便利。下面给出一个列重复的例子：

假设表 t 里面有 a b c d 四列，执行如下语句：

```
select a from t where b > 5
```

在该查询的过程中，t 表实际上只有 a, b 两列会被用到，而 c, d 的数据则显得多余。对应到该语句的查询计划，Selection 算子会用到 b 列，下面接着的 DataSource 算子会用到 a, b 两列，而剩下 c, d 两列则都可以裁剪掉，DataSource 算子在读数据时不需要将它们读进来。

出于上述考量，TiDB 会在逻辑优化阶段进行自上而下的扫描，裁剪不需要的列，减少资源浪费。该扫描过程称作“列裁剪”，对应逻辑优化规则中的 `columnPruner`。如果要关闭这个规则，可以在参照优化规则及表达式下推的黑名单中的关闭方法。

10.3.3.2.4 关联子查询去关联

子查询相关的优化中介绍了当没有关联列时，TiDB 是如何处理子查询的。由于为关联子查询解除关联依赖比较复杂，本文档中会介绍一些简单的场景以及这个优化规则的适用范围。

简介

以 `select * from t1 where t1.a < (select sum(t2.a)from t2 where t2.b = t1.b)` 为例，这里子查询 `t1.a ↪ < (select sum(t2.a)from t2 where t2.b = t1.b)` 中涉及了关联列上的条件 `t2.b=t1.b`，不过恰好由于

这是一个等值条件，因此可以将其等价的改写为 `select t1.* from t1, (select b, sum(a)sum_a from t2 group by b)t2 where t1.b = t2.b and t1.a < t2.sum_a;`。这样，一个关联子查询就被重新改写为 JOIN 的形式。

TiDB 之所以要进行这样的改写，是因为关联子查询每次子查询执行时都是要和它的外部查询结果绑定的。在上面的例子中，如果 `t1.a` 有一千万个值，那这个子查询就要被重复执行一千万次，因为 `t2.b=t1.b` 这个条件会随着 `t1.a` 值的不同而发生变化。当通过一些手段将关联依赖解除后，这个子查询就只需要被执行一次了。

限制

这种改写的弊端在于，在关联没有被解除时，优化器是可以使用关联列上的索引的。也就是说，虽然这个子查询可能被重复执行多次，但是每次都可以使用索引过滤数据。而解除关联的变换上，通常是会导致关联列的位置发生改变而导致虽然子查询只被执行了一次，但是单次执行的时间会比没有解除关联时的单次执行时间长。

因此，在外部的值比较少的情况下，不解除关联依赖反而可能对执行性能更有帮助。这时可以通过优化规则及表达式下推的黑名单中关闭子查询去关联优化规则的方式来关闭这个优化。

样例

```
create table t1(a int, b int);
create table t2(a int, b int, index idx(b));
explain select * from t1 where t1.a < (select sum(t2.a) from t2 where t2.b = t1.b);
```

```
+--
    ↪ --------------------------------+----------+-----------+---------------+--------------------------------
    ↪
| id                              | estRows  | task      | access object | operator info
    ↪                                                                                 |
+--
    ↪ --------------------------------+----------+-----------+---------------+--------------------------------
    ↪
| HashJoin_11                     | 9990.00  | root      |               | inner join, equal:[eq
    ↪ (test.t1.b, test.t2.b)], other cond:lt(cast(test.t1.a), Column#7)  |
| ├─HashAgg_23(Build)             | 7992.00  | root      |               | group by:test.t2.b,
    ↪ funcs:sum(Column#8)->Column#7, funcs:firstrow(test.t2.b)->test.t2.b |
| │ └─TableReader_24              | 7992.00  | root      |               | data:HashAgg_16
    ↪                                                                        |
| │   └─HashAgg_16                | 7992.00  | cop[tikv] |               | group by:test.t2.b,
    ↪ funcs:sum(test.t2.a)->Column#8                                        |
| │     └─Selection_22            | 9990.00  | cop[tikv] |               | not(isnull(test.t2.b
    ↪ ))                                                                    |
| │       └─TableFullScan_21      | 10000.00 | cop[tikv] | table:t2      | keep order:false,
    ↪ stats:pseudo                                                            |
| └─TableReader_15(Probe)         | 9990.00  | root      |               | data:Selection_14
    ↪                                                                         |
|   └─Selection_14                | 9990.00  | cop[tikv] |               | not(isnull(test.t1.b
    ↪ ))                                                                    |
```

```
|     └─TableFullScan_13           | 10000.00 | cop[tikv] | table:t1      | keep order:false,
    ↪ stats:pseudo                                                          |
+--
    ↪ ------------------------------+----------+-----------+---------------+--------------------------------
    ↪
```

上面是优化生效的情况，可以看到 HashJoin_11 是一个普通的 inner join。

接下来，关闭关联规则：

```
insert into mysql.opt_rule_blacklist values("decorrelate");
admin reload opt_rule_blacklist;
explain select * from t1 where t1.a < (select sum(t2.a) from t2 where t2.b = t1.b);
```

```
+--
    ↪ -------------------------------------+----------+-----------+-----------------------+----------------
    ↪
| id                                  | estRows  | task      | access object           |
    ↪ operator info                                                                 |
+--
    ↪ -------------------------------------+----------+-----------+-----------------------+----------------
    ↪
| Projection_10                       | 10000.00 | root      |                         | test.
    ↪ t1.a, test.t1.b                                                               |
| └─Apply_12                           | 10000.00 | root      |                         |
    ↪ CARTESIAN inner join, other cond:lt(cast(test.t1.a), Column#7)                |
|   ├─TableReader_14(Build)            | 10000.00 | root      |                         | data:
    ↪ TableFullScan_13                                                              |
|   │ └─TableFullScan_13               | 10000.00 | cop[tikv] | table:t1                | keep
    ↪ order:false, stats:pseudo                                                     |
|   └─MaxOneRow_15(Probe)              | 1.00     | root      |                         |
    ↪                                                                               |
|     └─HashAgg_27                     | 1.00     | root      |                         | funcs
    ↪ :sum(Column#10)->Column#7                                                     |
|       └─IndexLookUp_28               | 1.00     | root      |                         |
    ↪                                                                               |
|         ├─IndexRangeScan_25(Build)   | 10.00    | cop[tikv] | table:t2, index:idx(b) | range
    ↪ : decided by [eq(test.t2.b, test.t1.b)], keep order:false, stats:pseudo      |
|         └─HashAgg_17(Probe)          | 1.00     | cop[tikv] |                         | funcs
    ↪ :sum(test.t2.a)->Column#10                                                    |
|           └─TableRowIDScan_26        | 10.00    | cop[tikv] | table:t2                | keep
    ↪ order:false, stats:pseudo                                                     |
+--
    ↪ -------------------------------------+----------+-----------+-----------------------+----------------
    ↪
```

在执行了关闭关联规则的语句后，可以在 IndexRangeScan_25(Build) 的 operator info 中看到 range: decided by [eq(test.t2.b, test.t1.b)]。这部分信息就是关联依赖未被解除时，TiDB 使用关联条件进行索引范围查询的显示结果。

10.3.3.2.5 Max/Min 函数消除规则

在 SQL 中包含了 max/min 函数时，查询优化器会尝试使用 max/min 消除优化规则来将 max/min 聚合函数转换为 TopN 算子，从而能够有效地利用索引进行查询。

根据 select 语句中 max/min 函数的个数，这一优化规则有以下两种表现形式：

- 只有一个 max/min 函数时的优化规则
- 存在多个 max/min 函数时的优化规则

只有一个 max/min 函数时的优化规则

当一个 SQL 满足以下条件时，就会应用这个规则：

- 只有一个聚合函数，且为 max 或者 min 函数。
- 聚合函数没有相应的 group by 语句。

例如：

```
select max(a) from t
```

这时 max/min 消除优化规则会将其重写为：

```
select max(a) from (select a from t where a is not null order by a desc limit 1) t
```

这个新的 SQL 语句在 a 列存在索引（或 a 列是某个联合索引的前缀）时，能够利用索引只扫描一行数据来得到最大或者最小值，从而避免对整个表的扫描。

上述例子最终得到的执行计划如下：

```
mysql> explain select max(a) from t;
+------------------------------+---------+-----------+-------------------------+-------------------------------
    ↪
| id                           | estRows | task      | access object           | operator info
    ↪                        |
+------------------------------+---------+-----------+-------------------------+-------------------------------
    ↪
| StreamAgg_13                 | 1.00    | root      |                         | funcs:max(test.t
    ↪ .a)->Column#4       |
| └─Limit_17                   | 1.00    | root      |                         | offset:0, count
    ↪ :1                     |
|   └─IndexReader_27           | 1.00    | root      |                         | index:Limit_26
    ↪                        |
|     └─Limit_26               | 1.00    | cop[tikv] |                         | offset:0, count
    ↪ :1                     |
```

```
|       └─IndexFullScan_25     | 1.00    | cop[tikv] | table:t, index:idx_a(a) | keep order:true
    ↪ , desc, stats:pseudo |
+------------------------------+---------+-----------+-------------------------+-----------------------------
    ↪
5 rows in set (0.00 sec)
```

存在多个 max/min 函数时的优化规则

当一个 SQL 满足以下条件时，就会应用这个规则：

- 有多个聚合函数，且所有的聚合函数都是 max/min
- 聚合函数没有相应的 `group by` 语句。
- 每个 `max/min` 聚合函数参数中的列都有索引能够保序。

下面是一个简单的例子：

```
select max(a) - min(a) from t
```

优化规则会先检查 a 列是否存在索引能够为其保序，如果存在，这个 SQL 会先被重写为两个子查询的笛卡尔积：

```
select max_a - min_a
from
    (select max(a) as max_a from t) t1,
    (select min(a) as min_a from t) t2
```

这样，两个子句中的 `max/min` 函数就可以使用上述“只有一个 `max/min` 函数时的优化规则”分别进行优化，最终重写为：

```
select max_a - min_a
from
    (select max(a) as max_a from (select a from t where a is not null order by a desc limit 1) t)
        ↪  t1,
    (select min(a) as min_a from (select a from t where a is not null order by a asc limit 1) t)
        ↪ t2
```

同样的，如果 a 列能够使用索引保序，那这个优化只会扫描两行数据，避免了对整个表的扫描。但如果 a 列没有可以保序的索引，这个变换会使原本只需一次的全表扫描变成两次，因此这个规则就不会被应用。

最后得到的执行计划：

```
mysql> explain select max(a)-min(a) from t;
+-----------------------------------+---------+-----------+-------------------------+-----------------------------
    ↪
| id                                | estRows | task      | access object           | operator
    ↪ info                        |
+-----------------------------------+---------+-----------+-------------------------+-----------------------------
    ↪
```

```
| Projection_17                       | 1.00    | root      |                           | minus(
    ↪ Column#4, Column#5)->Column#6 |
| └─HashJoin_18                       | 1.00    | root      |                           | CARTESIAN
    ↪  inner join                |
|   ├─StreamAgg_45(Build)             | 1.00    | root      |                           | funcs:min
    ↪ (test.t.a)->Column#5       |
|   │ └─Limit_49                      | 1.00    | root      |                           | offset:0,
    ↪  count:1                   |
|   │   └─IndexReader_59              | 1.00    | root      |                           | index:
    ↪ Limit_58                      |
|   │     └─Limit_58                  | 1.00    | cop[tikv] |                           | offset:0,
    ↪  count:1                   |
|   │       └─IndexFullScan_57        | 1.00    | cop[tikv] | table:t, index:idx_a(a) | keep
    ↪ order:true, stats:pseudo       |
|   └─StreamAgg_24(Probe)             | 1.00    | root      |                           | funcs:max
    ↪ (test.t.a)->Column#4       |
|     └─Limit_28                      | 1.00    | root      |                           | offset:0,
    ↪  count:1                   |
|       └─IndexReader_38              | 1.00    | root      |                           | index:
    ↪ Limit_37                      |
|         └─Limit_37                  | 1.00    | cop[tikv] |                           | offset:0,
    ↪  count:1                   |
|           └─IndexFullScan_36        | 1.00    | cop[tikv] | table:t, index:idx_a(a) | keep
    ↪ order:true, desc, stats:pseudo |
+-------------------------------------+---------+-----------+-------------------------+-------------------------
    ↪
12 rows in set (0.01 sec)
```

10.3.3.2.6 谓词下推

本文档介绍 TiDB 逻辑优化规则中的谓词下推规则，旨在让读者对谓词下推形成理解，并了解常见的谓词下推适用及不适用的场景。

谓词下推将查询语句中的过滤表达式计算尽可能下推到距离数据源最近的地方，以尽早完成数据的过滤，进而显著地减少数据传输或计算的开销。

示例

以下通过一些例子对谓词下推优化进行说明，其中示例 1、2、3 为谓词下推适用的案例，示例 4、5、6 为谓词下推不适用的案例。

示例 1: 谓词下推到存储层

```
create table t(id int primary key, a int);
explain select * from t where a < 1;
+--
    ↪ ----------------------+----------+-----------+---------------+--------------------------------+
    ↪
```

```
| id                    | estRows  | task      | access object | operator info
    ↪                  |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| TableReader_7           | 3323.33  | root      |               | data:Selection_6
    ↪               |
| └─Selection_6           | 3323.33  | cop[tikv] |               | lt(test.t.a, 1)
    ↪                |
|   └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
3 rows in set (0.00 sec)
```

在该查询中，将谓词 `a < 1` 下推到 TiKV 上对数据进行过滤，可以减少由于网络传输带来的开销。

示例 2: 谓词下推到存储层

```
create table t(id int primary key, a int not null);
explain select * from t where a < substring('123', 1, 1);
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| id                    | estRows  | task      | access object | operator info
    ↪                  |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| TableReader_7           | 3323.33  | root      |               | data:Selection_6
    ↪               |
| └─Selection_6           | 3323.33  | cop[tikv] |               | lt(test.t.a, 1)
    ↪                |
|   └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
```

该查询与示例 1 中的查询生成了完成一样的执行计划，这是因为谓词 `a < substring('123', 1, 1)` 的 `substring` 的入参均为常量，因此可以提前计算，进而简化得到等价的谓词 `a < 1`。进一步的，可以将 `a < 1` 下推至 TiKV 上。

示例 3: 谓词下推到 join 下方

```
create table t(id int primary key, a int not null);
```

```
create table s(id int primary key, a int not null);
explain select * from t join s on t.a = s.a where t.a < 1;
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
| id                           | estRows  | task      | access object | operator info
    ↪                              |
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
| HashJoin_8                   | 4154.17  | root      |               | inner join, equal:[eq(
    ↪ test.t.a, test.s.a)] |
| ├─TableReader_15(Build)      | 3323.33  | root      |               | data:Selection_14
    ↪                            |
| │ └─Selection_14             | 3323.33  | cop[tikv] |               | lt(test.s.a, 1)
    ↪                              |
| │   └─TableFullScan_13       | 10000.00 | cop[tikv] | table:s       | keep order:false, stats:
    ↪ pseudo             |
| └─TableReader_12(Probe)      | 3323.33  | root      |               | data:Selection_11
    ↪                            |
|   └─Selection_11             | 3323.33  | cop[tikv] |               | lt(test.t.a, 1)
    ↪                              |
|     └─TableFullScan_10       | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo             |
+--
    ↪ ---------------------------+----------+-----------+---------------+--------------------------------
    ↪
7 rows in set (0.00 sec)
```

在该查询中，将谓词 `t.a < 1` 下推到 join 前进行过滤，可以减少 join 时的计算开销。

此外，这条 SQL 执行的是内连接，且 ON 条件是 `t.a = s.a`，可以由 `t.a < 1` 推导出谓词 `s.a < 1`，并将其下推至 join 运算前对 s 表进行过滤，可以进一步减少 join 时的计算开销。

示例 4: 存储层不支持的谓词无法下推

```
create table t(id int primary key, a int not null);
desc select * from t where substring('123', a, 1) = '1';
+--
    ↪ ----------------------+---------+-----------+---------------+----------------------------------+
    ↪
| id                      | estRows | task      | access object | operator info
    ↪                         |
+--
    ↪ ----------------------+---------+-----------+---------------+----------------------------------+
    ↪
| Selection_7             | 2.00    | root      |               | eq(substring("123", test.t.a,
```

```
    ↪ 1), "1") |
| └─TableReader_6         | 2.00    | root      |               | data:TableFullScan_5
    ↪                    |
|   └─TableFullScan_5     | 2.00    | cop[tikv] | table:t       | keep order:false, stats:pseudo
    ↪          |
+--
    ↪ ----------------------+---------+-----------+---------------+--------------------------------+
    ↪
```

在该查询中，存在谓词 `substring('123', a, 1)= '1'`。

从 explain 结果中可以看到，该谓词没有被下推到 TiKV 上进行计算，这是因为 TiKV coprocessor 中没有对 `substring` 内置函数进行支持，因此无法将其下推到 TiKV 上。

示例 5: 外连接中内表上的谓词不能下推

```
create table t(id int primary key, a int not null);
create table s(id int primary key, a int not null);
explain select * from t left join s on t.a = s.a where s.a is null;
+--
    ↪ ----------------------------+----------+-----------+---------------+--------------------------------
    ↪
| id                            | estRows  | task      | access object | operator info
    ↪                                    |
+--
    ↪ ----------------------------+----------+-----------+---------------+--------------------------------
    ↪
| Selection_7                   | 10000.00 | root      |               | isnull(test.s.a)
    ↪                                  |
| └─HashJoin_8                   | 12500.00 | root      |               | left outer join, equal
    ↪ :[eq(test.t.a, test.s.a)] |
|   ├─TableReader_13(Build)     | 10000.00 | root      |               | data:TableFullScan_12
    ↪                           |
|   │ └─TableFullScan_12        | 10000.00 | cop[tikv] | table:s       | keep order:false, stats
    ↪ :pseudo                  |
|   └─TableReader_11(Probe)     | 10000.00 | root      |               | data:TableFullScan_10
    ↪                           |
|     └─TableFullScan_10        | 10000.00 | cop[tikv] | table:t       | keep order:false, stats
    ↪ :pseudo                  |
+--
    ↪ ----------------------------+----------+-----------+---------------+--------------------------------
    ↪
6 rows in set (0.00 sec)
```

在该查询中，内表 s 上存在谓词 `s.a is null`。

从 explain 中可以看到，该谓词没有被下推到 join 前进行计算，这是因为外连接在不满足 on 条件时会对内表填充 NULL，而在该查询中 `s.a is null` 用来对 join 后的结果进行过滤，如果将其下推到 join 前在内表上进行过

滤，则下推前后不等价，因此不可进行下推。

示例 6: 谓词中包含用户变量时不能下推

```
create table t(id int primary key, a char);
set @a = 1;
explain select * from t where a < @a;
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| id                    | estRows  | task      | access object | operator info
    ↪                  |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
| Selection_5             | 8000.00  | root      |               | lt(test.t.a, getvar("a"))
    ↪      |
| └─TableReader_7         | 10000.00 | root      |               | data:TableFullScan_6
    ↪           |
|   └─TableFullScan_6     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo |
+--
    ↪ ----------------------+----------+-----------+---------------+------------------------------+
    ↪
3 rows in set (0.00 sec)
```

在该查询中，表 t 上存在谓词 a < @a，其中 @a 为值为 1 的用户变量。

从 explain 中可以看到，该谓词没有像示例 2 中一样，将谓词简化为 a < 1 并下推到 TiKV 上进行计算。这是因为，用户变量 @a 的值可能会某些场景下在查询过程中发生改变，且 TiKV 对于用户变量 @a 的值不可知，因此 TiDB 不会将 @a 替换为 1，且不会下推至 TiKV 上进行计算。

一个帮助理解的例子如下：

```
create table t(id int primary key, a int);
insert into t values(1, 1), (2,2);
set @a = 1;
select id, a, @a:=@a+1 from t where a = @a;
+----+------+----------+
| id | a    | @a:=@a+1 |
+----+------+----------+
|  1 |    1 | 2        |
|  2 |    2 | 3        |
+----+------+----------+
2 rows in set (0.00 sec)
```

可以从在该查询中看到，@a 的值会在查询过程中发生改变，因此如果将 a = @a 替换为 a = 1 并下推至 TiKV，则优化前后不等价。

10.3.3.2.7 分区裁剪

分区裁剪是只有当目标表为分区表时，才可以进行的一种优化方式。分区裁剪通过分析查询语句中的过滤条件，只选择可能满足条件的分区，不扫描匹配不上的分区，进而显著地减少计算的数据量。

例如：

```
CREATE TABLE t1 (
 id INT NOT NULL PRIMARY KEY,
 pad VARCHAR(100)
)
PARTITION BY RANGE COLUMNS(id) (
 PARTITION p0 VALUES LESS THAN (100),
 PARTITION p1 VALUES LESS THAN (200),
 PARTITION p2 VALUES LESS THAN (MAXVALUE)
);
INSERT INTO t1 VALUES (1, 'test1'),(101, 'test2'), (201, 'test3');
EXPLAIN SELECT * FROM t1 WHERE id BETWEEN 80 AND 120;
```

```
+--
    ↪ -------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
| id                          | estRows | task      | access object         | operator info
    ↪                                    |
+--
    ↪ -------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
| PartitionUnion_8            | 80.00   | root      |                       |
    ↪                                                |
| ├─TableReader_10            | 40.00   | root      |                       | data:
    ↪ TableRangeScan_9                             |
| │ └─TableRangeScan_9        | 40.00   | cop[tikv] | table:t1, partition:p0 | range:[80,120],
    ↪ keep order:false, stats:pseudo |
| └─TableReader_12            | 40.00   | root      |                       | data:
    ↪ TableRangeScan_11                            |
|   └─TableRangeScan_11       | 40.00   | cop[tikv] | table:t1, partition:p1 | range:[80,120],
    ↪ keep order:false, stats:pseudo |
+--
    ↪ -------------------------+---------+-----------+-----------------------+-----------------------------
    ↪
5 rows in set (0.00 sec)
```

分区裁剪的使用场景

分区表有 Range 分区和 hash 分区两种形式，分区裁剪对两种分区表也有不同的使用场景。

分区裁剪在 Hash 分区表上的应用

Hash 分区表上可以使用分区裁剪的场景

只有等值比较的查询条件能够支持 Hash 分区表的裁剪。

```
create table t (x int) partition by hash(x) partitions 4;
explain select * from t where x = 1;
```

```
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
| id                        | estRows  | task      | access object        | operator info
    ↪                   |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
| TableReader_8             | 10.00    | root      |                      | data:Selection_7
    ↪                |
| └─Selection_7             | 10.00    | cop[tikv] |                      | eq(test.t.x, 1)
    ↪                 |
|   └─TableFullScan_6       | 10000.00 | cop[tikv] | table:t, partition:p1 | keep order:false,
    ↪ stats:pseudo |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
```

在这条 SQL 中，由条件 `x = 1` 可以知道所有结果均在一个分区上。数值 1 在经过 Hash 后，可以确定其在分区 p1 中。因此只需要扫描分区 p1，而无需访问一定不会出现相关结果的 p2、p3、p4 分区。从执行计划来看，其中只出现了一个 TableFullScan 算子，且在 access object 中指定了 p1 分区，确认 partition pruning 生效了。

Hash 分区表上不能使用分区裁剪的场景

场景一

不能确定查询结果只在一个分区上的条件：如 in, between, `> < >= <=` 等查询条件，不能使用分区裁剪的优化。

```
create table t (x int) partition by hash(x) partitions 4;
explain select * from t where x > 2;
```

```
+--
    ↪ ---------------------------+----------+-----------+----------------------+-------------------------------
    ↪
| id                           | estRows  | task      | access object        | operator info
    ↪                   |
+--
    ↪ ---------------------------+----------+-----------+----------------------+-------------------------------
    ↪
| Union_10                     | 13333.33 | root      |                      |
    ↪                                |
```

```
| ├─TableReader_13              | 3333.33  | root      |                       | data:
    ↪ Selection_12              |
| | └─Selection_12              | 3333.33  | cop[tikv] |                       | gt(test.t.x, 2)
    ↪                 |
| |   └─TableFullScan_11        | 10000.00 | cop[tikv] | table:t, partition:p0 | keep order:false
    ↪ , stats:pseudo |
| ├─TableReader_16              | 3333.33  | root      |                       | data:
    ↪ Selection_15              |
| | └─Selection_15              | 3333.33  | cop[tikv] |                       | gt(test.t.x, 2)
    ↪                 |
| |   └─TableFullScan_14        | 10000.00 | cop[tikv] | table:t, partition:p1 | keep order:false
    ↪ , stats:pseudo |
| ├─TableReader_19              | 3333.33  | root      |                       | data:
    ↪ Selection_18              |
| | └─Selection_18              | 3333.33  | cop[tikv] |                       | gt(test.t.x, 2)
    ↪                 |
| |   └─TableFullScan_17        | 10000.00 | cop[tikv] | table:t, partition:p2 | keep order:false
    ↪ , stats:pseudo |
| └─TableReader_22              | 3333.33  | root      |                       | data:
    ↪ Selection_21              |
|   └─Selection_21              | 3333.33  | cop[tikv] |                       | gt(test.t.x, 2)
    ↪                 |
|     └─TableFullScan_20        | 10000.00 | cop[tikv] | table:t, partition:p3 | keep order:false
    ↪ , stats:pseudo |
+--
    ↪ ----------------------------+----------+-----------+-----------------------+-----------------------------
    ↪
```

在这条 SQL 中，`x > 2` 条件无法确定对应的 Hash Partition，所以不能使用分区裁剪。

场景二

由于分区裁剪的规则优化是在查询计划的生成阶段，对于执行阶段才能获取到过滤条件的场景，无法利用分区裁剪的优化。

```
create table t (x int) partition by hash(x) partitions 4;
explain select * from t2 where x = (select * from t1 where t2.x = t1.x and t2.x < 2);
```

```
+--
    ↪ ----------------------------------+----------+-----------+-----------------------+-------------------
    ↪
| id                                  | estRows  | task      | access object         | operator
    ↪  info                                |
+--
    ↪ ----------------------------------+----------+-----------+-----------------------+-------------------
    ↪
```

```
| Projection_13                         | 9990.00  | root      |                                | test.t2.
    ↪ x                                      |
| └─Apply_15                            | 9990.00  | root      |                                | inner
    ↪ join, equal:[eq(test.t2.x, test.t1.x)] |
|   ├─TableReader_18(Build)             | 9990.00  | root      |                                | data:
    ↪ Selection_17                           |
|   │ └─Selection_17                     | 9990.00  | cop[tikv] |                                | not(
    ↪ isnull(test.t2.x))                     |
|   │   └─TableFullScan_16               | 10000.00 | cop[tikv] | table:t2                       | keep
    ↪ order:false, stats:pseudo              |
|   └─Selection_19(Probe)               | 0.80     | root      |                                | not(
    ↪ isnull(test.t1.x))                     |
|     └─MaxOneRow_20                    | 1.00     | root      |                                |
    ↪                                        |
|       └─Union_21                      | 2.00     | root      |                                |
    ↪                                        |
|         ├─TableReader_24              | 2.00     | root      |                                | data:
    ↪ Selection_23                           |
|         │ └─Selection_23              | 2.00     | cop[tikv] |                                | eq(test
    ↪ .t2.x, test.t1.x), lt(test.t2.x, 2)    |
|         │   └─TableFullScan_22        | 2500.00  | cop[tikv] | table:t1, partition:p0         | keep
    ↪ order:false, stats:pseudo              |
|         └─TableReader_27              | 2.00     | root      |                                | data:
    ↪ Selection_26                           |
|           └─Selection_26              | 2.00     | cop[tikv] |                                | eq(test
    ↪ .t2.x, test.t1.x), lt(test.t2.x, 2)    |
|             └─TableFullScan_25        | 2500.00  | cop[tikv] | table:t1, partition:p1         | keep
    ↪ order:false, stats:pseudo              |
+--
    ↪ ---------------------------------+----------+-----------+------------------------+------------------
    ↪
```

这个查询每从 `t2` 读取一行，都会去分区表 `t1` 上进行查询，理论上这时会满足 `t1.x = val` 的过滤条件，但实际上由于分区裁剪只作用于查询计划生成阶段，而不是执行阶段，因而不会做裁剪。

分区裁剪在 Range 分区表上的应用

Range 分区表上可以使用分区裁剪的场景

场景一

等值比较的查询条件可以使用分区裁剪。

```
create table t (x int) partition by range (x) (
    partition p0 values less than (5),
    partition p1 values less than (10),
    partition p2 values less than (15)
    );
```

```
explain select * from t where x = 3;
```

```
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
| id                      | estRows  | task      | access object        | operator info
    ↪                  |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
| TableReader_8           | 10.00    | root      |                      | data:Selection_7
    ↪               |
| └─Selection_7           | 10.00    | cop[tikv] |                      | eq(test.t.x, 3)
    ↪                |
|   └─TableFullScan_6     | 10000.00 | cop[tikv] | table:t, partition:p0 | keep order:false,
    ↪ stats:pseudo |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------+
    ↪
```

使用 `in` 条件的等值比较查询条件也可以使用分区裁剪。

```
create table t (x int) partition by range (x) (
    partition p0 values less than (5),
    partition p1 values less than (10),
    partition p2 values less than (15)
    );
explain select * from t where x in(1,13);
```

```
+--
    ↪ --------------------------+----------+-----------+----------------------+------------------------------
    ↪
| id                          | estRows  | task      | access object        | operator info
    ↪                  |
+--
    ↪ --------------------------+----------+-----------+----------------------+------------------------------
    ↪
| Union_8                     | 40.00    | root      |                      |
    ↪                             |
| ├─TableReader_11            | 20.00    | root      |                      | data:Selection_10
    ↪               |
| │ └─Selection_10            | 20.00    | cop[tikv] |                      | in(test.t.x, 1,
    ↪ 13)            |
| │   └─TableFullScan_9       | 10000.00 | cop[tikv] | table:t, partition:p0 | keep order:false,
    ↪  stats:pseudo |
```

```
| └─TableReader_14            | 20.00    | root      |                       | data:Selection_13
    ↪                |
|   └─Selection_13            | 20.00    | cop[tikv] |                       | in(test.t.x, 1,
    ↪ 13)            |
|     └─TableFullScan_12      | 10000.00 | cop[tikv] | table:t, partition:p2 | keep order:false,
    ↪  stats:pseudo |
+--
    ↪ -------------------------+----------+-----------+-----------------------+-------------------------------
    ↪
```

在这条 SQL 中，由条件 `x in(1,13)` 可以知道所有结果只会分布在几个分区上。经过分析，发现所有 `x = 1` 的记录都在分区 `p0` 上，所有 `x = 13` 的记录都在分区 `p2` 上，因此只需要访问 `p0`、`p2` 这两个分区，

场景二

区间比较的查询条件如 `between, > < = >= <=` 可以使用分区裁剪。

```
create table t (x int) partition by range (x) (
    partition p0 values less than (5),
    partition p1 values less than (10),
    partition p2 values less than (15)
    );
explain select * from t where x between 7 and 14;
```

```
+--
    ↪ -------------------------+----------+-----------+-----------------------+-------------------------------
    ↪
| id                          | estRows  | task      | access object         | operator info
    ↪                     |
+--
    ↪ -------------------------+----------+-----------+-----------------------+-------------------------------
    ↪
| Union_8                     | 500.00   | root      |                       |
    ↪                                 |
| ├─TableReader_11            | 250.00   | root      |                       | data:Selection_10
    ↪                  |
| │ └─Selection_10            | 250.00   | cop[tikv] |                       | ge(test.t.x, 7),
    ↪ le(test.t.x, 14) |
| │   └─TableFullScan_9       | 10000.00 | cop[tikv] | table:t, partition:p1 | keep order:false,
    ↪  stats:pseudo    |
| └─TableReader_14            | 250.00   | root      |                       | data:Selection_13
    ↪                  |
|   └─Selection_13            | 250.00   | cop[tikv] |                       | ge(test.t.x, 7),
    ↪ le(test.t.x, 14) |
|     └─TableFullScan_12      | 10000.00 | cop[tikv] | table:t, partition:p2 | keep order:false,
    ↪  stats:pseudo    |
```

```
+--
    ↪ -------------------------+----------+-----------+----------------------+-------------------------------
    ↪
```

场景三

分区表达式为 fn(col) 的简单形式，查询条件是 > < = >= <= 之一，且 fn 是单调函数，可以使用分区裁剪。

关于 fn 函数，对于任意 x y，如果 x > y，则 fn(x)> fn(y)，那么这种是严格递增的单调函数。非严格递增的单调函数也可以符合分区裁剪要求，只要函数 fn 满足：对于任意 x y，如果 x > y，则 fn(x)>= fn(y)。理论上，所有满足单调条件（严格或者非严格）的函数都支持分区裁剪。目前，TiDB 支持的单调函数如下：

```
unix_timestamp
to_days
```

例如，分区表达式是 fn(col) 形式，fn 为我们支持的单调函数 to_days，就可以使用分区裁剪：

```
create table t (id datetime) partition by range (to_days(id)) (
    partition p0 values less than (to_days('2020-04-01')),
    partition p1 values less than (to_days('2020-05-01')));
explain select * from t where id > '2020-04-18';
```

```
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------
    ↪
| id                      | estRows  | task      | access object        | operator info
    ↪                             |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------
    ↪
| TableReader_8           | 3333.33  | root      |                      | data:Selection_7
    ↪                           |
| └─Selection_7           | 3333.33  | cop[tikv] |                      | gt(test.t.id,
    ↪ 2020-04-18 00:00:00.000000) |
|   └─TableFullScan_6     | 10000.00 | cop[tikv] | table:t, partition:p1 | keep order:false,
    ↪ stats:pseudo            |
+--
    ↪ ----------------------+----------+-----------+----------------------+-------------------------------
    ↪
```

Range 分区表上不能使用分区裁剪的场景

由于分区裁剪的规则优化是在查询计划的生成阶段，对于执行阶段才能获取到过滤条件的场景，无法利用分区裁剪的优化。

```
create table t1 (x int) partition by range (x) (
    partition p0 values less than (5),
    partition p1 values less than (10));
```

```
create table t2 (x int);
explain select * from t2 where x < (select * from t1 where t2.x < t1.x and t2.x < 2);
```

```
+--
    ↪ ----------------------------------+----------+-----------+-----------------------+------------------
    ↪
| id                                  | estRows  | task      | access object         | operator
    ↪  info                                          |
+--
    ↪ ----------------------------------+----------+-----------+-----------------------+------------------
    ↪
| Projection_13                       | 9990.00  | root      |                       | test.t2.
    ↪ x                                              |
| └─Apply_15                          | 9990.00  | root      |                       |
    ↪ CARTESIAN inner join, other cond:lt(test.t2.x, test.t1.x) |
|   ├─TableReader_18(Build)           | 9990.00  | root      |                       | data:
    ↪ Selection_17                                   |
|   │ └─Selection_17                  | 9990.00  | cop[tikv] |                       | not(
    ↪ isnull(test.t2.x))                             |
|   │   └─TableFullScan_16            | 10000.00 | cop[tikv] | table:t2              | keep
    ↪ order:false, stats:pseudo                      |
|   └─Selection_19(Probe)             | 0.80     | root      |                       | not(
    ↪ isnull(test.t1.x))                             |
|     └─MaxOneRow_20                  | 1.00     | root      |                       |
    ↪                                                |
|       └─Union_21                    | 2.00     | root      |                       |
    ↪                                                |
|         ├─TableReader_24            | 2.00     | root      |                       | data:
    ↪ Selection_23                                   |
|         │ └─Selection_23            | 2.00     | cop[tikv] |                       | lt(test
    ↪ .t2.x, 2), lt(test.t2.x, test.t1.x)            |
|         │   └─TableFullScan_22      | 2.50     | cop[tikv] | table:t1, partition:p0 | keep
    ↪ order:false, stats:pseudo                      |
|         └─TableReader_27            | 2.00     | root      |                       | data:
    ↪ Selection_26                                   |
|           └─Selection_26            | 2.00     | cop[tikv] |                       | lt(test
    ↪ .t2.x, 2), lt(test.t2.x, test.t1.x)            |
|             └─TableFullScan_25      | 2.50     | cop[tikv] | table:t1, partition:p1 | keep
    ↪ order:false, stats:pseudo                      |
+--
    ↪ ----------------------------------+----------+-----------+-----------------------+------------------
    ↪
14 rows in set (0.00 sec)
```

这个查询每从 `t2` 读取一行，都会去分区表 `t1` 上进行查询，理论上这时会满足 `t1.x > val` 的过滤条件，但实

际上由于分区裁剪只作用于查询计划生成阶段，而不是执行阶段，因而不会做裁剪。

10.3.3.2.8 TopN 和 Limit 下推

SQL 中的 LIMIT 子句在 TiDB 查询计划树中对应 Limit 算子节点，ORDER BY 子句在查询计划树中对应 Sort 算子节点，此外，我们会将相邻的 Limit 和 Sort 算子组合成 TopN 算子节点，表示按某个排序规则提取记录的前 N 项。从另一方面来说，Limit 节点等价于一个排序规则为空的 TopN 节点。

和谓词下推类似，TopN（及 Limit，下同）下推将查询计划树中的 TopN 计算尽可能下推到距离数据源最近的地方，以尽早完成数据的过滤，进而显著地减少数据传输或计算的开销。

如果要关闭这个规则，可参照优化规则及表达式下推的黑名单中的关闭方法。

示例

以下通过一些例子对 TopN 下推进行说明。

示例 1：下推到存储层 Coprocessor

```
create table t(id int primary key, a int not null);
explain select * from t order by a limit 10;
```

```
+---------------------------+----------+-----------+---------------+--------------------------------+
    ↪
| id                        | estRows  | task      | access object | operator info
    ↪                   |
+---------------------------+----------+-----------+---------------+--------------------------------+
    ↪
| TopN_7                    | 10.00    | root      |               | test.t.a, offset:0, count
    ↪ :10   |
| └─TableReader_15           | 10.00    | root      |               | data:TopN_14
    ↪                   |
|   └─TopN_14                | 10.00    | cop[tikv] |               | test.t.a, offset:0, count
    ↪ :10   |
|     └─TableFullScan_13     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo |
+---------------------------+----------+-----------+---------------+--------------------------------+
    ↪
4 rows in set (0.00 sec)
```

在该查询中，将 TopN 算子节点下推到 TiKV 上对数据进行过滤，每个 Coprocessor 只向 TiDB 传输 10 条记录。在 TiDB 将数据整合后，再进行最终的过滤。

示例 2：TopN 下推过 Join 的情况（排序规则仅依赖于外表中的列）

```
create table t(id int primary key, a int not null);
create table s(id int primary key, a int not null);
explain select * from t left join s on t.a = s.a order by t.a limit 10;
```

```
+---------------------------------+----------+-----------+--------------+--------------------------------
    ↪
| id                              | estRows  | task      | access object | operator info
    ↪                                    |
+---------------------------------+----------+-----------+--------------+--------------------------------
    ↪
| TopN_12                         | 10.00    | root      |               | test.t.a, offset:0,
    ↪ count:10                     |
| └─HashJoin_17                    | 12.50    | root      |               | left outer join,
    ↪ equal:[eq(test.t.a, test.s.a)] |
|   ├─TopN_18(Build)               | 10.00    | root      |               | test.t.a, offset:0,
    ↪ count:10                     |
|   │ └─TableReader_26             | 10.00    | root      |               | data:TopN_25
    ↪                                      |
|   │   └─TopN_25                  | 10.00    | cop[tikv] |               | test.t.a, offset:0,
    ↪ count:10                     |
|   │     └─TableFullScan_24       | 10000.00 | cop[tikv] | table:t       | keep order:false,
    ↪ stats:pseudo                   |
|   └─TableReader_30(Probe)        | 10000.00 | root      |               | data:
    ↪ TableFullScan_29                            |
|     └─TableFullScan_29           | 10000.00 | cop[tikv] | table:s       | keep order:false,
    ↪ stats:pseudo                   |
+---------------------------------+----------+-----------+--------------+--------------------------------
    ↪
8 rows in set (0.01 sec)
```

在该查询中，TopN 算子的排序规则仅依赖于外表 t 中的列，可以将 TopN 下推到 Join 之前进行一次计算，以减少 Join 时的计算开销。除此之外，TiDB 同样将 TopN 下推到了存储层中。

示例 3：TopN 不能下推过 Join 的情况

```
create table t(id int primary key, a int not null);
create table s(id int primary key, a int not null);
explain select * from t join s on t.a = s.a order by t.id limit 10;
```

```
+-------------------------------+----------+-----------+--------------+--------------------------------
    ↪
| id                            | estRows  | task      | access object | operator info
    ↪                               |
+-------------------------------+----------+-----------+--------------+--------------------------------
    ↪
| TopN_12                       | 10.00    | root      |               | test.t.id, offset:0,
    ↪ count:10              |
| └─HashJoin_16                  | 12500.00 | root      |               | inner join, equal:[eq(
    ↪ test.t.a, test.s.a)] |
```

```
|   ├─TableReader_21(Build)     | 10000.00 | root      |               | data:TableFullScan_20
    ↪                       |
|   │ └─TableFullScan_20        | 10000.00 | cop[tikv] | table:s       | keep order:false, stats
    ↪ :pseudo             |
|   └─TableReader_19(Probe)     | 10000.00 | root      |               | data:TableFullScan_18
    ↪                       |
|     └─TableFullScan_18        | 10000.00 | cop[tikv] | table:t       | keep order:false, stats
    ↪ :pseudo             |
+-------------------------------+----------+-----------+---------------+-----------------------------------
    ↪
6 rows in set (0.00 sec)
```

TopN 无法下推过 Inner Join。以上面的查询为例，如果先 Join 得到 100 条记录，再做 TopN 可以剩余 10 条记录。而如果在 TopN 之前就过滤到剩余 10 条记录，做完 Join 之后可能就剩下 5 条了，导致了结果的差异。

同理，TopN 无法下推到 Outer Join 的内表上。在 TopN 的排序规则涉及多张表上的列时，也无法下推，如 `t.a+s.a`。只有当 TopN 的排序规则仅依赖于外表上的列时，才可以下推。

示例 4：TopN 转换成 Limit 的情况

```
create table t(id int primary key, a int not null);
create table s(id int primary key, a int not null);
explain select * from t left join s on t.a = s.a order by t.id limit 10;
```

```
+---------------------------------+----------+-----------+---------------+-----------------------------------
    ↪
| id                              | estRows  | task      | access object | operator info
    ↪                                    |
+---------------------------------+----------+-----------+---------------+-----------------------------------
    ↪
| TopN_12                         | 10.00    | root      |               | test.t.id, offset:0,
    ↪ count:10                   |
| └─HashJoin_17                    | 12.50    | root      |               | left outer join,
    ↪ equal:[eq(test.t.a, test.s.a)] |
|   ├─Limit_21(Build)              | 10.00    | root      |               | offset:0, count:10
    ↪                               |
|   │ └─TableReader_31             | 10.00    | root      |               | data:Limit_30
    ↪                                    |
|   │   └─Limit_30                 | 10.00    | cop[tikv] |               | offset:0, count:10
    ↪                               |
|   │     └─TableFullScan_29       | 10.00    | cop[tikv] | table:t       | keep order:true,
    ↪ stats:pseudo                    |
|   └─TableReader_35(Probe)        | 10000.00 | root      |               | data:
    ↪ TableFullScan_34                            |
|     └─TableFullScan_34           | 10000.00 | cop[tikv] | table:s       | keep order:false,
    ↪ stats:pseudo                   |
```

```
+---------------------------------+----------+----------+--------------+-------------------------------------------------
    ↪
8 rows in set (0.00 sec)
```

在上面的查询中，TopN 首先推到了外表 t 上。然后因为它要对 t.id 进行排序，而 t.id 是表 t 的主键，可以直接按顺序读出 (keep order:true)，从而省略了 TopN 中的排序，将其简化为 Limit。

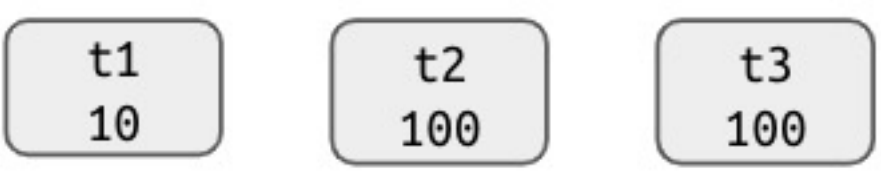
10.3.3.2.9 Join Reorder 算法简介

在实际的业务场景中，多个表的 Join 语句是很常见的，而 Join 的执行效率和各个表参与 Join 的顺序有关系。如 select * from t1, t2, t3 where t1.a=t2.a and t3.a=t2.a，这个 SQL 中可能的执行顺序有“t1 和 t2 先做 Join，然后再和 t3 做 Join”以及“t2 和 t3 先做 Join，然后再和 t1 做 Join”两种情况。根据 t1 和 t3 的数据量及数据分布，这两种执行顺序会有不同的性能表现。

因此优化器需要实现一种决定 Join 顺序的算法。目前 TiDB 中使用的算法是 Join Reorder 算法，又称贪心算法。

Join Reorder 算法实例

以三个表 t1、t2、t3 的 Join 为例。
序。



之后选定其中最小
一对。

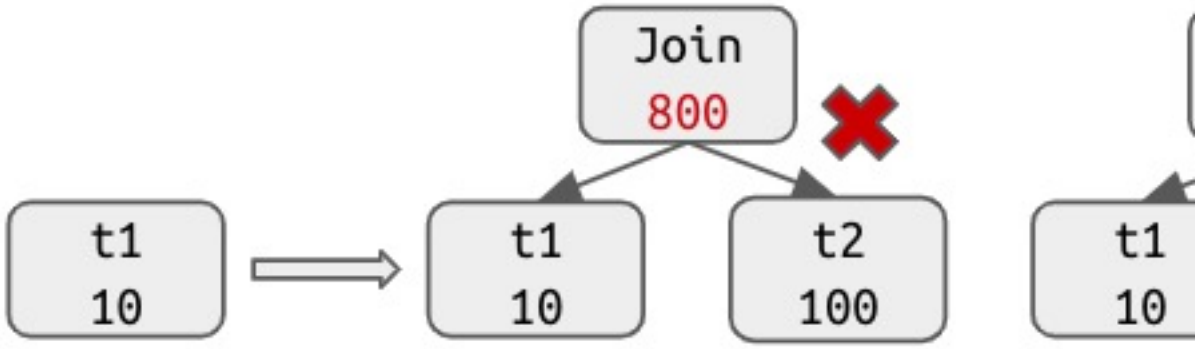


图 209: join-reorder-2

然后进入下一轮的选择，如果这时是四个表，那么就继续比较输出结果集的大小，进行选择。这里只有三个表，因此就直接得到了最终的 Join 结果。

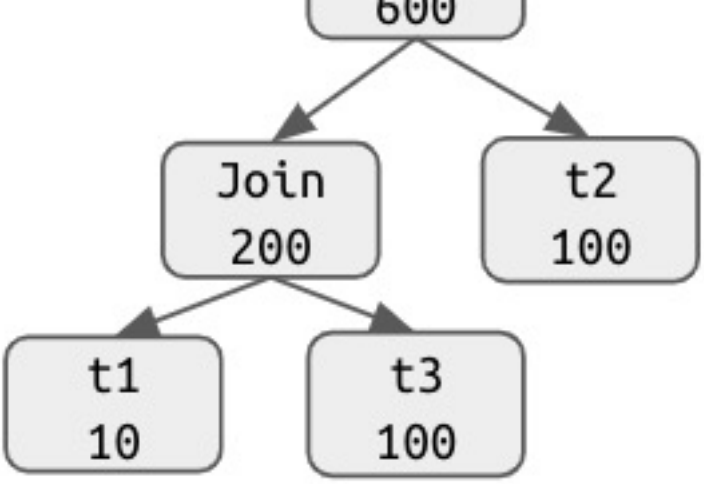


图 210: join-reorder-3

以上就是当前 TiDB 中使用的 Join reorder 算法。

Join reorder 算法限制

当前的 Join Reorder 算法存在如下限制

- 目前并不支持 Outer Join 的 Join Reorder
- 受结果集的计算算法所限并不会保证一定会选到合适的 Join order

目前 TiDB 中支持使用 STRAIGHT_JOIN 语法来强制指定一种 Join 顺序，参见语法元素说明。

10.3.3.3 物理优化

10.3.3.3.1 物理优化

物理优化是基于代价的优化，为上一阶段产生的逻辑执行计划制定物理执行计划。这一阶段中，优化器会为逻辑执行计划中的每个算子选择具体的物理实现。逻辑算子的不同物理实现有着不同的时间复杂度、资源消耗和物理属性等。在这个过程中，优化器会根据数据的统计信息来确定不同物理实现的代价，并选择整体代价最小的物理执行计划。

理解 TiDB 执行计划文档已对每个物理算子进行了一些介绍。在本章我们会重点介绍以下方面：

- 在索引的选择中会介绍 TiDB 在一张表有多个索引时，如何选择最优的索引进行表的访问。
- 在统计信息简介中会介绍 TiDB 收集了哪些统计信息来获得表的数据分布情况
- 在错误索引的解决方案中会介绍当发现 TiDB 索引选错时，你应该使用那些手段来让它使用正确的索引
- 在Distinct 优化中会介绍在物理优化中会做的一个有关 DISTINCT 关键字的优化，在这一小节中会介绍它的优缺点以及如何使用它。

10.3.3.3.2 索引的选择

从存储层读取数据是 SQL 计算过程中最为耗时的部分之一，TiDB 目前支持从不同的存储和不同的索引中读取数据，索引选择得是否合理将很大程度上决定一个查询的运行速度。

本章节将介绍 TiDB 如何选择索引去读入数据，以及相关的一些控制索引选择的方式。

读表

在介绍索引的选择之前，首先要了解 TiDB 有哪些读表的方式，这些方式的触发条件是什么，不同方式有什么区别，各有什么优劣。

读表算子

读表算子	触发条件	适用场景	说明
PointGet/BatchPoint	读表的范围是一个或多个单点范围	任何场景	如果能被触发，通常被认为是最快的算子，因为其直接调用 kvget 的接口进行计算，不走 coprocessor
TableReader	无	任何场景	从 TiKV 端直接扫描表数据，一般被认为是效率最低的算子，除非在 `_tidb_rowid` ↪ 这一列上存在范围查询，或者无其他可以选择的读表算子时，才会选择这个算子

读表算子	触发条件	适用场景	说明
TableReader	表在 TiFlash 节点上存在副本	需要读取的列比较少，但是需要计算的行很多	TiFlash 是列式存储，如果需要对少量的列和大量的行进行计算，一般会选择这个算子
IndexReader	表有一个或多个索引，且计算所需的列被包含在索引里	存在较小的索引上的范围查询，或者对索引列有顺序需求的时候	当存在多个索引的时候，会根据估算代价选择合理的索引
IndexLookupReader	表有一个或多个索引，且计算所需的列不完全被包含在索引里	同 IndexReader	因为计算列不完全被包含在索引里，所以读完索引后需要回表，这里会比 IndexReader 多一些开销

注意：

TableReader 是基于 `_tidb_rowid` 的索引，TiFlash 是列存索引，所以索引的选择即是读表算子的选择。

索引的选择

TiDB 基于规则或基于代价来选择索引。基于的规则包括前置规则和 Skyline-Pruning。在选择索引时，TiDB 会先尝试前置规则。如果存在索引满足某一条前置规则，则直接选择该索引。否则，TiDB 会采用 Skyline-Pruning 来排除不合适的索引，然后基于每个读表算子的代价估算，选择代价最小的索引。

基于规则选择

前置规则

TiDB 采用如下的启发式前置规则来选择索引：

- 规则 1：如果存在索引满足 “唯一性索引全匹配 + 不需要回表（即该索引生成的计划是 IndexReader）” 时，直接选择该索引。
- 规则 2：如果存在索引满足 “唯一性索引全匹配 + 需要回表（即该索引生成的计划是 IndexLookupReader）” 时，选择满足该条件且回表行数最小的索引作为候选索引。
- 规则 3：如果存在索引满足 “普通索引不需要回表 + 读取行数小于一定阈值” 时，选择满足该条件且读取行数最小的索引作为候选索引。
- 规则 4：如果规则 2 和 3 之中仅选出一条候选索引，则选择该候选索引。如果规则 2 和 3 均选出候选索引，则选择读取行数（读索引行数 + 回表行数）较小的索引。

上述规则中的 “索引全匹配” 指每个索引列上均存在等值条件。在执行 `EXPLAIN FORMAT = 'verbose' ...` 语句时，如果前置规则匹配了某一索引，TiDB 会输出一条 NOTE 级别的 warning 提示该索引匹配了前置规则。

在以下示例中，因为索引 `idx_b` 满足规则 2 中 “唯一性索引全匹配 + 需要回表” 的条件，TiDB 选择索引 `idx_b` 作为访问路径，`SHOW WARNING` 返回了索引 `idx_b` 命中前置规则的提示。

```
mysql> CREATE TABLE t(a INT PRIMARY KEY, b INT, c INT, UNIQUE INDEX idx_b(b));
Query OK, 0 rows affected (0.01 sec)

mysql> EXPLAIN FORMAT = 'verbose' SELECT b, c FROM t WHERE b = 3 OR b = 6;
+--
    ↪ -----------------+---------+---------+------+------------------------+-----------------------------+
    ↪
| id                | estRows | estCost | task | access object          | operator info
    ↪                |
+--
    ↪ -----------------+---------+---------+------+------------------------+-----------------------------+
    ↪
| Batch_Point_Get_5 | 2.00    | 8.80    | root | table:t, index:idx_b(b) | keep order:false, desc
    ↪ :false |
+--
    ↪ -----------------+---------+---------+------+------------------------+-----------------------------+
    ↪
1 row in set, 1 warning (0.00 sec)

mysql> SHOW WARNINGS;
```

```
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
| Level | Code | Message
    ↪                                                                                                      |
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
| Note  | 1105 | unique index idx_b of t is selected since the path only has point ranges with
    ↪ double scan |
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
1 row in set (0.00 sec)
```

Skyline-Pruning

Skyline-Pruning 是一个针对索引的启发式过滤规则，能降低错误估算导致选错索引的概率。Skyline-Pruning 从以下三个维度衡量一个索引的优劣：

- 索引的列涵盖了多少访问条件。“访问条件”指的是可以转化为某列范围的 where 条件，如果某个索引的列集合涵盖的访问条件越多，那么它在这个维度上更优。
- 选择该索引读表时，是否需要回表（即该索引生成的计划是 IndexReader 还是 IndexLookupReader）。不用回表的索引在这个维度上优于需要回表的索引。如果均需要回表，则比较索引的列涵盖了多少过滤条件。过滤条件指的是可以根据索引判断的 where 条件。如果某个索引的列集合涵盖的访问条件越多，则回表数量越少，那么它在这个维度上越优。
- 选择该索引是否能满足一定的顺序。因为索引的读取可以保证某些列集合的顺序，所以满足查询要求顺序的索引在这个维度上优于不满足的索引。

对于以上三种维度，如果索引 idx_a 在这三个维度上都不比 idx_b 差，且有一个维度比 idx_b 好，那么 TiDB 会优先选择 idx_a。在执行 EXPLAIN FORMAT = 'verbose' ... 语句时，如果 Skyline-Pruning 排除了某些索引，TiDB 会输出一条 NOTE 级别的 warning 提示哪些索引在 Skyline-Pruning 排除之后保留下来。

在以下示例中，索引 idx_b 和 idx_e 均劣于 idx_b_c，因而被 Skyline-Pruning 排除，SHOW WARNING 的返回结果显示了经过 Skyline-Pruning 后剩余的索引。

```
mysql> CREATE TABLE t(a INT PRIMARY KEY, b INT, c INT, d INT, e INT, INDEX idx_b(b), INDEX
    ↪ idx_b_c(b, c), INDEX idx_e(e));
Query OK, 0 rows affected (0.01 sec)

mysql> EXPLAIN FORMAT = 'verbose' SELECT * FROM t WHERE b = 2 AND c > 4;
+--
    ↪ ----------------------------+---------+---------+-----------+---------------------------+-----------
    ↪
| id                            | estRows | estCost | task      | access object             |
    ↪ operator info                                   |
```

```
+--
    ↪ ----------------------------+---------+---------+-----------+-----------------------------+------------
    ↪
| IndexLookUp_10                | 33.33   | 738.29  | root      |                             |
    ↪                                                   |
| ├─IndexRangeScan_8(Build)     | 33.33   | 2370.00 | cop[tikv] | table:t, index:idx_b_c(b, c) |
    ↪  range:(2 4,2 +inf], keep order:false, stats:pseudo |
| └─TableRowIDScan_9(Probe)     | 33.33   | 2370.00 | cop[tikv] | table:t                     |
    ↪  keep order:false, stats:pseudo                    |
+--
    ↪ ----------------------------+---------+---------+-----------+-----------------------------+------------
    ↪
3 rows in set, 1 warning (0.00 sec)

mysql> SHOW WARNINGS;
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
| Level | Code | Message
    ↪                                                                                                  |
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
| Note  | 1105 | [t,idx_b_c] remain after pruning paths for t given Prop{SortItems: [], TaskTp:
    ↪ rootTask} |
+--
    ↪ -----+------+-----------------------------------------------------------------------------------------+
    ↪
1 row in set (0.00 sec)
```

基于代价选择

在使用 Skyline-Pruning 规则排除了不合适的索引之后，索引的选择完全基于代价估算，读表的代价估算需要考虑以下几个方面：

- 索引的每行数据在存储层的平均长度。
- 索引生成的查询范围的行数量。
- 索引的回表代价。
- 索引查询时的范围数量。

根据这些因子和代价模型，优化器会选择一个代价最低的索引进行读表。

代价选择调优的常见问题

1. 估算的行数量不准确？

   一般是统计信息过期或者准确度不够造成的，可以重新执行 `analyze table` 或者修改 `analyze table` 的参数。

2. 统计信息准确，为什么读 TiFlash 更快，而优化器选择了 TiKV？

   目前区别 TiFlash 和 TiKV 的代价模型还比较粗糙，可以调小 tidb_opt_seek_factor 的值，让优化器倾向于选择 TiFlash。

3. 统计信息准确，某个索引要回表，但是它比另一个不用回表的索引实际执行更快，为什么选择了不用回表的索引？

   碰到这种情况，可能是代价估算时对于回表的代价计算得过大，可以调小 tidb_opt_network_factor，降低回表的代价。

控制索引的选择

通过Optimizer Hints 可以实现单条查询对索引选择的控制。

- USE_INDEX/IGNORE_INDEX 可以强制优化器使用/不使用某些索引。
- READ_FROM_STORAGE 可以强制优化器对于某些表选择 TiKV/TiFlash 的存储引擎进行查询。

10.3.3.3.3 统计信息简介

TiDB 使用统计信息来决定索引的选择。变量 tidb_analyze_version 用于控制所收集到的统计信息。目前 TiDB 中支持两种统计信息：tidb_analyze_version = 1 以及 tidb_analyze_version = 2。在 v5.3.0 及之后的版本中，该变量的默认值为 2，作为实验特性启用。如果从 v5.3.0 之前版本的集群升级至 v5.3.0 及之后的版本，tidb_analyze_version 的默认值不发生变化。

> 注意：
>
> 当 tidb_analyze_version = 2 时，如果执行 ANALYZE 语句后发生 OOM，请设置全局变量 tidb_analyze_version = 1，然后进行以下操作之一：
>
> - 如果 ANALYZE 语句是手动执行的，请手动 analyze 每张需要的表：

```
sql  select distinct(concat('ANALYZE ',table_schema, '.', table_name,';'))from
↪ information_schema.tables, mysql.stats_histograms where stats_ver = 2 and
↪ table_id = tidb_table_id ;
```

> - 如果 ANALYZE 语句是开启了自动 analyze 后 TiDB 自动执行的，请使用以下 SQL 语句生成 DROP STATS 的语句并执行：

```
sql  select distinct(concat('DROP STATS ',table_schema, '.', table_name,';'))from
↪  information_schema.tables, mysql.stats_histograms where stats_ver = 2 and
↪ table_id = tidb_table_id ;
```

两种版本中，TiDB 维护的统计信息如下：

信息	Version 1	Version 2
表的总行数	√	√
列的 Count-Min Sketch	√	×
索引的 Count-Min Sketch	√	×
列的 Top-N	√	√（改善了维护方式和精度）
索引的 Top-N	√（维护精度不足，会产生较大误差）	√（改善了维护方式和精度）
列的直方图	√	√（直方图中不包含 Top-N 中出现的值）
索引的直方图	√	√（直方图的桶中记录了各自的不同值的个数，且直方图不包
列的 NULL 值个数	√	√
索引的 NULL 值个数	√	√
列的平均长度	√	√
索引的平均长度	√	√

Version 2 的统计信息避免了 Version 1 中因为哈希冲突导致的在较大的数据量中可能产生的较大误差，并保持了大多数场景中的估算精度。

本文接下来将简单介绍其中出现的直方图和 Count-Min Sketch 以及 Top-N 这些数据结构，以及详细介绍统计信息的收集和维护。

直方图简介

直方图是一种对数据分布情况进行描述的工具，它会按照数据的值大小进行分桶，并用一些简单的数据来描述每个桶，比如落在桶里的值的个数。在 TiDB 中，会对每个表具体的列构建一个等深直方图，区间查询的估算便是借助该直方图来进行。

等深直方图，就是让落入每个桶里的值数量尽量相等。举个例子，比方说对于给定的集合 {1.6, 1.9, 1.9, 2.0, 2.4, 2.6, 2.7, 2.7, 2.8, 2.9, 3.4, 3.5}，并且生成 4 个桶，那么最终的等深直方图就会如下图所示，包含四个桶 [1.6, 1.9]，[2.0, 2.6]，[2.7, 2.8]，[2.9, 3.5]，其桶深均为 3。

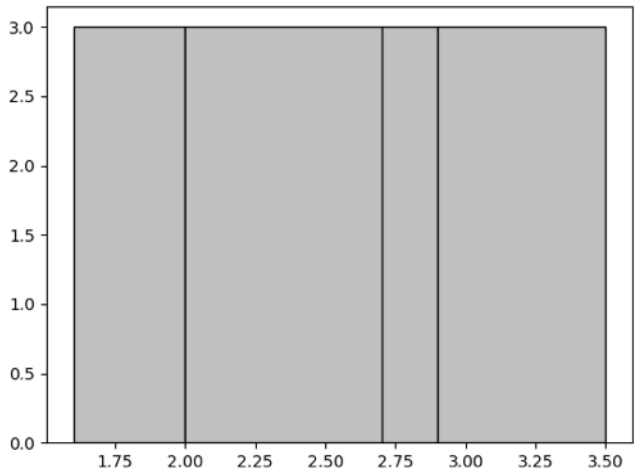


图 211: 等深直方图示例

在手动收集统计信息一节中有控制直方图桶数量上限的参数。当桶数量越多，直方图的估算精度就越高，不过也会同时增大统计信息的内存使用，可以视具体情况来做调整。

Count-Min Sketch

Count-Min Sketch 是一种哈希结构，当查询中出现诸如 `a = 1` 或者 `IN` 查询（如 `a in (1, 2, 3)`）这样的等值查询时，TiDB 便会使用这个数据结构来进行估算。

由于 Count-Min Sketch 是一个哈希结构，就有出现哈希碰撞的可能。当在 `EXPLAIN` 语句中发现等值查询的估算偏离实际值较大时，就可以认为是一个比较大的值和一个比较小的值被哈希到了一起。这时有以下两种手段来避免这个情况：

- 修改手动收集统计信息中提到的 `WITH NUM TOPN` 参数。TiDB 会将出现频率前 x 大的数据单独储存，之后的数据再储存到 Count-Min Sketch 中。因此可以调大这个值来避免一个比较大的值和一个比较小的值被哈希到一起。在 TiDB 中，这个参数的默认值是 20，最大可以设置为 1024。
- 修改统计信息的收集 -手动收集中提到的 `WITH NUM CMSKETCH DEPTH` 和 `WITH NUM CMSKETCH WIDTH` 两个参数，这两个参数会影响哈希的桶数和碰撞概率，可是适当调大来减少冲突概率，同时它会影响统计信息的内存使用，可以视具体情况来调整。在 TiDB 中，`DEPTH` 的默认值是 5，`WIDTH` 的默认值是 2048。

Top-N values

Top-N 即是这个列或者这个索引中，出现次数前 n 的值。TiDB 会记录前 n 个值的具体的值以及出现次数。

统计信息的收集

手动收集

可以通过执行 `ANALYZE` 语句来收集统计信息。

> 注意：
>
> 在 TiDB 中执行 `ANALYZE TABLE` 语句比在 MySQL 或 InnoDB 中耗时更长。InnoDB 采样的只是少量页面，但 TiDB 会完全重构一系列统计信息。适用于 MySQL 的脚本会误以为执行 `ANALYZE TABLE` 耗时较短。
>
> 如需更快的分析速度，可将 `tidb_enable_fast_analyze` 设置为 1 来打开快速分析功能。该参数的默认值为 0。
>
> 快速分析功能开启后，TiDB 会随机采样约 10000 行的数据来构建统计信息。因此在数据分布不均匀或者数据量比较少的情况下，统计信息的准确度会比较差。可能导致执行计划不优，比如选错索引。如果可以接受普通 `ANALYZE` 语句的执行时间，则推荐关闭快速分析功能。
>
> `tidb_enable_fast_analyze` 为实验性功能，目前与 `tidb_analyze_version=2` 的统计信息不完全匹配。因此开启 `tidb_enable_fast_analyze` 时需要将 `tidb_analyze_version` 的值设置为 1。

全量收集

可以通过以下几种语法进行全量收集。

收集 TableNameList 中所有表的统计信息：

```
ANALYZE TABLE TableNameList [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|CMSKETCH WIDTH]|[WITH NUM
    ↪ SAMPLES|WITH FLOATNUM SAMPLERATE];
```

- WITH NUM BUCKETS 用于指定生成直方图的桶数量上限。
- WITH NUM TOPN 用于指定生成 TOPN 数目的上限。
- WITH NUM CMSKETCH DEPTH 用于指定 CM Sketch 的长。
- WITH NUM CMSKETCH WIDTH 用于指定 CM Sketch 的宽。
- WITH NUM SAMPLES 用于指定采样的数目。
- WITH FLOAT_NUM SAMPLERATE 用于指定采样率。

WITH NUM SAMPLES 与 WITH FLOAT_NUM SAMPLERATE 这两种设置对应了两种不同的收集采样的算法。

- WITH NUM SAMPLES 指定了采样集的大小，在 TiDB 中是以蓄水池采样的方式实现。当表较大时，不推荐使用这种方式收集统计信息。因为蓄水池采样中间结果集会产生一定的冗余结果，会对内存等资源造成额外的压力。
- WITH FLOAT_NUM SAMPLERATE 是在 v5.3.0 中引入的采样方式，指定的采样率的大小，是取值范围 (0, 1] 的参数。在 TiDB 中是以伯努利采样的方式实现，更适合对较大的表进行采样，在收集效率和资源使用上更有优势。

在 v5.3.0 之前 TiDB 采用蓄水池采样的方式收集统计信息。自 v5.3.0 版本起，TiDB Version 2 的统计信息默认会选取伯努利采样的方式收集统计信息。若要重新使用蓄水池采样的方式采样，可以使用 WITH NUM SAMPLES 语句。

> 注意：
>
> 目前采样率基于自适应算法进行计算。当你通过 SHOW STATS_META 可以观察到一个表的行数时，可通过这个行数去计算采集 10 万行所对应的采样率。如果你观察不到这个值，可通过 TABLE_STORAGE_STATS 表的 TABLE_KEYS 列作为另一个参考来计算采样率。
>
> 通常情况下，STATS_META 相对 TABLE_KEYS 更可信，但是通过TiDB Lightning 等方式导入数据结束后，STATS_META 结果是 0。为了处理这个情况，你可以在 STATS_META 的结果远小于 TABLE_KEYS 的结果时，使用 TABLE_KEYS 计算采样率。

收集部分列的统计信息

执行 SQL 语句时，优化器在大多数情况下只会用到部分列（例如，WHERE、JOIN、ORDER BY、GROUP BY 子句中出现的列）的统计信息，这些被用到的列称为 PREDICATE COLUMNS。

如果一个表有很多列，收集所有列的统计信息会有较大的开销。为了降低开销，你可以只收集指定列或者 PREDICATE COLUMNS 的统计信息供优化器使用。

> 注意：
>
> 收集部分列的统计信息的功能仅适用于 tidb_analyze_version = 2 的情况。

- 如果要收集指定列的统计信息，请使用以下语法：

```
ANALYZE TABLE TableName COLUMNS ColumnNameList [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|
    ↪ CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

其中，`ColumnNameList` 表示指定列的名称列表。如果需要指定多列，请使用用逗号 , 分隔列名。例如，`ANALYZE table t columns a, b`。该语法除了收集指定表中指定列的统计信息，将同时收集该表中索引列的统计信息以及所有索引的统计信息。

> 注意：
>
> 该语法为全量收集。例如，在使用该语法收集了 a 列和 b 列的统计信息之后，如果还想要增加收集 c 列的统计信息，需要在语法中同时指定这三列 `ANALYZE table t columns a ↪ , b, c`，而不是只指定新增的那一列 `ANALYZE TABLE t COLUMNS c`。

- 如果要收集 `PREDICATE COLUMNS` 的统计信息，请进行以下操作：

> 警告：
>
> 收集 `PREDICATE COLUMNS` 的统计信息目前为实验特性，不建议在生产环境中使用。

1. 设置系统变量`tidb_enable_column_tracking` 的值为 `ON` 开启 TiDB 对 `PREDICATE COLUMNS` 的收集。开启后，TiDB 将每隔 `100 * stats-lease` 时间将 `PREDICATE COLUMNS` 信息写入系统表 `mysql. ↪ column_stats_usage`。
2. 在业务的查询模式稳定以后，使用以下语法收集 `PREDICATE COLUMNS` 的统计信息。

```
ANALYZE TABLE TableName PREDICATE COLUMNS [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|
    ↪ CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

该语法除了收集指定表中 `PREDICATE COLUMNS` 的统计信息，将同时收集该表中索引列的统计信息以及所有索引的统计信息。

> 注意：
>
> - 如果系统表 `mysql.column_stats_usage` 中没有关于该表的`PREDICATE COLUMNS` 记录，执行以上语句会收集该表中所有列的统计信息以及所有索引的统计信息。
> - 使用该语法收集统计信息后，当执行一种新的类型的 SQL 查询时，优化器可能会暂时使用旧的或者 pseudo 的列统计信息，然后在下一次收集统计信息的时候收集该列的统计信息。

- 如果要收集所有列的统计信息以及所有索引的统计信息，可以使用以下语法：

```
ANALYZE TABLE TableName ALL COLUMNS [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|CMSKETCH WIDTH]|[
    ↪ WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

如果要持久化 ANALYZE 语句中列的配置（包括 COLUMNS ColumnNameList、PREDICATE COLUMNS、ALL COLUMNS），请设置系统变量 tidb_persist_analyze_options 的值设置为 ON 以开启ANALYZE 配置持久化特性。开启 ANALYZE 配置持久化特性后：

- 当 TiDB 自动收集统计信息或者你手动执行 ANALYZE 语句收集统计信息但未指定列的配置时，TiDB 会继续沿用之前持久化的配置。
- 当多次手动执行 ANALYZE 语句并指定列的配置时，TiDB 会使用最新一次 ANALYZE 指定的配置项覆盖上一次记录的持久化配置。

如果你想查看一个表中哪些列是 PREDICATE COLUMNS，哪些列的统计信息已经被收集，可以使用以下语法：

```
SHOW COLUMN_STATS_USAGE [ShowLikeOrWhere];
```

SHOW COLUMN_STATS_USAGE 会输出 6 列，具体如下：

列名	说明
Db_name	数据库名
Table_name	表名
Partition_name	分区名
Column_name	列名
Last_used_at	该列统计信息在最近一次查询优化中被用到的时间
Last_analyzed_at	该列统计信息最近一次被收集的时间

在以下示例中，执行 ANALYZE TABLE t PREDICATE COLUMNS; 后，TiDB 将收集 b，c，d 列的统计信息，其中 b 列是 PREDICATE COLUMN，c 列和 d 列是索引列。

```
SET GLOBAL tidb_enable_column_tracking = ON;
Query OK, 0 rows affected (0.00 sec)

CREATE TABLE t (a INT, b INT, c INT, d INT, INDEX idx_c_d(c, d));
Query OK, 0 rows affected (0.00 sec)

-- 在此查询中优化器用到了 b 列的统计信息。
SELECT * FROM t WHERE b > 1;
Empty set (0.00 sec)

-- 等待一段时间（100 * stats-lease）后，TiDB 将收集的 `PREDICATE COLUMNS` 写入 mysql.
    ↪ column_stats_usage。
-- 指定 `last_used_at IS NOT NULL` 表示显示 TiDB 收集到的 `PREDICATE COLUMNS`。
SHOW COLUMN_STATS_USAGE WHERE db_name = 'test' AND table_name = 't' AND last_used_at IS NOT NULL;
+---------+------------+----------------+-------------+---------------------+------------------+
| Db_name | Table_name | Partition_name | Column_name | Last_used_at        | Last_analyzed_at |
+---------+------------+----------------+-------------+---------------------+------------------+
| test    | t          |                | b           | 2022-01-05 17:21:33 | NULL             |
+---------+------------+----------------+-------------+---------------------+------------------+
```

```
1 row in set (0.00 sec)

ANALYZE TABLE t PREDICATE COLUMNS;
Query OK, 0 rows affected, 1 warning (0.03 sec)

-- 指定 `last_analyzed_at IS NOT NULL` 表示显示收集过统计信息的列。
SHOW COLUMN_STATS_USAGE WHERE db_name = 'test' AND table_name = 't' AND last_analyzed_at IS NOT
    ↪ NULL;
+--
    ↪ -------+------------+----------------+-------------+---------------------+-------------------+
    ↪
| Db_name | Table_name | Partition_name | Column_name | Last_used_at        | Last_analyzed_at
    ↪    |
+--
    ↪ -------+------------+----------------+-------------+---------------------+-------------------+
    ↪
| test    | t          |                | b           | 2022-01-05 17:21:33 | 2022-01-05 17:23:06
    ↪  |
| test    | t          |                | c           | NULL                | 2022-01-05 17:23:06
    ↪  |
| test    | t          |                | d           | NULL                | 2022-01-05 17:23:06
    ↪  |
+--
    ↪ -------+------------+----------------+-------------+---------------------+-------------------+
    ↪
3 rows in set (0.00 sec)
```

收集索引的统计信息

如果要收集 TableName 中 IndexNameList 里所有索引的统计信息，请使用以下语法：

```
ANALYZE TABLE TableName INDEX [IndexNameList] [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|CMSKETCH
    ↪ WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

当 IndexNameList 为空时，该语法将收集 TableName 中所有索引的统计信息。

> **注意：**
>
> 为了保证前后统计信息的一致性，当设置 `tidb_analyze_version=2` 时，该语句也会收集整个表的统计信息（包括所有列和所有索引的统计信息）而不限于索引的统计信息。

收集分区的统计信息

- 如果要收集 TableName 中所有的 PartitionNameList 中分区的统计信息，请使用以下语法：

```
ANALYZE TABLE TableName PARTITION PartitionNameList [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|
    ↪ CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

- 如果要收集 TableName 中所有的 PartitionNameList 中分区的索引统计信息，请使用以下语法：

```
ANALYZE TABLE TableName PARTITION PartitionNameList INDEX [IndexNameList] [WITH NUM BUCKETS|
    ↪ TOPN|CMSKETCH DEPTH|CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

- 当收集分区的统计信息时，如果只收集部分列的统计信息，请使用以下语法：

> 警告：
>
> 收集 PREDICATE COLUMNS 的统计信息目前为实验特性，不建议在生产环境中使用。

```
ANALYZE TABLE TableName PARTITION PartitionNameList [COLUMNS ColumnNameList|PREDICATE
    ↪ COLUMNS|ALL COLUMNS] [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|CMSKETCH WIDTH]|[WITH NUM
    ↪  SAMPLES|WITH FLOATNUM SAMPLERATE];
```

增量收集

对于类似时间列这样的单调不减列，在进行全量收集后，可以使用增量收集来单独分析新增的部分，以提高分析的速度。

> 注意：
>
> 1. 目前只有索引提供了增量收集的功能
> 2. 使用增量收集时，必须保证表上只有插入操作，且应用方需要保证索引列上新插入的值是单调不减的，否则会导致统计信息不准，影响 TiDB 优化器选择合适的执行计划

可以通过以下几种语法进行增量收集。

增量收集 TableName 中所有的 IndexNameList 中的索引列的统计信息：

```
ANALYZE INCREMENTAL TABLE TableName INDEX [IndexNameList] [WITH NUM BUCKETS|TOPN|CMSKETCH DEPTH|
    ↪ CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

增量收集 TableName 中所有的 PartitionNameList 中分区的索引列统计信息：

```
ANALYZE INCREMENTAL TABLE TableName PARTITION PartitionNameList INDEX [IndexNameList] [WITH NUM
    ↪ BUCKETS|TOPN|CMSKETCH DEPTH|CMSKETCH WIDTH]|[WITH NUM SAMPLES|WITH FLOATNUM SAMPLERATE];
```

自动更新

在发生增加，删除以及修改语句时，TiDB 会自动更新表的总行数以及修改的行数。这些信息会定期持久化下来，更新的周期是 20 * stats-lease，stats-lease 的默认值是 3s，如果将其指定为 0，那么将不会自动更新。

和统计信息自动更新相关的三个系统变量如下：

系统变量名	默认值	功能
tidb_auto_analyze_ratio	0.5	自动更新阈值
tidb_auto_analyze_start_time	00:00 +0000	一天中能够进行自动更新的开始时间
tidb_auto_analyze_end_time	23:59 +0000	一天中能够进行自动更新的结束时间

当某个表 tbl 的修改行数与总行数的比值大于 tidb_auto_analyze_ratio，并且当前时间在 tidb_auto_analyze_start_time 和 tidb_auto_analyze_end_time 之间时，TiDB 会在后台执行 ANALYZE TABLE tbl 语句自动更新这个表的统计信息。

在 v5.0 版本之前，执行查询语句时，TiDB 会以 feedback-probability 的概率收集反馈信息，并将其用于更新直方图和 Count-Min Sketch。从 v5.0 版本起，该功能默认关闭，暂不建议开启此功能。

控制 ANALYZE 并发度

执行 ANALYZE 语句的时候，你可以通过一些参数来调整并发度，以控制对系统的影响。

tidb_build_stats_concurrency

目前 ANALYZE 执行的时候会被切分成一个个小的任务，每个任务只负责某一个列或者索引。tidb_build_stats_concurrency 可以控制同时执行的任务的数量，其默认值是 4。

tidb_distsql_scan_concurrency

在执行分析普通列任务的时候，tidb_distsql_scan_concurrency 可以用于控制一次读取的 Region 数量，其默认值是 15。

tidb_index_serial_scan_concurrency

在执行分析索引列任务的时候，tidb_index_serial_scan_concurrency 可以用于控制一次读取的 Region 数量，其默认值是 1。

ANALYZE 配置持久化

从 v5.4.0 起，TiDB 支持 ANALYZE 配置持久化功能，方便后续收集统计信息时沿用已有配置项。

TiDB 支持持久化的配置项包括：

配置项	对应的 ANALYZE 语法
直方图桶数	WITH NUM BUCKETS
TopN 个数	WITH NUM TOPN

配置项	对应的 ANALYZE 语法
采样数	WITH NUM SAMPLES
采样率	WITH FLOATNUM SAMPLERATE
ANALYZE 的列的类型	AnalyzeColumnOption ::= (‘ALL COLUMNS’ | ‘PREDICATE COLUMNS’ | ‘COLUMNS’ ColumnNameList)
ANALYZE 的列	ColumnNameList ::= Identifier (‘,’ Identifier)*

开启 ANALYZE 配置持久化功能

ANALYZE 配置持久化功能默认开启（系统变量 tidb_analyze_version 为默认值 2，tidb_persist_analyze_options 为默认值 ON），用于记录手动执行 ANALYZE 语句时指定的持久化配置项。记录后，当 TiDB 下一次自动更新统计信息或者你手动收集统计信息但未指定配置项时，TiDB 会按照记录的配置项收集统计信息。

多次手动执行 ANALYZE 语句并指定持久化配置项时，TiDB 会使用最新一次 ANALYZE 指定的配置项覆盖上一次记录的持久化配置。

关闭 ANALYZE 配置持久化功能

要关闭 ANALYZE 配置持久化功能，请设置系统变量 tidb_persist_analyze_options 为 OFF。由于 ANALYZE 配置持久化功能在 tidb_analyze_version = 1 的情况下不适用，因此设置 tidb_analyze_version = 1 同样会达到关闭配置持久化的效果。

关闭 ANALYZE 配置持久化功能后，已持久化的配置记录不会被清除。因此，当再次开启该功能时，TiDB 会继续使用之前记录的持久化配置收集统计信息。

> 注意：
>
> 当再次开启 ANALYZE 配置持久化功能时，如果之前记录的持久化配置项已经不适用当前的数据，请手动执行 ANALYZE 语句并指定新的持久化配置项。

查看 ANALYZE 状态

在执行 ANALYZE 时，可以通过 SQL 语句来查看当前 ANALYZE 的状态。

语法如下：

```
SHOW ANALYZE STATUS [ShowLikeOrWhere];
```

该语句会输出 ANALYZE 的状态，可以通过使用 ShowLikeOrWhere 来筛选需要的信息。

目前 SHOW ANALYZE STATUS 会输出 7 列，具体如下：

列名	说明
table_schema	数据库名
table_name	表名
partition_name	分区名
job_info	任务具体信息。如果分析索引则会包含索引名
row_count	已经分析的行数
start_time	任务开始执行的时间
state	任务状态，包括 pending（等待）、running（正在执行）、finished（执行成功）和 failed（执行失败）

统计信息的查看

你可以通过一些语句来查看统计信息的状态。

表的元信息

你可以通过 SHOW STATS_META 来查看表的总行数以及修改的行数等信息。

语法如下：

其中，ShowLikeOrWhereOpt 部分的语法图为：

```
SHOW STATS_META [ShowLikeOrWhere];
```

目前 SHOW STATS_META 会输出 6 列，具体如下：

列名	说明
db_name	数据库名
table_name	表名
partition_name	分区名
update_time	更新时间
modify_count	修改的行数
row_count	总行数

> 注意：
>
> 在 TiDB 根据 DML 语句自动更新总行数以及修改的行数时，update_time 也会被更新，因此并不能认为 update_time 是最近一次发生 Analyze 的时间。

表的健康度信息

通过 SHOW STATS_HEALTHY 可以查看表的统计信息健康度，并粗略估计表上统计信息的准确度。当 modify_count >= row_count 时，健康度为 0；当 modify_count < row_count 时，健康度为 (1 - modify_count/row_count) * 100。

SHOW STATS_HEALTHY 的语法图为：



图 212: ShowStatsHealthy

其中，ShowLikeOrWhereOpt 部分的语法图为：

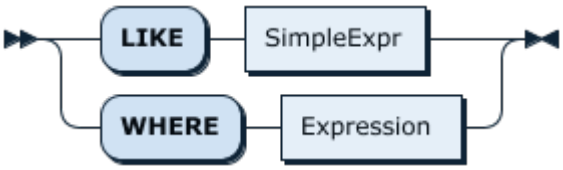


图 213: ShowLikeOrWhereOpt

目前，SHOW STATS_HEALTHY 会输出 4 列，具体如下：

列名	说明
db_name	数据库名
table_name	表名
partition_name	分区名
healthy	健康度

列的元信息

你可以通过 SHOW STATS_HISTOGRAMS 来查看列的不同值数量以及 NULL 数量等信息。

语法如下：

```
SHOW STATS_HISTOGRAMS [ShowLikeOrWhere];
```

该语句会输出所有列的不同值数量以及 NULL 数量等信息，你可以通过 ShowLikeOrWhere 来筛选需要的信息。

目前 SHOW STATS_HISTOGRAMS 会输出 10 列，具体如下：

列名	说明
db_name	数据库名
table_name	表名
partition_name	分区名
column_name	根据 is_index 来变化：is_index 为 0 时是列名，为 1 时是索引名
is_index	是否是索引列
update_time	更新时间
distinct_count	不同值数量
null_count	NULL 的数量
avg_col_size	列平均长度
correlation	该列与整型主键的皮尔逊系数，表示两列之间的关联程度

直方图桶的信息

你可以通过 SHOW STATS_BUCKETS 来查看直方图每个桶的信息。

语法如下：

```
SHOW STATS_BUCKETS [ShowLikeOrWhere];
```

语法图：

SHOW STATS_BUCKETS:



图 214: SHOW STATS_BUCKETS

该语句会输出所有桶的信息，你可以通过 ShowLikeOrWhere 来筛选需要的信息。

目前 SHOW STATS_BUCKETS 会输出 11 列，具体如下：

列名	说明
db_name	数据库名
table_name	表名
partition_name	分区名
column_name	根据 is_index 来变化：is_index 为 0 时是列名，为 1 时是索引名

列名	说明
is_index	是否是索引列
bucket_id	桶的编号
count	所有落在这个桶及之前桶中值的数量
repeats	最大值出现的次数
lower_bound	最小值
upper_bound	最大值
ndv	当前桶内不同值的个数。当 `tidb_analyze_version = 1` 时，该值恒为 0，没有实际意义。

Top-N 信息

你可以通过 `SHOW STATS_TOPN` 来查看当前 TiDB 中收集的 Top-N 值的信息。

语法如下：

```
SHOW STATS_TOPN [ShowLikeOrWhere];
```

目前 `SHOW STATS_TOPN` 会输出 7 列，具体如下：

列名	说明
db_name	数据库名
table_name	表名
partition_name	分区名
column_name	根据 is_index 来变化：is_index 为 0 时是列名，为 1 时是索引名
is_index	是否是索引列
value	该列的值
count	该值出现的次数

删除统计信息

可以通过执行 `DROP STATS` 语句来删除统计信息。

语法如下：

```
DROP STATS TableName;
```

该语句会删除 TableName 中所有的统计信息。

统计信息的加载

默认情况下，列的统计信息占用空间大小不同，TiDB 对统计信息的加载方式也会不同。

- 对于 count、distinctCount、nullCount 等占用空间较小的统计信息，只要有数据更新，TiDB 就会自动将对应的统计信息加载进内存供 SQL 优化阶段使用。
- 对于直方图、TopN、CMSketch 等占用空间较大的统计信息，为了确保 SQL 执行的性能，TiDB 会按需进行异步加载。例如，对于直方图，只有当某条 SQL 语句的优化阶段使用到了某列的直方图统计信息时，TiDB 才会将该列的直方图信息加载到内存。按需异步加载的优势是统计信息加载不会影响到 SQL 执行的性能，但在 SQL 优化时有可能使用不完整的统计信息。

从 v5.4.0 开始，TiDB 引入了统计信息同步加载的特性，支持执行当前 SQL 语句时将直方图、TopN、CMSketch 等占用空间较大的统计信息同步加载到内存，提高该 SQL 语句优化时统计信息的完整性。

> **警告：**
>
> 统计信息同步加载目前为实验性特性，不建议在生产环境中使用。

统计信息同步加载特性默认关闭。要开启该特性，请将系统变量 `tidb_stats_load_sync_wait` 的值设置为 SQL 优化等待加载列的完整统计信息的超时时间（单位为毫秒）。该变量默认值为 0，代表未开启。

开启该特性后，你可以进一步配置该特性：

- 通过修改系统变量`tidb_stats_load_pseudo_timeout` 的值控制 SQL 优化等待超时后的行为。该变量默认值为 OFF，代表超时后 SQL 执行失败。当设置该变量为 ON 时，整个 SQL 优化过程不会使用任何列上的直方图、TopN 或 CMSketch，而是退回使用 pseudo 的统计信息。
- 通过修改 TiDB 配置项`stats-load-concurrency` 的值控制统计信息同步加载可以并发处理的最大列数。该配置项的默认值为 5。
- 通过修改 TiDB 配置项`stats-load-queue-size` 的值设置统计信息同步加载最多可以缓存多少列的请求。该配置项的默认值为 1000。

统计信息的导入导出

导出统计信息

统计信息的导出接口如下。

通过以下接口可以获取数据库 `${db_name}` 中的表 `${table_name}` 的 json 格式的统计信息：

```
http://${tidb-server-ip}:${tidb-server-status-port}/stats/dump/${db_name}/${table_name}
```

通过以下接口可以获取数据库 `${db_name}` 中的表 `${table_name}` 在指定时间上的 json 格式统计信息。指定的时间应在 GC SafePoint 之后。

```
http://${tidb-server-ip}:${tidb-server-status-port}/stats/dump/${db_name}/${table_name}/${
    ↪ yyyyMMddHHmmss}
```

通过以下接口可以获取数据库 `${db_name}` 中的表 `${table_name}` 在指定时间上的 json 格式统计信息。指定的时间应在 GC SafePoint 之后。

```
http://${tidb-server-ip}:${tidb-server-status-port}/stats/dump/${db_name}/${table_name}/${yyyy-MM
    ↪ -dd HH:mm:ss}
```

导入统计信息

导入的统计信息一般是通过统计信息导出接口得到的 json 文件。

语法如下：

```
LOAD STATS 'file_name';
```

`file_name` 为要导入的统计信息的文件名。

另请参阅

- DROP STATS

10.3.3.3.4 错误索引的解决方案

在观察到某个查询的执行速度达不到预期时，可能是它的索引使用有误，这时就需要通过一些手段来解决。通常可以先使用表的健康度信息来查看统计信息的健康度。根据健康度可以分为以下两种情况处理。

健康度较低

这意味着距离 TiDB 上次 ANALYZE 很久了。这时可以先使用 ANALYZE 命令对统计信息进行更新。更新之后如果索引的使用上还是错误的，可以查看下一小节。

健康度接近 100%

这时意味着刚刚结束 ANALYZE 命令或者结束后不久。这时可能和 TiDB 对行数的估算逻辑有关。

对于等值查询，错误索引可能是由Count-Min Sketch 引起的。这时可以先检查是不是这种特殊情况，然后进行对应的处理。

如果经过检查发现不是上面的可能情况，可以使用Optimizer Hints 中提到的 USE_INDEX 或者 `use index` 来强制选择索引。同时也可以使用执行计划管理中提到的方式来非侵入地更改查询的行为。

其他情况

除去上述情况外，也存在因为数据的更新导致现有所有索引都不再适合的情况。这时就需要对条件和数据分布进行分析，查看是否有新的索引可以加快查询速度，然后使用 `ADD INDEX` 命令增加新的索引。

10.3.3.3.5 Distinct 优化

本文档介绍可用于 DISTINCT 的优化，包括简单 DISTINCT 和聚合函数 DISTINCT 的优化。

简单 DISTINCT

通常简单的 DISTINCT 会被优化成 GROUP BY 来执行。例如：

```
mysql> explain select DISTINCT a from t;
+--
    ↪ ----------------------+---------+-----------+---------------+-------------------------------------------
    ↪
| id                    | estRows | task      | access object | operator info
    ↪                                        |
+--
    ↪ ----------------------+---------+-----------+---------------+-------------------------------------------
    ↪
| HashAgg_6             | 2.40    | root      |               | group by:test.t.a, funcs:
    ↪ firstrow(test.t.a)->test.t.a |
| └─TableReader_11      | 3.00    | root      |               | data:TableFullScan_10
    ↪                                  |
```

```
|   └─TableFullScan_10     | 3.00    | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo                        |
+--
    ↪ ------------------------+---------+-----------+---------------+-------------------------------------
    ↪
3 rows in set (0.00 sec)
```

聚合函数 DISTINCT

通常来说，带有 DISTINCT 的聚合函数会单线程的在 TiDB 侧执行。使用系统变量 tidb_opt_distinct_agg_push_down ↪ 或者 TiDB 的配置项 distinct-agg-push-down 控制优化器是否执行带有 DISTINCT 的聚合函数（比如 select ↪ count(distinct a)from t）下推到 Coprocessor 的优化操作。

在以下示例中，tidb_opt_distinct_agg_push_down 开启前，TiDB 需要从 TiKV 读取所有数据，并在 TiDB 侧执行 disctinct。tidb_opt_distinct_agg_push_down 开启后，distinct a 被下推到了 Coprocessor，在 HashAgg_5 里新增了一个 group by 列 test.t.a。

```
mysql> desc select count(distinct a) from test.t;
+--
    ↪ ----------------------+----------+-----------+---------------+-------------------------------------
    ↪
| id                      | estRows  | task      | access object | operator info
    ↪                             |
+--
    ↪ ----------------------+----------+-----------+---------------+-------------------------------------
    ↪
| StreamAgg_6             | 1.00     | root      |               | funcs:count(distinct test.t.a)
    ↪ ->Column#4 |
| └─TableReader_10        | 10000.00 | root      |               | data:TableFullScan_9
    ↪                        |
|   └─TableFullScan_9     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo           |
+--
    ↪ ----------------------+----------+-----------+---------------+-------------------------------------
    ↪
3 rows in set (0.01 sec)

mysql> set session tidb_opt_distinct_agg_push_down = 1;
Query OK, 0 rows affected (0.00 sec)

mysql> desc select count(distinct a) from test.t;
+--
    ↪ ------------------------+----------+-----------+---------------+-------------------------------------
    ↪
| id                        | estRows  | task      | access object | operator info
    ↪                             |
+--
```

```
    ↪ ------------------------+----------+-----------+---------------+--------------------------------
    ↪
| HashAgg_8                 | 1.00     | root      |               | funcs:count(distinct test.t.
    ↪ a)->Column#3 |
| └─TableReader_9           | 1.00     | root      |               | data:HashAgg_5
    ↪                             |
|   └─HashAgg_5             | 1.00     | cop[tikv] |               | group by:test.t.a,
    ↪                              |
|     └─TableFullScan_7     | 10000.00 | cop[tikv] | table:t       | keep order:false, stats:
    ↪ pseudo           |
+--
    ↪ ------------------------+----------+-----------+---------------+--------------------------------
    ↪
4 rows in set (0.00 sec)
```

10.3.3.4 执行计划缓存

TiDB 支持对 Prepare / Execute 请求的执行计划缓存。其中包括以下两种形式的预处理语句：

- 使用 COM_STMT_PREPARE 和 COM_STMT_EXECUTE 的协议功能；
- 执行 Prepare / Execute SQL 语句查询；

TiDB 优化器对这两类查询的处理是一样的：Prepare 时将参数化的 SQL 查询解析成 AST（抽象语法树），每次 Execute 时根据保存的 AST 和具体的参数值生成执行计划。

当开启执行计划缓存后，每条 Prepare 语句的第一次 Execute 会检查当前查询是否可以使用执行计划缓存，如果可以则将生成的执行计划放进一个由 LRU 链表构成的缓存中；在后续的 Execute 中，会先从缓存中获取执行计划，并检查是否可用，如果获取和检查成功则跳过生成执行计划这一步，否则重新生成执行计划并放入缓存中。

在当前版本中，当 Prepare 语句符合以下条件任何一条，查询或者计划不会被缓存：

- SELECT、UPDATE、INSERT、DELETE、Union、Intersect、Except 以外的 SQL 语句；
- 访问分区表、临时表或访问表中包含生成列的查询；
- 包含子查询的查询，如 `select * from t where a > (select ...)`；
- 包含 ignore_plan_cache 这一 Hint 的查询，例如 `select /*+ ignore_plan_cache()*/ * from t`；
- 包含除 ? 外其他变量（即系统变量或用户自定义变量）的查询，例如 `select * from t where a>? and b>@x`；
- 查询包含无法被缓存函数。目前不能被缓存的函数有：database()、current_user、current_role、user、connection_id、last_insert_id、row_count、version、like；
- ? 在 Limit 后的查询，如 `Limit ?` 或者 `Limit 10, ?`，此时 ? 的具体值对查询性能影响较大，故不缓存；
- ? 直接在 Order By 后的查询，如 `Order By ?`，此时 ? 表示根据 Order By 后第几列排序，排序列不同的查询使用同一个计划可能导致错误结果，故不缓存；如果是普通表达式，如 `Order By a+?` 则会缓存；
- ? 紧跟在 Group by 后的查询，如 `Group By ?`，此时 ? 表示根据 Group By 后第几列聚合，聚合列不同的查询使用同一个计划可能导致错误结果，故不缓存；如果是普通表达式，如 `Group By a+?` 则会缓存；

- ? 出现在窗口函数 Window Frame 定义中的查询，如 (partition by year order by sale rows ? preceding)；如果 ? 出现在窗口函数的其他位置，则会缓存；
- 用参数进行 int 和 string 比较的查询，如 c_int >= ? 或者 c_int in (?, ?)等，其中 ? 为字符串类型，如 set @x='123'；此时为了保证结果和 MySQL 兼容性，需要每次对参数进行调整，故不会缓存；
- 会访问 TiFlash 的计划不会被缓存；
- 大部分情况下计划中含有 TableDual 的计划将将不会被缓存，除非当前执行的 Prepare 语句不含参数，则对应的 TableDual 计划可以被缓存。

LRU 链表是设计成 session 级别的缓存，因为 Prepare / Execute 不能跨 session 执行。LRU 链表的每个元素是一个 key-value 对，value 是执行计划，key 由如下几部分组成：

- 执行 Execute 时所在数据库的名字；
- Prepare 语句的标识符，即紧跟在 PREPARE 关键字后的名字；
- 当前的 schema 版本，每条执行成功的 DDL 语句会修改 schema 版本；
- 执行 Execute 时的 SQL Mode；
- 当前设置的时区，即系统变量 time_zone 的值；
- 系统变量 sql_select_limit 的值；

key 中任何一项变动（如切换数据库，重命名 Prepare 语句，执行 DDL，或修改 SQL Mode / time_zone 的值），或 LRU 淘汰机制触发都会导致 Execute 时无法命中执行计划缓存。

成功从缓存中获取到执行计划后，TiDB 会先检查执行计划是否依然合法，如果当前 Execute 在显式事务里执行，并且引用的表在事务前序语句中被修改，而缓存的执行计划对该表访问不包含 UnionScan 算子，则它不能被执行。

在通过合法性检测后，会根据当前最新参数值，对执行计划的扫描范围做相应调整，再用它执行获取数据。

关于执行计划缓存和查询性能有几点值得注意：

- 不管计划是否已经被缓存，都会受到 SQL Binding 的影响。对于没有被缓存的计划，即在第一次执行 Execute 时，会受到已有 SQL Binding 的影响；而对于已经缓存的计划，如果有新的 SQL Binding 被创建产生，则原有已经被缓存的计划会失效。
- 已经被缓存的计划不会受到统计信息更新、优化规则和表达式下推黑名单更新的影响，仍然会使用已经保存在缓存中的计划。
- 重启 TiDB 实例时（如不停机滚动升级 TiDB 集群），Prepare 信息会丢失，此时执行 execute stmt ... 可能会遇到 Prepared Statement not found 的错误，此时需要再执行一次 prepare stmt ...。
- 考虑到不同 Execute 的参数会不同，执行计划缓存为了保证适配性会禁止一些和具体参数值密切相关的激进查询优化手段，导致对特定的一些参数值，查询计划可能不是最优。比如查询的过滤条件为 where a > ? and a < ?，第一次 Execute 时参数分别为 2 和 1，考虑到这两个参数下次执行时可能会是 1 和 2，优化器不会生成对当前参数最优的 TableDual 执行计划。
- 如果不考虑缓存失效和淘汰，一份执行计划缓存会对应各种不同的参数取值，理论上也会导致某些取值下执行计划非最优。比如查询过滤条件为 where a < ?，假如第一次执行 Execute 时用的参数值为 1，此时优化器生成最优的 IndexScan 执行计划放入缓存，在后续执行 Exeucte 时参数变为 10000，此时 TableScan 可能才是更优执行计划，但由于执行计划缓存，执行时还是会使用先前生成的 IndexScan。因此执行计划缓存更适用于查询较为简单（查询编译耗时占比较高）且执行计划较为固定的业务场景。

目前执行计划缓存功能默认关闭，可以通过打开配置文件中prepare-plan-cache 项启用这项功能。

> **注意：**
>
> 执行计划缓存功能仅针对 Prepare / Execute 请求，对普通查询无效。

在开启了执行计划缓存功能后，可以通过 session 级别的系统变量 last_plan_from_cache 查看上一条 Execute 语句是否使用了缓存的执行计划，例如：

```
MySQL [test]> create table t(a int);
Query OK, 0 rows affected (0.00 sec)

MySQL [test]> prepare stmt from 'select * from t where a = ?';
Query OK, 0 rows affected (0.00 sec)

MySQL [test]> set @a = 1;
Query OK, 0 rows affected (0.00 sec)

-- 第一次 execute 生成执行计划放入缓存
MySQL [test]> execute stmt using @a;
Empty set (0.00 sec)

MySQL [test]> select @@last_plan_from_cache;
+------------------------+
| @@last_plan_from_cache |
+------------------------+
| 0                      |
+------------------------+
1 row in set (0.00 sec)

-- 第二次 execute 命中缓存
MySQL [test]> execute stmt using @a;
Empty set (0.00 sec)

MySQL [test]> select @@last_plan_from_cache;
+------------------------+
| @@last_plan_from_cache |
+------------------------+
| 1                      |
+------------------------+
1 row in set (0.00 sec)
```

如果发现某一组 Prepare / Execute 由于执行计划缓存导致了非预期行为，可以通过 SQL Hint ignore_plan_cache() 让该组语句不使用缓存。还是用上述的 stmt 为例：

```
MySQL [test]> prepare stmt from 'select /*+ ignore_plan_cache() */ * from t where a = ?';
```

```
Query OK, 0 rows affected (0.00 sec)

MySQL [test]> set @a = 1;
Query OK, 0 rows affected (0.00 sec)

MySQL [test]> execute stmt using @a;
Empty set (0.00 sec)

MySQL [test]> select @@last_plan_from_cache;
+------------------------+
| @@last_plan_from_cache |
+------------------------+
| 0                      |
+------------------------+
1 row in set (0.00 sec)

MySQL [test]> execute stmt using @a;
Empty set (0.00 sec)

MySQL [test]> select @@last_plan_from_cache;
+------------------------+
| @@last_plan_from_cache |
+------------------------+
| 0                      |
+------------------------+
1 row in set (0.00 sec)
```

10.3.4 控制执行计划

10.3.4.1 控制执行计划

SQL 性能调优的前两个章节介绍了如何理解 TiDB 的执行计划以及 TiDB 如何生成一个执行计划。本章节将介绍当你确定了执行计划所存在的问题时，可以使用哪些手段来控制执行计划的生成。本章节主要包括以下三方面内容：

- Optimizer Hints中，我们会介绍如何使用 Hint 来指导 TiDB 生成执行计划。
- 但是使用 Hint 会侵入性地更改 SQL，在一些场景下并不能简单的插入 Hint。在执行计划管理中，我们会介绍 TiDB 如何使用另一种语法来非侵入地控制执行计划的生成，同时还会介绍后台自动对执行计划进行演进的手段，来减轻诸如版本升级等原因造成的执行计划不稳定，而导致集群性能下降的问题。
- 最后在优化规则及表达式下推的黑名单中，我们会介绍黑名单的使用。

10.3.4.2 Optimizer Hints

TiDB 支持 Optimizer Hints 语法，它基于 MySQL 5.7 中介绍的类似 comment 的语法，例如 `/*+ HINT_NAME(t1, t2)*/`。当 TiDB 优化器选择的不是最优查询计划时，建议使用 Optimizer Hints。

> **注意：**
>
> MySQL 命令行客户端在 5.7.7 版本之前默认清除了 Optimizer Hints。如果需要在这些早期版本的客户端中使用 Hint 语法，需要在启动客户端时加上 `--comments` 选项，例如 `mysql -h 127.0.0.1 -P 4000 -uroot --comments`。

10.3.4.2.1 语法

Optimizer Hints 不区分大小写，通过 `/*+ ... */` 注释的形式跟在 SELECT、UPDATE 或 DELETE 关键字的后面。INSERT 关键字后不支持 Optimizer Hints。

多个不同的 Hint 之间需用逗号隔开，例如：

```
SELECT /*+ USE_INDEX(t1, idx1), HASH_AGG(), HASH_JOIN(t1) */ count(*) FROM t t1, t t2 WHERE t1.a = t2.b;
```

可以通过 `Explain`/`Explain Analyze` 语句的输出，来查看 Optimizer Hints 对查询执行计划的影响。

如果 Optimizer Hints 包含语法错误或不完整，查询语句不会报错，而是按照没有 Optimizer Hints 的情况执行。如果 Hint 不适用于当前语句，TiDB 会返回 Warning，用户可以在查询结束后通过 `Show Warnings` 命令查看具体信息。

> **注意：**
>
> 如果注释不是跟在指定的关键字后，会被当作是普通的 MySQL comment，注释不会生效，且不会上报 warning。

TiDB 目前支持的 Optimizer Hints 根据生效范围的不同可以划分为两类：第一类是在查询块范围生效的 Hint，例如`/*+ HASH_AGG()*/`；第二类是在整个查询范围生效的 Hint，例如`/*+ MEMORY_QUOTA(1024 MB)*/`。

每条语句中每一个查询和子查询都对应着一个不同的查询块，每个查询块有自己对应的名字。以下面这条语句为例：

```
SELECT * FROM (SELECT * FROM t) t1, (SELECT * FROM t) t2;
```

该查询语句有 3 个查询块，最外面一层 SELECT 所在的查询块的名字为 `sel_1`，两个 SELECT 子查询的名字依次为 `sel_2` 和 `sel_3`。其中数字序号根据 SELECT 出现的位置从左到右计数。如果分别用 DELETE 和 UPDATE 查询替代第一个 SELECT 查询，则对应的查询块名字分别为 `del_1` 和 `upd_1`。

10.3.4.2.2 查询块范围生效的 Hint

这类 Hint 可以跟在查询语句中任意 SELECT、UPDATE 或 DELETE 关键字的后面。通过在 Hint 中使用查询块名字可以控制 Hint 的生效范围，以及准确标识查询中的每一个表（有可能表的名字或者别名相同），方便明确 Hint 的参数指向。若不显式地在 Hint 中指定查询块，Hint 默认作用于当前查询块。以如下查询为例：

```
SELECT /*+ HASH_JOIN(@sel_1 t1@sel_1, t3) */ * FROM (SELECT t1.a, t1.b FROM t t1, t t2 WHERE t1.a
    ↪  = t2.a) t1, t t3 WHERE t1.b = t3.b;
```

该 Hint 在 `sel_1` 这个查询块中生效，参数分别为 `sel_1` 中的 `t1` 表（`sel_2` 中也有一个 `t1` 表）和 `t3` 表。

如上例所述，在 Hint 中使用查询块名字的方式有两种：第一种是作为 Hint 的第一个参数，与其他参数用空格隔开。除 QB_NAME 外，本节所列的所有 Hint 除自身明确列出的参数外都有一个隐藏的可选参数 @QB_NAME，通过使用这个参数可以指定该 Hint 的生效范围；第二种在 Hint 中使用查询块名字的方式是在参数中的某一个表名后面加 @QB_NAME，用以明确指出该参数是哪个查询块中的表。

> 注意：
>
> Hint 声明的位置必须在指定生效的查询块之中或之前，不能是在之后的查询块中，否则无法生效。

QB_NAME

当查询语句是包含多层嵌套子查询的复杂语句时，识别某个查询块的序号和名字很可能会出错，Hint QB_NAME 可以方便我们使用查询块。QB_NAME 是 Query Block Name 的缩写，用于为某个查询块指定新的名字，同时查询块原本默认的名字依然有效。例如：

```
SELECT /*+ QB_NAME(QB1) */ * FROM (SELECT * FROM t) t1, (SELECT * FROM t) t2;
```

这条 Hint 将最外层 SELECT 查询块的命名为 QB1，此时 QB1 和默认名称 `sel_1` 对于这个查询块来说都是有效的。

> 注意：
>
> 上述例子中，如果指定的 QB_NAME 为 `sel_2`，并且不给原本 `sel_2` 对应的第二个查询块指定新的 QB_NAME，则第二个查询块的默认名字 `sel_2` 会失效。

MERGE_JOIN(t1_name [, tl_name …])

`MERGE_JOIN(t1_name [, tl_name ...])` 提示优化器对指定表使用 Sort Merge Join 算法。这个算法通常会占用更少的内存，但执行时间会更久。当数据量太大，或系统内存不足时，建议尝试使用。例如：

```
SELECT /*+ MERGE_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

> 注意：
>
> MERGE_JOIN 的别名是 TIDB_SMJ，在 3.0.x 及之前版本仅支持使用该别名；之后的版本同时支持使用这两种名称，但推荐使用 MERGE_JOIN。

INL_JOIN(t1_name [, tl_name …])

`INL_JOIN(t1_name [, tl_name ...])` 提示优化器对指定表使用 Index Nested Loop Join 算法。这个算法可能会在某些场景更快，消耗更少系统资源，有的场景会更慢，消耗更多系统资源。对于外表经过 WHERE 条件过滤后结果集较小（小于 1 万行）的场景，可以尝试使用。例如：

```
SELECT /*+ INL_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

`INL_JOIN()` 中的参数是建立查询计划时内表的候选表，比如 `INL_JOIN(t1)` 只会考虑使用 t1 作为内表构建查询计划。表如果指定了别名，就只能使用表的别名作为 `INL_JOIN()` 的参数；如果没有指定别名，则用表的本名作为其参数。比如在 `SELECT /*+ INL_JOIN(t1)*/ * FROM t t1, t t2 WHERE t1.a = t2.b;` 中，`INL_JOIN()` 的参数只能使用 t 的别名 t1 或 t2，不能用 t。

> 注意：
>
> `INL_JOIN` 的别名是 `TIDB_INLJ`，在 3.0.x 及之前版本仅支持使用该别名；之后的版本同时支持使用这两种名称，但推荐使用 `INL_JOIN`。

INL_HASH_JOIN

`INL_HASH_JOIN(t1_name [, tl_name])` 提示优化器使用 Index Nested Loop Hash Join 算法。该算法与 Index Nested Loop Join 使用条件完全一样，两者的区别是 `INL_JOIN` 会在连接的内表上建哈希表，而 `INL_HASH_JOIN` 会在连接的外表上建哈希表，后者对于内存的使用是有固定上限的，而前者使用的内存使用取决于内表匹配到的行数。

HASH_JOIN(t1_name [, tl_name …])

`HASH_JOIN(t1_name [, tl_name ...])` 提示优化器对指定表使用 Hash Join 算法。这个算法多线程并发执行，执行速度较快，但会消耗较多内存。例如：

```
SELECT /*+ HASH_JOIN(t1, t2) */ * FROM t1, t2 WHERE t1.id = t2.id;
```

> 注意：
>
> `HASH_JOIN` 的别名是 `TIDB_HJ`，在 3.0.x 及之前版本仅支持使用该别名；之后的版本同时支持使用这两种名称，推荐使用 `HASH_JOIN`。

HASH_AGG()

`HASH_AGG()` 提示优化器对指定查询块中所有聚合函数使用 Hash Aggregation 算法。这个算法多线程并发执行，执行速度较快，但会消耗较多内存。例如：

```
SELECT /*+ HASH_AGG() */ count(*) FROM t1, t2 WHERE t1.a > 10 GROUP BY t1.id;
```

STREAM_AGG()

STREAM_AGG() 提示优化器对指定查询块中所有聚合函数使用 Stream Aggregation 算法。这个算法通常会占用更少的内存，但执行时间会更久。数据量太大，或系统内存不足时，建议尝试使用。例如：

```
SELECT /*+ STREAM_AGG() */ count(*) FROM t1, t2 WHERE t1.a > 10 GROUP BY t1.id;
```

USE_INDEX(t1_name, idx1_name [, idx2_name …])

USE_INDEX(t1_name, idx1_name [, idx2_name ...]) 提示优化器对指定表仅使用给出的索引。

下面例子的效果等价于 SELECT * FROM t t1 use index(idx1, idx2);：

```
SELECT /*+ USE_INDEX(t1, idx1, idx2) */ * FROM t1;
```

> 注意：
>
> 当该 Hint 中只指定表名，不指定索引名时，表示不考虑使用任何索引，而是选择全表扫。

IGNORE_INDEX(t1_name, idx1_name [, idx2_name …])

IGNORE_INDEX(t1_name, idx1_name [, idx2_name ...]) 提示优化器对指定表忽略给出的索引。

下面例子的效果等价于 SELECT * FROM t t1 ignore index(idx1, idx2);：

```
SELECT /*+ IGNORE_INDEX(t1, idx1, idx2) */ * FROM t t1;
```

AGG_TO_COP()

AGG_TO_COP() 提示优化器将指定查询块中的聚合函数下推到 coprocessor。如果优化器没有下推某些适合下推的聚合函数，建议尝试使用。例如：

```
SELECT /*+ AGG_TO_COP() */ sum(t1.a) FROM t t1;
```

LIMIT_TO_COP()

LIMIT_TO_COP() 提示优化器将指定查询块中的 Limit 和 TopN 算子下推到 coprocessor。优化器没有下推 Limit 或者 TopN 算子时建议尝试使用该提示。例如：

```
SELECT /*+ LIMIT_TO_COP() */ * FROM t WHERE a = 1 AND b > 10 ORDER BY c LIMIT 1;
```

READ_FROM_STORAGE(TIFLASH[t1_name [, tl_name …]], TIKV[t2_name [, tl_name …]])

READ_FROM_STORAGE(TIFLASH[t1_name [, tl_name ...]], TIKV[t2_name [, tl_name ...]]) 提示优化器从指定的存储引擎来读取指定的表，目前支持的存储引擎参数有 TIKV 和 TIFLASH。如果为表指定了别名，就只能使用表的别名作为 READ_FROM_STORAGE() 的参数；如果没有指定别名，则用表的本名作为其参数。例如：

```
SELECT /*+ READ_FROM_STORAGE(TIFLASH[t1], TIKV[t2]) */ t1.a FROM t t1, t t2 WHERE t1.a = t2.a;
```

> **注意：**
>
> 如果需要提示优化器使用的表不在同一个数据库内，需要显式指定数据库名。例如 `SELECT /*+ READ_FROM_STORAGE(TIFLASH[test1.t1,test2.t2])*/ t1.a FROM test1.t t1, test2.t t2 WHERE t1.a = t2.a;`。

USE_INDEX_MERGE(t1_name, idx1_name [, idx2_name …])

`USE_INDEX_MERGE(t1_name, idx1_name [, idx2_name ...])` 提示优化器通过 index merge 的方式来访问指定的表，其中索引列表为可选参数。若显式地指出索引列表，会尝试在索引列表中选取索引来构建 index merge。若不给出索引列表，会尝试在所有可用的索引中选取索引来构建 index merge。例如：

```
SELECT /*+ USE_INDEX_MERGE(t1, idx_a, idx_b, idx_c) */ * FROM t1 WHERE t1.a > 10 OR t1.b > 10;
```

当对同一张表有多个 `USE_INDEX_MERGE` Hint 时，优化器会从这些 Hint 指定的索引列表的并集中尝试选取索引。

> **注意：**
>
> `USE_INDEX_MERGE` 的参数是索引名，而不是列名。对于主键索引，索引名为 `primary`。

目前该 Hint 生效的条件较为苛刻，包括：

- 如果查询有除了全表扫以外的单索引扫描方式可以选择，优化器不会选择 index merge；

10.3.4.2.3 查询范围生效的 Hint

这类 Hint 只能跟在语句中第一个 `SELECT`、`UPDATE` 或 `DELETE` 关键字的后面，等同于在当前这条查询运行时对指定的系统变量进行修改，其优先级高于现有系统变量的值。

> **注意：**
>
> 这类 Hint 虽然也有隐藏的可选变量 `@QB_NAME`，但就算指定了该值，Hint 还是会在整个查询范围生效。

NO_INDEX_MERGE()

`NO_INDEX_MERGE()` 会关闭优化器的 index merge 功能。

下面的例子不会使用 index merge：

```
SELECT /*+ NO_INDEX_MERGE() */ * FROM t WHERE t.a > 0 or t.b > 0;
```

除了 Hint 外，系统变量 tidb_enable_index_merge 也能决定是否开启该功能。

> 注意：
>
> - NO_INDEX_MERGE 优先级高于 USE_INDEX_MERGE，当这两类 Hint 同时存在时，USE_INDEX_MERGE 不会生效。
> - 当存在子查询时，NO_INDEX_MERGE 放在最外层才能生效。

USE_TOJA(boolean_value)

参数 boolean_value 可以是 TRUE 或者 FALSE。USE_TOJA(TRUE) 会开启优化器尝试将 in (subquery) 条件转换为 join 和 aggregation 的功能。相对地，USE_TOJA(FALSE) 会关闭该功能。

下面的例子会将 in (SELECT t2.a FROM t2)subq 转换为等价的 join 和 aggregation：

```
SELECT /*+ USE_TOJA(TRUE) */ t1.a, t1.b FROM t1 WHERE t1.a in (SELECT t2.a FROM t2) subq;
```

除了 Hint 外，系统变量 tidb_opt_insubq_to_join_and_agg 也能决定是否开启该功能。

MAX_EXECUTION_TIME(N)

MAX_EXECUTION_TIME(N) 把语句的执行时间限制在 N 毫秒以内，超时后服务器会终止这条语句的执行。

下面的 Hint 设置了 1000 毫秒（即 1 秒）超时：

```
SELECT /*+ MAX_EXECUTION_TIME(1000) */ * FROM t1 inner join t2 WHERE t1.id = t2.id;
```

除了 Hint 之外，系统变量 global.max_execution_time 也能对语句执行时间进行限制。

MEMORY_QUOTA(N)

MEMORY_QUOTA(N) 用于限制语句执行时的内存使用。该 Hint 支持 MB 和 GB 两种单位。内存使用超过该限制时会根据当前设置的内存超限行为来打出一条 log 或者终止语句的执行。

下面的 Hint 设置了 1024 MB 的内存限制：

```
SELECT /*+ MEMORY_QUOTA(1024 MB) */ * FROM t;
```

除了 Hint 外，系统变量 tidb_mem_quota_query 也能限制语句执行的内存使用。

READ_CONSISTENT_REPLICA()

READ_CONSISTENT_REPLICA() 会开启从数据一致的 TiKV follower 节点读取数据的特性。

下面的例子会从 follower 节点读取数据：

```
SELECT /*+ READ_CONSISTENT_REPLICA() */ * FROM t;
```

除了 Hint 外，环境变量 tidb_replica_read 设为 'follower' 或者 'leader' 也能决定是否开启该特性。

IGNORE_PLAN_CACHE()

IGNORE_PLAN_CACHE() 提示优化器在处理当前 prepare 语句时不使用 plan cache。

该 Hint 用于在prepare-plan-cache 开启的场景下临时对某类查询禁用 plan cache。

以下示例强制该 prepare 语句不使用 plan cache：

```
prepare stmt FROM 'SELECT  /*+ IGNORE_PLAN_CACHE() */ * FROM t WHERE t.id = ?';
```

NTH_PLAN(N)

NTH_PLAN(N) 提示优化器选用在物理优化阶段搜索到的第 N 个物理计划。N 必须是正整数。

如果指定的 N 超出了物理优化阶段的搜索范围，TiDB 会返回 warning，并根据不存在该 Hint 时一样的策略选择最优物理计划。

该 Hint 在启用 cascades planner 的情况下不会生效。

以下示例会强制优化器在物理阶段选择搜索到的第 3 个物理计划：

```
SELECT /*+ NTH_PLAN(3) */ count(*) from t where a > 5;
```

> 注意：
>
> NTH_PLAN(N) 主要用于测试用途，并且在未来不保证其兼容性，请谨慎使用。

10.3.4.3 执行计划管理 (SPM)

执行计划管理，又称 SPM (SQL Plan Management)，是通过执行计划绑定，对执行计划进行人为干预的一系列功能，包括执行计划绑定、自动捕获绑定、自动演进绑定等。

10.3.4.3.1 执行计划绑定 (SQL Binding)

执行计划绑定是 SPM 的基础。在优化器 Hints 中介绍了可以通过 Hint 的方式选择指定的执行计划，但有时需要在不修改 SQL 语句的情况下干预执行计划的选择。执行计划绑定功能使得可以在不修改 SQL 语句的情况下选择指定的执行计划。

创建绑定

```
CREATE [GLOBAL | SESSION] BINDING FOR BindableStmt USING BindableStmt;
```

该语句可以在 GLOBAL 或者 SESSION 作用域内为 SQL 绑定执行计划。目前支持的可创建执行计划绑定的 SQL 类型 (BindableStmt) 包括：SELECT，DELETE，UPDATE 和带有 SELECT 子查询的 INSERT/REPLACE。

其中，有两类特定的语法由于语法冲突不能创建执行计划绑定，例如：

```
-- 类型一：使用 `join` 关键字但不通过 `using` 关键字指定关联列的笛卡尔积
create global binding for
    select * from t t1 join t t2
using
    select * from t t1 join t t2;
```

```
-- 类型二：包含了 `using` 关键字的 `delete` 语句
create global binding for
    delete from t1 using t1 join t2 on t1.a = t2.a
using
    delete from t1 using t1 join t2 on t1.a = t2.a;
```

可以通过等价的 SQL 改写绕过这个语法冲突的问题。例如，上述两个例子可以改写为：

```
-- 类型一的第一种改写：为 `join` 关键字添加 `using` 子句
create global binding for
    select * from t t1 join t t2 using (a)
using
    select * from t t1 join t t2 using (a);

-- 类型一的第二种改写：去掉 `join` 关键字
create global binding for
    select * from t t1, t t2
using
    select * from t t1, t t2;

-- 类型二的改写：去掉 `delete` 语句中的 `using` 关键字
create global binding for
    delete t1 from t1 join t2 on t1.a = t2.a
using
    delete t1 from t1 join t2 on t1.a = t2.a;
```

> **注意：**
>
> 在对带 SELECT 子查询的 INSERT/REPLACE 语句创建执行计划绑定时，需要将想要绑定的优化器 Hints 指定在 SELECT 子查询中，而不是 INSERT/REPLACE 关键字后，不然优化器 Hints 不会生效。

例如：

```
-- Hint 能生效的用法
create global binding for
    insert into t1 select * from t2 where a > 1 and b = 1
using
    insert into t1 select /*+ use_index(@sel_1 t2, a) */ * from t2 where a > 1 and b = 1;

-- Hint 不能生效的用法
create global binding for
    insert into t1 select * from t2 where a > 1 and b = 1
using
    insert /*+ use_index(@sel_1 t2, a) */ into t1 select * from t2 where a > 1 and b = 1;
```

如果在创建执行计划绑定时不指定作用域，隐式作用域 SESSION 会被使用。TiDB 优化器会将被绑定的 SQL 进行“标准化”处理，然后存储到系统表中。在处理 SQL 查询时，只要“标准化”后的 SQL 和系统表中某个被绑定的 SQL 语句一致，并且系统变量 tidb_use_plan_baselines 的值为 on（其默认值为 on），即可使用相应的优化器 Hint。如果存在多个可匹配的执行计划，优化器会从中选择代价最小的一个进行绑定。

标准化：把 SQL 中的常量变成变量参数，对空格和换行符等做标准化处理，并对查询引用到的表显式指定数据库。例如：

```
select * from t where a >    1
-- 以上语句标准化后如下：
select * from test . t where a > ?
```

> 注意：
>
> 在进行标准化的时候，被逗号 , 连接起来的多个常量会被标准化为 ... 而不是 ?。
>
> 例如：
>
> ```
> select * from t limit 10
> select * from t limit 10, 20
> select * from t where a in (1)
> select * from t where a in (1,2,3)
> -- 以上语句标准化后如下：
> select * from test . t limit ?
> select * from test . t limit ...
> select * from test . t where a in ( ? )
> select * from test . t where a in ( ... )
> ```
>
> 因此包含单个常量的 SQL 语句和包含被逗号连接起来多个常量的 SQL 语句，在被绑定时会被 TiDB 视作不同的 SQL 语句，需要分别创建绑定。

值得注意的是，如果一条 SQL 语句在 GLOBAL 和 SESSION 作用域内都有与之绑定的执行计划，因为优化器在遇到 SESSION 绑定时会忽略 GLOBAL 绑定的执行计划，该语句在 SESSION 作用域内绑定的执行计划会屏蔽掉语句在 GLOBAL 作用域内绑定的执行计划。

例如：

```
-- 创建一个 global binding，指定其使用 sort merge join
create global binding for
    select * from t1, t2 where t1.id = t2.id
using
    select /*+ merge_join(t1, t2) */ * from t1, t2 where t1.id = t2.id;

-- 从该 SQL 的执行计划中可以看到其使用了 global binding 中指定的 sort merge join
explain select * from t1, t2 where t1.id = t2.id;
```

```
-- 创建另一个 session binding，指定其使用 hash join
create binding for
    select * from t1, t2 where t1.id = t2.id
using
    select /*+ hash_join(t1, t2) */ * from t1, t2 where t1.id = t2.id;

-- 从该 SQL 的执行计划中可以看到其使用了 session binding 中指定的 hash join，而不是 global
    ↪ binding 中指定的 sort merge join
explain select * from t1, t2 where t1.id = t2.id;
```

第一个 select 语句在执行时优化器会通过 GLOBAL 作用域内的绑定为其加上 sm_join(t1, t2) hint，explain 出的执行计划中最上层的节点为 MergeJoin。而第二个 select 语句在执行时优化器则会忽视 GLOBAL 作用域内的绑定而使用 SESSION 作用域内的绑定为该语句加上 hash_join(t1, t2) hint，explain 出的执行计划中最上层的节点为 HashJoin。

每个标准化的 SQL 只能同时有一个通过 CREATE BINDING 创建的绑定。对相同的标准化 SQL 创建多个绑定时，会保留最后一个创建的绑定，之前的所有绑定（创建的和演进出来的）都会被删除。但 session 绑定和 global 绑定仍然允许共存，不受这个逻辑影响。

另外，创建绑定时，TiDB 要求 session 处于某个数据库上下文中，也就是执行过 use ${database} 或者客户端连接时指定了数据库。

需要注意的是原始 SQL 和绑定 SQL 在参数化以及去掉 Hint 后文本必须相同，否则创建会失败，例如：

```
CREATE BINDING FOR SELECT * FROM t WHERE a > 1 USING SELECT * FROM t use index(idx) WHERE a > 2;
```

可以创建成功，因为原始 SQL 和绑定 SQL 在参数化以及去掉 Hint 后文本都是 select * from test . t where ↪ a > ?，而

```
CREATE BINDING FOR SELECT * FROM t WHERE a > 1 USING SELECT * FROM t use index(idx) WHERE b > 2;
```

则不可以创建成功，因为原始 SQL 在经过处理后是 select * from test . t where a > ?，而绑定 SQL 在经过处理后是 select * from test . t where b > ?。

> 注意：
>
> 对于 PREPARE/EXECUTE 语句组，或者用二进制协议执行的查询，创建执行计划绑定的对象应当是查询语句本身，而不是 PREPARE/EXECUTE 语句。

删除绑定

```
DROP [GLOBAL | SESSION] BINDING FOR BindableStmt;
```

该语句可以在 GLOBAL 或者 SESSION 作用域内删除指定的执行计划绑定，在不指定作用域时默认作用域为 SESSION。

一般来说，SESSION 作用域的绑定主要用于测试或在某些特殊情况下使用。若需要集群中所有的 TiDB 进程都生效，则需要使用 GLOBAL 作用域的绑定。SESSION 作用域对 GLOBAL 作用域绑定的屏蔽效果会持续到该 SESSION 结束。

承接上面关于 SESSION 绑定屏蔽 GLOBAL 绑定的例子，继续执行：

```
-- 删除 session 中创建的 binding
drop session binding for select * from t1, t2 where t1.id = t2.id;

-- 重新查看该 SQL 的执行计划
explain select * from t1,t2 where t1.id = t2.id;
```

在这里 SESSION 作用域内被删除掉的绑定会屏蔽 GLOBAL 作用域内相应的绑定，优化器不会为 `select` 语句添加 `sm_join(t1, t2)` hint，`explain` 给出的执行计划中最上层节点并不被 hint 固定为 MergeJoin，而是由优化器经过代价估算后自主进行选择。

> **注意：**
>
> 执行 DROP GLOBAL BINDING 会删除当前 tidb-server 实例缓存中的绑定，并将系统表中对应行的状态修改为 'deleted'。该语句不会直接删除系统表中的记录，因为其他 tidb-server 实例需要读取系统表中的 'deleted' 状态来删除其缓存中对应的绑定。对于这些系统表中状态为 'deleted' 的记录，后台线程每隔 100 个 `bind-info-lease`（默认值为 `3s`，合计 `300s`）会触发一次对 `update_time` 在 10 个 `bind-info-lease` 以前的绑定（确保所有 tidb-server 实例已经读取过这个 'deleted' 状态并更新完缓存）的回收清除操作。

查看绑定

```
SHOW [GLOBAL | SESSION] BINDINGS [ShowLikeOrWhere];
```

该语句会按照绑定更新时间由新到旧的顺序输出 GLOBAL 或者 SESSION 作用域内的执行计划绑定，在不指定作用域时默认作用域为 SESSION。目前 SHOW BINDINGS 会输出 8 列，具体如下：

列名	说明
original_sql	参数化后的原始 SQL
bind_sql	带 Hint 的绑定 SQL
default_db	默认数据库名
status	状态，包括 using（正在使用）、deleted（已删除）、invalid（无效）、rejected（演进时被拒绝）和 pending verify（等待演进验证）

列名	说明
create_time	创建时间
update_time	更新时间
charset	字符集
collation	排序规则
source	创建方式，包括 manual（由 `create [global] binding` 生成）、capture（由 tidb 自动创建生成）和 evolve（由 tidb 自动演进生成）

排查绑定

```
SELECT @@[SESSION.]last_plan_from_binding;
```

该语句使用系统变量`last_plan_from_binding`显示上一条执行的语句所使用的执行计划是否来自 binding 的执行计划。

另外，当使用 `explain format = 'verbose'` 语句查看一条 SQL 语句的查询计划时，如果该 SQL 语句使用了 binding，`explain` 语句会输出 warning 警告。此时可以通过查看 warning 了解该 SQL 语句使用了哪一条 binding。

```
-- 创建一个 global binding

create global binding for
    select * from t
using
    select * from t;

-- 使用 explain format = 'verbose' 语句查看 SQL 的执行计划，通过查看 warning 信息确认查询所使用的 binding

explain format = 'verbose' select * from t;
show warnings;
```

10.3.4.3.2 自动捕获绑定 (Baseline Capturing)

通过将 `tidb_capture_plan_baselines` 的值设置为 on（其默认值为 `off`）可以打开自动捕获绑定功能。

> 注意：

> 自动绑定功能依赖于Statement Summary，因此在使用自动绑定之前需打开 Statement Summary 开关。

开启自动绑定功能后，每隔 `bind-info-lease`（默认值为 `3s`）会遍历一次 Statement Summary 中的历史 SQL 语句，并为至少出现两次的 SQL 语句自动捕获绑定。绑定的执行计划为 Statement Summary 中记录执行这条语句时使用的执行计划。

对于以下几种 SQL 语句，TiDB 不会自动捕获绑定：

- EXPLAIN 和 EXPLAIN ANALYZE 语句；
- TiDB 内部执行的 SQL 语句，比如统计信息自动加载使用的 SELECT 查询；
- 存在手动创建的执行计划绑定的 SQL 语句；

对于 PREPARE/EXECUTE 语句组，或通过二进制协议执行的查询，TiDB 会为真正的查询（而不是 PREPARE/EXECUTE 语句）自动捕获绑定。

> 注意：
>
> 由于 TiDB 存在一些内嵌 SQL 保证一些功能的正确性，所以自动捕获绑定时会默认屏蔽内嵌 SQL。

10.3.4.3.3 自动演进绑定 (Baseline Evolution)

自动演进绑定，在 TiDB 4.0 版本引入，是执行计划管理的重要功能之一。

由于某些数据变更后，原先绑定的执行计划可能是一个不优的计划。为了解决该问题，引入自动演进绑定功能来自动优化已经绑定的执行计划。

另外自动演进绑定还可以一定程度上避免统计信息改动后，对执行计划带来的抖动。

使用方式

通过以下语句可以开启自动演进绑定功能：

```
set global tidb_evolve_plan_baselines = on;
```

`tidb_evolve_plan_baselines` 的默认值为 `off`。

> 警告：
>
> 自动演进功能目前为实验特性，存在未知风险，不建议在生产环境中使用。
>
> 此变量开关已强制关闭，直到自动演进成为正式功能 GA (Generally Available)。如果你尝试打开开关，会产生报错。如果你已经在生产环境中使用了此功能，请尽快将它禁用。如发现 binding 状态不如预期，请与 PingCAP 的技术支持联系获取相关支持。

在打开自动演进功能后，如果优化器选出的最优执行计划不在之前绑定的执行计划之中，会将其记录为待验证的执行计划。每隔 bind-info-lease（默认值为 3s），会选出一个待验证的执行计划，将其和已经绑定的执行计划中代价最小的比较实际运行时间。如果待验证的运行时间更优的话（目前判断标准是运行时间小于等于已绑定执行计划运行时间的 2/3），会将其标记为可使用的绑定。以下示例描述上述过程。

假如有表 t 定义如下：

```
create table t(a int, b int, key(a), key(b));
```

在表 t 上进行如下查询：

```
select * from t where a < 100 and b < 100;
```

表上满足条件 a < 100 的行很少。但由于某些原因，优化器没能选中使用索引 a 这个最优执行计划，而是误选了速度慢的全表扫，那么用户首先可以通过如下语句创建一个绑定：

```
create global binding for select * from t where a < 100 and b < 100 using select * from t use
    ↪ index(a) where a < 100 and b < 100;
```

当以上查询语句再次执行时，优化器会在刚创建绑定的干预下选择使用索引 a，进而降低查询时间。

假如随着在表中进行插入和修改，表中满足条件 a < 100 的行变得越来越多，而满足条件 b < 100 的行变得越来越少，这时再在绑定的干预下使用索引 a 可能就不是最优了。

绑定的演进可以解决这类问题。当优化器感知到表数据变化后，会对这条查询生成使用索引 b 的执行计划。但由于绑定的存在，这个执行计划不会被采纳和执行，不过它会被存在后台的演进列表里。在演进过程中，如果它被验证为执行时间明显低于使用索引 a 的执行时间（即当前绑定的执行计划），那么索引 b 会被加入到可用的绑定列表中。在此之后，当这条查询再次被执行时，优化器首先生成使用索引 b 的执行计划，并确认它在绑定列表中，所以会采纳它并执行，进而可以在数据变化后降低这条查询的执行时间。

为了减少自动演进对集群的影响，可以通过设置 tidb_evolve_plan_task_max_time 来限制每个执行计划运行的最长时间，其默认值为 600s。实际在验证执行计划时，计划的最长运行时间还会被限制为不超过已验证执行计划的运行时间的两倍；通过 tidb_evolve_plan_task_start_time 和 tidb_evolve_plan_task_end_time 可以限制运行演进任务的时间窗口，默认值分别为 00:00 +0000 和 23:59 +0000。

注意事项

由于自动演进绑定会自动地创建新的绑定，当查询的环境发生变动时，自动创建的绑定可能会有多种行为的选择。这里列出一些注意事项：

- 自动演进只会对存在至少一个 global 绑定的标准化 SQL 进行演进。
- 由于创建新的绑定会删除之前所有绑定（对于一条标准化 SQL），自动演进的绑定也会在手动重新创建绑定后被删除。
- 所有和计算过程相关的 hint，在演进时都会被保留。计算过程相关的 hint 有如下几种：

Hint	说明
memory_quota	查询过程最多可以使用多少内存
use_toja	优化器是否考虑把子查询转化为 join
use_cascades	是否使用 cascades 优化器

Hint	说明
no_index_merge	优化器是否考虑将 index merge 作为一个读表选项
read_consistent_replica	是否强制读表时使用 follower read
max_execution_time	查询过程最多消耗多少时间

- `read_from_storage` 是一个非常特别的 hint，因为它指定了读表时选择从 TiKV 读还是从 TiFlash 读。由于 TiDB 提供隔离读的功能，当隔离条件变化时，这个 hint 对演进出来的执行计划影响很大，所以当最初创建的绑定中存在这个 hint，TiDB 会无视其所有演进的绑定。

10.3.4.3.4 升级检查 (Upgrade Checklist)

执行计划管理功能 (SPM) 在版本升级过程中可能会出现一些兼容性问题导致升级失败，你需要在版本升级前做一些检查，确保版本顺利升级。

- 当你尝试从 v5.2 以前的版本（即 v4.0、v5.0、v5.1）升级到当前版本，需要注意在升级前检查自动演进的开关 `tidb_evolve_plan_baselines` 是否已经关闭。如果尚未关闭，则需要将其关闭后再进行升级。具体操作如下所示：

```
-- 在待升级的版本上检查自动演进的开关 `tidb_evolve_plan_baselines` 是否关闭。

select @@global.tidb_evolve_plan_baselines;

-- 如果演进的开关 `tidb_evolve_plan_baselines` 尚未关闭，则需要将其关闭。

set global tidb_evolve_plan_baselines = off;
```

- 当你尝试从 v4.0 版本升级到当前版本，需要注意在升级前检查所有可用绑定对应的查询语句在新版本中是否存在语法错误。如果存在语法错误，则需要删除对应的绑定。

  具体操作如下所示：

```
-- 在待升级的版本上检查现有可用绑定对应的查询语句。

select bind_sql from mysql.bind_info where status = 'using';

-- 将上一条查询得到的结果，在新版本的测试环境中进行验证。

bind_sql_0;
bind_sql_1;
...

-- 如果报错信息是语法错误（ERROR 1064 (42000): You have an error in your SQL syntax），
    ↪ 则需要删除对应的绑定。
-- 如果是其他错误，如未找到表，则表示语法兼容，不需要进行额外的处理。
```

10.3.4.4 优化规则与表达式下推的黑名单

本文主要介绍优化规则的黑名单与表达式下推的黑名单。

10.3.4.4.1 优化规则黑名单

优化规则黑名单是针对优化规则的调优手段之一，主要用于手动禁用一些优化规则。

重要的优化规则

优化规则	规则名称	简介
列裁剪	column_prune	对于上层算子不需要的列，不在下层算子输出该列，减少计算
子查询去关联	decorrelate	尝试对相关子查询进行改写，将其转换为普通 join 或 aggregation 计算
聚合消除	aggregation_eliminate	尝试消除执行计划中的某些不必要的聚合算子
投影消除	projection_eliminate	消除执行计划中不必要的投影算子
最大最小消除	max_min_eliminate	改写聚合中的 max/min 计算，转化为 `order by + limit 1`
谓词下推	predicate_push_down	尝试将执行计划中过滤条件下推到离数据源更近的算子上
外连接消除	outer_join_eliminate	尝试消除执行计划中不必要的 left join 或者 right join
分区裁剪	partition_processor	将分区表查询改成为用 union all，并裁剪掉不满足过滤条件的分区
聚合下推	aggregation_push_down	尝试将执行计划中的聚合算子下推到更底层的计算节点
TopN 下推	topn_push_down	尝试将执行计划中的 TopN 算子下推到离数据源更近的算子上
Join 重排序	join_reorder	对多表 join 确定连接顺序

禁用优化规则

当某些优化规则在一些特殊查询中的优化结果不理想时，可以使用优化规则黑名单禁用一些优化规则。

使用方法

> **注意：**
>
> 以下操作都需要数据库的 super privilege 权限。每个优化规则都有各自的名字，比如列裁剪的名字是 “column_prune”。所有优化规则的名字都可以在重要的优化规则表格中第二列查到。

- 如果你想禁用某些规则，可以在 `mysql.opt_rule_blacklist` 表中写入规则的名字，例如：

```
INSERT INTO mysql.opt_rule_blacklist VALUES("join_reorder"), ("topn_push_down");
```

执行以下 SQL 语句可让禁用规则立即生效，包括相应 TiDB Server 的所有旧链接：

```
ADMIN reload opt_rule_blacklist;
```

> **注意：**
>
> `admin reload opt_rule_blacklist` 只对执行该 SQL 语句的 TiDB server 生效。若需要集群中所有 TiDB server 生效，需要在每台 TiDB server 上执行该 SQL 语句。

- 需要解除一条规则的禁用时，需要删除表中禁用该条规则的相应数据，再执行 `admin reload`：

```
DELETE FROM mysql.opt_rule_blacklist WHERE name IN ("join_reorder", "topn_push_down");
admin reload opt_rule_blacklist;
```

10.3.4.4.2 表达式下推黑名单

表达式下推黑名单是针对表达式下推的调优手段之一，主要用于对于某些存储类型手动禁用一些表达式。

已支持下推的表达式

表达式分类	具体操作
逻辑运算	AND (&&), OR (||), NOT (!)
比较运算	<, <=, =, != (<>), >, >=, <=>, IN(), IS NULL, LIKE, IS TRUE, IS FALSE, COALESCE()
数值运算	+, -, *, /, ABS(), CEIL(), CEILING(), FLOOR()
控制流运算	CASE, IF(), IFNULL()
JSON 运算	JSON_TYPE(json_val), JSON_EXTRACT(json_doc, path[, path] ⋯), JSON_UNQUOTE(json_val), JSON_OBJECT(key, val[, key, val] ⋯), JSON_ARRAY([val[, val] ⋯]), JSON_MERGE(json_doc, json_doc[, json_doc] ⋯), JSON_SET(json_doc, path, val[, path, val] ⋯), JSON_INSERT(json_doc, path, val[, path, val] ⋯), JSON_REPLACE(json_doc, path, val[, path, val] ⋯), JSON_REMOVE(json_doc, path[, path] ⋯)
日期运算	DATE_FORMAT()

禁止特定表达式下推

当函数的计算过程由于下推而出现异常时，可通过黑名单功能禁止其下推来快速恢复业务。具体而言，你可以将上述支持的函数或运算符名加入黑名单 `mysql.expr_pushdown_blacklist` 中，以禁止特定表达式下推。

`mysql.expr_pushdown_blacklist` 的 schema 如下：

```
DESC mysql.expr_pushdown_blacklist;
```

```
+------------+--------------+------+------+-------------------+-------+
| Field      | Type         | Null | Key  | Default           | Extra |
+------------+--------------+------+------+-------------------+-------+
| name       | char(100)    | NO   |      | NULL              |       |
| store_type | char(100)    | NO   |      | tikv,tiflash,tidb |       |
| reason     | varchar(200) | YES  |      | NULL              |       |
+------------+--------------+------+------+-------------------+-------+
3 rows in set (0.00 sec)
```

以上结果字段解释如下：

- `name`：禁止下推的函数名。

- `store_type`：用于指明希望禁止该函数下推到哪些组件进行计算。组件可选 `tidb`、`tikv` 和 `tiflash`。`store_type` 不区分大小写，如果需要禁止向多个存储引擎下推，各个存储之间需用逗号隔开。
  - `store_type` 为 `tidb` 时表示在读取 TiDB 内存表时，是否允许该函数在其他 TiDB Server 上执行。
  - `store_type` 为 `tikv` 时表示是否允许该函数在 TiKV Server 的 Coprocessor 模块中执行。
  - `store_type` 为 `tiflash` 时表示是否允许该函数在 TiFlash Server 的 Coprocessor 模块中执行。
- `reason`：用于记录该函数被加入黑名单的原因。

使用方法

加入黑名单

如果要将一个或多个函数或运算符加入黑名单，执行以下步骤：

1. 向 `mysql.expr_pushdown_blacklist` 插入对应的函数名或运算符名以及希望禁止下推的存储引擎集合。
2. 执行 `admin reload expr_pushdown_blacklist;`。

移出黑名单

如果要将一个或多个函数及运算符移出黑名单，执行以下步骤：

1. 从 `mysql.expr_pushdown_blacklist` 表中删除对应的函数名或运算符名。
2. 执行 `admin reload expr_pushdown_blacklist;`。

> **注意：**
>
> `admin reload expr_pushdown_blacklist` 只对执行该 SQL 语句的 TiDB server 生效。若需要集群中所有 TiDB server 生效，需要在每台 TiDB server 上执行该 SQL 语句。

表达式黑名单用法示例

以下示例首先将运算符 < 及 > 加入黑名单，然后将运算符 > 从黑名单中移出。

黑名单是否生效可以从 `explain` 结果中进行观察（参见EXPLAIN 简介）。

1. 对于以下 SQL 语句，`where` 条件中的 `a < 2` 和 `a > 2` 可以下推到 TiKV 进行计算。

```
EXPLAIN SELECT * FROM t WHERE a < 2 AND a > 2;
```

```
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| id                    | estRows  | task      | access object | operator info
    ↪                     |
```

```
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| TableReader_7           | 0.00     | root      |              | data:Selection_6
    ↪                    |
| └─Selection_6           | 0.00     | cop[tikv] |              | gt(ssb_1.t.a, 2), lt(
    ↪ ssb_1.t.a, 2) |
|   └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t      | keep order:false, stats:
    ↪ pseudo      |
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
3 rows in set (0.00 sec)
```

2. 往 `mysql.expr_pushdown_blacklist` 表中插入禁用表达式，并且执行 `admin reload expr_pushdown_blacklist` ↪ 。

```
INSERT INTO mysql.expr_pushdown_blacklist VALUES('<','tikv',''), ('>','tikv','');
```

```
Query OK, 2 rows affected (0.01 sec)
Records: 2  Duplicates: 0  Warnings: 0
```

```
ADMIN reload expr_pushdown_blacklist;
```

```
Query OK, 0 rows affected (0.00 sec)
```

3. 重新观察执行计划，发现表达式下推黑名单生效，`where` 条件中的 `<` 和 `>` 没有被下推到 TiKV Coprocessor 上。

```
EXPLAIN SELECT * FROM t WHERE a < 2 and a > 2;
```

```
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| id                      | estRows  | task      | access object | operator info
    ↪                      |
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
| Selection_7             | 10000.00 | root      |              | gt(ssb_1.t.a, 2), lt(
    ↪ ssb_1.t.a, 2) |
| └─TableReader_6         | 10000.00 | root      |              | data:TableFullScan_5
    ↪                |
|   └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t      | keep order:false, stats:
    ↪ pseudo      |
```

```
+--
    ↪ ----------------------+----------+-----------+--------------+--------------------------------+
    ↪
3 rows in set (0.00 sec)
```

4. 将某一表达式 (> 大于) 禁用规则从黑名单表中删除，并且执行 admin reload expr_pushdown_blacklist ↪ 。

```
DELETE FROM mysql.expr_pushdown_blacklist WHERE name = '>';
```

```
Query OK, 1 row affected (0.01 sec)
```

```
ADMIN reload expr_pushdown_blacklist;
```

```
Query OK, 0 rows affected (0.00 sec)
```

5. 重新观察执行计划，可以看到只有 > 表达式被重新下推到 TiKV Coprocessor，< 表达式仍然被禁用下推。

```
EXPLAIN SELECT * FROM t WHERE a < 2 AND a > 2;
```

```
+--
    ↪ ------------------------+----------+-----------+---------------+--------------------------------+
    ↪
| id                        | estRows  | task      | access object | operator info
    ↪                  |
+--
    ↪ ------------------------+----------+-----------+---------------+--------------------------------+
    ↪
| Selection_8               | 0.00     | root      |               | lt(ssb_1.t.a, 2)
    ↪               |
| └─TableReader_7           | 0.00     | root      |               | data:Selection_6
    ↪               |
|   └─Selection_6           | 0.00     | cop[tikv] |               | gt(ssb_1.t.a, 2)
    ↪               |
|     └─TableFullScan_5     | 10000.00 | cop[tikv] | table:t       | keep order:false,
    ↪ stats:pseudo |
+--
    ↪ ------------------------+----------+-----------+---------------+--------------------------------+
    ↪
4 rows in set (0.00 sec)
```

11 最佳实践

11.1 TiDB 最佳实践

本文档总结使用 TiDB 时的一些最佳实践，主要涉及 SQL 使用和 OLAP/OLTP 优化技巧，特别是一些 TiDB 专有的优化开关。

建议先阅读讲解 TiDB 原理的三篇文章（讲存储，说计算，谈调度），再来看这篇文章。

11.1.1 前言

数据库是一个通用的基础组件，在开发过程中会考虑到多种目标场景，在具体的业务场景中，需要根据业务的实际情况对数据的参数或者使用方式进行调整。

TiDB 是一个兼容 MySQL 协议和语法的分布式数据库，但是由于其内部实现，特别是支持分布式存储以及分布式事务，使得一些使用方法和 MySQL 有所区别。

11.1.2 基本概念

TiDB 的最佳实践与其实现原理密切相关，建议读者先了解一些基本的实现机制，包括 Raft、分布式事务、数据分片、负载均衡、SQL 到 KV 的映射方案、二级索引的实现方法、分布式执行引擎。下面会做一点简单的介绍，更详细的信息可以参考 PingCAP 公众号以及知乎专栏的一些文章。

11.1.2.1 Raft

Raft 是一种一致性协议，能提供强一致的数据复制保证，TiDB 最底层用 Raft 来同步数据。每次写入都要写入多数副本，才能对外返回成功，这样即使丢掉少数副本，也能保证系统中还有最新的数据。比如最大 3 副本的话，每次写入 2 副本才算成功，任何时候，只丢失一个副本的情况下，存活的两个副本中至少有一个具有最新的数据。

相比 Master-Slave 方式的同步，同样是保存三副本，Raft 的方式更为高效，写入的延迟取决于最快的两个副本，而不是最慢的那个副本。所以使用 Raft 同步的情况下，异地多活成为可能。在典型的两地三中心场景下，每次写入只需要本数据中心以及离得近的一个数据中心写入成功就能保证数据的一致性，而并不需要三个数据中心都写成功。但是这并不意味着在任何场景都能构建跨机房部署的业务，当写入量比较大时候，机房之间的带宽和延迟成为关键因素，如果写入速度超过机房之间的带宽，或者是机房之间延迟过大，整个 Raft 同步机制依然无法很好的运转。

11.1.2.2 分布式事务

TiDB 提供完整的分布式事务，事务模型是在 Google Percolator 的基础上做了一些优化。具体的实现可以参考《Percolator 和 TiDB 事务算法》这篇文章。本文档只讨论以下几点：

- 乐观锁

  TiDB 的乐观事务模型，只有在真正提交的时候，才会做冲突检测。如果有冲突，则需要重试。这种模型在冲突严重的场景下，会比较低效，因为重试之前的操作都是无效的，需要重复做。举一个比较极端的例子，就是把数据库当做计数器用，如果访问的并发度比较高，那么一定会有严重的冲突，导致大

量的重试甚至是超时。但是如果访问冲突并不十分严重，那么乐观锁模型具备较高的效率。在冲突严重的场景下，推荐使用悲观锁，或在系统架构层面解决问题，比如将计数器放在 Redis 中。

- 悲观锁

  TiDB 的悲观事务模式，悲观事务的行为和 MySQL 基本一致，在执行阶段就会上锁，先到先得，避免冲突情况下的重试，可以保证有较多冲突的事务的成功率。悲观锁同时解决了希望通过 `select for update` 对数据提前锁定的场景。但如果业务场景本身冲突较少，乐观锁的性能会更有优势。

- 事务大小限制

  由于分布式事务要做两阶段提交，并且底层还需要做 Raft 复制，如果一个事务非常大，会使得提交过程非常慢，并且会卡住下面的 Raft 复制流程。为了避免系统出现被卡住的情况，我们对事务的大小做了限制：

  - 单个事务包含的 SQL 语句不超过 5000 条（默认）
    * 单条 KV entry 不超过 6MB（默认）
    * KV entry 的总大小不超过 10G

  在 Google 的 Cloud Spanner 上面，也有类似的限制。

11.1.2.3 数据分片

TiKV 自动将底层数据按照 Key 的 Range 进行分片。每个 Region 是一个 Key 的范围，从 `StartKey` 到 EndKey 的左闭右开区间。Region 中的 Key-Value 总量超过一定值，就会自动分裂。这部分用户不需要担心。

11.1.2.4 负载均衡

PD 会根据整个 TiKV 集群的状态，对集群的负载进行调度。调度是以 Region 为单位，以 PD 配置的策略为调度逻辑，自动完成。

11.1.2.5 SQL on KV

TiDB 自动将 SQL 结构映射为 KV 结构。具体的可以参考《三篇文章了解 TiDB 技术内幕 - 说计算》这篇文档。简单来说，TiDB 执行了以下操作：

- 一行数据映射为一个 KV，Key 以 `TableID` 构造前缀，以行 ID 为后缀
- 一条索引映射为一个 KV，Key 以 `TableID+IndexID` 构造前缀，以索引值构造后缀

可以看到，对于一个表中的数据或者索引，会具有相同的前缀，这样在 TiKV 的 Key 空间内，这些 Key-Value 会在相邻的位置。那么当写入量很大，并且集中在一个表上面时，就会造成写入的热点，特别是连续写入的数据中某些索引值也是连续的（比如 update time 这种按时间递增的字段），会在很少的几个 Region 上形成写入热点，成为整个系统的瓶颈。同样，如果所有的数据读取操作也都集中在很小的一个范围内（比如在连续的几万或者十几万行数据上），那么可能造成数据的访问热点。

11.1.2.6 二级索引

TiDB 支持完整的二级索引，并且是全局索引，很多查询可以通过索引来优化。如果利用好二级索引，对业务非常重要，很多 MySQL 上的经验在 TiDB 这里依然适用，不过 TiDB 还有一些自己的特点，需要注意，这一节主要讨论在 TiDB 上使用二级索引的一些注意事项。

- 二级索引是否越多越好

  二级索引能加速查询，但是要注意新增一个索引是有副作用的。上一节介绍了索引的存储模型，那么每增加一个索引，在插入一条数据的时候，就要新增一个 Key-Value，所以索引越多，写入越慢，并且空间占用越大。另外过多的索引也会影响优化器运行时间，并且不合适的索引会误导优化器。所以索引并不是越多越好。

- 对哪些列建索引比较合适

  上文提到，索引很重要但不是越多越好，因此需要根据具体的业务特点创建合适的索引。原则上需要对查询中需要用到的列创建索引，目的是提高性能。下面几种情况适合创建索引：

  - 区分度比较大的列，通过索引能显著地减少过滤后的行数
  - 有多个查询条件时，可以选择组合索引，注意需要把等值条件的列放在组合索引的前面

  这里举一个例子，假设常用的查询是 `select * from t where c1 = 10 and c2 = 100 and c3 > 10`，那么可以考虑建立组合索引 `Index cidx (c1, c2, c3)`，这样可以用查询条件构造出一个索引前缀进行 Scan。

- 通过索引查询和直接扫描 Table 的区别

  TiDB 实现了全局索引，所以索引和 Table 中的数据并不一定在一个数据分片上。通过索引查询的时候，需要先扫描索引，得到对应的行 ID，然后通过行 ID 去取数据，所以可能会涉及到两次网络请求，会有一定的性能开销。

  如果查询涉及到大量的行，那么扫描索引是并发进行，只要第一批结果已经返回，就可以开始去取 Table 的数据，所以这里是一个并行 + Pipeline 的模式，虽然有两次访问的开销，但是延迟并不会很大。

  以下情况不会涉及到两次访问的问题：

  - 索引中的列已经满足了查询需求。比如 Table `t` 上面的列 `c` 有索引，查询是 `select c from t where c > 10;`，这个时候，只需要访问索引，就可以拿到所需要的全部数据。这种情况称之为覆盖索引 (Covering Index)。所以如果很关注查询性能，可以将部分不需要过滤但是需要在查询结果中返回的列放入索引中，构造成组合索引，比如这个例子：`select c1, c2 from t where c1 > 10;`，要优化这个查询可以创建组合索引 `Index c12 (c1, c2)`。
  - 表的 Primary Key 是整数类型。在这种情况下，TiDB 会将 Primary Key 的值当做行 ID，所以如果查询条件是在 PK 上面，那么可以直接构造出行 ID 的范围，直接扫描 Table 数据，获取结果。

- 查询并发度

  数据分散在很多 Region 上，所以 TiDB 在做查询的时候会并发进行，默认的并发度比较保守，因为过高的并发度会消耗大量的系统资源，且对于 OLTP 类型的查询，往往不会涉及到大量的数据，较低的并发度已经可以满足需求。对于 OLAP 类型的 Query，往往需要较高的并发度。所以 TiDB 支持通过 System Variable 来调整查询并发度。

  - tidb_distsql_scan_concurrency

  在进行扫描数据的时候的并发度，这里包括扫描 Table 以及索引数据。

- tidb_index_lookup_size

如果是需要访问索引获取行 ID 之后再访问 Table 数据，那么每次会把一批行 ID 作为一次请求去访问 Table 数据，这个参数可以设置 Batch 的大小，较大的 Batch 会使得延迟增加，较小的 Batch 可能会造成更多的查询次数。这个参数的合适大小与查询涉及的数据量有关。一般不需要调整。

- tidb_index_lookup_concurrency

如果是需要访问索引获取行 ID 之后再访问 Table 数据，每次通过行 ID 获取数据时候的并发度通过这个参数调节。

- 通过索引保证结果顺序

索引除了可以用来过滤数据之外，还能用来对数据排序，首先按照索引的顺序获取行 ID，然后再按照行 ID 的返回顺序返回行的内容，这样可以保证返回结果按照索引列有序。前面提到了扫索引和获取 Row 之间是并行 + Pipeline 模式，如果要求按照索引的顺序返回 Row，那么这两次查询之间的并发度设置的太高并不会降低延迟，所以默认的并发度比较保守。可以通过 tidb_index_serial_scan_concurrency 变量进行并发度调整。

- 逆序索引

目前 TiDB 支持对索引进行逆序 Scan，目前速度比顺序 Scan 慢一些，通常情况下慢 20%，在数据频繁修改造成版本较多的情况下，会慢的更多。如果可能，建议避免对索引的逆序 Scan。

11.1.3 场景与实践

上一节我们讨论了一些 TiDB 基本的实现机制及其对使用带来的影响，本节我们从具体的使用场景出发，谈一些更为具体的操作实践。我们以从部署到支撑业务这条链路为序，进行讨论。

11.1.3.1 部署

在部署之前请务必阅读TiDB 部署建议以及对硬件的需求。

推荐通过TiUP 部署 TiDB 集群，这个工具可以部署、停止、销毁、升级整个集群，非常方便易用。非常不推荐手动部署，后期的维护和升级会很麻烦。

11.1.3.2 导入数据

为了提高导入数据期间的写入性能，可以对 TiKV 的参数进行调优，具体的文档查看TiKV 性能参数调优。

11.1.3.3 写入

上面提到了 TiDB 对单个事务的大小有限制，这层限制是在 KV 层面，反映在 SQL 层面的话，简单来说一行数据会映射为一个 KV entry，每多一个索引，也会增加一个 KV entry。

> 注意：
>
> 对事务的大小限制，要考虑 TiDB 做编码以及事务额外 Key 的开销，在使用的时候，建议每个事务的行数不超过 200 行，且单行数据小于 100k，否则可能性能不佳。

建议无论是 Insert，Update 还是 Delete 语句，都通过分 Batch 或者是加 Limit 的方式限制。

在删除大量数据的时候，建议使用 `Delete from t where xx limit 5000;` 这样的方案，通过循环来删除，用 `Affected Rows == 0` 作为循环结束条件。

如果一次删除的数据量非常大，这种循环的方式会越来越慢，因为每次删除都是从前向后遍历，前面的删除之后，短时间内会残留不少删除标记（后续会被 GC 清理掉），影响后面的 Delete 语句。如果有可能，建议把 Where 条件细化。举个例子，假设要删除 2017-05-26 当天的所有数据，那么可以这样做：

```
for i from 0 to 23:
    while affected_rows > 0:
        delete from t where insert_time >= i:00:00 and insert_time < (i+1):00:00 limit 5000;
        affected_rows = select affected_rows()
```

上面是一段伪代码，意思就是要把大块的数据拆成小块删除，以避免删除过程中前面的 Delete 语句影响后面的 Delete 语句。

11.1.3.4 查询

看业务的查询需求以及具体的语句，可以参考 TiDB 专用系统变量和语法这篇文档。可以通过 SET 语句控制 SQL 执行的并发度，另外通过 Hint 控制 Join 物理算子选择。

另外 MySQL 标准的索引选择 Hint 语法，也可以用，通过 `Use Index/Ignore Index hint` 控制优化器选择索引。

如果是个 OLTP 和 OLAP 混合类型的业务，可以把 TP 请求和 AP 请求发送到不同的 tidb-server 上，这样能够减小 AP 业务对于 TP 业务的影响。承载 AP 业务的 tidb-server 推荐使用高配的机器，比如 CPU 核数比较多，内存比较大。

但彻底的隔离 OLTP 和 OLAP，推荐将 OLAP 的业务跑在 TiFlash 上。TiFlash 是列存引擎，在 OLAP 的分析查询场景上，性能极具亮点，TiFlash 可以在存储层上做到物理隔离，并可做到一致性读取。

11.1.3.5 监控和日志

Metrics 系统是了解系统状态的最佳方法，建议所有的用户都部署监控系统。

TiDB 使用 Grafana + Prometheus 监控系统状态。如果使用 TiUP 部署集群，那么会自动部署和配置监控系统。

监控系统中的监控项很多，大部分是给 TiDB 开发者查看的内容，如果没有对源代码比较深入的了解，并没有必要了解这些监控项。我们会精简出一些和业务相关或者是系统关键组件状态相关的监控项，放在一个独立的 overview 面板中，供用户使用。

除了监控之外，查看日志也是了解系统状态的常用方法。TiDB 的三个组件 tidb-server/tikv-server/pd-server 都有一个 `--log-file` 的参数。如果启动的时候设置了这个参数，那么日志会保存着参数所设置的文件的位置，另外会自动的按天对 Log 文件做归档。如果没有设置 `--log-file` 参数，日志会输出在 `stderr` 中。

从 4.0 版本开始，从解决易用性的角度出发，提供了TiDB Dashboard UI 系统，通过浏览器访问 `http://PD_IP:PD_PORT/dashboard` 即可打开 TiDB Dashboard。TiDB Dashboard 可以提供集群状态、性能分析、流量可视化、SQL 诊断、日志搜索等功能。

11.1.3.6 文档

了解一个系统或者解决使用中的问题最好的方法是阅读文档，明白实现原理。TiDB 有大量的官方文档，希望大家在遇到问题的时候能先尝试通过文档或者搜索 Issue list 寻找解决方案。官方文档查看 docs-cn。如果希望阅读英文文档，可以查看 docs。

其中的 FAQ 和故障诊断章节建议大家仔细阅读。另外 TiDB 还有一些不错的工具，也有配套的文档，具体的见各项工具的 GitHub 页面。

除了文档之外，还有很多不错的文章介绍 TiDB 的各项技术细节内幕，大家可以关注下面这些文章发布渠道：

- 公众号：微信搜索 PingCAP
- 知乎专栏：TiDB 的后花园
- 官方博客

11.1.4 TiDB 的最佳适用场景

简单来说，TiDB 适合具备下面这些特点的场景：

- 数据量大，单机保存不下
- 不希望做 Sharding 或者懒得做 Sharding
- 访问模式上没有明显的热点
- 需要事务、需要强一致、需要灾备
- 希望 Real-Time HTAP，减少存储链路

11.2 三节点混合部署的最佳实践

在对性能要求不高且需要控制成本的场景下，将 TiDB、TiKV、PD 混合部署在三台机器上是一个可行的方案。

本文以 TPC-C 作为工作负载，提供一些在三节点混合部署场景下部署和参数调整的建议。

11.2.1 部署环境和测试方法

三台物理机，每台机器有 16 个 CPU 核心且内存为 32 GB。每节点（台机器）上混合部署 1 TiDB（实例）+ 1 TiKV（实例）+ 1 PD（实例）。

由于 PD 和 TiKV 都会存储信息到磁盘，磁盘的写入读取延迟会直接影响到 PD 和 TiKV 的服务延迟，为了防止 PD 和 TiKV 对磁盘资源的争抢导致相互影响，推荐 PD 和 TiKV 采用不同的磁盘。

在 TiUP bench 中使用 TPC-C 5000 Warehouse 数据，每次以 `terminals` 参数为 128 的并发度测试 12 小时。主要观察集群性能的稳定性指标。

下图是默认参数配置下，12 小时内集群的 QPS 监控，可以看倒有比较明显的抖动。



图 215: QPS with default config

调整参数后，稳定性得到了改善。



图 216: QPS with modified config

11.2.2 参数调整

上图中出现抖动的主要原因是，默认的线程池和后台任务资源分配针对的是资源比较充足的机器。在混合部署的场景下，因为可用资源需要被多个组件共享，所以需要通过配置参数进行限制。

本次测试最终采用的配置如下：

```
tikv:
    readpool.unified.max-thread-count: 6
    server.grpc-concurrency: 2
    storage.scheduler-worker-pool-size: 2
    gc.max-write-bytes-per-sec: 300K
    rocksdb.max-background-jobs: 3
    rocksdb.max-sub-compactions: 1
    rocksdb.rate-bytes-per-sec: “200M”

  tidb:
    performance.committer-concurrency: 4
    performance.max-procs: 8
```

接下来会分别介绍这几个参数的意义和调整方法。

11.2.2.1 TiKV 线程池大小配置

本节的配置项会调整一些与前台业务有关的线程池的资源配比。缩减这些线程池会损失一些性能，但在混合部署场景下可用资源有限，本身也难以达到很高的性能，所以选择牺牲一小部分性能去换取整体的稳定性。

可以先以默认配置为基础运行一次实际负载的测试，观察各线程池的实际使用量，再修改这些配置项，缩减利用率不高的线程池大小。

11.2.2.1.1 readpool.unified.max-thread-count

该参数默认取值为机器线程数的 80%，因为是混合部署场景，你需要手动计算并指定该值。可以先设置成期望 TiKV 使用的 CPU 线程数的 80%。

11.2.2.1.2 server.grpc-concurrency

该参数默认为 4。因为现有部署方案的 CPU 资源有限，实际请求数也不会很多。可以把该参数值调低，后续观察监控面板保持其使用率在 80% 以下即可。

本次测试最后选择设置该参数值为 2，通过 gRPC poll CPU 面板观察，利用率正好在 80% 左右。

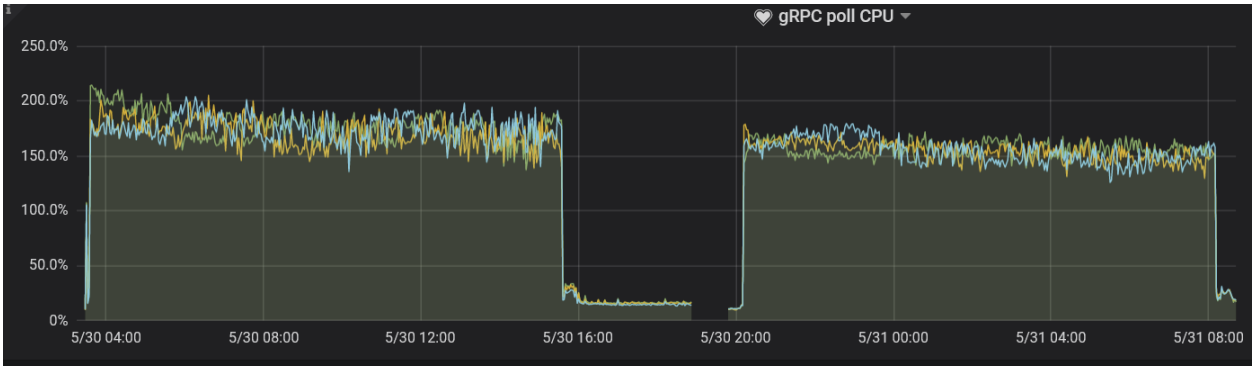


图 217: gRPC Pool CPU

11.2.2.1.3 storage.scheduler-worker-pool-size

在 TiKV 检测到机器 CPU 数大于或等于 16 时，该参数值默认为 8。CPU 数小于 16 时，该参数值默认为 4。它主要用于将复杂的事务请求转化为简单的 key-value 读写。但是 scheduler 线程池本身不进行任何写操作。

一般来说该线程池的利用率保持在 50% - 75% 之间是比较好的。和 gRPC 线程池情况类似，混合部署时该参数默认取值偏大，资源利用不充分。本次测试最后选择取值为 2，通过 Scheduler worker CPU 面板观察，利用率比较符合最佳实践。



图 218: Scheduler Worker CPU

11.2.2.2 TiKV 后台任务资源配置

除前台任务之外，TiKV 上还会持续地有后台任务进行数据整理和过期数据的清除。默认配置为这些后台任务分配了比较多的资源以应对大流量的写入。

混合部署场景下，默认配置就不是很合适了，需要通过以下参数对这些后台任务地资源使用量进行限制。

11.2.2.2.1 `rocksdb.max-background-jobs` 和 `rocksdb.max-sub-compactions`

RocksDB 线程池是进行 Compact 和 Flush 任务的线程池，默认大小为 8。这明显超出了实际可以使用的资源，需要限制。`rocksdb.max-sub-compactions` 是单个 compaction 任务的子任务并发数，默认值为 3，在写入流量不大的情况下可以进行限制。

这次测试最终将 `rocksdb.max-background-jobs` 设置为 3，将 `rocksdb.max-sub-compactions` 设置为 1。在 12 小时的 TPC-C 负载下没有发生 Write Stall，根据实际负载进行这两项参数的优化时，可以逐步调低这两个配置，并通过监控观察。

- 如果遇到了 Write Stall，可以先调大 `rocksdb.max-background-jobs` 的取值。
- 如果还是存在问题，可将 `rocksdb.max-sub-compactions` 设置为 2 或者 3。

11.2.2.2.2 `rocksdb.rate-bytes-per-sec`

该参数用于限制后台 compaction 任务的磁盘流量。默认配置下没有进行任何限制。为了避免 compaction 挤占前台服务资源，可以根据硬盘的顺序读写速度进行调整，为正常的服务保留足够多的磁盘带宽。

类似 compaction 线程池的调整方法，调整该参数值后，也根据是否存在 Write Stall 来判断取值是否合理。

11.2.2.2.3 `gc.max_write_bytes_per_sec`

因为 TiDB 使用的是 MVCC 的模型，TiKV 还需要周期性地在后台清除旧版本的数据。当可用资源有限的时候，这个操作会引起周期性的性能抖动。可以用 `gc.max_write_bytes_per_sec` 来限制这一操作的资源使用。



图 219: GC Impact

除了在配置文件中设置该参数之外，还可以通过 tikv-ctl 动态调节，为调整该参数提供便利：

```
tiup ctl tikv --host=${ip:port} modify-tikv-config -n gc.max_write_bytes_per_sec -v ${limit}
```

> 注意：
>
> 对于更新频繁的业务场景，限制 GC 流量可能会导致 MVCC 版本堆积，进而影响读取性能。所以这个参数的取值需要进行多次尝试，在性能抖动和性能衰退之间找一个比较平衡的取值。

11.2.2.3 TiDB 参数调整

一般通过系统变量调整 TiDB 执行算子的优化参数，例如 `tidb_hash_join_concurrency`、`tidb_index_lookup_join_concurrency` 等。

本测试中没有对这些参数进行调整。如果实际的业务负载测试中，出现执行算子消耗过多 CPU 资源的情况，可以针对业务场景对特定的算子资源使用量进行限制。这部分内容可以参考 TiDB 系统变量文档。

11.2.2.3.1 `performance.max-procs`

该参数控制着整个 Go 进程能使用的 CPU 核心数量，默认情况下为当前机器或者 cgroups 的 CPU 数量。

Go 运行时会定期使用一定比例的线程进行 GC 等后台工作。在混合部署模式下，如果不对这一参数进行限制，GC 等后台操作会过多占用 CPU 资源。

11.3 TiDB 高并发写入场景最佳实践

在 TiDB 的使用过程中，一个典型场景是高并发批量写入数据到 TiDB。本文阐述了该场景中的常见问题，旨在给出一个业务的最佳实践，帮助读者避免因使用 TiDB 不当而影响业务开发。

11.3.1 目标读者

本文假设你已对 TiDB 有一定的了解，推荐先阅读 TiDB 原理相关的三篇文章（讲存储，说计算，谈调度），以及 TiDB Best Practice。

11.3.2 高并发批量插入场景

高并发批量插入的场景通常出现在业务系统的批量任务中，例如清算以及结算等业务。此类场景存在以下特点：

- 数据量大
- 需要短时间内将历史数据入库
- 需要短时间内读取大量数据

这就对 TiDB 提出了以下挑战：

- 写入/读取能力是否可以线性水平扩展
- 随着数据持续大并发写入，数据库性能是否稳定不衰减

对于分布式数据库来说，除了本身的基础性能外，最重要的就是充分利用所有节点能力，避免让单个节点成为瓶颈。

11.3.3 TiDB 数据分布原理

如果要解决以上挑战，需要从 TiDB 数据切分以及调度的原理开始讲起。这里只作简单说明，详情可参阅谈调度。

TiDB 以 Region 为单位对数据进行切分，每个 Region 有大小限制（默认 96M）。Region 的切分方式是范围切分。每个 Region 会有多副本，每一组副本，称为一个 Raft Group。每个 Raft Group 中由 Leader 负责执行这块数据的读 & 写（TiDB 支持Follower-Read）。Leader 会自动地被 PD 组件均匀调度在不同的物理节点上，用以均分读写压力。

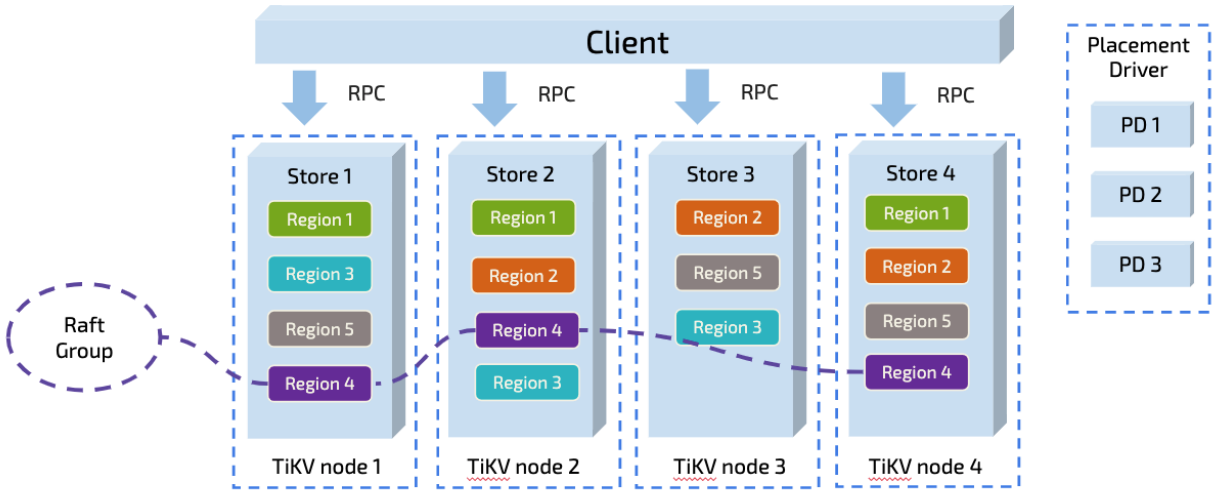


图 220: TiDB 数据概览

从原理上来说，只要没有业务上的写入热点（即业务写入没有 AUTO_INCREMENT 的主键和单调递增的索引，更多细节可参阅 TiDB 正确使用方式），依靠这个架构，TiDB 不仅具备线性扩展的读写能力，也能够充分利用分布式资源。从这一点看，TiDB 尤其适合高并发批量写入场景的业务。

但理论场景和实际情况往往存在不同。以下实例说明了热点是如何产生的。

11.3.4 热点产生的实例

以下为一张示例表：

```
CREATE TABLE IF NOT EXISTS TEST_HOTSPOT(
      id         BIGINT PRIMARY KEY,
      age        INT,
      user_name  VARCHAR(32),
      email      VARCHAR(128)
)
```

这个表的结构非常简单，除了 id 为主键以外，没有额外的二级索引。将数据写入该表的语句如下，id 通过随机数离散生成：

```
INSERT INTO TEST_HOTSPOT(id, age, user_name, email) values(%v, %v, '%v', '%v');
```

负载是短时间内密集地执行以上写入语句。

以上操作看似符合理论场景中的 TiDB 最佳实践，业务上没有热点产生。只要有足够的机器，就可以充分利用 TiDB 的分布式能力。要验证是否真的符合最佳实践，可以在实验环境中进行测试。

部署拓扑 2 个 TiDB 节点，3 个 PD 节点，6 个 TiKV 节点。请忽略 QPS，因为测试只是为了阐述原理，并非 benchmark。

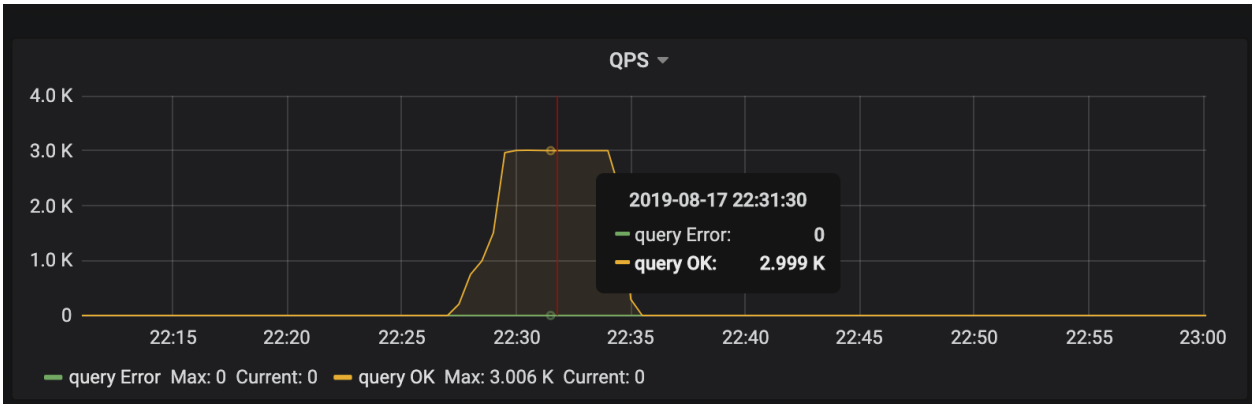


图 221: QPS1

客户端在短时间内发起了“密集”的写入，TiDB 收到的请求是 3K QPS。理论上，压力应该均摊给 6 个 TiKV 节点。但是从 TiKV 节点的 CPU 使用情况上看，存在明显的写入倾斜（tikv - 3 节点是写入热点）：

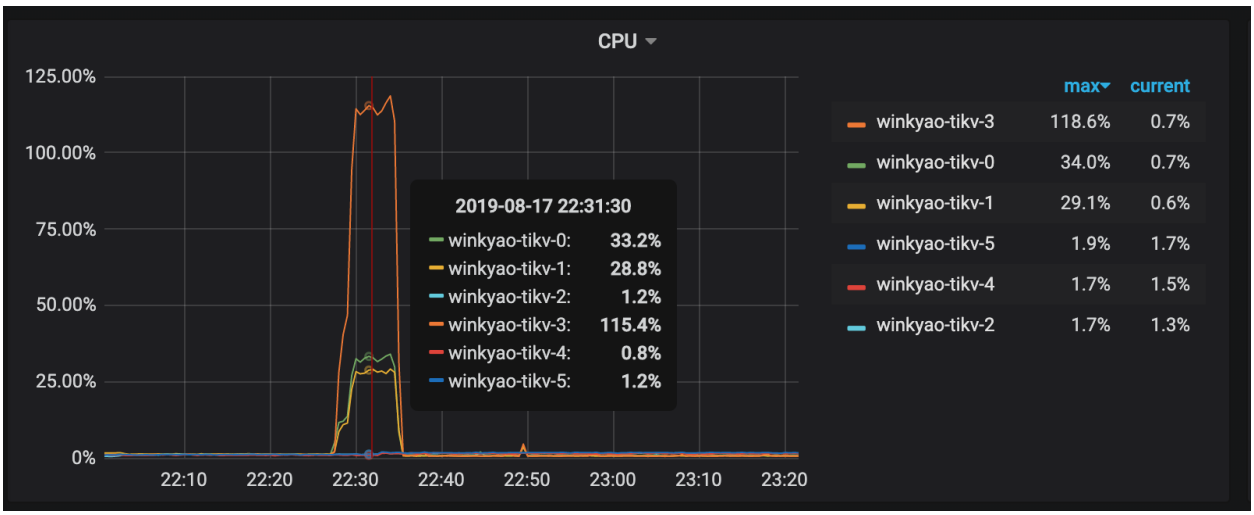


图 222: QPS2

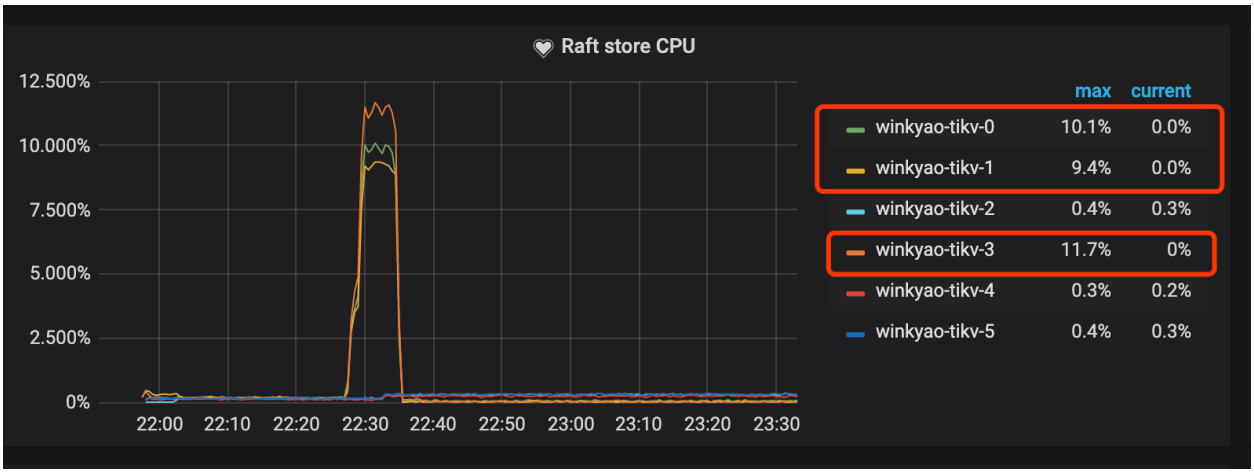


图 223: QPS3

Raft store CPU 为 `raftstore` 线程的 CPU 使用率，通常代表写入的负载。在这个场景下 tikv-3 为 Raft Leader，tikv-0 和 tikv-1 是 Raft 的 Follower，其他的 TiKV 节点的负载几乎为空。

从 PD 的监控中也可以证明热点的产生：

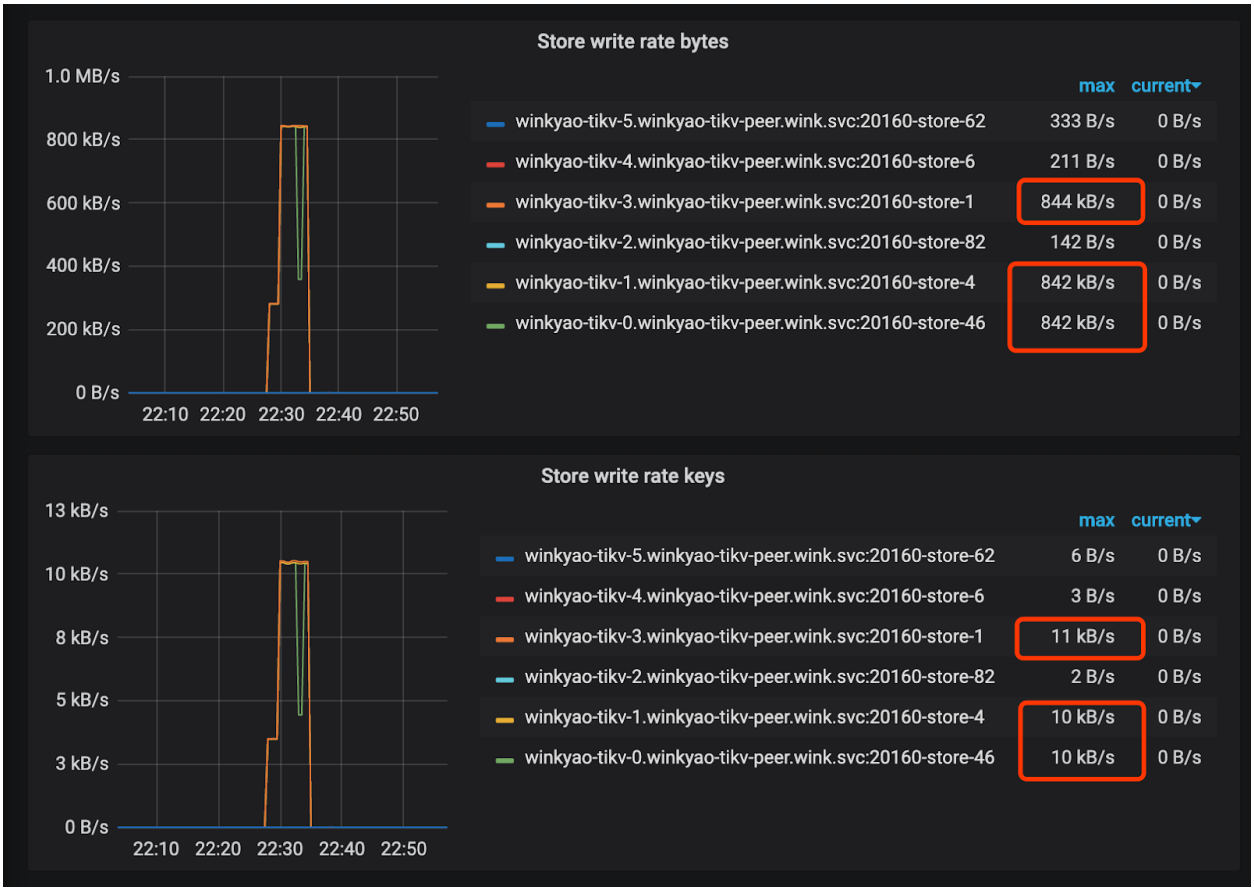


图 224: QPS4

11.3.5 热点问题产生的原因

以上测试并未达到理论场景中最佳实践，因为刚创建表的时候，这个表在 TiKV 中只会对应为一个 Region，范围是：

```
[CommonPrefix + TableID, CommonPrefix + TableID + 1)
```

短时间内大量数据会持续写入到同一个 Region 上。

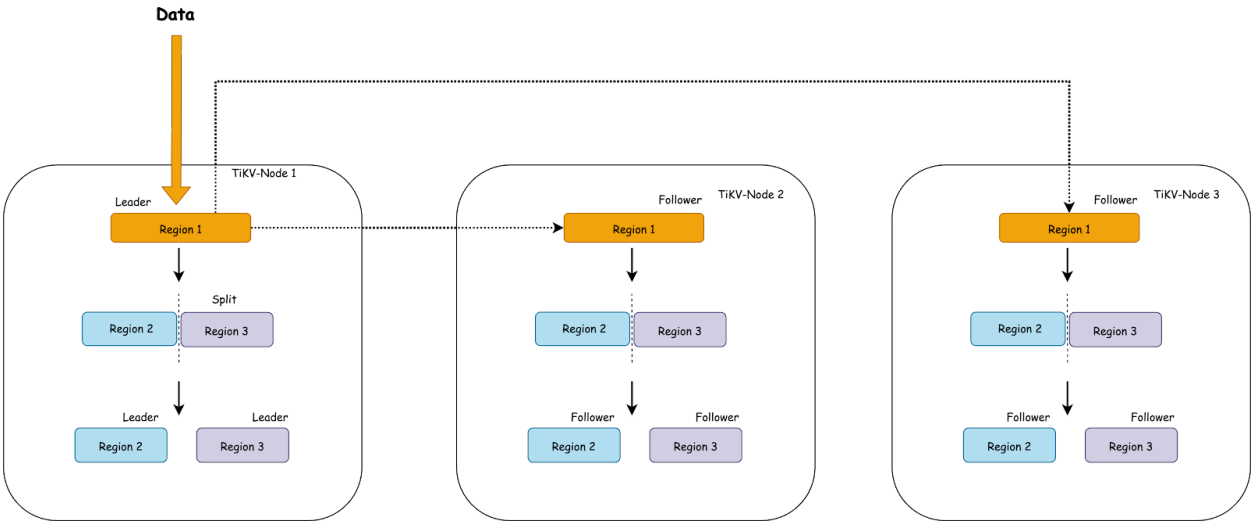


图 225: TiKV Region 分裂流程

上图简单描述了这个过程，随着数据持续写入，TiKV 会将一个 Region 切分为多个。但因为首先发起选举的是原 Leader 所在的 Store，所以新切分好的两个 Region 的 Leader 很可能还会在原 Store 上。新切分好的 Region 2，3 上，也会重复之前发生在 Region 1 上的过程。也就是压力会密集地集中在 TiKV-Node 1 上。

在持续写入的过程中，PD 发现 Node 1 中产生了热点，会将 Leader 均分到其他的 Node 上。如果 TiKV 的节点数多于副本数的话，TiKV 会尽可能将 Region 迁移到空闲的节点上。这两个操作在数据插入的过程中，也能在 PD 监控中得到印证：

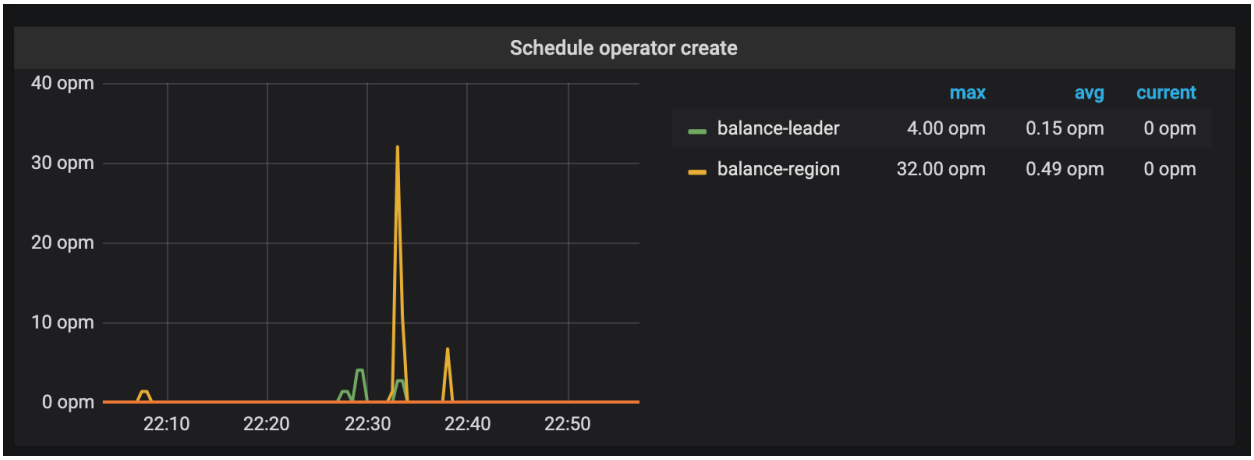


图 226: QPS5

在持续写入一段时间后，整个集群会被 PD 自动地调度成一个压力均匀的状态，到那个时候整个集群的能力才会真正被利用起来。在大多数情况下，以上热点产生的过程是没有问题的，这个阶段属于表 Region 的预热阶段。

但是对于高并发批量密集写入场景来说，应该避免这个阶段。

11.3.6 热点问题的规避方法

为了达到场景理论中的最佳性能，可跳过这个预热阶段，直接将 Region 切分为预期的数量，提前调度到集群的各个节点中。

TiDB 在 v3.0.x 以及 v2.1.13 后支持一个叫 Split Region 的新特性。这个特性提供了新的语法：

```
SPLIT TABLE table_name [INDEX index_name] BETWEEN (lower_value) AND (upper_value) REGIONS
    ↪ region_num
```

```
SPLIT TABLE table_name [INDEX index_name] BY (value_list) [, (value_list)]
```

但是 TiDB 并不会自动提前完成这个切分操作。原因如下：

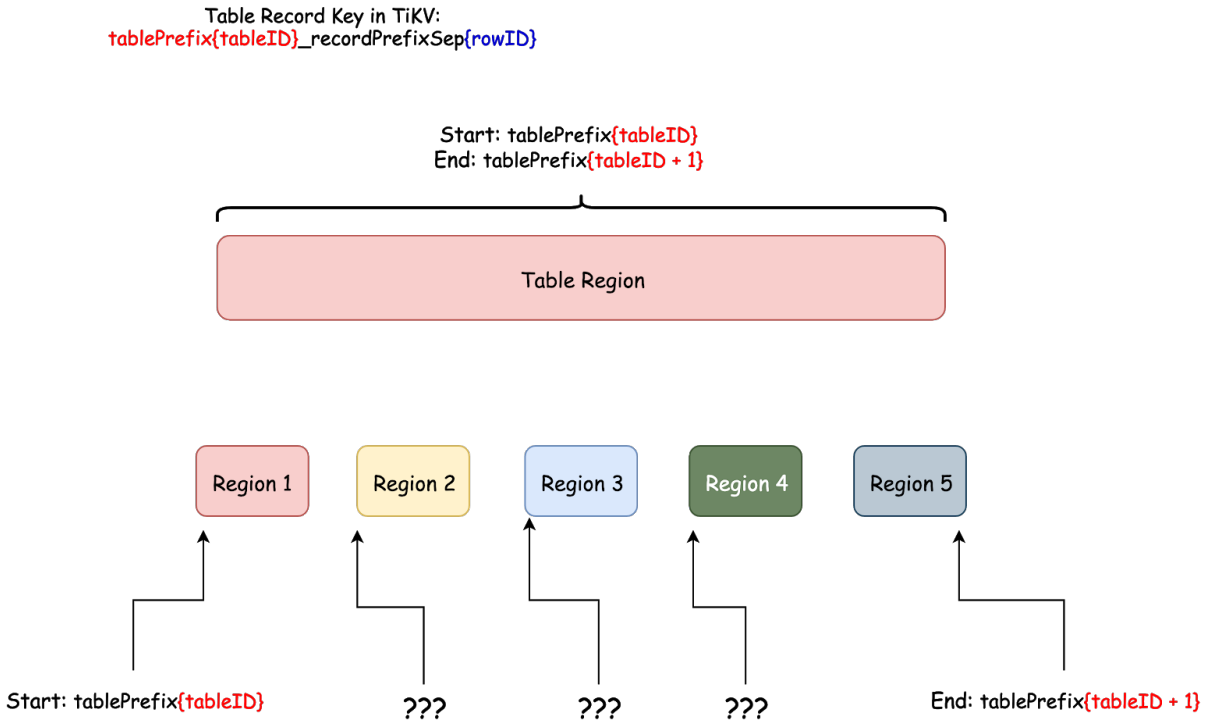


图 227: Table Region Range

从上图可知，根据行数据 key 的编码规则，行 ID (rowID) 是行数据中唯一可变的部分。在 TiDB 中，rowID 是一个 Int64 整型。但是用户不一定能将 Int64 整型范围均匀切分成需要的份数，然后均匀分布在不同的节点上，还需要结合实际情况。

如果行 ID 的写入是完全离散的，那么上述方式是可行的。如果行 ID 或者索引有固定的范围或者前缀（例如，只在 [2000w, 5000w) 的范围内离散插入数据），这种写入依然在业务上不产生热点，但是如果按上面的方式进行切分，那么有可能一开始数据仍只写入到某个 Region 上。

作为一款通用数据库，TiDB 并不对数据的分布作假设，所以开始只用一个 Region 来对应一个表。等到真实数据插入进来以后，TiDB 自动根据数据的分布来作切分。这种方式是较通用的。

所以 TiDB 提供了 Split Region 语法，专门针对短时批量写入场景作优化。基于以上案例，下面尝试用 Split Region 语法提前切散 Region，再观察负载情况。

由于测试的写入数据在正数范围内完全离散，所以用以下语句，在 Int64 空间内提前将表切分为 128 个 Region：

```
SPLIT TABLE TEST_HOTSPOT BETWEEN (0) AND (9223372036854775807) REGIONS 128;
```

切分完成以后，可以通过 SHOW TABLE test_hotspot REGIONS; 语句查看打散的情况。如果 SCATTERING 列值全部为 0，代表调度成功。

也可以通过 table-regions.py 脚本，查看 Region 的分布。目前分布已经比较均匀了：

```
[root@172.16.4.4 scripts]# python table-regions.py --host 172.16.4.3 --port 31453 test
    ↪ test_hotspot
[RECORD - test.test_hotspot] - Leaders Distribution:
  total leader count: 127
  store: 1, num_leaders: 21, percentage: 16.54%
  store: 4, num_leaders: 20, percentage: 15.75%
  store: 6, num_leaders: 21, percentage: 16.54%
  store: 46, num_leaders: 21, percentage: 16.54%
  store: 82, num_leaders: 23, percentage: 18.11%
  store: 62, num_leaders: 21, percentage: 16.54%
```

再重新运行写入负载：

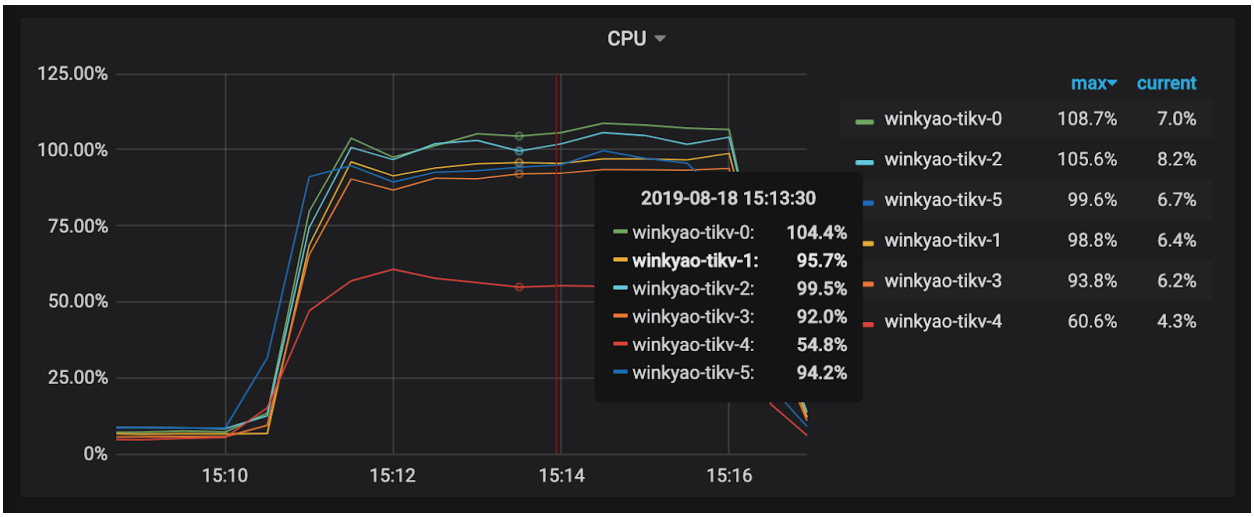


图 228: QPS6

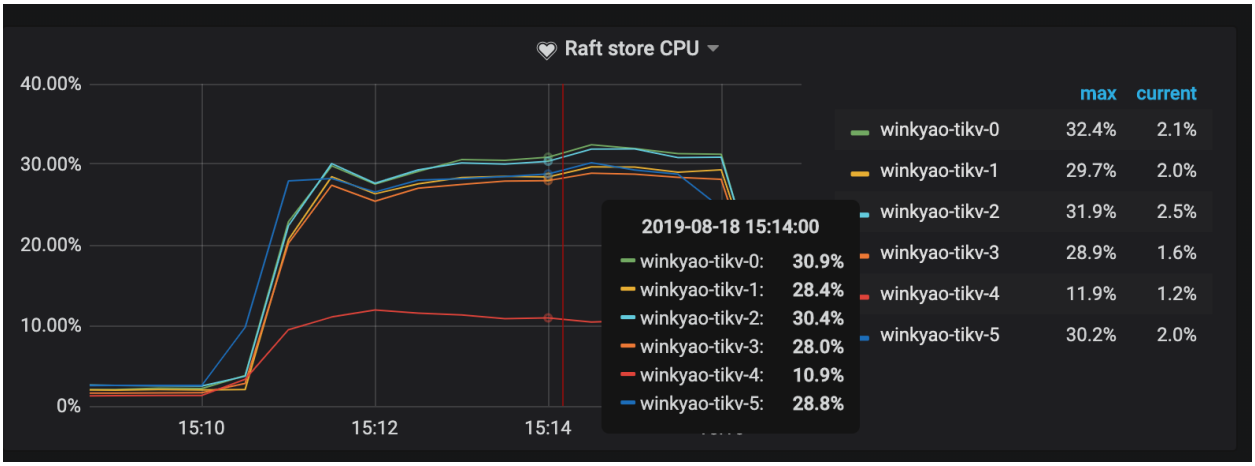


图 229: QPS7

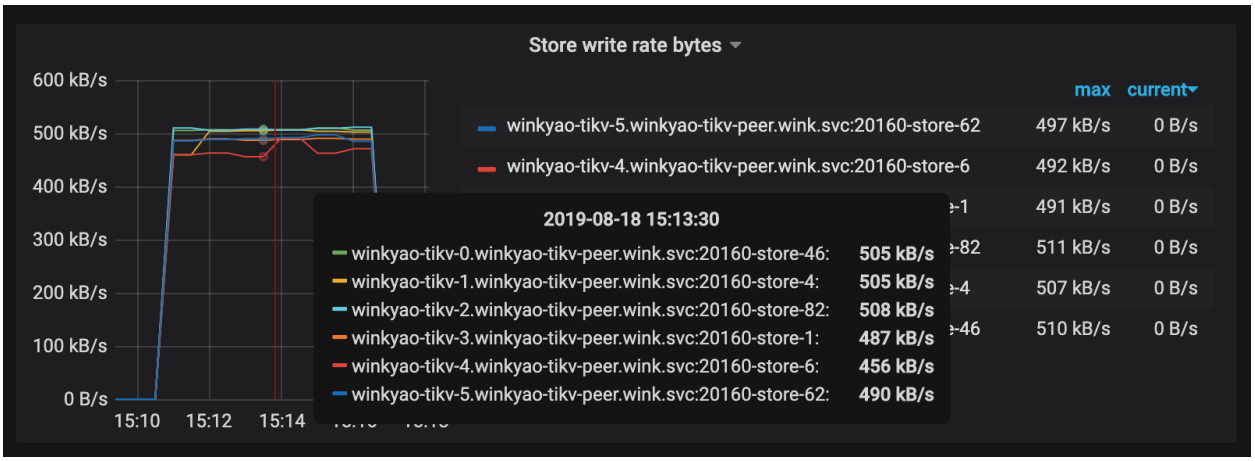


图 230: QPS8

可以看到已经消除了明显的热点问题了。

本示例仅为一个简单的表，还有索引热点的问题需要考虑。读者可参阅 Split Region 文档来了解如何预先切散索引相关的 Region。

11.3.6.1 更复杂的热点问题

问题一：

如果表没有主键或者主键不是整数类型，而且用户也不想自己生成一个随机分布的主键 ID 的话，TiDB 内部有一个隐式的 `_tidb_rowid` 列作为行 ID。在不使用 `SHARD_ROW_ID_BITS` 的情况下，`_tidb_rowid` 列的值基本也为单调递增，此时也会有写热点存在（参阅 `SHARD_ROW_ID_BITS` 的详细说明）。

要避免由 `_tidb_rowid` 带来的写入热点问题，可以在建表时，使用 `SHARD_ROW_ID_BITS` 和 `PRE_SPLIT_REGIONS` 这两个建表选项（参阅`PRE_SPLIT_REGIONS` 的详细说明）。

SHARD_ROW_ID_BITS 用于将 _tidb_rowid 列生成的行 ID 随机打散。PRE_SPLIT_REGIONS 用于在建完表后预先进行 Split region。

> **注意：**
>
> PRE_SPLIT_REGIONS 的值必须小于或等于 SHARD_ROW_ID_BITS。

以下全局变量会影响 PRE_SPLIT_REGIONS 的行为，需要特别注意：

- tidb_scatter_region：该变量用于控制建表完成后是否等待预切分和打散 Region 完成后再返回结果。如果建表后有大批量写入，需要设置该变量值为 1，表示等待所有 Region 都切分和打散完成后再返回结果给客户端。否则未打散完成就进行写入会对写入性能影响有较大的影响。

示例：

```
create table t (a int, b int) SHARD_ROW_ID_BITS = 4 PRE_SPLIT_REGIONS=3;
```

- SHARD_ROW_ID_BITS = 4 表示 tidb_rowid 的值会随机分布成 16 (16=2^4) 个范围区间。
- PRE_SPLIT_REGIONS=3 表示建完表后提前切分出 8 (2^3) 个 Region。

开始写数据进表 t 后，数据会被写入提前切分好的 8 个 Region 中，这样也避免了刚开始建表完后因为只有一个 Region 而存在的写热点问题。

问题二：

如果表的主键为整数类型，并且该表使用了 AUTO_INCREMENT 来保证主键唯一性（不需要连续或递增）的表而言，由于 TiDB 直接使用主键行值作为 _tidb_rowid，此时无法使用 SHARD_ROW_ID_BITS 来打散热点。

要解决上述热点问题，可以利用 AUTO_RANDOM 列属性（参阅 AUTO_RANDOM 的详细说明），将 AUTO_INCREMENT 改为 AUTO_RANDOM，插入数据时让 TiDB 自动为整型主键列分配一个值，消除行 ID 的连续性，从而达到打散热点的目的。

11.3.7 参数配置

TiDB 2.1 版本中在 SQL 层引入了 latch 机制，用于在写入冲突比较频繁的场景中提前发现事务冲突，减少 TiDB 和 TiKV 事务提交时写写冲突导致的重试。通常，跑批场景使用的是存量数据，所以并不存在事务的写入冲突。可以把 TiDB 的 latch 功能关闭，以减少为细小对象分配内存：

```
[txn-local-latches]
enabled = false
```

11.4 在三数据中心下就近读取数据

在三数据中心模式下，Region 的三个副本都会隔离在各个数据中心里。然而在强一致读的要求下，TiDB 的每一个查询都需要访问对应数据的 Leader 副本，而查询的来源可能和 Leader 所在的数据中心不一致，这就会引起跨数据中心的数据访问，从而造成访问的延迟上升。本文主要介绍使用 Stale Read 功能，以牺牲数据实时性的方式，避免跨数据中心的访问，从而降低访问的延迟。

11.4.1 部署三数据中心的 TiDB 集群

部署三数据中心的方法，参考同城多数据中心部署 TiDB。

请注意，如果 TiKV 和 TiDB 都有 `labels` 配置项，在给 TiKV 和 TiDB 配置标签时，同一个数据中心下的 TiKV 和 TiDB 应该配置相同的 zone 标签。假设 TiKV 和 TiDB 都在 `dc-1` 数据中心下，那么两者都需要配置如下标签：

```
[labels]
zone=dc-1
```

11.4.2 使用 Stale Read 就近读取数据

Stale Read 为用户提供了一种读取历史数据的一种机制。使用 Stale Read 功能，你能从指定时间点或时间范围内读取对应的历史数据，从而避免数据同步带来延迟。在部分跨数据中心部署的场景中使用 Stale Read 功能，通过牺牲一定的实时性，TiDB 可就近访问对应数据所在当前中心的副本，避免跨数据中心的网络延迟，降低整体查询的访问延迟。

当 TiDB 收到 Stale Read 查询时，假如对应的 TiDB 配置了 zone 标签，就会将请求发送到对应数据副本所在 TiKV 拥有相同的 zone 标签的节点上。

如何使用 Stale Read 查询，参考使用 AS OF TIMESTAMP 语法读取历史数据。

11.5 使用 Grafana 监控 TiDB 的最佳实践

使用 TiUP 部署 TiDB 集群时，如果在拓扑配置中添加了 Grafana 和 Prometheus，会部署一套Grafana + Prometheus 的监控平台，用于收集和展示 TiDB 集群各个组件和机器的 metric 信息。本文主要介绍使用 TiDB 监控的最佳实践，旨在帮助 TiDB 用户高效利用丰富的 metric 信息来分析 TiDB 的集群状态或进行故障诊断。

11.5.1 监控架构

Prometheus 是一个拥有多维度数据模型和灵活查询语句的时序数据库。Grafana 是一个开源的 metric 分析及可视化系统。

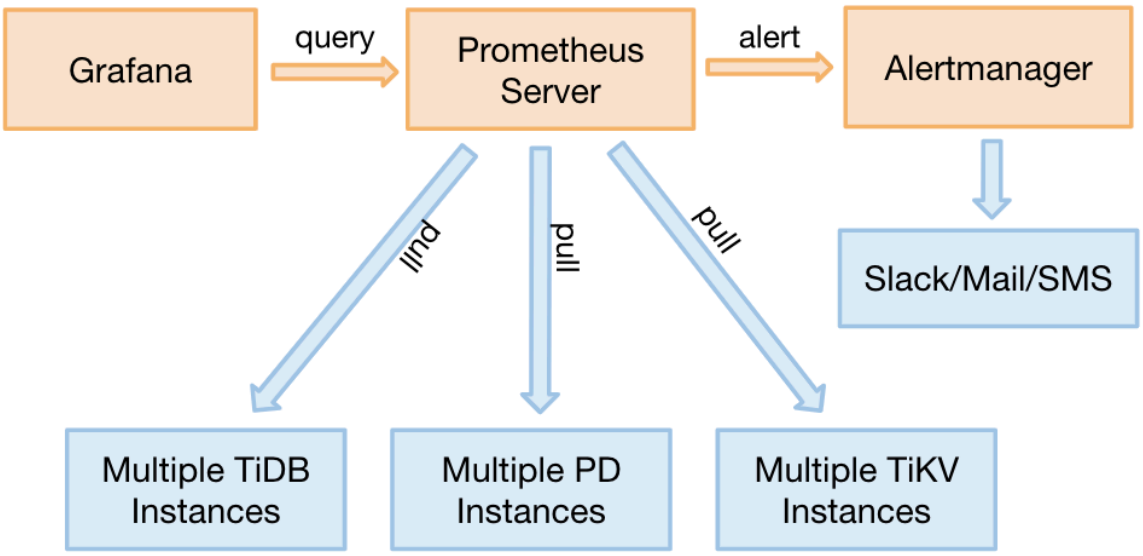


图 231: TiDB 监控整体架构

从 TiDB 2.1.3 版本开始，监控可以支持 pull，这是一个非常好的调整，它有以下几个优点：

- 如果 Prometheus 需要迁移，无需重启整个 TiDB 集群。调整前，因为组件要调整 push 的目标地址，迁移 Prometheus 需要重启整个集群。
- 支持部署 2 套独立的 Grafana + Prometheus 的监控平台（非 HA），防止监控的单点。
- 去掉了 Pushgateway 这个单点组件。

11.5.2 监控数据的来源与展示

TiDB 的 3 个核心组件（TiDB server、TiKV server 和 PD server）可以通过 HTTP 接口来获取 metric 数据。这些 metric 均是从程序代码中上传的，默认端口如下：

组件	端口
TiDB server	10080
TiKV server	20181
PD server	2379

下面以 TiDB server 为例，展示如何通过 HTTP 接口查看一个语句的 QPS 数据：

```
curl http://__tidb_ip__:10080/metrics |grep tidb_executor_statement_total
```

```
## 可以看到实时 QPS 数据，并根据不同 type 对 SQL 语句进行了区分，value 是 counter 类型的累计值（
    ↪ 科学计数法）。
tidb_executor_statement_total{type="Delete"} 520197
```

```
tidb_executor_statement_total{type="Explain"} 1
tidb_executor_statement_total{type="Insert"} 7.20799402e+08
tidb_executor_statement_total{type="Select"} 2.64983586e+08
tidb_executor_statement_total{type="Set"} 2.399075e+06
tidb_executor_statement_total{type="Show"} 500531
tidb_executor_statement_total{type="Use"} 466016
```

这些数据会存储在 Prometheus 中，然后在 Grafana 上进行展示。在面板上点击鼠标右键会出现 Edit 按钮（或直接按 E 键），如下图所示：

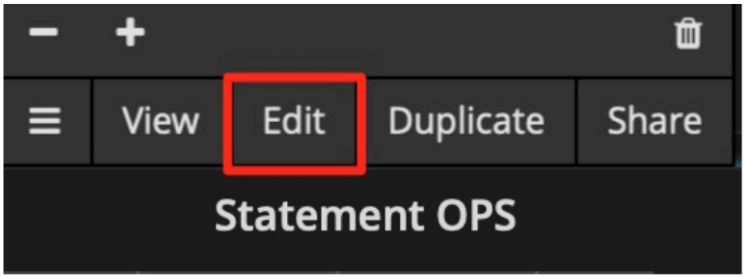


图 232: Metrics 面板的编辑入口

点击 Edit 按钮之后，在 Metrics 面板上可以看到利用该 metric 的 query 表达式。面板上一些细节的含义如下：

- `rate[1m]`：表示 1 分钟的增长速率，只能用于 counter 类型的数据。
- `sum`：表示 value 求和。
- `by type`：表示将求和后的数据按 metric 原始值中的 type 进行分组。
- `Legend format`：表示指标名称的格式。
- `Resolution`：默认打点步长是 15s，Resolution 表示是否将多个样本数据合并成一个点。

Metrics 面板中的表达式如下：



图 233: Metric 面板中的表达式

Prometheus 支持很多表达式与函数，更多表达式请参考 Prometheus 官网页面。

11.5.3 Grafana 使用技巧

本小节介绍高效利用 Grafana 监控分析 TiDB 指标的七个技巧。

11.5.3.1 技巧 1：查看所有维度并编辑表达式

在监控数据的来源与展示一节的示例中，数据是按照 type 进行分组的。如果你想知道是否还能按其它维度分组，并快速查看还有哪些维度，可采用以下技巧：在 query 的表达式上只保留指标名称，不做任何计算，Legend format 也留空。这样就能显示出原始的 metric 数据。比如，下图能看到有 3 个维度（instance、job 和 type）：

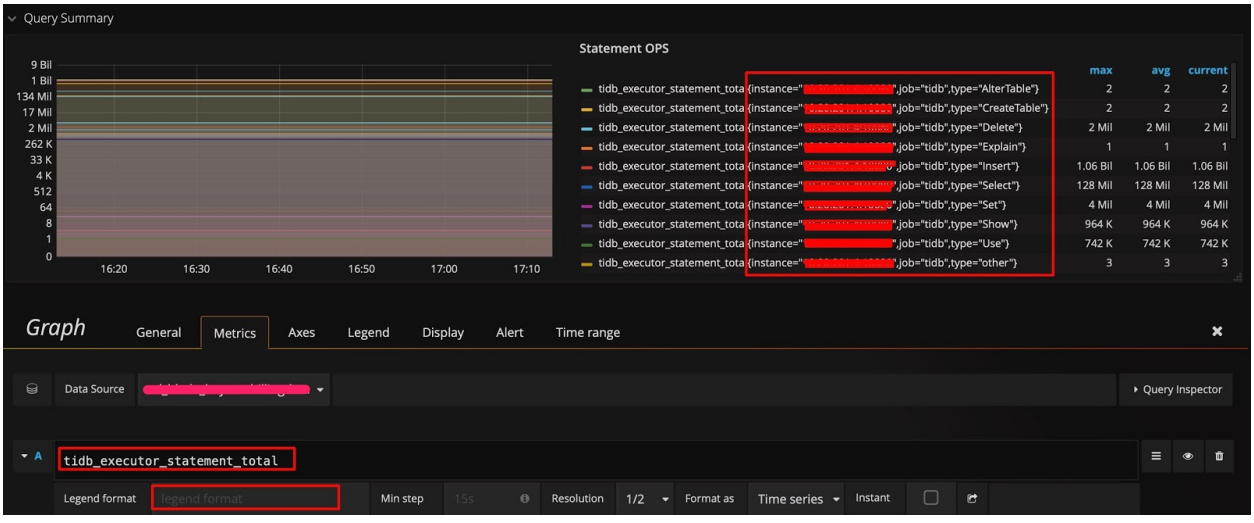


图 234: 编辑表达式并查看所有维度

然后调整表达式，在原有的 type 后面加上 instance 这个维度，在 Legend format 处增加 {{instance}}，就可以看到每个 TiDB server 上执行的不同类型 SQL 语句的 QPS 了。如下图所示：

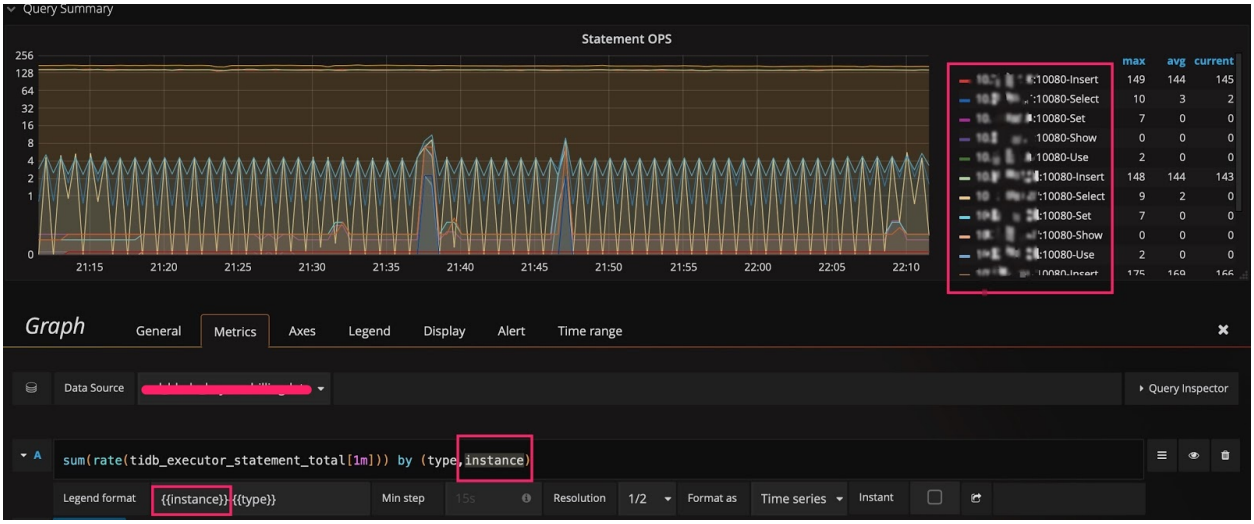


图 235: 给表达式增加一个 instance 维度

11.5.3.2 技巧 2：调整 Y 轴标尺的计算方式

以 Query Duration 指标为例，默认的比例尺采用 2 的对数计算，显示上会将差距缩小。为了观察到明显的变化，可以将比例尺改为线性，从下面两张图中可以看到显示上的区别，明显发现那个时刻有个 SQL 语句运行较慢。

当然也不是所有场景都适合用线性，比如观察 1 个月的性能趋势，用线性可能就会有很多噪点，不好观察。

标尺默认的比例尺为 2 的对数：



图 236: 标尺默认的比例尺为 2 的对数

将标尺的比例尺调整为线性：

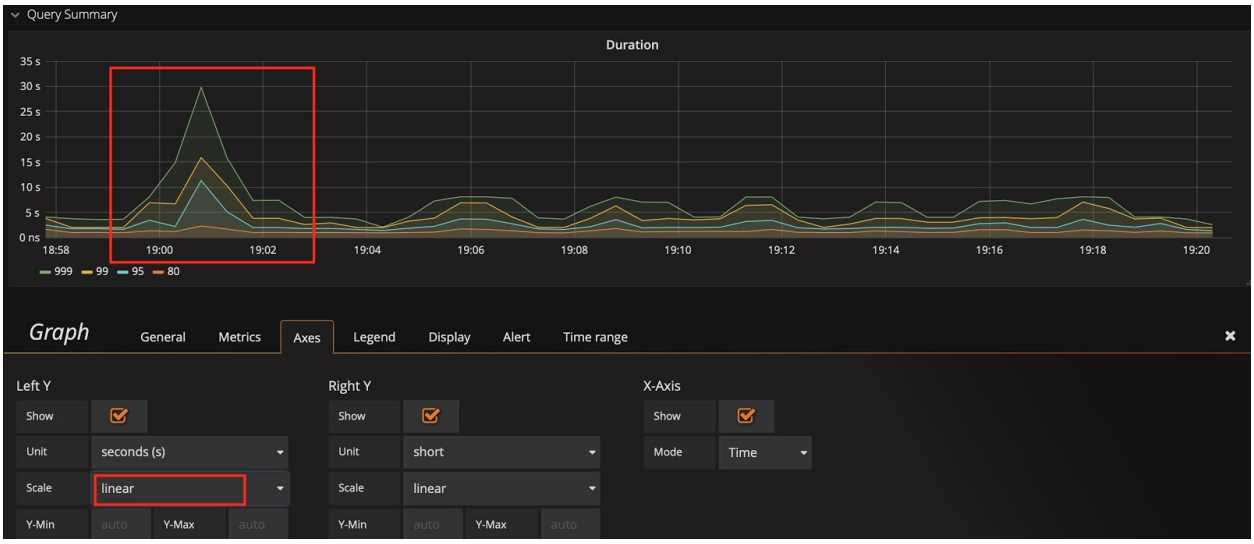


图 237: 调整标尺的比例尺为线性

> **建议：**
>
> 结合技巧 1，会发现这里还有一个 `sql_type` 的维度，可以立刻分析出是 `SELECT` 慢还是 `UPDATE` 慢；并且可以分析出是哪个 instance 上的语句慢。

11.5.3.3 技巧 3：调整 Y 轴基线，放大变化

有时已经用了线性比例尺，却还是看不出变化趋势。比如下图中，在扩容后想观察 `Store size` 的实时变化效果，但由于基数较大，观察不到微弱的变化。这时可以将 Y 轴最小值从 0 改为 `auto`，将上部放大。观察下面两张图的区别，可以看出数据已开始迁移了。

基线默认为 0：

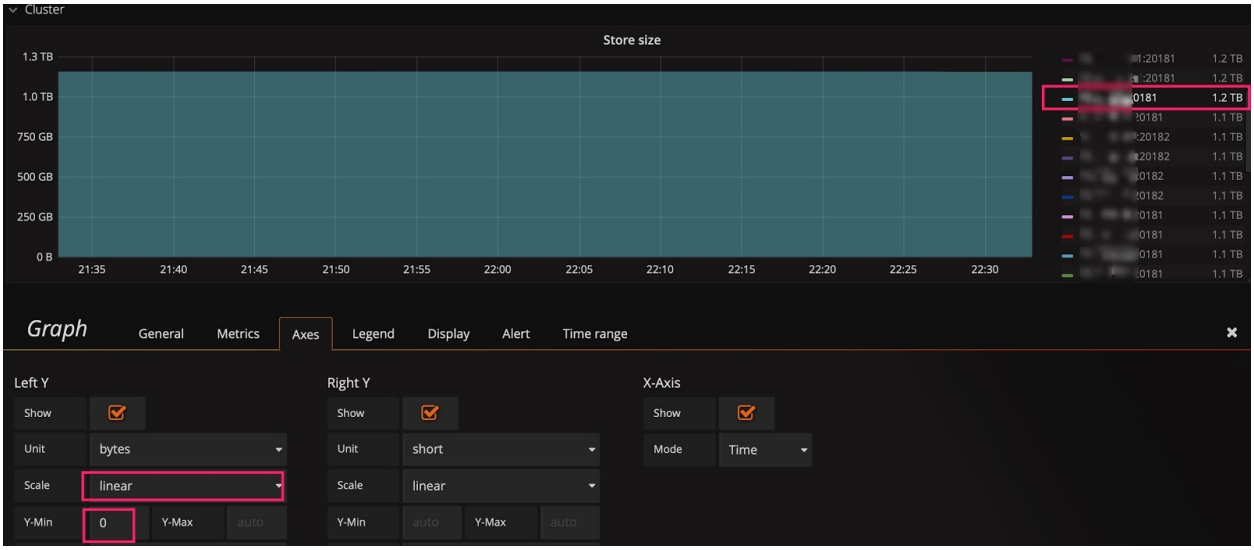


图 238: 基线默认为 0

将基线调整为 auto：



图 239: 调整基线为 auto

11.5.3.4 技巧 4：标尺联动

在 Settings 面板中，有一个 Graph Tooltip 设置项，默认使用 Default。

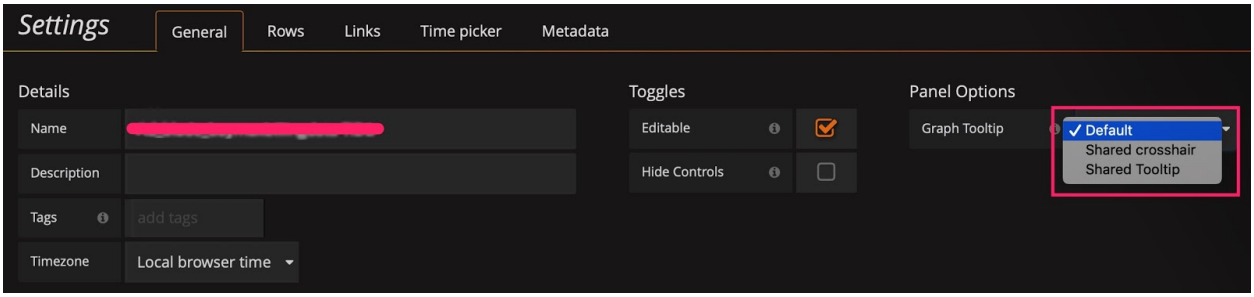


图 240: 图形展示工具

下面将图形展示工具分别调整为 Shared crosshair 和 Shared Tooltip 看看效果。可以看到标尺能联动展示了，方便排查问题时确认 2 个指标的关联性。

将图形展示工具调整为 Shared crosshair：

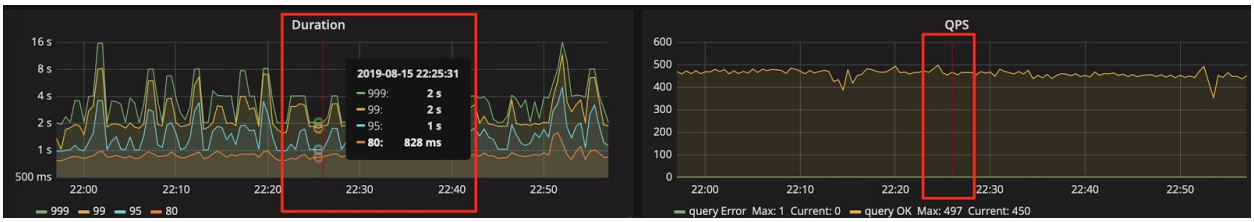


图 241: 调整图形展示工具为 Shared crosshair

将图形展示工具调整为 Shared Tooltip：

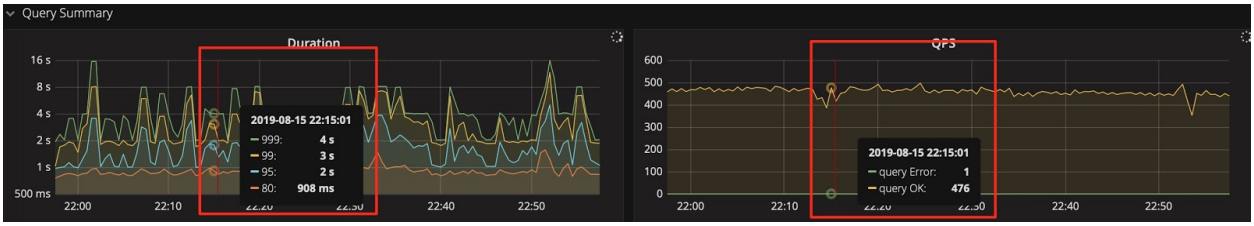


图 242: 调整图形展示工具为 Shared Tooltip

11.5.3.5 技巧 5：手动输入 ip：端口号查看历史信息

PD 的 dashboard 只展示当前 leader 的 metric 信息，而有时想看历史上 PD leader 当时的状况，但是 instance 下拉列表中已不存在这个成员了。此时，可以手动输入 ip:2379 来查看当时的数据。

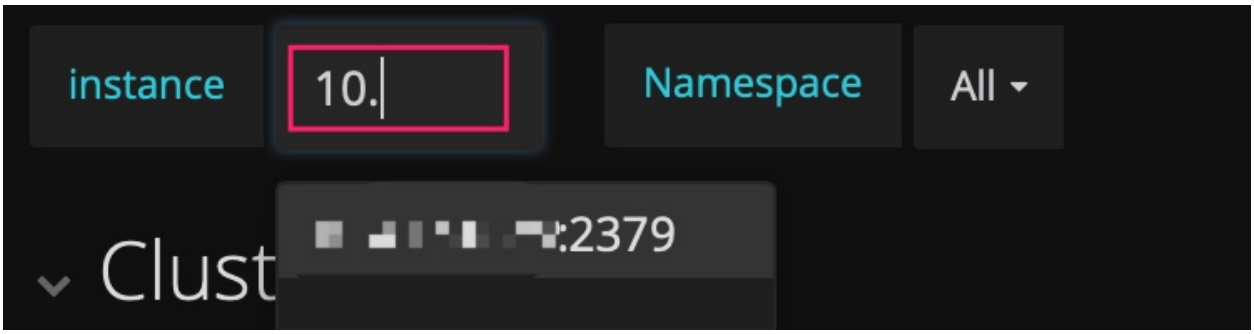


图 243: 查看历史 metric 信息

11.5.3.6 技巧 6：巧用 Avg 函数

通常默认图例中只有 Max 和 Current 函数。当指标波动较大时，可以增加 Avg 等其它汇总函数的图例，来看一段时间的整体趋势。

增加 Avg 等汇总函数：

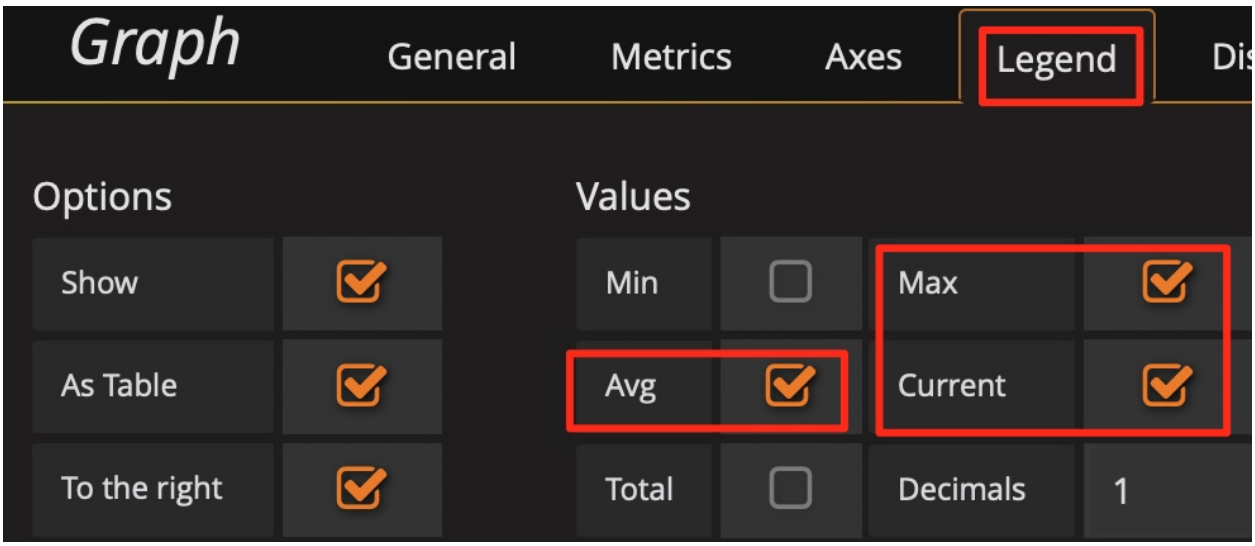


图 244: 增加 Avg 等汇总函数

然后查看整体趋势：

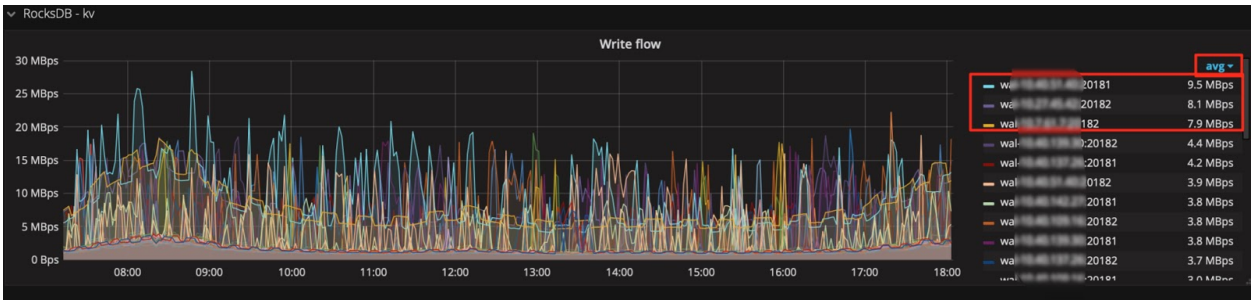


图 245: 增加 Avg 函数查看整体趋势

11.5.3.7 技巧 7：使用 Prometheus 的 API 接口获得表达式的结果

Grafana 通过 Prometheus 的接口获取数据，你也可以用该接口来获取数据，这个用法还可以衍生出许多功能：

- 自动获取集群规模、状态等信息。
- 对表达式稍加改动给报表提供数据，如统计每天的 QPS 总量、每天的 QPS 峰值和每天的响应时间。
- 将重要的指标进行定期健康巡检。

Prometheus 的 API 接口如下：

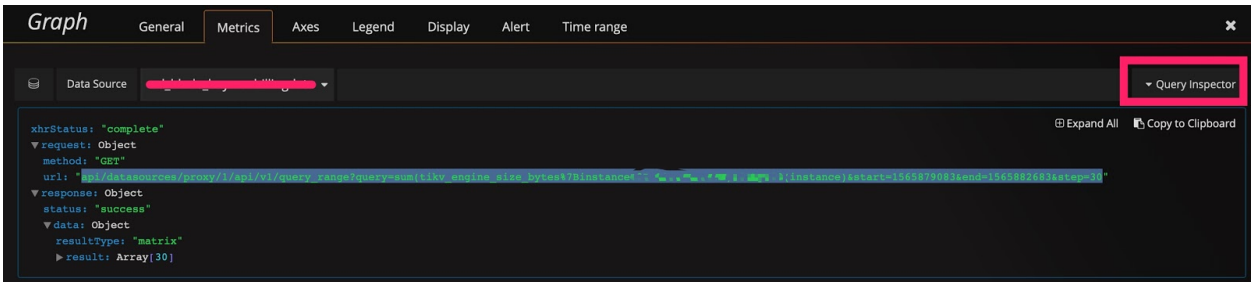


图 246: Prometheus 的 API 接口

```
curl -u user:pass 'http://__grafana_ip__:3000/api/datasources/proxy/1/api/v1/query_range?query=
    ↪ sum(tikv_engine_size_bytes%7Binstancexxxxxxxxx20181%22%7D)%20by%20(instance)&start
    ↪ =1565879269&end=1565882869&step=30' |python -m json.tool
```

```
{
    "data": {
        "result": [
            {
                "metric": {
                    "instance": "xxxxxxxxxx:20181"
                },
                "values": [
                    [
```

```
                    1565879269,
                    "1006046235280"
                ],
                [
                    1565879299,
                    "1006057877794"
                ],
                [
                    1565879329,
                    "1006021550039"
                ],
                [
                    1565879359,
                    "1006021550039"
                ],
                [
                    1565882869,
                    "1006132630123"
                ]
            ]
        }
    ],
    "resultType": "matrix"
},
"status": "success"
}
```

11.5.4 总结

Grafana + Prometheus 监控平台是一套非常强大的组合工具，用好这套工具可以为分析节省很多时间，提高效率，更重要的是，我们可以更容易发现问题。在运维 TiDB 集群，尤其是数据量大的情况下，这套工具能派上大用场。

11.6 PD 调度策略最佳实践

本文将详细介绍 PD 调度系统的原理，并通过几个典型场景的分析和处理方式，分享调度策略的最佳实践和调优方式，帮助大家在使用过程中快速定位问题。本文假定你对 TiDB，TiKV 以及 PD 已经有一定的了解，相关核心概念如下：

- Leader/Follower/Learner
- Operator
- Operator Step
- Pending/Down
- Region/Peer/Raft Group

- Region Split
- Scheduler
- Store

> 注意：
>
> 本文内容基于 TiDB 3.0 版本，更早的版本 (2.x) 缺少部分功能的支持，但是基本原理类似，也可以以本文作为参考。

11.6.1 PD 调度原理

该部分介绍调度系统涉及到的相关原理和流程。

11.6.1.1 调度流程

宏观上来看，调度流程大体可划分为 3 个部分：

1. 信息收集

   TiKV 节点周期性地向 PD 上报 `StoreHeartbeat` 和 `RegionHeartbeat` 两种心跳消息：

   - `StoreHeartbeat` 包含 Store 的基本信息、容量、剩余空间和读写流量等数据。
   - `RegionHeartbeat` 包含 Region 的范围、副本分布、副本状态、数据量和读写流量等数据。

   PD 梳理并转存这些信息供调度进行决策。

2. 生成调度

   不同的调度器从自身的逻辑和需求出发，考虑各种限制和约束后生成待执行的 Operator。这里所说的限制和约束包括但不限于：

   - 不往处于异常状态中（如断连、下线、繁忙、空间不足或在大量收发 snapshot 等）的 Store 添加副本
   - Balance 时不选择状态异常的 Region
   - 不尝试把 Leader 转移给 Pending Peer
   - 不尝试直接移除 Leader
   - 不破坏 Region 各种副本的物理隔离
   - 不破坏 Label property 等约束

3. 执行调度

   调度执行的步骤为：

   a. Operator 先进入一个由 `OperatorController` 管理的等待队列。

   b. `OperatorController` 会根据配置以一定的并发量从等待队列中取出 Operator 并执行。执行的过程就是依次把每个 Operator Step 下发给对应 Region 的 Leader。

   c. 标记 Operator 为 “finish” 或 “timeout” 状态，然后从执行列表中移除。

11.6.1.2 负载均衡

Region 负载均衡调度主要依赖 balance-leader 和 balance-region 两个调度器。二者的调度目标是将 Region 均匀地分散在集群中的所有 Store 上，但它们各有侧重：balance-leader 关注 Region 的 Leader，目的是分散处理客户端请求的压力；balance-region 关注 Region 的各个 Peer，目的是分散存储的压力，同时避免出现爆盘等状况。

balance-leader 与 balance-region 有着相似的调度流程：

1. 根据不同 Store 的对应资源量的情况分别打分。
2. 不断从得分较高的 Store 选择 Leader 或 Peer 迁移到得分较低的 Store 上。

两者的分数计算上也有一定差异：balance-leader 比较简单，使用 Store 上所有 Leader 所对应的 Region Size 加和作为得分。因为不同节点存储容量可能不一致，计算 balance-region 得分会分以下三种情况：

- 当空间富余时使用数据量计算得分（使不同节点数据量基本上均衡）
- 当空间不足时由使用剩余空间计算得分（使不同节点剩余空间基本均衡）
- 处于中间态时则同时考虑两个因素做加权和当作得分

此外，为了应对不同节点可能在性能等方面存在差异的问题，还可为 Store 设置负载均衡的权重。leader-weight 和 region-weight 分别用于控制 Leader 权重以及 Region 权重（默认值都为 “1”）。假如把某个 Store 的 leader-weight 设为 “2”，调度稳定后，则该节点的 Leader 数量约为普通节点的 2 倍；假如把某个 Store 的 region-weight 设为 “0.5”，那么调度稳定后该节点的 Region 数量约为其他节点的一半。

11.6.1.3 热点调度

热点调度对应的调度器是 hot-region-scheduler。从 TiDB v3.0 版本开始，统计热点 Region 的方式为：

1. 根据 Store 上报的信息，统计出持续一段时间读或写流量超过一定阈值的 Region。
2. 用与负载均衡类似的方式把这些 Region 分散开来。

对于写热点，热点调度会同时尝试打散热点 Region 的 Peer 和 Leader；对于读热点，由于只有 Leader 承载读压力，热点调度会尝试将热点 Region 的 Leader 打散。

11.6.1.4 集群拓扑感知

让 PD 感知不同节点分布的拓扑是为了通过调度使不同 Region 的各个副本尽可能分散，保证高可用和容灾。PD 会在后台不断扫描所有 Region，当发现 Region 的分布不是当前的最优化状态时，会生成调度以替换 Peer，将 Region 调整至最佳状态。

负责这个检查的组件叫 replicaChecker（跟 Scheduler 类似，但是不可关闭）。它依赖于 location-labels 配置项来进行调度。比如配置 [zone,rack,host] 定义了三层的拓扑结构：集群分为多个 zone（可用区），每个 zone 下有多个 rack（机架），每个 rack 下有多个 host（主机）。PD 在调度时首先会尝试将 Region 的 Peer 放置在不同的 zone，假如无法满足（比如配置 3 副本但总共只有 2 个 zone）则保证放置在不同的 rack；假如 rack 的数量也不足以保证隔离，那么再尝试 host 级别的隔离，以此类推。

11.6.1.5 缩容及故障恢复

缩容是指预备将某个 Store 下线，通过命令将该 Store 标记为 “Offline” 状态，此时 PD 通过调度将待下线节点上的 Region 迁移至其他节点。

故障恢复是指当有 Store 发生故障且无法恢复时，有 Peer 分布在对应 Store 上的 Region 会产生缺少副本的状况，此时 PD 需要在其他节点上为这些 Region 补副本。

这两种情况的处理过程基本上是一样的。`replicaChecker` 检查到 Region 存在异常状态的 Peer 后，生成调度在健康的 Store 上创建新副本替换异常的副本。

11.6.1.6 Region merge

Region merge 指的是为了避免删除数据后大量小甚至空的 Region 消耗系统资源，通过调度把相邻的小 Region 合并的过程。Region merge 由 `mergeChecker` 负责，其过程与 `replicaChecker` 类似：PD 在后台遍历，发现连续的小 Region 后发起调度。

具体来说，当某个新分裂出来的 Region 存在的时间超过配置项`split-merge-interval` 的值（默认 1h）后，如果出现以下任意情况，该 Region 会触发 Region merge 调度：

- 该 Region 大小小于配置项`max-merge-region-size` 的值（默认 20MiB）
- 该 Region 中 key 的数量小于配置项`max-merge-region-keys` 的值（默认 200000）

11.6.2 查询调度状态

你可以通过观察 PD 相关的 Metrics 或使用 pd-ctl 工具等方式查看调度系统状态。更具体的信息可以参考PD 监控和 PD Control。

11.6.2.1 Operator 状态

Grafana PD/Operator 页面展示了 Operator 的相关统计信息。其中比较重要的有：

- Schedule Operator Create：Operator 的创建情况
- Operator finish duration：Operator 执行耗时的情况
- Operator Step duration：不同 Operator Step 执行耗时的情况

查询 Operator 的 pd-ctl 命令有：

- `operator show`：查询当前调度生成的所有 Operator
- `operator show [admin | leader | region]`：按照类型查询 Operator

11.6.2.2 Balance 状态

Grafana PD/Statistics - Balance 页面展示了负载均衡的相关统计信息，其中比较重要的有：

- Store Leader/Region score：每个 Store 的得分
- Store Leader/Region count：每个 Store 的 Leader/Region 数量
- Store available：每个 Store 的剩余空间

使用 pd-ctl 的 `store` 命令可以查询 Store 的得分、数量、剩余空间和 weight 等信息。

11.6.2.3 热点调度状态

Grafana PD/Statistics - hotspot 页面展示了热点 Region 的相关统计信息，其中比较重要的有：

- Hot write Region’s leader/peer distribution：写热点 Region 的 Leader/Peer 分布情况
- Hot read Region’s leader distribution：读热点 Region 的 Leader 分布情况

使用 pd-ctl 同样可以查询上述信息，可以使用的命令有：

- `hot read`：查询读热点 Region 信息
- `hot write`：查询写热点 Region 信息
- `hot store`：按 Store 统计热点分布情况
- `region topread [limit]`：查询当前读流量最大的 Region
- `region topwrite [limit]`：查询当前写流量最大的 Region

11.6.2.4 Region 健康度

Grafana PD/Cluster/Region health 面板展示了异常 Region 的相关统计信息，包括 Pending Peer、Down Peer、Offline Peer，以及副本数过多或过少的 Region。

通过 pd-ctl 的 `region check` 命令可以查看具体异常的 Region 列表：

- `region check miss-peer`：缺副本的 Region
- `region check extra-peer`：多副本的 Region
- `region check down-peer`：有副本状态为 Down 的 Region
- `region check pending-peer`：有副本状态为 Pending 的 Region

11.6.3 调度策略控制

使用 pd-ctl 可以从以下三个方面来调整 PD 的调度策略。更具体的信息可以参考 PD Control。

11.6.3.1 启停调度器

pd-ctl 支持动态创建和删除 Scheduler，你可以通过这些操作来控制 PD 的调度行为，如：

- `scheduler show`：显示当前系统中的 Scheduler
- `scheduler remove balance-leader-scheduler`：删除（停用）balance region 调度器
- `scheduler add evict-leader-scheduler 1`：添加移除 Store 1 的所有 Leader 的调度器

11.6.3.2 手动添加 Operator

PD 支持直接通过 pd-ctl 来创建或删除 Operator，如：

- `operator add add-peer 2 5`：在 Store 5 上为 Region 2 添加 Peer
- `operator add transfer-leader 2 5`：将 Region 2 的 Leader 迁移至 Store 5
- `operator add split-region 2`：将 Region 2 拆分为 2 个大小相当的 Region
- `operator remove 2`：取消 Region 2 当前待执行的 Operator

11.6.3.3 调度参数调整

使用 pd-ctl 执行 `config show` 命令可以查看所有的调度参数，执行 `config set {key} {value}` 可以调整对应参数的值。常见调整如下：

- `leader-schedule-limit`：控制 Transfer Leader 调度的并发数
- `region-schedule-limit`：控制增删 Peer 调度的并发数
- `enable-replace-offline-replica`：开启节点下线的调度
- `enable-location-replacement`：开启调整 Region 隔离级别相关的调度
- `max-snapshot-count`：每个 Store 允许的最大收发 Snapshot 的并发数

11.6.4 典型场景分析与处理

该部分通过几个典型场景及其应对方式说明 PD 调度策略的最佳实践。

11.6.4.1 Leader/Region 分布不均衡

PD 的打分机制决定了一般情况下，不同 Store 的 Leader Count 和 Region Count 不能完全说明负载均衡状态，所以需要从 TiKV 的实际负载或者存储空间占用来判断是否有负载不均衡的状况。

确认 Leader/Region 分布不均衡后，首先观察不同 Store 的打分情况。

如果不同 Store 的打分是接近的，说明 PD 认为此时已经是均衡状态了，可能的原因有：

- 存在热点导致负载不均衡。可以参考热点分布不均匀中的解决办法进行分析处理。
- 存在大量空 Region 或小 Region，因此不同 Store 的 Leader 数量差别特别大，导致 Raftstore 负担过重。此时需要开启Region Merge 并尽可能加速合并。
- 不同 Store 的软硬件环境存在差异。可以酌情调整 `leader-weight` 和 `region-weight` 来控制 Leader/Region 的分布。
- 其他不明原因。仍可以通过调整 `leader-weight` 和 `region-weight` 来控制 Leader/Region 的分布。

如果不同 Store 的分数差异较大，需要进一步检查 Operator 的相关 Metrics，特别关注 Operator 的生成和执行情况，这时大体上又分两种情况：

- 生成的调度是正常的，但是调度的速度很慢。可能的原因有：
  - 调度速度受限于 limit 配置。PD 默认配置的 limit 比较保守，在不对正常业务造成显著影响的前提下，可以酌情将 `leader-schedule-limit` 或 `region-schedule-limit` 调大一些。此外，`max-pending-peer-count` 以及 `max-snapshot-count` 限制也可以放宽。
  - 系统中同时运行有其他的调度任务产生竞争，导致 balance 速度上不去。这种情况下如果 balance 调度的优先级更高，可以先停掉其他的调度或者限制其他调度的速度。例如 Region 没均衡的情况下做下线节点操作，下线的调度与 Region Balance 会抢占 `region-schedule-limit` 配额，此时你可以调小 `replica-schedule-limit` 以限制下线调度的速度，或者设置 `enable-replace-offline-replica = false` 来暂时关闭下线流程。
  - 调度执行得太慢。可以通过 Operator step duration 进行判断。通常不涉及到收发 Snapshot 的 Step（比如 `TransferLeader`，`RemovePeer`，`PromoteLearner` 等）的完成时间应该在毫秒级，涉及到 Snapshot 的 Step（如 `AddLearner`，`AddPeer` 等）的完成时间为数十秒。如果耗时明显过高，可能是 TiKV 压力过大或者网络等方面的瓶颈导致的，需要具体情况具体分析。

- 没能生成对应的 balance 调度。可能的原因有：
  - 调度器未被启用。比如对应的 Scheduler 被删除了，或者 limit 被设置为 “0”。
  - 由于其他约束无法进行调度。比如系统中有 `evict-leader-scheduler`，此时无法把 Leader 迁移至对应的 Store。再比如设置了 Label property，也会导致部分 Store 不接受 Leader。
  - 集群拓扑的限制导致无法均衡。比如 3 副本 3 数据中心的集群，由于副本隔离的限制，每个 Region 的 3 个副本都分别分布在不同的数据中心，假如这 3 个数据中心的 Store 数不一样，最后调度就会收敛在每个数据中心均衡，但是全局不均衡的状态。

11.6.4.2 节点下线速度慢

这个场景需要从 Operator 相关的 Metrics 入手，分析 Operator 的生成执行情况。

如果调度在正常生成，只是速度很慢，可能的原因有：

- 调度速度受限于 limit 配置。可以适当调大 `replica-schedule-limit`，`max-pending-peer-count` 以及 `max-snapshot-count` 限制也可以放宽。
- 系统中同时运行有其他的调度任务产生竞争。处理方法参考Leader/Region 分布不均衡。
- 下线单个节点时，由于待操作的 Region 有很大一部分（3 副本配置下约 1/3）的 Leader 都集中在下线的节点上，下线速度会受限于这个单点生成 Snapshot 的速度。你可以通过手动给该节点添加一个 `evict-leader-scheduler` 调度器迁走 Leader 来加速。

如果没有对应的 Operator 调度生成，可能的原因有：

- 下线调度被关闭，或者 `replica-schedule-limit` 被设为 “0”。
- 找不到节点来转移 Region。例如相同 Label 的替代节点可用容量都不足 20%，PD 为了避免爆盘的风险会停止调度。这种情况需要添加更多节点，或者删除一些数据释放空间。

11.6.4.3 节点上线速度慢

目前 PD 没有对节点上线特殊处理。节点上线实际上是依靠 balance region 机制来调度的，所以参考Leader/Region 分布不均衡中的排查步骤即可。

11.6.4.4 热点分布不均匀

热点调度的问题大体上可以分为以下几种情况：

- 从 PD 的 Metrics 能看出来有不少 hot Region，但是调度速度跟不上，不能及时地把热点 Region 分散开来。

解决方法：调大 `hot-region-schedule-limit` 并减少其他调度器的 limit 配额，从而加快热点调度的速度。还可调小 `hot-region-cache-hits-threshold` 使 PD 对更快响应流量的变化。

- 单一 Region 形成热点，比如大量请求频繁 scan 一个小表，这个可以从业务角度或者 Metrics 统计的热点信息看出来。由于单 Region 热点现阶段无法使用打散的手段来消除，需要确认热点 Region 后手动添加 `split-region` 调度将这样的 Region 拆开。

- 从 PD 的统计来看没有热点，但是从 TiKV 的相关 Metrics 可以看出部分节点负载明显高于其他节点，成为整个系统的瓶颈。这是因为目前 PD 统计热点 Region 的维度比较单一，仅针对流量进行分析，在某些场景下无法准确定位热点。例如部分 Region 有大量的点查请求，从流量上来看并不显著，但是过高的 QPS 导致关键模块达到瓶颈。

  解决方法：首先从业务层面确定形成热点的 table，然后添加 `scatter-range-scheduler` 调度器使这个 table 的所有 Region 均匀分布。TiDB 也在其 HTTP API 中提供了相关接口来简化这个操作，具体可以参考 TiDB HTTP API 文档。

11.6.4.5 Region Merge 速度慢

Region Merge 速度慢也很有可能是受到 limit 配置的限制（`merge-schedule-limit` 及 `region-schedule-limit`），或者是与其他调度器产生了竞争。具体来说，可有如下处理方式：

- 假如已经从相关 Metrics 得知系统中有大量的空 Region，这时可以通过把 `max-merge-region-size` 和 `max-merge-region-keys` 调整为较小值来加快 Merge 速度。这是因为 Merge 的过程涉及到副本迁移，所以 Merge 的 Region 越小，速度就越快。如果生成 Merge Operator 的速度很快，想进一步加快 Region Merge 过程，还可以把 `patrol-region-interval` 调整为“10ms”(从 v5.3.0 起，此配置项默认值为“10ms”)，这个能加快巡检 Region 的速度，但是会消耗更多的 CPU 资源。
- 创建过大量表后（包括执行 `Truncate Table` 操作）又清空了。此时如果开启了 split table 特性，这些空 Region 是无法合并的，此时需要调整以下参数关闭这个特性：
  - TiKV: 将 `split-region-on-table` 设为 `false`，该参数不支持动态修改。
  - PD: 使用 PD Control，根据集群情况选择性地设置以下参数。
    * 如果集群中不存在 TiDB 实例，将`key-type` 的值设置为 `raw` 或 `txn`。此时，无论 `enable-cross-table-merge` 设置如何，PD 均可以跨表合并 Region。该参数支持动态修改。

```
config set key-type txn
```

    * 如果集群中存在 TiDB 实例，将 `key-type` 的值设置为 `table`。此时将 `enable-cross-table-merge` 设置为 `true`，可以使 PD 跨表合并 Region。该参数支持动态修改。

```
config set enable-cross-table-merge true
```

  如果修改未生效，请参阅FAQ - 修改 TiKV/PD 的 toml 配置文件后没有生效。

> 注意：
> 在开启 `placement-rules` 后，请合理切换 `key-type`，避免无法正常解码 key。

- 对于 3.0.4 和 2.1.16 以前的版本，Region 中 Key 的个数 (`approximate_keys`) 在特定情况下（大部分发生在删表之后）统计不准确，造成 keys 的统计值很大，无法满足 `max-merge-region-keys` 的约束。你可以通过调大 `max-merge-region-keys` 来避免这个问题。

11.6.4.6 TiKV 节点故障处理策略

没有人工介入时，PD 处理 TiKV 节点故障的默认行为是，等待半小时之后（可通过 `max-store-down-time` 配置调整），将此节点设置为 Down 状态，并开始为涉及到的 Region 补充副本。

实践中，如果能确定这个节点的故障是不可恢复的，可以立即做下线处理，这样 PD 能尽快补齐副本，降低数据丢失的风险。与之相对，如果确定这个节点是能恢复的，但可能半小时之内来不及，则可以把 `max-store-down-time` 临时调整为比较大的值，这样能避免超时之后产生不必要的副本补充，造成资源浪费。

自 v5.2.0 起，TiKV 引入了慢节点检测机制。通过对 TiKV 中的请求进行采样，计算出一个范围在 1~100 的分数。当分数大于等于 80 时，该 TiKV 节点会被设置为 slow 状态。可以通过添加`evict-slow-store-scheduler` 来针对慢节点进行对应的检测和调度。当检测到有且只有一个 TiKV 节点为慢节点，并且该 TiKV 的 slow score 到达上限（默认 100）时，将节点上的 leader 驱逐（其作用类似于 `evict-leader-scheduler`）。

11.7 海量 Region 集群调优最佳实践

在 TiDB 的架构中，所有数据以一定 key range 被切分成若干 Region 分布在多个 TiKV 实例上。随着数据的写入，一个集群中会产生上百万个甚至千万个 Region。单个 TiKV 实例上产生过多的 Region 会给集群带来较大的负担，影响整个集群的性能表现。

本文将介绍 TiKV 核心模块 Raftstore 的工作流程，海量 Region 导致性能问题的原因，以及优化性能的方法。

11.7.1 Raftstore 的工作流程

一个 TiKV 实例上有多个 Region。Region 消息是通过 Raftstore 模块驱动 Raft 状态机来处理的。这些消息包括 Region 上读写请求的处理、Raft log 的持久化和复制、Raft 的心跳处理等。但是，Region 数量增多会影响整个集群的性能。为了解释这一点，需要先了解 TiKV 的核心模块 Raftstore 的工作流程。

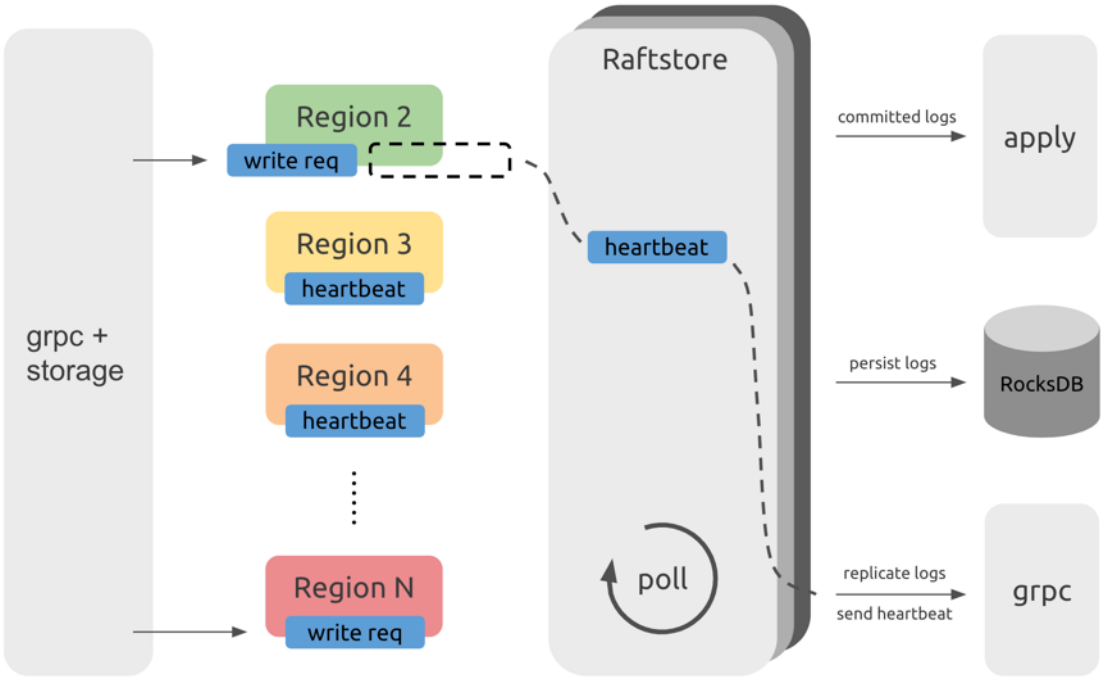


图 247: 图 1 Raftstore 处理流程示意图

> 注意：
>
> 该图仅为示意，不代表代码层面的实际结构。

上图是 Raftstore 处理流程的示意图。如图所示，从 TiDB 发来的请求会通过 gRPC 和 storage 模块变成最终的 KV 读写消息，并被发往相应的 Region，而这些消息并不会被立即处理而是被暂存下来。Raftstore 会轮询检查每个 Region 是否有需要处理的消息。如果 Region 有需要处理的消息，那么 Raftstore 会驱动 Raft 状态机去处理这些消息，并根据这些消息所产生的状态变更去进行后续操作。例如，在有写请求时，Raft 状态机需要将日志落盘并且将日志发送给其他 Region 副本；在达到心跳间隔时，Raft 状态机需要将心跳信息发送给其他 Region 副本。

11.7.2 性能问题

从 Raftstore 处理流程示意图可以看出，需要依次处理各个 Region 的消息。那么在 Region 数量较多的情况下，Raftstore 需要花费一些时间去处理大量 Region 的心跳，从而带来一些延迟，导致某些读写请求得不到及时处理。如果读写压力较大，Raftstore 线程的 CPU 使用率容易达到瓶颈，导致延迟进一步增加，进而影响性能表现。

通常在有负载的情况下，如果 Raftstore 的 CPU 使用率达到了 85% 以上，即可视为达到繁忙状态且成为了瓶颈，同时 `propose wait duration` 可能会高达百毫秒级别。

> 注意：
>
> - Raftstore 的 CPU 使用率是指单线程的情况。如果是多线程 Raftstore，可等比例放大使用率。
> - 由于 Raftstore 线程中有 I/O 操作，所以 CPU 使用率不可能达到 100%。

11.7.2.1 性能监控

可在 Grafana 的 TiKV 面板下查看相关的监控 metrics：

- Thread-CPU 下的 `Raft store CPU`

  参考值：低于 `raftstore.store-pool-size * 85%`。

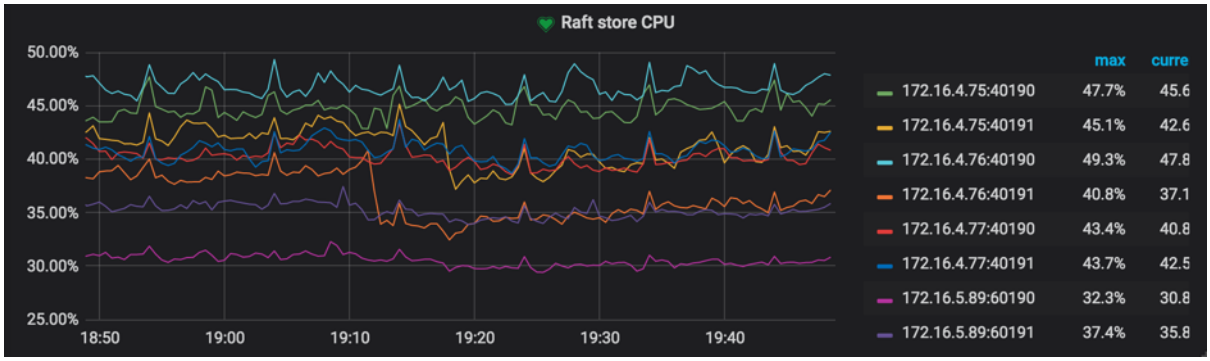


图 248: 图 2 查看 Raftstore CPU

- Raft Propose 下的 `Propose wait duration`

  `Propose wait duration` 是从发送请求给 Raftstore，到 Raftstore 真正开始处理请求之间的延迟时间。如果该延迟时间较长，说明 Raftstore 比较繁忙或者处理 append log 比较耗时导致 Raftstore 不能及时处理请求。

  参考值：低于 50-100ms。

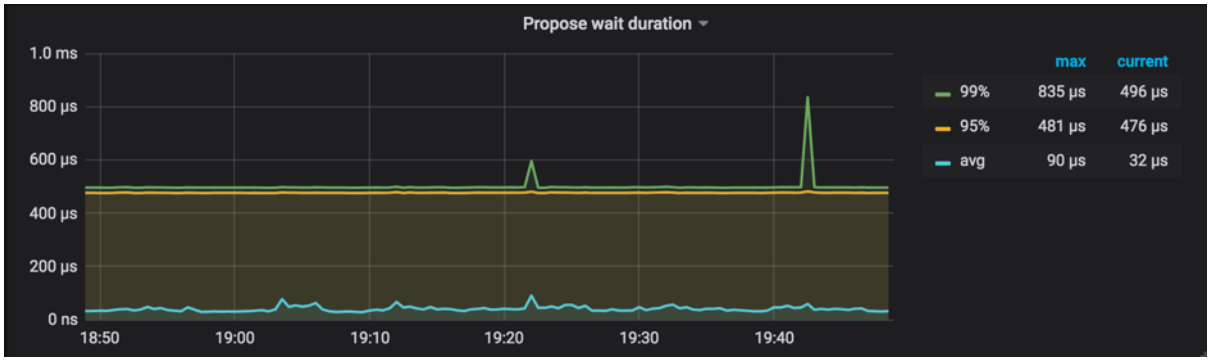


图 249: 图 3 查看 Propose wait duration

11.7.3 性能优化方法

找到性能问题的根源后，可从以下两个方向来解决性能问题：

- 减少单个 TiKV 实例的 Region 数
- 减少单个 Region 的消息数

11.7.3.1 方法一：增加 TiKV 实例

如果 I/O 资源和 CPU 资源都比较充足，可在单台机器上部署多个 TiKV 实例，以减少单个 TiKV 实例上的 Region 个数；或者增加 TiKV 集群的机器数。

11.7.3.2 方法二：调整 `raft-base-tick-interval`

除了减少 Region 个数外，还可以通过减少 Region 单位时间内的消息数量来减小 Raftstore 的压力。例如，在 TiKV 配置中适当调大 `raft-base-tick-interval`：

```
[raftstore]
raft-base-tick-interval = "2s"
```

`raft-base-tick-interval` 是 Raftstore 驱动每个 Region 的 Raft 状态机的时间间隔，也就是每隔该时长就需要向 Raft 状态机发送一个 tick 消息。增加该时间间隔，可以有效减少 Raftstore 的消息数量。

需要注意的是，该 tick 消息的间隔也决定了 `election timeout` 和 `heartbeat` 的间隔。示例如下：

```
raft-election-timeout = raft-base-tick-interval * raft-election-timeout-ticks
raft-heartbeat-interval = raft-base-tick-interval * raft-heartbeat-ticks
```

如果 Region Follower 在 `raft-election-timeout` 间隔内未收到来自 Leader 的心跳，就会判断 Leader 出现故障而发起新的选举。`raft-heartbeat-interval` 是 Leader 向 Follower 发送心跳的间隔，因此调大 `raft-base-tick-interval` 可以减少单位时间内 Raft 发送的网络消息，但也会让 Raft 检测到 Leader 故障的时间更长。

11.7.3.3 方法三：提高 Raftstore 并发数

从 v3.0 版本起，Raftstore 已经扩展为多线程，极大降低了 Raftstore 线程成为瓶颈的可能性。

TiKV 默认将 `raftstore.store-pool-size` 配置为 2。如果 Raftstore 出现瓶颈，可以根据实际情况适当调高该参数值，但不建议设置过高以免引入不必要的线程切换开销。

11.7.3.4 方法四：开启 Hibernate Region 功能

在实际情况中，读写请求并不会均匀分布到每个 Region 上，而是集中在少数的 Region 上。那么可以尽量减少暂时空闲的 Region 的消息数量，这也就是 Hibernate Region 的功能。无必要时可不进行 `raft-base-tick`，即不驱动空闲 Region 的 Raft 状态机，那么就不会触发这些 Region 的 Raft 产生心跳信息，极大地减小了 Raftstore 的工作负担。

Hibernate Region 在 TiKV master 分支上默认开启。可根据实际情况和需求来配置此功能的开启和关闭，请参阅配置 Hibernate Region。

11.7.3.5 方法五：开启 Region Merge

> 注意：
>
> 从 TiDB v3.0 开始，`Region Merge` 默认开启。

开启 `Region Merge` 也能减少 Region 的个数。与 `Region Split` 相反，`Region Merge` 是通过调度把相邻的小 Region 合并的过程。在集群中删除数据或者执行 `Drop Table/Truncate Table` 语句后，可以将小 Region 甚至空 Region 进行合并以减少资源的消耗。

通过 pd-ctl 设置以下参数即可开启 `Region Merge`：

```
>> pd-ctl config set max-merge-region-size 20
>> pd-ctl config set max-merge-region-keys 200000
>> pd-ctl config set merge-schedule-limit 8
```

详情请参考如何配置 Region Merge（英文）和 PD 配置文件描述。

同时，默认配置的 `Region Merge` 的参数设置较为保守，可以根据需求参考PD 调度策略最佳实践中提供的方法加快 `Region Merge` 过程的速度。

11.7.4 其他问题和解决方案

11.7.4.1 切换 PD Leader 的速度慢

PD 需要将 Region Meta 信息持久化在 etcd 上，以保证切换 PD Leader 节点后 PD 能快速继续提供 Region 路由服务。随着 Region 数量的增加，etcd 出现性能问题，使得 PD 在切换 Leader 时从 etcd 获取 Region Meta 信息的速度较慢。在百万 Region 量级时，从 etcd 获取信息的时间可能需要十几秒甚至几十秒。

因此从 v3.0 版本起，PD 默认开启配置项 `use-region-storage`，将 Region Meta 信息存在本地的 LevelDB 中，并通过其他机制同步 PD 节点间的信息。

11.7.4.2 PD 路由信息更新不及时

在 TiKV 中，pd-worker 模块将 Region Meta 信息定期上报给 PD，在 TiKV 重启或者切换 Region Leader 时需要通过统计信息重新计算 Region 的 `approximate size/keys`。因此在 Region 数量较多的情况下，pd-worker 单线程可能成为瓶颈，造成任务得不到及时处理而堆积起来。因此 PD 不能及时获取某些 Region Meta 信息以致路由信息更新不及时。该问题不会影响实际的读写，但可能导致 PD 调度不准确以及 TiDB 更新 Region cache 时需要多几次 round-trip。

可在 TiKV Grafana 面板中查看 Task 下的 Worker pending tasks 来确定 pd-worker 是否有任务堆积。通常来说，pending tasks 应该维持在一个比较低的值。

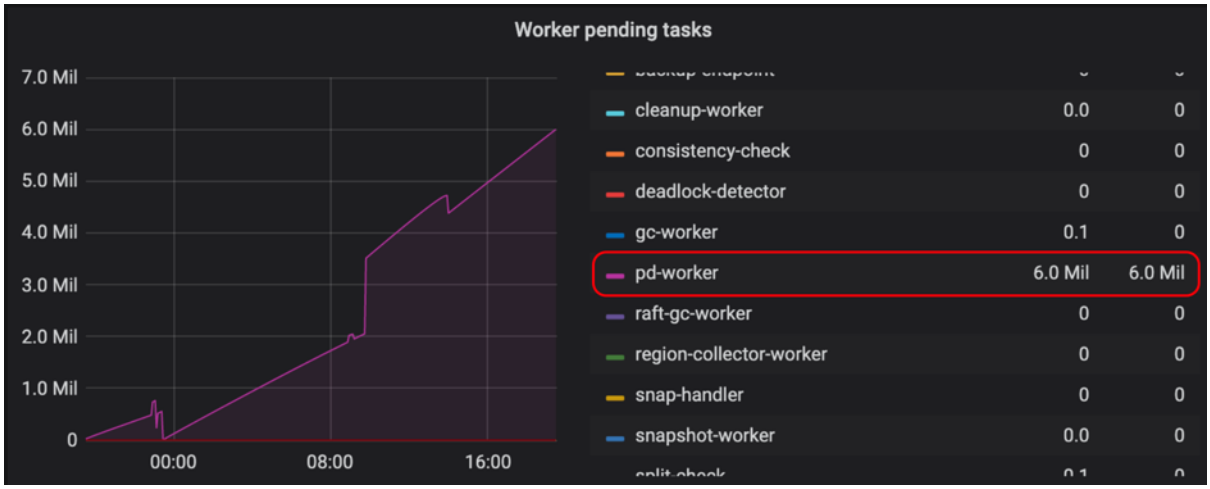


图 250: 图 4 查看 pd-worker

老版本 TiDB (< v3.0.5) pd-worker 的效率有一些缺陷，如果碰到类似问题，建议升级至最新版本。

11.7.4.3 Prometheus 查询 metrics 的速度慢

在大规模集群中，随着 TiKV 实例数的增加，Prometheus 查询 metrics 时的计算压力较大，导致 Grafana 查看 metrics 的速度较慢。从 v3.0 版本起设置了一些 metrics 的预计算，让这个问题有所缓解。

11.8 通过拓扑 label 进行副本调度

> **注意：**
>
> TiDB 在 v5.3.0 中引入了实验特性 Placement Rules in SQL。使用该功能，你可以更方便地配置表和分区的位置。在未来版本中，Placement Rules in SQL 可能取代通过 PD 配置放置规则的功能。

为了提升 TiDB 集群的高可用性和数据容灾能力，我们推荐让 TiKV 节点尽可能在物理层面上分散，例如让 TiKV 节点分布在不同的机架甚至不同的机房。PD 调度器根据 TiKV 的拓扑信息，会自动在后台通过调度使得 Region 的各个副本尽可能隔离，从而使得数据容灾能力最大化。

要让这个机制生效，需要在部署时进行合理配置，把集群的拓扑信息（特别是 TiKV 的位置）上报给 PD。阅读本章前，请先确保阅读TiUP 部署方案。

11.8.1 根据集群拓扑配置 labels

11.8.1.1 设置 TiKV 的 labels 配置

TiKV 支持在命令行参数或者配置文件中以键值对的形式绑定一些属性，我们把这些属性叫做标签 (label)。TiKV 在启动后，会将自身的标签上报给 PD，因此我们可以使用标签来标识 TiKV 节点的地理位置。

比如集群的拓扑结构分成三层：机房 (zone) -> 机架 (rack) -> 主机 (host)，就可以使用这 3 个标签来设置 TiKV 的位置。

使用命令行参数的方式：

```
tikv-server --labels zone=<zone>,rack=<rack>,host=<host>
```

使用配置文件的方式：

```
[server]
labels = "zone=<zone>,rack=<rack>,host=<host>"
```

11.8.1.2 设置 PD 的 location-labels 配置

根据前面的描述，标签可以是用来描述 TiKV 属性的任意键值对，但 PD 无从得知哪些标签是用来标识地理位置的，而且也无从得知这些标签的层次关系。因此，PD 也需要一些配置来使得 PD 理解 TiKV 节点拓扑。

PD 上的配置叫做 location-labels，在集群初始化之前，可以通过 PD 的配置文件进行配置。

```
[replication]
location-labels = ["zone", "rack", "host"]
```

如果需要在 PD 集群初始化完成后进行配置，则需要使用 pd-ctl 工具进行在线更改：

```
pd-ctl config set location-labels zone,rack,host
```

其中，location-labels 配置是一个字符串数组，每一项与 TiKV 的 labels 的 key 是对应的，且其中每个 key 的顺序代表了不同标签的层次关系。

> 注意：
>
> 必须同时配置 PD 的 location-labels 和 TiKV 的 labels 参数，否则 PD 不会根据拓扑结构进行调度。

11.8.1.3 设置 PD 的 isolation-level 配置

在配置了 location-labels 的前提下，用户可以还通过 isolation-level 配置来进一步加强对 TiKV 集群的拓扑隔离要求。假设按照上面的说明通过 location-labels 将集群的拓扑结构分成三层：机房 (zone) -> 机架 (rack) -> 主机 (host)，并对 isolation-level 作如下配置

```
[replication]
isolation-level = "zone"
```

当 PD 集群初始化完成后，需要使用 pd-ctl 工具进行在线更改：

```
pd-ctl config set isolation-level zone
```

其中，isolation-level 配置是一个字符串，需要与 location-labels 的其中一个 key 对应。该参数限制 TiKV 拓扑集群的最小且强制隔离级别要求。

> 注意：
>
> isolation-level 默认情况下为空，即不进行强制隔离级别限制，若要对其进行设置，必须先配置 PD 的 location-labels 参数，同时保证 isolation-level 的值一定为 location-labels 中的一个。

11.8.1.4 使用 TiUP 进行配置（推荐）

如果使用 TiUP 部署集群，可以在初始化配置文件中统一进行 location 相关配置。TiUP 会负责在部署时生成对应的 TiKV 和 PD 配置文件。

下面的例子定义了 zone/host 两层拓扑结构。集群的 TiKV 分布在三个 zone，每个 zone 内有两台主机，其中 z1 每台主机部署两个 TiKV 实例，z2 和 z3 每台主机部署 1 个实例。以下例子中 tikv-n 代表第 n 个 TiKV 节点的 IP 地址。

```
server_configs:
  pd:
    replication.location-labels: ["zone", "host"]

tikv_servers:
## z1
  - host: tikv-1
    config:
      server.labels:
        zone: z1
        host: h1
   - host: tikv-2
    config:
      server.labels:
        zone: z1
```

```
        host: h1
  - host: tikv-3
    config:
      server.labels:
        zone: z1
        host: h2
  - host: tikv-4
    config:
      server.labels:
        zone: z1
        host: h2
## z2
  - host: tikv-5
    config:
      server.labels:
        zone: z2
        host: h1
   - host: tikv-6
    config:
      server.labels:
        zone: z2
        host: h2
## z3
  - host: tikv-7
    config:
      server.labels:
        zone: z3
        host: h1
  - host: tikv-8
    config:
      server.labels:
        zone: z3
        host: h2
```

详情参阅TiUP 跨数据中心部署拓扑。

11.8.2 基于拓扑 label 的 PD 调度策略

PD 在副本调度时，会按照 label 层级，保证同一份数据的不同副本尽可能分散。

下面以上一节的拓扑结构为例分析。

假设集群副本数设置为 3 (max-replicas=3)，因为总共有 3 个 zone，PD 会保证每个 Region 的 3 个副本分别放置在 z1/z2/z3，这样当任何一个数据中心发生故障时，TiDB 集群依然是可用的。

假如集群副本数设置为 5 (max-replicas=5)，因为总共只有 3 个 zone，在这一层级 PD 无法保证各个副本的隔离，此时 PD 调度器会退而求其次，保证在 host 这一层的隔离。也就是说，会出现一个 Region 的多个副本分布

在同一个 zone 的情况，但是不会出现多个副本分布在同一台主机。

在 5 副本配置的前提下，如果 z3 出现了整体故障或隔离，并且 z3 在一段时间后仍然不能恢复（由 max-store -down-time 控制），PD 会通过调度补齐 5 副本，此时可用的主机只有 4 个了，故而无法保证 host 级别的隔离，于是可能出现多个副本被调度到同一台主机的情况。

但假如 isolation-level 设置不为空，值为 zone，这样就规定了 Region 副本在物理层面上的最低隔离要求，也就是说 PD 一定会保证同一 Region 的副本分散于不同的 zone 之上。即便遵循此隔离限制会无法满足 max-replicas 的多副本要求，PD 也不会进行相应的调度。例如，当前存在 TiKV 集群的三个机房 z1/z2/z3，在三副本的设置下，PD 会将同一 Region 的三个副本分别分散调度至这三个机房。若此时 z1 整个机房发生了停电事故并在一段时间后仍然不能恢复，PD 会认为 z1 上的 Region 副本不再可用。但由于 isolation-level 设置为了 zone，PD 需要严格保证不同的 Region 副本不会落到同一 zone 上。此时的 z2 和 z3 均已存在副本，则 PD 在 isolation-level 的最小强制隔离级别限制下便不会进行任何调度，即使此时仅存在两个副本。

类似地，isolation-level 为 rack 时，最小隔离级别便为同一机房的不同 rack。在此设置下，如果能在 zone 级别保证隔离，会首先保证 zone 级别的隔离。只有在 zone 级别隔离无法完成时，才会考虑避免出现在同一 zone 同一 rack 的调度，并以此类推。

总的来说，PD 能够根据当前的拓扑结构使得集群容灾能力最大化。所以如果用户希望达到某个级别的容灾能力，就需要根据拓扑结构在对应级别提供多于副本数 (max-replicas) 的机器。同时 TiDB 也提供了诸如 isolation-level 这样的强制隔离级别设置，以便更灵活地根据场景来控制对数据的拓扑隔离级别。

11.9 高可用容灾方案

11.9.1 同城多数据中心部署 TiDB

作为 NewSQL 数据库，TiDB 兼顾了传统关系型数据库的优秀特性、NoSQL 数据库可扩展性以及跨数据中心场景下的高可用。本文档旨在介绍同城多数据中心部署 TiDB 方案。

11.9.1.1 了解 Raft 协议

Raft 是一种分布式一致性算法，在 TiDB 集群的多种组件中，PD 和 TiKV 都通过 Raft 实现了数据的容灾。Raft 的灾难恢复能力通过如下机制实现：

- Raft 成员的本质是日志复制和状态机。Raft 成员之间通过复制日志来实现数据同步；Raft 成员在不同条件下切换自己的成员状态，其目标是选出 leader 以提供对外服务。
- Raft 是一个表决系统，它遵循多数派协议，在一个 Raft Group 中，某成员获得大多数投票，它的成员状态就会转变为 leader。也就是说，当一个 Raft Group 还保有大多数节点 (majority) 时，它就能够选出 leader 以提供对外服务。

遵循 Raft 可靠性的特点，放到现实场景中：

- 想克服任意 1 台服务器 (host) 的故障，应至少提供 3 台服务器。
- 想克服任意 1 个机柜 (rack) 的故障，应至少提供 3 个机柜。
- 想克服任意 1 个数据中心（dc，又称机房）的故障，应至少提供 3 个数据中心。
- 想应对任意 1 个城市的灾难场景，应至少规划 3 个城市用于部署。

可见，原生 Raft 协议对于偶数副本的支持并不是很友好，考虑跨城网络延迟影响，或许同城三数据中心是最适合部署 Raft 的高可用及容灾方案。

11.9.1.2 同城三数据中心方案

同城三数据中心方案，即同城存有三个机房部署 TiDB 集群，同城三数据中心间的数据同步通过集群自身内部（Raft 协议）完成。同城三数据中心可同时对外进行读写服务，任意中心发生故障不影响数据一致性。

11.9.1.2.1 简易架构图

集群 TiDB、TiKV 和 PD 组件分别分布在 3 个不同的数据中心，这是最常规且高可用性最高的方案。

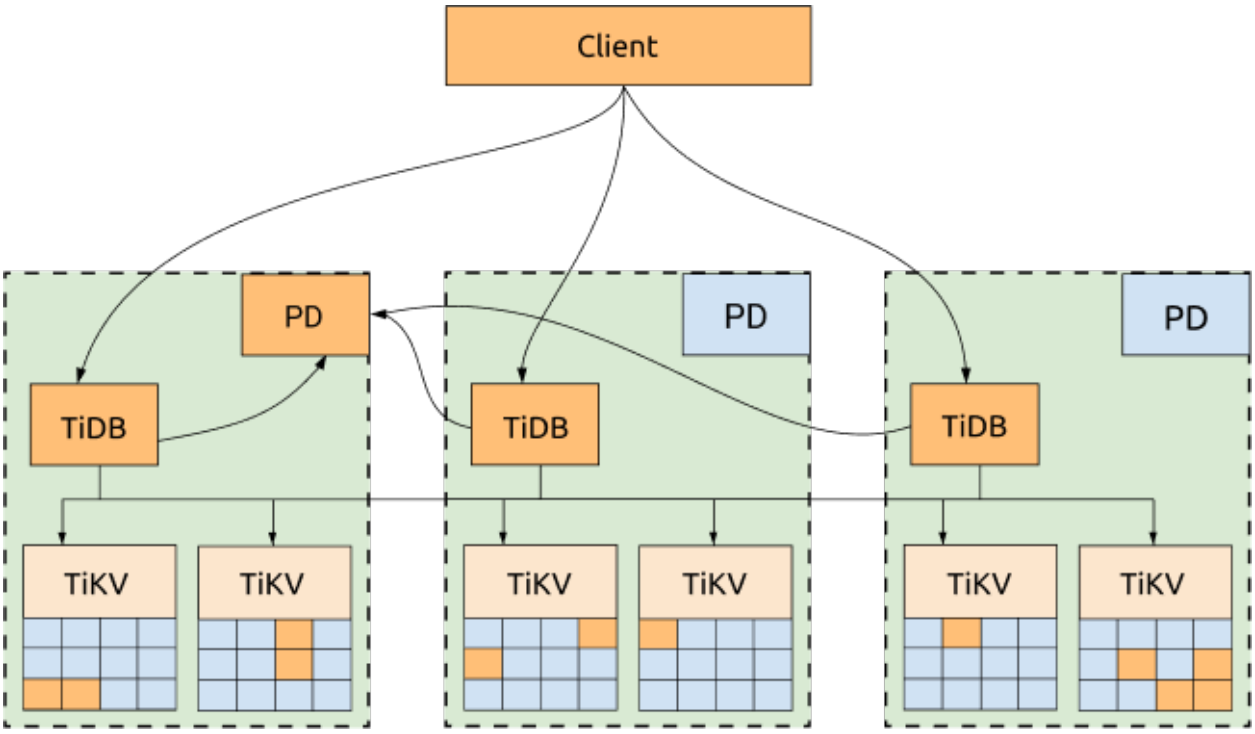


图 251: 三中心部署

优点：

- 所有数据的副本分布在三个数据中心，具备高可用和容灾能力
- 任何一个数据中心失效后，不会产生任何数据丢失 (RPO = 0)
- 任何一个数据中心失效后，其他两个数据中心会自动发起 leader election，并在合理长的时间内（通常情况 20s 以内）自动恢复服务

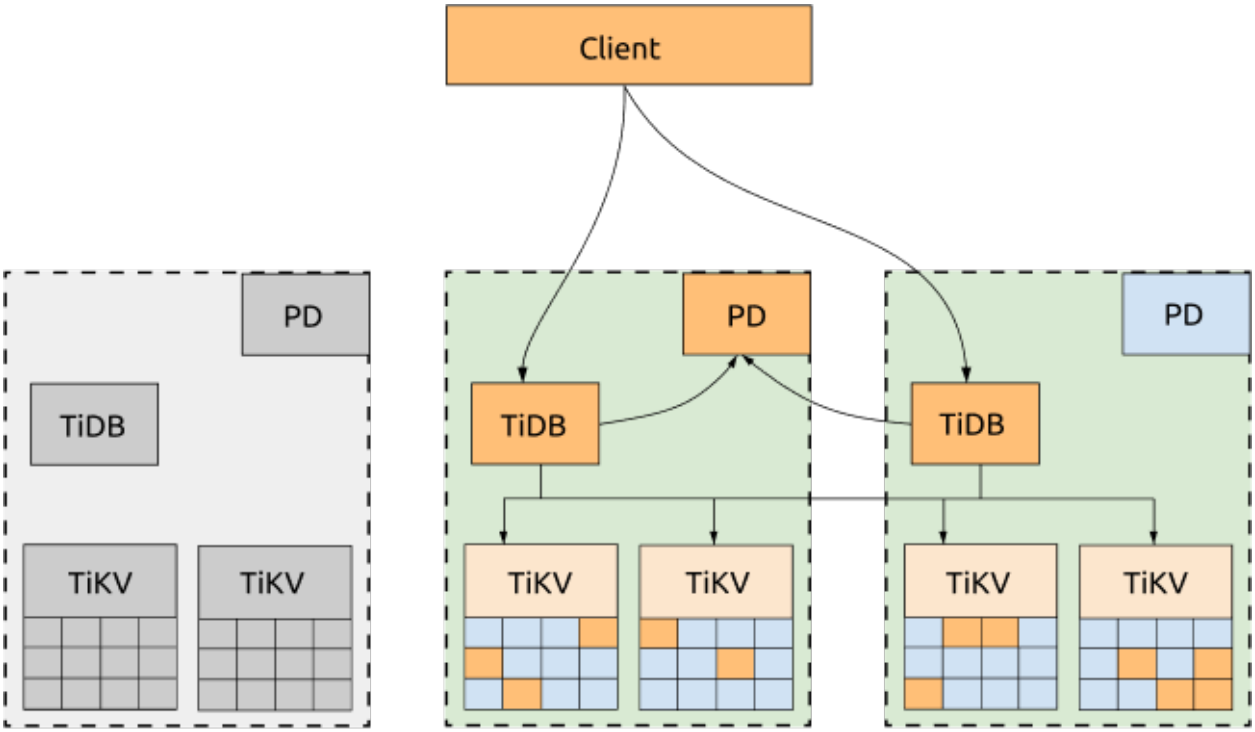


图 252: 三中心部署容灾

缺点：

性能受网络延迟影响。具体影响如下：

- 对于写入的场景，所有写入的数据需要同步复制到至少 2 个数据中心，由于 TiDB 写入过程使用两阶段提交，故写入延迟至少需要 2 倍数据中心间的延迟。
- 对于读请求来说，如果数据 leader 与发起读取的 TiDB 节点不在同一个数据中心，也会受网络延迟影响。
- TiDB 中的每个事务都需要向 PD leader 获取 TSO，当 TiDB 与 PD leader 不在同一个数据中心时，它上面运行的事务也会因此受网络延迟影响，每个有写入的事务会获取两次 TSO。

11.9.1.2.2 架构优化图

如果不需要每个数据中心同时对外提供服务，可以将业务流量全部派发到一个数据中心，并通过调度策略把 Region leader 和 PD leader 都迁移到同一个数据中心。这样一来，不管是从 PD 获取 TSO，还是读取 Region，都不会受数据中心间网络的影响。当该数据中心失效后，PD leader 和 Region leader 会自动在其它数据中心选出，只需要把业务流量转移至其他存活的数据中心即可。

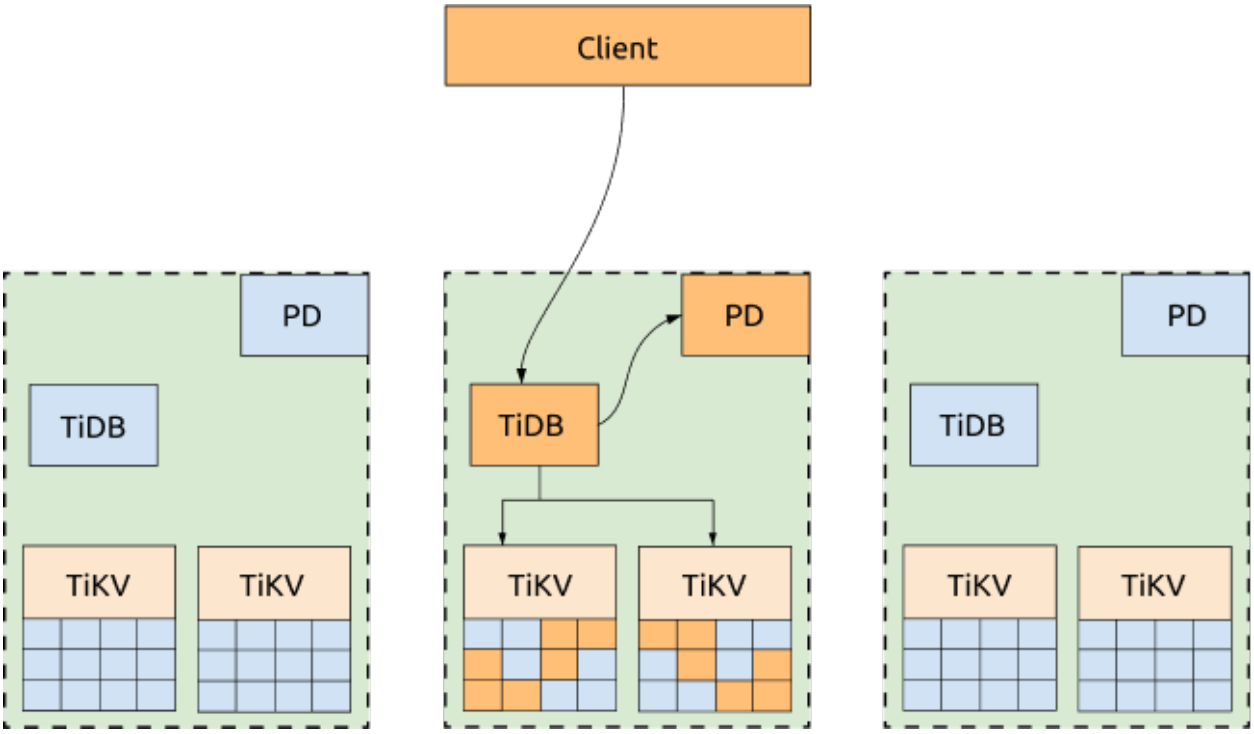


图 253: 三中心部署读性能优化

优点：

集群 TSO 获取能力以及读取性能有所提升。具体调度策略设置模板参照如下：

```
-- 其他机房统一驱逐 leader 到业务流量机房
config set label-property reject-leader LabelName labelValue

-- 迁移 PD leader 并设置优先级
member leader transfer pdName1
member leader_priority pdName1 5
member leader_priority pdName2 4
member leader_priority pdName3 3
```

> 注意：
>
> TiDB 5.2 及以上版本默认不支持 `label-property` 配置。若要设置副本策略，请使用Placement Rules。

缺点：

- 写入场景仍受数据中心网络延迟影响，这是因为遵循 Raft 多数派协议，所有写入的数据需要同步复制到至少 2 个数据中心

- TiDB Server 数据中心级别单点
- 业务流量纯走单数据中心，性能受限于单数据中心网络带宽压力
- TSO 获取能力以及读取性能受限于业务流量数据中心集群 PD、TiKV 组件是否正常，否则仍受跨数据中心网络交互影响

11.9.1.2.3 样例部署图

样例拓扑架构

下面假设某城存有 IDC1、IDC2、IDC3 三机房，机房 IDC 中存有两套机架，每个机架存有三台服务器，不考虑混布以及单台机器多实例部署下，同城三数据中心架构集群（3 副本）部署参考如下：

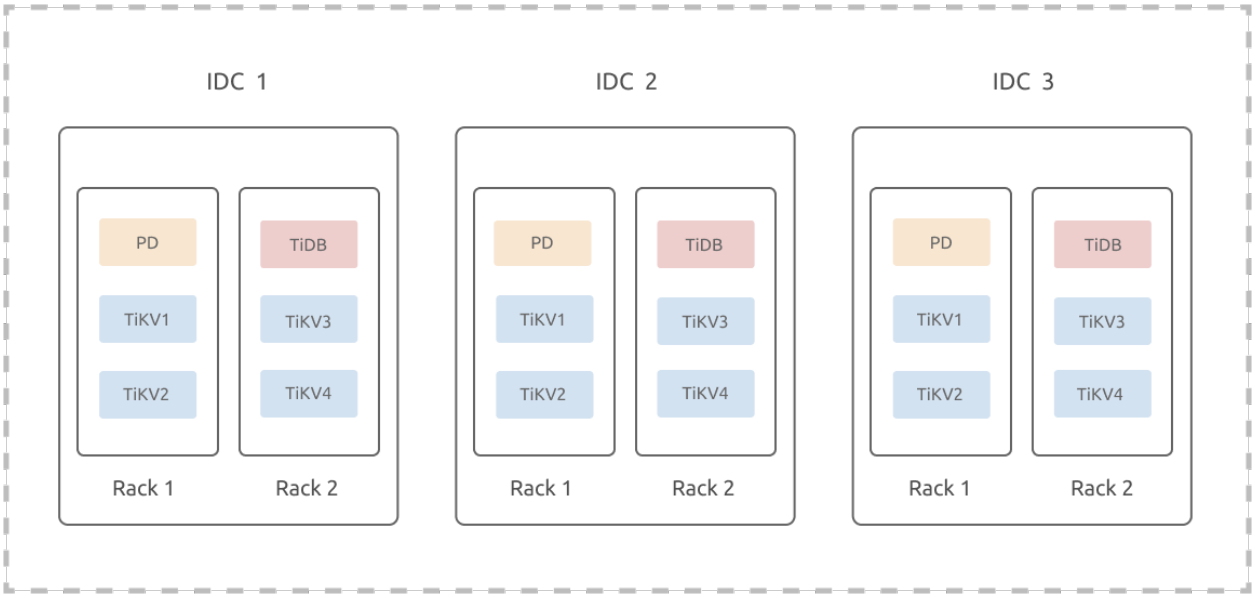


图 254: 同城三数据中心集群部署

TiKV Labels 简介

TiKV 是一个 Multi-Raft 系统，其数据按 Region（默认 96M）切分，每个 Region 的 3 个副本构成了一个 Raft Group。假设一个 3 副本 TiDB 集群，由于 Region 的副本数与 TiKV 实例数量无关，则一个 Region 的 3 个副本只会被调度到其中 3 个 TiKV 实例上，也就是说即使集群扩容 N 个 TiKV 实例，其本质仍是一个 3 副本集群。

由于 3 副本的 Raft Group 只能容忍 1 副本故障，当集群被扩容到 N 个 TiKV 实例时，这个集群依然只能容忍一个 TiKV 实例的故障。2 个 TiKV 实例的故障可能会导致某些 Region 丢失多个副本，整个集群的数据也不再完整，访问到这些 Region 上的数据的 SQL 请求将会失败。而 N 个 TiKV 实例中同时有两个发生故障的概率是远远高于 3 个 TiKV 中同时有两个发生故障的概率的，也就是说 Multi-Raft 系统集群扩容 TiKV 实例越多，其可用性是逐渐降低的。

正因为 Multi-Raft TiKV 系统局限性，Labels 标签应运而出，其主要用于描述 TiKV 的位置信息。Label 信息随着部署或滚动更新操作刷新到 TiKV 的启动配置文件中，启动后的 TiKV 会将自己最新的 Label 信息上报给 PD，PD 根据用户登记的 Label 名称（也就是 Label 元信息），结合 TiKV 的拓扑进行 Region 副本的最优调度，从而提高系统可用性。

TiKV Labels 样例规划

针对 TiKV Labels 标签，你需要根据已有的物理资源、容灾能力容忍度等方面进行设计与规划，进而提升系统的可用性和容灾能力。并根据已规划的拓扑架构，在集群初始化配置文件中进行配置（此处省略其他非重点项）：

```
server_configs:
  pd:
    replication.location-labels: ["zone","dc","rack","host"]

tikv_servers:
  - host: 10.63.10.30
    config:
      server.labels: { zone: "z1", dc: "d1", rack: "r1", host: "30" }
  - host: 10.63.10.31
    config:
      server.labels: { zone: "z1", dc: "d1", rack: "r1", host: "31" }
  - host: 10.63.10.32
    config:
      server.labels: { zone: "z1", dc: "d1", rack: "r2", host: "32" }
  - host: 10.63.10.33
    config:
      server.labels: { zone: "z1", dc: "d1", rack: "r2", host: "33" }

  - host: 10.63.10.34
    config:
      server.labels: { zone: "z2", dc: "d1", rack: "r1", host: "34" }
  - host: 10.63.10.35
    config:
      server.labels: { zone: "z2", dc: "d1", rack: "r1", host: "35" }
  - host: 10.63.10.36
    config:
      server.labels: { zone: "z2", dc: "d1", rack: "r2", host: "36" }
  - host: 10.63.10.37
    config:
      server.labels: { zone: "z2", dc: "d1", rack: "r2", host: "37" }

  - host: 10.63.10.38
    config:
      server.labels: { zone: "z3", dc: "d1", rack: "r1", host: "38" }
  - host: 10.63.10.39
    config:
      server.labels: { zone: "z3", dc: "d1", rack: "r1", host: "39" }
  - host: 10.63.10.40
    config:
      server.labels: { zone: "z3", dc: "d1", rack: "r2", host: "40" }
  - host: 10.63.10.41
    config:
```

```
server.labels: { zone: "z3", dc: "d1", rack: "r2", host: "41" }
```

本例中，zone 表示逻辑可用区层级，用于控制副本的隔离（当前集群 3 副本）。

不直接采用 dc，rack，host 三层 Label 结构的原因是考虑到将来可能发生 dc（数据中心）的扩容，假设新扩容的 dc 编号是 d2，d3，d4，则只需在对应可用区下扩容 dc；rack 扩容只需在对应 dc 下扩容即可。

如果直接采用 dc，rack，host 三层 Label 结构，那么扩容 dc 操作可能需重打 Label，TiKV 数据整体需要 Rebalance。

11.9.1.2.4 高可用和容灾分析

同城多数据中心方案提供的保障是，任意一个数据中心故障时，集群能自动恢复服务，不需要人工介入，并能保证数据一致性。注意，各种调度策略都是用于帮助性能优化的，当发生故障时，调度机制总是第一优先考虑可用性而不是性能。

11.9.2 两地三中心部署

本文档简要介绍两地三中心部署的架构模型及配置。

11.9.2.1 简介

两地三中心架构，即生产数据中心、同城灾备中心、异地灾备中心的高可用容灾方案。在这种模式下，两个城市的三个数据中心互联互通，如果一个数据中心发生故障或灾难，其他数据中心可以正常运行并对关键业务或全部业务实现接管。相比同城多中心方案，两地三中心具有跨城级高可用能力，可以应对城市级自然灾害。

TiDB 分布式数据库通过 Raft 算法原生支持两地三中心架构的建设，并保证数据库集群数据的一致性和高可用性。而且因同城数据中心网络延迟相对较小，可以把业务流量同时派发到同城两个数据中心，并通过控制 Region Leader 和 PD Leader 分布实现同城数据中心共同负载业务流量的设计。

11.9.2.2 架构

本文以北京和西安为例，阐述 TiDB 分布式数据库两地三中心架构的部署模型。

本例中，北京有两个机房 IDC1 和 IDC2，异地西安一个机房 IDC3。北京同城两机房之间网络延迟低于 3 ms，北京与西安之间的网络使用 ISP 专线，延迟约 20 ms。

下图为集群部署架构图，具体如下：

- 集群采用两地三中心部署方式，分别为北京 IDC1，北京 IDC2，西安 IDC3；
- 集群采用 5 副本模式，其中 IDC1 和 IDC2 分别放 2 个副本，IDC3 放 1 个副本；TiKV 按机柜打 Label，既每个机柜上有一份副本。
- 副本间通过 Raft 协议保证数据的一致性和高可用，对用户完全透明。

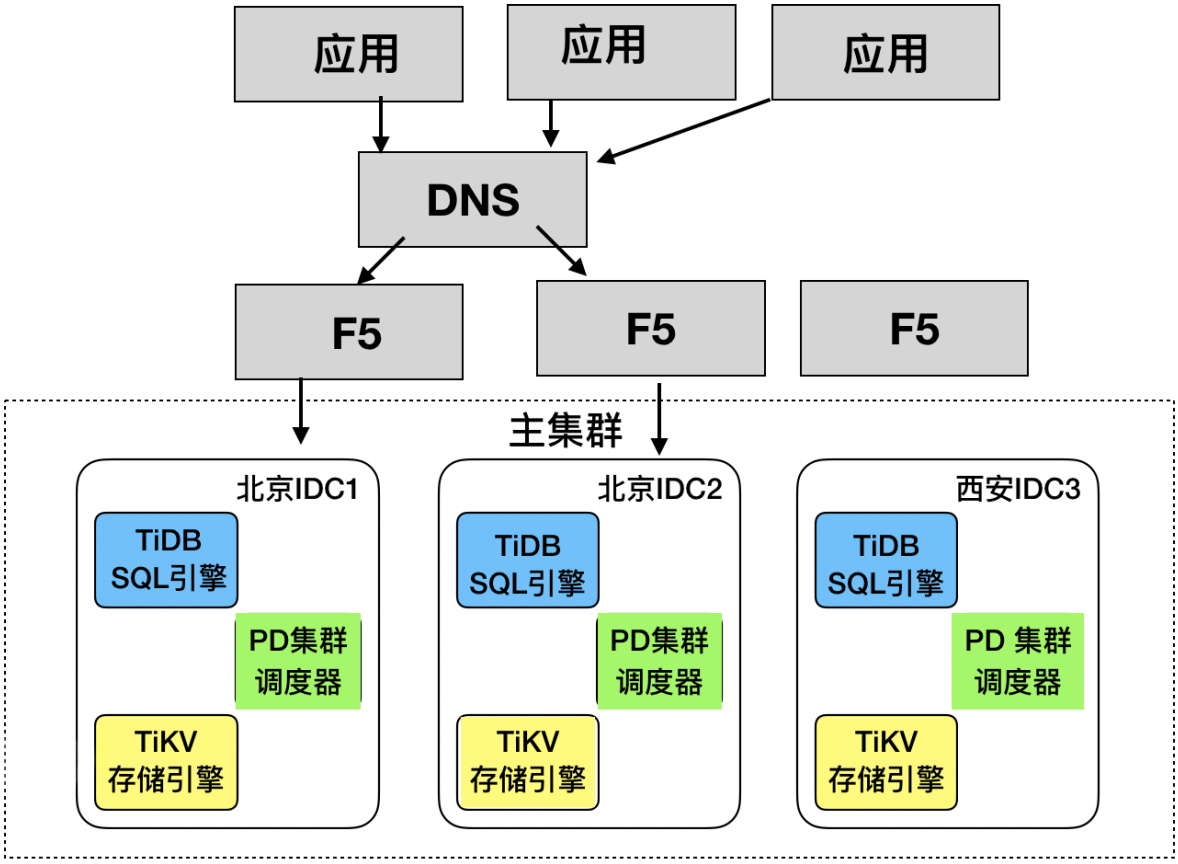


图 255: 两地三中心集群架构图

该架构具备高可用能力，同时通过 PD 调度限制了 Region Leader 尽量只出现在同城的两个数据中心，这相比于三数据中心，即 Region Leader 分布不受限制的方案有以下优缺点：

- 优点
  - Region Leader 都在同城低延迟机房，数据写入速度更优。
  - 两中心可同时对外提供服务，资源利用率更高。
  - 可保证任一数据中心失效后，服务可用并且不发生数据丢失。
- 缺点
  - 因为数据一致性是基于 Raft 算法实现，当同城两个数据中心同时失效时，因为异地灾备中心只剩下一份副本，不满足 Raft 算法大多数副本存活的要求。最终将导致集群暂时不可用，需要从一副本恢复集群，只会丢失少部分还没同步的热数据。这种情况出现的概率是比较小的。
  - 由于使用到了网络专线，导致该架构下网络设施成本较高。
  - 两地三中心需设置 5 副本，数据冗余度增加，增加空间成本。

11.9.2.2.1 详细示例

北京、西安两地三中心配置详解：

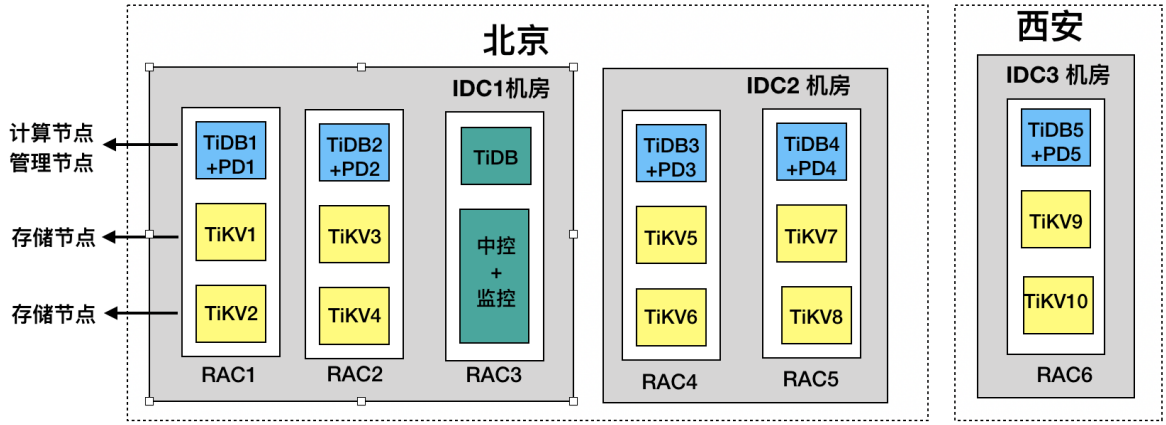


图 256: 两地三中心配置详图

- 如上图所示，北京有两个机房 IDC1 和 IDC2，机房 IDC1 中有三套机架 RAC1、 RAC2、 RAC3，机房 IDC2 有机架 RAC4、 RAC5；西安机房 IDC3 有机架 RAC6。
- 如上图中 RAC1 机架所示，TiDB、 PD 服务部署在同一台服务器上，还有两台 TiKV 服务器；每台 TiKV 服务器部署 2 个 TiKV 实例 (tikv-server)，RAC2、 RAC4、 RAC5、 RAC6 类似。
- 机架 RAC3 上安放 TiDB Server 及中控 + 监控服务器。部署 TiDB Server，用于日常管理维护、备份使用。中控 + 监控服务器上部署 Prometheus、 Grafana 以及恢复工具；
- 另可增加备份服务器，其上部署 Drainer，Drainer 以输出 file 文件的方式将 binlog 数据保存到指定位置，实现增量备份的目的。

11.9.2.3 配置

11.9.2.3.1 示例

以下为一个 `tiup topology.yaml` 文件示例：

```
### # Global variables are applied to all deployments and used as the default value of
### # the deployments if a specific deployment value is missing.
global:
  user: "tidb"
  ssh_port: 22
  deploy_dir: "/data/tidb_cluster/tidb-deploy"
  data_dir: "/data/tidb_cluster/tidb-data"

server_configs:
  tikv:
    server.grpc-compression-type: gzip
```

```
  pd:
    replication.location-labels:  ["dc","zone","rack","host"]
    schedule.tolerant-size-ratio: 20.0

pd_servers:
  - host: 10.63.10.10
    name: "pd-10"
  - host: 10.63.10.11
    name: "pd-11"
  - host: 10.63.10.12
    name: "pd-12"
  - host: 10.63.10.13
    name: "pd-13"
  - host: 10.63.10.14
    name: "pd-14"

tidb_servers:
  - host: 10.63.10.10
  - host: 10.63.10.11
  - host: 10.63.10.12
  - host: 10.63.10.13
  - host: 10.63.10.14

tikv_servers:
  - host: 10.63.10.30
    config:
      server.labels: { dc: "1", zone: "1", rack: "1", host: "30" }
  - host: 10.63.10.31
    config:
      server.labels: { dc: "1", zone: "2", rack: "2", host: "31" }
  - host: 10.63.10.32
    config:
      server.labels: { dc: "2", zone: "3", rack: "3", host: "32" }
  - host: 10.63.10.33
    config:
      server.labels: { dc: "2", zone: "4", rack: "4", host: "33" }
  - host: 10.63.10.34
    config:
      server.labels: { dc: "3", zone: "5", rack: "5", host: "34" }
      raftstore.raft-min-election-timeout-ticks: 1000
      raftstore.raft-max-election-timeout-ticks: 1200

monitoring_servers:
  - host: 10.63.10.60
```

```
grafana_servers:
  - host: 10.63.10.60

alertmanager_servers:
  - host: 10.63.10.60
```

11.9.2.3.2 Labels 设计

在两地三中心部署方式下，对于 Labels 的设计需要充分考虑到系统的可用性和容灾能力，建议根据部署的物理结构来定义 DC、ZONE、RACK、HOST 四个等级。

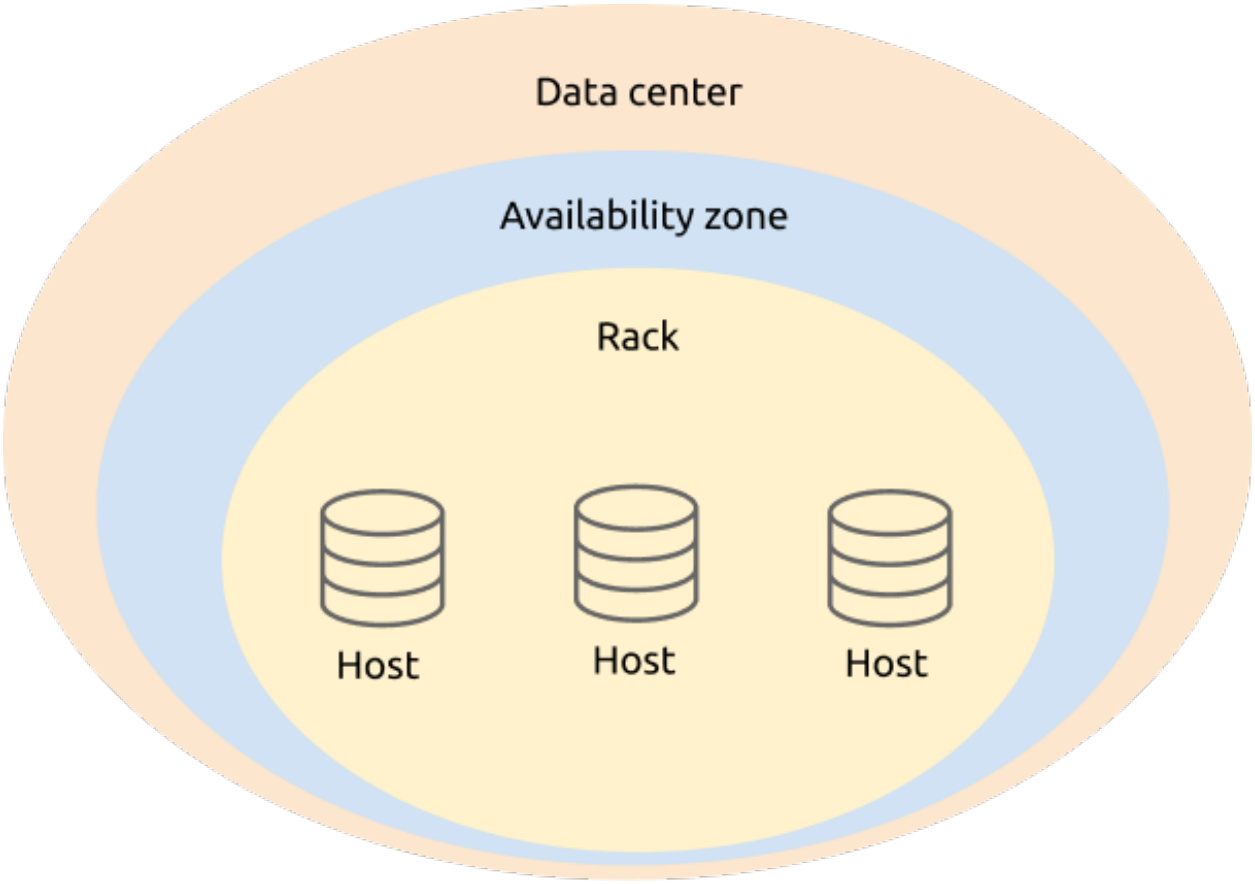


图 257: Label 逻辑定义图

PD 设置中添加 TiKV label 的等级配置。

```
server_configs:
  pd:
    replication.location-labels:  ["dc","zone","rack","host"]
```

tikv_servers 设置基于 TiKV 真实物理部署位置的 Label 信息，方便 PD 进行全局管理和调度。

```
tikv_servers:
  - host: 10.63.10.30
    config:
      server.labels: { dc: "1", zone: "1", rack: "1", host: "30" }
  - host: 10.63.10.31
    config:
      server.labels: { dc: "1", zone: "2", rack: "2", host: "31" }
  - host: 10.63.10.32
    config:
      server.labels: { dc: "2", zone: "3", rack: "3", host: "32" }
  - host: 10.63.10.33
    config:
      server.labels: { dc: "2", zone: "4", rack: "4", host: "33" }
  - host: 10.63.10.34
    config:
      server.labels: { dc: "3", zone: "5", rack: "5", host: "34" }
```

11.9.2.3.3 参数配置优化

在两地三中心的架构部署中，从性能优化的角度，除了常规参数配置外，还需要对集群中相关组件参数进行调整。

- 启用 TiKV gRPC 消息压缩。由于涉及到集群中的数据在网络中传输，需要开启 gRPC 消息压缩，降低网络流量。

```
server.grpc-compression-type: gzip
```

- 调整 PD balance 缓冲区大小，提高 PD 容忍度，因为 PD 会根据节点情况计算出各个对象的 score 作为调度的依据，当两个 store 的 leader 或 Region 的得分差距小于指定倍数的 Region size 时，PD 会认为此时 balance 达到均衡状态。

```
schedule.tolerant-size-ratio: 20.0
```

- 异地 DC3 TiKV 节点网络优化，单独修改异地 TiKV 此参数，拉长异地副本参与选举的时间，尽量避免异地 TiKV 中的副本参与 Raft 选举。

```
raftstore.raft-min-election-timeout-ticks: 1000
raftstore.raft-max-election-timeout-ticks: 1200
```

- 调度设置。在集群启动后，通过 `tiup ctl pd` 工具进行调度策略修改。修改 TiKV Raft 副本数按照安装时规划好的副本数进行设置，在本例中为 5 副本。

```
config set max-replicas 5
```

- 禁止向异地机房调度 Raft Leader，当 Raft Leader 在异地数据中心时，会造成不必要的本地数据中心与异地数据中心间的网络消耗，同时由于网络带宽和延迟的影响，也会对 TiDB 的集群性能产生影响。需要禁用异地中心的 Raft leader 的调度。

```
config set label-property reject-leader dc 3
```

> **注意：**
>
> TiDB 5.2 及以上版本默认不支持 `label-property` 配置。若要设置副本策略，请使用Placement Rules。

- 设置 PD 的优先级，为了避免出现异地数据中心的 PD 成为 Leader，可以将本地数据中心的 PD 优先级调高（数字越大，优先级越高），将异地的 PD 优先级调低。

```
member leader_priority PD-10 5
member leader_priority PD-11 5
member leader_priority PD-12 5
member leader_priority PD-13 5
member leader_priority PD-14 1
```

11.9.3 同城两中心自适应同步模式部署

本文介绍同城两中心的部署模式，相关架构、示例配置、副本方法、启用该模式的方法。

11.9.3.1 简介

TiDB 在 on-premises 部署的场景下，通常采用多中心部署方案，以保证高可用和容灾能力。多中心部署方案包括两地三中心部署模式、同城三中心部署模式等多种部署模式。本文介绍同城两中心部署方案，即在一座城市部署两个数据中心，成本更低，同样能满足高可用和容灾要求。该部署方案采用自适应同步模式，即 Data Replication Auto Synchronous，简称 DR Auto-Sync。

同城两中心部署方案下，两个数据中心距离在 50 km 以内，通常位于同一个城市或两个相邻城市（例如北京和廊坊），数据中心间的网络连接延迟小于 1.5 ms，带宽大于 10 Gbps。

11.9.3.2 部署架构

本文以某城市为例，城里有两个数据中心 IDC1 和 IDC2，分为位于城东和城西。

下图为集群部署架构图，具体如下：

- 集群采用同城两种中心部署方案，主数据中心 IDC1 在城东，从数据中心 IDC2 在城西。
- 集群采用推荐的 4 副本模式，其中 IDC1 中放 2 个 Voter 副本，IDC2 中放 1 个 Voter 副本 + 1 个 Learner 副本；TiKV 按机房的实际情况打上合适的 Label。
- 副本间通过 Raft 协议保证数据的一致性和高可用，对用户完全透明。

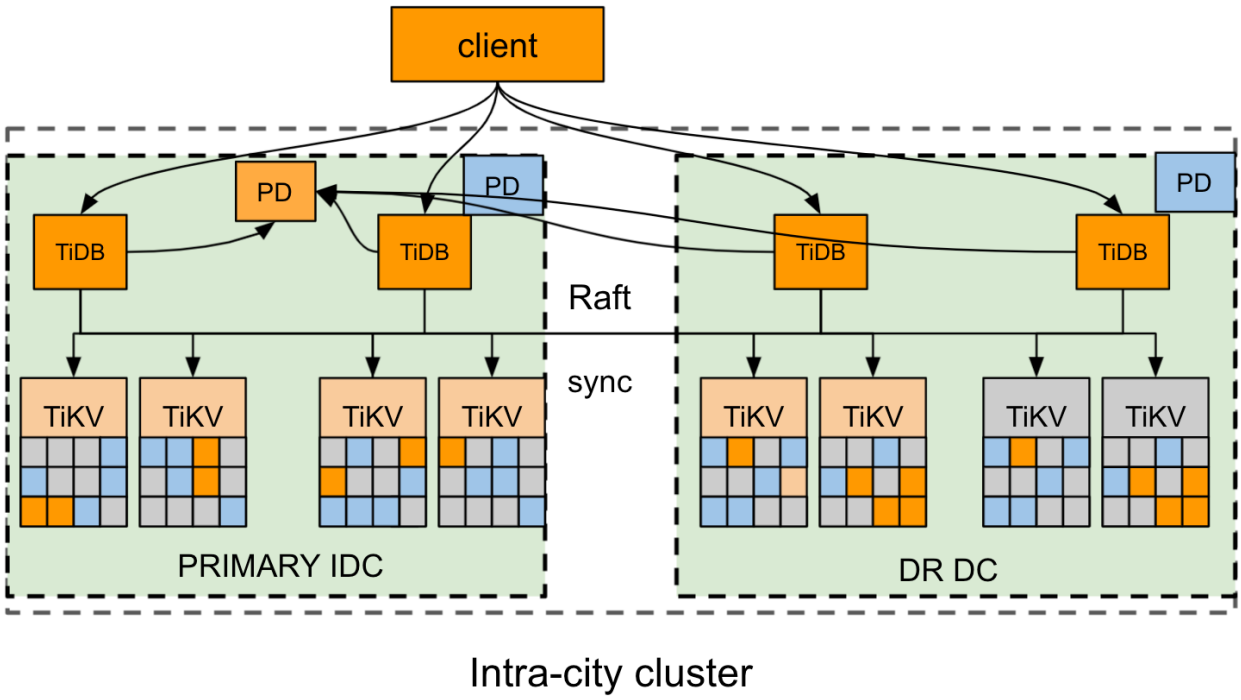


图 258: 同城两中心集群架构图

该部署方案定义了三种状态来控制和标示集群的同步状态，该状态约束了 TiKV 的同步方式。集群的复制模式可以自动在三种状态之间自适应切换。要了解切换过程，请参考状态转换。

- sync：同步复制模式，此时从数据中心 (DR) 至少有一个副本与主数据中心 (PRIMARY) 进行同步，Raft 算法保证每条日志 (log) 按 Label 同步复制到 DR。
- async：异步复制模式，此时不保证 DR 与 PRIMARY 完全同步，Raft 算法使用经典的 majority 方式复制日志。
- sync-recover：恢复同步，此时不保证 DR 与 PRIMARY 完全同步，Raft 逐步切换成 Label 复制，切换成功后汇报给 PD。

11.9.3.3 配置

11.9.3.3.1 示例

以下 `tiup topology.yaml` 示例拓扑文件为同城两中心典型的拓扑配置：

```
### # Global variables are applied to all deployments and used as the default value of
### # the deployments if a specific deployment value is missing.
global:
  user: "tidb"
  ssh_port: 22
  deploy_dir: "/data/tidb_cluster/tidb-deploy"
  data_dir: "/data/tidb_cluster/tidb-data"
server_configs:
```

```
  pd:
    replication.location-labels:  ["zone","rack","host"]
pd_servers:
  - host: 10.63.10.10
    name: "pd-10"
  - host: 10.63.10.11
    name: "pd-11"
  - host: 10.63.10.12
    name: "pd-12"
tidb_servers:
  - host: 10.63.10.10
  - host: 10.63.10.11
  - host: 10.63.10.12
tikv_servers:
  - host: 10.63.10.30
    config:
      server.labels: { zone: "east", rack: "east-1", host: "30" }
  - host: 10.63.10.31
    config:
      server.labels: { zone: "east", rack: "east-2", host: "31" }
  - host: 10.63.10.32
    config:
      server.labels: { zone: "west", rack: "west-1", host: "32" }
  - host: 10.63.10.33
    config:
      server.labels: { zone: "west", rack: "west-2", host: "33" }
monitoring_servers:
  - host: 10.63.10.60
grafana_servers:
  - host: 10.63.10.60
alertmanager_servers:
  - host: 10.63.10.60
```

11.9.3.3.2 Placement Rules 规划

为了按照规划的集群拓扑进行部署，你需要使用Placement Rules 来规划集群副本的放置位置。以 4 副本（2 Voter 副本在主数据中心，1 Voter 和 1 Learner 在从数据中心）的部署方式为例，可使用 Placement Rules 进行如下副本配置：

```
cat rule.json
[
  {
    "group_id": "pd",
    "id": "zone-east",
    "start_key": "",
```

```
  "end_key": "",
  "role": "voter",
  "count": 2,
  "label_constraints": [
    {
      "key": "zone",
      "op": "in",
      "values": [
        "east"
      ]
    }
  ],
  "location_labels": [
    "zone",
    "rack",
    "host",
  ]
},
{
  "group_id": "pd",
  "id": "zone-west",
  "start_key": "",
  "end_key": "",
  "role": "voter",
  "count": 1,
  "label_constraints": [
    {
      "key": "zone",
      "op": "in",
      "values": [
        "west"
      ]
    }
  ],
  "location_labels": [
    "zone",
    "rack",
    "host"
  ]
},
{
  "group_id": "pd",
  "id": "zone-west",
  "start_key": "",
  "end_key": "",
```

```
    "role": "learner",
    "count": 1,
    "label_constraints": [
      {
        "key": "zone",
        "op": "in",
        "values": [
          "west"
        ]
      }
    ],
    "location_labels": [
      "zone",
      "rack",
      "host"
    ]
  }
]
```

如果需要使用 `rule.json` 中的配置，你可以通过如下命令覆盖原有的配置：

```
pd-ctl config placement-rules rule-bundle save --in="rule.json"
```

11.9.3.3.3 启用自适应同步模式

副本的复制模式由 PD 节点控制。如果要使用 DR Auto-sync 自适应同步模式，需要在部署集群前先配置好 PD 的配置文件，如下所示：

```
[replication-mode]
replication-mode = "dr-auto-sync"
[replication-mode.dr-auto-sync]
label-key = "zone"
primary = "east"
dr = "west"
primary-replicas = 2
dr-replicas = 1
wait-store-timeout = "1m"
wait-sync-timeout = "1m"
```

在以上配置文件中：

- `replication-mode` 为待启用的复制模式，以上示例中设置为 `dr-auto-sync`。默认使用 majority 算法。
- `label-key` 用于区分不同的数据中心，需要和 Placement Rules 相匹配。其中主数据中心为 “east”，从数据中心为 “west”。
- `primary-replicas` 是在主数据中心 Voter 副本的数量。
- `dr-replicas` 是在从数据中心 Voter 副本的数量。

- `wait-store-timeout` 是当出现网络隔离或者故障时，切换到异步复制模式的等待时间。如果超过这个时间还没恢复，则自动切换到异步复制模式。默认时间为 60 秒。

如果需要检查当前集群的复制状态，可以通过以下 API 获取：

```
curl http://pd_ip:pd_port/pd/api/v1/replication_mode/status
```

```
{
  "mode": "dr-auto-sync",
  "dr-auto-sync": {
    "label-key": "zone",
    "state": "sync"
  }
}
```

状态转换

简单来讲，集群的复制模式可以自动在三种状态之间自适应的切换：

- 当集群一切正常时，会进入同步复制模式来最大化地保障灾备机房的数据完整性。
- 当机房网络断连或灾备机房发生整体故障时，在经过一段提前设置好的保护窗口之后，集群会进入异步复制状态，来保障业务的可用性。
- 当机房网络重连或灾备机房整体恢复之后，灾备机房的 TiKV 节点会重新加入到集群，逐步同步数据并最终转为同步复制模式。

状态转换的细节过程如下：

1. 初始化：集群在初次启动时处于 sync（同步复制）模式，PD 会下发信息给 TiKV，所有 TiKV 节点会严格按照 sync 模式的要求进行工作。
2. 同步切异步：PD 通过定时检查 TiKV 的心跳信息来判断 TiKV 是否宕机或断连。如果宕机数超过 PRIMARY/DR 各自副本的数量 `primary-replicas` 和 `dr-replicas`，意味着无法完成同步复制，需要切换状态。当宕机时间超过了 `wait-store-timeout` 设定的时间，PD 将集群状态切换成 async（异步复制）模式。然后 PD 再将 async 状态下发到所有 TiKV 节点，TiKV 的复制模式由双中心同步方式转为原生的 Raft 大多数落实方式 (majority)。
3. 异步切同步：PD 通过定时检查 TiKV 的心跳信息来判断 TiKV 是否恢复连接，如果宕机数小于 PRIMARY/DR 各自副本的数量，意味着可以切回同步了。PD 会将集群复制状态先切换至 sync-recover，再将该状态下发给所有 TiKV 节点。TiKV 的所有 Region 逐步切换成双机房同步复制模式，切换成功后通过心跳将状态同步信息给 PD。PD 记录 TiKV 上 Region 的状态并统计恢复进度。当 TiKV 的所有 Region 都完成了同步复制模式的切换，PD 将集群复制状态切换为 sync。

11.9.3.3.4 灾难恢复

本节介绍同城两中心部署提供的容灾恢复方案。

当处于同步复制状态的集群发生了灾难，可进行 RPO = 0 的数据恢复：

- 如果主数据中心发生故障，丢失了大多数 Voter 副本，但是从数据中心有完整的数据，可在从数据中心恢复数据。此时需要人工介入，通过专业工具恢复（恢复方式请联系 TiDB 团队）。
- 如果从数据中心发生故障，丢失了少数 Voter 副本，能自动切换成 async 异步复制模式。

当不处于同步复制状态的集群发生了灾难，不能保证满足 RPO = 0 进行数据恢复：

- 如果丢失了大多数 Voter 副本，需要人工介入，通过专业工具恢复（恢复方式请联系 TiDB 团队）。

12 运维工具

12.1 TiDB 工具功能概览

TiDB 提供了丰富的工具，可以帮助你进行部署运维、数据管理（例如，数据迁移、备份恢复、数据校验）、在 TiKV 上运行 Spark SQL。请根据需要选择适用的工具。

12.1.1 部署运维工具

TiDB 提供了 TiUP 和 TiDB Operator 两种部署运维工具，满足你在不同系统环境下的部署运维需求。

12.1.1.1 在物理机或虚拟机上部署运维 TiDB - TiUP

TiUP 是在物理机或虚拟机上的 TiDB 包管理器，管理着 TiDB 的众多的组件，如 TiDB、PD、TiKV 等。当你想要运行 TiDB 生态中任何组件时，只需要执行一行 TiUP 命令即可。

TiUP cluster 是 TiUP 提供的使用 Golang 编写的集群管理组件，通过 TiUP cluster 组件就可以进行日常的运维工作，包括部署、启动、关闭、销毁、弹性扩缩容、升级 TiDB 集群，以及管理 TiDB 集群参数。

基本信息：

- 术语及核心概念
- 使用 TiUP 部署 TiDB 集群
- TiUP 组件管理
- 适用 TiDB 版本：v4.0 及以上

12.1.1.2 在 Kubernetes 上部署运维 TiDB - TiDB Operator

TiDB Operator 是 Kubernetes 上的 TiDB 集群自动运维系统，提供包括部署、升级、扩缩容、备份恢复、配置变更的 TiDB 全生命周期管理。借助 TiDB Operator，TiDB 可以无缝运行在公有云或私有部署的 Kubernetes 集群上。

基本信息：

- TiDB Operator 架构
- 在 Kubernetes 上部署运维 TiDB 快速上手
- 适用 TiDB 版本：v2.1 及以上

12.1.2 数据管理工具

TiDB 提供了丰富的数据管理工具，例如数据迁移、导入导出、备份恢复、增量同步、数据校验等。

12.1.2.1 数据迁入 - TiDB Data Migration (DM)

TiDB Data Migration (DM) 是将 MySQL/MariaDB 数据迁移到 TiDB 的工具，支持全量数据的迁移和增量数据的复制。

基本信息：

- TiDB DM 的输入：MySQL/MariaDB
- TiDB DM 的输出：TiDB 集群
- 适用 TiDB 版本：所有版本
- Kubernetes 支持：使用 TiDB Operator 在 Kubernetes 上部署 TiDB DM。

如果数据量在 TB 级别以下，推荐直接使用 TiDB DM 迁移 MySQL/MariaDB 数据到 TiDB（迁移的过程包括全量数据的导出导入和增量数据的复制）。

如果数据量在 TB 级别，推荐的迁移步骤如下：

1. 使用Dumpling 导出 MySQL/MariaDB 全量数据。
2. 使用TiDB Lightning 将全量导出数据导入 TiDB 集群。
3. 使用 TiDB DM 复制 MySQL/MariaDB 增量数据到 TiDB。

> 注意：
>
> - 原 Syncer 工具已停止维护，不再推荐使用，相关场景请使用 TiDB DM 的增量复制模式进行替代。

12.1.2.2 全量导出 - Dumpling

Dumpling 是一个用于从 MySQL/TiDB 进行全量逻辑导出的工具。

基本信息：

- Dumpling 的输入：MySQL/TiDB 集群
- Dumpling 的输出：SQL/CSV 文件
- 适用 TiDB 版本：所有版本
- Kubernetes 支持：尚未支持

> 注意：
>
> PingCAP 之前维护的 Mydumper 工具 fork 自 Mydumper project，针对 TiDB 的特性进行了优化。Mydumper 已经被Dumpling 工具取代，并使用 Go 语言编写，支持更多针对 TiDB 特性的优化。建议切换到 Dumpling。

12.1.2.3 全量导入 - TiDB Lightning

TiDB Lightning 是一个用于将全量数据导入到 TiDB 集群的工具。

使用 TiDB Lightning 导入数据到 TiDB 时，有三种模式：

- `local` 模式：TiDB Lightning 将数据解析为有序的键值对，并直接将其导入 TiKV。这种模式一般用于导入大量的数据（TB 级别）到新集群，但在数据导入过程中集群无法提供正常的服务。
- `importer` 模式：和 `local` 模式类似，但是需要部署额外的组件 `tikv-importer` 协助完成键值对的导入。对于 4.0 以上的目标集群，请优先使用 `local` 模式进行导入。
- `tidb` 模式：以 TiDB/MySQL 作为后端，这种模式相比 `local` 和 `importer` 模式的导入速度较慢，但是可以在线导入，同时也支持将数据导入到 MySQL。

基本信息：

- TiDB Lightning 的输入：
  - Dumpling 输出文件
  - 其他格式兼容的 CSV 文件
- 适用 TiDB 版本：v2.1 及以上
- Kubernetes 支持：使用 TiDB Lightning 快速恢复 Kubernetes 上的 TiDB 集群数据

> 注意：
>
> 原 Loader 工具已停止维护，不再推荐使用。相关场景请使用 TiDB Lightning 的 `tidb` 模式进行替代，详细信息请参考TiDB Lightning TiDB-backend 文档。

12.1.2.4 备份和恢复 - Backup & Restore

Backup & Restore (BR) 是一个对 TiDB 进行分布式备份和恢复的工具，可以高效地对大数据量的 TiDB 集群进行数据备份和恢复。

基本信息：

- 备份输出和恢复输入的文件类型：SST + `backupmeta` 文件
- 适用 TiDB 版本：v4.0 及以上
- Kubernetes 支持：使用 BR 工具备份 TiDB 集群数据到兼容 S3 的存储，使用 BR 工具恢复 S3 兼容存储上的备份数据

12.1.2.5 TiDB 增量数据同步 - TiCDC

TiCDC 是一款通过拉取 TiKV 变更日志实现的 TiDB 增量数据同步工具，具有将数据还原到与上游任意 TSO 一致状态的能力，同时提供开放数据协议 (TiCDC Open Protocol)，支持其他系统订阅数据变更。

基本信息：

- TiCDC 的输入：TiDB 集群
- TiCDC 的输出：TiDB 集群、 MySQL、 Kafka、 Apache Pulsar、 Confluent
- 适用 TiDB 版本：v4.0.6 及以上

12.1.2.6 TiDB 增量日志同步 - TiDB Binlog

TiDB Binlog 是收集 TiDB 的增量 binlog 数据，并提供准实时同步和备份的工具。该工具可用于 TiDB 集群间的增量数据同步，如将其中一个 TiDB 集群作为另一个 TiDB 集群的从集群。

基本信息：

- TiDB Binlog 的输入：TiDB 集群
- TiDB Binlog 的输出：TiDB 集群、 MySQL、 Kafka 或者增量备份文件
- 适用 TiDB 版本：v2.1 及以上
- Kubernetes 支持：TiDB Binlog 运维文档，Kubernetes 上的 TiDB Binlog Drainer 配置

12.1.2.7 数据校验 - sync-diff-inspector

sync-diff-inspector 是一个用于校验 MySQL/TiDB 中两份数据是否一致的工具。该工具还提供了修复数据的功能，可用于修复少量不一致的数据。

基本信息：

- sync-diff-inspector 的输入：TiDB、 MySQL
- sync-diff-inspector 的输出：TiDB、 MySQL
- 适用 TiDB 版本：所有版本

12.1.3 OLAP 分析工具

TiDB 提供了 OLAP 分析工具 TiSpark。通过 TiSpark，你可以像使用原生 Spark 一样查询 TiDB 表。

12.1.3.1 使用 Spark 查询 TiKV 数据源 - TiSpark

TiSpark 是 PingCAP 为解决用户复杂 OLAP 需求而推出的产品。它借助 Spark 平台，同时融合 TiKV 分布式集群的优势，和 TiDB 一起为用户一站式解决 HTAP (Hybrid Transactional/Analytical Processing) 的需求。

12.2 TiDB 工具适用场景

本文档从生态工具的适用场景出发，介绍部分常见场景下的生态工具选择。

12.2.1 在物理机或虚拟机上部署运维 TiDB

当需要在物理机或虚拟机上部署运维 TiDB 时，你可以先安装TiUP，再通过 TiUP 管理 TiDB 的众多组件，如 TiDB、PD、 TiKV 等。

12.2.2 在 Kubernetes 上部署运维 TiDB

当需要在 Kubernetes 上部署运维 TiDB 时，你可以先创建 Kubernetes 集群，部署TiDB Operator，然后使用 TiDB Operator 部署运维 TiDB 集群。

12.2.3 从 CSV 导入数据到 TiDB

当需要将其他工具导出的格式兼容的 CSV files 导入到 TiDB 时，可使用TiDB Lightning。

12.2.4 从 MySQL/Aurora 导入全量数据

当需要从 MySQL/Aurora 导入全量数据时，可先使用Dumpling 将数据导出为 SQL dump files，然后再使用TiDB Lightning 将数据导入到 TiDB 集群。

12.2.5 从 MySQL/Aurora 迁移数据

当既需要从 MySQL/Aurora 导入全量数据，又需要迁移增量数据时，可使用TiDB Data Migration (DM) 完成从 Amazon Aurora 迁移数据到 TiDB。

如果全量数据量较大（TB 级别），则可先使用Dumpling 与TiDB Lightning 完成全量数据的迁移，再使用 DM 完成增量数据的迁移。

12.2.6 TiDB 集群备份与恢复

当需要对 TiDB 集群进行备份或在之后对 TiDB 集群进行恢复时，可使用BR。

另外，BR 也可以对 TiDB 的数据进行增量备份和增量恢复。

12.2.7 迁出数据到 TiDB

当需要将 TiDB 集群的数据迁出到其他 TiDB 集群时，可使用Dumpling 从 TiDB 将全量数据导出为 SQL dump files，然后再使用TiDB Lightning 将数据导入到 TiDB。

如果还需要执行增量数据的迁移，则可使用TiDB Binlog。

12.2.8 TiDB 增量数据订阅

当需要订阅 TiDB 增量数据的变更时，可使用TiDB Binlog。

12.3 TiDB 工具下载

本页面汇总了 TiDB 工具官方维护版本的下载链接。

12.3.1 TiUP

TiUP 安装过程十分简洁，无论是 Darwin 还是 Linux 操作系统，执行一行命令即可安装成功。详情请参考安装 TiUP。

12.3.2 TiDB Operator

TiDB Operator 运行在 Kubernetes 集群。在搭建好 Kubernetes 集群后，你可以选择在线或者离线部署 TiDB Operator。详情请参考在 Kubernetes 上部署 TiDB Operator。

12.3.3 TiDB Binlog

如需下载最新版本的TiDB Binlog，直接下载 TiDB 安装包即可，因为 TiDB Binlog 包含在 TiDB 安装包中。

安装包	操作系统	架构	SHA256 校验和
https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -{version}- ↪ linux ↪ -amd64 ↪ .tar ↪ .gz (TiDB Binlog)	Linux	amd64	https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -{version}- ↪ linux ↪ -amd64 ↪ .sha256 ↪

注意：

下载链接中的 {version} 为 TiDB 的版本号。例如，v5.4.0 版本的下载链接为 https://download ↪ .pingcap.org/tidb-v5.4.0-linux-amd64.tar.gz。

12.3.4 TiDB Lightning

使用下表中的链接下载TiDB Lightning：

安装包	操作系统	架构	SHA256 校验和
https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -toolkit ↪ -{version}- ↪ linux ↪ -amd64 ↪ .tar ↪ .gz	Linux	amd64	https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -toolkit ↪ -{version}- ↪ linux ↪ -amd64 ↪ .sha256 ↪

注意：

下载链接中的 {version} 为 TiDB Lightning 的版本号。例如，v5.4.0 版本的下载链接为 https://download.pingcap.org/tidb-toolkit-v5.4.0-linux-amd64.tar.gz。

12.3.5 备份和恢复 (BR) 工具

使用下表中的链接下载BR 工具：

安装包	操作系统	架构	SHA256 校验和
https ↪ :// ↪ download ↪ . ↪ pingcap ↪ . ↪ org ↪ / ↪ tidb ↪ - ↪ toolkit ↪ -{ ↪ version ↪ }- ↪ linux ↪ - ↪ amd64 ↪ . ↪ tar ↪ . ↪ gz	Linux	amd64	https ↪ :// ↪ download ↪ . ↪ pingcap ↪ . ↪ org ↪ / ↪ tidb ↪ - ↪ toolkit ↪ -{ ↪ version ↪ }- ↪ linux ↪ - ↪ amd64 ↪ . ↪ sha256 ↪

注意：

下载链接中的 {version} 为 BR 的版本号。例如，v5.4.0 版本的下载链接为 https://download ↪ .pingcap.org/tidb-toolkit-v5.4.0-linux-amd64.tar.gz。

12.3.6 TiDB DM (Data Migration)

使用下表中的链接下载DM：

安装包	操作系统	架构	SHA256 校验和
https ↪ ://download ↪ .pingcap ↪ .org ↪ /dm ↪ -{version}- ↪ linux ↪ - ↪ amd64 ↪ .tar ↪ .gz	Linux	amd64	https ↪ ://download ↪ .pingcap ↪ .org ↪ /dm ↪ -{version}- ↪ linux ↪ - ↪ amd64 ↪ .sha256 ↪

> **注意：**
>
> 下载链接中的 `{version}` 为 DM 的版本号。例如，`v5.4.0` 版本的下载链接为 `https://download.pingcap.org/dm-v5.4.0-linux-amd64.tar.gz`。可以通过 DM Release 查看当前已发布版本。

12.3.7 Dumpling

使用下表中的链接下载Dumpling:

安装包	操作系统	架构	SHA256 校验和
https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -toolkit ↪ -{version} ↪ -linux ↪ -amd64 ↪ .tar ↪ .gz	Linux	amd64	https ↪ ://download ↪ .pingcap ↪ .org ↪ /tidb ↪ -toolkit ↪ -{version} ↪ -linux ↪ -amd64 ↪ .sha256 ↪

注意：

- 下载链接中的 `{version}` 为 Dumpling 的版本号。例如，`v5.4.0` 版本的下载链接为 `https ↪ ://download.pingcap.org/tidb-toolkit-v5.4.0-linux-amd64.tar.gz`。可以通过 TiDB Release 查看当前已发布版本。
- Dumpling 已支持 arm64 linux，将下载链接中的 amd64 替换为 arm64，即表示 arm64 版 Dumpling。

12.3.8 sync-diff-inspector

使用下表中的链接下载sync-diff-inspector:

安装包	操作系统	架构	SHA256 校验和
tidb-enterprise-tools-nightly-linux-amd64.tar.gz	Linux	amd64	tidb-enterprise-tools-nightly-linux-amd64.sha256

12.3.9 TiCDC

TiCDC 的下载方式见TiCDC 安装部署文档。

12.4 TiUP

12.4.1 TiUP 文档地图

12.4.1.1 使用文档

- TiUP 概览：对 TiUP 进行整体介绍，如何安装和基本的用法以及相关术语
- TiUP 术语：解释使用 TiUP 过程中可能用到的术语，了解 TiUP 的核心概念
- TiUP 组件管理：详细介绍 TiUP 所有命令，如何使用 TiUP 下载、更新、删除组件
- TiUP FAQ：TiUP 使用过程中的常见问题，包含 TiUP 第三方组件的 FAQ
- TiUP 故障排查：如果在使用 TiUP 过程中遇到问题，可以参考故障排查文档的解决方案
- TiUP 参考手册：TiUP 详细参考手册，包含各类命令、组件、镜像。

12.4.1.2 资源

- AskTUG TiUP 主题
- TiUP Issues：TiUP Github Issues 列表

12.4.2 TiUP 简介

在各种系统软件和应用软件的安装管理中，包管理器均有着广泛的应用，包管理工具的出现大大简化了软件的安装和升级维护工作。例如，几乎所有使用 RPM 的 Linux 都会使用 Yum 来进行包管理，而 Anaconda 则可以非常方便地管理 python 的环境和相关软件包。

在早期的 TiDB 生态中，没有专门的包管理工具，使用者只能通过相应的配置文件和文件夹命名来手动管理，如 Prometheus 等第三方监控报表工具甚至需要额外的特殊管理，这样大大提升了运维管理难度。

从 TiDB 4.0 版本开始，TiUP 作为新的工具，承担着包管理器的角色，管理着 TiDB 生态下众多的组件，如 TiDB、PD、TiKV 等。用户想要运行 TiDB 生态中任何组件时，只需要执行 TiUP 一行命令即可，相比以前，极大地降低了管理难度。

12.4.2.1 安装 TiUP

TiUP 安装过程十分简洁，无论是 Darwin 还是 Linux 操作系统，执行一行命令即可安装成功：

```
curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
```

该命令将 TiUP 安装在 `$HOME/.tiup` 文件夹下，之后安装的组件以及组件运行产生的数据也会放在该文件夹下。同时，它还会自动将 `$HOME/.tiup/bin` 加入到 Shell Profile 文件的 PATH 环境变量中，这样你就可以直接使用 TiUP 了。

例如，你可以查看 TiUP 的版本：

```
tiup --version
```

> **注意：**
>
> TiUP 默认会收集使用情况信息，并将这些信息分享给 PingCAP 用于改善产品。若要了解所收集的信息详情及如何禁用该行为，请参见遥测。

12.4.2.2 TiUP 生态介绍

TiUP 的直接功能是作为 TiDB 生态中的包管理器，但这并不是它的最终使命。TiUP 的愿景是将 TiDB 生态中所有工具的使用门槛降到极致，这个仅仅靠包管理功能是做不到的，还需要引入一些额外的包来丰富这个系统，它们一起加入到 TiUP 生态中，让 TiDB 的世界变得更简单。

TiUP 系列文档的主要内容就是介绍 TiUP 及这些包的功能和使用方式。

在 TiUP 生态中，你可以通过在任何命令后加上 `--help` 的方式来获得帮助信息，比如通过以下命令获取 TiUP 本身的帮助信息:

```
tiup --help
```

```
TiUP is a command-line component management tool that can help to download and install
TiDB platform components to the local system. You can run a specific version of a component via
"tiup <component>[:version]". If no version number is specified, the latest version installed
locally will be used. If the specified component does not have any version installed locally,
the latest stable version will be downloaded from the repository.

Usage:
  tiup [flags] <command> [args...]
  tiup [flags] <component> [args...]

Available Commands:
  install     Install a specific version of a component
  list        List the available TiDB components or versions
  uninstall   Uninstall components or versions of a component
  update      Update tiup components to the latest version
```

```
  status      List the status of instantiated components
  clean       Clean the data of instantiated components
  mirror      Manage a repository mirror for TiUP components
  help        Help about any command or component

Components Manifest:
  use "tiup list" to fetch the latest components manifest

Flags:
  -B, --binary <component>[:version]   Print binary path of a specific version of a component <
      ↪ component>[:version]
                                       and the latest version installed will be selected if no
                                           ↪ version specified
      --binpath string                 Specify the binary path of component instance
  -h, --help                           help for tiup
      --skip-version-check             Skip the strict version check, by default a version must
          ↪ be a valid SemVer string
  -T, --tag string                     Specify a tag for component instance
  -v, --version                        version for tiup

Component instances with the same "tag" will share a data directory ($TIUP_HOME/data/$tag):
  $ tiup --tag mycluster playground

Examples:
  $ tiup playground                    # Quick start
  $ tiup playground nightly            # Start a playground with the latest nightly version
  $ tiup install <component>[:version] # Install a component of specific version
  $ tiup update --all                  # Update all installed components to the latest version
  $ tiup update --nightly              # Update all installed components to the nightly version
  $ tiup update --self                 # Update the "tiup" to the latest version
  $ tiup list                          # Fetch the latest supported components list
  $ tiup status                        # Display all running/terminated instances
  $ tiup clean <name>                  # Clean the data of running/terminated instance (Kill
      ↪ process if it's running)
  $ tiup clean --all                   # Clean the data of all running/terminated instances

Use "tiup [command] --help" for more information about a command.
```

输出的帮助信息较长，你可以只关注两部分：

- 可用的命令
  - install：用于安装组件
  - list：查看可用组件列表
  - uninstall：卸载组件
  - update：更新组件版本

  - status：查看组件运行记录
  - clean：清除组件运行记录
  - mirror：从官方镜像克隆一个私有镜像
  - help：输出帮助信息

- 可用的组件

  - playground：在本机启动集群
  - client：连接本机的集群
  - cluster：部署用于生产环境的集群
  - bench：对数据库进行压力测试

> 注意：
>
> - 可用的组件会持续增加，以 `tiup list` 输出结果为准。
> - 组件的可用版本列表也会持续增加，以 `tiup list <component>` 输出结果为准。

命令和组件的区别在于，命令是 TiUP 自带的，用于进行包管理的操作。而组件是 TiUP 通过包管理操作安装的独立组件包。比如执行 `tiup list` 命令，TiUP 会直接运行自己内部的代码，而执行 `tiup playground` 命令则会先检查本地有没有叫做 playground 的组件包，若没有则先从镜像上下载过来，然后运行这个组件包。

12.4.3 TiUP 术语及核心概念

本文主要说明 TiUP 的重要术语和核心概念。

12.4.3.1 TiUP 组件

TiUP 程序只包含少数几个命令，用来下载、更新、卸载组件。TiUP 通过各种组件来扩展其功能。组件是一个可以运行的程序或脚本，通过 `tiup <component>` 运行组件时，TiUP 会添加一组环境变量，并为该程序创建好对应的数据目录，然后运行该程序。

通过运行 `tiup <component>` 命令，你可以运行支持的 TiUP 组件，其中运行的逻辑为：

1. 如果用户通过 `tiup <component>[:version]` 运行指定某个组件的特定版本：

   - 组件在本地未安装任何版本，则从镜像服务器下载最新稳定版本
   - 组件在本地安装有其他版本，但是没有用户指定的版本，则从镜像服务器下载用户指定版本
   - 如果本地已经安装指定版本，则设置环境变量来运行已经安装的版本

2. 如果用户通过 `tiup <component>` 运行某个组件，且未指定任何版本：

   - 组件在本地未安装任何版本，则从镜像服务器下载最新稳定版本
   - 如果本地已经安装部分版本，则设置环境变量来运行已经安装的版本中的最新版本

12.4.3.2 TiUP 镜像仓库

TiUP 的所有组件都从镜像仓库 (mirrors) 下载，镜像仓库包含各个组件的 TAR 包以及对应的元信息（版本、入口启动文件、校验和）。TiUP 默认使用 PingCAP 官方的镜像仓库。用户可以通过 `TIUP_MIRRORS` 环境变量自定义镜像仓库。

镜像仓库可以是本地文件目录或在线 HTTP 服务器：

1. `TIUP_MIRRORS=/path/to/local tiup list`
2. `TIUP_MIRRORS=https://private-mirrors.example.com tiup list`

12.4.4 使用 TiUP 命令管理组件

TiUP 主要通过以下一些命令来管理组件：

- list：查询组件列表，用于了解可以安装哪些组件，以及这些组件可选哪些版本
- install：安装某个组件的特定版本
- update：升级某个组件到最新的版本
- uninstall：卸载组件
- status：查看组件运行状态
- clean：清理组件实例
- help：打印帮助信息，后面跟其他 TiUP 命令则是打印该命令的使用方法

本文介绍常用的组件管理操作及相应命令。

12.4.4.1 查询组件列表

你可以使用 `tiup list` 命令来查询组件列表。该命令用法如下：

- `tiup list`：查看当前有哪些组件可以安装
- `tiup list ${component}`：查看某个组件有哪些版本可以安装

你也可以在命令中组合使用以下参数 (flag)：

- `--installed`：查看本地已经安装了哪些组件，或者已经安装了某个组件的哪些版本
- `--all`：显式隐藏的组件
- `--verbose`：显式所有列（安装的版本、支持的平台）

示例一：查看当前已经安装的所有组件

```
tiup list --installed
```

示例二：从服务器获取 TiKV 所有可安装版本组件列表

```
tiup list tikv
```

12.4.4.2 安装组件

你可以使用 `tiup install` 命令来安装组件。该命令的用法如下：

- `tiup install <component>`：安装指定组件的最新稳定版
- `tiup install <component>:[version]`：安装指定组件的指定版本

示例一：使用 TiUP 安装最新稳定版的 TiDB

```
tiup install tidb
```

示例二：使用 TiUP 安装 nightly 版本的 TiDB

```
tiup install tidb:nightly
```

示例三：使用 TiUP 安装 v5.4.0 版本的 TiKV

```
tiup install tikv:v5.4.0
```

12.4.4.3 升级组件

在官方组件提供了新版之后，你可以使用 `tiup update` 命令来升级组件。除了以下几个参数，该命令的用法基本和 `tiup install` 相同：

- `--all`：升级所有组件
- `--nightly`：升级至 nightly 版本
- `--self`：升级 TiUP 自己至最新版本
- `--force`：强制升级至最新版本

示例一：升级所有组件至最新版本

```
tiup update --all
```

示例二：升级所有组件至 nightly 版本

```
tiup update --all --nightly
```

示例三：升级 TiUP 至最新版本

```
tiup update --self
```

12.4.4.4 运行组件

安装完成之后，你可以使用 `tiup <component>` 命令来启动相应的组件：

```
tiup [flags] <component>[:version] [args...]

Flags:
  -T, --tag string                    为组件实例指定 tag
```

该命令需要提供一个组件的名字以及可选的版本，若不提供版本，则使用该组件已安装的最新稳定版。

在组件启动之前，TiUP 会先为它创建一个目录，然后将组件放到该目录中运行。组件会将所有数据生成在该目录中，目录的名字就是该组件运行时指定的 tag 名称。如果不指定 tag，则会随机生成一个 tag 名称，并且在实例终止时自动删除工作目录。

如果想要多次启动同一个组件并复用之前的工作目录，就可以在启动时用 `--tag` 指定相同的名字。指定 tag 后，在实例终止时就不会自动删除工作目录，方便下次启动时复用。

示例一：运行 v5.4.0 版本的 TiDB

```
tiup tidb:v5.4.0
```

示例二：指定 tag 运行 TiKV

```
tiup --tag=experiment tikv
```

12.4.4.4.1 查询组件运行状态

你可以使用 `tiup status` 命令来查看组件的运行状态：

```
tiup status
```

运行该命令会得到一个实例列表，每行一个实例。列表中包含这些列：

- Name：实例的 tag 名称
- Component：实例的组件名称
- PID：实例运行的进程 ID
- Status：实例状态，`RUNNING` 表示正在运行，`TERM` 表示已经终止
- Created Time：实例的启动时间
- Directory：实例的工作目录，可以通过 `--tag` 指定
- Binary：实例的可执行程序，可以通过 `--binpath` 指定
- Args：实例的运行参数

12.4.4.4.2 清理组件实例

你可以使用 `tiup clean` 命令来清理组件实例，并删除工作目录。如果在清理之前实例还在运行，会先 kill 相关进程。该命令用法如下：

```
tiup clean [tag] [flags]
```

支持以下参数：

- `--all`：清除所有的实例信息

其中 tag 表示要清理的实例 tag，如果使用了 `--all` 则不传递 tag。

示例一：清理 tag 名称为 `experiment` 的组件实例

```
tiup clean experiment
```

示例二：清理所有组件实例

```
tiup clean --all
```

12.4.4.4.3 卸载组件

TiUP 安装的组件会占用本地磁盘空间，如果不想保留过多老版本的组件，可以先查看当前安装了哪些版本的组件，然后再卸载某个组件。

你可以使用 `tiup uninstall` 命令来卸载某个组件的所有版本或者特定版本，也支持卸载所有组件。该命令用法如下：

```
tiup uninstall [component][:version] [flags]
```

支持的参数：

- `--all`：卸载所有的组件或版本
- `--self`：卸载 TiUP 自身

component 为要卸载的组件名称，version 为要卸载的版本，这两个都可以省略，省略任何一个都需要加上 `--all` 参数：

- 若省略版本，加 `--all` 表示卸载该组件所有版本
- 若版本和组件都省略，则加 `--all` 表示卸载所有组件及其所有版本

示例一：卸载 v5.4.0 版本的 TiDB

```
tiup uninstall tidb:v5.4.0
```

示例二：卸载所有版本的 TiKV

```
tiup uninstall tikv --all
```

示例三：卸载所有已经安装的组件

```
tiup uninstall --all
```

12.4.5 TiUP FAQ

12.4.5.1 TiUP 是否可以不使用官方镜像源？

TiUP 支持通过环境变量 `TIUP_MIRRORS` 指定镜像源，镜像源的地址可以是一个本地目录或 HTTP 服务器地址。如果用户的环境不能访问网络，可以建立自己的离线镜像源使用 TiUP。

如果在使用非官方镜像之后想要切回官方镜像可以采取以下任一措施：

- 将 `TIUP_MIRRORS` 变量设置成官方镜像的地址：`https://tiup-mirrors.pingcap.com`。
- 先确保 `TIUP_MIRRORS` 变量没有设置，再使用执行 `tiup mirror set https://tiup-mirrors.pingcap.com` 命令。

12.4.5.2 如何将自己编写的组件放入 TiUP 镜像仓库？

TiUP 暂时不支持外部开发的组件，但是 TiUP Team 已经制定了 TiUP 组件开发规范，同时正在开发 tiup-publish 组件，完成 tiup-publish 组件后，开发者可以通过 `tiup publish <comp> <version>` 将自己开发的组件发布到 TiUP 的官方镜像仓库。

12.4.5.3 tiup-playground 和 tiup-cluster 有什么区别？

TiUP Playground 组件主要定位是快速上手和搭建单机的开发环境，支持 Linux/MacOS，要运行一个指定版本的 TiUP 集群更加简单。TiUP Cluster 组件主要是部署生产环境集群，通常是一个大规模的集群，还包含运维相关操作。

12.4.5.4 怎么样编写 tiup-cluster 组件的拓扑文件？

可以参考拓扑文件的样例，样例中包含了：

1. 两地三中心
2. 最小部署拓扑
3. 完整拓扑文件

可以根据自己的需求选择不同的模板，进行编辑。

12.4.5.5 同一个主机是否可以部署多个实例？

同一个主机可以使用 TiUP Cluster 部署多个实例，但是需要配置不同的端口和目录信息，否则可能导致目录以及端口冲突。

12.4.5.6 是否可以检测同一个集群内的端口和目录冲突？

同一个集群的端口和目录冲突会在部署和扩容的时候进行检测，如果有目录和端口冲突，本次部署或扩容会中断。

12.4.5.7 是否可以检测不同集群的端口和目录冲突？

如果不同集群是由同一个 TiUP 中控机部署的，会在部署和扩容时进行检测，如果属于不同的 TiUP 中控机，目前不支持检测。

12.4.5.8 集群部署期间，TiUP 收到报错 `ssh: handshake failed: read tcp 10.10.10.34:38980 -> 10.10.10.34:3600: read: connection reset by peer`

该报错可能是因为 TiUP 默认并发超过 ssh 默认最大连接数导致，可尝试加大 ssh 默认连接数，然后重启 sshd 服务解决：

```
vi /etc/ssh/sshd_config
MaxSessions 1000
MaxStartups 1000
```

12.4.6 TiUP 故障排查

本文介绍 TiUP 使用过程中一些常见的故障及排查方式，如果本文不包含你目前遇到的问题，可以通过以下方式求助：

1. Github Issues 新建一个 Issue。
2. 在 AskTUG 提交你的问题。

12.4.6.1 1. TiUP 命令故障排查

12.4.6.1.1 1.1 使用 `tiup list` 看不到最新的组件列表

TiUP 并不会每次都从镜像服务器更新最新的组件列表，可以通过 `tiup list` 来强制刷新组件列表。

12.4.6.1.2 1.2 使用 `tiup list <component>` 看不到一个组件的最新版本信息

同 1.1 一样，组件的版本信息只会在本地无缓存的情况下从镜像服务器获取，可以通过 `tiup list <component>` 刷新组件列表。

12.4.6.1.3 1.3 下载组件的过程中中断

如果下载组件的过程中网络中断，可能是由于网络不稳定导致的，可以尝试重新下载，如果多次不能成功下载，请反馈到 Github Issues，可能是由于 CDN 服务器导致的。

12.4.6.1.4 1.4 下载组件过程中出现 checksum 错误

由于 CDN 会有短暂的缓存时间，导致新的 checksum 文件和组件包不匹配，建议过 5 分钟后重试，如果依然不匹配，请反馈到 Github Issues。

12.4.6.2 2. TiUP Cluster 组件故障排查

12.4.6.2.1 2.1 部署过程中提示 `unable to authenticate, attempted methods [none publickey]`

由于部署时会向远程主机上传组件包，以及进行初始化，这个过程需要连接到远程主机，该错误是由于找不到连接到远程主机的 SSH 私钥导致的。请确认你是否通过 `tiup cluster deploy -i identity_file` 指定该私钥。

1. 如果没有指定 `-i` 参数，可能是由于 TiUP 没有自动找到私钥路径，建议通过 `-i` 显式指定私钥路径。
2. 如果指定了 `-i` 参数，可能是由于指定的私钥不能登录，可以通过手动执行 `ssh -i identity_file user@remote` 命令来验证。
3. 如果是通过密码登录远程主机，请确保指定了 `-p` 参数，同时输入了正确的登录密码。

12.4.6.2.2 2.2 使用 TiUP Cluster 升级中断

为了避免用户误用，TiUP Cluster 不支持指定部分节点升级，所以升级失败之后，需要重新进行升级操作，包括升级过程中的幂等操作。

升级操作会分为以下几步：

1. 首先备份所有节点的老版本组件
2. 分发新的组件到远程
3. 滚动重启所有组件

如果升级操作在滚动重启时中断，可以不用重复进行 `tiup cluster upgrade` 操作，而是通过 `tiup cluster restart -N <node1> -N <node2>` 来重启未完成重启的节点。如果同一组件的未重启节点数量比较多，也可以通过 `tiup cluster restart -R <component>` 来重启某一个类型的组件。

12.4.6.2.3 2.3 升级发现 node_exporter-9100.service/blackbox_exporter-9115.service 不存在

这种情况可能是由于之前的集群是由 TiDB Ansible 迁移过来的，且之前 TiDB Ansible 未部署 exporter 导致的。要解决这种情况，可以暂时通过手动从其他节点复制缺少的文件到新的节点。后续我们会在迁移过程中补全缺失的组件。

12.4.7 TiUP 命令参考手册

12.4.7.1 TiUP

TiUP 在 TiDB 生态中承担包管理器的功能，管理着 TiDB 生态下众多的组件，如 TiDB、PD、TiKV 等。

12.4.7.1.1 语法

```
tiup [flags] <command> [args...]        # 执行命令
#### or
tiup [flags] <component> [args...]      # 运行组件
```

使用 `help` 命令可以获取特定命令的信息，每个命令的摘要都显示了其参数及其用法。必须参数显示在尖括号中，可选参数显示在方括号中。

`<command>` 代表命令名字，支持的命令列表请参考下方命令清单，`<component>` 代表组件名，支持的组件列表请参考下方组件清单。

12.4.7.1.2 选项

-B, –binary

打印指定组件的二进制文件路径：

- 执行 `tiup -B/--binary <component>` 将打印已安装的 `<component>` 组件的最新稳定版路径，若 `<component>` 组件未安装，则报错
- 执行 `tiup -B/--binary <component>:<version>` 将打印已经安装的 `<component>` 组件的 `<version>` 版本所在的路径，若该版本未安装，则报错

- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

> **注意：**
>
> 该选项只能用于 `tiup [flags] <component> [args...]` 格式的命令。

–binpath (string)

指定要执行的组件的路径：执行一个组件时，如果不想使用 TiUP 镜像中的二进制文件，可以使用该参数使用自定义路径的二进制文件替换之。

> **注意：**
>
> 该选项只能用于 `tiup [flags] <component> [args...]` 格式的命令。

–skip-version-check

> **注意：**
>
> 该选项自 `v1.3.0` 版本起已废弃。

- 跳过版本号合法性检查，默认指定的版本号只能是 Semantic Version。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-T, –tag (string)

对启动的组件指定一个 `tag`：有的组件在执行过程中需要使用磁盘存储，TiUP 会分配一个临时目录作为该组件本次执行的存储目录，如果希望分配固定目录，可以用 `-T/--tag` 来指定目录名字，这样多次执行使用同样的 `tag` 就能读写到同一批文件。

-v, –version

打印 TiUP 的版本

–help

打印帮助信息

12.4.7.1.3 命令清单

TiUP 包含众多的命令，这些命令又包含了许多子命令，具体命令及其子命令的说明请参考对应的链接：

- install：安装组件
- list：查看组件列表
- uninstall：卸载组件
- update：升级已安装的组件
- status：查看组件运行状态
- clean：清理组件数据目录
- mirror：镜像管理
- telemetry：遥测开关
- completion：TiUP 命令补全
- env：查看 TiUP 相关环境变量
- help：查看特定命令或组件的帮助文档

12.4.7.1.4 组件清单

- cluster：生产环境 TiDB 集群管理
- dm：生产环境 DM 集群管理

12.4.7.2 TiUP 命令

12.4.7.2.1 tiup clean

命令 `tiup clean` 用于清除组件运行过程中产生的数据。

语法

```
tiup clean [name] [flags]
```

`[name]` 取值为status 命令输出的 `Name` 字段。若省略 `[name]`，则必须配合 `--all` 使用。

选项

–all

- 清除所有运行记录。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

```
Clean instance of `%s`, directory: %s
```

<< 返回上一页 - TiUP 命令清单

12.4.7.2.2 tiup completion

为了降低使用成本，TiUP 提供了 `tiup completion` 命令用于生成命令行自动补全的配置文件。目前支持 bash 和 zsh 两种 shell 的命令补全。

如果是 bash，需要提前安装好 bash-completion：

- 在 macOS 上的安装方式为：如果 bash 版本小于 4.1，执行 `brew install bash-completion`；如果 bash 版本大于等于 4.1，则执行 `brew install bash-completion@2`。
- 在 Linux 上的安装方式为：使用包管理器安装 `bash-completion` 包，例如执行 `yum install bash-completion` 或者 `apt install bash-completion`。

语法

```
tiup completion <shell>
```

`<shell>` 为 shell 类型，目前支持 bash 和 zsh。

使用方式

bash

将自动补全代码写入一个文件，并且在 `.bash_profile` 中对其执行 `source` 命令：

```
tiup completion bash > ~/.tiup.completion.bash

printf "
##### tiup shell completion
source '$HOME/.tiup.completion.bash'
" >> $HOME/.bash_profile

source $HOME/.bash_profile
```

zsh

```
tiup completion zsh > "${fpath[1]}/_tiup"
```

<< 返回上一页 - TiUP 命令清单

12.4.7.2.3 tiup env

TiUP 为用户提供了灵活的定制化接口，其中一部分是使用环境变量来实现的，命令 `tiup env` 用于查询 TiUP 支持用户自定义的环境变量以及它们此时的值。

语法

```
tiup env [name1...N]
```

`[name1...N]` 用于查看指定的环境变量，若不指定，则默认查看所有支持的环境变量。

选项

无

输出

- 若未指定 `[name1...N]`，则输出 “{key}” = “{value}” 列表
- 若指定了 `[name1...N]`，则按顺序输出 “{value}” 列表

以上输出中若 `value` 为空则代表未设置环境变量的值，此时 TiUP 会使用默认值。

<< 返回上一页 - TiUP 命令清单

12.4.7.2.4 tiup help

TiUP 命令行界面为用户提供了丰富的帮助信息，用户可以通过 `help` 命令或者 `--help` 参数查看。`tiup help <command>` 命令等价于 `tiup <command> --help`。

语法

```
tiup help [command]
```

`[command]` 用于指定要查看哪个命令的帮助信息，若不指定，则查看 TiUP 自身的帮助信息。

选项

无

输出

`[command]` 或 TiUP 的帮助信息。

<< 返回上一页 - TiUP 命令清单

12.4.7.2.5 tiup install

命令 `tiup install` 用于组件安装，它会从镜像仓库中下载指定版本的组件包，并在本地的 TiUP 数据目录中解压，以便后续使用。另外，当 TiUP 需要运行一个镜像仓库中不存在的组件时，会尝试先下载该组件，再自动运行，若仓库中不存在会报错。

语法

```
tiup install <component1>[:version] [component2...N] [flags]
```

`<component1>` 和 `<component2>` 代表组件名字，`[version]` 代表一个可选的版本号，若不加 `version`，则安装指定组件的最新稳定版本。`[component2...N]` 表示可同时指定多个组件或同一个组件的多个版本。

选项

无

输出

- 正常情况下输出组件的下载信息
- 若组件不存在则报错 `The component "%s" not found`
- 若版本不存在则报错 `version %s not supported by component %s`

<< 返回上一页 - TiUP 命令清单

12.4.7.2.6 tiup list

命令 `tiup list` 用于查询镜像中可用的组件列表。

语法

```
tiup list [component] [flags]
```

`[component]` 是可选的组件名称。若指定，则列出该组件的所有版本；若不指定，则列出所有组件列表。

选项

–all

- 显示所有组件。默认只显示非隐藏组件。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–installed

- 只显示已经安装的组件或版本。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–verbose

- 在组件列表中显示已安装的版本列表。默认组件列表不显示当前已安装的版本。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 若未指定 `[component]`
  - 若指定 –verbose：输出 组件名 `(Name)`、已安装版本 `(Installed)`、组件管理员 `(Owner)`、组件描述 `(Description)` 构成的组件信息列表
  - 若不指定 –verbose：输出 组件名 `(Name)`、组件管理员 `(Owner)`、组件描述 `(Description)` 构成的组件信息列表
- 若指定 `[component]`
  - 若 `[component]` 存在：输出 版本 `(Version)`、是否已安装 `(Installed)`、发布时间 `(Release)`、支持的平台 `(Platforms)` 构成的版本信息列表
  - 若 `[component]` 不存在：报错 `failed to fetch component: unknown component`

<< 返回上一页 - TiUP 命令清单

12.4.7.2.7 tiup mirror

tiup mirror

在 TiUP 中，镜像是一个非常重要的概念，目前 TiUP 支持两种形式的镜像：

- 本地镜像：即 TiUP 客户端和镜像在同一台机器上，客户端通过文件系统访问镜像
- 远程镜像：即 TiUP 客户端和镜像不在同一台机器上，客户端通过网络访问镜像

命令 `tiup mirror` 用于管理镜像，提供了创建镜像，组件发布，密钥管理等多种功能。

语法

```
tiup mirror <command> [flags]
```

`<command>` 代表子命令，支持的子命令列表请参考下方命令清单。

选项

无

命令清单

- genkey: 生成私钥文件
- sign: 使用私钥文件对特定文件进行签名
- init: 创建一个空的镜像
- set: 设置当前镜像
- grant: 为当前镜像引入新的组件管理员
- publish: 向当前镜像推送新的组件
- modify: 修改当前镜像中的组件属性
- rotate: 更新当前镜像中的根证书
- clone: 从已有镜像克隆一个新的镜像
- merge: 合并镜像

<< 返回上一页 - TiUP 命令清单

tiup mirror clone

命令 `tiup mirror clone` 用于克隆一个已经存在的镜像或克隆部分组件生成一个新的镜像。新旧镜像的组件相同，但使用的签名密钥不同。

语法

```
tiup mirror clone <target-dir> [global version] [flags]
```

- `<target-dir>` 是本地存放克隆下来的镜像的路径，如果不存在则会自动创建。
- 若指定了 `[global version]` 参数，TiUP 会尝试克隆指定版本的所有组件。若某些组件没有指定的版本，则克隆其最新版本。

选项

-f, –full

- 是否克隆整个镜像。指定该选项后会从目标镜像完整克隆所有组件的所有版本，此时其他指定的选项将失效。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-a, –arch

- 仅克隆能在指定平台上运行的组件。
- 数据类型：STRINGS
- 该参数接受以逗号分隔的多个平台名称，例如 `amd64,arm64`。如果未指定该选项，默认克隆 AMD64 和 ARM64 平台的组件，即 `amd64,arm64`。

-o, –os

- 仅克隆能在指定操作系统上运行的组件。
- 数据类型：STRINGS
- 该参数接受以逗号分隔的多个操作系统名称，例如 `linux,darwin`。如果未指定该选项，默认克隆 Linux 和 Darwin 系统的组件，即 `linux,darwin`。

–prefix

- 匹配版本时是否前缀匹配。默认情况下必须严格匹配指定的版本才会下载，指定该选项之后，仅前缀匹配指定的版本也会被下载。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–{component}（strings，默认为空）

指定要克隆的 `{component}` 组件的版本列表。`{component}` 为组件名，可选的组件名可执行`tiup list --all` 查看。

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror genkey

在 TiUP 镜像的定义中，有三类角色：

- 镜像管理员：拥有 `root.json`、`index.json`、`snapshot.json` 以及 `timestamp.json` 的修改权限
- 组件管理员：拥有相关组件的修改权限
- 普通用户：可以下载并使用组件

由于修改文件需要相关的管理员进行签名，因此管理员必须拥有自己的私钥。命令 `tiup mirror genkey` 就是用于生成私钥的。

> **警告：**
>
> 请勿通过网络传输私钥。

语法

```
tiup mirror genkey [flags]
```

选项

-n, –name

- 密钥的名字，该名字决定最终生成的文件名。生成的私钥文件路径为：`${TIUP_HOME}/keys/{name}.json`，其中 `TIUP_HOME` 为 TiUP 的家目录，默认路径为 `$HOME/.tiup`，`name` 为 `-n/--name` 指定的密钥名字。
- 数据类型：`STRING`
- 如果不指定该选项，密钥名默认为 `private`。

-p, –public

- 显示当前私钥对应的公钥，当前私钥名字由 `-n/--name` 选项指定。
- 当指定了 `-p/--public` 时，不会创建新的私钥。若 `-n/--name` 指定的私钥不存在，则报错。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–save

- 将公钥信息储存为文件放置于当前目录，文件名称为 `{hash-prefix}-public.json`，其中 `hash-prefix` 为该密钥 ID 的前 16 位。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 若未指定 `-p/--public`：
  - 若指定的密钥已存在：`Key already exists, skipped`
  - 若指定的密钥不存在：`private key have been write to ${TIUP_HOME}/keys/{name}.json`
- 若指定 `-p/--public`：
  - 若指定的密钥不存在：`Error: open ${TIUP_HOME}/keys/{name}.json: no such file or directory`
  - 若指定的密钥存在：输出该密钥对应的公钥内容

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror grant

命令 `tiup mirror grant` 用来向当前镜像中引入一个组件管理员。

组件管理员可以使用其密钥发布新的组件，也可以对其之前发布的组件作出修改。添加组件管理员时，待添加的组件管理员需要先将其公钥发送给镜像管理员。

> **注意：**
>
> 该命令仅支持在当前镜像为本地镜像时使用。

语法

```
tiup mirror grant <id> [flags]
```

<id> 为该组件管理员的 ID，该 ID 需要在整个镜像中唯一，建议使用符合正则 `^[a-z\d](?:[a-z\d]|-(?=[a-z\d])){0,38}$` 的字符串。

选项

-k, –key（string，默认 ${TIUP_HOME}/keys/private.json）

指定引入的组件管理员的密钥。该密钥可以是公钥也可以是私钥。如果传入私钥，会被转换成对应的公钥储存在镜像中。

一个密钥只能被一个组件管理员使用。

-n, –name（string，默认 <id>）

指定组件管理员的名字，该名字会展示在组件列表的 Owner 字段上。若未指定 -n/--name 则使用 <id> 作为组件管理员名字。

输出

- 若执行成功：无输出
- 若管理员 ID 重复：`Error: owner %s exists`
- 若密钥已被其他管理员使用：`Error: key %s exists`

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror init

命令 `tiup mirror init` 用于初始化一个空的镜像。初始化的镜像不包含任何组件和组件管理员，仅生成以下文件：

```
+ <mirror-dir>                              # 镜像根目录
|-- root.json                               # 镜像根证书
|-- 1.index.json                            # 组件/用户索引
|-- snapshot.json                           # 镜像最新快照
|-- timestamp.json                          # 镜像最新时间戳
|--+ keys                                   # 镜像私钥（可移动到其他位置）
   |-- {hash1..hashN}-root.json             # 根证书私钥
   |-- {hash}-index.json                    # 索引私钥
   |-- {hash}-snapshot.json                 # 快照私钥
   |-- {hash}-timestamp.json                # 时间戳私钥
```

以上文件的具体作用及内容格式请参考镜像说明。

语法

```
tiup mirror init <path> [flags]
```

<path> 为本地目录路径，可以为相对路径。TiUP 会以此路径为镜像文件存放路径，在其中生成文件。若该目录已存在，则必须保证为空，若该目录不存在，则 TiUP 会自动创建。

选项

-k, –key-dir（string，默认 {path}/keys）

指定生成私钥文件的目录。若指定的文件目录不存在，则会自动创建。

输出

- 若成功：无输出
- 若 <path> 不为空：`Error: the target path '%s' is not an empty directory`
- 若 <path> 不是目录：`Error: fdopendir: not a directory`

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror merge

命令 `tiup mirror merge` 用于合并一个或多个镜像到当前镜像。

执行此命令需要满足几个条件：

- 目标镜像的所有组件的管理员 ID 必须在当前镜像中存在。
- 执行该命令用户的 `${TIUP_HOME}/keys` 目录中有上述管理员 ID 在当前镜像中对应的所有私钥（可以使用命令`tiup mirror set` 将当前镜像切换成目前有权限修改的镜像）。

语法

```
tiup mirror merge <mirror-dir-1> [mirror-dir-N] [flags]
```

- `<mirror-dir-1>`：要合并到当前镜像的第一个镜像
- `[mirror-dir-N]`：要合并到当前镜像的第 N 个镜像

选项

无

输出

- 成功：无输出
- 当前镜像缺失目标镜像某个组件的管理员，或 `${TIUP_HOME}/keys` 缺失该管理员的私钥：`Error: missing owner keys for owner %s on component %s`

tiup mirror modify

命令 `tiup mirror modify` 用于修改已发布的组件。只有合法的组件管理员才可以修改组件，且只能修改其自己发布的组件。组件发布方式参考publish 命令。

语法

```
tiup mirror modify <component>[:version] [flags]
```

各个参数解释如下：

- `<component>`：组件名称
- `[version]`：想要修改的版本，若不指定，则表示修改整个组件

选项

-k, –key（string，默认 ${TIUP_HOME}/keys/private.json）

组件管理员的私钥，客户端需要使用该私钥对组件信息 (`{component}.json`) 进行签名。

–yank

- 将指定组件或指定版本标记为不可用：
  - 标记组件不可用之后 `tiup list` 将看不到该组件，也无法安装该组件的新版本。
  - 标记版本不可用之后 `tiup list <component>` 将看不到该版本，也无法安装该版本。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–hide

- 将该组件设置为隐藏，隐藏之后该组件将不在 `tiup list` 的列表中显示，但是可通过执行 `tiup list --all` 查看。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

> 注意：
>
> 该选项只能应用于组件上，无法应用于组件的版本上。

输出

- 若成功：无输出
- 若该组件管理员无权修改目标组件：

- 若使用远程镜像：`Error: The server refused, make sure you have access to this component`
- 若使用本地镜像：`Error: the signature is not correct`

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror publish

命令 `tiup mirror publish` 用于发布新组件，或已有组件的新版本。只有有权限的组件管理员才可以发布组件。引入组件管理员的方式可参考grant 命令。

语法

```
tiup mirror publish <comp-name> <version> <tarball> <entry> [flags]
```

各个参数解释如下：

- `<comp-name>`：组件名，如 `tidb`，建议使用符合正则 `^[a-z\d](?:[a-z\d]|-(?=[a-z\d])){0,38}$` 的字符串
- `<version>`：当前正在发布的版本，版本号需要符合 Semantic Versioning
- `<tarball>`：`.tar.gz` 包的本地路径，需要将组件的可执行文件及依赖放在该包中，由 TiUP 上传到镜像
- `<entry>`：组件的可执行文件在 `<tarball>` 中的位置

选项

-k, –key（string，默认 ${TIUP_HOME}/keys/private.json）

组件管理员的私钥，客户端需要使用该私钥对组件信息 (`{component}.json`) 进行签名。

–arch（string，默认 ${GOARCH}）

该 `<tarlball>` 中的二进制文件运行的平台，一个 `<tarball>` 只能选以下三个平台之一：

- `amd64`：表示在 amd64 架构的机器上运行
- `arm64`：表示在 arm64 架构的机器上运行
- `any`：表示可以在以上两种架构的机器上运行（比如脚本）

> 注意：
>
> 若 `--arch` 指定为 any，则 `--os` 也必须指定为 any。

–os（string，默认 ${GOOS}）

该 `<tarlball>` 中的二进制文件运行的操作系统，一个 `<tarball>` 只能选以下三个操作系统之一：

- `linux`：表示在 Linux 操作系统上运行
- `darwin`：表示在 Darwin 操作系统上运行
- `any`：表示可以在以上两种操作系统上运行（比如脚本）

> 注意：
>
> 若 --os 指定为 any，则 --arch 也必须指定为 any。

–desc（string，默认为空）

该组件的描述信息。

–hide

- 是否为隐藏组件。若为隐藏组件，则不在 tiup list 的列表中显示，但在 tiup list --all 的列表中会显示。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

- 若成功：无输出
- 若该组件管理员无权修改目标组件：
  - 若使用远程镜像：Error: The server refused, make sure you have access to this component
  - 若使用本地镜像：Error: the signature is not correct

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror rotate

TiUP 的镜像中有一个非常重要的文件：root.json，里面记录了整个系统需要使用的公钥，是 TiUP 信任链的基础，它的内容主要包含几个部分：

- N 个管理员的签名，对于官方镜像，N 为 5，默认初始化的镜像 N 为 3
- 用于验证以下文件的公钥：
  - root.json
  - index.json
  - snapshot.json
  - timestamp.json
- 过期时间，对于官方镜像，为 root.json 创建时间后延一年

关于镜像的详细介绍可以参考镜像说明。

在某些情况下，用户需要更新 root.json:

- 更换镜像的密钥
- 更新证书过期时间

更新 root.json 内容之后，必须由管理员对其进行重新签名，否则客户端会拒绝，更新流程如下：

1. 更新 root.json 的内容
2. N 个管理员对新的 root.json 进行签名
3. 更新 snapshot.json，记录新的 root.json 的 version
4. 对新的 snapshot.json 进行签名
5. 更新 timestamp.json，记录新的 snapshot.json 的 hash
6. 对新的 timestamp.json 进行签名

TiUP 使用命令 `tiup mirror rotate` 来自动化以上流程。

> 注意：
>
> - 经测试，小于 TiUP v1.5.0 的版本无法正确获得新的 root.json #983。
> - 使用此功能前请确保所有的 TiUP 客户端升级到了 v1.5.0 或以上版本。

语法

```
tiup mirror rotate [flags]
```

该命令会启动一个编辑器，修改其内容为目标值（比如将 `expires` 字段的值向后推移），然后需要将 `version` 字段加一并保存。保存之后会启动一个临时的 http 服务器，等待 N 个不同的镜像管理员签名。

镜像管理员签名的方式参考sign 命令。

选项

–addr（string，默认 0.0.0.0:8080）

临时服务器的监听地址，需要确保该地址可以被其他镜像管理员访问，这样管理员才能使用sign 命令签名。

输出

- 各个镜像管理员当前的签名状态

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror set

命令 `tiup mirror set` 用于切换当前镜像，支持本地文件系统和远程网络两种镜像。

官方镜像为 `https://tiup-mirrors.pingcap.com`。

语法

```
tiup mirror set <mirror-addr> [flags]
```

`<mirror-addr>` 为镜像地址，可以有两种形式：

- 网络地址：以 http 或者 https 开头，如 `http://172.16.5.5:8080`，`https://tiup-mirrors.pingcap.com` 等
- 本地文件路径：镜像目录的绝对路径，比如 `/path/to/local-tiup-mirror`

选项

-r, –root（string，默认 `{mirror-dir}/root.json`）

指定根证书。

每个镜像的根证书不相同，而根证书是镜像安全性最关键的一环，在使用网络镜像时，可能遭受中间人攻击，为了避免此类攻击，推荐手动将根网络镜像的根证书下载到本地：

```
wget <mirror-addr>/root.json -O /path/to/local/root.json
```

然后进行人工查验，认定无误之后，再通过手工指定根证书的方式切换镜像：

```
tiup mirror set <mirror-addr> -r /path/to/local/root.json
```

在这种操作方式下，如果中间人在 `wget` 之前攻击了镜像，用户可发现根证书不正确。如果在 `wget` 之后攻击了镜像，TiUP 会发现镜像和根证书不符。

输出

无

<< 返回上一页 - TiUP Mirror 命令清单

tiup mirror sign

命令 `tiup mirror sign` 用于对镜像中定义的元信息文件（*.json）进行签名，这些文件可能储存在本地文件系统，也可以放在远端使用 http 协议提供签名入口。

语法

```
tiup mirror sign <manifest-file> [flags]
```

`<manifest-file>` 为被签名的文件地址，可以有两种地址：

- 网络地址：http 或者 https 开头，如 `http://172.16.5.5:8080/rotate/root.json`
- 本地文件路径：相对路径或绝对路径均可

如果是网络地址，该地址必须提供以下功能：

- 支持以 `http get` 访问，此时应当返回被签名文件的完整内容（包含 signatures 字段）
- 支持以 `http post` 访问，客户端会在 `http get` 返回的内容的 signatures 字段中加上本次的签名 POST 到该地址

选项

-k, –key (string, 默认 ${TIUP_HOME}/keys/private.json)

指定用于签名的私钥位置。

–timeout (int，默认 10)

通过网络签名时网络的访问超时时间，单位为秒。

> **注意：**
>
> 只有当 `<manifest-file>` 为网络地址时该选项有效。

输出

- 成功：无输出
- 文件已被指定的 key 签名过：`Error: this manifest file has already been signed by specified key`
- 文件不是合法的 manifest：`Error: unmarshal manifest: %s`

<< 返回上一页 - TiUP Mirror 命令清单

12.4.7.2.8 tiup status

使用命令 `tiup status` 可查看组件的运行信息：通过 `tiup [flags] <component> [args...]` 运行组件之后，可以通过该命令查看组件的运行信息。

> **注意：**
>
> 只能查询到以下两种组件的信息：
>
> - 尚在运行的组件
> - 通过 `tiup -T/--tag` 指定 tag 运行的组件

语法

```
tiup status [flags]
```

选项

无

输出

由以下字段构成的表格：

- Name: 通过 `-T/--tag` 指定的 Tag 名字，若未指定，则为随机字符串
- Component: 运行的组件
- PID: 对应的进程 ID
- Status: 组件运行状态
- Created Time: 启动时间
- Directory: 数据目录
- Binary: 二进制文件路径
- Args: 启动参数

<< 返回上一页 - TiUP 命令清单

12.4.7.2.9 tiup telemetry

TiDB、TiUP 及 TiDB Dashboard 默认会收集使用情况信息，并将这些信息分享给 PingCAP 用于改善产品，例如，通过这些使用信息，PingCAP 可以了解常见的 TiDB 集群操作，从而确定新功能优先级。

当 TiUP 遥测功能开启时，执行 TiUP 命令时会将使用情况信息分享给 PingCAP，包括（但不限于）：

- 随机生成的遥测标示符
- TiUP 命令的执行情况，如命令执行是否成功、命令执行耗时等
- 使用 TiUP 进行部署的情况，如部署的目标机器硬件信息、组件版本号、修改过的部署配置名称等

以下信息不会被分享：

- 集群的准确名字
- 集群的拓扑结构
- 集群的配置文件

TiUP 使用命令 `tiup telemetry` 来控制遥测。

> 注意：
>
> 该功能默认是打开的。

语法

```
tiup telemetry <command>
```

`<command>` 代表子命令，支持的子命令列表请参考下方命令一节。

命令

status

命令 `tiup telemetry status` 查看当前的遥测设置，输出以下信息：

- `status`: 当前是否开启遥测 `(enable|disable)`
- `uuid`: 随机生成的遥测标示符

reset

命令 `tiup telemetry reset` 重置当前的遥测标示符，以一个新的随机标识符代替之。

enable

命令 `tiup telemetry enable` 启用遥测。

disable

命令 `tiup telemetry disable` 停用遥测。

<< 返回上一页 - TiUP 命令清单

12.4.7.2.10 tiup uninstall

命令 `tiup uninstall` 用于卸载已安装的组件。

语法

```
tiup uninstall <component1>:<version> [component2...N] [flags]
```

- `<component1>` 表示要卸载的组件名字
- `<version>` 表示要卸载的版本，如果省略，则表示卸载该组件的全部已安装版本，因为安全原因，省略 `<version>` 时必须加上选项 `--all` 明确表示需要卸载该组件的所有版本
- `[component2...N]` 表示可指定卸载多个组件或版本

选项

–all

- 卸载指定组件的全部已安装版本，省略 `<version>` 时使用。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–self

- 卸载 TiUP 自身：删除所有从镜像上下载过来的数据，但会保留 TiUP 及其组件产生的数据，数据存放在 `TIUP_HOME` 环境变量指定的目录中，若未设置过 `TIUP_HOME`，则默认值为 `~/.tiup/`。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 正常退出：`Uninstalled component "%s" successfully!`
- 若未指定 `<version>` 也未指定 `--all`：报错 `Use "tiup uninstall tidbx --all" if you want to remove all versions.`

<< 返回上一页 - TiUP 命令清单

12.4.7.2.11 tiup update

命令 `tiup update` 用于升级已安装的组件或者自身。

语法

```
tiup update [component1][:version] [component2..N] [flags]
```

- `[component1]` 表示要升级的组件名字
- `[version]` 表示要升级的版本，如果省略，则表示升级到该组件的最新稳定版本

- `[component2...N]` 表示可指定升级多个组件或版本。如果一个组件也不指定：即 `[component1][:version ] [component2..N]` 为空，则需要配合使用 `--all` 选项或 `--self` 选项。

升级操作不会删除旧的版本，仍然可以在执行时指定旧版本使用。

选项

–all

- 若未指定任何组件，则必须指定该选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–force

- 若指定的组件版本已经安装，则默认跳过升级操作，指定该参数可强制升级已安装版本。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–nightly

- 将指定组件升级到 nightly 版本。使用该参数的命令等价于 `tiup update <component>:nightly`。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–self

- 升级 TiUP 自身。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 升级成功：`Updated successfully!`
- 目标版本不存在：`Error: version %s not supported by component %s`

<< 返回上一页 - TiUP 命令清单

12.4.7.3 TiUP Cluster 命令

12.4.7.3.1 TiUP Cluster

TiUP Cluster 是 TiUP 提供的使用 Golang 编写的集群管理组件，通过 TiUP Cluster 组件就可以进行日常的运维工作，包括部署、启动、关闭、销毁、弹性扩缩容、升级 TiDB 集群、管理 TiDB 集群参数。

语法

```
tiup cluster [command] [flags]
```

`[command]` 代表命令名字，支持的命令列表请参考下方命令清单。

选项

–ssh (string，默认 builtin)

指定 SSH 客户端连接远端（部署 TiDB 服务的机器）执行命令，支持以下值：

- builtin：使用 tiup-cluster 内置的 easyssh 客户端
- system：使用当前操作系统默认的 SSH 客户端
- none：不使用 ssh 客户端，这种方式只支持部署到当前机器

–ssh-timeout（uint，默认 5）

设置 SSH 连接超时时间，单位为秒。

–wait-timeout（uint，默认 120）

运维过程中涉及到很多操作：指定 systemctl 启动/停止服务，等待端口上线/下线等，每个操作可能会消耗数秒。`--wait-timeout` 用于设置每个步骤的最长等待时间（单位为秒），超时后报错退出。

-y, –yes

- 跳过所有风险操作的二次确认，除非是使用脚本调用 TiUP，否则不推荐使用。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-v, –version

- 输出 TiUP Cluster 当前版本信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出相关命令的帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

命令清单

- import：导入 ansible 部署的集群

- template：输出拓扑模版
- check：部署前后的集群检查
- deploy：根据指定拓扑部署集群
- list：查询已部署的集群列表
- display：展示指定集群状态
- start：启动指定集群
- stop：停止指定集群
- restart：重启指定集群
- scale-in：缩容指定集群
- scale-out：扩容指定集群
- upgrade：升级指定集群
- prune：销毁指定集群中状态为 Tombstone 的实例
- edit-config：修改指定集群配置
- reload：重载指定集群配置
- patch：替换已部署集群的某个服务
- rename：重命名集群
- clean：删除指定集群数据
- destroy：销毁指定集群
- audit：查询集群操作审计日志
- replay：重试指定命令
- enable：开启指定集群或服务开机自启动
- disable：关闭指定集群或服务开机自启动
- help：输出帮助信息

<< 返回上一页 - TiUP 组件清单

12.4.7.3.2 tiup cluster audit

命令 `tiup cluster audit` 可以用于查看历史上对所有集群执行了什么命令，以及每个命令的执行日志。

语法

```
tiup cluster audit [audit-id] [flags]
```

- 若不填写 `[audit-id]` 则按时间倒序输出操作记录的表格，第一列为 audit-id
- 若填写 `[audit-id]` 则查看指定的 audit-id 的执行日志

选项

-h, –help

- 输出帮助信息。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 若指定 `[audit-id]`，则输出对应的执行日志
- 若不指定 `[audit-id]` 则输出含有以下字段的表格：
  - ID：该条记录对应的 audit-id
  - Time：该条记录对应的命令执行时间
  - Command：该条记录对应的命令

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.3 tiup cluster check

对于严肃的生产环境，在正式上线之前需要进行一系列检查，来确保集群拥有最好的表现。为了简化人工检查的步骤，TiUP Cluster 提供了 check 子命令，用于检查指定集群的机器硬件和软件环境是否满足正常运行条件。

检查项列表

操作系统版本

检查部署机操作系统发行版和版本：目前仅支持部署在 CentOS 7 的操作系统上，之后随兼容性改进可能支持更多系统版本。

CPU EPOLLEXCLUSIVE

检查部署机 CPU 是否支持 EPOLLEXCLUSIVE。

numactl

检查部署机是否安装 numactl，若用户配置绑核，则必须安装 numactl。

系统时间

检查部署器系统时间是否同步：将部署机系统时间与中控机对比，偏差超出某一阈值（500ms）后报错。

时间同步服务

检查部署机是否配置了时间同步服务：即 ntpd 是否在运行

Swap 分区

检查部署机是否启用 Swap 分区：建议禁用 Swap 分区

内核参数

检查各项内核参数的值：

- net.ipv4.tcp_tw_recycle: 0
- net.ipv4.tcp_syncookies: 0
- net.core.somaxconn: 32768
- vm.swappiness: 0
- vm.overcommit_memory: 0 或 1
- fs.file-max: 1000000

THP（透明大页）

检查部署机是否启用透明大页：建议禁用透明大页。

系统限制

检查 /etc/security/limits.conf 中各项 limit 值：

```
<deploy-user>    soft    nofile    1000000
<deploy-user>    hard    nofile    1000000
<deploy-user>    soft    stack     10240
```

其中 <deploy-user> 为部署、运行 TiDB 集群的用户，最后一列的数值为要求达到的最小值。

SELinux

检查 SELinux 是否启用：建议用户禁用 SELinux。

防火墙

检查 FirewallD 服务是否启用：建议用户禁用 FirewallD 或为 TiDB 集群各服务添加允许规则。

irqbalance

检查 irqbalance 服务是否启用：建议用户启用 irqbalance 服务。

磁盘挂载参数

检查 ext4 分区的挂载参数：确保挂载参数包含 nodelalloc,noatime 选项。

端口占用

检查部署机上是否已有进程占用了端口：检查拓扑中定义的端口（包括自动补全的默认端口）在部署机上是否已被占用。

> 注意：
>
> 端口占用检查假设集群尚未启动，如果检查的是已经部署并启动的集群，那么端口占用检查一定会失败，因为端口确实被占用了。

CPU 核心数

检查部署机 CPU 信息：建议生产集群 CPU 逻辑核心数 >= 16

> 注意：
>
> 默认不检查 CPU 核心数，需要通过选项 --enable-cpu 启用。

内存大小

检查部署机的内存大小：建议生产集群总内存容量 >= 32Gb。

> **注意：**
>
> 默认不检查内存大小，需要通过选项 `--enable-mem` 启用。

fio 磁盘性能测试

使用 fio 测试 data_dir 所在磁盘的性能，包括三个测试项目：

- fio_randread_write_latency
- fio_randread_write
- fio_randread

> **注意：**
>
> 默认不进行 fio 磁盘性能测试，需要通过选项 `--enable-disk` 启用。

语法

```
tiup cluster check <topology.yml | cluster-name> [flags]
```

若集群尚未部署，需要传递将用于部署集群的topology.yml 文件，tiup-cluster 会根据该文件的内容连接到对应机器去检查。若集群已经部署，则可以使用集群的名字 `<cluster-name>` 作为检查对象。

> **注意：**
>
> 若传递的是集群名字，则需要配合 `--cluster` 选项使用。

选项

–apply

- 尝试自动修复失败的检查项，目前仅会尝试修复以下项目：
  - SELinux
  - 防火墙
  - irqbalance
  - 内核参数
  - 系统 Limits
  - THP（透明大页）
- 数据类型：`BOOLEAN`

- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–cluster

tiup-cluster 支持对未部署的集群进行检查，也支持对已部署的集群进行检查，命令格式：

```
tiup cluster check <topology.yml | cluster-name> [flags]
```

若选择的格式为 `tiup cluster check <cluster-name>` 则必须加上该选项：`tiup cluster check <cluster-name> --cluster`。

该选项的数据类型为 BOOLEAN。该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-N, –node

- 指定要检查的节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为`tiup-component-cluster-display` 命令返回的集群状态表格的第一列。
- 数据类型：`STRINGS`
- 如果不指定该选项，默认检查所有节点，即 `[]`。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将检查它们的交集中的服务。

-R, –role

- 指定要检查的角色。该选项的值为以逗号分割的节点角色列表，角色为`tiup-component-cluster-display` 命令返回的集群状态表格的第二列。
- 数据类型：`STRINGS`
- 如果不指定该选项，默认检查所有角色。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将检查它们的交集中的服务。

–enable-cpu

- 默认情况下 tiup-cluster 不检查 CPU 核心数，该选项用于启用 CPU 核心数检查。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–enable-disk

- 默认情况下 tiup-cluster 不进行 fio 磁盘性能测试，该选项用于启用 fio 磁盘性能测试。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–enable-mem

默认情况下 tiup-cluster 不检查内存大小，该选项用于启用内存大小检查。

-u, –user（string，默认为当前执行命令的用户）

指定连接目标机器的用户名，该用户在目标机器上需要有免密 sudo root 的权限。

> 注意：
>
> 仅当 `--cluster` 选项为 false 时该选项有效，否则该值固定为部署集群时拓扑文件中指定的用户名。

-i, –identity_file（string，默认 ~/.ssh/id_rsa）

指定连接目标机器的密钥文件。

> 注意：
>
> 仅当 `--cluster` 选项为 false 时该选项有效，否则该值固定为 `${TIUP_HOME}/storage/cluster/clusters/<cluster-name>/ssh/id_rsa`

-p, –password

- 在连接目标机器时使用密码登录：
  - 对于指定了 `--cluster` 的集群，密码为部署集群时拓扑文件中指定的用户的密码
  - 对于未指定 `--cluster` 的集群，密码为 `-u/--user` 参数指定的用户的密码
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

输出含有以下字段的表格：

- Node：目标节点
- Check：检查项
- Result：检查结果（Pass/Warn/Fail）
- Message：结果描述

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.4 tiup cluster clean

在测试环境中，有时候需要将集群重置回刚部署的状态，即删除所有数据，命令 `tiup cluster clean` 可以很方便的做到这一点：它会停止集群，然后删除集群上的数据。手工重启集群之后，就能得到一个全新的集群了。

> 警告：
>
> 该命令一定会先停止集群（即使选择只清理日志也是），生产环境请勿使用。

语法

```
tiup cluster clean <cluster-name> [flags]
```

`<cluster-name>` 为要清理的集群。

选项

–all

- 同时清理数据和日志，等价于同时指定 `--data` 和 `--log`，若不指定该选项，则必须至少指定以下选项之一：
  - –data：清理数据
  - –log：清理日志
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–data

该选项开启数据清理，若不指定该选项，也不指定 `--all`，则不清理数据。

–log

- 该选项开启日志清理，若不指定该选项，也不指定 `--all`，则不清理日志。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–ignore-node（StringArray，默认为空）

指定不需要清理的节点，如需指定多个，重复使用多次该选项：`--ignore-node <node-A> --ignore-node <node-B>`。

–ignore-role（StringArray，默认为空）

指定不需要清理的角色，如需指定多个，重复使用多次该选项：`--ignore-role <role-A> --ignore-role <role-B>`。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.5 tiup cluster deploy

命令 `tiup cluster deploy` 用于部署一个全新的集群。

语法

```
tiup cluster deploy <cluster-name> <version> <topology.yaml> [flags]
```

- `<cluster-name>` 表示新集群的名字，不能和现有集群同名
- `<version>` 为要部署的 TiDB 集群版本号，如 `v5.4.0`
- `<topology.yaml>` 为事先编写好的拓扑文件

选项

-u, –user（string，默认为当前执行命令的用户）

指定连接目标机器的用户名，该用户在目标机器上需要有免密 sudo root 的权限。

-i, –identity_file（string，默认 ~/.ssh/id_rsa）

指定连接目标机器的密钥文件。

-p, –password

- 在连接目标机器时使用密码登录，不可和 `-i/--identity_file` 同时使用。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–ignore-config-check

- 在组件二进制文件部署之后，TiUP 会对 TiDB，TiKV 和 PD 组件执行配置检查，检查方式为 `<binary> --config-check <config-file>`，其中 `<binary>` 为部署的二进制文件的路径，`<config-file>` 为根据用户配置生成的配置文件。如果想要跳过该项检查，可以使用该选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–no-labels

- 当两个或多个 TiKV 部署到同一台机器时，会存在一个风险：由于 PD 无法感知集群的拓扑结构，可能将一个 Region 的多个副本调度到一台物理机上的不同 TiKV，这样这台物理机就成为了单点。为了避免这种情况，用户可以通过 label 来告诉 PD 不要将相同的 Region 调度到同一台机器上（配置方式参考通过拓扑 label 进行副本调度）。
- 但是对于测试环境，可能并不在意是否将一个 Region 的副本调度到了同一台机器上，这个时候可以使用 `--no-labels` 来绕过检查。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–skip-create-user

- 在部署集群时，tiup-cluster 会先检查拓扑文件中指定的用户名是否存在，如果不存在就会创建一个。指定 `--skip-create-user` 选项后不再检查用户是否存在，直接跳过创建步骤。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

部署日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.6 tiup cluster destroy

当业务下线之后，如果想将集群占有的机器释放出来让给其他业务使用，需要清理掉集群上的数据以及部署的二进制文件。`tiup cluster destroy` 命令会执行以下操作销毁集群：

- 停止集群
- 对于每个服务，删除其日志目录，部署目录，数据目录
- 如果各个服务的数据目录/部署目录的父目录是由 tiup-cluster 创建的，也一并删除

语法

```
tiup cluster destroy <cluster-name> [flags]
```

<cluster-name> 为要销毁的集群名字。

选项

–force

- 在某些情况下，有可能集群中的某些节点已经宕机，导致无法通过 SSH 连接到节点进行操作，这个时候可以通过 --force 选项忽略这些错误。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

–retain-node-data（StringArray，默认为空）

指定需要保留数据的节点，如需指定多个，重复使用多次该选项：--retain-node-data <node-A> --retain-node-data <node-B>。

–retain-role-data（StringArray，默认为空）

指定需要保留数据的角色，如需指定多个，重复使用多次该选项：--retain-role-data <role-A> --retain-role-data <role-B>。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.7 tiup cluster disable

命令 tiup cluster disable 用于关闭集群服务所在机器重启之后的自启动，该命令会到指定的节点上去执行 systemctl disable <service> 来关闭服务的自启动。

语法

```
tiup cluster disable <cluster-name> [flags]
```

<cluster-name> 为要关闭自启的集群。

选项

-N, –node

- 指定要关闭自启的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为tiup-component-cluster-display 命令返回的集群状态表格的第一列。

- 数据类型：STRINGS
- 如果不指定该选项，默认关闭所有节点的自启。

> 注意：
>
> 若同时指定了 -R, --role，那么将关闭它们的交集中的服务自启。

-R, –role

- 指定要关闭自启的角色，该选项的值为以逗号分割的节点角色列表，角色为tiup-component-cluster-display 命令返回的集群状态表格的第二列。
- 数据类型：STRINGS
- 如果不指定该选项，默认关闭所有角色的自启。

> 注意：
>
> 若同时指定了 -N, --node，那么将关闭它们的交集中的服务自启。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.8 tiup cluster display

如果想查看集群中每个组件的运行状态，逐一登录到各个机器上查看显然很低效。因此，tiup-cluster 提供了 tiup cluster display 命令来高效完成这件工作。

语法

```
tiup cluster display <cluster-name> [flags]
```

<cluster-name> 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。

选项

–dashboard

- 默认情况会展示整个集群的所有节点信息，加上该选项后仅展示 dashboard 的信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-N, –node（strings，默认为 []，表示所有节点）

指定要查询的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，如果不确定要查询节点的 ID，不指定此选项，输出会显示所有节点的 ID 和状态信息。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将查询它们的交集中的服务状态。

-R, –role（strings，默认为 []，表示所有角色）

指定要查询的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，如果不确定要查询节点的角色，不指定此选项，输出会显示所有节点的角色和状态信息。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将查询它们的交集中的服务状态。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 集群名称
- 集群版本
- SSH 客户端类型
- Dashboard 地址
- 含有以下字段的表格：
  - ID：节点 ID，由 `IP:PORT` 构成
  - Role：该节点部署的服务角色（如 TiDB、TiKV 等）
  - Host：该节点对应的机器 IP
  - Ports：服务占用的端口号
  - OS/Arch：该节点的操作系统和机器架构
  - Status：该节点服务当前的状态
  - Data Dir：服务的数据目录，- 表示没有数据目录
  - Deploy Dir：服务的部署目录

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.9 tiup cluster edit-config

在部署集群之后，如果需要再调整集群服务的配置，则可以使用命令 `tiup cluster edit-config`，它会启动一个编辑器（默认为 $EDITOR 环境变量指定的值，当 EDITOR 环境变量不存在时，使用 vi 打开）允许用户修改指定集群的拓扑文件。

> 注意：
>
> - 修改配置时不能增删机器，增删机器属于集群扩容和集群缩容的功能。
> - 执行完该命令后配置只是在中控机上修改了，要应用配置需要执行 `tiup cluster reload` 命令来重新加载。

语法

```
tiup cluster edit-config <cluster-name> [flags]
```

`<cluster-name>` 代表要操作的集群名。

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 正常情况无输出
- 若修改了不能修改的字段，则保存文件时报错并提示用户重新编辑，不能修改的字段参考拓扑文件中的相关描述

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.10 tiup cluster enable

命令 `tiup cluster enable` 用于设置集群服务在机器重启后的自启动，该命令会到指定的节点上去执行 `systemctl enable <service>` 来开启服务的自启。

> 注意：
>
> 集群全部关机重启时，服务启动的顺序由节点的操作系统启动顺序决定。重启顺序不正确时，在某些情况下，重启后的集群仍然不能提供服务（比如先启动了 TiKV 但未启动 PD，systemd 重启 TiKV 多次仍未找到 PD 则会放弃）。

语法

```
tiup cluster enable <cluster-name> [flags]
```

<cluster-name> 为要启用自启的集群。

选项

-N, –node

- 指定要开启自启的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为tiup-component-cluster-display 命令返回的集群状态表格的第一列。
- 数据类型：STRINGS
- 如果不指定该选项，默认开启所有节点的自启。

> 注意：
>
> 若同时指定了 -R, --role，那么将开启它们的交集中的服务自启。

-R, –role

- 指定要开启自启的角色，该选项的值为以逗号分割的节点角色列表，角色为tiup-component-cluster-display 命令返回的集群状态表格的第二列。
- 数据类型：STRINGS
- 如果不指定该选项，默认开启所有角色的自启。

> 注意：
>
> 若同时指定了 -N, --node，那么将开启它们的交集中的服务自启。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.11 tiup cluster help

tiup-cluster 在命令行界面为用户提供了丰富的帮助信息，这些帮助信息可以通过 help 命令或者 --help 参数获得。tiup cluster help <command> 基本等价于 tiup cluster <command> --help。

语法

```
tiup cluster help [command] [flags]
```

[command] 用于指定要查看哪个命令的帮助信息，若不指定，则查看 tiup-cluster 自身的帮助信息。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

[command] 或 tiup-cluster 的帮助信息。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.12 tiup cluster import

在 TiDB 4.0 以前的版本，集群多是通过 TiDB Ansible 部署的。TiUP Cluster 提供了 import 命令用于让这些集群过渡到使用 tiup-cluster 组件管理。

注意：

- TiDB Ansible 配置导入到 TiUP 中管理后，不能再通过 TiDB Ansible 对集群进行操作，否则可能因元信息不一致造成冲突
- 如果使用 TiDB Ansible 部署的集群存在以下情况之一，暂不支持导入：
  - 启用了 TLS 加密功能的集群
  - 纯 KV 集群（没有 TiDB 实例的集群）
  - 启用了 Kafka 的集群
  - 启用了 Spark 的集群
  - 启用了 TiDB Lightning/ Importer 的集群
  - 仍使用老版本 push 的方式收集监控指标（从 v3.0 起默认为 pull 模式，如果没有特意调整过则可以支持）
  - 在 inventory.ini 配置文件中单独为机器的 node_exporter / blackbox_exporter 通过 node_exporter_port/blackbox_exporter_port 设置了非默认端口（在 group_vars 目录中统一配置的可以兼容）
- 如果使用 TiDB Ansible 部署的集群中有部分节点未部署监控，应当先使用 TiDB Ansible 在 inventory.ini 文件的 monitored_servers 分组中补充对应节点的信息，并通过 deploy.yaml playbook 将补充的监控组件部署完整。否则在集群数据导入 TiUP 后进行其他运维操作时，可能会因监控组件缺失而出错。

语法

```
tiup cluster import [flags]
```

选项

-d, –dir string（string, 默认当前目录）

指定 TiDB Ansible 所在目录。

–ansible-config（string，默认 “./ansible.cfg”）

指定 Ansible 的配置文件路径。

–inventory string（string，默认 “inventory.ini”）

指定 Ansible inventory 文件的名字。

–no-backup

- 默认情况下，导入成功之后会将 `--dir` 指定的目录里所有内容备份到 `${TIUP_HOME}/.tiup/storage/cluster/clusters/{cluster-name}/ansible-backup` 下。该选项用于禁用默认的备份步骤，如果该目录下有多个 inventory 文件（部署了多个集群），推荐禁用默认备份。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–rename（string，默认为空）

重命名导入的集群。默认集群名为 inventory 中指定的 cluster_name。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

导入过程的日志信息。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.13 tiup cluster list

tiup-cluster 支持使用同一个中控机部署多套集群，而命令 `tiup cluster list` 可以查看当前登录的用户使用该中控机部署了哪些集群。

> 注意：
>
> 部署的集群数据默认放在 `~/.tiup/storage/cluster/clusters/` 目录下，因此在同一台中控机上，当前登录用户无法查看其他用户部署的集群。

语法

```
tiup cluster list [flags]
```

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

输出含有以下字段的表格：

- Name：集群名字
- User：部署用户
- Version：集群版本
- Path：集群部署数据在中控机上的路径
- PrivateKey：连接集群的私钥所在路径

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.14 tiup cluster patch

在集群运行过程中，如果需要动态替换某个服务的二进制文件（即替换过程中保持集群可用），那么可以使用 tiup cluster patch 命令，它会完成以下几件事情：

- 将用于替换的二进制包上传到目标机器
- 如果目标服务是 TiKV、TiFlash 或者 TiDB Binlog 之类的存储服务，则先通过 API 下线节点
- 停止目标服务
- 解压二进制包，替换服务
- 启动目标服务

语法

```
tiup cluster patch <cluster-name> <package-path> [flags]
```

- <cluster-name> 代表要操作的集群名
- <package-path> 为用于替换的二进制包，其打包方式如下：
  - 确定当前要替换的组件名称 ${component} (tidb, tikv, pd…) 以及其版本 ${version} (v4.0.0, v4.0.1 …)，以及其运行的平台 ${os} (linux) 和 ${arch} (amd64, arm64)
  - 下载当前的组件包：wget https://tiup-mirrors.pingcap.com/${component}-${version}-${os}-${arch}.tar.gz -O /tmp/${component}-${version}-${os}-${arch}.tar.gz
  - 建立临时打包目录：mkdir -p /tmp/package && cd /tmp/package

- 解压原来的二进制包：`tar xf /tmp/${component}-${version}-${os}-${arch}.tar.gz`
- 查看临时打包目录中的文件结构：`find .`
- 将要替换的二进制文件或配置文件复制到临时目录的对应位置
- 重新打包 `tar czf /tmp/${component}-hotfix-${os}-${arch}.tar.gz *`
- 通过以上步骤之后，`/tmp/${component}-hotfix-${os}-${arch}.tar.gz` 就可以用于 patch 命令了

选项

–overwrite

- 对某个组件（比如 TiDB，TiKV）进行 patch 后，如果要在该集群扩容该组件，tiup-cluster 会默认使用 patch 前的版本。如果希望后续扩容的时候也使用 patch 之后的版本，需要指定 `--overwrite` 选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–transfer-timeout（uint，默认 300）

在重启 PD 或 TiKV 时，会先将被重启节点的 leader 迁移到其他节点，迁移过程会需要一定时间，可以通过设置 `--transfer-timeout` 设置最长等待时间（单位为秒），超时之后会跳过等待直接重启服务。

> 注意：
> 若出现跳过等待直接重启的情况，服务性能可能会出现抖动。

-N, –node（strings，默认为 []，未选中任何节点）

指定要替换的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
> 若同时指定了 `-R, --role`，那么将替换它们的交集中的服务。

-R, –role（strings，默认为 []，未选中任何角色）

指定要替换的角色，该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> 注意：
> 若同时指定了 `-N, --node`，那么将替换它们的交集中的服务。

–offline

- 声明当前集群处于停止状态。指定该选项时，TiUP Cluster 仅原地替换集群组件的二进制文件，不执行迁移 Leader 以及重启服务等操作。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.15 tiup cluster prune

在缩容集群时，对于某些组件，并不会立即停止服务并删除数据，而是需要等数据调度完成之后，用户手动执行 `tiup cluster prune` 命令清理。

语法

```
tiup cluster prune <cluster-name> [flags]
```

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

清理过程的日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.16 tiup cluster reload

在修改集群配置之后，需要通过 `tiup cluster reload` 命令让集群重新加载配置才会生效，该命令会将中控机的配置发布到服务运行的远端机器，并按照升级的流程按顺序重启服务，重启过程中集群仍然可用。

语法

```
tiup cluster reload <cluster-name> [flags]
```

<cluster-name> 代表要操作的集群名。

选项

–force

- 忽略重新加载过程中的错误，强制 reload。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

–transfer-timeout（uint，默认 300）

在重启 PD 或 TiKV 时，会先将被重启节点的 leader 迁移到其他节点，迁移过程会需要一定时间，可以通过设置 --transfer-timeout 设置最长等待时间（单位为秒），超时之后会跳过等待直接重启服务。

> 注意：
>
> 若出现跳过等待直接重启的情况，服务性能可能会出现抖动。

–ignore-config-check

- 在组件二进制文件部署之后，TiUP 会对 TiDB，TiKV 和 PD 组件执行配置检查，检查方式为 <binary> --config-check <config-file>，其中 <binary> 为部署的二进制文件的路径，<config-file> 为根据用户配置生成的配置文件。如果想要跳过该项检查，可以使用该选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

-N, –node（strings，默认为 []，表示所有节点）

指定要重启的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
>
> - 若同时指定了 -R, --role，那么将重启它们的交集中的服务
> - 若指定了选项 --skip-restart，则该选项无效

-R, –role（strings，默认为 []，表示所有角色）

指定要重启的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

注意：

1. 若同时指定了 -N，--node，那么将重启它们的交集中的服务
2. 若指定了选项 --skip-restart，则该选项无效

–skip-restart

- 命令 tiup cluster reload 会执行两个操作：
  - 刷新所有节点配置
  - 重启指定节点
- 该选项指定后仅刷新配置，不重启任何节点，这样刷新的配置也不会应用，需要等对应服务下次重启才会生效。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.17 tiup cluster rename

集群名字是部署集群时指定的，在集群部署之后，如果想更改集群名字，可以通过 tiup cluster rename 命令来实现。

注意：

如果配置了 grafana_servers 的 dashboard_dir 字段，在执行 tiup cluster rename 命令进行集群重命名后，需要额外做以下操作：

- 对于本地的 dashboards 目录中的 *.json 文件，将 datasource 字段的值更新为新的集群名（这是因为 datasource 是以集群名命名的）
- 执行 tiup cluster reload -R grafana 命令

语法

```
tiup cluster rename <old-cluster-name> <new-cluster-name> [flags]
```

- <old-cluster-name> 老的集群名
- <new-cluster-name> 新的集群名

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-cluster 的执行日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.18 tiup cluster replay

对集群进行升级或重启等操作时，操作有可能因为环境的原因而偶然失败。这时如果重新进行操作，需要从头开始执行所有步骤。如果集群规模较大，会耗费较长时间。此时可以使用 tiup cluster replay 命令重试刚才失败的命令，并且跳过已经成功的步骤。

语法

```
tiup cluster replay <audit-id> [flags]
```

- <audit-id> 代表要重试的命令对应的 audit-id。使用 tiup cluster audit 可查看历史命令及其 audit-id。

选项

-h, –help

输出帮助信息。

输出

<audit-id> 对应的命令的输出。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.19 tiup cluster restart

命令 tiup cluster restart 用于重启指定集群的所有或部分服务。

> **注意：**
>
> 重启过程中会有一段时间服务不可用。

语法

```
tiup cluster restart <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要重启的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> **注意：**
>
> 若同时指定了 `-R, --role`，那么将重启它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要重启的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> **注意：**
>
> 若同时指定了 `-N, --node`，那么将重启它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

重启服务的日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.20 tiup cluster scale-in

`tiup cluster scale-in` 命令用于集群缩容，缩容即下线服务，最终会将指定的节点从集群中移除，并删除遗留的相关文件。

下线特殊处理

由于 TiKV，TiFlash 和 TiDB Binlog 组件的下线是异步的（需要先通过 API 执行移除操作）并且下线过程耗时较长（需要持续观察节点是否已经下线成功），所以对 TiKV，TiFlash 和 TiDB Binlog 组件做了特殊处理：

- 对 TiKV，TiFlash 及 TiDB Binlog 组件的操作:
  - tiup-cluster 通过 API 将其下线后直接退出而不等待下线完成
  - 执行 `tiup cluster display` 查看下线节点的状态，等待其状态变为 Tombstone
  - 执行 `tiup cluster prune` 命令清理 Tombstone 节点，该命令会执行以下操作：
    * 停止已经下线掉的节点的服务
    * 清理已经下线掉的节点的相关数据文件
    * 更新集群的拓扑，移除已经下线掉的节点
- 对其他组件的操作
  - 下线 PD 组件时，会通过 API 将指定节点从集群中删除掉（这个过程很快），然后停掉指定 PD 的服务并且清除该节点的相关数据文件
  - 下线其他组件时，直接停止并且清除节点的相关数据文件

语法

```
tiup cluster scale-in <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。

选项

-N, –node（strings，无默认值，必须非空）

选择要缩容的节点，若缩容多个节点，以逗号分割。

–force

- 在某些情况下，有可能被缩容的节点宿主机已经宕机，导致无法通过 SSH 连接到节点进行操作，这个时候可以通过 `--force` 选项强制将其从集群中移除。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

> 警告：
>
> 使用该选项强制移除正在服务和下线中的 TiKV / TiFlash 节点时，这些节点会被直接删除，不等待数据调度完成，因此这个场景下存在较大概率丢失数据的风险。不建议对未宕机的节点使用该选项。

–transfer-timeout（uint，默认 300）

在缩容 PD 或 TiKV 时，会先将被缩容节点的 leader 迁移到其他节点，迁移过程会需要一定时间，可以通过设置 `--transfer-timeout` 设置最长等待时间（单位为秒），超时之后会跳过等待直接缩容服务。

> 注意：
>
> 若出现跳过等待直接缩容的情况，服务性能可能会出现抖动。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

缩容日志

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.21 tiup cluster scale-out

`tiup cluster scale-out` 命令用于集群扩容，扩容的内部逻辑与部署类似，tiup-cluster 组件会先建立新节点的 SSH 连接，在目标节点上创建必要的目录，然后执行部署并且启动服务。其中 PD 节点的扩容会通过 join 方式加入到集群中，并且会更新与 PD 有关联的服务的配置；其他服务直接启动加入到集群中。

语法

```
tiup cluster scale-out <cluster-name> <topology.yaml> [flags]
```

- `<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看
- `<topology.yaml>` 为事先编写好的扩容拓扑文件，该文件应当仅包含扩容部分的拓扑

选项

-u, –user（string，默认为当前执行命令的用户）

指定连接目标机器的用户名，该用户在目标机器上需要有免密 sudo root 的权限。

-i, –identity_file（string，默认 ~/.ssh/id_rsa）

指定连接目标机器的密钥文件。

-p, –password

- 在连接目标机器时使用密码登录，不可和 `-i/--identity_file` 同时使用。
- 数据类型：BOOLEAN

- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–no-labels

- 当两个或多个 TiKV 部署到同一台机器时，会存在一个风险：由于 PD 无法感知集群的拓扑结构，可能将一个 Region 的多个副本调度到一台物理机上的不同 TiKV，这样这台物理机就成为了单点。为了避免这种情况，用户可以通过 label 来指定 PD 不要将相同的 Region 调度到同一台机器上（配置方式参考通过拓扑 label 进行副本调度）。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

但是对于测试环境，可能并不在意是否将一个 Region 的副本调度到了同一台机器上，这个时候可以使用 `--no-labels` 来绕过检查。

–skip-create-user

在扩容集群时，tiup-cluster 会先检查拓扑文件中指定的用户名是否存在，如果不存在就会创建一个。指定 `--skip-create-user` 选项后不再检查用户是否存在，直接跳过创建步骤。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

扩容日志

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.22 tiup cluster start

命令 `tiup cluster start` 用于启动指定集群的所有或部分服务。

语法

```
tiup cluster start <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要启动的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将启动它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要启动的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将启动它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

启动日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.23 tiup cluster stop

命令 `tiup cluster stop` 用于停止指定集群的所有服务或部分服务。

> 注意：
>
> 核心服务停止后集群将无法提供服务。

语法

```
tiup cluster stop <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要停止的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将停止它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要停止的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将停止它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

停止服务的日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.24 tiup cluster template

部署集群之前，需要准备一份集群的拓扑文件。TiUP 内置了拓扑文件的模版，用户可以通过修改该模版来生成最终的拓扑文件。使用 `tiup cluster template` 命令可以输出 TiUP 内置的模版内容。

语法

```
tiup cluster template [flags]
```

如果不指定该选项，输出的默认模版包含以下实例：

- 3 个 PD 实例
- 3 个 TiKV 实例
- 1 个 TiDB 实例
- 1 个 Prometheus 实例
- 1 个 Grafana 实例
- 1 个 Alertmanager 实例

选项

–full

- 输出详细的拓扑模版，该模版会以注释的形式带上可配置的参数。在命令中添加该选项，可开启该选项。
- 如果不指定该选项，默认输出最简单的拓扑模版。

–multi-dc

- 输出多数据中心的拓扑模版。在命令中添加该选项，可开启该选项。
- 如果不指定该选项，默认输出单地单机房的拓扑模版。

-h, –help

输出帮助信息。

输出

根据指定选项输出拓扑模版，可重定向到拓扑文件中用于部署。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.3.25 tiup cluster upgrade

命令 `tiup cluster upgrade` 用于将指定集群升级到特定版本。

语法

```
tiup cluster upgrade <cluster-name> <version> [flags]
```

- `<cluster-name>` 为要操作的集群名字，如果忘记集群名字可通过集群列表查看。
- `<version>` 为要升级到的目标版本，目前仅允许升级到比当前集群更高的版本，不允许升级到比当前集群更低的版本，即不允许降级。同时也不允许升级成 nightly 版本

选项

–force

- 升级集群需要保证集群目前是启动的，在某些情况下，可能希望在集群未启动的状态下升级，这时候可以使用 `--force` 忽略升级过程的错误，强制替换二进制文件并启动集群。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

> 注意：
>
> 对正在提供服务的集群强制升级可能导致集群服务不可用。对于未启动的集群，升级成功后会自动启动集群。

–transfer-timeout（uint，默认 300）

在升级 PD 或 TiKV 时，会先将被升级节点的 leader 迁移到其他节点，迁移过程会需要一定时间，可以通过设置 `--transfer-timeout` 设置最长等待时间（单位为秒），超时之后会跳过等待直接升级服务。

> **注意：**
>
> 若出现跳过等待直接升级的情况，服务性能可能会出现抖动。

–ignore-config-check

- 在二进制文件更新之后，TiUP 会对 TiDB，TiKV 和 PD 组件执行配置检查，检查方式为 `<binary> --config-check <config-file>`，其中 `<binary>` 为新部署的二进制文件的路径，`<config-file>` 为根据用户配置生成的配置文件。如果想要跳过该项检查，可以使用该选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

–offline

- 声明当前集群处于停止状态。指定该选项时，TiUP Cluster 仅原地替换集群组件的二进制文件，不执行迁移 Leader 以及重启服务等操作。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

升级服务的日志。

<< 返回上一页 - TiUP Cluster 命令清单

12.4.7.4 TiUP DM 命令

12.4.7.4.1 TiUP DM

类似于TiUP Cluster 对 TiDB 集群的管理，TiUP DM 是用于对 DM 集群的日常运维工作，包括部署、启动、关闭、销毁、弹性扩缩容、升级 DM 集群、管理 DM 集群参数。

语法

```
tiup dm [command] [flags]
```

`[command]` 代表命令名字，支持的命令列表请参考下方命令清单。

选项

–ssh (string，默认 builtin)

指定 SSH 客户端连接远端（部署 TiDB 服务的机器）执行命令，支持以下几个值：

- builtin：使用 tiup-cluster 内置的 easyssh 客户端
- system：使用当前操作系统默认的 ssh 客户端
- none：不使用 ssh 客户端，这种方式只支持部署到当前机器

–ssh-timeout（uint，默认 5）

设置 ssh 连接超时时间，单位为秒。

–wait-timeout（uint，默认 120）

运维过程中涉及到很多操作：指定 systemctl 启动/停止服务，等待端口上线/下线等，每个操作可能会消耗数秒，`--wait-timeout` 用于设置每个步骤的最长等待时间（单位为秒），超时后报错退出。

-y, –yes

- 跳过所有风险操作的二次确认，除非是使用脚本调用 TiUP，否则不推荐使用。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-v, –version

- 输出 tiup-dm 当前版本信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出相关命令的帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

命令清单

- import：导入使用 DM-Ansible 部署的 DM v1.0 集群
- template：输出拓扑模版
- deploy：根据指定拓扑部署集群
- list：查询已部署的集群列表
- display：展示指定集群状态
- start：启动指定集群
- stop：停止指定集群
- restart：重启指定集群
- scale-in：缩容指定集群
- scale-out：扩容指定集群
- upgrade：升级指定集群
- prune：销毁指定集群中状态为 Tombstone 的实例
- edit-config：修改指定集群配置
- reload：重载指定集群配置

- patch：替换已部署集群的某个服务
- destroy：销毁指定集群
- audit：查询集群操作审计日志
- replay：重试指定命令
- enable：开启指定集群或服务开机自启动
- disable：关闭指定集群或服务开机自启动
- help：输出帮助信息

<< 返回上一页 - TiUP 组件清单

12.4.7.4.2 tiup dm audit

命令 `tiup dm audit` 可以用于查看执行在所有集群上的历史命令，以及每个命令的执行日志。

语法

```
tiup dm audit [audit-id] [flags]
```

- 若不填写 `[audit-id]` 则按时间倒序输出操作记录的表格，第一列为 audit-id
- 若填写 `[audit-id]` 则查看指定的 audit-id 的执行日志

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 若指定 `[audit-id]`，则输出对应的执行日志
- 若不指定 `[audit-id]` 则输出含有以下字段的表格：
  - ID：该条记录对应的 audit-id
  - Time：该条记录对应的命令执行时间
  - Command：该条记录对应的命令

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.3 tiup dm deploy

命令 `tiup dm deploy` 用于部署一个全新的集群。

语法

```
tiup dm deploy <cluster-name> <version> <topology.yaml> [flags]
```

- <cluster-name> 表示新集群的名字，不能和现有集群同名
- <version> 为要部署的 DM 集群版本号，如 v2.0.0
- <topology.yaml> 为事先编写好的拓扑文件

选项

-u, –user（string，默认为当前执行命令的用户）

指定连接目标机器的用户名，该用户在目标机器上需要有免密 sudo root 的权限。

-i, –identity_file（string，默认 ~/.ssh/id_rsa）

指定连接目标机器的密钥文件。

-p, –password

- 在连接目标机器时使用密码登录，不可和 -i/--identity_file 同时使用。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

部署日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.4 tiup dm destroy

当业务下线之后，如果想将集群占有的机器释放出来让给其他业务使用，需要清理掉集群上的数据以及部署的二进制文件。tiup dm destroy 命令会执行以下操作销毁集群：

- 停止集群
- 对于每个服务，删除其日志目录，部署目录，数据目录
- 如果各个服务的数据目录/部署目录的父目录是由 tiup-dm 创建的，也一并删除

语法

```
tiup dm destroy <cluster-name> [flags]
```

<cluster-name> 为要销毁的集群名字。

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 false。在命令中添加该选项，并传入 true 值或不传值，均可开启此功能。

输出

tiup-dm 的执行日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.5 tiup dm disable

集群服务所在的机器重启之后会自启动，命令 tiup dm disable 用于关闭该自启动。该命令会在指定的节点上执行 systemctl disable <service> 来关闭服务的自启动。

语法

```
tiup dm disable <cluster-name> [flags]
```

<cluster-name> 为要关闭自启的集群。

选项

-N, –node

- 指定要关闭自启的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。
- 数据类型：STRINGS
- 如果不指定该选项，默认关闭所有节点的自启。

> 注意：
>
> 若同时指定了 -R, --role，那么将关闭它们的交集中的服务自启。

-R, –role

- 指定要关闭自启的角色，该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。
- 数据类型：STRINGS
- 如果不指定该选项，默认关闭所有角色的自启。

> 注意：
>
> 若同时指定了 -N, --node，那么将关闭它们的交集中的服务自启。

-h, –help

输出帮助信息。

输出

tiup-dm 的执行日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.6 tiup dm display

如果想查看集群中每个组件的运行状态，逐一登录到各个机器上查看显然很低效。因此，tiup-dm 提供了 `tiup dm display` 命令来高效完成这件工作。

语法

```
tiup dm display <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要查询的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，如果不确定要查询节点的 ID，不指定此选项，输出会显示所有节点的 ID 和状态信息。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将查询它们的交集中的服务状态。

-R, –role（strings，默认为 []，表示所有角色）

指定要查询的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，如果不确定要查询节点的角色，不指定此选项，输出会显示所有节点的角色和状态信息。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将查询它们的交集中的服务状态。

-h, –help

- 输出帮助信息。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 集群名称
- 集群版本
- SSH 客户端类型
- 含有以下字段的表格：
  - ID：节点 ID，由 IP:PORT 构成
  - Role：该节点部署的服务角色（如 TiDB 和 TiKV 等）
  - Host：该节点对应的机器 IP
  - Ports：服务占用的端口号
  - OS/Arch：该节点的操作系统和机器架构
  - Status：该节点服务当前的状态
  - Data Dir：服务的数据目录，- 表示没有数据目录
  - Deploy Dir：服务的部署目录

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.7 tiup dm edit-config

在部署集群之后，如果需要再调整集群服务的配置，则可以使用命令 `tiup dm edit-config`，它会启动一个编辑器（默认为 `$EDITOR` 环境变量指定的值，当 `EDITOR` 环境变量不存在时，使用 `vi` 打开）允许用户修改指定集群的拓扑文件。

> 注意：
>
> - 修改配置时不能增删机器，增删机器属于集群扩容和集群缩容的功能。
> - 执行完该命令后配置只是在中控机上修改了，要应用配置需要执行 `tiup dm reload` 命令来重新加载。

语法

```
tiup dm edit-config <cluster-name> [flags]
```

`<cluster-name>` 代表要操作的集群名。

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

- 正常情况无输出

- 若修改了不能修改的字段，则保存文件时报错并提示用户重新编辑，不能修改的字段参考拓扑文件中的相关描述

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.8 tiup dm enable

命令 `tiup dm enable` 用于设置集群服务在机器重启后的自启动，该命令会到指定的节点上去执行 `systemctl enable <service>` 来开启服务的自启。

语法

```
tiup dm enable <cluster-name> [flags]
```

`<cluster-name>` 为要启用自启的集群。

选项

-N, –node

- 指定要开启自启的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。
- 数据类型：`STRINGS`
- 如果不指定该选项，默认开启所有节点的自启。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将开启它们的交集中的服务自启。

-R, –role

- 指定要开启自启的角色，该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。
- 数据类型：`STRINGS`
- 如果不指定该选项，默认开启所有角色的自启。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将开启它们的交集中的服务自启。

-h, –help

输出帮助信息。

输出

tiup-dm 的执行日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.9 tiup dm help

tiup-dm 在命令行界面为用户提供了丰富的帮助信息，这些帮助信息可以通过 `help` 命令或者 `--help` 参数获得。基本上，`tiup dm help <command>` 等价于 `tiup dm <command> --help`。

语法

```
tiup dm help [command] [flags]
```

`[command]` 用于指定要查看哪个命令的帮助信息，若不指定，则查看 tiup-dm 自身的帮助信息。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

`[command]` 或 tiup-dm 的帮助信息。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.10 tiup dm import

在 DM 1.0 版本，集群基本是通过 TiDB Ansible 部署的，TiUP DM 提供了 `import` 命令导入 v1.0 的集群并重新部署 v2.0 的集群。

> **注意：**
>
> - 不支持导入 v1.0 集群中的 DM Portal 组件
> - 导入前请先停止原集群
> - 对于需要升级到 v2.0 的数据迁移任务，请不要执行 `stop-task`
> - 仅支持导入到 v2.0.0-rc.2 或更高版本
> - `import` 命令用于将 DM v1.0 集群导入到全新的 v2.0 集群。如果需要将数据迁移任务导入到已有的 v2.0 集群，请参考TiDB Data Migration 1.0.x 到 2.0.x+ 手动升级
> - 部分组件生成的部署目录会跟原集群不一样，具体可以使用 `display` 命令查看
> - 导入前运行 `tiup update --self && tiup update dm` 确认升级 TiUP DM 组件到最新版本
> - 导入后集群中仅会有一个 DM-master 节点，可参考扩容节点对 DM-master 进行扩容

语法

```
tiup dm import [flags]
```

选项

-v, –cluster-version（string，required）

重新部署的版本号，必须指定 v2.0.0-rc.2 或更高版本。

-d, –dir string（string, 默认当前目录）

指定 TiDB Ansible 所在目录。

–inventory string（string，默认 “inventory.ini”）

指定 Ansible inventory 文件的名字。

–rename（string，默认为空）

重命名导入的集群。默认集群名为 inventory 中指定的 `cluster_name`。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

导入过程的日志信息。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.11 tiup dm list

tiup-dm 支持使用同一个中控机部署多套集群，而命令 `tiup dm list` 可以查看当前登录的用户使用该中控机部署了哪些集群。

> 注意：
>
> 部署的集群数据默认放在 `~/.tiup/storage/dm/clusters/` 目录下，因此在同一台中控机上，当前登录用户无法查看其他用户部署的集群。

语法

```
tiup dm list [flags]
```

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

输出含有以下字段的表格：

- Name：集群名字
- User：部署用户
- Version：集群版本
- Path：集群部署数据在中控机上的路径
- PrivateKey：连接集群的私钥所在路径

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.12 tiup dm patch

在集群运行过程中，如果需要动态替换某个服务的二进制文件（即替换过程中保持集群可用），那么可以使用 `tiup dm patch` 命令，它会完成以下几件事情：

- 将用于替换的二进制包上传到目标机器
- 通过 API 下线节点
- 停止目标服务
- 解压二进制包，替换服务
- 启动目标服务

语法

```
tiup dm patch <cluster-name> <package-path> [flags]
```

- `<cluster-name>` 代表要操作的集群名
- `<package-path>` 为用于替换的二进制包

准备条件

执行 `tiup dm patch` 命令前，需要进行以下操作准备用于替换的二进制包：

- 确定当前要替换的组件名称 `${component}` (dm-master，dm-worker 等) 以及其版本 `${version}` (v2.0.0，v2.0.1 等)，以及其运行的平台 `${os}` (linux) 和 `${arch}` (amd64, arm64)
- 下载当前的组件包：`wget https://tiup-mirrors.pingcap.com/${component}-${version}-${os}-${arch}.tar.gz -O /tmp/${component}-${version}-${os}-${arch}.tar.gz`
- 建立临时打包目录：`mkdir -p /tmp/package && cd /tmp/package`
- 解压原来的二进制包：`tar xf /tmp/${component}-${version}-${os}-${arch}.tar.gz`
- 查看临时打包目录中的文件结构：`find .`
- 将要替换的二进制文件或配置文件复制到临时目录的对应位置
- 重新打包 `tar czf /tmp/${component}-hotfix-${os}-${arch}.tar.gz *`

完成以上操作后，`/tmp/${component}-hotfix-${os}-${arch}.tar.gz` 就可以作为 `<package-path>` 用于 patch 命令中。

选项

–overwrite

- 对某个组件（比如 TiDB，TiKV）进行 patch 之后，该集群扩容该组件时，tiup-dm 默认会用原来的版本。如果希望后续扩容的时候也使用 patch 之后的版本的话，就需要指定 `--overwrite` 选项。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-N, –node（strings，默认为 []，未选中任何节点）

指定要替换的节点，该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
>
> 若同时指定了 `-R, --role`，那么将替换它们的交集中的服务。

-R, –role（strings，默认为 []，未选中任何角色）

指定要替换的角色，该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> 注意：
>
> 若同时指定了 `-N, --node`，那么将替换它们的交集中的服务。

–offline

声明当前集群处于离线状态。指定该选项时，TiUP DM 仅原地替换集群组件的二进制文件，不重启服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

tiup-dm 的执行日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.13 tiup dm prune

在缩容集群后，etcd 中仍然会有少量元信息不会被清理，通常不会有问题，如果确实需要清理，可以手动执行 `tiup dm prune` 命令清理。

语法

```
tiup dm prune <cluster-name> [flags]
```

选项

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

清理过程的日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.14 tiup dm reload

在修改集群配置之后，需要通过 `tiup dm reload` 命令让集群重新加载配置才会生效，该命令会将中控机的配置发布到服务运行的远端机器，并按照升级的流程按顺序重启服务，重启过程中集群仍然可用。

语法

```
tiup dm reload <cluster-name> [flags]
```

`<cluster-name>` 代表要操作的集群名。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要重启的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> 注意：
>
> - 若同时指定了 `-R, --role`，那么将重启它们的交集中的服务
> - 若指定了选项 `--skip-restart`，则该选项无效

-R, –role（strings，默认为 []，表示所有角色）

指定要重启的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> 注意：
>
> - 若同时指定了 `-N, --node`，那么将重启它们的交集中的服务
> - 若指定了选项 `--skip-restart`，则该选项无效

–skip-restart

- 命令 `tiup dm reload` 会执行两个操作：
  - 刷新所有节点配置
  - 重启指定节点
- 该选项指定后仅刷新配置，不重启任何节点，这样刷新的配置也不会应用，需要等对应服务下次重启才会生效。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

tiup-dm 的执行日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.15 tiup dm replay

对集群进行升级或重启等操作时，操作有可能因为环境的原因而偶然失败。这时如果重新进行操作，需要从头开始执行所有步骤。如果集群规模较大，会耗费较长时间。此时可以使用 `tiup dm replay` 命令，命令重试刚才失败的命令，并且跳过已经成功的步骤。

语法

```
tiup dm replay <audit-id> [flags]
```

- `<audit-id>` 代表要重试的命令对应的 `audit-id`。使用`tiup dm audit` 可查看历史命令及其 `audit-id`。

选项

-h, –help

输出帮助信息。

输出

`<audit-id>` 对应的命令的输出。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.16 tiup dm restart

命令 `tiup dm restart` 用于重启指定集群的所有或部分服务。

> **注意：**
>
> 重启过程中会有一段时间服务不可用。

语法

```
tiup dm restart <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要重启的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> **注意：**
>
> 若同时指定了 `-R, --role`，那么将重启它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要重启的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> **注意：**
>
> 若同时指定了 `-N, --node`，那么将重启它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：`BOOLEAN`
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

重启服务的日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.17 tiup dm scale-in

`tiup dm scale-in` 命令用于集群缩容，缩容即下线服务，最终会将指定的节点从集群中移除，并删除遗留的相关文件。

语法

```
tiup dm scale-in <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。

选项

-N, –node（strings，无默认值，必须非空）

选择要缩容的节点，若缩容多个节点，以逗号分割。

–force

- 在某些情况下，有可能被缩容的节点宿主机已经宕机，导致无法通过 SSH 连接到节点进行操作，这个时候可以通过 `--force` 选项强制将其从集群中移除。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

缩容日志

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.18 tiup dm scale-out

`tiup dm scale-out` 命令用于集群扩容，扩容的内部逻辑与部署类似，tiup-dm 组件会先建立新节点的 SSH 连接，在目标节点上创建必要的目录，然后执行部署并且启动服务。

语法

```
tiup dm scale-out <cluster-name> <topology.yaml> [flags]
```

- `<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表
- `<topology.yaml>` 为事先编写好的扩容拓扑文件，该文件应当仅包含扩容部分的拓扑

选项

-u, –user（string，默认为当前执行命令的用户）

指定连接目标机器的用户名，该用户在目标机器上需要有免密 sudo root 的权限。

-i, –identity_file（string，默认 ~/.ssh/id_rsa）

指定连接目标机器的密钥文件。

-p, –password

- 在连接目标机器时使用密码登录，不可和 `-i/--identity_file` 同时使用。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

扩容日志

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.19 tiup dm start

命令 `tiup dm start` 用于启动指定集群的所有或部分服务。

语法

```
tiup dm start <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要启动的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> **注意：**
>
> 若同时指定了 `-R, --role`，那么将启动它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要启动的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> **注意：**
>
> 若同时指定了 `-N, --node`，那么将启动它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

启动日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.20 tiup dm stop

命令 `tiup dm stop` 用于停止指定集群的所有服务或部分服务。

> **注意：**
>
> 核心服务停止后集群将无法提供服务。

语法

```
tiup dm stop <cluster-name> [flags]
```

`<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。

选项

-N, –node（strings，默认为 []，表示所有节点）

指定要停止的节点，不指定则表示所有节点。该选项的值为以逗号分割的节点 ID 列表，节点 ID 为集群状态表格的第一列。

> **注意：**
>
> 若同时指定了 `-R, --role`，那么将停止它们的交集中的服务。

-R, –role（strings，默认为 []，表示所有角色）

指定要停止的角色，不指定则表示所有角色。该选项的值为以逗号分割的节点角色列表，角色为集群状态表格的第二列。

> **注意：**
>
> 若同时指定了 `-N, --node`，那么将停止它们的交集中的服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

停止服务的日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.21 tiup dm template

部署集群之前，需要准备一份集群的拓扑文件。TiUP 内置了拓扑文件的模版，用户可以通过修改模版来生成最终的拓扑文件。使用 `tiup dm template` 命令可以输出 TiUP 内置的模版内容。

语法

```
tiup dm template [flags]
```

如果不指定该选项，输出的默认模版包含以下实例：

- 3 个 DM-master 实例
- 3 个 DM-worker 实例
- 1 个 Prometheus 实例
- 1 个 Grafana 实例
- 1 个 Alertmanager 实例

选项

–full

- 输出详细的拓扑模版，该模版会以注释的形式带上可配置的参数。在命令中添加该选项，可开启该选项。
- 如果不指定该选项，默认输出最简单的拓扑模版。

-h, –help

输出帮助信息。

输出

根据指定选项输出拓扑模版到标准输出，可重定向到拓扑文件中用于部署。

<< 返回上一页 - TiUP DM 命令清单

12.4.7.4.22 tiup dm upgrade

命令 `tiup dm upgrade` 用于将指定集群升级到特定版本。

语法

```
tiup dm upgrade <cluster-name> <version> [flags]
```

- `<cluster-name>` 为要操作的集群名字，如果忘记集群名字可查看集群列表。
- `<version>` 为要升级到的目标版本，目前仅允许升级到比当前集群更高的版本，不允许升级到比当前集群更低的版本，即不允许降级。同时也不允许升级成 nightly 版本

选项

–offline

声明当前集群处于离线状态。指定该选项时，TiUP DM 仅原地替换集群组件的二进制文件，不重启服务。

-h, –help

- 输出帮助信息。
- 数据类型：BOOLEAN
- 该选项默认关闭，默认值为 `false`。在命令中添加该选项，并传入 `true` 值或不传值，均可开启此功能。

输出

升级服务的日志。

<< 返回上一页 - TiUP DM 命令清单

12.4.8 通过 TiUP 部署 TiDB 集群的拓扑文件配置

通过 TiUP 部署或扩容 TiDB 集群时，需要提供一份拓扑文件（示例）来描述集群拓扑。

同样，修改集群配置也是通过编辑拓扑文件来实现的，区别在于修改配置时仅允许修改部分字段。本文档介绍拓扑文件的各个区块以各区块中的各字段。

12.4.8.1 文件结构

一个通过 TiUP 部署的 TiDB 集群拓扑文件可能包含以下区块：

- global：集群全局配置，其中一些是集群的默认值，可以在实例里面单独配置
- monitored：监控服务配置，即 blackbox exporter 和 node exporter，每台机器上都会部署一个 node exporter 和一个 blackbox exporter
- server_configs：组件全局配置，可单独针对每个组件配置，若在实例中存在同名配置项，那么以实例中配置的为准
- pd_servers：PD 实例的配置，用来指定 PD 组件部署到哪些机器上
- tidb_servers：TiDB 实例的配置，用来指定 TiDB 组件部署到哪些机器上
- tikv_servers：TiKV 实例的配置，用来指定 TiKV 组件部署到哪些机器上
- tiflash_servers：TiFlash 实例的配置，用来指定 TiFlash 组件部署到哪些机器上
- pump_servers：Pump 实例的配置，用来指定 Pump 组件部署到哪些机器上
- drainer_servers：Drainer 实例的配置，用来指定 Drainer 组件部署到哪些机器上
- cdc_servers：CDC 实例的配置，用来指定 CDC 组件部署到哪些机器上
- tispark_masters：TiSpark Master 实例的配置，用来指定 TiSpark Master 组件部署到哪台机器上，仅允许部署一个 TiSpark Master 节点
- tispark_workers：TiSpark Worker 实例的配置，用来指定 TiSpark Worker 组件部署到哪些机器上
- monitoring_servers：用来指定 Prometheus 和 NGMonitoring 部署在哪些机器上，TiUP 支持部署多台 Prometheus 实例，但真实投入使用的只有第一个
- grafana_servers：Grafana 实例的配置，用来指定 Grafana 部署在哪台机器上
- alertmanager_servers：Alertemanager 实例的配置，用来指定 Alertmanager 部署在哪些机器上

12.4.8.1.1 global

`global` 区块为集群的全局配置，包含以下字段：

- `user`：以什么用户来启动部署的集群，默认值：“tidb”，如果 `<user>` 字段指定的用户在目标机器上不存在，会自动尝试创建
- `group`：自动创建用户时指定用户所属的用户组，默认和 `<user>` 字段值相同，若指定的组不存在，则自动创建
- `ssh_port`：指定连接目标机器进行操作的时候使用的 SSH 端口，默认值：22
- `enable_tls`：是否对集群启用 TLS，启用之后组件之间、客户端与组件之间都必须使用生成的 TLS 证书链接，启用后无法关闭，默认值：false
- `deploy_dir`：每个组件的部署目录，默认值：“deploy”。其应用规则如下：
  - 如果在实例级别配置了绝对路径的 `deploy_dir`，那么实际部署目录为该实例设定的 `deploy_dir`
  - 对于每个实例，如果用户未配置 `deploy_dir`，其默认值为相对路径 `<component-name>-<component-port>`
  - 如果 `global.deploy_dir` 为绝对路径，那么组件会部署到 `<global.deploy_dir>/<instance.deploy_dir>` 目录
  - 如果 `global.deploy_dir` 为相对路径，那么组件会部署到 `/home/<global.user>/<global.deploy_dir>/<instance.deploy_dir>` 目录
- `data_dir`：数据目录，默认值：“data”。其应用规则如下：
  - 如果在实例级别配置了绝对路径的 `data_dir`，那么实际数据目录为该实例设定的 `data_dir`

  - 对于每个实例，如果用户未配置 data_dir，其默认值为 <global.data_dir>
  - 如果 data_dir 为相对路径，那么组件数据将放到 <deploy_dir>/<data_dir> 中，其中 <deploy_dir> 的计算规则请参考 deploy_dir 字段的应用规则
- log_dir：日志目录，默认值：“log”。其应用规则如下：
  - 如果在实例级别配置了绝对路径的 log_dir，那么实际日志目录为该实例设定的 log_dir
  - 对于每个实例，如果用户未配置 log_dir，其默认值为 <global.log_dir>
  - 如果 log_dir 为相对路径，那么组件日志将放到 <deploy_dir>/<log_dir> 中，其中 <deploy_dir> 的计算规则请参考 deploy_dir 字段的应用规则
- os：目标机器的操作系统，该字段决定了向目标机器推送适配哪个操作系统的组件，默认值：linux
- arch：目标机器的 CPU 架构，该字段决定了向目标机器推送哪个平台的二进制包，支持 amd64 和 arm64，默认值：amd64
- resource_control：运行时资源控制，该字段下所有配置都将写入 systemd 的 service 文件中，默认无限制。支持控制的资源如下：
  - memory_limit: 限制运行时最大内存，例如 “2G” 表示最多使用 2GB 内存
  - cpu_quota：限制运行时最大 CPU 占用率，例如 “200%”
  - io_read_bandwidth_max：读磁盘 I/O 的最大带宽，例如：“/dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0 100M”
  - io_write_bandwidth_max：写磁盘 I/O 的最大带宽，例如：“/dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0 100M”
  - limit_core：控制 core dump 的大小

global 配置示例：

```
global:
  user: "tidb"
  resource_control:
    memory_limit: "2G"
```

上述配置指定使用 tidb 用户启动集群，同时限制每个组件运行时最多只能使用 2GB 内存。

12.4.8.1.2 monitored

monitored 用于配置目标机上的监控服务：node_exporter 和 blackbox_exporter。包含以下字段：

- node_exporter_port：node_exporter 的服务端口，默认值：9100
- blackbox_exporter_port：blackbox_exporter 的服务端口，默认值：9115
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成

monitored 配置示例：

```
monitored:
  node_exporter_port: 9100
  blackbox_exporter_port: 9115
```

上述配置指定了 node_exporter 使用 9100 端口，blackbox_exporter 使用 9115 端口。

12.4.8.1.3 server_configs

server_configs 用于配置服务，生成各组件的配置文件，类似 global 区块，该区块内的配置可以在具体的实例中被覆盖。主要包含以下字段：

- tidb：TiDB 服务的相关配置，支持的完整配置请参考 TiDB 配置文件描述
- tikv：TiKV 服务的相关配置，支持的完整配置请参考 TiKV 配置文件描述
- pd：PD 服务的相关配置，支持的完整配置请参考 PD 配置文件描述
- tiflash：TiFlash 服务的相关配置，支持的完整配置请参考 TiFlash 配置参数
- tiflash_learner：每个 TiFlash 中内置了一个特殊的 TiKV，该配置项用于配置这个特殊的 TiKV，一般不建议修改这个配置项下的内容
- pump：Pump 服务的相关配置，支持的完整配置请参考TiDB Binlog 配置说明
- drainer：Drainer 服务的相关配置，支持的完整配置请参考TiDB Binlog 配置说明
- cdc：CDC 服务的相关配置，支持的完整配置请参考TiCDC 安装部署

server_configs 配置示例：

```
server_configs:
  tidb:
    run-ddl: true
    lease: "45s"
    split-table: true
    token-limit: 1000
  tikv:
    log-level: "info"
    readpool.unified.min-thread-count: 1
```

上述配置指定了 TiDB 和 TiKV 的全局配置。

12.4.8.1.4 pd_servers

pd_servers 指定了将 PD 的服务部署到哪些机器上，同时可以指定每台机器上的服务配置。pd_servers 是一个数组，每个数组的元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- listen_host：当机器上有多个 IP 时，可以指定服务的监听 IP，默认为 0.0.0.0
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- name：指定该 PD 实例的名字，不同实例的名字必须唯一，否则无法部署
- client_port：指定 PD 的客户端连接端口，默认值：2379
- peer_port：指定 PD 之间互相通信的端口，默认值：2380
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl。在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”

- config：该字段配置规则和 server_configs 里的 pd 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 pd 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制。如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- listen_host
- name
- client_port
- peer_port
- deploy_dir
- data_dir
- log_dir
- arch
- os

pd_servers 配置示例：

```
pd_servers:
  - host: 10.0.1.11
    config:
      schedule.max-merge-region-size: 20
      schedule.max-merge-region-keys: 200000
  - host: 10.0.1.12
```

上述配置指定了将 PD 部署到 10.0.1.11 和 10.0.1.12，并针对 10.0.1.11 的 PD 进行一些特殊配置。

12.4.8.1.5 tidb_servers

tidb_servers 指定了将 TiDB 服务部署到哪些机器上，同时可以指定每台机器上的服务配置。tidb_servers 是一个数组，每个数组的元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- listen_host：当机器上有多个 IP 时，可以指定服务的监听 IP，默认为 0.0.0.0
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：TiDB 服务的监听端口，用于提供给 MySQL 客户端连接，默认值：4000
- status_port：TiDB 状态服务的监听端口，用于外部通过 http 请求查看 TiDB 服务的状态，默认值：10080
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成

- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 tidb 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 tidb 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- listen_host
- port
- status_port
- deploy_dir
- log_dir
- arch
- os

tidb_servers 配置示例：

```
tidb_servers:
  - host: 10.0.1.14
    config:
      log.level: warn
      log.slow-query-file: tidb-slow-overwrited.log
  - host: 10.0.1.15
```

12.4.8.1.6 tikv_servers

tikv_servers 约定了将 TiKV 服务部署到哪些机器上，同时可以指定每台机器上的服务配置。tikv_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- listen_host：当机器上有多个 IP 时，可以指定服务的监听 IP，默认为 0.0.0.0
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：TiKV 服务的监听端口，默认值：20160
- status_port：TiKV 状态服务的监听端口，默认值：20180
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成

- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 tikv 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 tikv 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- listen_host
- port
- status_port
- deploy_dir
- data_dir
- log_dir
- arch
- os

tikv_servers 配置示例：

```
tikv_servers:
  - host: 10.0.1.14
    config:
      server.labels: { zone: "zone1", host: "host1" }
  - host: 10.0.1.15
    config:
      server.labels: { zone: "zone1", host: "host2" }
```

12.4.8.1.7 tiflash_servers

tiflash_servers 约定了将 TiFlash 服务部署到哪些机器上，同时可以指定每台机器上的服务配置。该区块是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- tcp_port：TiFlash TCP 服务的端口，默认 9000
- http_port：TiFlash HTTP 服务的端口，默认 8123

- flash_service_port：TiFlash 提供服务的端口，TiDB 通过该端口从 TiFlash 读数据，默认 3930
- metrics_port：TiFlash 的状态端口，用于输出 metric 数据，默认 8234
- flash_proxy_port：内置 TiKV 的端口，默认 20170
- flash_proxy_status_port：内置 TiKV 的状态端口，默认为 20292
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成，TiFlash 的数据目录支持多个，采用逗号分割
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- tmp_path: TiFlash 临时文件的存放路径，默认使用 [path 或者 storage.latest.dir 的第一个目录] + “/tmp”
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 tiflash 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 tiflash 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- learner_config：每个 TiFlash 中内置了一个特殊的 TiKV 模块，该配置项用于配置这个特殊的 TiKV 模块，一般不建议修改这个配置项下的内容
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上字段中，data_dir 在部署完成之后只能新增目录，而下列字段部署完成之后不能再修改：

- host
- tcp_port
- http_port
- flash_service_port
- flash_proxy_port
- flash_proxy_status_port
- metrics_port
- deploy_dir
- log_dir
- tmp_path
- arch
- os

tiflash_servers 配置示例：

```
tiflash_servers:
  - host: 10.0.1.21
  - host: 10.0.1.22
```

12.4.8.1.8 pump_servers

pump_servers 约定了将 TiDB Binlog 组件的 Pump 服务部署到哪些机器上，同时可以指定每台机器上的服务配置。pump_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：Pump 服务的监听端口，默认 8250
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 pump 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 pump 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- data_dir
- log_dir
- arch
- os

pump_servers 配置示例：

```
pump_servers:
  - host: 10.0.1.21
    config:
      gc: 7
  - host: 10.0.1.22
```

12.4.8.1.9 drainer_servers

drainer_servers 约定了将 TiDB Binlog 组件的 Drainer 服务部署到哪些机器上，同时可以指定每台机器上的服务配置，drainer_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：Drainer 服务的监听端口，默认 8249
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- commit_ts：Drainer 启动的时候会去读取 checkpoint，如果读取不到，就会使用该字段做为初次启动开始的同步时间点，该字段默认为 -1（从 PD 总获取最新时间戳作为 commit_ts）
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 drainer 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 drainer 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- data_dir
- log_dir
- commit_ts
- arch
- os

drainer_servers 配置示例：

```
drainer_servers:
  - host: 10.0.1.21
    config:
      syncer.db-type: "mysql"
      syncer.to.host: "127.0.0.1"
      syncer.to.user: "root"
      syncer.to.password: ""
      syncer.to.port: 3306
      syncer.ignore-table:
        - db-name: test
          tbl-name: log
        - db-name: test
```

```
        tbl-name: audit
```

12.4.8.1.10 cdc_servers

cdc_servers 约定了将 TiCDC 服务部署到哪些机器上，同时可以指定每台机器上的服务配置，cdc_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：TiCDC 服务的监听端口，默认 8300
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录。若不指定该字段或指定为相对目录，数据目录则按照 global 中配置的 data_dir 生成。
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- gc-ttl：TiCDC 在 PD 设置的服务级别 GC safepoint 的 TTL (Time To Live) 时长，单位为秒，默认值为 86400，即 24 小时
- tz：TiCDC 服务使用的时区。TiCDC 在内部转换 timestamp 等时间数据类型和向下游同步数据时使用该时区，默认为进程运行本地时区。
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config：该字段配置规则和 server_configs 里的 cdc 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- data_dir
- log_dir
- arch
- os

cdc_servers 配置示例：

```
cdc_servers:
  - host: 10.0.1.20
    gc-ttl: 86400
```

```
    data_dir: "/cdc-data"
  - host: 10.0.1.21
    gc-ttl: 86400
    data_dir: "/cdc-data"
```

12.4.8.1.11 tispark_masters

tispark_masters 约定了将 TiSpark 的 master 节点部署到哪些机器上，同时可以指定每台机器上的服务配置，tispark_masters 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- listen_host：当机器上有多个 IP 时，可以指定服务的监听 IP，默认为 0.0.0.0
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：Spark 的监听端口，节点之前通讯用，默认 7077
- web_port：Spark 的网页端口，提供网页服务，可查看任务情况，默认 8080
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- java_home：指定要使用的 JRE 环境所在的路径。该参数对应系统环境变量 JAVA_HOME
- spark_config：用于配置 TiSpark 服务，生成配置文件并下发到 host 指定的机器
- spark_env：配置 Spark 启动时的环境变量
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- listen_host
- port
- web_port
- deploy_dir
- arch
- os

tispark_masters 配置示例：

```
tispark_masters:
  - host: 10.0.1.21
    spark_config:
      spark.driver.memory: "2g"
      spark.eventLog.enabled: "False"
      spark.tispark.grpc.framesize: 2147483647
      spark.tispark.grpc.timeout_in_sec: 100
      spark.tispark.meta.reload_period_in_sec: 60
      spark.tispark.request.command.priority: "Low"
      spark.tispark.table.scan_concurrency: 256
```

```
    spark_env:
      SPARK_EXECUTOR_CORES: 5
      SPARK_EXECUTOR_MEMORY: "10g"
      SPARK_WORKER_CORES: 5
      SPARK_WORKER_MEMORY: "10g"
  - host: 10.0.1.22
```

12.4.8.1.12 tispark_workers

tispark_workers 约定了将 TiSpark 的 worker 节点部署到哪些机器上，同时可以指定每台机器上的服务配置，tispark_workers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- listen_host：当机器上有多个 IP 时，可以指定服务的监听 IP，默认为 0.0.0.0
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：Spark 的监听端口，节点之前通讯用，默认 7077
- web_port：Spark 的网页端口，提供网页服务，可查看任务情况，默认 8080
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- java_home：指定要使用的 JRE 环境所在的路径。该参数对应系统环境变量 JAVA_HOME
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- listen_host
- port
- web_port
- deploy_dir
- arch
- os

tispark_workers 配置示例：

```
tispark_workers:
  - host: 10.0.1.22
  - host: 10.0.1.23
```

12.4.8.1.13 monitoring_servers

monitoring_servers 约定了将 Prometheus 服务部署到哪台机器上，同时可以指定这台机器上的服务配置，monitoring_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，必填项

- ng_port：指定连接 NGMonitoring 的 SSH 端口，在 TiUP v1.7.0 引入，支持 TiDB v5.3.0 及以上版本的持续性能分析和 Top SQL 功能。
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：指定 Prometheus 提供服务的端口，默认值：9090
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- storage_retention：Prometheus 监控数据保留时间，默认 “30d”
- rule_dir：该字段指定一个本地目录，该目录中应当含有完整的 *.rules.yml 文件，这些文件会在集群配置初始化阶段被传输到目标机器上，作为 Prometheus 的规则
- remote_config：用于支持将 Prometheus 数据写到远端，或从远端读取数据，该字段下有两个配置：
  - remote_write：参考 Prometheus <remote_write> 文档
  - remote_read：参考 Prometheus <remote_read> 文档
- external_alertmanagers：若配置了 external_alertmanagers，Prometheus 会将配置行为报警通知到集群外的 Alertmanager。该字段为一个数组，数组的元素为每个外部的 Alertmanager，由 host 和 web_port 字段构成
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- data_dir
- log_dir
- arch
- os

monitoring_servers 配置示例：

```
monitoring_servers:
  - host: 10.0.1.11
    rule_dir: /local/rule/dir
    remote_config:
      remote_write:
      - queue_config:
          batch_send_deadline: 5m
```

```
        capacity: 100000
        max_samples_per_send: 10000
        max_shards: 300
      url: http://127.0.0.1:8003/write
    remote_read:
    - url: http://127.0.0.1:8003/read
  external_alertmanagers:
  - host: 10.1.1.1
    web_port: 9093
  - host: 10.1.1.2
    web_port: 9094
```

12.4.8.1.14 grafana_servers

grafana_servers 约定了将 Grafana 服务部署到哪台机器上，同时可以指定这台机器上的服务配置，grafana_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：指定 Grafana 提供服务的端口，默认值：3000
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- username：Grafana 登录界面的用户名
- password：Grafana 对应的密码
- dashboard_dir：该字段指定一个本地目录，该目录中应当含有完整的 dashboard(*.json) 文件，这些文件会在集群配置初始化阶段被传输到目标机器上，作为 Grafana 的 dashboards
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

> 注意：
>
> 如果配置了 grafana_servers 的 dashboard_dir 字段，在执行 tiup cluster rename 命令进行集群重命名后，需要完成以下操作：
>
> 1. 对于本地的 dashboards 目录中的 *.json 文件，将 datasource 字段的值更新为新的集群名（这是因为 datasource 是以集群名命名的）
> 2. 执行 tiup cluster reload -R grafana 命令

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- arch
- os

grafana_servers 配置示例：

```
grafana_servers:
  - host: 10.0.1.11
    dashboard_dir: /local/dashboard/dir
```

12.4.8.1.15 alertmanager_servers

alertmanager_servers 约定了将 Alertmanager 服务部署到哪些机器上，同时可以指定这台机器上的服务配置，alertmanager_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- web_port：指定 Alertmanager 提供网页服务的端口，默认值：9093
- cluster_port：指定 Alertmanger 和其他 Alertmanager 通讯的端口，默认值：9094
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点的 ID，例如 “0,1”
- config_file：该字段指定一个本地文件，该文件会在集群配置初始化阶段被传输到目标机器上，作为 Alertmanager 的配置
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- web_port
- cluster_port
- deploy_dir
- data_dir
- log_dir
- arch

- os

alertmanager_servers 配置示例：

```
alertmanager_servers:
  - host: 10.0.1.11
    config_file: /local/config/file
  - host: 10.0.1.12
    config_file: /local/config/file
```

12.4.9 通过 TiUP 部署 DM 集群的拓扑文件配置

在部署或扩容 TiDB Data Migration (DM) 集群时，需要提供一份拓扑文件来描述集群拓扑，同样，修改配置也是通过编辑拓扑文件来实现的，区别在于修改配置时仅允许修改部分字段。

拓扑文件示例参考。

12.4.9.1 文件结构

一个 DM 集群的拓扑文件可能包含以下区块：

- global：集群全局配置，其中一些是集群的默认值，可以在实例里面单独配置
- server_configs：组件全局配置，可单独针对每个组件配置，若在实例中存在同名配置项，那么以实例中配置的为准
- master_servers：DM master 实例的配置，用来指定 DM 组件的 master 服务部署到哪些机器上
- worker_servers：DM worker 实例的配置，用来指定 DM 组件的 worker 服务部署到哪些机器上
- monitoring_servers：用来指定 Prometheus 部署在哪机器上，TiUP 支持部署多台 Prometheus 实例，但真实投入使用的只有第一个
- grafana_servers：Grafana 实例的配置，用来指定 Grafana 部署在哪台机器上
- alertmanager_servers：Alertemanager 实例的配置，用来指定 Alertmanager 部署在哪些机器上

12.4.9.1.1 global

global 区块为集群的全局配置，包含以下字段：

- user：以什么用户来启动部署的集群，默认值：“tidb”，如果 <user> 字段指定的用户在目标机器上不存在，会自动尝试创建
- group：自动创建用户时指定用户所属的用户组，默认和 <user> 字段值相同，若指定的组不存在，则自动创建
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，默认值：22
- deploy_dir：每个组件的部署目录，默认值：“deploy”，其应用规则如下：
  - 如果在实例级别配置了绝对路径的 deploy_dir，那么实际部署目录为该实例设定的 deploy_dir
  - 对于每个实例，如果用户未配置 deploy_dir，其默认值为相对路径 <component-name>-<component-port>
  - 如果 global.deploy_dir 为绝对路径，那么组件会部署到 <global.deploy_dir>/<instance.deploy_dir> 目录

  - 如果 global.deploy_dir 为相对路径，那么组件会部署到 /home/<global.user>/<global.deploy_dir>/<instance.deploy_dir> 目录

- data_dir：数据目录，默认值：“data”，其应用规则如下：
  - 如果在实例级别配置了绝对路径的 data_dir，那么实际数据目录为该实例设定的 data_dir
  - 对于每个实例，如果用户未配置 data_dir，其默认值为 <global.data_dir>
  - 如果 data_dir 为相对路径，那么组件数据将放到 <deploy_dir>/<data_dir> 中，其中 <deploy_dir> 的计算规则请参考 deploy_dir 字段的应用规则
- log_dir：数据目录，默认值：“log”，其应用规则如下：
  - 如果在实例级别配置了绝对路径的 log_dir，那么实际日志目录为该实例设定的 log_dir
  - 对于每个实例，如果用户未配置 log_dir，其默认值为 <global.log_dir>
  - 如果 log_dir 为相对路径，那么组件日志将放到 <deploy_dir>/<log_dir> 中，其中 <deploy_dir> 的计算规则请参考 deploy_dir 字段的应用规则
- os：目标机器的操作系统，该字段决定了向目标机器推送适配哪个操作系统的组件，默认值：linux
- arch：目标机器的 CPU 架构，该字段决定了向目标机器推送哪个平台的二进制包，支持 amd64 和 arm64，默认值：amd64
- resource_control：运行时资源控制，该字段下所有配置都将写入 systemd 的 service 文件中，默认不限制，支持控制的资源：
  - memory_limit: 限制运行时最大内存，例如“2G”表示最多使用 2GB 内存
  - cpu_quota：限制运行时最大 CPU 占用率，例如“200%”
  - io_read_bandwidth_max：读磁盘 IO 的最大带宽，例如：“/dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0 100M”
  - io_write_bandwidth_max：写磁盘 IO 的最大带宽，例如：“/dev/disk/by-path/pci-0000:00:1f.2-scsi-0:0:0:0 100M”
  - limit_core：控制 core dump 大小

global 配置示例：

```
global:
  user: "tidb"
  resource_control:
    memory_limit: "2G"
```

上述配置指定使用 tidb 用户启动集群，同时限制每个组件运行时最多只能使用 2GB 内存。

12.4.9.1.2 server_configs

server_configs 用于配置服务，生成各组件的配置文件，类似 global 区块，该区块内的配置可以在具体的实例中被覆盖。主要包含以下字段：

- master：DM master 服务的相关配置，支持的完整配置请参考DM-master 配置文件介绍
- worker：DM worker 服务的相关配置，支持的完整配置请参考DM-worker 配置文件介绍

server_configs 配置示例：

```
server_configs:
  master:
    log-level: info
    rpc-timeout: "30s"
    rpc-rate-limit: 10.0
    rpc-rate-burst: 40
  worker:
    log-level: info
```

12.4.9.2 master_servers

master_servers 约定了将 DM 组件的 master 节点部署到哪些机器上，同时可以指定每台机器上的服务配置，master_servers 是一个数组，每个数组的元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- name：指定该 DM master 实例的名字，不同实例的名字必须唯一，否则无法部署
- port：指定 DM master 提供给服务的端口，默认 8261
- peer_port：指定 DM master 之间互相通信的端口，默认值：8291
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点 ID，比如 “0,1”
- config：该字段配置规则和 server_configs 里的 master 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 master 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control
- v1_source_path：从 v1.0.x 升级时，可指定该目录，该目录中应当存放有 V1 的源的配置文件

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- name
- port
- peer_port
- deploy_dir
- data_dir

- log_dir
- arch
- os
- v1_source_path

master_servers 配置示例：

```
master_servers:
  - host: 10.0.1.11
    name: master1
    ssh_port: 22
    port: 8261
    peer_port: 8291
    deploy_dir: "/dm-deploy/dm-master-8261"
    data_dir: "/dm-data/dm-master-8261"
    log_dir: "/dm-deploy/dm-master-8261/log"
    numa_node: "0,1"
    # The following configs are used to overwrite the `server_configs.master` values.
    config:
      log-level: info
      rpc-timeout: "30s"
      rpc-rate-limit: 10.0
      rpc-rate-burst: 40
  - host: 10.0.1.18
    name: master2
  - host: 10.0.1.19
    name: master3
```

12.4.9.3 worker_servers

worker_servers 约定了将 DM 组件的 worker 节点部署到哪些机器上，同时可以指定每台机器上的服务配置，worker_servers 是一个数组，每个数组的元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- name：指定该 DM worker 实例的名字，不同实例的名字必须唯一，否则无法部署
- port：指定 DM worker 提供给服务的端口，默认 8262
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点 ID，比如 “0,1”

- config：该字段配置规则和 server_configs 里的 worker 配置规则相同，若配置了该字段，会将该字段内容和 server_configs 里的 worker 内容合并（若字段重叠，以本字段内容为准），然后生成配置文件并下发到 host 指定的机器
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- name
- port
- deploy_dir
- data_dir
- log_dir
- arch
- os

worker_servers 配置示例：

```
worker_servers:
  - host: 10.0.1.12
    ssh_port: 22
    port: 8262
    deploy_dir: "/dm-deploy/dm-worker-8262"
    log_dir: "/dm-deploy/dm-worker-8262/log"
    numa_node: "0,1"
    # config is used to overwrite the `server_configs.worker` values
    config:
      log-level: info
  - host: 10.0.1.19
```

12.4.9.3.1 monitoring_servers

monitoring_servers 约定了将 Prometheus 服务部署到哪台机器上，同时可以指定这台机器上的服务配置，monitoring_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：指定 Prometheus 提供服务的端口，默认值：9090
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成

- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点 ID，比如 “0,1”
- storage_retention：Prometheus 监控数据保留时间，默认 “15d”
- rule_dir：该字段指定一个本地目录，该目录中应当含有完整的 *.rules.yml 文件，这些文件会在集群配置初始化阶段被传输到目标机器上，作为 Prometheus 的规则
- remote_config：用于支持将 Prometheus 数据写到远端，或从远端读取数据，该字段下有两个配置：
  - remote_write：参考 Prometheus <remote_write> 文档
  - remote_read：参考 Prometheus <remote_read> 文档
- external_alertmanagers：若配置了 external_alertmanagers，Prometheus 会将配置行为报警通知到集群外的 Alertmanager。该字段为一个数组，数组的元素为每个外部的 Alertmanager，由 host 和 web_port 字段构成
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- data_dir
- log_dir
- arch
- os

monitoring_servers 配置示例：

```
monitoring_servers:
  - host: 10.0.1.11
    rule_dir: /local/rule/dir
    remote_config:
      remote_write:
      - queue_config:
          batch_send_deadline: 5m
          capacity: 100000
          max_samples_per_send: 10000
          max_shards: 300
        url: http://127.0.0.1:8003/write
      remote_read:
      - url: http://127.0.0.1:8003/read
    external_alertmanagers:
```

```
  - host: 10.1.1.1
    web_port: 9093
  - host: 10.1.1.2
    web_port: 9094
```

12.4.9.3.2 grafana_servers

grafana_servers 约定了将 Grafana 服务部署到哪台机器上，同时可以指定这台机器上的服务配置，grafana_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- port：指定 Grafana 提供服务的端口，默认值：3000
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- username：Grafana 登录界面的用户名
- password：Grafana 对应的密码
- dashboard_dir：该字段指定一个本地目录，该目录中应当含有完整的 dashboard(*.json) 文件，这些文件会在集群配置初始化阶段被传输到目标机器上，作为 Grafana 的 dashboards
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global中的 resource_control

> 注意：
>
> 如果配置了 grafana_servers 的 dashboard_dir 字段，在执行 tiup cluster rename 命令进行集群重命名后，需要完成以下操作：
>
> 1. 在本地的 dashboards 目录中，将 datasource 字段的值更新为新的集群名（datasource 是以集群名命名的）
> 2. 执行 tiup cluster reload -R grafana 命令

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- port
- deploy_dir
- arch
- os

grafana_servers 配置示例：

```
grafana_servers:
  - host: 10.0.1.11
    dashboard_dir: /local/dashboard/dir
```

12.4.9.3.3 alertmanager_servers

alertmanager_servers 约定了将 Alertmanager 服务部署到哪些机器上，同时可以指定这台机器上的服务配置，alertmanager_servers 是一个数组，每个数组元素包含以下字段：

- host：指定部署到哪台机器，字段值填 IP 地址，不可省略
- ssh_port：指定连接目标机器进行操作的时候使用的 SSH 端口，若不指定，则使用 global 区块中的 ssh_port
- web_port：指定 Alertmanager 提供网页服务的端口，默认值：9093
- cluster_port：指定 Alertmanger 和其他 Alertmanager 通讯的端口，默认值：9094
- deploy_dir：指定部署目录，若不指定，或指定为相对目录，则按照 global 中配置的 deploy_dir 生成
- data_dir：指定数据目录，若不指定，或指定为相对目录，则按照 global 中配置的 data_dir 生成
- log_dir：指定日志目录，若不指定，或指定为相对目录，则按照 global 中配置的 log_dir 生成
- numa_node：为该实例分配 NUMA 策略，如果指定了该参数，需要确保目标机装了 numactl，在指定该参数的情况下会通过 numactl 分配 cpubind 和 membind 策略。该字段参数为 string 类型，字段值填 NUMA 节点 ID，比如 “0,1”
- config_file：该字段指定一个本地文件，该文件会在集群配置初始化阶段被传输到目标机器上，作为 Alertmanager 的配置
- os：host 字段所指定的机器的操作系统，若不指定该字段，则默认为 global 中的 os
- arch：host 字段所指定的机器的架构，若不指定该字段，则默认为 global 中的 arch
- resource_control：针对该服务的资源控制，如果配置了该字段，会将该字段和 global 中的 resource_control 内容合并（若字段重叠，以本字段内容为准），然后生成 systemd 配置文件并下发到 host 指定机器。resource_control 的配置规则同 global 中的 resource_control

以上所有字段中，部分字段部署完成之后不能再修改。如下所示：

- host
- web_port
- cluster_port
- deploy_dir
- data_dir
- log_dir
- arch
- os

alertmanager_servers 配置示例：

```
alertmanager_servers:
  - host: 10.0.1.11
    config_file: /local/config/file
```

```
  - host: 10.0.1.12
    config_file: /local/config/file
```

12.4.10 TiUP 镜像参考指南

TiUP 镜像是 TiUP 的组件仓库，存放了一系列的组件和这些组件的元信息。镜像有两种存在形式：

- 本地磁盘上的目录：用于服务本地的 TiUP 客户端，文档中将称之为本地镜像
- 基于远程的磁盘目录启动的 HTTP 镜像：服务远程的 TiUP 客户端，文档中将称之为远程镜像

12.4.10.1 镜像的创建与更新

镜像可以通过以下两种方式创建：

- 通过命令 `tiup mirror init` 从零生成
- 通过命令 `tiup mirror clone` 从已有镜像克隆

在创建镜像之后，可以通过 `tiup mirror` 相关命令来给镜像添加组件或删除组件，无论是通过何种方式更新镜像，TiUP 都不会从镜像中删除任何文件，而是通过增加文件并分配新版本号的方式更新。

12.4.10.2 镜像结构

一个典型的镜像目录结构如下：

```
+ <mirror-dir>                                  # 镜像根目录
|-- root.json                                   # 镜像根证书
|-- {2..N}.root.json                            # 镜像根证书
|-- {1..N}.index.json                           # 组件/用户索引
|-- {1..N}.{component}.json                     # 组件元信息
|-- {component}-{version}-{os}-{arch}.tar.gz    # 组件二进制包
|-- snapshot.json                               # 镜像最新快照
|-- timestamp.json                              # 镜像最新时间戳
|--+ commits                                    # 镜像更新日志（可删除）
   |--+ commit-{ts1..tsN}
      |-- {N}.root.json
      |-- {N}.{component}.json
      |-- {N}.index.json
      |-- {component}-{version}-{os}-{arch}.tar.gz
      |-- snapshot.json
      |-- timestamp.json
|--+ keys                                       # 镜像私钥（可移动到其他位置）
   |-- {hash1..hashN}-root.json                 # 根证书私钥
   |-- {hash}-index.json                        # 索引私钥
   |-- {hash}-snapshot.json                     # 快照私钥
   |-- {hash}-timestamp.json                    # 时间戳私钥
```

注意：

- commits 目录是在更新镜像过程中产生的日志，用于回滚镜像，磁盘空间不足时可以定期删除旧的文件夹
- keys 文件夹中存放的私钥较敏感，建议单独妥善保管

12.4.10.2.1 根证书

在 TiUP 镜像中，根证书用于存放其他元信息文件的公钥，每次获取到任何元信息文件（*.json）都需要根据其文件类型（root，index，snapshot，timestamp）在当前已安装的 root.json 中找到对应的公钥，然后用公钥验证其签名是否合法。

根证书文件格式如下：

```
{
    "signatures": [                                         # 每个元信息文件有一系列的签名，
        ↪ 签名由该文件对应的几个私钥签出
        {
            "keyid": "{id-of-root-key-1}",                  # 第一个参与签名私钥的 ID，该 ID
                ↪ 由私钥对应的公钥内容哈希得到
            "sig": "{signature-by-root-key-1}"              # 该私钥对此文件 signed
                ↪ 部分签名的结果
        },
        ...
        {
            "keyid": "{id-of-root-key-N}",                  # 第 N 个参与签名私钥的 ID
            "sig": "{signature-by-root-key-N}"              # 该私钥对此文件 signed
                ↪ 部分签名的结果
        }
    ],
    "signed": {                                             # 被签名的部分
        "_type": "root",                                    # 该字段说明本文件的类型，root.
            ↪ json 的类型就是 root
        "expires": "{expiration-date-of-this-file}",        # 该文件的过期时间，
            ↪ 过期后客户端会拒绝此文件
        "roles": {                                          # root.json 中的 roles
            ↪ 用来记录对各个元文件签名的密钥
            "{role:index,root,snapshot,timestamp}": {       # 涉及的元文件类型包括 index,
                ↪ root, snapshot, timestamp
                "keys": {                                   # 只有 keys
                    ↪ 中记录的密钥签名才是合法的
                    "{id-of-the-key-1}": {                  # 用于签名 {role} 的第 1 个密钥
                        ↪ ID
                        "keytype": "rsa",                   # 密钥类型，目前固定为 rsa
```

```
                    "keyval": {                                 # 密钥的 payload
                        "public": "{public-key-content}"        # 表示公钥内容
                    },
                    "scheme": "rsassa-pss-sha256"               # 目前固定为 rsassa-pss-sha256
                },
                "{id-of-the-key-N}": {                          # 用于签名 {role} 的第 N 个密钥
                    ↪ ID
                    "keytype": "rsa",
                    "keyval": {
                        "public": "{public-key-content}"
                    },
                    "scheme": "rsassa-pss-sha256"
                }
            },
            "threshold": {N},                                   # threshold 指示该元文件需要至少
                ↪ N 个密钥签名
            "url": "/{role}.json"                               # url 是指该文件的获取地址，对于
                ↪ index 文件，需要在前面加上版本号，即 /{N}.index.json
        }
    },
    "spec_version": "0.1.0",                                    # 本文件遵循的规范版本，
        ↪ 未来变更文件结构需要升级版本号，目前为 0.1.0
    "version": {N}                                              # 本文件的版本号，
        ↪ 每次更新文件需要创建一个新的 {N+1}.root.json，并将其 version 设置为 N + 1
  }
}
```

12.4.10.2.2 索引

索引文件记录了镜像中所有的组件以及组件的所有者信息。

其格式如下：

```
{
    "signatures": [                                             # 该文件的签名
        {
            "keyid": "{id-of-index-key-1}",                     # 第一个参与签名的 key 的 ID
            "sig": "{signature-by-index-key-1}",                # 该私钥对此文件 signed
                ↪ 部分签名的结果
        },
        ...
        {
            "keyid": "{id-of-root-key-N}",                      # 第 N 个参与签名私钥的 ID
            "sig": "{signature-by-root-key-N}"                  # 该私钥对此文件 signed
                ↪ 部分签名的结果
        }
```

```
],
"signed": {
    "_type": "index",                                       # 指示该文件类型
    "components": {                                         # 组件列表
        "{component1}": {                                   # 第一个组件的名称
            "hidden": {bool},                               # 是否是隐藏组件
            "owner": "{owner-id}",                          # 组件管理员 ID
            "standalone": {bool},                           # 该组件是否可独立运行
            "url": "/{component}.json",                     # 获取组件的地址，需要加上版本号
                ↪ : /{N}.{component}.json
            "yanked": {bool}                                # 该组件是否已被标记为删除
        },
        ...
        "{componentN}": {                                   # 第 N 个组件的名称
            ...
        },
    },
    "default_components": ["{component1}.."{componentN}"], # 镜像必须包含的默认组件，
        ↪ 该字段目前固定为空（未启用）
    "expires": "{expiration-date-of-this-file}",            # 该文件的过期时间，
        ↪ 过期后客户端会拒绝此文件
    "owners": {
        "{owner1}": {                                       # 第一个属主的 ID
            "keys": {                                       # 只有 keys
                ↪ 中记录的密钥签名才是合法的
                "{id-of-the-key-1}": {                      # 该属主的第一个密钥
                    "keytype": "rsa",                       # 密钥类型，目前固定为 rsa
                    "keyval": {                             # 密钥的 payload
                        "public": "{public-key-content}"    # 表示公钥内容
                    },
                    "scheme": "rsassa-pss-sha256"           # 目前固定为 rsassa-pss-sha256
                },
                ...
                "{id-of-the-key-N}": {                      # 该属主的第 N 个密钥
                    ...
                }
            },
            "name": "{owner-name}",                         # 该属主的名字
            "threshod": {N}                                 #
                ↪ 指示该属主拥有的组件必须含有至少 N 个合法签名
        },
        ...
        "{ownerN}": {                                       # 第 N 个属主的 ID
            ...
        }
```

```
        }
        "spec_version": "0.1.0",                                    # 本文件遵循的规范版本，
            ↪ 未来变更文件结构需要升级版本号，目前为 0.1.0
        "version": {N}                                              # 本文件的版本号，
            ↪ 每次更新文件需要创建一个新的 {N+1}.index.json，并将其 version 设置为 N + 1
    }
}
```

12.4.10.2.3 组件

组件元信息文件记录了特定组件的平台以及版本信息。

其格式如下：

```
{
    "signatures": [                                                 # 该文件的签名
        {
            "keyid": "{id-of-index-key-1}",                         # 第一个参与签名的 key 的 ID
            "sig": "{signature-by-index-key-1}",                    # 该私钥对此文件 signed
                ↪ 部分签名的结果
        },
        ...
        {
            "keyid": "{id-of-root-key-N}",                          # 第 N 个参与签名私钥的 ID
            "sig": "{signature-by-root-key-N}"                      # 该私钥对此文件 signed
                ↪ 部分签名的结果
        }
    ],
    "signed": {
        "_type": "component",                                       # 指示该文件类型
        "description": "{description-of-the-component}",            # 该组件的描述信息
        "expires": "{expiration-date-of-this-file}",                # 该文件的过期时间，
            ↪ 过期后客户端会拒绝此文件
        "id": "{component-id}",                                     # 该组件的 ID，具有全局唯一性
        "nightly": "{nightly-cursor}",                              # nightly 游标，值为最新的
            ↪ nightly 的版本号（如 v5.0.0-nightly-20201209）
        "platforms": {                                              # 该组件支持的平台（如 darwin/
            ↪ amd64，linux/arm64 等）
            "{platform-pair-1}": {
                "{version-1}": {                                    # Semantic Version 版本号（如 v1
                    ↪ .0.0 等）
                    "dependencies": null,                           # 用于约定组件之间的依赖关系，
                        ↪ 该字段尚未使用，固定为 null
                    "entry": "{entry}",                             # 入口二进制文件位于 tar
                        ↪ 包的相对路径
```

```
                    "hashs": {                                      # tar 包的 checksum，我们使用
                        ↪ sha256 和 sha512
                        "sha256": "{sum-of-sha256}",
                        "sha512": "{sum-of-sha512}",
                    },
                    "length": {length-of-tar},                      # tar 包的长度
                    "released": "{release-time}",                   # 该版本的 release 时间
                    "url": "{url-of-tar}",                          # tar 包的下载地址
                    "yanked": {bool}                                # 该版本是否已被禁用
                }
            },
            ...
            "{platform-pair-N}": {
                ...
            }
        },
        "spec_version": "0.1.0",                                    # 本文件遵循的规范版本，
            ↪ 未来变更文件结构需要升级版本号，目前为 0.1.0
        "version": {N}                                              # 本文件的版本号，
            ↪ 每次更新文件需要创建一个新的 {N+1}.{component}.json，并将其 version 设置为 N + 1
}
```

12.4.10.2.4 快照

快照文件记录了各个元文件当前的版本号。

其格式如下：

```
{
    "signatures": [                                                 # 该文件的签名
        {
            "keyid": "{id-of-index-key-1}",                         # 第一个参与签名的 key 的 ID
            "sig": "{signature-by-index-key-1}",                    # 该私钥对此文件 signed
                ↪ 部分签名的结果
        },
        ...
        {
            "keyid": "{id-of-root-key-N}",                          # 第 N 个参与签名私钥的 ID
            "sig": "{signature-by-root-key-N}"                      # 该私钥对此文件 signed
                ↪ 部分签名的结果
        }
    ],
    "signed": {
        "_type": "snapshot",                                        # 指示该文件类型
        "expires": "{expiration-date-of-this-file}",                # 该文件的过期时间，
            ↪ 过期后客户端会拒绝此文件
```

```
        "meta": {                                               # 其他元文件的信息
            "/root.json": {
                "length": {length-of-json-file},                # root.json 的长度
                "version": {version-of-json-file}               # root.json 的 version
            },
            "/index.json": {
                "length": {length-of-json-file},
                "version": {version-of-json-file}
            },
            "/{component-1}.json": {
                "length": {length-of-json-file},
                "version": {version-of-json-file}
            },
            ...
            "/{component-N}.json": {
                ...
            }
        },
        "spec_version": "0.1.0",                                # 本文件遵循的规范版本，
            ↪ 未来变更文件结构需要升级版本号，目前为 0.1.0
        "version": 0                                            # 本文件的版本号，固定为 0
    }
```

12.4.10.2.5 时间戳

时间戳文件记录了当前快照 checksum。

其文件格式如下：

```
{
    "signatures": [                                             # 该文件的签名
        {
            "keyid": "{id-of-index-key-1}",                     # 第一个参与签名的 key 的 ID
            "sig": "{signature-by-index-key-1}",                # 该私钥对此文件 signed
                ↪ 部分签名的结果
        },
        ...
        {
            "keyid": "{id-of-root-key-N}",                      # 第 N 个参与签名私钥的 ID
            "sig": "{signature-by-root-key-N}"                  # 该私钥对此文件 signed
                ↪ 部分签名的结果
        }
    ],
    "signed": {
        "_type": "timestamp",                                   # 指示该文件类型
```

```
    "expires": "{expiration-date-of-this-file}",           # 该文件的过期时间，
        ↪ 过期后客户端会拒绝此文件
    "meta": {                                              # snapshot.json 的信息
        "/snapshot.json": {
            "hashes": {
                "sha256": "{sum-of-sha256}"                # snapshot.json 的 sha256
            },
            "length": {length-of-json-file}                # snapshot.json 的长度
        }
    },
    "spec_version": "0.1.0",                               # 本文件遵循的规范版本，
        ↪ 未来变更文件结构需要升级版本号，目前为 0.1.0
    "version": {N}                                         # 本文件的版本号，
        ↪ 每次更新文件需要覆盖 timestamp.json，并将其 version 设置为 N + 1
```

12.4.10.3 客户端工作流程

客户端通过以下逻辑保证从镜像下载到的文件是安全的：

- 客户端安装时随 binary 附带了一个 root.json
- 客户端运行时以已有的 root.json 为基础，做如下操作：
  1. 获取 root.json 中的 version，记为 N
  2. 向镜像请求 {N+1}.root.json，若成功，使用 root.json 中记录的公钥验证该文件是否合法
- 向镜像请求 timestamp.json，并使用 root.json 中记录的公钥验证该文件是否合法
- 检查 timestamp.json 中记录的 snapshot.json 的 checksum 和本地的 snapshot.json 的 checksum 是否吻合
  - 若不吻合，则向镜像请求最新的 snapshot.json 并使用 root.json 中记录的公钥验证该文件是否合法
- 对于 index.json 文件，从 snapshot.json 中获取其版本号 N，并请求 {N}.index.json，然后使用 root.json 中记录的公钥验证该文件是否合法
- 对于组件（如 tidb.json，tikv.json），从 snapshot.json 中获取其版本号 N，并请求 {N}.{component}.json，然后使用 index.json 中记录的公钥验证该文件是否合法
- 对于组件 tar 文件，从 {component}.json 中获取其 url 及 checksum，请求 url 得到 tar 包，并验证 checksum 是否正确

12.4.11 TiUP 组件文档

12.4.11.1 本地快速部署 TiDB 集群

TiDB 集群是由多个组件构成的分布式系统，一个典型的 TiDB 集群至少由 3 个 PD 节点、3 个 TiKV 节点和 2 个 TiDB 节点构成。对于想要快速体验 TiDB 的用户来说，手工部署这么多组件是非常耗时且麻烦的事情。本文介绍 TiUP 中的 playground 组件，以及如何通过 playground 组件快速搭建一套本地的 TiDB 测试环境。

12.4.11.1.1 playground 组件介绍

playground 组件的基本用法：

```
tiup playground ${version} [flags]
```

如果直接执行 `tiup playground` 命令，则 TiUP playground 会使用本地安装的 TiDB/TiKV/PD 组件或者安装这些组件的稳定版本，来启动一个由 1 个 TiKV、1 个 TiDB、1 个 PD 和 1 个 TiFlash 实例构成的集群。该命令实际做了以下事情：

- 因为该命令没有指定 playground 的版本，TiUP 会先查找已安装的 playground 的最新版本，假设已安装的 playground 最新版为 v1.9.0，则该命令相当于 tiup playground:v1.9.0
- 如果 playground 从未安装过任何版本的 TiDB/TiKV/PD 组件，TiUP 会先安装这些组件的最新稳定版，然后再启动运行这些组件的实例
- 因为该命令没有指定 TiDB/PD/TiKV 各组件的版本，默认情况下，它会使用各组件的最新发布版本，假设当前为 v5.4.0，则该命令相当于 tiup playground:1.9.0 v5.4.0
- 因为该命令也没有指定各组件的个数，默认情况下，它会启动由 1 个 TiDB、1 个 TiKV、1 个 PD 和 1 个 TiFlash 实例构成的最小化集群
- 在依次启动完各个 TiDB 组件后，playground 会提醒集群启动成功，并告诉你一些有用的信息，譬如如何通过 MySQL 客户端连接集群、如何访问TiDB Dashboard 等

playground 的命令行参数说明：

```
Flags:
      --db int                    设置集群中 TiDB 节点的数量（默认为1）
      --db.host host              指定 TiDB 的监听地址
      --db.port int               指定 TiDB 的端口号
      --db.binpath string         指定 TiDB 二进制文件的位置（开发调试用，可忽略）
      --db.config string          指定 TiDB 的配置文件（开发调试用，可忽略）
      --db.timeout int            指定 TiDB 最长等待超时时间，单位为秒。若配置为 0，则永不超时。
      --drainer int               设置集群中 Drainer 数据
      --drainer.binpath string    指定 Drainer 二进制文件的位置（开发调试用，可忽略）
      --drainer.config string     指定 Drainer 的配置文件
  -h, --help                      打印帮助信息
      --host string               设置每个组件的监听地址（默认为 127.0.0.1），
          ↪ 如果要提供给别的电脑访问，可设置为 0.0.0.0
      --kv int                    设置集群中 TiKV 节点的数量（默认为1）
      --kv.binpath string         指定 TiKV 二进制文件的位置（开发调试用，可忽略）
      --kv.config string          指定 TiKV 的配置文件（开发调试用，可忽略）
      --mode string               指定 playground 的运行模式，取值选项为 'tidb'（默认）和 'tikv-
          ↪ slim'
      --pd int                    设置集群中 PD 节点的数量（默认为1）
      --pd.host host              指定 PD 的监听地址
      --pd.binpath string         指定 PD 二进制文件的位置（开发调试用，可忽略）
      --pd.config string          指定 PD 的配置文件（开发调试用，可忽略）
      --pump int                  设置集群中 Pump 节点的数量（非 0 的时候 TiDB 会开启 TiDB Binlog
          ↪ ）
```

```
      --pump.binpath string      指定 Pump 二进制文件的位置（开发调试用，可忽略）
      --pump.config string       指定 Pump 的配置文件（开发调试用，可忽略）
  -T, --tag string               设置 playground 的 tag 信息
      --ticdc int                设置集群中 TiCDC 节点的数量（默认为 0）
      --ticdc.binpath string     指定 TiCDC 二进制文件的位置（开发调试用，可忽略）
      --ticdc.config string      指定 TiCDC 的配置文件（开发调试用，可忽略）
      --tiflash int              设置集群中 TiFlash 节点的数量（默认为 1）
      --tiflash.binpath string   指定 TiFlash 的二进制文件位置（开发调试用，可忽略）
      --tiflash.config string    指定 TiFlash 的配置文件（开发调试用，可忽略）
      --tiflash.timeout int      指定 TiFlash 最长等待超时时间，单位为秒，若配置为 0，则永不超时
          ↪ 。
  -v, --version                  显示 playground 的版本号
      --without-monitor          设置不使用 Prometheus 和 Grafana 的监控功能。若不添加此参数，
          ↪ 则默认开启监控功能。
```

12.4.11.1.2 使用示例

查看可用的 TiDB 版本

```
tiup list tidb
```

启动一个指定版本的 TiDB 集群

```
tiup playground ${version}
```

将 `${version}` 替换为所需的版本号。

启动一个每日构建版的 TiDB 集群

```
tiup playground nightly
```

nightly 就是这个集群的版本号，这里指定为每日构建版本。

覆盖 PD 的默认配置

首先，你需要复制 PD 的配置模版。假设你将复制的配置文件放置在 `~/config/pd.toml`，按需修改一些内容后，执行以下命令可以覆盖 PD 的默认配置：

```
tiup playground --pd.config ~/config/pd.toml
```

替换默认的二进制文件

默认启动 playground 时，各个组件都是使用官方镜像组件包中的二进制文件启动的，如果本地编译了一个临时的二进制文件想要放入集群中测试，可以使用 `--{comp}.binpath` 这个参数替换，例如执行以下命令替换 TiDB 的二进制文件：

```
tiup playground --db.binpath /xx/tidb-server
```

启动多个组件实例

默认情况下各启动一个 TiDB、TiKV 和 PD 实例，如果希望启动多个，可以加上如下参数：

```
tiup playground --db 3 --pd 3 --kv 3
```

12.4.11.1.3 快速连接到由 playground 启动的 TiDB 集群

TiUP 提供了 `client` 组件，用于自动寻找并连接 playground 在本地启动的 TiDB 集群，使用方式为：

```
tiup client
```

该命令会在控制台上提供当前机器上由 playground 启动的 TiDB 集群列表，选中需要连接的 TiDB 集群，点击回车后，可以打开一个自带的 MySQL 客户端以连接 TiDB。

12.4.11.1.4 查看已启动集群的信息

```
tiup playground display
```

可以看到如下信息：

```
Pid    Role     Uptime
---    ----     ------
84518  pd       35m22.929404512s
84519  tikv     35m22.927757153s
84520  pump     35m22.92618275s
86189  tidb     exited
86526  tidb     34m28.293148663s
86190  drainer  35m19.91349249s
```

12.4.11.1.5 扩容集群

扩容集群的命令行参数与启动集群的相似。以下命令可以扩容两个 TiDB：

```
tiup playground scale-out --db 2
```

12.4.11.1.6 缩容集群

可在 `tiup playground scale-in` 命令中指定 `pid`，以缩容对应的实例。可以通过 `tiup playground display` 命令查看 `pid`。

```
tiup playground scale-in --pid 86526
```

12.4.11.2 使用 TiUP 部署运维 TiDB 线上集群

本文重在介绍如何使用 TiUP 的 cluster 组件，如果需要线上部署的完整步骤，可参考使用 TiUP 部署 TiDB 集群。

与 playground 组件用于部署本地测试集群类似，cluster 组件用于快速部署生产集群。对比 playground，cluster 组件提供了更强大的生产集群管理功能，包括对集群的升级、缩容、扩容甚至操作、审计等。

cluster 组件的帮助文档如下：

```
tiup cluster
```

```
Starting component `cluster`: /home/tidb/.tiup/components/cluster/v1.9.0/cluster
Deploy a TiDB cluster for production

Usage:
  tiup cluster [flags]
  tiup [command]

Available Commands:
  deploy      部署集群
  start       启动已部署的集群
  stop        停止集群
  restart     重启集群
  scale-in    集群缩容
  scale-out   集群扩容
  clean       清理数据
  destroy     销毁集群
  upgrade     升级集群
  exec        在集群的一个或多个机器上执行命令
  display     获取集群信息
  list        获取集群列表
  audit       查看集群操作日志
  import      导入一个由 TiDB Ansible 部署的集群
  edit-config 编辑 TiDB 集群的配置
  reload      用于必要时重载集群配置
  patch       使用临时的组件包替换集群上已部署的组件
  help        打印 Help 信息

Flags:
  -h, --help              帮助信息
      --native-ssh        使用系统默认的 SSH 客户端
      --wait-timeout int  等待操作超时的时间
      --ssh-timeout int   SSH 连接超时时间
  -y, --yes               跳过所有的确认步骤
```

12.4.11.2.1 部署集群

部署集群的命令为 `tiup cluster deploy`，一般用法为：

```
tiup cluster deploy <cluster-name> <version> <topology.yaml> [flags]
```

该命令需要提供集群的名字、集群使用的 TiDB 版本，以及一个集群的拓扑文件。

拓扑文件的编写可参考示例。以一个最简单的拓扑为例，将下列文件保存为 `/tmp/topology.yaml`：

> **注意：**
>
> TiUP Cluster 组件的部署和扩容拓扑是使用 yaml 语法编写，所以需要注意缩进。

```
---

pd_servers:
  - host: 172.16.5.134
    name: pd-134
  - host: 172.16.5.139
    name: pd-139
  - host: 172.16.5.140
    name: pd-140

tidb_servers:
  - host: 172.16.5.134
  - host: 172.16.5.139
  - host: 172.16.5.140

tikv_servers:
  - host: 172.16.5.134
  - host: 172.16.5.139
  - host: 172.16.5.140

grafana_servers:
  - host: 172.16.5.134

monitoring_servers:
  - host: 172.16.5.134
```

TiUP 默认部署在 amd64 架构上运行的 binary，若目标机器为 arm64 架构，可以在拓扑文件中进行配置：

```
global:
  arch: "arm64"           # 让所有机器默认使用 arm64 的 binary

tidb_servers:
  - host: 172.16.5.134
    arch: "amd64"         # 这台机器会使用 amd64 的 binary
  - host: 172.16.5.139
    arch: "arm64"         # 这台机器会使用 arm64 的 binary
  - host: 172.16.5.140    # 没有配置 arch 字段的机器，会使用 global 中的默认值，这个例子中是
      ↪ arm64
```

```
...
```

假如我们想要使用 TiDB 的 v5.4.0 版本，集群名字为 prod-cluster，则执行以下命令：

```
tiup cluster deploy -p prod-cluster v5.4.0 /tmp/topology.yaml
```

执行过程中会再次确认拓扑结构并提示输入目标机器上的 root 密码（-p 表示使用密码）：

```
Please confirm your topology:
TiDB Cluster: prod-cluster
TiDB Version: v5.4.0
Type        Host          Ports        Directories
----        ----          -----        -----------
pd          172.16.5.134  2379/2380    deploy/pd-2379,data/pd-2379
pd          172.16.5.139  2379/2380    deploy/pd-2379,data/pd-2379
pd          172.16.5.140  2379/2380    deploy/pd-2379,data/pd-2379
tikv        172.16.5.134  20160/20180  deploy/tikv-20160,data/tikv-20160
tikv        172.16.5.139  20160/20180  deploy/tikv-20160,data/tikv-20160
tikv        172.16.5.140  20160/20180  deploy/tikv-20160,data/tikv-20160
tidb        172.16.5.134  4000/10080   deploy/tidb-4000
tidb        172.16.5.139  4000/10080   deploy/tidb-4000
tidb        172.16.5.140  4000/10080   deploy/tidb-4000
prometheus  172.16.5.134  9090         deploy/prometheus-9090,data/prometheus-9090
grafana     172.16.5.134  3000         deploy/grafana-3000
Attention:
    1. If the topology is not what you expected, check your yaml file.
    1. Please confirm there is no port/directory conflicts in same host.
Do you want to continue? [y/N]:
```

输入密码后 tiup-cluster 便会下载需要的组件并部署到对应的机器上，当看到以下提示时说明部署成功：

```
Deployed cluster `prod-cluster` successfully
```

12.4.11.2.2 查看集群列表

集群部署成功后，可以通过 tiup cluster list 命令在集群列表中查看该集群：

```
tiup cluster list
```

```
Starting /root/.tiup/components/cluster/v1.9.0/cluster list
Name          User  Version    Path                                            PrivateKey
----          ----  -------    ----                                            ----------
prod-cluster  tidb  v5.4.0    /root/.tiup/storage/cluster/clusters/prod-cluster  /root/.tiup/
    ↪ storage/cluster/clusters/prod-cluster/ssh/id_rsa
```

12.4.11.2.3 启动集群

集群部署成功后，可以执行以下命令启动该集群。如果忘记了部署的集群的名字，可以使用 `tiup cluster list` 命令查看。

```
tiup cluster start prod-cluster
```

12.4.11.2.4 检查集群状态

如果想查看集群中每个组件的运行状态，逐一登录到各个机器上查看显然很低效。因此，TiUP 提供了 `tiup cluster display` 命令，用法如下：

```
tiup cluster display prod-cluster
```

```
Starting /root/.tiup/components/cluster/v1.9.0/cluster display prod-cluster
TiDB Cluster: prod-cluster
TiDB Version: v5.4.0
ID                  Role        Host          Ports        Status     Data Dir
    ↪ Deploy Dir
--                  ----        ----          -----        ------     --------
    ↪ ----------
172.16.5.134:3000   grafana     172.16.5.134  3000         Up         -
    ↪ deploy/grafana-3000
172.16.5.134:2379   pd          172.16.5.134  2379/2380    Up|L       data/pd-2379
    ↪ deploy/pd-2379
172.16.5.139:2379   pd          172.16.5.139  2379/2380    Up|UI      data/pd-2379
    ↪ deploy/pd-2379
172.16.5.140:2379   pd          172.16.5.140  2379/2380    Up         data/pd-2379
    ↪ deploy/pd-2379
172.16.5.134:9090   prometheus  172.16.5.134  9090         Up         data/prometheus-9090
    ↪ deploy/prometheus-9090
172.16.5.134:4000   tidb        172.16.5.134  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.139:4000   tidb        172.16.5.139  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.140:4000   tidb        172.16.5.140  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.134:20160  tikv        172.16.5.134  20160/20180  Up         data/tikv-20160
    ↪ deploy/tikv-20160
172.16.5.139:20160  tikv        172.16.5.139  20160/20180  Up         data/tikv-20160
    ↪ deploy/tikv-20160
172.16.5.140:20160  tikv        172.16.5.140  20160/20180  Up         data/tikv-20160
    ↪ deploy/tikv-20160
```

Status 列用 Up 或者 Down 表示该服务是否正常。对于 PD 组件，同时可能会带有 |L 表示该 PD 是 Leader，|UI 表示该 PD 运行着TiDB Dashboard。

12.4.11.2.5 缩容节点

> **注意：**
>
> 本节只展示缩容命令的语法示例，线上扩缩容具体步骤可参考使用 TiUP 扩容缩容 TiDB 集群。

缩容即下线服务，最终会将指定的节点从集群中移除，并删除遗留的相关文件。

由于 TiKV 和 TiDB Binlog 组件的下线是异步的（需要先通过 API 执行移除操作）并且下线过程耗时较长（需要持续观察节点是否已经下线成功），所以对 TiKV 和 TiDB Binlog 组件做了特殊处理：

- 对 TiKV 及 Binlog 组件的操作
  - TiUP cluster 通过 API 将其下线后直接退出而不等待下线完成
  - 等之后再执行集群操作相关的命令时，会检查是否存在已经下线完成的 TiKV 或者 Binlog 节点。如果不存在，则继续执行指定的操作；如果存在，则执行如下操作：
    1. 停止已经下线掉的节点的服务
    2. 清理已经下线掉的节点的相关数据文件
    3. 更新集群的拓扑，移除已经下线掉的节点
- 对其他组件的操作
  - 下线 PD 组件时，会通过 API 将指定节点从集群中删除掉（这个过程很快），然后停掉指定 PD 的服务并且清除该节点的相关数据文件
  - 下线其他组件时，直接停止并且清除节点的相关数据文件

缩容命令的基本用法：

```
tiup cluster scale-in <cluster-name> -N <node-id>
```

它需要指定至少两个参数，一个是集群名字，另一个是节点 ID。节点 ID 可以参考上一节使用 `tiup cluster display` 命令获取。

比如想缩容 172.16.5.140 上的 TiKV 节点，可以执行：

```
tiup cluster scale-in prod-cluster -N 172.16.5.140:20160
```

通过 `tiup cluster display` 可以看到该 TiKV 已经被标记为 `Offline`：

```
tiup cluster display prod-cluster
```

```
Starting /root/.tiup/components/cluster/v1.9.0/cluster display prod-cluster
TiDB Cluster: prod-cluster
TiDB Version: v5.4.0
ID                  Role        Host          Ports        Status     Data Dir
    ↪ Deploy Dir
--                  ----        ----          -----        ------     --------
    ↪ ----------
```

```
172.16.5.134:3000   grafana     172.16.5.134  3000         Up         -
    ↪ deploy/grafana-3000
172.16.5.134:2379   pd          172.16.5.134  2379/2380    Up|L       data/pd-2379
    ↪ deploy/pd-2379
172.16.5.139:2379   pd          172.16.5.139  2379/2380    Up|UI      data/pd-2379
    ↪ deploy/pd-2379
172.16.5.140:2379   pd          172.16.5.140  2379/2380    Up         data/pd-2379
    ↪ deploy/pd-2379
172.16.5.134:9090   prometheus  172.16.5.134  9090         Up         data/prometheus-9090
    ↪ deploy/prometheus-9090
172.16.5.134:4000   tidb        172.16.5.134  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.139:4000   tidb        172.16.5.139  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.140:4000   tidb        172.16.5.140  4000/10080   Up         -
    ↪ deploy/tidb-4000
172.16.5.134:20160  tikv        172.16.5.134  20160/20180  Up         data/tikv-20160
    ↪ deploy/tikv-20160
172.16.5.139:20160  tikv        172.16.5.139  20160/20180  Up         data/tikv-20160
    ↪ deploy/tikv-20160
172.16.5.140:20160  tikv        172.16.5.140  20160/20180  Offline    data/tikv-20160
    ↪ deploy/tikv-20160
```

待 PD 将其数据调度到其他 TiKV 后，该节点会被自动删除。

12.4.11.2.6 扩容节点

> **注意：**
>
> 本节只用于展示扩容命令的语法示例，线上扩缩容可参考使用 TiUP 扩容缩容 TiDB 集群。

扩容的内部逻辑与部署类似，TiUP cluster 组件会先保证节点的 SSH 连接，在目标节点上创建必要的目录，然后执行部署并且启动服务。其中 PD 节点的扩容会通过 join 方式加入到集群中，并且会更新与 PD 有关联的服务的配置；其他服务直接启动加入到集群中。所有服务在扩容时都会做正确性验证，最终返回是否扩容成功。

例如，在集群 `tidb-test` 中扩容一个 TiKV 节点和一个 PD 节点：

1. 新建 scale.yaml 文件，添加新增的 TiKV 和 PD 节点 IP：

> **注意：**
>
> 需要新建一个拓扑文件，文件中只写入扩容节点的描述信息，不要包含已存在的节点。

```
---

pd_servers:
  - host: 172.16.5.140

tikv_servers:
  - host: 172.16.5.140
```

2. 执行扩容操作。TiUP cluster 根据 scale.yaml 文件中声明的端口、目录等信息在集群中添加相应的节点：

```
tiup cluster scale-out tidb-test scale.yaml
```

执行完成之后可以通过 `tiup cluster display tidb-test` 命令检查扩容后的集群状态。

12.4.11.2.7 滚动升级

> **注意：**
>
> 本节只用于展示命令的语法示例，线上升级请参考使用 TiUP 升级 TiDB。

滚动升级功能借助 TiDB 的分布式能力，升级过程中尽量保证对前端业务透明、无感知。升级时会先检查各个组件的配置文件是否合理，如果配置有问题，则报错退出；如果配置没有问题，则工具会逐个节点进行升级。其中对不同节点有不同的操作。

不同节点的操作

- 升级 PD 节点
  - 优先升级非 Leader 节点
  - 所有非 Leader 节点升级完成后再升级 Leader 节点
    * 工具会向 PD 发送一条命令将 Leader 迁移到升级完成的节点上
    * 当 Leader 已经切换到其他节点之后，再对旧的 Leader 节点做升级操作
  - 同时升级过程中，若发现有不健康的节点，工具会中止本次升级并退出，此时需要由人工判断、修复后再执行升级。
- 升级 TiKV 节点
  - 先在 PD 中添加一个迁移对应 TiKV 上 Region leader 的调度，通过迁移 Leader 确保升级过程中不影响前端业务
  - 等待迁移 Leader 完成之后，再对该 TiKV 节点进行升级更新
  - 等更新后的 TiKV 正常启动之后再移除迁移 Leader 的调度
- 升级其他服务
  - 正常停止服务再更新

升级操作

升级命令参数如下：

```
Usage:
  tiup cluster upgrade <cluster-name> <version> [flags]

Flags:
      --force                 在不 transfer leader 的情况下强制升级（危险操作）
  -h, --help                  帮助手册
      --transfer-timeout int  transfer leader 的超时时间

Global Flags:
      --native-ssh        使用系统默认的 SSH 客户端
      --wait-timeout int  等待操作超时的时间
      --ssh-timeout int   SSH 连接的超时时间
  -y, --yes               跳过所有的确认步骤
```

例如，把集群升级到 v5.4.0 的命令为：

```
tiup cluster upgrade tidb-test v5.4.0
```

12.4.11.2.8 更新配置

如果想要动态更新组件的配置，TiUP cluster 组件为每个集群保存了一份当前的配置，如果想要编辑这份配置，则执行 `tiup cluster edit-config <cluster-name>` 命令。例如：

```
tiup cluster edit-config prod-cluster
```

然后 TiUP cluster 组件会使用 vi 打开配置文件供编辑（如果你想要使用其他编辑器，请使用 `EDITOR` 环境变量自定义编辑器，例如 `export EDITOR=nano`），编辑完之后保存即可。此时的配置并没有应用到集群，如果想要让它生效，还需要执行：

```
tiup cluster reload prod-cluster
```

该操作会将配置发送到目标机器，重启集群，使配置生效。

> 注意：
>
> 对于监控组件，可以通过执行 `tiup cluster edit-config` 命令在对应实例上添加自定义配置路径来进行配置自定义，例如：

```
---

grafana_servers:
```

```
  - host: 172.16.5.134
    dashboard_dir: /path/to/local/dashboards/dir

monitoring_servers:
  - host: 172.16.5.134
    rule_dir: /path/to/local/rules/dir

alertmanager_servers:
  - host: 172.16.5.134
    config_file: /path/to/local/alertmanager.yml
```

路径内容格式如下：

- grafana_servers 的 dashboard_dir 字段指定的文件夹中应当含有完整的 *.json 文件。
- monitoring_servers 的 rule_dir 字段定义的文件夹中应当含有完整的 *.rules.yml 文件。
- alertmanager_servers 的 config_file 格式请参考 Alertmanager 配置模板。

在执行 tiup reload 时，TiUP 会将中控机上对应的配置上传到目标机器对应的配置目录中，上传之前会删除目标机器中已有的旧配置文件。如果想要修改某一个配置文件，请确保将所有的（包含未修改的）配置文件都放在同一个目录中。例如，要修改 Grafana 的 tidb.json 文件，可以先将 Grafana 的 dashboards 目录中所有的 *.json 文件拷贝到本地目录中，再修改 tidb.json 文件。否则最终的目标机器上将缺失其他的 JSON 文件。

> 注意：
>
> 如果配置了 grafana_servers 的 dashboard_dir 字段，在执行 tiup cluster rename 命令进行集群重命名后，需要完成以下操作：
>
> 1. 在本地的 dashboards 目录中，将集群名修改为新的集群名。
> 2. 在本地的 dashboards 目录中，将 datasource 更新为新的集群名（datasource 是以集群名命名的）。
> 3. 执行 tiup cluster reload -R grafana 命令。

12.4.11.2.9 更新组件

常规的升级集群可以使用 upgrade 命令，但是在某些场景下（例如 Debug），可能需要用一个临时的包替换正在运行的组件，此时可以用 patch 命令：

```
tiup cluster patch --help
```

```
Replace the remote package with a specified package and restart the service

Usage:
  tiup cluster patch <cluster-name> <package-path> [flags]
```

```
Flags:
  -h, --help                    帮助信息
  -N, --node strings            指定被替换的节点
      --overwrite               在未来的 scale-out 操作中使用当前指定的临时包
  -R, --role strings            指定被替换的服务类型
      --transfer-timeout int    transfer leader 的超时时间

Global Flags:
      --native-ssh        使用系统默认的 SSH 客户端
      --wait-timeout int  等待操作超时的时间
      --ssh-timeout int   SSH 连接的超时时间
  -y, --yes               跳过所有的确认步骤
```

例如，有一个 TiDB 的 hotfix 包放在 `/tmp/tidb-hotfix.tar.gz`，如果此时想要替换集群上的所有 TiDB，则可以执行：

```
tiup cluster patch test-cluster /tmp/tidb-hotfix.tar.gz -R tidb
```

或者只替换其中一个 TiDB：

```
tiup cluster patch test-cluster /tmp/tidb-hotfix.tar.gz -N 172.16.4.5:4000
```

12.4.11.2.10 导入 TiDB Ansible 集群

> **注意：**
>
> TiUP cluster 组件对 TiSpark 的支持目前为实验性特性，暂不支持导入启用了 TiSpark 组件的集群。

在 TiUP 之前，一般使用 TiDB Ansible 部署 TiDB 集群，import 命令用于将这部分集群过渡给 TiUP 接管。import 命令用法如下：

```
tiup cluster import --help
```

```
Import an exist TiDB cluster from TiDB-Ansible

Usage:
  tiup cluster import [flags]

Flags:
  -d, --dir string         TiDB-Ansible 的目录，默认为当前目录
  -h, --help               import 的帮助信息
      --inventory string   inventory 文件的名字 (默认为 "inventory.ini")
      --no-backup          不备份 Ansible 目录，用于存在多个 inventory 文件的 Ansible 目录
  -r, --rename NAME        重命名被导入的集群
```

```
Global Flags:
      --native-ssh        使用系统默认的 SSH 客户端
      --wait-timeout int  等待操作超时的时间
      --ssh-timeout int   SSH 连接的超时时间
  -y, --yes               跳过所有的确认步骤
```

例如，导入一个 TiDB Ansible 集群：

```
cd tidb-ansible
tiup cluster import
```

或者

```
tiup cluster import --dir=/path/to/tidb-ansible
```

12.4.11.2.11 查看操作日志

操作日志的查看可以借助 audit 命令，其用法如下：

```
Usage:
  tiup cluster audit [audit-id] [flags]

Flags:
  -h, --help   help for audit
```

在不使用 [audit-id] 参数时，该命令会显示执行的命令列表，如下：

```
tiup cluster audit
```

```
Starting component `cluster`: /home/tidb/.tiup/components/cluster/v1.9.0/cluster audit
ID      Time                       Command
--      ----                       -------
4BLhr0  2022-03-01T13:25:09+08:00  /home/tidb/.tiup/components/cluster/v1.9.0/cluster deploy test
    ↪  v5.4.0 /tmp/topology.yaml
4BKWjF  2022-02-28T23:36:57+08:00  /home/tidb/.tiup/components/cluster/v1.9.0/cluster deploy test
    ↪  v5.4.0 /tmp/topology.yaml
4BKVwH  2022-02-28T23:02:08+08:00  /home/tidb/.tiup/components/cluster/v1.9.0/cluster deploy test
    ↪  v5.4.0 /tmp/topology.yaml
4BKKH1  2022-02-28T16:39:04+08:00  /home/tidb/.tiup/components/cluster/v1.9.0/cluster destroy
    ↪ test
4BKKDx  2022-02-28T16:36:57+08:00  /home/tidb/.tiup/components/cluster/v1.9.0/cluster deploy test
    ↪  v5.4.0 /tmp/topology.yaml
```

第一列为 audit-id，如果想看某个命令的执行日志，则传入这个 audit-id：

```
tiup cluster audit 4BLhr0
```

12.4.11.2.12 在集群节点机器上执行命令

exec 命令可以很方便地到集群的机器上执行命令，其使用方式如下：

```
Usage:
  tiup cluster exec <cluster-name> [flags]

Flags:
      --command string    需要执行的命令（默认是 "ls"）
  -h, --help              帮助信息
  -N, --node strings      指定要执行的节点 ID （节点 id 通过 display 命令获取）
  -R, --role strings      指定要执行的 role
      --sudo              是否使用 root（默认为 false）
```

例如，如果要到所有的 TiDB 节点上执行 `ls /tmp`：

```
tiup cluster exec test-cluster --command='ls /tmp'
```

12.4.11.2.13 集群控制工具 (controllers)

在 TiUP 之前，我们用 tidb-ctl、tikv-ctl、pd-ctl 等工具操控集群，为了方便下载和使用，TiUP 将它们集成到了统一的组件 ctl 中：

```
Usage:
  tiup ctl {tidb/pd/tikv/binlog/etcd} [flags]

Flags:
  -h, --help   help for tiup
```

这个命令和之前的命令对应关系为：

```
tidb-ctl [args] = tiup ctl tidb [args]
pd-ctl [args] = tiup ctl pd [args]
tikv-ctl [args] = tiup ctl tikv [args]
binlogctl [args] = tiup ctl bindlog [args]
etcdctl [args] = tiup ctl etcd [args]
```

例如，以前查看 store 的命令为 `pd-ctl -u http://127.0.0.1:2379 store`，集成到 TiUP 中的命令为：

```
tiup ctl pd -u http://127.0.0.1:2379 store
```

12.4.11.2.14 部署机环境检查

使用 check 子命令可以对部署机的环境进行一系列检查，并输出检查结果。通过执行 check 子命令，可以发现常见的不合理配置或不支持情况。命令参数列表如下：

```
Usage:
  tiup cluster check <topology.yml | cluster-name> [flags]
```

```
Flags:
      --apply                 Try to fix failed checks
      --cluster               Check existing cluster, the input is a cluster name.
      --enable-cpu            Enable CPU thread count check
      --enable-disk           Enable disk IO (fio) check
      --enable-mem            Enable memory size check
  -h, --help                  help for check
  -i, --identity_file string  The path of the SSH identity file. If specified, public key
      ↪ authentication will be used.
  -p, --password              Use password of target hosts. If specified, password
      ↪ authentication will be used.
      --user string           The user name to login via SSH. The user must has root (or sudo)
          ↪ privilege.
```

默认情况下，此功能用于在部署前进行环境检查，通过指定 `--cluster` 参数切换模式，也可以用于对已部署集群的部署机进行检查，例如：

```
#### check deploy servers before deploy
tiup cluster check topology.yml --user tidb -p

#### check deploy servers of an existing cluster
tiup cluster check <cluster-name> --cluster
```

其中，CPU 线程数检查、内存大小检查和磁盘性能检查三项默认关闭，对于生产环境，建议将此三项检测开启并确保通过，以获得最佳性能。

- CPU：线程数大于等于 16 为通过检查
- 内存：物理内存总大小大于等于 32 GB 为通过检查
- 磁盘：对 `data_dir` 所在分区执行 `fio` 测试并记录结果

在运行检测时，若指定了 `--apply` 参数，程序将尝试对其中未通过的项目自动修复。自动修复仅限于部分可通过修改配置或系统参数调整的项目，其它未修复的项目需要根据实际情况手工处理。

环境检查不是部署集群的必需流程。对于生产环境建议在部署前执行环境检查并通过所有检测项。如果未通过全部检查项，也可能正常部署和运行集群，但可能无法获得最佳性能表现。

12.4.11.2.15 使用中控机系统自带的 SSH 客户端连接集群

在以上所有操作中，涉及到对集群机器的操作都是通过 TiUP 内置的 SSH 客户端连接集群执行命令，但是在某些场景下，需要使用系统自带的 SSH 客户端来对集群执行操作，比如：

- 使用 SSH 插件来做认证
- 使用定制的 SSH 客户端

此时可以通过命令行参数 `--native-ssh` 启用系统自带命令行：

- 部署集群: `tiup cluster deploy <cluster-name> <version> <topo> --native-ssh`
- 启动集群: `tiup cluster start <cluster-name> --native-ssh`
- 升级集群: `tiup cluster upgrade ... --native-ssh`

所有涉及集群操作的步骤都可以加上 `--native-ssh` 来使用系统自带的客户端。

也可以使用环境变量 `TIUP_NATIVE_SSH` 来指定是否使用本地 SSH 客户端，避免每个命令都需要添加 `--native-ssh` 参数：

```
export TIUP_NATIVE_SSH=true
#### 或者
export TIUP_NATIVE_SSh=1
#### 或者
export TIUP_NATIVE_SSH=enable
```

若环境变量和 `--native-ssh` 同时指定，则以 `--native-ssh` 为准。

> 注意：
>
> 在部署集群的步骤中，若需要使用密码的方式连接 (-p)，或者密钥文件设置了 passphrase，则需要保证中控机上安装了 sshpass，否则连接时会报错。

12.4.11.2.16 迁移中控机与备份

TiUP 相关的数据都存储在用户 home 目录的 `.tiup` 目录下，若要迁移中控机只需要拷贝 `.tiup` 目录到对应目标机器即可。

1. 在原机器 home 目录下执行 `tar czvf tiup.tar.gz .tiup`。
2. 把 `tiup.tar.gz` 拷贝到目标机器 home 目录。
3. 在目标机器 home 目录下执行 `tar xzvf tiup.tar.gz`。
4. 添加 `.tiup` 目录到 PATH 环境变量。

   如使用 bash 并且是 `tidb` 用户，在 `~/.bashrc` 中添加 `export PATH=/home/tidb/.tiup/bin:$PATH` 后执行 `source ~/.bashrc`，根据使用的 shell 与用户做相应调整。

> 注意：
>
> 为了避免中控机磁盘损坏等异常情况导致 TiUP 数据丢失，建议定时备份 `.tiup` 目录。

12.4.11.3 搭建私有镜像

在构建私有云时，通常会使用隔离的网络环境，此时无法访问 TiUP 的官方镜像。因此，TiUP 提供了构建私有镜像的方案，它主要由 mirror 指令来实现，该方案也可用于离线部署。

12.4.11.3.1 mirror 指令介绍

mirror 指令的帮助文档如下：

```
tiup mirror --help
```

```
The 'mirror' command is used to manage a component repository for TiUP, you can use
it to create a private repository, or to add new component to an existing repository.
The repository can be used either online or offline.
It also provides some useful utilities to help managing keys, users and versions
of components or the repository itself.

Usage:
  tiup mirror <command> [flags]

Available Commands:
  init        Initialize an empty repository
  sign        Add signatures to a manifest file
  genkey      Generate a new key pair
  clone       Clone a local mirror from remote mirror and download all selected components
  publish     Publish a component

Flags:
  -h, --help          help for mirror
      --repo string   Path to the repository

Global Flags:
      --skip-version-check   Skip the strict version check, by default a version must be a valid
          ↪ SemVer string

Use "tiup mirror [command] --help" for more information about a command.
```

构建本地镜像将会使用 tiup mirror clone 指令，其基本用法如下：

```
tiup mirror clone <target-dir> [global-version] [flags]
```

- target-dir：指需要把克隆下来的数据放到哪个目录里。
- global-version：用于为所有组件快速设置一个共同的版本。

tiup mirror clone 命令提供了很多可选参数，日后可能会提供更多。但这些参数其实可以分为四类：

1. 克隆时是否使用前缀匹配方式匹配版本

   如果指定了 --prefix 参数，则会才用前缀匹配方式匹配克隆的版本号。例：指定 --prefix 时，填写版本“v5.0.0”将会匹配“v5.0.0-rc”，“v5.0.0”

2. 是否全量克隆

   如果指定了 --full 参数，则会完整地克隆官方镜像。

> **注意：**
>
> 如果既不指定 `--full` 参数，又不指定 `global-version` 或克隆的 component 版本，那么 TiUP 就只会克隆一些元信息。

3. 限定只克隆特定平台的包

   如果只想克隆某个平台的包，那么可以使用 `--os` 和 `--arch` 来限定：

   - 只想克隆 linux 平台的，则执行 `tiup mirror clone <target-dir> [global-version] --os=linux`
   - 只想克隆 amd64 架构的，则执行 `tiup mirror clone <target-dir> [global-version] --arch=amd64`
   - 只想克隆 linux/amd64 的，则执行 `tiup mirror clone <target-dir> [global-version] --os=linux --arch=amd64`

4. 限定只克隆组件的特定版本

   如果只想克隆某个组件的某一个版本而不是所有版本，则使用 `--<component>=<version>` 来限定，例如：

   - 只想克隆 TiDB 的 v5.4.0 版本，则执行 `tiup mirror clone <target-dir> --tidb v5.4.0`
   - 只想克隆 TiDB 的 v5.4.0 版本，以及 TiKV 的所有版本，则执行 `tiup mirror clone <target-dir> --tidb v5.4.0 --tikv all`
   - 克隆一个集群的所有组件的 v5.4.0 版本，则执行 `tiup mirror clone <target-dir> v5.4.0`

12.4.11.3.2 使用示例

使用 `tiup mirror clone` 命令克隆的仓库可以在主机之间共享。可以通过 SCP、NFS 共享文件，也可以通过 HTTP 或 HTTPS 协议使用仓库。用 `tiup mirror set <location>` 命令来指定仓库的位置。

参考使用 TiUP 部署 TiDB 集群安装 TiUP 离线镜像，部署并启动 TiDB 集群。

12.4.11.4 使用 TiUP bench 组件压测 TiDB

在测试数据库性能时，经常需要对数据库进行压测，为了满足这一需求，TiUP 集成了 bench 组件。目前，TiUP bench 组件提供 TPC-C 和 TPC-H 两种压测的 workload，其命令参数如下，更多信息请参考 TPC 官方网站。

```
tiup bench
```

```
Starting component `bench`: /home/tidb/.tiup/components/bench/v1.5.0/bench
Benchmark database with different workloads

Usage:
  tiup bench [command]

Available Commands:
  help        帮助信息
  tpcc        以 TPC-C 作为 workload 压测
  tpch        以 TPC-H 作为 workload 压测
```

```
Flags:
      --count int           总执行次数，0 表示无限次
  -D, --db string           被压测数据库名称（默认为 "test"）
  -d, --driver string       数据库驱动：mysql
      --dropdata            在 prepare 之前清除历史数据
  -h, --help                bench 命令自身的帮助信息
  -H, --host string         数据库的主机地址（默认 "127.0.0.1"）
      --ignore-error        忽略压测时数据库报出的错误
      --interval duration   两次报告输出时间的间隔（默认 10s）
      --isolation int       隔离级别 0: Default, 1: ReadUncommitted,
                            2: ReadCommitted, 3: WriteCommitted, 4: RepeatableRead,
                            5: Snapshot, 6: Serializable, 7: Linerizable
      --max-procs int       runtime.GOMAXPROCS
  -p, --password string     数据库密码
  -P, --port int            数据库端口（默认 4000）
      --pprof string        pprof 地址
      --silence             压测过程中不打印错误信息
      --summary             只打印 Summary
  -T, --threads int         压测并发线程数（默认 16）
      --time duration       总执行时长（默认 2562047h47m16.854775807s）
  -U, --user string         压测时使用的数据库用户（默认 "root"）
```

下文分别介绍如何使用 TiUP 运行 TPC-C 测试和 TPC-H 测试。

12.4.11.4.1 使用 TiUP 运行 TPC-C 测试

TiUP bench 组件支持如下运行 TPC-C 测试的命令和参数：

```
Available Commands:
  check       检查数据一致性
  cleanup     清除数据
  prepare     准备数据
  run         开始压测

Flags:
      --check-all         运行所有的一致性检测
  -h, --help              tpcc 的帮助信息
      --output string     准备数据时生成 csv 文件的目录
      --parts int         分区仓库 的数量(默认 1)
      --tables string     指定用于生成文件的表，多个表用逗号分割，只有设置了 output 时才有效。
          ↪ 默认生成所有的表
      --warehouses int    仓库的数量（默认 10）
```

TPC-C 测试步骤

1. 通过 HASH 使用 4 个分区创建 4 个仓库：

```
tiup bench tpcc --warehouses 4 --parts 4 prepare
```

2. 运行 TPC-C 测试：

```
tiup bench tpcc --warehouses 4 run
```

3. 清理数据：

```
tiup bench tpcc --warehouses 4 cleanup
```

4. 检查一致性：

```
tiup bench tpcc --warehouses 4 check
```

5. 生成 CSV 文件：

```
tiup bench tpcc --warehouses 4 prepare --output data
```

6. 为指定的表生成 CSV 文件：

```
tiup bench tpcc --warehouses 4 prepare --output data --tables history,orders
```

7. 开启 pprof：

```
tiup bench tpcc --warehouses 4 prepare --output data --pprof :10111
```

12.4.11.4.2 使用 TiUP 运行 TPC-H 测试

TiUP bench 组件支持如下运行 TPC-H 测试的命令和参数：

```
Available Commands:
  cleanup     清除数据
  prepare     准备数据
  run         开始压测

Flags:
      --check             检查输出数据，只有 scale 因子为 1 时有效
  -h, --help              tpch 的帮助信息
      --queries string    所有的查询语句（默认 "q1,q2,q3,q4,q5,q6,q7,q8,q9,q10,q11,q12,q13,q14,q15
          ↪ ,q16,q17,q18,q19,q20,q21,q22")
      --sf int            scale 因子
```

12.4.11.4.3 TPC-H 测试步骤

1. 准备数据：

```
tiup bench tpch --sf=1 prepare
```

2. 运行 TPC-H 测试，根据是否检查结果执行相应命令：

   - 检查结果：

   ```
   tiup bench tpch --sf=1 --check=true run
   ```

   - 不检查结果：

   ```
   tiup bench tpch --sf=1 run
   ```

3. 清理数据：

```
tiup bench tpch cleanup
```

12.5 TiDB Operator

TiDB Operator 是 Kubernetes 上的 TiDB 集群自动运维系统，提供包括部署、升级、扩缩容、备份恢复、配置变更的 TiDB 全生命周期管理。借助 TiDB Operator，TiDB 可以无缝运行在公有云或私有部署的 Kubernetes 集群上。

TiDB Operator 的文档目前独立于 TiDB 文档，文档名称为 TiDB in Kubernetes 用户文档。要访问 TiDB Operator 的文档，请点击以下链接：

- TiDB in Kubernetes 用户文档

12.6 使用 Dumpling 导出数据

使用数据导出工具 Dumpling，你可以把存储在 TiDB 或 MySQL 中的数据导出为 SQL 或 CSV 格式，用于逻辑全量备份。Dumpling 也支持将数据导出到 Amazon S3 中。

下图展示了使用 Dumpling 导出数据的场景。

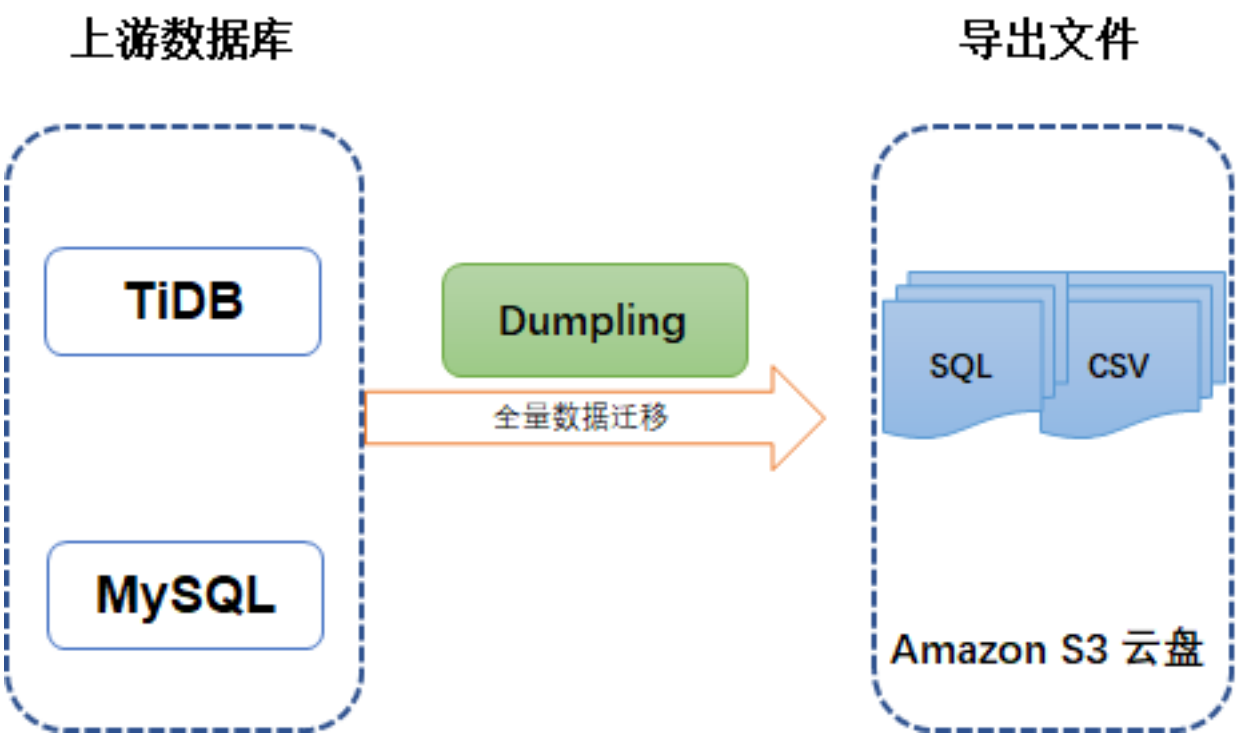


图 259: Dumpling 导出数据

要快速了解 Dumpling 的基本功能，建议先观看下面的培训视频（时长 28 分钟）。注意本视频只作为功能介绍、学习参考，具体操作步骤和最新功能，请以文档内容为准。

你可以通过下列任意方式获取 Dumpling：

- TiUP 执行 `tiup install dumpling` 命令。获取后，使用 `tiup dumpling ...` 命令运行 Dumpling。
- 下载包含 Dumpling 的tidb-toolkit 安装包。

更多详情，可以使用 –help 选项查看，或参考Dumpling 主要选项表。

使用 Dumpling 时，需要在已经启动的集群上执行导出命令。

TiDB 还提供了其他工具，你可以根据需要选择使用：

- 如果需要直接备份 SST 文件（键值对），或者对延迟不敏感的增量备份，请使用备份工具BR。
- 如果需要实时的增量备份，请使用TiCDC。
- 所有的导出数据都可以用TiDB Lightning 导回到 TiDB。

> 注意：
>
> PingCAP 之前维护的 Mydumper 工具 fork 自 mydumper project，针对 TiDB 的特性进行了优化。关于 Mydumper 的更多信息，请参考 v4.0 版 Mydumper 使用文档。Mydumper 目前已经不再开发新功能，其绝大部分功能已经被Dumpling 取代，请切换到 Dumpling。

相比 Mydumper，Dumpling 做了如下改进：

- 支持导出多种数据形式，包括 SQL/CSV。
- 支持全新的 table-filter，筛选数据更加方便。
- 支持导出到 Amazon S3 云盘。
- 针对 TiDB 进行了更多优化：
  - 支持配置 TiDB 单条 SQL 内存限制。
  - 针对 TiDB v4.0.0 及更新版本支持自动调整 TiDB GC 时间。
  - 使用 TiDB 的隐藏列 `_tidb_rowid` 优化了单表内数据的并发导出性能。
  - 对于 TiDB 可以设置 tidb_snapshot 的值指定备份数据的时间点，从而保证备份的一致性，而不是通过 `FLUSH TABLES WITH READ LOCK` 来保证备份一致性。

12.6.1 从 TiDB/MySQL 导出数据

12.6.1.1 需要的权限

- SELECT
- RELOAD
- LOCK TABLES
- REPLICATION CLIENT
- PROCESS

12.6.1.2 导出为 SQL 文件

本文假设在 `127.0.0.1:4000` 有一个 TiDB 实例，并且这个 TiDB 实例中有无密码的 root 用户。

Dumpling 默认导出数据格式为 SQL 文件。也可以通过设置 `--filetype sql` 导出数据到 SQL 文件：

```
dumpling \
  -u root \
  -P 4000 \
  -h 127.0.0.1 \
  --filetype sql \
  -t 8 \
  -o /tmp/test \
  -r 200000 \
  -F 256MiB
```

以上命令中：

- `-h`、`-P`、`-u` 分别代表地址、端口、用户。如果需要密码验证，可以使用 `-p $YOUR_SECRET_PASSWORD` 将密码传给 Dumpling。
- `-o` 用于选择存储导出文件的目录，支持本地文件路径或外部存储 URL 格式。
- `-t` 用于指定导出的线程数。增加线程数会增加 Dumpling 并发度提高导出速度，但也会加大数据库内存消耗，因此不宜设置过大。一般不超过 64。

- `-r` 用于指定单个文件的最大行数，指定该参数后 Dumpling 会开启表内并发加速导出，同时减少内存使用。当上游为 TiDB 且版本为 v3.0 或更新版本时，该参数大于 0 表示使用 TiDB region 信息划分表内并发，具体取值将不再生效。
- `-F` 选项用于指定单个文件的最大大小，单位为 `MiB`，可接受类似 `5GiB` 或 `8KB` 的输入。如果你想使用 TiDB Lightning 将该文件加载到 TiDB 实例中，建议将 `-F` 选项的值保持在 256 MiB 或以下。

> **注意：**
>
> 如果导出的单表大小超过 10 GB，强烈建议使用 `-r` 和 `-F` 参数。

12.6.1.3 导出为 CSV 文件

假如导出数据的格式是 CSV（使用 `--filetype csv` 即可导出 CSV 文件），还可以使用 `--sql <SQL>` 导出指定 SQL 选择出来的记录，例如，导出 `test.sbtest1` 中所有 `id < 100` 的记录：

```
./dumpling \
  -u root \
  -P 4000 \
  -h 127.0.0.1 \
  -o /tmp/test \
  --filetype csv \
  --sql 'select * from `test`.`sbtest1` where id < 100'
```

> **注意：**
>
> - `--sql` 选项仅仅可用于导出 CSV 的场景。
> - 该命令将在要导出的所有表上执行 `select * from <table-name> where id < 100` 语句。如果部分表没有指定的字段，那么导出会失败。
> - Dumpling 导出不区分字符串与关键字。如果导入的数据是 Boolean 类型的 `true` 和 `false`，导出时会被转换为 1 和 0。

12.6.1.4 输出文件格式

- `metadata`：此文件包含导出的起始时间，以及 master binary log 的位置。

```
cat metadata
```

```
Started dump at: 2020-11-10 10:40:19
SHOW MASTER STATUS:
        Log: tidb-binlog
        Pos: 420747102018863124
```

```
Finished dump at: 2020-11-10 10:40:20
```

- `{schema}-schema-create.sql`：创建 schema 的 SQL 文件。

```
cat test-schema-create.sql
```

```
CREATE DATABASE `test` /*!40100 DEFAULT CHARACTER SET utf8mb4 */;
```

- `{schema}.{table}-schema.sql`：创建 table 的 SQL 文件

```
cat test.t1-schema.sql
```

```
CREATE TABLE `t1` (
  `id` int(11) DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_bin;
```

- `{schema}.{table}.{0001}.{sql|csv}`：数据源文件

```
cat test.t1.0.sql
```

```
/*!40101 SET NAMES binary*/;
INSERT INTO `t1` VALUES
(1);
```

- `*-schema-view.sql`、`*-schema-trigger.sql`、`*-schema-post.sql`：其他导出文件

12.6.1.5 导出到 Amazon S3 云盘

Dumpling 在 v4.0.8 及更新版本支持导出到 Amazon S3 云盘。如果需要将数据备份到 Amazon S3 后端存储，那么需要在 `-o` 参数中指定 Amazon S3 的存储路径。

可以参照 AWS 官方文档 - 如何创建 S3 存储桶在指定的 `Region` 区域中创建一个 S3 桶 `Bucket`。如有需要，还可以参照 AWS 官方文档 - 创建文件夹在 Bucket 中创建一个文件夹 `Folder`。

将有权限访问该 Amazon S3 后端存储的账号的 `SecretKey` 和 `AccessKey` 作为环境变量传入 Dumpling 节点。

```
export AWS_ACCESS_KEY_ID=${AccessKey}
export AWS_SECRET_ACCESS_KEY=${SecretKey}
```

Dumpling 同时还支持从 `~/.aws/credentials` 读取凭证文件。更多 Dumpling 存储配置可以参考外部存储。

在进行 Dumpling 备份时，显式指定参数 `--s3.region`，即表示 Amazon S3 存储所在的区域，例如 `ap-northeast-1`。

```
./dumpling \
  -u root \
  -P 4000 \
  -h 127.0.0.1 \
  -r 200000 \
  -o "s3://${Bucket}/${Folder}" \
  --s3.region "${region}"
```

12.6.1.6 筛选导出的数据

12.6.1.6.1 使用 --where 选项筛选数据

默认情况下，Dumpling 会导出排除系统数据库（包括 mysql、sys、INFORMATION_SCHEMA、PERFORMANCE_SCHEMA、METRICS_SCHEMA 和 INSPECTION_SCHEMA）外所有其他数据库。你可以使用 --where <SQL where expression> 来指定要导出的记录。

```
./dumpling \
  -u root \
  -P 4000 \
  -h 127.0.0.1 \
  -o /tmp/test \
  --where "id < 100"
```

上述命令将会导出各个表的 id < 100 的数据。注意 --where 参数无法与 --sql 一起使用。

12.6.1.6.2 使用 --filter 选项筛选数据

Dumpling 可以通过 --filter 指定 table-filter 来筛选特定的库表。table-filter 的语法与 .gitignore 相似，详细语法参考表库过滤。

```
./dumpling \
  -u root \
  -P 4000 \
  -h 127.0.0.1 \
  -o /tmp/test \
  -r 200000 \
  --filter "employees.*" \
  --filter "*.WorkOrder"
```

上述命令将会导出 employees 数据库的所有表，以及所有数据库中的 WorkOrder 表。

12.6.1.6.3 使用 -B 或 -T 选项筛选数据

Dumpling 也可以通过 -B 或 -T 选项导出特定的数据库/数据表。

> 注意：
>
> - --filter 选项与 -T 选项不可同时使用。
> - -T 选项只能接受完整的 库名.表名 形式，不支持只指定表名。例：Dumpling 无法识别 -T WorkOrder。

例如通过指定：

- `-B employees` 导出 `employees` 数据库
- `-T employees.WorkOrder` 导出 `employees.WorkOrder` 数据表

12.6.1.7 通过并发提高 Dumpling 的导出效率

默认情况下，导出的文件会存储到 `./export-<current local time>` 目录下。常用选项如下：

- `-t` 用于指定导出的线程数。增加线程数会增加 Dumpling 并发度提高导出速度，但也会加大数据库内存消耗，因此不宜设置过大。
- `-r` 选项用于指定单个文件的最大记录数，或者说，数据库中的行数。开启后 Dumpling 会开启表内并发，提高导出大表的速度。当上游为 TiDB 且版本为 v3.0 或更新版本时，该参数大于 0 表示使用 TiDB region 信息划分表内并发，具体取值将不再生效。

利用以上选项可以提高 Dumpling 的导出速度。

12.6.1.8 调整 Dumpling 的数据一致性选项

> 注意：
>
> 数据一致性选项的默认值为 auto。在大多数场景下，你不需要调整该选项。

Dumpling 通过 `--consistency <consistency level>` 标志控制导出数据“一致性保证”的方式。在使用 snapshot 来保证一致性的时候，可以使用 `--snapshot` 选项指定要备份的时间戳。还可以使用以下的一致性级别：

- `flush`：使用 `FLUSH TABLES WITH READ LOCK` 短暂地中断备份库的 DML 和 DDL 操作、保证备份连接的全局一致性和记录 POS 信息。所有的备份连接启动事务后释放该锁。推荐在业务低峰或者 MySQL 备份库上进行全量备份。
- `snapshot`：获取指定时间戳的一致性快照并导出。
- `lock`：为待导出的所有表上读锁。
- `none`：不做任何一致性保证。
- `auto`：对 MySQL 使用 `flush`，对 TiDB 使用 `snapshot`。

操作完成之后，你可以在 `/tmp/test` 查看导出的文件：

```
$ ls -lh /tmp/test | awk '{print $5 "\t" $9}'

140B  metadata
66B   test-schema-create.sql
300B  test.sbtest1-schema.sql
190K  test.sbtest1.0.sql
300B  test.sbtest2-schema.sql
190K  test.sbtest2.0.sql
300B  test.sbtest3-schema.sql
190K  test.sbtest3.0.sql
```

12.6.1.9 导出 TiDB 的历史数据快照

Dumping 可以通过 --snapshot 指定导出某个 tidb_snapshot 的数据。

--snapshot 选项可设为 TSO（SHOW MASTER STATUS 输出的 Position 字段）或有效的 datetime 时间（YYYY-MM-DD hh:mm:ss 形式），例如：

```
./dumpling --snapshot 417773951312461825
./dumpling --snapshot "2020-07-02 17:12:45"
```

即可导出 TSO 为 417773951312461825 或 2020-07-02 17:12:45 时的 TiDB 历史数据快照。

12.6.1.10 控制导出 TiDB 大表时的内存使用

Dumping 导出 TiDB 较大单表时，可能会因为导出数据过大导致 TiDB 内存溢出 (OOM)，从而使连接中断导出失败。可以通过以下参数减少 TiDB 的内存使用。

- 设置 -r 参数，可以划分导出数据区块减少 TiDB 扫描数据的内存开销，同时也可开启表内并发提高导出效率。当上游为 TiDB 且版本为 v3.0 或更新版本时，该参数大于 0 表示使用 TiDB region 信息划分表内并发，具体取值将不再生效。
- 调小 --tidb-mem-quota-query 参数到 8589934592 (8GB) 或更小。可控制 TiDB 单条查询语句的内存使用。
- 调整 --params "tidb_distsql_scan_concurrency=5" 参数，即设置导出时的 session 变量 tidb_distsql_scan_concurrency 从而减少 TiDB scan 操作的并发度。

12.6.1.11 导出大规模数据时的 TiDB GC 设置

如果导出的 TiDB 版本为 v4.0.0 或更新版本，并且 Dumpling 可以访问 TiDB 集群的 PD 地址，Dumpling 会自动配置延长 GC 时间且不会对原集群造成影响。

其他情况下，假如导出的数据量非常大，可以提前调长 GC 时间，以避免因为导出过程中发生 GC 导致导出失败：

```
SET GLOBAL tidb_gc_life_time = '720h';
```

操作结束之后，再恢复 GC 时间为默认值 10m：

```
SET GLOBAL tidb_gc_life_time = '10m';
```

12.6.2 Dumpling 主要选项表

主要选项	用途	默认值
-V 或 –version	输出 Dumpling 版本并直接退出	
-B 或 –database	导出指定数据库	
-T 或 –tables-list	导出指定数据表	
-f 或 –filter	导出能匹配模式的表，语法可参考table-filter	[*.*,!/^(mysql|sys|INFORMATION_SCHEMA|PERFORMANCE_SCHEMA|METRICS_SCHEMA|INSPECTION_SCHEMA)$/.*] (导出除系统库外的所有库表)

主要选项	用途	默认值
–case-sensitive	table-filter 是否大小写敏感	false，大小写不敏感
-h 或 –host	连接的数据库主机的地址	“127.0.0.1”
-t 或 –threads	备份并发线程数	4

主要选项	用途	默认值
-r 或 –rows	将 table 划分成 row 行数据，一般针对大表操作并发生成多个文件。当上游为 TiDB 且版本为 v3.0 或更新版本时，该参数大于 0 表示使用 TiDB region 信息划分表内并发，具体取值将不再生效。	

主要选项	用途	默认值
-L 或 –logfile	日志输出地址，为空时会输出到控制台	“”
–loglevel	日志级别 {debug,info,warn,error,dpanic,panic,fatal}	“info”
–logfmt	日志输出格式 {text,json}	“text”
-d 或 –no-data	不导出数据，适用于只导出 schema 场景	
–no-header	导出 csv 格式的 table 数据，不生成 header	
-W 或 –no-views	不导出 view	true
-m 或 –no-schemas	不导出 schema，只导出数据	

主要选项	用途	默认值
-s 或 –statement-size	控制 INSERT SQL 语句的大小，单位 bytes	
-F 或 –filesize	将 table 数据划分出来的文件大小，需指明单位 (如 128B, 64 KiB , 32 MiB , 1.5 GiB)	
–filetype	导出文件类型 (csv/sql)	“sql”

主要选项	用途	默认值
-o 或 –output	导出本地文件路径或外部存储 URL	“./export-${time}”
-S 或 –sql	根据指定的 sql 导出数据，该选项不支持并发导出	

主要选项	用途	默认值
–consistency	flush: dump 前用 FTWRL snap-shot: 通过 TSO 来指定 dump 某个快照时间点的 TiDB 数据 lock: 对需要 dump 的所有表执行 lock ↪ ↪ tables ↪ ↪ read ↪ 命令 none: 不加锁 dump，无法保证一致性 auto: 对 MySQL 使用 –consistency flush；对 TiDB	“auto”

主要选项	用途	默认值
–snapshot	snapshot tso，只在 consistency=snapshot 下生效	
–where	对备份的数据表通过 where 条件指定范围	
-p 或 –password	连接的数据库主机的密码	
-P 或 –port	连接的数据库主机的端口	4000
-u 或 –user	连接的数据库主机的用户名	“root”
–dump-empty-database	导出空数据库的建库语句	true

主要选项	用途	默认值
–ca	用于 TLS 连接的 certifi-cate au-thor-ity 文件的地址	
–cert	用于 TLS 连接的 client certifi-cate 文件的地址	
–key	用于 TLS 连接的 client pri-vate key 文件的地址	
–csv-delimiter	csv 文件中字符类型变量的定界符	’ “’

主要选项	用途	默认值
–csv-separator	csv 文件中各值的分隔符，如果数据中可能有逗号，建议源文件导出时分隔符使用非常见组合字符	‘,’
–csv-null-value	csv 文件空值的表示	“\N”
–escape-backslash	使用反斜杠 (\) 来转义导出文件中的特殊字符	true

主要选项	用途	默认值
–output-filename-template	以 golang tem-plate 格式表示的数据文件名格式支持 {{. ↪ DB ↪ }} ↪ 、 {{. ↪ Table ↪ }} ↪ 、 {{. ↪ Index ↪ }} ↪ 三个参数分别表示数据文件的库名、表名、分块 ID	‘{{.DB}}.{{.Table}}.{{.Index}}’

主要选项	用途	默认值
–status-addr	Dumpling 的服务地址，包含了 Prometheus 拉取 metrics 信息及 pprof 调试的地址	“:8281”

主要选项	用途	默认值
–tidb-mem-quota-query	单条 dumpling 命令导出 SQL 语句的内存限制，单位为 byte。对于 v4.0.10 或以上版本，若不设置该参数，默认使用 TiDB 中的 mem-quota-query 配置项值作为内存限制值。对于 v4.0.10 以下版本，该参数值默认为 32 GB	34359738368

主要选项	用途	默认值
–params	为需导出的数据库连接指定 session 变量，可接受的格式: “character_set_client=latin1,character_set_connection=latin1”	

12.7 TiDB Lightning

12.7.1 TiDB Lightning 简介

TiDB Lightning 是一个将全量数据高速导入到 TiDB 集群的工具，可在此下载。

TiDB Lightning 有以下两个主要的使用场景：

- 迅速导入大量新数据。
- 恢复所有备份数据。

目前，TiDB Lightning 支持：

- 导入Dumpling输出的数据、CSV 文件或Amazon Aurora 生成的 Apache Parquet 文件。
- 从本地盘或Amazon S3 云盘读取数据。

要快速了解 Lightning 的基本原理和使用方法，建议先观看下面的培训视频（时长 28 分钟）。注意本视频只为学习参考，具体操作步骤和最新功能，请以文档内容为准。

12.7.1.1 TiDB Lightning 整体架构

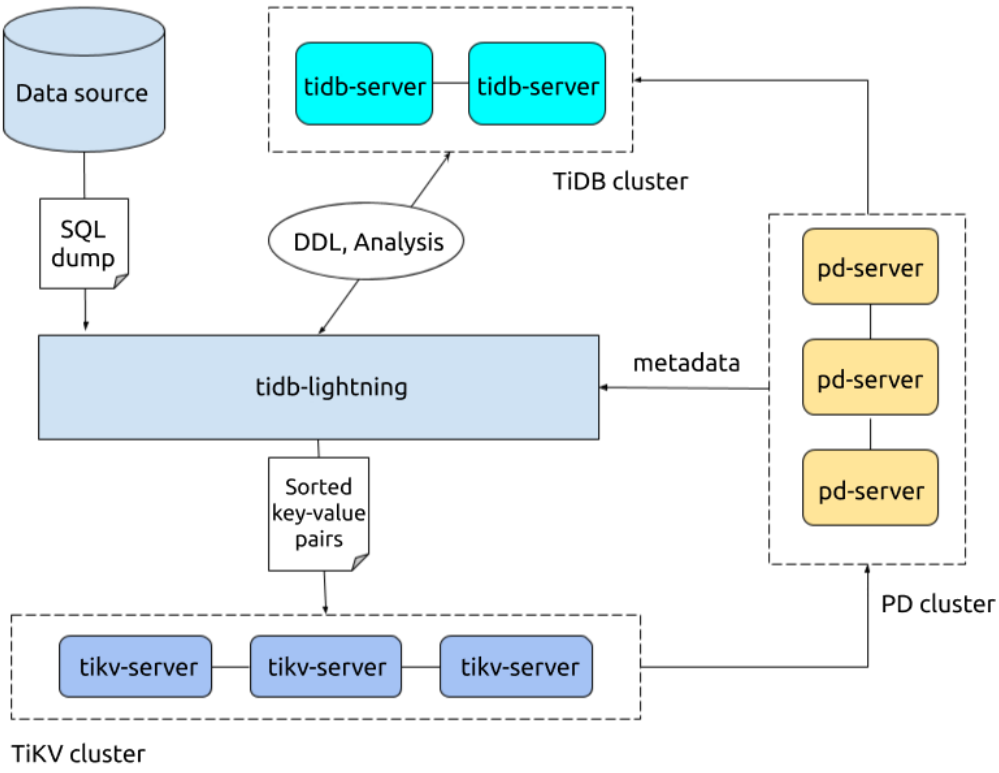


图 260: TiDB Lightning 整体架构

TiDB Lightning 整体工作原理如下：

1. 在导入数据之前，`tidb-lightning` 会自动将 TiKV 集群切换为“导入模式” (import mode)，优化写入效率并停止自动压缩。
2. `tidb-lightning` 会在目标数据库建立架构和表，并获取其元数据。
3. 每张表都会被分割为多个连续的区块，这样来自大表 (200 GB+) 的数据就可以用增量方式并行导入。
4. `tidb-lightning` 会为每一个区块准备一个“引擎文件 (engine file)”来处理键值对。`tidb-lightning` 会并发读取 SQL dump，将数据源转换成与 TiDB 相同编码的键值对，然后将这些键值对排序写入本地临时存储文件中。
5. 当一个引擎文件数据写入完毕时，`tidb-lightning` 便开始对目标 TiKV 集群数据进行分裂和调度，然后导入数据到 TiKV 集群。

   引擎文件包含两种：数据引擎与索引引擎，各自又对应两种键值对：行数据和次级索引。通常行数据在数据源里是完全有序的，而次级索引是无序的。因此，数据引擎文件在对应区块写入完成后会被立即上传，而所有的索引引擎文件只有在整张表所有区块编码完成后才会执行导入。

6. 整张表相关联的所有引擎文件完成导入后，tidb-lightning 会对比本地数据源及下游集群的校验和 (checksum)，确保导入的数据无损，然后让 TiDB 分析 (ANALYZE) 这些新增的数据，以优化日后的操作。同时，tidb-lightning 调整 AUTO_INCREMENT 值防止之后新增数据时发生冲突。

   表的自增 ID 是通过行数的上界估计值得到的，与表的数据文件总大小成正比。因此，最后的自增 ID 通常比实际行数大得多。这属于正常现象，因为在 TiDB 中自增 ID 不一定是连续分配的。

7. 在所有步骤完毕后，tidb-lightning 自动将 TiKV 切换回“普通模式” (normal mode)，此后 TiDB 集群可以正常对外提供服务。

如果需要导入的目标集群是 v3.x 或以下的版本，需要使用 Importer-backend 来完成数据的导入。在这个模式下，tidb-lightning 需要将解析的键值对通过 gRPC 发送给 tikv-importer 并由 tikv-importer 完成数据的导入。

TiDB Lightning 还支持使用 TiDB-backend 作为后端导入数据：tidb-lightning 将数据转换为 INSERT 语句，然后直接在目标集群上执行这些语句。详见TiDB Lightning Backends。

12.7.1.2 使用限制

- TiDB Lightning 与 TiFlash 一起使用时需要注意：

  无论是否已为一张表创建 TiFlash 副本，你都可以使用 TiDB Lightning 导入数据至该表。但该场景下，TiDB Lightning 导入数据耗费的时间更长，具体取决于 TiDB Lightning 部署机器的网卡带宽、TiFlash 节点的 CPU 及磁盘负载及 TiFlash 副本数等因素。

- TiDB Lightning 与 TiDB 一起使用时需要注意：

  TiDB Lightning 在 v5.4.0 之前不支持导入 charset=GBK 的表。

- 如果源数据是 Apache Parquet 文件，TiDB Lightning 目前只支持由 Amazon Aurora 生成的 Apache Parquet 文件。

12.7.2 前置需求

12.7.2.1 前置检查

从 TiDB 5.3.0 开始，TiDB Lightning 增加了前置检查功能，可以在执行迁移前检查配置。默认开启。该功能会自动进行一些磁盘空间和执行配置的常规检查，主要目的是确保后续的整个导入过程顺利。

下表介绍了各检查项和详细解释。

检查项	支持版本	解释
集群版本/状态是否正常	>= 5.3.0	检查配置中集群是否可以连接，当后端模式是 Local 的时候，还会检查 TiKV/PD/TiFlash 版本是否支持 Local 导入模式。

检查项	支持版本	解释
是否有权限读取数据	>= 5.3.0	检查当从云存储（Amazon S3）读取数据的时候，是否有对应的权限，确保不会因权限缺失导致导入中断。

检查项	支持版本	解释
导入空间是否足够	>= 5.3.0	检查 TiKV 集群是否有足够空间导入数据。检查时会对数据源进行采样，通过采样结果预估索引大小占比。由于估算中考虑了索引，因此可能会出现尽管数据源大小低于本地盘可用空间，但依然无法通过检测的情况。当后端是 Local 的时

检查项	支持版本	解释
Region 分布状态	>= 5.3.0	检查 TiKV 集群的 Region 分布是否均匀，以及是否存在大量空 region，如果空 Region 的数量大于 max(1000, 表的数量 * 3)，即大于“1000”和“3 倍表数量”二者中的较大者，TiDB Lightning 无法执行导入。

检查项	支持版本	解释
数据文件是否有大 CSV 文件	>= 5.3.0	当备份文件中出现大于 10 GiB 的 CSV 文件且无法进行自动切分 (Strict-For-mat=false) 的时候，会导致导入性能下降。该检查的目的是提醒用户确保数据格式的情况下，开启自动切分 CSV 功能。

检查项	支持版本	解释
是否可以从断点恢复	>= 5.3.0	该检查是确保断点恢复过程中，不会出现对源文件和数据库中 schema 进行修改，导致导入错误数据的情况。

检查项	支持版本	解释
是否可以导入数据到已存在的数据表中	>= 5.3.0	当导入到已创建好的数据表中时，该检查尽可能的检查此次导入的源文件是否和已存在的数据表匹配。检查列数是否匹配，如果源文件存在列名，则检查列名是否匹配。当源文件存在缺省列，则检查缺省列在数据表中是否存在

检查项	支持版本	解释
导入的目标表是否为空	>= 5.3.1	如果导入的目标表不为空，则 TiDB Lightning 会自动报错退出；如果开启了并行导入模式 (incremental-import = true)，则会跳过此检查项。

12.7.2.2 TiDB Lightning 使用前提

使用 TiDB Lightning 导入数据前，先检查环境是否满足要求，这有助于减少导入过程的错误，避免导入失败的情况。

12.7.2.2.1 下游数据库权限要求

TiDB Lightning 导入数据时，根据导入方式和启用特性等，需要下游数据库用户具备不同的权限，可参考下表：

```
<td></td>
<td>特性</td>
<td>作用域</td>
<td>所需权限</td>
<td>备注</td>
```

```
<td rowspan="2">必需</td>
<td rowspan="2">基本功能</td>
```

```
<td>目标 table</td>
<td>CREATE,SELECT,INSERT,UPDATE,DELETE,DROP,ALTER</td>
<td>DROP 仅 tidb-lightning-ctl 在执行 checkpoint-destroy-all 时需要</td>
```

```
<td>目标 database</td>
<td>CREATE</td>
<td></td>
```

```
<td rowspan="4">必需</td>
<td>tidb-backend</td>
<td>information_schema.columns</td>
<td>SELECT</td>
<td></td>
```

```
<td  rowspan="3">local-backend</td>
<td>mysql.tidb</td>
<td>SELECT</td>
<td></td>
```

```
<td>-</td>
<td>SUPER</td>
<td></td>
```

```
<td>-</td>
<td>RESTRICTED_VARIABLES_ADMIN,RESTRICTED_TABLES_ADMIN</td>
<td>当目标 TiDB 开启 SEM</td>
```

```
<td>推荐</td>
<td>冲突检测，max-error</td>
<td>lightning.task-info-schema-name 配置的 schema</td>
<td>SELECT,INSERT,UPDATE,DELETE,CREATE,DROP</td>
<td>如不需要，该值必须设为""</td>
```

```
<td>可选</td>
<td>并行导入</td>
<td>lightning.meta-schema-name 配置的 schema</td>
<td>SELECT,INSERT,UPDATE,DELETE,CREATE,DROP</td>
<td>如不需要，该值必须设为""</td>
```

```
<td>可选</td>
<td>checkpoint.driver = “mysql” </td>
<td>checkpoint.schema 设置</td>
<td>SELECT,INSERT,UPDATE,DELETE,CREATE,DROP</td>
<td>使用数据库而非文件形式存放 checkpoint 信息时需要</td>
```

12.7.2.2.2 下游数据库所需空间

目标 TiKV 集群必须有足够空间接收新导入的数据。除了标准硬件配置以外，目标 TiKV 集群的总存储空间必须大于 数据源大小 × 副本数量 × 2。例如集群默认使用 3 副本，那么总存储空间需为数据源大小的 6 倍以上。公式中的 2 倍可能难以理解，其依据是以下因素的估算空间占用：

- 索引会占据额外的空间
- RocksDB 的空间放大效应

目前无法精确计算 Dumpling 从 MySQL 导出的数据大小，但你可以用下面 SQL 语句统计信息表的 data_length 字段估算数据量：

统计所有 schema 大小，单位 MiB，注意修改 ${schema_name}

```
select table_schema,sum(data_length)/1024/1024 as data_length,sum(index_length)/1024/1024 as
    ↪ index_length,sum(data_length+index_length)/1024/1024 as sum from information_schema.
    ↪ tables where table_schema = "${schema_name}" group by table_schema;
```

统计最大单表，单位 MiB，注意修改 ${schema_name}

```
select table_name,table_schema,sum(data_length)/1024/1024 as data_length,sum(index_length)
    ↪ /1024/1024 as index_length,sum(data_length+index_length)/1024/1024 as sum from
    ↪ information_schema.tables where table_schema = "${schema_name}" group by table_name,
    ↪ table_schema order by sum  desc limit 5;
```

12.7.2.2.3 TiDB Lightning 运行时资源要求

操作系统：本文档示例使用的是若干新的、纯净版 CentOS 7 实例，你可以在本地虚拟化一台主机，或在供应商提供的平台上部署一台小型的云虚拟主机。TiDB Lightning 运行过程中，默认会占满 CPU，建议单独部署在一台主机上。如果条件不允许，你可以将 TiDB Lightning 和其他组件（比如`tikv-server`）部署在同一台机器上，然后设置`region-concurrency` 配置项的值为逻辑 CPU 数的 75%，以限制 TiDB Lightning 对 CPU 资源的使用。

内存和 CPU：因为 TiDB Lightning 对计算机资源消耗较高，建议分配 64 GB 以上的内存以及 32 核以上的 CPU，而且确保 CPU 核数和内存（GB）比为 1:2 以上，以获取最佳性能。

存储空间：配置项 `sorted-kv-dir` 设置排序的键值对的临时存放地址，目标路径需要是一个空目录，至少需要数据源最大单表的空间。建议与 `data-source-dir` 使用不同的存储设备，独占 IO 会获得更好的导入性能，且建议优先考虑配置闪存等高性能存储介质。

网络：建议使用带宽 >=10Gbps 的网卡。

12.7.3 主要功能

12.7.3.1 TiDB Lightning 断点续传

大量的数据导入一般耗时数小时至数天，长时间运行的进程会有一定机率发生非正常中断。如果每次重启都从头开始，就会浪费掉之前已成功导入的数据。为此，TiDB Lightning 提供了“断点续传”的功能，即使 `tidb-lightning` 崩溃，在重启时仍然接着之前的进度继续工作。

本文主要介绍 TiDB Lightning 断点续传的启用与配置、断点的存储，以及断点续传的控制。

12.7.3.1.1 断点续传的启用与配置

```
[checkpoint]
#### 启用断点续传。
#### 导入时，TiDB Lightning 会记录当前进度。
#### 若 TiDB Lightning 或其他组件异常退出，在重启时可以避免重复再导入已完成的数据。
enable = true

#### 存储断点的方式
####  - file：存放在本地文件系统（要求 v2.1.1 或以上）
####  - mysql：存放在兼容 MySQL 的数据库服务器
driver = "file"

#### 存储断点的架构名称（数据库名称）
#### 仅在 driver = "mysql" 时生效
#### schema = "tidb_lightning_checkpoint"

#### 断点的存放位置
#### # 若 driver = "file"，此参数为断点信息存放的文件路径。
#### 如果不设置该参数则默认为 `/tmp/CHECKPOINT_SCHEMA.pb`
#### # 若 driver = "mysql"，此参数为数据库连接参数（DSN），格式为“用户:密码@tcp(地址:端口)/”。
#### 默认会重用 [tidb] 设置目标数据库来存储断点。
#### 为避免加重目标集群的压力，建议另外使用一个兼容 MySQL 的数据库服务器。
#### dsn = "/tmp/tidb_lightning_checkpoint.pb"

#### 导入成功后是否保留断点。默认为删除。
#### 保留断点可用于调试，但有可能泄漏数据源的元数据。
#### keep-after-success = false
```

12.7.3.1.2 断点的存储

TiDB Lightning 支持两种存储方式：本地文件或 MySQL 数据库。

- 若 `driver = "file"`，断点会存放在一个本地文件，其路径由 dsn 参数指定。由于断点会频繁更新，建议将这个文件放到写入次数不受限制的盘上，例如 RAM disk。
- 若 `driver = "mysql"`，断点可以存放在任何兼容 MySQL 5.7 或以上的数据库中，包括 MariaDB 和 TiDB。在没有选择的情况下，默认会存在目标数据库里。

目标数据库在导入期间会有大量的操作，若使用目标数据库来存储断点会加重其负担，甚至有可能造成通信超时丢失数据。因此，强烈建议另外部署一台兼容 MySQL 的临时数据库服务器。此数据库也可以安装在 `tidb-lightning` 的主机上。导入完毕后可以删除。

12.7.3.1.3 断点续传的控制

若 tidb-lightning 因不可恢复的错误而退出（例如数据出错），重启时不会使用断点，而是直接报错离开。为保证已导入的数据安全，这些错误必须先解决掉才能继续。使用 tidb-lightning-ctl 工具可以标示已经恢复。

--checkpoint-error-destroy

```
tidb-lightning-ctl --checkpoint-error-destroy='`schema`.`table`'
```

该命令会让失败的表从头开始整个导入过程。选项中的架构和表名必须以反引号 (`) 包裹，而且区分大小写。

- 如果导入 `schema`.`table` 这个表曾经出错，这条命令会：
    1. 从目标数据库移除 (DROP) 这个表，清除已导入的数据。
    2. 将断点重设到 “未开始” 的状态。
- 如果 `schema`.`table` 没有出错，则无操作。

传入 “all” 会对所有表进行上述操作。这是最方便、安全但保守的断点错误解决方法：

```
tidb-lightning-ctl --checkpoint-error-destroy=all
```

--checkpoint-error-ignore

```
tidb-lightning-ctl --checkpoint-error-ignore='`schema`.`table`' &&
tidb-lightning-ctl --checkpoint-error-ignore=all
```

如果导入 `schema`.`table` 这个表曾经出错，这条命令会清除出错状态，如同没事发生过一样。传入 “all” 会对所有表进行上述操作。

> **注意：**
>
> 除非确定错误可以忽略，否则不要使用这个选项。如果错误是真实的话，可能会导致数据不完全。启用校验和 (CHECKSUM) 可以防止数据出错被忽略。

--checkpoint-remove

```
tidb-lightning-ctl --checkpoint-remove='`schema`.`table`' &&
tidb-lightning-ctl --checkpoint-remove=all
```

无论是否有出错，把表的断点清除。

--checkpoint-dump

```
tidb-lightning-ctl --checkpoint-dump=output/directory
```

将所有断点备份到传入的文件夹，主要用于技术支持。此选项仅于 driver = "mysql" 时有效。

12.7.3.2 表库过滤

TiDB 生态工具默认情况下作用于所有数据库，但实际使用中，往往只需要作用于其中的部分子集。例如，用户只想处理 `foo*` 和 `bar*` 形式的表，而无需对其他表进行操作。

从 TiDB 4.0 起，所有 TiDB 生态系统工具都使用一个通用的过滤语法来定义子集。本文档介绍如何使用表库过滤功能。

12.7.3.2.1 使用表库过滤

命令行

在命令行中使用多个 `-f` 或 `--filter` 参数，即可在 TiDB 生态工具中应用表库过滤规则。每个过滤规则均采用 `db.table` 形式，支持通配符（详情见下一节）。以下为各个工具中的使用示例：

- BR：

```
./br backup full -f 'foo*.*' -f 'bar*.*' -s 'local:///tmp/backup'
#                ^~~~~~~~~~~~~~~~~~~~~~
./br restore full -f 'foo*.*' -f 'bar*.*' -s 'local:///tmp/backup'
#                 ^~~~~~~~~~~~~~~~~~~~~~
```

- Dumpling：

```
./dumpling -f 'foo*.*' -f 'bar*.*' -P 3306 -o /tmp/data/
#          ^~~~~~~~~~~~~~~~~~~~~~
```

- TiDB Lightning：

```
./tidb-lightning -f 'foo*.*' -f 'bar*.*' -d /tmp/data/ --backend tidb
#                ^~~~~~~~~~~~~~~~~~~~~~
```

TOML 配置文件

在 TOML 文件中，表库过滤规则以字符串数组的形式指定。以下为各个工具中的使用示例：

- TiDB Lightning：

```
[mydumper]
filter = ['foo*.*', 'bar*.*']
```

- TiCDC：

```
[filter]
rules = ['foo*.*', 'bar*.*']

[[sink.dispatchers]]
matcher = ['db1.*', 'db2.*', 'db3.*']
dispatcher = 'ts'
```

12.7.3.2.2 表库过滤语法

直接使用表名

每条表库过滤规则由“库”和“表”组成，两部分之间以英文句号 (.) 分隔。只有表名与规则完全相符的表才会被接受。

```
db1.tbl1
db2.tbl2
db3.tbl3
```

表名只由有效的标识符组成，例如：

- 数字（0 到 9）
- 字母（a 到 z，A 到 Z）
- $
- _
- 非 ASCII 字符（U+0080 到 U+10FFFF）

其他 ASCII 字符均为保留字。部分标点符号有特殊含义，详情见下一节。

使用通配符

表名的两个部分均支持使用通配符（详情见 fnmatch(3)）。

- *：匹配零个或多个字符。
- ?：匹配一个字符。
- [a-z]：匹配“a”和“z”之间的一个字符。
- [!a-z]：匹配不在“a”和“z”之间的一个字符。

```
db[0-9].tbl[0-9a-f][0-9a-f]
data.*
*.backup_*
```

此处，“字符”指的是一个 Unicode 码位，例如：

- U+00E9 “é” 是 1 个字符。
- U+0065，U+0301 “é” 是 2 个字符。
- U+1F926 U+1F3FF U+200D U+2640 U+FE0F “🤦🏿‍♀” 是 5 个字符。

使用文件导入

如需导入一个文件作为过滤规则，请在规则的开头加上一个“@”来指定文件名。库表过滤解析器将导入文件中的每一行都解析为一条额外的过滤规则。

例如，config/filter.txt 文件有以下内容：

```
employees.*
*.WorkOrder
```

以下两条表库过滤命令是等价的：

```
./dumpling -f '@config/filter.txt'
./dumpling -f 'employees.*' -f '*.WorkOrder'
```

导入的文件里不能使用过滤规则导入另一个文件。

注释与空行

导入的过滤规则文件中，每一行开头和结尾的空格都会被去除。此外，空行（空字符串）也将被忽略。

行首的 # 表示该行是注释，会被忽略。而不在行首的 # 则会被认为是语法错误。

```
#### 这是一行注释
db.table   # 这一部分不是注释，且可能引起错误
```

排除规则

在一条过滤规则的开头加上 !，则表示符合这条规则的表不会被 TiDB 生态工具处理。通过应用排除规则，库表过滤可以作为屏蔽名单来使用。

```
*.*
#^ 注意：必须先添加 *.* 规则来包括所有表
!*.Password
!employees.salaries
```

转义字符

如果需要将特殊字符转化为标识符，可以在特殊字符前加上反斜杠 \。

```
db\.with\.dots.*
```

为了简化语法并向上兼容，不支持下列字符序列：

- 在行尾去除空格后使用 \（使用 [ ] 来匹配行尾的空格）。
- 在 \ 后使用数字或字母（[0-9a-zA-Z]）。特别是类似 C 的转义序列，如 \0、\r、\n、\t 等序列，目前在表库过滤规则中无意义。

引号包裹的标识符

除了 \ 之外，还可以用 " 和 ` 来控制特殊字符。

```
"db.with.dots"."tbl\1"
`db.with.dots`.`tbl\2`
```

也可以通过输入两次引号，将引号包含在标识符内。

```
"foo""bar".`foo``bar`
#### 等价于：
foo\"bar.foo\`bar
```

用引号包裹的标识符不可以跨越多行。

用引号只包裹标识符的一部分是无效的，例如：

```
"this is "invalid*.*
```

正则表达式

如果你需要使用较复杂的过滤规则，可以将每个匹配模型写为正则表达式，以 `/` 为分隔符：

```
/^db\d{2,}$/./^tbl\d{2,}$/
```

这类正则表示使用 Go dialect。只要标识符中有一个子字符串与正则表达式匹配，则视为匹配该模型。例如，`/b/` 匹配 `db01`。

> 注意：
>
> 正则表达式中的每一个 `/` 都需要转义为 `\/`，包括在 `[...]` 里面的 `/`。不允许在 `\Q...\E` 之间放置一个未转义的 `/`。

12.7.3.2.3 使用多个过滤规则

当表的名称与过滤列表中所有规则均不匹配时，默认情况下这些表被忽略。

要建立一个屏蔽名单，必须使用显式的 `*.*` 作为第一条过滤规则，否则所有表均被排除。

```
#### 所有表均被过滤掉
./dumpling -f '!*.Password'

#### 只有 “Password” 表被过滤掉，其余表仍保留
./dumpling -f '*.*' -f '!*.Password'
```

如果一个表的名称与过滤列表中的多个规则匹配，则以最后匹配的规则为准。例如：

```
#### rule 1
employees.*
#### rule 2
!*.dep*
#### rule 3
*.departments
```

过滤结果如下：

表名	规则 1	规则 2	规则 3	结果
irrelevant.table				默认（拒绝）
employees.employees	√			规则 1（接受）
employees.dept_emp	√	√		规则 2（拒绝）

表名	规则 1	规则 2	规则 3	结果
employees.departments	√	√	√	规则 3（接受）
else.departments		√	√	规则 3（接受）

注意：

在 TiDB 生态工具的默认配置中，系统库总是被排除。系统库有以下六个：

- INFORMATION_SCHEMA
- PERFORMANCE_SCHEMA
- METRICS_SCHEMA
- INSPECTION_SCHEMA
- mysql
- sys

12.7.3.3 CSV 支持与限制

本文介绍如何使用 TiDB Lightning 从 CSV 文件迁移数据到 TiDB。关于如何从 MySQL 生成 CSV 文件，可以参考使用 Dumpling 导出到 CSV 文件。

TiDB Lightning 支持读取 CSV（逗号分隔值）的数据源，以及其他定界符格式如 TSV（制表符分隔值）。

12.7.3.3.1 文件名

包含整张表的 CSV 文件需命名为 `db_name.table_name.csv`，该文件会被解析为数据库 `db_name` 里名为 `table_name` 的表。

如果一个表分布于多个 CSV 文件，这些 CSV 文件命名需加上文件编号的后缀，如 `db_name.table_name.003.csv`。数字部分不需要连续但必须递增，并用零填充。

文件扩展名必须为 `*.csv`，即使文件的内容并非逗号分隔。

12.7.3.3.2 表结构

CSV 文件是没有表结构的。要导入 TiDB，就必须为其提供表结构。可以通过以下任一方法实现：

- 创建包含 DDL 语句 `CREATE TABLE` 的文件 `db_name.table_name-schema.sql` 以及包含 `CREATE DATABASE` DDL 语句的文件 `db_name-schema-create.sql`。
- 首先在 TiDB 中直接创建空表，然后在 `tidb-lightning.toml` 中设置 `[mydumper] no-schema = true`。

12.7.3.3.3 配置

CSV 格式可在 `tidb-lightning.toml` 文件中 `[mydumper.csv]` 下配置。大部分设置项在 MySQL `LOAD DATA` 语句中都有对应的项目。

```
[mydumper.csv]
#### 字段分隔符，支持一个或多个字符，默认值为 ','。如果数据中可能有逗号，
    ↪ 建议源文件导出时分隔符使用非常见组合字符例如'|+|'。
separator = ','
#### 引用定界符，设置为空表示字符串未加引号。
delimiter = '"'
#### 行尾定界字符，支持一个或多个字符。设置为空（默认值）表示 "\n"（换行）和 "\r\n" （回车+换行
    ↪ ），均表示行尾。
terminator = ""
#### CSV 文件是否包含表头。
#### 如果为 true，首行将会被跳过。
header = true
#### CSV 是否包含 NULL。
#### 如果为 true，CSV 文件的任何列都不能解析为 NULL。
not-null = false
#### 如果 `not-null` 为 false（即 CSV 可以包含 NULL），
#### 为以下值的字段将会被解析为 NULL。
null = '\N'
#### 是否解析字段内的反斜线转义符。
backslash-escape = true
#### 是否移除以分隔符结束的行。
trim-last-separator = false
```

对于诸如 `separator`，`delimiter` 和 `terminator` 等取值为字符串的配置项，如果需要设置的字符串中包含特殊字符，可以通过使用反斜杠 `\` 转义的方式进行输入，输入的转义序列必须被包含在一对双引号 `"` 之间。例如，设置 `separator = "\u001f"` 表示使用 ASCII 字符 0X1F 作为字符串定界符。另外，也可以使用单引号字符串 `'...'` 禁止对字符进行转义。另外，设置 `separator = '\n'` 表示使用两个字符 `\` + `n` 作为字符串定界符，而不是转义后的换行符 `\n`。

更多详细的内容请参考 TOML v1.0.0 标准。

`separator`

- 指定字段分隔符。
- 可以为一个或多个字符，不能为空。
- 常用值：
  - CSV 用 `','`
  - TSV 用 `"\t"`
  - “ŏ001”表示使用 ASCII 字符 0x01
- 对应 LOAD DATA 语句中的 `FIELDS TERMINATED BY` 项。

`delimiter`

- 指定引用定界符。
- 如果 `delimiter` 为空，所有字段都会被取消引用。
- 常用值：
  - `'"'` 使用双引号引用字段，和 RFC 4180 一致。
  - `''` 不引用
- 对应 LOAD DATA 语句中的 `FIELDS ENCLOSED BY` 项。

`terminator`

- 指定行尾定界符。
- 如果 `terminator` 为空，表示“\n”（换行）和“\r\n”（回车 + 换行），均表示行尾。
- 对应 LOAD DATA 语句中的 `LINES TERMINATED BY` 项。

`header`

- 是否所有 CSV 文件都包含表头行。
- 如为 true，第一行会被用作列名。如为 false，第一行并无特殊性，按普通的数据行处理。

`not-null` 和 `null`

- `not-null` 决定是否所有字段不能为空。
- 如果 `not-null` 为 false，设定了 `null` 的字符串会被转换为 SQL NULL 而非具体数值。
- 引用不影响字段是否为空。

  例如有如下 CSV 文件：

```
A,B,C
\N,"\N",
```

  在默认设置（`not-null = false; null = '\N'`）下，列 A and B 导入 TiDB 后都将会转换为 NULL。列 C 是空字符串 `''`，但并不会解析为 NULL。

`backslash-escape`

- 是否解析字段内的反斜线转义符。
- 如果 `backslash-escape` 为 true，下列转义符会被识别并转换。

转义符	转换为
\0	空字符 (U+0000)
\b	退格 (U+0008)
\n	换行 (U+000A)
\r	回车 (U+000D)
\t	制表符 (U+0009)
\Z	Windows EOF (U+001A)

其他情况下（如 \"）反斜线会被移除，仅在字段中保留其后面的字符（"），这种情况下，保留的字符仅作为普通字符，特殊功能（如界定符）都会失效。

- 引用不会影响反斜线转义符的解析与否。
- 对应 LOAD DATA 语句中的 FIELDS ESCAPED BY '\' 项。

trim-last-separator

- 将 separator 字段当作终止符，并移除尾部所有分隔符。

  例如有如下 CSV 文件：

```
A,,B,,
```

- 当 trim-last-separator = false，该文件会被解析为包含 5 个字段的行 ('A', '', 'B', '', '')。
- 当 trim-last-separator = true，该文件会被解析为包含 3 个字段的行 ('A', '', 'B')。
- 此配置项已被弃用，建议使用兼容性更好的 terminator。

  如果有如下旧的配置：

```
separator = ','
trim-last-separator = true
```

  建议修改为：

```
separator = ','
terminator = ",\n" # 请根据文件实际使用的换行符指定为 ",\n" 或 ",\r\n"
```

不可配置项

TiDB Lightning 并不完全支持 LOAD DATA 语句中的所有配置项。例如：

- 不可使用行前缀（LINES STARTING BY）。
- 不可跳过表头（IGNORE n LINES）。如有表头，必须是有效的列名。

12.7.3.3.4 设置 strict-format 启用严格格式

导入文件的大小统一约为 256 MB 时，TiDB Lightning 可达到最佳工作状态。如果导入单个 CSV 大文件，TiDB Lightning 只能使用一个线程来处理，这会降低导入速度。

要解决此问题，可先将 CSV 文件分割为多个文件。对于通用格式的 CSV 文件，在没有读取整个文件的情况下无法快速确定行的开始和结束位置。因此，默认情况下 TiDB Lightning 不会自动分割 CSV 文件。但如果你确定待导入的 CSV 文件符合特定的限制要求，则可以启用 strict-format 设置。启用后，TiDB Lightning 会将单个 CSV 大文件分割为单个大小为 256 MB 的多个文件块进行并行处理。

```
[mydumper]
strict-format = true
```

严格格式的 CSV 文件中，每个字段仅占一行，即必须满足以下条件之一：

- 分隔符为空；
- 每个字段不包含 `terminator` 对应的字符串。在默认配置下，对应每个字段不包含 CR (`\r`) 或 LF (`\n`)。

如果 CSV 文件不是严格格式但 `strict-format` 被误设为 `true`，跨多行的单个完整字段会被分割成两部分，导致解析失败，甚至不报错地导入已损坏的数据。

12.7.3.3.5 通用配置

CSV

默认设置已按照 RFC 4180 调整。

```
[mydumper.csv]
separator = ',' # 如果数据中可能有逗号，建议源文件导出时分隔符使用非常见组合字符例如'|+|'
delimiter = '"'
header = true
not-null = false
null = '\N'
backslash-escape = true
```

示例内容：

```
ID,Region,Count
1,"East",32
2,"South",\N
3,"West",10
4,"North",39
```

TSV

```
[mydumper.csv]
separator = "\t"
delimiter = ''
header = true
not-null = false
null = 'NULL'
backslash-escape = false
```

示例内容：

```
ID    Region    Count
1     East      32
2     South     NULL
3     West      10
4     North     39
```

TPC-H DBGEN

```
[mydumper.csv]
separator = '|'
delimiter = ''
terminator = "|\n"
header = false
not-null = true
backslash-escape = false
```

示例内容：

```
1|East|32|
2|South|0|
3|West|10|
4|North|39|
```

12.7.3.4 TiDB Lightning 后端

TiDB Lightning 的后端决定 `tidb-lightning` 组件将如何把将数据导入到目标集群中。目前，TiDB Lightning 支持以下后端：

- Local-backend
- Importer-backend
- TiDB-backend

以上几种后端导入数据的区别如下：

- Local-backend：`tidb-lightning` 先将数据编码成键值对并排序存储在本地临时目录，然后将这些键值对以 SST 文件的形式上传到各个 TiKV 节点，然后由 TiKV 将这些 SST 文件 Ingest 到集群中。和 `Importer-backend` 原理相同，不过不依赖额外的 `tikv-importer` 组件。
- Importer-backend：`tidb-lightning` 先将 SQL 或 CSV 数据编码成键值对，由 `tikv-importer` 对写入的键值对进行排序，然后把这些键值对 Ingest 到 TiKV 节点中。
- TiDB-backend：`tidb-lightning` 先将数据编码成 `INSERT` 语句，然后直接在 TiDB 节点上运行这些 SQL 语句进行数据导入。

后端	Local-backend	Importer-backend	TiDB-backend
速度	快 (~500 GB/小时)	快 (~400 GB/小时)	慢 (~50 GB/小时)
资源使用率	高	高	低
占用网络带宽	高	中	低
导入时是否满足 ACID	否	否	是
目标表	必须为空	必须为空	可以不为空
额外组件	无	`tikv-importer`	无
支持 TiDB 集群版本	>= v4.0.0	全部	全部
是否影响 TiDB 对外提供服务	是	是	否

注意：

- 使用多个 TiDB Lightning 向同一目标导入时，禁止混用不同的 backend，例如，不可同时使用 Local-backend 和 TiDB-backend 导入同一 TiDB 集群。
- 默认情况下，不应同时启动多个 TiDB Lightning 实例导入同一 TiDB 集群，而应考虑使用并行导入特性。

12.7.3.4.1 如何选择后端模式

- 如果导入的目标集群为 v4.0 或以上版本，请优先考虑使用 Local-backend 模式。Local-backend 部署更简单并且性能也较其他两个模式更高
- 如果目标集群为 v3.x 或以下，建议使用 Importer-backend 模式
- 如果需要导入的集群为生产环境线上集群，或需要导入的表中已包含有数据，则最好使用 TiDB-backend 模式

12.7.3.4.2 TiDB Lightning Local-backend

自 TiDB 4.0.3 版本起，TiDB Lightning 引入了 Local-backend 特性。该特性支持导入数据到 v4.0.0 以上的 TiDB 集群。

部署和配置 TiDB Lightning

TiDB Lightning Local-backend 模式的部署方法见TiDB Lightning 部署与执行。

12.7.3.4.3 TiDB Lightning TiDB-backend

注意：

从 TiDB v4.0 起，PingCAP 不再维护 Loader 工具。从 v5.0 起，不再提供 Loader 的文档。Loader 的功能已经完全被 TiDB Lightning 的 TiDB backend 功能取代，强烈建议切换到 TiDB Lightning。

部署和配置 TiDB Lightning

使用 TiDB-backend 时，你无需部署 `tikv-importer`。与标准部署过程相比，部署 TiDB-backend 时有如下不同：

- 可以跳过所有涉及 `tikv-importer` 的步骤。
- 必须更改相应配置申明使用的是 TiDB-backend。

硬件需求

使用 TiDB-backend 时，TiDB Lightning 的速度仅受限于 TiDB 执行 SQL 语句的速度。因此，即使是低配的机器也足够发挥出最佳性能。推荐的硬件配置如下：

- 16 逻辑核 CPU
- 足够储存整个数据源的 SSD 硬盘，读取速度越快越好
- 千兆网卡

手动部署

手动部署时，你无需下载和配置 tikv-importer，TiDB Lightning 可在此下载。

在运行 tidb-lightning 之前，在配置文件中加上如下几行：

```
[tikv-importer]
backend = "tidb"
```

或者在用命令行启动 tidb-lightning 时，传入参数 --backend tidb。

配置说明与示例

```
#### tidb-lightning 任务配置

[lightning]
#### 启动之前检查集群是否满足最低需求
check-requirements = true

#### 每张表被切分成一个用于存储索引的“索引引擎”和若干存储行数据的“数据引擎”
#### 这两项设置控制两种引擎文件的最大并发数
#### 控制同时允许导入的最大表数量，对于 TiDB-backend，默认值为 CPU 数
index-concurrency = 40

#### 控制同时允许导入的最大“数据引擎”数，默认值为 CPU 数，本配置应大于或等于 index-concurrency
table-concurrency = 40

#### 执行 SQL 语句的并发数。默认与逻辑 CPU 的数量相同
#### TiDB-backend 的瓶颈通常不在 CPU，可以根据下游集群的实际负载调大此配置以优化写入速度
#### 在调整此配置时，建议将 index-concurrency 和 table-concurrency 也调整成相同的值
region-concurrency = 40

#### 日志相关的配置
#### 输出日志级别
level = "info"

#### 日志输出的文件。如果为空（默认），则会输出至 /tmp/lightning.log.{timestamp}；
    ↪ 如果希望输出至系统标准输出，请设置为 "-"
#### file = "tidb-lightning.log"

[checkpoint]
#### 启用断点续传
#### 导入数据时，TiDB Lightning 会记录当前表导入的进度
#### 若 TiDB Lightning 或其他组件异常退出，在重启时也可以避免重复再导入已完成的数据
```

```
enable = true

#### 存储断点的方式
####  - file（默认）：存放在本地文件系统（要求 v2.1.1 或以上）
####  - mysql：存放在兼容 MySQL 的数据库服务器
driver = "file"

#### 存储断点的数据库名称
#### 仅在 driver = "mysql" 时生效
#### schema = "tidb_lightning_checkpoint"

#### 断点的存放位置
#### # 若 driver = "file"，此参数为断点信息存放的文件路径
#### 如果不设置该参数则默认为 `/tmp/CHECKPOINT_SCHEMA.pb`
#### # 若 driver = "mysql"，此参数为数据库连接参数（DSN），格式为“用户:密码@tcp(地址:端口)/”
#### 默认会重用 [tidb] 设置目标数据库来存储断点
#### 为避免加重目标集群的压力，建议另外使用一个兼容 MySQL 的数据库服务器
#### dsn = "/tmp/tidb_lightning_checkpoint.pb"

#### 导入成功后是否保留断点。默认为删除
#### 保留断点可用于调试，但有可能泄漏数据源的元数据
#### keep-after-success = false

[tikv-importer]
#### 后端模式，对于 TiDB-backend 请设置为 “tidb”
backend = "tidb"

#### 对于插入重复数据时执行的操作：
#### - replace：新数据替代已有数据
#### - ignore：保留已有数据，忽略新数据
#### - error：中止导入并报错
#### on-duplicate = "replace"

[mydumper]
#### 设置文件读取的区块大小(默认为 64 KiB)，确保该值比数据源的最长字符串长。
#### read-block-size = "64KiB"

#### （源数据文件）单个导入区块大小的最小值（默认为 100 GiB）。
#### TiDB Lightning 根据该值将一张大表分割为多个数据引擎文件。
#### batch-size = "100GiB"

#### 本地源数据目录或外部存储 URL
data-source-dir = "/data/my_database"

#### “严格”格式的导入数据可加快处理速度。
```

```
#### strict-format = true 要求：
#### 在 CSV 文件的所有记录中，每条数据记录的值不可包含字符换行符（U+000A 和 U+000D，即 \r 和 \n）
#### 甚至被引号包裹的字符换行符都不可包含，即换行符只可用来分隔行。
#### 导入数据源为严格格式时，TiDB Lightning 会快速定位大文件的分割位置进行并行处理。
#### 但是如果输入数据为非严格格式，可能会将一条完整的数据分割成两部分，导致结果出错。
#### 为保证数据安全而非追求处理速度，默认值为 false。
#### strict-format = false

#### 如果 strict-format = true，TiDB Lightning 会将 CSV 大文件分割为多个文件块进行并行处理。max-
    ↪ region-size 是分割后每个文件块的最大大小。
#### max-region-size = "256MiB" # Byte（默认是 256 MiB）

#### 只导入与该通配符规则相匹配的表，默认会过滤掉系统表。详情见表库过滤章节。
#### filter = ['*.*', '!mysql.*', '!sys.*', '!INFORMATION_SCHEMA.*', '!PERFORMANCE_SCHEMA.*', '!
    ↪ METRICS_SCHEMA.*', '!INSPECTION_SCHEMA.*']

#### 配置 CSV 文件的解析方式（如果源文件中不包含 CSV 文件可不设置此项）。
[mydumper.csv]
#### 字段分隔符，应为单个 ASCII 字符。
separator = ','

#### 引用定界符，可为单个 ASCII 字符或空字符串。
delimiter = '"'

#### CSV 文件是否包含表头。
#### 如果 header = true，将跳过首行。
header = true

#### CSV 文件是否包含 NULL。
#### 如果 not-null = true，CSV 所有列都不能解析为 NULL。
not-null = false

#### 如果 not-null = false（即 CSV 可以包含 NULL），
#### 为以下值的字段将会被解析为 NULL。
null = '\N'

#### 是否对字段内“\“进行转义
backslash-escape = true

#### 如果有行以分隔符结尾，删除尾部分隔符。
trim-last-separator = false

[tidb]
#### 目标集群的信息。tidb-server 的地址，填一个即可。
host = "172.16.31.1"
```

```
port = 4000
user = "root"
password = ""

#### 解析和执行 SQL 语句的默认 SQL 模式。
#### sql-mode = "ONLY_FULL_GROUP_BY,NO_ENGINE_SUBSTITUTION"

#### SQL 连接是否使用 TLS。可选值为：
####  * ""            - 如果填充了 [tidb.security] 部分，则强制使用 TLS（与 "cluster" 情况相同
    ↪ ），否则与 "false" 情况相同
####  * "false"       - 禁用 TLS
####  * "cluster"     - 强制使用 TLS，并使用 [tidb.security] 部分中指定的 CA 验证服务器的证书
####  * "skip-verify" - 强制使用 TLS，但不验证服务器的证书（不安全！）
####  * "preferred"   - 与 "skip-verify" 相同，但是如果服务器不支持 TLS，则会退回到未加密的连接
tls = ""

#### 指定证书和密钥用于 TLS 连接 MySQL。
#### [tidb.security]

#### CA 的公钥证书。设置为空字符串可禁用 SQL 的 TLS。
#### ca-path = "/path/to/ca.pem"

#### 该服务的公钥证书。默认为 `security.cert-path` 的副本
#### cert-path = "/path/to/lightning.pem"

#### 此服务的私钥。默认为 `security.key-path` 的副本
#### key-path = "/path/to/lightning.key"

#### 设置周期性后台操作。
#### 支持的单位：h（时）、m（分）、s（秒）。
[cron]

#### 在日志中打印导入进度的持续时间。
log-progress = "5m"
```

解决冲突

TiDB-backend 支持导入到已填充的表（非空表）。但是，新数据可能会与旧数据的唯一键冲突。你可以通过使用如下任务配置来控制遇到冲突时的默认行为：

```
[tikv-importer]
backend = "tidb"
on-duplicate = "replace" # 或者 “error”、“ignore”
```

设置	冲突时默认行为	对应 SQL 语句
replace	新数据替代旧数据	`REPLACE INTO ...`
ignore	保留旧数据，忽略新数据	`INSERT IGNORE INTO ...`
error	中止导入	`INSERT INTO ...`

从 Loader 迁移到 TiDB Lightning TiDB-backend

当需要将数据导入到 TiDB 集群时，TiDB Lightning TiDB-backend 可以完全取代 Loader。下表说明了如何将 Loader 的配置迁移到TiDB Lightning 配置中：

Loader

TiDB Lightning

```
#### 日志级别
log-level = "info"
#### 日志的输出目录
log-file = "loader.log"
#### Prometheus
status-addr = ":8272"
#### 线程数
pool-size = 16
```

```
[lightning]
#### 日志级别
level = "info"
#### 日志的输出目录。如果未指定该位置目录，默认为执行命令的所在目录。
file = "tidb-lightning.log"
#### Prometheus
pprof-port = 8289
#### 并发度（最好使用默认设置）
#region-concurrency = 16
```

```
#### 断点数据库名
checkpoint-schema = "tidb_loader"
```

```
[checkpoint]
#### 断点存储
enable = true
schema = "tidb_lightning_checkpoint"
#### 断点默认存储在本地的文件系统，这样更高效。但你也可以
#### 选择将断点存储在目标数据库中，设置如下：
#### driver = "mysql"
```

```
[tikv-importer]
#### 使用 TiDB-backend
backend = "tidb"
```

```
#### 数据源目录
dir = "/data/export/"
```

```
[mydumper]
#### 数据源目录
data-source-dir = "/data/export"
```

```
[db]
#### TiDB 连接参数
host = "127.0.0.1"
port = 4000
user = "root"
password = ""
#sql-mode = ""
```

```
[tidb]
#### TiDB 连接参数
host = "127.0.0.1"
port = 4000
#### 在 TiDB-backend 模式下，该参数为可选参数
#### status-port = 10080
user = "root"
password = ""
#sql-mode = ""
```

```
#### [[route-rules]]
#### Table routes
#### schema-pattern = "shard_db_*"
#### table-pattern = "shard_table_*"
#### target-schema = "shard_db"
#### target-table = "shard_table"
```

```
#### [[routes]]
#### Table routes
#### schema-pattern = "shard_db_*"
#### table-pattern = "shard_table_*"
#### target-schema = "shard_db"
#### target-table = "shard_table"
```

12.7.3.4.4 TiDB Lightning Importer-backend

部署 Importer-backend

本节介绍 TiDB Lightning 使用 Importer 模式的部署方式：手动部署。

硬件需求

tidb-lightning 和 tikv-importer 这两个组件皆为资源密集程序，建议各自单独部署。

为了优化效能，建议硬件配置如下：

- tidb-lightning
  - 32+ 逻辑核 CPU
  - 足够储存整个数据源的 SSD 硬盘，读取速度越快越好
  - 使用万兆网卡，带宽需 300 MB/s 以上
  - 运行过程默认会占满 CPU 资源，因此建议将 tidb-lightning 部署到一台单独的机器上。条件不允许的情况下可以和其他组件（比如 tidb-server）部署在同一台机器上，然后通过配置 region-concurrency 限制 tidb-lightning 使用 CPU 资源。
- tikv-importer
  - 32+ 逻辑核 CPU
  - 40 GB+ 内存
  - 1 TB+ SSD 硬盘，IOPS 越高越好（要求 ≥8000）
    - 硬盘必须大于最大的 N 个表的大小总和，其中 N = max(index-concurrency, table-concurrency)。
  - 使用万兆网卡，带宽需 300 MB/s 以上
  - 运行过程中 CPU、I/O 和网络带宽资源都可能占满，建议单独部署。

如果机器充裕的话，可以部署多套 tidb-lightning + tikv-importer，然后将源数据以表为粒度进行切分，并发导入。

手动部署 TiDB Lightning

第 1 步：部署 TiDB 集群

在开始数据导入之前，需先部署一套要进行导入的 TiDB 集群 (版本要求 2.0.9 以上)，建议使用最新版。部署方法可参考使用 TiUP 部署 TiDB 集群。

第 2 步：下载 TiDB Lightning 安装包

在工具下载页面下载 TiDB Lightning 安装包（需选择与 TiDB 集群相同的版本）。

第 3 步：启动 tikv-importer

1. 从安装包上传 bin/tikv-importer。
2. 配置 tikv-importer.toml。

```
# TiKV Importer 配置文件模版

# 日志文件。
log-file = "tikv-importer.log"
# 日志等级：trace、debug、info、warn、error、off。
log-level = "info"

# 状态服务器的监听地址。
status-server-address = "0.0.0.0:8286"

[server]
# tikv-importer 监听的地址，tidb-lightning 需要连到这个地址进行数据写入。
addr = "0.0.0.0:8287"

[metric]
# 给 Prometheus 客户端的推送任务名称。
job = "tikv-importer"
# 给 Prometheus 客户端的推送间隔。
interval = "15s"
# Prometheus Pushgateway 地址。
address = ""

[import]
# 存储引擎文档（engine file）的文件夹路径。
import-dir = "/mnt/ssd/data.import/"
```

上面仅列出了 tikv-importer 的基本配置。完整配置请参考tikv-importer 配置说明。

3. 运行 tikv-importer。

```
nohup ./tikv-importer -C tikv-importer.toml > nohup.out &
```

第 4 步：启动 tidb-lightning

1. 从安装包上传 bin/tidb-lightning 及 bin/tidb-lightning-ctl。
2. 将数据源写入到同样的机器。
3. 配置 tidb-lightning.toml。对于没有出现在下述模版中的配置，TiDB Lightning 给出配置错误的提醒并退出。

```
[lightning]

# 转换数据的并发数，默认为逻辑 CPU 数量，不需要配置。
# 混合部署的情况下可以配置为逻辑 CPU 的 75% 大小。
# region-concurrency =
```

```
# 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
# tikv-importer 的监听地址，需改成 tikv-importer 服务器的实际地址。
addr = "172.16.31.10:8287"

[mydumper]
# 源数据目录。
data-source-dir = "/data/my_database"

[tidb]
# 目标集群的信息。tidb-server 的监听地址，填一个即可。
host = "172.16.31.1"
port = 4000
user = "root"
password = ""
# 表架构信息在从 TiDB 的“状态端口”获取。
status-port = 10080
```

上面仅列出了 tidb-lightning 的基本配置信息。完整配置信息请参考tidb-lightning 配置说明。

4. 运行 tidb-lightning。如果直接在命令行中用 nohup 启动程序，可能会因为 SIGHUP 信号而退出，建议把 nohup 放到脚本里面，如：

```
#!/bin/bash
nohup ./tidb-lightning -config tidb-lightning.toml > nohup.out &
```

12.7.3.5 TiDB Lightning 并行导入

TiDB Lightning 的Local 后端模式从 v5.3.0 版本开始支持单表或多表数据的并行导入。通过支持同步启动多个实例，并行导入不同的单表或多表的不同数据，使 TiDB Lightning 具备水平扩展的能力，可大大降低导入大量数据所需的时间。

在技术实现上，TiDB Lightning 通过在目标 TiDB 中记录各个实例以及每个导入表导入数据的元信息，协调不同实例的 Row ID 分配范围、全局 Checksum 的记录和 TiKV 及 PD 的配置变更与恢复。

TiDB Lightning 并行导入可以用于以下场景：

- 并行导入分库分表的数据。在该场景中，来自上游多个数据库实例中的多个表，分别由不同的 TiDB Lightning 实例并行导入到下游 TiDB 数据库中。
- 并行导入单表的数据。在该场景中，存放在某个目录中或云存储（如 Amazon S3）中的多个单表文件，分别由不同的 TiDB Lightning 实例并行导入到下游 TiDB 数据库中。该功能为 v5.3.0 版本引入的新功能。

注意：

- 并行导入只支持初始化 TiDB 的空表，不支持导入数据到已有业务写入的数据表，否则可能会导致数据不一致的情况。
- 并行导入一般用于 local-backend 模式。
- 使用多个 TiDB Lightning 向同一目标导入时，禁止混用不同的 backend，例如，不可同时使用 Local-backend 和 TiDB-backend 导入同一 TiDB 集群。

下图展示了并行导入分库分表的工作流程。在该场景中，你可以使用多个 TiDB Lightning 实例导入 MySQL 的分表到下游的 TiDB 集群。

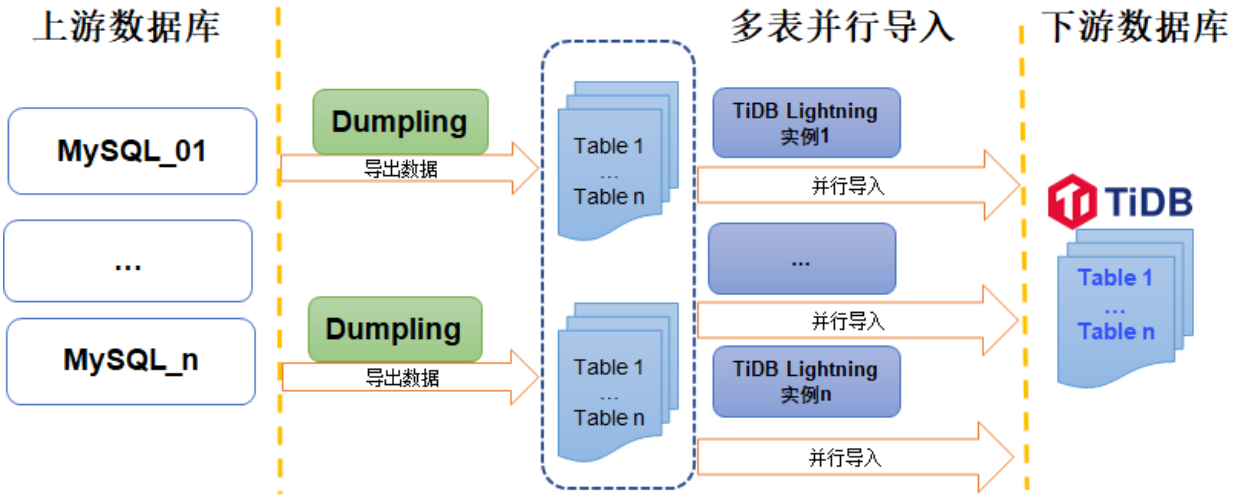


图 261: 并行导入分库分表

下图展示了并行导入单表的工作流程。在该场景中，你可以使用多个 TiDB Lightning 实例，将单个表中的数据拆分后，并行导入到下游的 TiDB 集群。

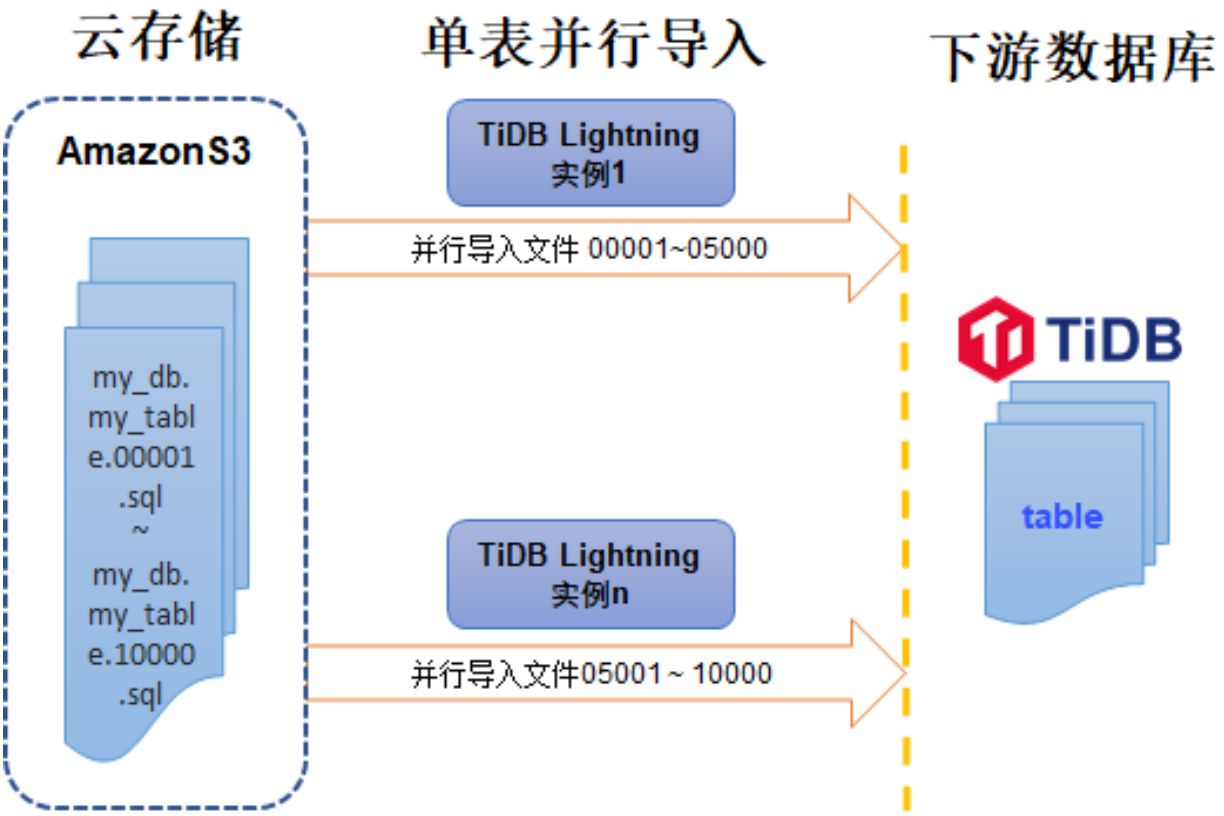


图 262: 并行导入单表

12.7.3.5.1 使用说明

使用 TiDB Lightning 并行导入无须额外配置。TiDB Lightning 在启动时，会在下游 TiDB 中注册元信息，并自动检测是否有其他实例向目标集群导入数据。如果有，则自动进入并行导入模式。

但是在并行导入时，需要注意以下情况：

- 解决主键或者唯一索引的冲突
- 导入性能优化

解决主键或者唯一索引的冲突

在使用Local 后端模式并行导入时，需要确保多个 TiDB Lightning 的数据源之间，以及它们和 TiDB 的目标表中的数据没有主键或者唯一索引的冲突，并且导入的目标表不能有其他应用进行数据写入。否则，TiDB Lightning 将无法保证导入结果的正确性，并且导入完成后相关的数据表将处于数据索引不一致的状态。

导入性能优化

由于 TiDB Lightning 需要将生成的 Key-Value 数据上传到对应 Region 的每一个副本所在的 TiKV 节点，其导入速度受目标集群规模的限制。在通常情况下，建议确保目标 TiDB 集群中的 TiKV 实例数量与 TiDB Lightning 的实例数量大于 n:1 (n 为 Region 的副本数量)。同时，在使用 TiDB Lightning 并行导入模式时，为达到最优性能，建议进行如下限制：

- 每个 TiDB Lightning 部署在单独的机器上面。TiDB Lightning 默认会消耗所有的 CPU 资源，在单台机器上面部署多个实例并不能提升性能。
- 每个 TiDB Lightning 实例导入的源文件总大小不超过 5 TiB
- TiDB Lightning 实例的总数量不超过 10 个

在使用 TiDB Lightning 并行导入分库分表数据的时候，请根据数据量大小选择使用的 TiDB Lightning 实例数量。

- 如果 MySQL 数据量小于 2 TiB，可以使用 1 个 TiDB Lightning 实例进行并行导入
- 如果 MySQL 数据量超过 2 TiB，并且 MySQL 实例总数小于 10 个，建议每个 MySQL 实例对应 1 个 TiDB Lightning 实例，而且并行 TiDB Lightning 实例数量不要超过 10 个
- 如果 MySQL 数据量超过 2 TiB，并且 MySQL 实例总数超过 10 个，建议这些 MySQL 实例导出的数据平均分配 5 到 10 个 TiDB Lightning 实例进行导入

接下来，本文档将以两个并行导入的示例，详细介绍了不同场景下并行导入的操作步骤：

- 示例 1：使用 Dumpling + TiDB Lightning 并行导入分库分表数据至 TiDB
- 示例 2：使用 TiDB Lightning 并行导入单表数据

使用限制

TiDB Lightning 在运行时，需要独占部分资源，因此如果需要在单台机器上面部署多个 TiDB Lightning 实例 (不建议生产环境使用) 或多台机器共享磁盘存储时，需要注意如下使用限制：

- 每个 TiDB Lightning 实例的 tikv-importer.sorted-kv-dir 必须设置为不同的路径。多个实例共享相同的路径会导致非预期的行为，可能导致导入失败或数据出错。
- 每个 TiDB Lightning 的 checkpoint 需要分开存储。checkpoint 的详细配置见TiDB Lightning 断点续传。
  - 如果设置 checkpoint.driver = “file” (默认值)，需要确保每个实例设置的 checkpoint 的路径不同。
  - 如果设置 checkpoint.driver = “mysql”, 需要为每个实例设置不同的 schema。
- 每个 TiDB Lightning 的 log 文件应该设置为不同的路径。共享同一个 log 文件将不利于日志的查询和排查问题。
- 如果开启Web 界面 或 Debug API, 需要为每个实例的 `lightning.status-addr` 设置不同地址，否则，TiDB Lightning 进程会由于端口冲突无法启动。

12.7.3.5.2 示例 1：使用 Dumpling + TiDB Lightning 并行导入分库分表数据至 TiDB

在本示例中，假设上游为包含 10 个分表的 MySQL 集群，总共大小为 10 TiB。使用 5 个 TiDB Lightning 实例并行导入，每个实例导入 2 TiB 数据，预计可以将总导入时间 (不包含 Dumpling 导出的耗时) 由约 40 小时降至约 10 小时。

假设上游的库名为 `my_db`，每个分表的名字为 `my_table_01` ~ `my_table_10`，需要合并导入至下游的 `my_db.my_table` 表中。下面介绍具体的操作步骤。

第 1 步：使用 Dumpling 导出数据

在部署 TiDB Lightning 的 5 个节点上面分别导出两个分表的数据：

- 如果两个分表位于同一个 MySQL 实例中，可以直接使用 Dumpling 的 `--filter` 参数一次性导出。此时在使用 TiDB Lightning 导入时，指定 `data-source-dir` 为 Dumpling 数据导出的目录即可；
- 如果两个分表的数据分布在不同的 MySQL 节点上，则需要使用 Dumpling 分别导出，两次导出数据需要放置在同一父目录下不同子目录里，然后在使用 TiDB Lightning 导入时，`data-source-dir` 指定为此父级目录。

使用 Dumpling 导出数据的步骤，请参考Dumpling。

第 2 步：配置 TiDB Lightning 的数据源

创建 `tidb-lightning.toml` 配置文件，并加入如下内容：

```
[lightning]
status-addr = ":8289"

[mydumper]
#### 设置为 Dumpling 导出数据的路径，如果 Dumpling 执行了多次并分属不同的目录，
    ↪ 请将多次导出的数据置放在相同的父目录下并指定此父目录即可。
data-source-dir = "/path/to/source-dir"

[tikv-importer]
#### 是否允许向已存在数据的表导入数据。默认值为 false。
#### 当使用并行导入模式时，由于多个 TiDB Lightning 实例同时导入一张表，因此此开关必须设置为 true
    ↪ 。
incremental-import = true
#### "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
#### "tidb"：TB 级以下数据量也可以采用 "tidb" 后端模式，下游 TiDB 可正常提供服务。
backend = "local"

#### 设置本地排序数据的路径
sorted-kv-dir = "/path/to/sorted-dir"

#### 设置分库分表合并规则
[[routes]]
schema-pattern = "my_db"
table-pattern = "my_table_*"
target-schema = "my_db"
target-table = "my_table"
```

如果数据源存放在 Amazon S3 或 GCS 等外部存储中，请参考外部存储。

第 3 步：开启 TiDB Lightning 进行数据导入

在使用 TiDB Lightning 并行导入时，对每个 TiDB Lightning 节点机器配置的需求与非并行导入模式相同，每个 TiDB Lightning 节点需要消耗相同的资源，建议部署在不同的机器上。详细的部署步骤，请参考TiDB Lightning 部署与执行

依次在每台机器上面运行 TiDB Lightning。如果直接在命令行中用 nohup 启动程序，可能会因为 SIGHUP 信号而退出，建议把 nohup 放到脚本里面，如：

```
#### !/bin/bash
nohup tiup tidb-lightning -config tidb-lightning.toml > nohup.out &
```

在并行导入的场景下，TiDB Lightning 在启动任务之后，会自动进行下列检查：

- 检查本地盘空间以及 TiKV 集群是否有足够空间导入数据。检查时会对数据源进行采样，通过采样结果预估索引大小占比。由于估算中考虑了索引，因此可能会出现尽管数据源大小低于本地盘可用空间，但依然无法通过检测的情况。
- 检查 TiKV 集群的 region 分布是否均匀，以及是否存在大量空 region，如果空 region 的数量大于 max(1000, 表的数量 * 3)，即大于“1000”和“3 倍表数量”二者中的最大者，则无法执行导入。
- 检查数据源导入数据是否有序，并且根据检查结果自动调整 `mydumper.batch-size` 的大小。因此 `mydumper.batch-size` 配置不再对用户开放。

你也可以通过 `lightning.check-requirements` 配置来关闭检查，执行强制导入。更多详细检查内容，可以查看Lightning 执行前检查项。

第 4 步：查看进度

开始导入后，可以通过以下任意方式查看进度：

- 通过 grep 日志关键字 `progress` 查看进度，默认 5 分钟更新一次。
- 通过监控面板查看进度。详情请参见TiDB Lightning 监控。

等待所有的 TiDB Lightning 运行结束，则整个导入完成。

12.7.3.5.3 示例 2：使用 TiDB Lightning 并行导入单表数据

TiDB Lightning 也支持并行导入单表的数据。例如，将存放在 Amazon S3 中的多个单表文件，分别由不同的 TiDB Lightning 实例并行导入到下游 TiDB 数据库中。该方法可以加快整体导入速度。关于更多远端存储信息，请参考TiDB Lightning 支持的远端存储。

> **注意：**
>
> 在本地环境下，可以使用 Dumpling 的 `--filesize` 或 `--where` 参数，预先将单表的数据划分成不同的部分导出至多台机器的本地磁盘，此时依然可以使用并行导入功能，其配置与示例 1 相同。

假设通过 Dumpling 导出的源文件存放在 Amazon S3 云存储中，数据文件为 `my_db.my_table.00001.sql` ~ `my_db.my_table.10000.sql` 共计 10000 个 SQL 文件。如果希望使用 2 个 TiDB Lightning 实例加速导入，则需要在配置文件中增加如下设置：

```
[[mydumper.files]]
#### db schema 文件
pattern = '(?i)^(?:[^/]*/)*my_db-schema-create\.sql'
```

```
schema = "my_db"
type = "schema-schema"

[[mydumper.files]]
#### table schema 文件
pattern = '(?i)^(?:[^/]*/)*my_db\.my_table-schema\.sql'
schema = "my_db"
table = "my_table"
type = "table-schema"

[[mydumper.files]]
#### 只导入 00001~05000 这些数据文件并忽略其他文件
pattern = '(?i)^(?:[^/]*/)*my_db\.my_table\.(0[0-4][0-9][0-9][0-9]|05000)\.sql'
schema = "my_db"
table = "my_table"
type = "sql"
```

另外一个实例的配置修改为只导入 05001 ~ 10000 数据文件即可。

其他步骤请参考示例 1 中的相关步骤。

12.7.3.5.4 错误处理

部分 TiDB Lightning 节点异常终止

在并行导入过程中，如果一个或多个 TiDB Lightning 节点异常终止，需要首先根据日志中的报错明确异常退出的原因，然后根据错误类型做不同处理：

1. 如果是正常退出 (如手动 Kill 等)，或内存溢出被操作系统终止等，可以在适当调整配置后直接重启 TiDB Lightning，无须任何其他操作。
2. 如果是不影响数据正确性的报错，如网络超时，可以在每一个失败的节点上使用 tidb-lightning-ctl 工具清除断点续传源数据中记录的错误，然后重启这些异常的节点，从断点位置继续导入，详见checkpoint-error-ignore。
3. 如果是影响数据正确性的报错，如 checksum mismatched，表示源文件中有非法的数据，则需要在每一个失败的节点上使用 tidb-lightning-ctl 工具，清除失败的表中已导入的数据及断点续传相关的源数据，详见checkpoint-error-destroy。此命令会删除下游导入失败表中已导入的数据，因此，应在所有 TiDB Lightning 节点（包括任务正常结束的）重新配置和导入失败的表的数据，可以配置 filters 参数只导入报错失败的表。

导入过程中报错 “Target table is calculating checksum. Please wait until the checksum is finished and try again”

在部分并行导入场景，如果表比较多或者一些表的数据量比较小，可能会出现当一个或多个任务开始处理某个表的时候，此表对应的其他任务已经完成，并开始数据一致性校验。此时，由于数据一致性校验不支持写入其他数据，对应的任务会返回 “Target table is calculating checksum. Please wait until the checksum is finished and try again” 错误。等校验任务完成，重启这些失败的任务，报错会消失，数据的正确性也不会受影响。

12.7.3.6 TiDB Lightning 错误处理功能

从 TiDB 5.4.0 开始，你可以配置 TiDB Lightning 以跳过诸如无效类型转换、唯一键冲突等错误，让导入任务持续进行，就如同出现错误的行数据不存在一样。你可以依据生成的报告，手动修复这些错误。该功能适用于以下场景：

- 要导入的数据有少许错误
- 手动定位错误比较困难
- 如果遇到错误就重启 TiDB Lightning，代价太大

本文介绍了类型错误处理功能 (`lightning.max-error`) 和重复问题处理功能 (`tikv-importer.duplicate-resolution`) 的使用方法，以及保存这些错误的数据库 (`lightning.task-info-schema-name`)，并提供了一个示例。

12.7.3.6.1 类型错误 (Type error)

> 警告:
>
> TiDB Lightning 类型错误处理功能（`lightning.max-error`）是实验特性。不建议在生产环境中仅依赖该功能处理相关错误。

你可以通过修改配置项 `lightning.max-error` 来增加数据类型相关的容错数量。如果设置为 N，那么 TiDB Lightning 允许数据源中出现 N 个错误，而且会跳过这些错误，一旦超过这个错误数就会退出。默认值为 0，表示不允许出现错误。

这些错误会被记录到数据库中。在导入完成后，你可以查看数据库中的数据，手动进行处理。请参见错误报告。

```
[lightning]
max-error = 0
```

该配置对下列错误有效：

- 无效值。例如：在 INT 列设置了 `'Text'`
- 数字溢出。例如：在 TINYINT 列设置了 500
- 字符串溢出。例如: 在 VARCHAR(5) 列中设置了'非常长的文字'
- 零日期时间，如 `'0000-00-00'` 和 `'2021-12-00'`
- 在 NOT NULL 列中设置了 NULL
- 生成的列表达式求值失败
- 列计数不匹配。行中数值的数量和列的数量不一致
- `on-duplicate = "error"` 时，TiDB 后端的唯一键/主键冲突
- 其他 SQL 错误

下列错误是致命错误，不能通过配置 `max-error` 跳过：

- 原始 CSV、SQL 或者 Parquet 文件中的语法错误，例如未闭合的引号
- I/O、网络、或系统权限错误

在 Local 后端模式下，唯一键/主键的冲突是单独处理的。相关内容将在接下来的章节进行介绍。

12.7.3.6.2 Local-backend 模式下解决重复问题

Local-backend 模式下，TiDB Lightning 导入数据时先将数据转换成 KV 对数组（KV pairs），然后批量添加到 TiKV 中。与 TiDB-backend 模式不同，TiDB Lightning 在 Local-backend 模式下直到任务结束才会检测重复行。因此，Local-backend 模式下的重复错误不是通过 `max-error` 进行控制，而是通过 `duplicate-resolution` 配置项进行控制的。你可以通过配置该参数的行为，来决定如何处理有冲突的数据。

```
[tikv-importer]
duplicate-resolution = 'none'
```

`duplicate-resolution` 有以下三个选项：

- ‘none’：不对重复数据进行检测。如果唯一键或主键冲突确实存在，那么导入的表格里会出现不一致的数据和索引，checksum 检查的时候会失败。
- ‘record’：检测重复数据，但不会对重复数据进行修复。如果唯一键或主键冲突确实存在，那么导入的表格里会出现不一致的数据和索引，checksum 检查的时候会失败。
- ‘remove’：检测重复数据，并且删除全部重复行。导入的表格会保持一致，但是重复的行会被忽略，只能通过手动方式添加回来。

TiDB Lightning 只能检测数据源的重复项，不能解决运行 TiDB Lightning 之前的存量数据的冲突问题。

12.7.3.6.3 错误报告

所有错误都会写入下游 TiDB 集群 `lightning_task_info` 数据库中的表中。在导入完成后，你可以根据数据库中记录的内容，手动进行处理。

你可以使用 `lightning.task-info-schema-name` 配置更改数据库名称。

```
[lightning]
task-info-schema-name = 'lightning_task_info'
```

在此数据库中，TiDB Lightning 创建了 3 个表：

```
CREATE TABLE syntax_error_v1 (
    task_id     bigint NOT NULL,
    create_time datetime(6) NOT NULL DEFAULT now(6),
    table_name  varchar(261) NOT NULL,
    path        varchar(2048) NOT NULL,
    offset      bigint NOT NULL,
    error       text NOT NULL,
    context     text
);
CREATE TABLE type_error_v1 (
```

```
    task_id     bigint NOT NULL,
    create_time datetime(6) NOT NULL DEFAULT now(6),
    table_name  varchar(261) NOT NULL,
    path        varchar(2048) NOT NULL,
    offset      bigint NOT NULL,
    error       text NOT NULL,
    row_data    text NOT NULL
);
CREATE TABLE conflict_error_v1 (
    task_id     bigint NOT NULL,
    create_time datetime(6) NOT NULL DEFAULT now(6),
    table_name  varchar(261) NOT NULL,
    index_name  varchar(128) NOT NULL,
    key_data    text NOT NULL,
    row_data    text NOT NULL,
    raw_key     mediumblob NOT NULL,
    raw_value   mediumblob NOT NULL,
    raw_handle  mediumblob NOT NULL,
    raw_row     mediumblob NOT NULL,
    KEY (task_id, table_name)
);
```

type_error_v1 记录由 `max-error` 配置项管理的所有类型错误 (Type error)。每个错误一行。

conflict_error_v1 记录所有后端中的唯一键/主键冲突。每对冲突有两行。

列名	语法	类型	冲突	说明
task_id	√	√	√	生成此错误的 TiDB Lightning 任务 ID
create_table	√	√	√	记录错误的时间
table_name	√	√	√	包含错误的表的名称，格式为 ``db`.`tbl``
path	√	√		包含错误文件的路径
offset	√	√		文件中发现错误的字节位置
error	√	√		错误信息
context	√			围绕错误的文本
index_name			√	冲突中唯一键的名称。主键冲突为 'PRIMARY'
key_data			√	导致错误的行的格式化键句柄。该内容仅供人参考，机器不可读
row_data		√	√	导致错误的格式化行数据。该内容仅供人参考，机器不可读
raw_key			√	冲突的 KV 对的键
raw_value			√	冲突的 KV 对的值
raw_handle			√	冲突行的行句柄
raw_row			√	冲突行的编码值

注意：

错误报告记录的是文件偏移量，不是行号或列号，因为行号或列号的获取效率很低。你可以使用下列命令在字节位实现快速跳转（以 183 为例）：

- shell：输出前面几行

```
head -c 183 file.csv | tail
```

- shell，输出后面几行

```
tail -c +183 file.csv | head
```

- vim：`:goto 183` 或 `183go`

12.7.3.6.4 示例

在该示例中，我们准备了一个包含一些已知错误的数据源。以下是处理这些错误的具体步骤：

1. 准备数据库和表结构：

```
mkdir example && cd example
echo 'CREATE SCHEMA example;' > example-schema-create.sql
echo 'CREATE TABLE t(a TINYINT PRIMARY KEY, b VARCHAR(12) NOT NULL UNIQUE);' > example.t-
    ↪ schema.sql
```

2. 准备数据：

```
cat <<EOF > example.t.1.sql
    INSERT INTO t (a, b) VALUES
    (0, NULL),              -- 列不为空
    (1, 'one'),
    (2, 'two'),
    (40, 'forty'),          -- 与下面的 `40` 冲突
    (54, 'fifty-four'),     -- 与下面的 ``'fifty-four'`` 冲突
    (77, 'seventy-seven'),  -- 字符串长度超过 12 个字符
    (600, 'six hundred'),   -- 数字超出了 TINYINT 数据类型支持的范围
    (40, 'fourty'),         -- 与上面的 `40` 冲突
    (42, 'fifty-four');     -- 与上面的 ``'fifty-four'`` 冲突
EOF
```

3. 配置 TiDB Lightning，启用严格 SQL 模式，使用 Local 后端模式进行导入，通过删除解决重复项，并最多跳过 10 个错误：

```
cat <<EOF > config.toml
    [lightning]
    max-error = 10
    [tikv-importer]
```

```
    backend = 'local'
    sorted-kv-dir = '/tmp/lightning-tmp/'
    duplicate-resolution = 'remove'
    [mydumper]
    data-source-dir = '.'
    [tidb]
    host = '127.0.0.1'
    port = 4000
    user = 'root'
    password = ''
    sql-mode = 'STRICT_TRANS_TABLES,NO_ZERO_IN_DATE,NO_ZERO_DATE'
EOF
```

4. 运行 TiDB Lightning。因为已跳过所有错误，该命令执行完会成功退出：

```
tiup tidb-lightning -c config.toml
```

5. 验证导入的表仅包含两个正常行：

```
$ mysql -u root -h 127.0.0.1 -P 4000 -e 'select * from example.t'
+---+-----+
| a | b   |
+---+-----+
| 1 | one |
| 2 | two |
+---+-----+
```

6. 检查 type_error_v1 表是否捕获了涉及类型转换的三行：

```
$ mysql -u root -h 127.0.0.1 -P 4000 -e 'select * from lightning_task_info.type_error_v1;' -
    ↪ E
*************************** 1. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.620090
 table_name: `example`.`t`
       path: example.t.1.sql
     offset: 46
      error: failed to cast value as varchar(12) CHARACTER SET utf8mb4 COLLATE utf8mb4_bin
          ↪ for column `b` (#2): [table:1048]Column 'b' cannot be null
   row_data: (0,NULL)
*************************** 2. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.627496
 table_name: `example`.`t`
       path: example.t.1.sql
     offset: 183
```

```
      error: failed to cast value as varchar(12) CHARACTER SET utf8mb4 COLLATE utf8mb4_bin
          ↪ for column `b` (#2): [types:1406]Data Too Long, field len 12, data len 13
   row_data: (77,'seventy-seven')
*************************** 3. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.629929
 table_name: `example`.`t`
       path: example.t.1.sql
     offset: 253
      error: failed to cast value as tinyint(4) for column `a` (#1): [types:1690]constant
          ↪ 600 overflows tinyint
   row_data: (600,'six hundred')
```

7. 检查 `conflict_error_v1` 表是否捕获了具有唯一键/主键冲突的四行：

```
$ mysql -u root -h 127.0.0.1 -P 4000 -e 'select * from lightning_task_info.conflict_error_v1
    ↪ ;' --binary-as-hex -E
*************************** 1. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.669601
 table_name: `example`.`t`
 index_name: PRIMARY
   key_data: 40
   row_data: (40, "forty")
    raw_key: 0x7480000000000000C15F728000000000000028
  raw_value: 0x8000010000000020500666F727479
 raw_handle: 0x7480000000000000C15F728000000000000028
    raw_row: 0x8000010000000020500666F727479
*************************** 2. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.674798
 table_name: `example`.`t`
 index_name: PRIMARY
   key_data: 40
   row_data: (40, "fourty")
    raw_key: 0x7480000000000000C15F728000000000000028
  raw_value: 0x8000010000000020600666F75727479
 raw_handle: 0x7480000000000000C15F728000000000000028
    raw_row: 0x8000010000000020600666F75727479
*************************** 3. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.680332
 table_name: `example`.`t`
 index_name: b
   key_data: 54
   row_data: (54, "fifty-four")
```

```
    raw_key: 0x7480000000000000C15F698000000000000001016669667479 2D666FFF7572000000000000F9
  raw_value: 0x0000000000000036
 raw_handle: 0x7480000000000000C15F728000000000000036
    raw_row: 0x800001000000020A0066696674792D666F7572
*************************** 4. row ***************************
    task_id: 1635888701843303564
create_time: 2021-11-02 21:31:42.681073
 table_name: `example`.`t`
 index_name: b
   key_data: 42
   row_data: (42, "fifty-four")
    raw_key: 0x7480000000000000C15F698000000000000001016669667479 2D666FFF7572000000000000F9
  raw_value: 0x000000000000002A
 raw_handle: 0x7480000000000000C15F72800000000000002A
    raw_row: 0x800001000000020A0066696674792D666F7572
```

12.7.3.7 TiDB Lightning Web 界面

TiDB Lightning 支持在网页上查看导入进度或执行一些简单任务管理，这就是 TiDB Lightning 的服务器模式。本文将介绍服务器模式下的 Web 界面和一些常见操作。

启用服务器模式的方式有如下几种：

1. 在启动 `tidb-lightning` 时加上命令行参数 `--server-mode`。

```
./tidb-lightning --server-mode --status-addr :8289
```

2. 在配置文件中设置 `lightning.server-mode`。

```
[lightning]
server-mode = true
status-addr = ':8289'
```

TiDB Lightning 启动后，可以访问 `http://127.0.0.1:8289` 来管理程序（实际的 URL 取决于你的 `status-addr` 设置）。

服务器模式下，TiDB Lightning 不会立即开始运行，而是通过用户在 web 页面提交（多个）任务来导入数据。

12.7.3.7.1 TiDB Lightning Web 首页

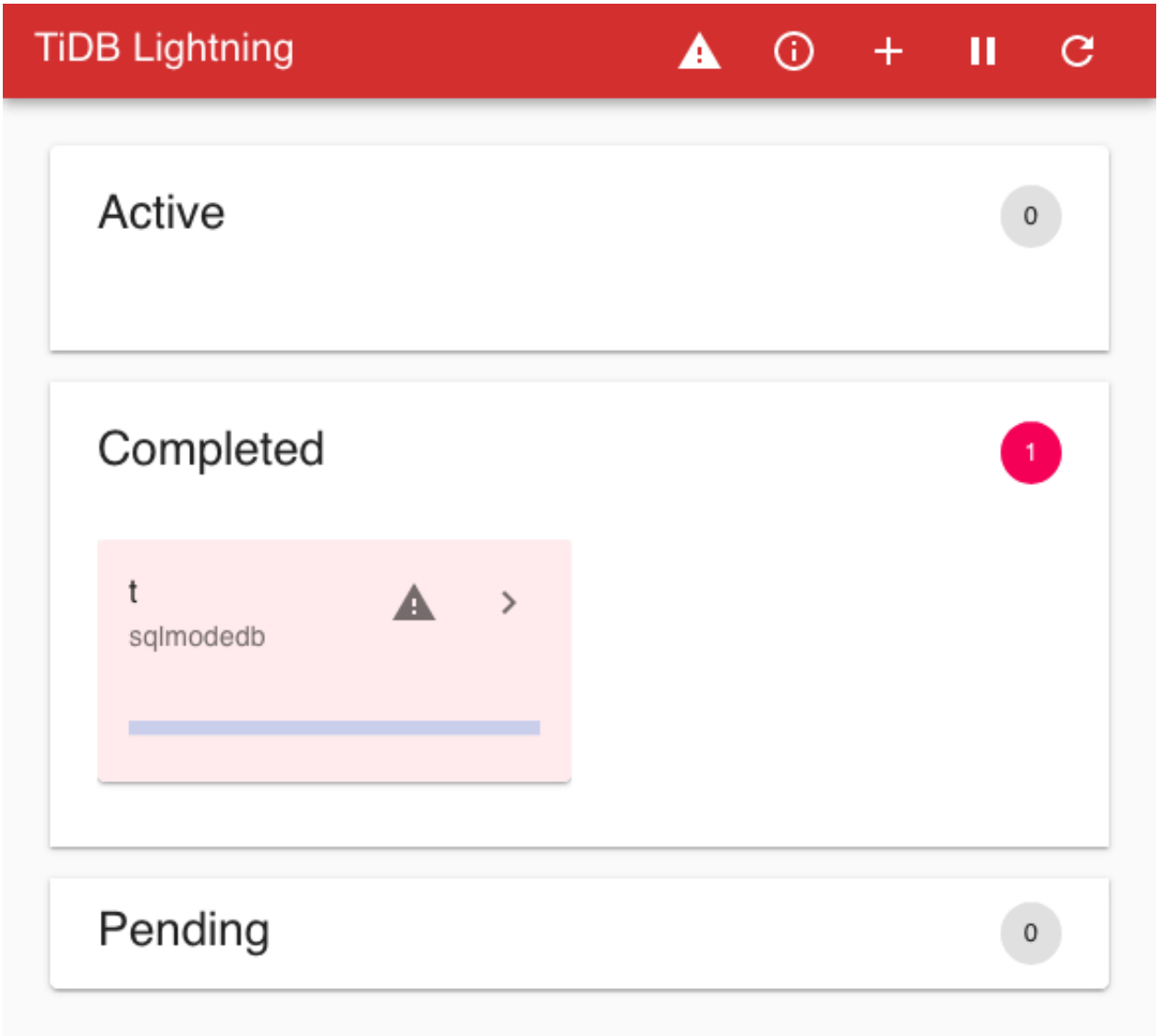


图 263: TiDB Lightning Web 首页

标题栏上图标所对应的功能，从左到右依次为：

图标	功能
“TiDB Lightning”	点击即返回首页
□	显示前一个任务的所有错误信息
ⓘ	列出当前及队列中的任务，可能会出现一个标记提示队列中任务的数量
+	提交单个任务
□/▸	暂停/继续当前操作
□	设置网页自动刷新

标题栏下方的三个面板显示了不同状态下的所有表：

- Active：当前正在导入这些表
- Completed：这些表导入成功或失败
- Pending：这些表还没有被处理

每个面板都包含用于描述表状态的卡片。

12.7.3.7.2 提交任务

点击标题栏的 + 图标提交任务。

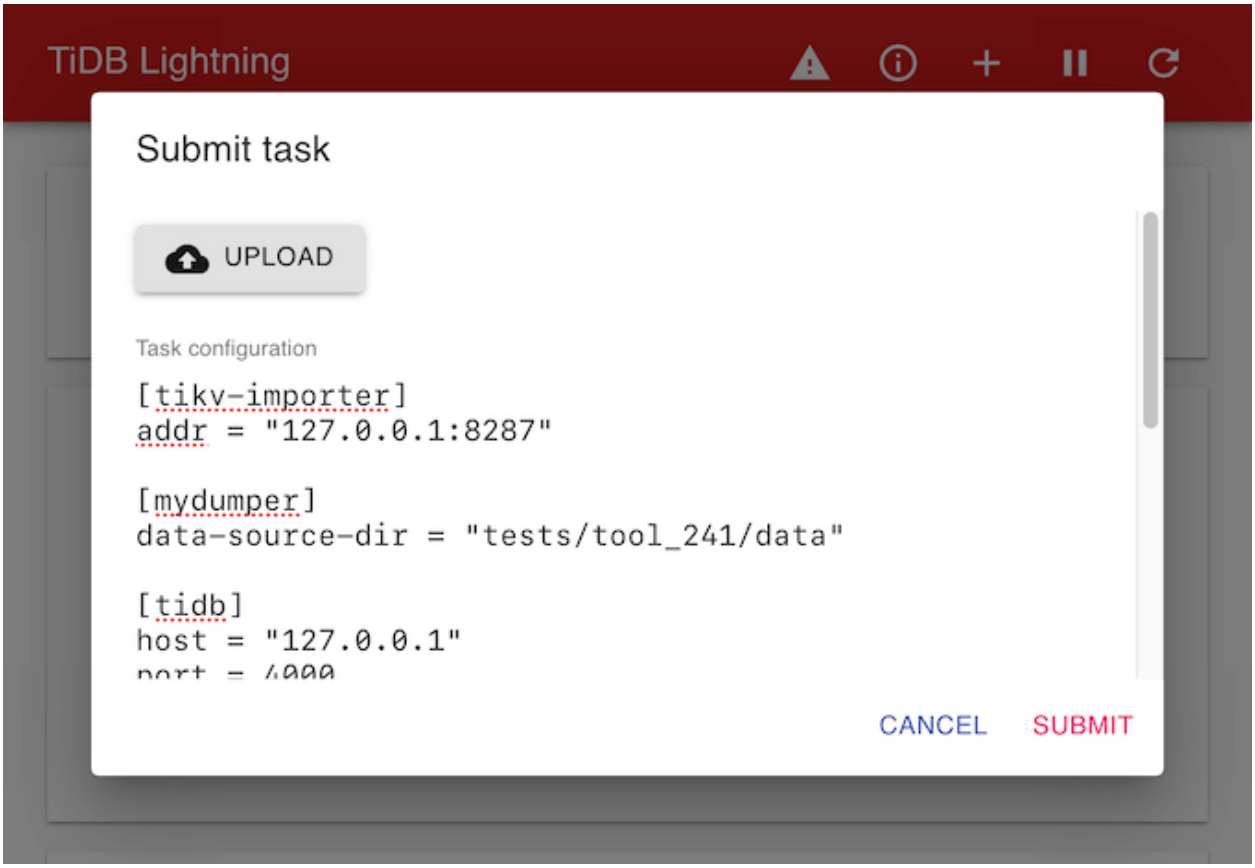


图 264: 提交任务对话框

任务 (task) 为 TOML 格式的文件，具体参考TiDB Lightning 任务配置。你也可以点击 UPLOAD 上传一个本地的 TOML 文件。

点击 SUBMIT 运行任务。如果当前有任务正在运行，新增任务会加入队列并在当前任务结束后执行。

12.7.3.7.3 查看导入进度

点击首页表格卡片上的 > 图标，查看表格导入的详细进度。



图 265: 表格导入进度

该页显示每张表的引擎文件的导入过程。

点击标题栏上的 TiDB Lightning 返回首页。

12.7.3.7.4 管理任务

单击标题栏上的 ⓘ 图标来管理当前及队列中的任务。



图 266: 任务管理页面

每个任务都是依据提交时间来标记。点击该任务将显示 JSON 格式的配置文件。

点击任务上的 ⋮ 可以对该任务进行管理。你可以立即停止任务，或重新排序队列中的任务。

12.7.4 TiDB Lightning 教程

TiDB Lightning 是一个将全量数据高速导入到 TiDB 集群的工具，目前支持 SQL 或 CSV 输出格式的数据源。你可以在以下两种场景下使用 TiDB Lightning：

- 迅速导入大量新数据。
- 备份恢复所有数据。

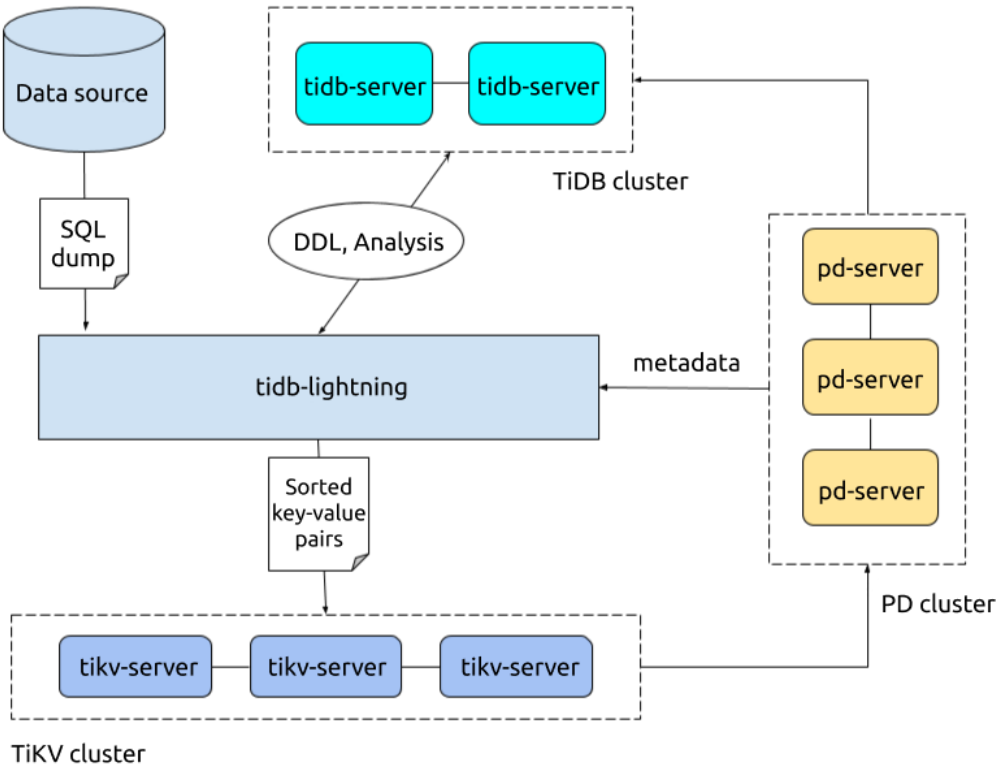


图 267: TiDB Lightning 整体架构

本教程假设使用的是若干新的、纯净版 CentOS 7 实例，你可以（使用 VMware、VirtualBox 及其他工具）在本地虚拟化或在供应商提供的平台上部署一台小型的云虚拟主机。因为 TiDB Lightning 对计算机资源消耗较高，建议分配 16 GB 以上的内存以及 32 核以上的 CPU 以获取最佳性能。

> **警告：**
>
> 本教程中的部署方法只适用于测试及功能体验，并不适用于生产或开发环境。

12.7.4.1 准备全量备份数据

我们使用dumpling 从 MySQL 导出数据，如下：

```
./bin/dumpling -h 127.0.0.1 -P 3306 -u root -t 16 -F 256MB -B test -f 'test.t[12]' -o /data/
    ↪ my_database/
```

其中：

- `-B test`：从 `test` 数据库导出。
- `-f test.t[12]`：只导出 `test.t1` 和 `test.t2` 这两个表。
- `-t 16`：使用 16 个线程导出数据。
- `-F 256MB`：将每张表切分成多个文件，每个文件大小约为 256 MB。

这样全量备份数据就导出到了 `/data/my_database` 目录中。

12.7.4.2 部署 TiDB Lightning

12.7.4.2.1 第 1 步：部署 TiDB 集群

在开始数据导入之前，需先部署一套要进行导入的 TiDB 集群。本教程以 TiDB v5.0.0 版本为例，具体部署方法可参考使用 TiUP 部署 TiDB 集群。

12.7.4.2.2 第 2 步：下载 TiDB Lightning 安装包

通过以下链接获取 TiDB Lightning 安装包（TiDB Lightning 完全兼容较低版本的 TiDB 集群，建议选择最新稳定版本）：

- v5.3.0: tidb-toolkit-v5.3.0-linux-amd64.tar.gz

12.7.4.2.3 第 3 步：启动 `tidb-lightning`

1. 将安装包里的 `bin/tidb-lightning` 及 `bin/tidb-lightning-ctl` 上传至部署 TiDB Lightning 的服务器。
2. 将数据源也上传到同样的服务器。
3. 配置 `tidb-lightning.toml`。

```
[lightning]
# 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
# 选择使用的 local 后端
backend = "local"
# 设置排序的键值对的临时存放地址，目标路径需要是一个空目录
sorted-kv-dir = "/mnt/ssd/sorted-kv-dir"

[mydumper]
# 源数据目录。
data-source-dir = "/data/my_datasource/"

# 配置通配符规则，默认规则会过滤 mysql、 sys、 INFORMATION_SCHEMA、 PERFORMANCE_SCHEMA、
    ↪ METRICS_SCHEMA、 INSPECTION_SCHEMA 系统数据库下的所有表
```

```
# 若不配置该项，导入系统表时会出现“找不到 schema”的异常
filter = ['*.*', '!mysql.*', '!sys.*', '!INFORMATION_SCHEMA.*', '!PERFORMANCE_SCHEMA.*', '!
    ↪ METRICS_SCHEMA.*', '!INSPECTION_SCHEMA.*']
[tidb]
# 目标集群的信息
host = "172.16.31.2"
port = 4000
user = "root"
password = "rootroot"
# 表架构信息在从 TiDB 的“状态端口”获取。
status-port = 10080
# 集群 pd 的地址
pd-addr = "172.16.31.3:2379"
```

4. 配置合适的参数运行 tidb-lightning。如果直接在命令行中用 nohup 启动程序，可能会因为 SIGHUP 信号而退出，建议把 nohup 放到脚本里面，如：

```
#!/bin/bash
nohup ./tidb-lightning -config tidb-lightning.toml > nohup.out &
```

12.7.4.2.4 第 4 步：检查数据

导入完毕后，TiDB Lightning 会自动退出。若导入成功，日志的最后一行会显示 tidb lightning exit。

如果出错，请参见TiDB Lightning 常见问题。

12.7.4.3 总结

本教程对 TiDB Lightning 进行了简单的介绍，并快速部署了一套简单的 TiDB Lightning 集群，将全量备份数据导入到 TiDB 集群中。

关于 TiDB Lightning 的详细功能和使用，参见TiDB Lightning 简介。

12.7.5 TiDB Lightning 部署与执行

本文主要介绍 TiDB Lightning 使用 Local-backend 进行数据导入的硬件需求，以及手动部署 TiDB Lightning 的方式。

如果使用 Local-backend 进行数据导入，TiDB Lightning 运行后，TiDB 集群将无法正常对外提供服务。如果你不希望 TiDB 集群的对外服务受到影响，可以参考TiDB Lightning TiDB-backend 中的硬件需求与部署方式进行数据导入。

12.7.5.1 注意事项

在使用 TiDB Lightning 前，需注意以下事项：

- 若 tidb-lightning 崩溃，集群会留在“导入模式”。若忘记转回“普通模式”，集群会产生大量未压缩的文件，继而消耗 CPU 并导致延迟。此时，需要使用 tidb-lightning-ctl 手动将集群转回“普通模式”：

```
bin/tidb-lightning-ctl --switch-mode=normal
```

12.7.5.2 硬件需求

tidb-lightning为资源密集程序，为了优化效能，建议硬件配置如下：

- 32+ 逻辑核 CPU
- 20GB+ 内存
- 足够储存整个数据源的 SSD 硬盘，读取速度越快越好
- 使用万兆网卡，带宽需要 1GB/s 以上
- 运行过程默认会占满 CPU，建议单独部署。条件不允许的情况下可以和其他组件（比如 tikv-server）部署在同一台机器上，然后通过配置 region-concurrency 限制 tidb-lightning 使用 CPU 资源。

> 注意：
>
> - tidb-lightning 是 CPU 密集型程序，如果和其它程序混合部署，需要通过 region-concurrency 限制 tidb-lightning 的 CPU 实际占用核数，否则会影响其他程序的正常运行。建议将混合部署机器上 75% 的 CPU 资源分配给 tidb-lightning。例如，机器为 32 核，则 tidb-lightning 的 region-concurrency 可设为“24”。

此外，目标 TiKV 集群必须有足够空间接收新导入的数据。除了标准硬件配置以外，目标 TiKV 集群的总存储空间必须大于 数据源大小 × 副本数量 × 2。例如集群默认使用 3 副本，那么总存储空间需为数据源大小的 6 倍以上。

12.7.5.3 导出数据

使用dumpling 从 MySQL 导出数据，如下：

```
./bin/dumpling -h 127.0.0.1 -P 3306 -u root -t 16 -F 256MB -B test -f 'test.t[12]' -o /data/
    ↪ my_database/
```

其中：

- -B test：从 test 数据库导出。
- -f test.t[12]：只导出 test.t1 和 test.t2 这两个表。
- -t 16：使用 16 个线程导出数据。
- -F 256MB：将每张表切分成多个文件，每个文件大小约为 256 MB。

如果数据源是 CSV 文件，请参考CSV 支持获取配置信息。

12.7.5.4 部署 TiDB Lightning

本节介绍 TiDB Lightning 的部署方式：手动部署。

12.7.5.4.1 手动部署 TiDB Lightning

第 1 步：部署 TiDB 集群

在开始数据导入之前，需先部署一套要进行导入的 TiDB 集群，建议使用最新版本进行部署。部署方法可参考使用 TiUP 部署 TiDB 集群。

第 2 步：下载 TiDB Lightning 安装包

在工具下载页面下载 TiDB Lightning 安装包（TiDB Lightning 完全兼容较低版本的 TiDB 集群，建议选择最新稳定版本）。

第 3 步：启动 `tidb-lightning`

1. 从安装包上传 `bin/tidb-lightning` 及 `bin/tidb-lightning-ctl`。
2. 将数据源写入到同样的机器。
3. 配置 `tidb-lightning.toml`。对于没有出现在下述模版中的配置，TiDB Lightning 给出配置错误的提醒并退出。`sorted-kv-dir`需要设置为一个空的目录，并且确保所在的磁盘有较多空闲的空间。

```
[lightning]

# 转换数据的并发数，默认为逻辑 CPU 数量，不需要配置。
# 混合部署的情况下可以配置为逻辑 CPU 的 75% 大小。
# region-concurrency =

# 日志
level = "info"
file = "tidb-lightning.log"

[tikv-importer]
# backend 设置为 local 模式
backend = "local"
# 设置本地临时存储路径
sorted-kv-dir = "/mnt/ssd/sorted-kv-dir"

[mydumper]
# 源数据目录。
data-source-dir = "/data/my_database"

[tidb]
# 目标集群的信息。tidb-server 的监听地址，填一个即可。
host = "172.16.31.1"
port = 4000
user = "root"
password = ""
# 表架构信息在从 TiDB 的“状态端口”获取。
status-port = 10080
# pd-server 的地址，填一个即可
```

```
pd-addr = "172.16.31.4:2379"
```

上面仅列出了 tidb-lightning 的基本配置信息。完整配置信息请参考tidb-lightning 配置说明。

4. 运行 tidb-lightning。如果直接在命令行中用 nohup 启动程序，可能会因为 SIGHUP 信号而退出，建议把 nohup 放到脚本里面，如：

```
#!/bin/bash
nohup ./tidb-lightning -config tidb-lightning.toml > nohup.out &
```

12.7.5.5 升级 TiDB Lightning

你可以通过替换二进制文件升级 TiDB Lightning，无需其他配置。重启 TiDB Lightning 的具体操作参见FAQ。

如果当前有运行的导入任务，推荐任务完成后再升级 TiDB Lightning。否则，你可能需要从头重新导入，因为无法保证断点可以跨版本工作。

12.7.6 TiDB Lightning 配置参数

你可以使用配置文件或命令行配置 TiDB Lightning。本文主要介绍 TiDB Lightning 的全局配置、任务配置和 TiKV Importer 的配置，以及如何使用命令行进行参数配置。

12.7.6.1 配置文件

TiDB Lightning 的配置文件分为“全局”和“任务”两种类别，二者在结构上兼容。只有当服务器模式开启时，全局配置和任务配置才会有区别；默认情况下，服务器模式为禁用状态，此时 TiDB Lightning 只会执行一个任务，且全局和任务配置使用同一配置文件。

12.7.6.1.1 TiDB Lightning 全局配置

```
##### tidb-lightning 全局配置

[lightning]
### 用于拉取 web 界面和 Prometheus 监控项的 HTTP 端口。设置为 0 时为禁用状态。
status-addr = ':8289'

### 切换为服务器模式并使用 web 界面
### 详情参见“TiDB Lightning web 界面”文档
server-mode = false

### 日志
level = "info"
file = "tidb-lightning.log"
max-size = 128 # MB
max-days = 28
max-backups = 14
```

12.7.6.1.2 TiDB Lightning 任务配置

```
##### tidb-lightning 任务配置

[lightning]
### 启动之前检查集群是否满足最低需求。
### check-requirements = true

### 引擎文件的最大并行数。
### 每张表被切分成一个用于存储索引的“索引引擎”和若干存储行数据的“数据引擎”。
### 这两项设置控制两种引擎文件的最大并发数。
### 这两项设置的值会影响 tikv-importer 的内存和磁盘用量。
### 两项数值之和不能超过 tikv-importer 的 max-open-engines 的设定。
index-concurrency = 2
table-concurrency = 6

### 数据的并发数。默认与逻辑 CPU 的数量相同。
### 混合部署的情况下可以将其大小配置为逻辑 CPU 数的 75%，以限制 CPU 的使用。
### region-concurrency =

### I/O 最大并发数。I/O 并发量太高时，会因硬盘内部缓存频繁被刷新
### 而增加 I/O 等待时间，导致缓存未命中和读取速度降低。
### 对于不同的存储介质，此参数可能需要调整以达到最佳效率。
io-concurrency = 5

### TiDB Lightning 停止迁移任务之前能容忍的最大非严重 (non-fatal errors) 错误数。
### 在忽略非严重错误所在的行数据之后，迁移任务可以继续执行。
### 将该值设置为 N，表示 TiDB Lightning 会在遇到第 (N+1) 个错误时停止迁移任务。
### 被忽略的行会被记录到位于目标集群的 "task info" 数据库中。最大错误数可以通过下面参数配置。
max-error = 0
### 参数 task-info-schema-name 指定用于存储 TiDB Lightning 执行结果的数据库。
### 要关闭该功能，需要将该值设置为空字符串。
### task-info-schema-name = 'lightning_task_info'

### 在并行导入模式下，在目标集群保存各个 TiDB Lightning 实例元信息的 schema 名字，默认为 "
    ↪ lightning_metadata"。
### 如果未开启并行导入模式，无须设置此配置项。
### **注意：**
### - 对于参与同一批并行导入的每个 TiDB Lightning 实例，该参数设置的值必须相同，
    ↪ 否则将无法确保导入数据的正确性。
### - 如果开启并行导入模式，需要确保导入使用的用户（对于 tidb.user 配置项）
    ↪ 有权限创建和访问此配置对应的库。
### - TiDB Lightning 在导入完成后会删除此 schema，因此不要使用已存在的库名配置该参数。
meta-schema-name = "lightning_metadata"

[security]
```

```
### 指定集群中用于 TLS 连接的证书和密钥。
### CA 的公钥证书。如果留空，则禁用 TLS。
### ca-path = "/path/to/ca.pem"
### 此服务的公钥证书。
### cert-path = "/path/to/lightning.pem"
### 该服务的密钥。
### key-path = "/path/to/lightning.key"

[checkpoint]
### 是否启用断点续传。
### 导入数据时，TiDB Lightning 会记录当前表导入的进度。
### 所以即使 TiDB Lightning 或其他组件异常退出，在重启时也可以避免重复再导入已完成的数据。
enable = true
### 存储断点的数据库名称。
schema = "tidb_lightning_checkpoint"
### 存储断点的方式。
###  - file：存放在本地文件系统。
###  - mysql：存放在兼容 MySQL 的数据库服务器。
driver = "file"

### dsn 是数据源名称 (data source name)，表示断点的存放位置。
### 若 driver = "file"，则 dsn 为断点信息存放的文件路径。
#若不设置该路径，则默认存储路径为“/tmp/CHECKPOINT_SCHEMA.pb”。
### 若 driver = "mysql"，则 dsn 为“用户:密码@tcp(地址:端口)/”格式的 URL。
### 若不设置该 URL，则默认会使用 [tidb] 部分指定的 TiDB 服务器来存储断点。
### 为减少目标 TiDB 集群的压力，建议指定另一台兼容 MySQL 的数据库服务器来存储断点。
### dsn = "/tmp/tidb_lightning_checkpoint.pb"

### 所有数据导入成功后是否保留断点。设置为 false 时为删除断点。
### 保留断点有利于进行调试，但会泄漏关于数据源的元数据。
### keep-after-success = false

[tikv-importer]
### "local"：默认使用该模式，适用于 TB 级以上大数据量，但导入期间下游 TiDB 无法对外提供服务。
### "tidb"：TB 级以下数据量也可以采用 "tidb" 后端模式，下游 TiDB 可正常提供服务。
### backend = "local"
### 是否允许向已存在数据的表导入数据。默认值为 false。
### 当使用并行导入模式时，由于多个 TiDB Lightning 实例同时导入一张表，因此此开关必须设置为 true。
### incremental-import = false
### 当后端是 “importer” 时，tikv-importer 的监听地址（需改为实际地址）。
addr = "172.16.31.10:8287"
### 当后端是 “tidb” 时，插入重复数据时执行的操作。
### - replace：新数据替代已有数据
### - ignore：保留已有数据，忽略新数据
### - error：中止导入并报错
```

```
### on-duplicate = "replace"
### 当后端模式为 'local' 时，设置是否检测和解决重复的记录（唯一键冲突）。
### 目前支持三种解决方法：
###  - record：仅将重复记录添加到目的 TiDB 中的 `lightning_task_info.conflict_error_v1` 表中。
    ↪ 注意，该方法要求目的 TiKV 的版本为 v5.2.0 或更新版本。如果版本过低，则会启用下面的 'none'
    ↪  模式。
###  - none：不检测重复记录。该模式是三种模式中性能最佳的，但是可能会导致目的 TiDB
    ↪ 中出现数据不一致的情况。
###  - remove：记录所有的重复记录，和 'record' 模式相似。但是会删除所有的重复记录，以确保目的
    ↪ TiDB 中的数据状态保持一致。
### duplicate-resolution = 'none'
### 当后端是 “local” 时，一次请求中发送的 KV 数量。
### send-kv-pairs = 32768
### 当后端是 “local” 时，本地进行 KV 排序的路径。如果磁盘性能较低（如使用机械盘），建议设置成与
    ↪  `data-source-dir` 不同的磁盘，这样可有效提升导入性能。
### sorted-kv-dir = ""
### 当后端是 “local” 时，TiKV 写入 KV 数据的并发度。当 TiDB Lightning 和 TiKV
    ↪ 直接网络传输速度超过万兆的时候，可以适当增加这个值。
### range-concurrency = 16

[mydumper]
### 设置文件读取的区块大小，确保该值比数据源的最长字符串长。
read-block-size = "64KiB" # 默认值

### 引擎文件需按顺序导入。由于并行处理，多个数据引擎几乎在同时被导入，
### 这样形成的处理队列会造成资源浪费。因此，为了合理分配资源，TiDB Lightning
### 稍微增大了前几个区块的大小。该参数也决定了比例系数，即在完全并发下
### “导入”和“写入”过程的持续时间比。这个值可以通过计算 1 GiB 大小的
### 单张表的（导入时长/写入时长）得到。在日志文件中可以看到精确的时间。
### 如果“导入”更快，区块大小的差异就会更小；比值为 0 时则说明区块大小一致。
### 取值范围为（0 <= batch-import-ratio < 1）。
batch-import-ratio = 0.75

### 本地源数据目录或外部存储 URL
data-source-dir = "/data/my_database"
### 如果 no-schema = true，那么 TiDB Lightning 假设目标 TiDB 集群上
### 已有表结构，并且不会执行 `CREATE TABLE` 语句。
no-schema = false
### 指定包含 `CREATE TABLE` 语句的表结构文件的字符集。只支持下列选项：
###  - utf8mb4：表结构文件必须使用 UTF-8 编码，否则会报错。
###  - gb18030：表结构文件必须使用 GB-18030 编码，否则会报错。
###  - auto：自动判断文件编码是 UTF-8 还是 GB-18030，两者皆非则会报错（默认）。
###  - binary：不尝试转换编码。
character-set = "auto"
```

```
### 指定源数据文件的字符集，Lightning 会在导入过程中将源文件从指定的字符集转换为 UTF-8 编码。
### 该配置项目前仅用于指定 CSV 文件的字符集。只支持下列选项：
###  - utf8mb4：源数据文件使用 UTF-8 编码。
###  - GB18030：源数据文件使用 GB-18030 编码。
###  - GBK：源数据文件使用 GBK 编码（GBK 编码是对 GB-2312 字符集的拓展，也被称为 Code Page
    ↪ 936）。
###  - binary：不尝试转换编码（默认）。
### 留空此配置将默认使用 "binary"，即不尝试转换编码。
### 需要注意的是，Lightning 不会对源数据文件的字符集做假定，仅会根据此配置对数据进行转码并导入。
### 如果字符集设置与源数据文件的实际编码不符，可能会导致导入失败、导入缺失或导入数据乱码。
data-character-set = "binary"
### 指定在源数据文件的字符集转换过程中，出现不兼容字符时的替换字符。
### 此项不可与字段分隔符、引用界定符和换行符号重复。
### 默认值为 "\uFFFD"，即 UTF-8 编码中的 "error" Rune 或 Unicode replacement character。
### 改变默认值可能会导致潜在的源数据文件解析性能下降。
data-invalid-char-replace = "\uFFFD"

### “严格”格式的导入数据可加快处理速度。
### strict-format = true 要求：
### 在 CSV 文件的所有记录中，每条数据记录的值不可包含字符换行符（U+000A 和 U+000D，即 \r 和 \n）
### 甚至被引号包裹的字符换行符都不可包含，即换行符只可用来分隔行。
### 导入数据源为严格格式时，TiDB Lightning 会快速定位大文件的分割位置进行并行处理。
### 但是如果输入数据为非严格格式，可能会将一条完整的数据分割成两部分，导致结果出错。
### 为保证数据安全而非追求处理速度，默认值为 false。
strict-format = false

### 如果 strict-format = true，TiDB Lightning 会将 CSV 大文件分割为多个文件块进行并行处理。max-
    ↪ region-size 是分割后每个文件块的最大大小。
### max-region-size = "256MiB" # 默认值

### 只导入与该通配符规则相匹配的表。详情见相应章节。
filter = ['*.*', '!mysql.*', '!sys.*', '!INFORMATION_SCHEMA.*', '!PERFORMANCE_SCHEMA.*', '!
    ↪ METRICS_SCHEMA.*', '!INSPECTION_SCHEMA.*']

### 配置 CSV 文件的解析方式。
[mydumper.csv]
### 字段分隔符，支持一个或多个字符，默认值为 ','。
separator = ','
### 引用定界符，设置为空表示字符串未加引号。
delimiter = '"'
### 行尾定界字符，支持一个或多个字符。设置为空（默认值）表示 "\n"（换行）和 "\r\n" （回车+换行
    ↪ ），均表示行尾。
terminator = ""
### CSV 文件是否包含表头。
### 如果 header = true，将跳过首行。
```

```
header = true
### CSV 文件是否包含 NULL。
### 如果 not-null = true，CSV 所有列都不能解析为 NULL。
not-null = false
### 如果 not-null = false（即 CSV 可以包含 NULL），
### 为以下值的字段将会被解析为 NULL。
null = '\N'
### 是否对字段内“\“进行转义
backslash-escape = true
### 如果有行以分隔符结尾，删除尾部分隔符。
trim-last-separator = false

### [[mydumper.files]]
### 解析 AWS Aurora parquet 文件所需的表达式
### pattern = '(?i)^(?:[^/]*/)*([a-z0-9_]+)\.([a-z0-9_]+)/(?:[^/]*/)*(?:[a-z0-9\-_.]+\.(parquet))
    ↪ $'
### schema = '$1'
### table = '$2'
### type = '$3'
###
### 设置分库分表合并规则，将 my_db1 中的 table1、table2 两个表，以及 my_db2 中的 table3、table4
    ↪ 两个表，共计 2 个数据库中的 4 个表都导入到目的数据库 my_db 中的 table5 表中。
### [[routes]]
### schema-pattern = "my_db1"
### table-pattern = "table[1-2]"
### target-schema = "my_db"
### target-table = "table5"
###
### [[routes]]
### schema-pattern = "my_db2"
### table-pattern = "table[3-4]"
### target-schema = "my_db"
### target-table = "table5"

[tidb]
### 目标集群的信息。tidb-server 的地址，填一个即可。
host = "172.16.31.1"
port = 4000
user = "root"
password = ""
### 表结构信息从 TiDB 的“status-port”获取。
status-port = 10080
### pd-server 的地址，填一个即可。
pd-addr = "172.16.31.4:2379"
### tidb-lightning 引用了 TiDB 库，并生成产生一些日志。
```

```
### 设置 TiDB 库的日志等级。
log-level = "error"

### 设置 TiDB 会话变量，提升 Checksum 和 Analyze 的速度。
### 各参数定义可参阅”控制 Analyze 并发度“文档
build-stats-concurrency = 20
distsql-scan-concurrency = 100
index-serial-scan-concurrency = 20
checksum-table-concurrency = 16

### 解析和执行 SQL 语句的默认 SQL 模式。
sql-mode = "ONLY_FULL_GROUP_BY,NO_ENGINE_SUBSTITUTION"
### `max-allowed-packet` 设置数据库连接允许的最大数据包大小，
### 对应于系统参数中的 `max_allowed_packet`。 如果设置为 0，
### 会使用下游数据库 global 级别的 `max_allowed_packet`。
max-allowed-packet = 67_108_864

### SQL 连接是否使用 TLS。可选值为：
###  * ""            - 如果填充了 [tidb.security] 部分，则强制使用 TLS（与 "cluster" 情况相同），
    ↪ 否则与 "false" 情况相同
###  * "false"       - 禁用 TLS
###  * "cluster"     - 强制使用 TLS 并使用 [tidb.security] 部分中指定的 CA 验证服务器的证书
###  * "skip-verify" - 强制使用 TLS，但不验证服务器的证书（不安全！）
###  * "preferred"   - 与 "skip-verify" 相同，但是如果服务器不支持 TLS，则会退回到未加密的连接
### tls = ""
### 指定证书和密钥用于 TLS 连接 MySQL。
### 默认为 [security] 部分的副本。
### [tidb.security]
### CA 的公钥证书。设置为空字符串可禁用 SQL 的 TLS。
### ca-path = "/path/to/ca.pem"
### 该服务的公钥证书。默认为 `security.cert-path` 的副本
### cert-path = "/path/to/lightning.pem"
### 此服务的私钥。默认为 `security.key-path` 的副本
### key-path = "/path/to/lightning.key"

### 数据导入完成后，tidb-lightning 可以自动执行 Checksum、Compact 和 Analyze 操作。
### 在生产环境中，建议这将些参数都设为 true。
### 执行的顺序为：Checksum -> Compact -> Analyze。
[post-restore]
### 配置是否在导入完成后对每一个表执行 `ADMIN CHECKSUM TABLE <table>` 操作来验证数据的完整性。
### 可选的配置项：
### - "required"（默认）。在导入完成后执行 CHECKSUM 检查，如果 CHECKSUM 检查失败，则会报错退出。
### - "optional"。在导入完成后执行 CHECKSUM 检查，如果报错，会输出一条 WARN 日志并忽略错误。
### - "off"。导入结束后不执行 CHECKSUM 检查。
### 默认值为 "required"。从 v4.0.8 开始，checksum 的默认值由此前的 "true" 改为 "required"。
```

```
### 注意：考虑到与旧版本的兼容性，依然可以在本配置项设置 `true` 和 `false` 两个布尔值，其效果与
    ↪ "required" 和 `off` 相同。
checksum = "required"
### 配置是否在 CHECKSUM 结束后对所有表逐个执行 `ANALYZE TABLE <table>` 操作。
### 此配置的可选配置项与 `post-restore` 相同，但默认值为 "optional"。
analyze = "optional"

### 如果设置为 true，会在导入每张表后执行一次 level-1 Compact。
### 默认值为 false。
level-1-compact = false

### 如果设置为 true，会在导入过程结束时对整个 TiKV 集群执行一次 full Compact。
### 默认值为 false。
compact = false


### 设置周期性后台操作。
### 支持的单位：h（时）、m（分）、s（秒）。
[cron]
### TiDB Lightning 自动刷新导入模式状态的持续时间，该值应小于 TiKV 对应的设定值。
switch-mode = "5m"
### 在日志中打印导入进度的持续时间。
log-progress = "5m"
```

12.7.6.1.3 TiKV Importer 配置参数

```
### TiKV Importer 配置文件模版

### 日志文件
log-file = "tikv-importer.log"
### 日志等级：trace, debug, info, warn, error 和 off
log-level = "info"

### 状态服务器的监听地址。
### Prometheus 可以从这个地址抓取监控指标。
status-server-address = "0.0.0.0:8286"

[server]
### tikv-importer 的监听地址，tidb-lightning 需要连到这个地址进行数据写入。
addr = "0.0.0.0:8287"
### gRPC 服务器的线程池大小。
grpc-concurrency = 16

[metric]
```

```
### 当使用 Prometheus Pushgateway 时会涉及相关设置。通常可以通过 Prometheus 从
    ↪ 状态服务器地址中抓取指标。
### 给 Prometheus 客户端推送的 job 名称。
job = "tikv-importer"
### 给 Prometheus 客户端推送的间隔。
interval = "15s"
### Prometheus Pushgateway 的地址。
address = ""

[rocksdb]
### background job 的最大并发数。
max-background-jobs = 32

[rocksdb.defaultcf]
### 数据在刷新到硬盘前能存于内存的容量上限。
write-buffer-size = "1GB"
### 内存中写缓冲器的最大数量。
max-write-buffer-number = 8

### 各个压缩层级使用的算法。
### 第 0 层的算法用于压缩 KV 数据。
### 第 6 层的算法用于压缩 SST 文件。
### 第 1 至 5 层的算法目前尚未使用。
compression-per-level = ["lz4", "no", "no", "no", "no", "no", "lz4"]

[rocksdb.writecf]
### 同上
compression-per-level = ["lz4", "no", "no", "no", "no", "no", "lz4"]

[security]
### TLS 证书的路径。空字符串表示禁用安全连接。
### ca-path = ""
### cert-path = ""
### key-path = ""

[import]
### 存储引擎文件的文件夹路径
import-dir = "/mnt/ssd/data.import/"
### 处理 RPC 请求的线程数
num-threads = 16
### 导入 job 的并发数。
num-import-jobs = 24
### 预处理 Region 最长时间。
### max-prepare-duration = "5m"
### 把要导入的数据切分为这个大小的 Region。
```

```
#region-split-size = "512MB"
### 设置 stream-channel-window 的大小。
### channel 满了之后 stream 会处于阻塞状态。
### stream-channel-window = 128
### 同时打开引擎文档的最大数量。
max-open-engines = 8
### Importer 上传至 TiKV 的最大速度（字节/秒）。
### upload-speed-limit = "512MB"
### 目标存储可用空间比率（store_available_space/store_capacity）的最小值。
### 如果目标存储空间的可用比率低于该值，Importer 将会暂停上传 SST
### 来为 PD 提供足够时间进行 Regions 负载均衡。
min-available-ratio = 0.05
```

12.7.6.2 命令行参数

12.7.6.2.1 tidb-lightning

使用 tidb-lightning 可以对下列参数进行配置：

参数	描述	对应配置项
–config file	从 file 读取全局设置。如果没有指定则使用默认设置。	
-V	输出程序的版本	
-d directory	读取数据的本地目录或外部存储 URL	mydumper ↪ . ↪ data ↪ - ↪ source ↪ - ↪ dir

参数	描述	对应配置项
-L level	日志的等级：debug、info、warn、error 或 fatal (默认为 info)	lightning.log.level
-f rule	表库过滤的规则 (可多次指定)	mydumper.filter
–backend backend	选择后端的模式：local、importer 或 tidb	tikv-importer.backend
–log-file file	日志文件路径 (默认值为 /tmp/lightning.log.{timestamp}, 设置为 '-' 表示日志输出到终端)	lightning.log.file

参数	描述	对应配置项
–status-addr ip:port	TiDB Lightning 服务器的监听地址	lightning.status-port
–importer host:port	TiKV Importer 的地址	tikv-importer.addr
–pd-urls host:port	PD endpoint 的地址	tidb.pd-addr
–tidb-host host	TiDB Server 的 host	tidb.host
–tidb-port port	TiDB Server 的端口（默认为 4000）	tidb.port
–tidb-status port	TiDB Server 的状态端口的（默认为 10080）	tidb.status-port
–tidb-user user	连接到 TiDB 的用户名	tidb.user
–tidb-password pass-word	连接到 TiDB 的密码	tidb.password

参数	描述	对应配置项
–no-schema	忽略表结构文件，直接从 TiDB 中获取表结构信息	mydumper ↪ .no ↪ - ↪ schema ↪
–enable-checkpoint bool	是否启用断点 (默认值为 true)	checkpoint ↪ . ↪ enable ↪
–analyze level	导入后分析表信息，可选值为 required、optional (默认值)、off	post-restore ↪ . ↪ analyze ↪
–checksum level	导入后比较校验和，可选值为 required (默认值)、optional、off	post-restore ↪ . ↪ checksum ↪
–check-requirement bool	开始之前检查集群版本兼容性 (默认值为 true)	lightning ↪ . ↪ check ↪ - ↪ requirements ↪

参数	描述	对应配置项
–ca file	TLS 连接的 CA 证书路径	security ↪ .ca ↪ - ↪ path ↪
–cert file	TLS 连接的证书路径	security ↪ . ↪ cert ↪ - ↪ path ↪
–key file	TLS 连接的私钥路径	security ↪ . ↪ key ↪ - ↪ path ↪
–server-mode	在服务器模式下启动 TiDB Lightning	lightning ↪ . ↪ server ↪ - ↪ mode ↪

如果同时对命令行参数和配置文件中的对应参数进行更改，命令行参数将优先生效。例如，在 `cfg.toml` 文件中，不管对日志等级做出什么修改，运行 `./tidb-lightning -L debug --config cfg.toml` 命令总是将日志级别设置为 “debug”。

12.7.6.2.2 `tidb-lightning-ctl`

使用 `tidb-lightning-ctl` 可以对下列参数进行配置：

参数	描述
–compact	执行 full compact
–switch-mode mode	将每个 TiKV Store 切换到指定模式（normal 或 import）
–fetch-mode	打印每个 TiKV Store 的当前模式
–import-engine uuid	将 TiKV Importer 上关闭的引擎文件导入到 TiKV 集群
–cleanup-engine uuid	删除 TiKV Importer 上的引擎文件
–checkpoint-dump folder	将当前的断点以 CSV 格式存储到文件夹中
–checkpoint-error-destroy tablename	删除断点，如果报错则删除该表
–checkpoint-error-ignore tablename	忽略指定表中断点的报错
–checkpoint-remove tablename	无条件删除表的断点

tablename 必须是`db`.`tbl` 中的限定表名（包括反引号），或关键词 all。

此外，上表中所有 tidb-lightning 的参数也适用于 tidb-lightning-ctl。

12.7.6.2.3 tikv-importer

使用 tikv-importer 可以对下列参数进行配置：

参数	描述	对应配置项
-C, –config file	从 file 读取配置。如果没有指定，则使用默认设置	
-V, –version	输出程序的版本	
-A, –addr ip:port	TiKV Importer 服务器的监听地址	server.addr
–status-server ip:port	状态服务器的监听地址	status-server-address
–import-dir dir	引擎文件的存储目录	import.import-dir
–log-level level	日志的等级：trace、debug、info、warn、error 或 off	log-level
–log-file file	日志文件路径	log-file

12.7.7 TiDB Lightning 监控告警

tidb-lightning 支持使用 Prometheus 采集监控指标 (metrics)。本文主要介绍 TiDB Lightning 的监控配置与监控指标。

12.7.7.1 监控配置

如果是手动部署 TiDB Lightning，则参照以下步骤进行配置。

只要 Prometheus 能发现 tidb-lightning 和 tikv-importer 的监控地址，就能收集对应的监控指标。

监控的端口可在 tidb-lightning.toml 中配置：

```
[lightning]
### 用于调试和 Prometheus 监控的 HTTP 端口。输入 0 关闭。
pprof-port = 8289

...
```

监控的端口也可在 tikv-importer.toml 配置：

```
### 状态服务器的监听地址
status-server-address = '0.0.0.0:8286'
```

配置 Prometheus 后，tidb-lightning 才能发现服务器。配置方法如下，将服务器地址直接添加至 scrape_configs 部分：

```
...
scrape_configs:
  - job_name: 'lightning'
    static_configs:
```

```
    - targets: ['192.168.20.10:8289']
- job_name: 'tikv-importer
  static_configs:
    - targets: ['192.168.20.9:8286']
```

12.7.7.2 Grafana 面板

Grafana 的可视化面板可以让你在网页上监控 Prometheus 指标。

使用 TiUP 部署 TiDB 集群时，需要在拓扑配置中添加 Grafana 和 Prometheus，TiUP 会同时部署一套 Grafana + Prometheus 的监控系统。此时需先导入面板的 JSON 文件。

12.7.7.2.1 第一行：速度面板

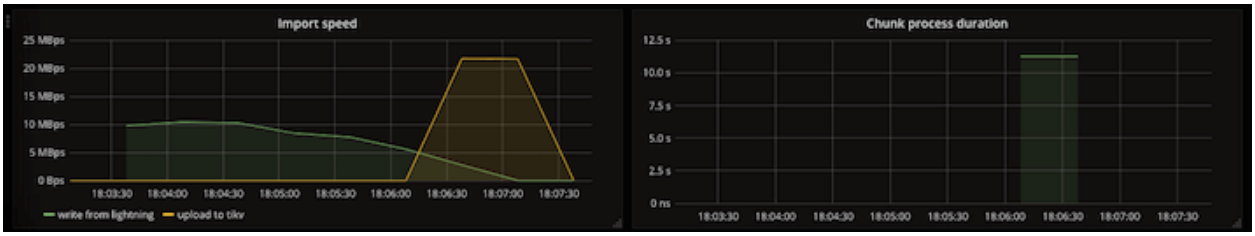


图 268: 第一行速度面板

面板名称	序列	描述
Import speed	write from lightning	从 TiDB Lightning 向 TiKV Importer 发送键值对的速度，取决于每个表的复杂性
Import speed	upload to tikv	从 TiKV Importer 上传 SST 文件到所有 TiKV 副本的总体速度

面板名称	序列	描述
Chunk process duration		完全编码单个数据文件所需的平均时间

有时导入速度会降到 0，这是为了平衡其他部分的速度，属于正常现象。

12.7.7.2.2 第二行：进度面板

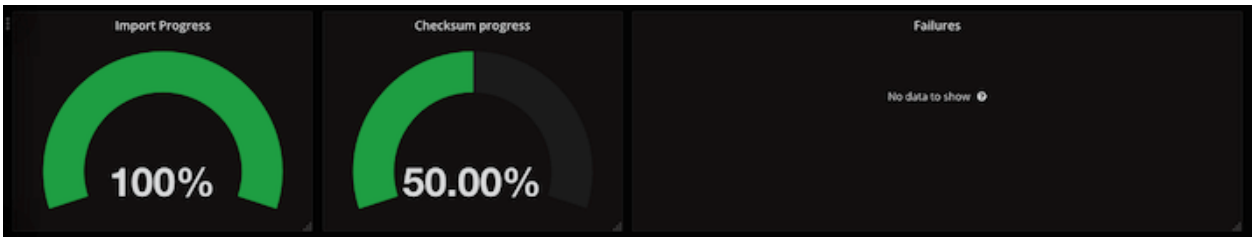


图 269: 第二行进度面板

面板名称	描述
Import progress	已编码的文件所占百分比
Checksum progress	已导入的表所占百分比
Failures	导入失败的表的数量以及故障点，通常为空

12.7.7.2.3 第三行：资源使用面板

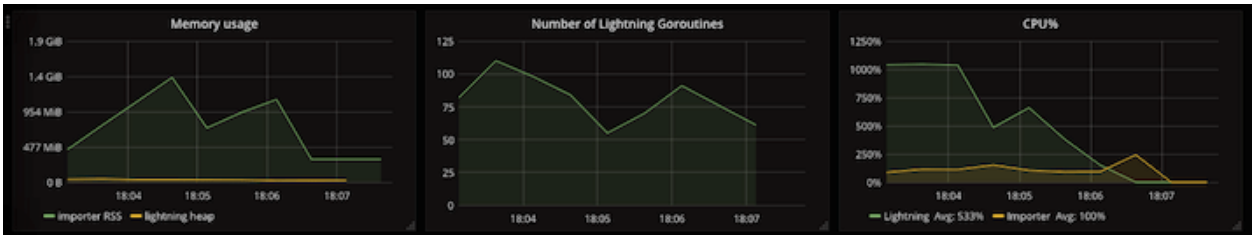


图 270: 第三行资源使用面板

面板名称	描述
Memory usage	每个服务占用的内存

面板名称	描述
Number of Lightning Goroutines	TiDB Lightning 使用的运行中的 goroutines 数量
CPU%	每个服务使用的逻辑 CPU 数量

12.7.7.2.4 第四行：配额使用面板

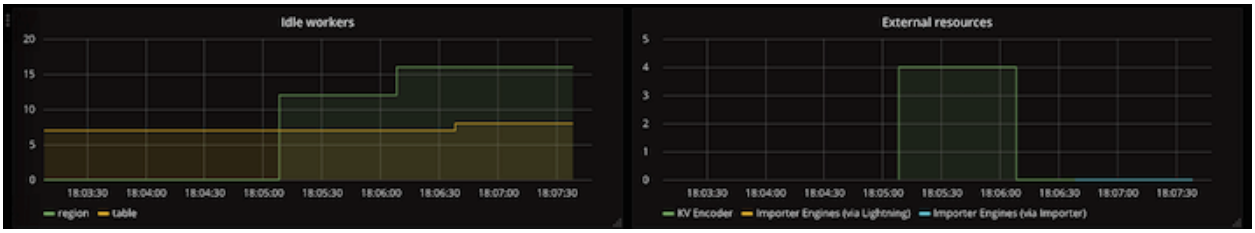


图 271: 第四行配额使用面板

面板名称	序列	描述
Idle workers	io	未使用的 `io-concurrency` 的数量，通常接近配置值（默认为 5），接近 0 时表示磁盘运行太慢

面板名称	序列	描述
Idle workers	closed-engine	已关闭但未清理的引擎数量，通常接近 `index-concurrency` 与 `table-concurrency` 的和（默认为 8），接近 0 时表示 TiDB Lightning 比 TiKV Importer 快，导致 TiDB Lightning 延迟
Idle workers	table	未使用的 `table-concurrency` 的数量，通常为 0，直到进程结束
Idle workers	index	未使用的 `index-concurrency` 的数量，通常为 0，直到进程结束

面板名称	序列	描述
Idle workers	region	未使用的 region-concurrency 的数量，通常为 0，直到进程结束
External resources	KV Encoder	已激活的 KV encoder 的数量，通常与 region-concurrency 的数量相同，直到进程结束
External resources	Importer Engines	打开的引擎文件数量，不应超过 max-open-engines 的设置

12.7.7.2.5 第五行：读取速度面板

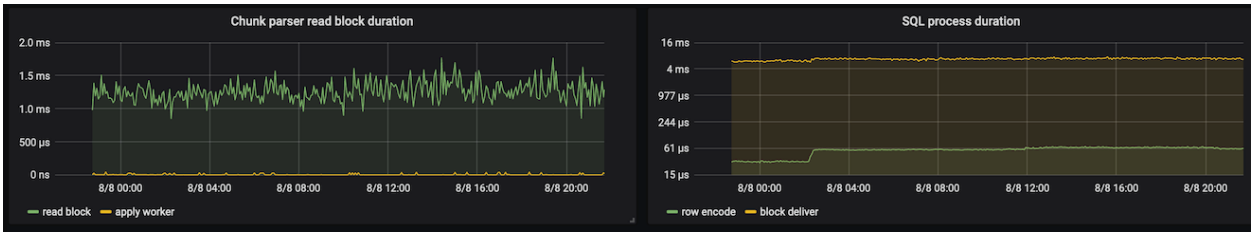


图 272: 第五行读取速度面板

面板名称	序列	描述
Chunk parser read block duration	read block	读取一个字节块来准备解析时所消耗的时间
Chunk parser read block duration	apply worker	等待 `io-concurrency` 空闲所消耗的时间
SQL process duration	row encode	解析和编码单行所消耗的时间
SQL process duration	block deliver	将一组键值对发送到 TiKV Importer 所消耗的时间

如果上述项的持续时间过长，则表示 TiDB Lightning 使用的磁盘运行太慢或 I/O 太忙。

12.7.7.2.6 第六行：存储空间面板

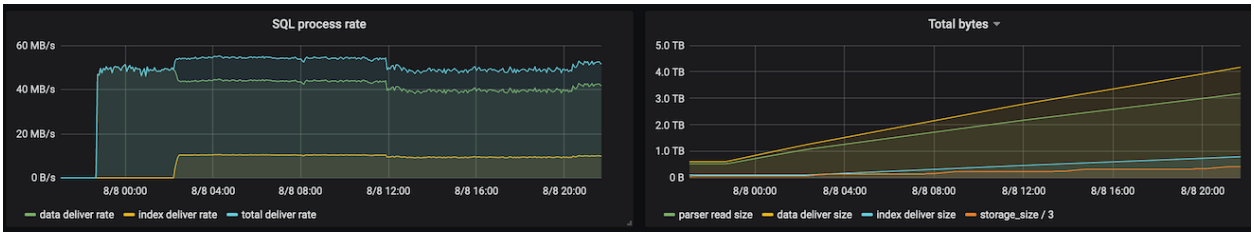


图 273: 第六行存储空间面板

面板名称	序列	描述
SQL process rate	data deliver rate	向 TiKV Importer 发送数据键值对的速度
SQL process rate	index deliver rate	向 TiKV Importer 发送索引键值对的速度
SQL process rate	total deliver rate	发送数据键值对及索引键值对的速度之和
Total bytes	parser read size	TiDB Lightning 正在读取的字节数
Total bytes	data deliver size	已发送到 TiKV Importer 的数据键值对的字节数
Total bytes	index deliver size	已发送到 TiKV Importer 的索引键值对的字节数
Total bytes	storage_size/3	TiKV 集群占用的存储空间大小的 1/3（3 为默认的副本数量）

12.7.7.2.7 第七行：导入速度面板



图 274: 第七行导入速度面板

面板名称	序列	描述
Delivery duration	Range delivery	将一个 range 的键值对上传到 TiKV 集群所消耗的时间
Delivery duration	SST delivery	将单个 SST 文件上传到 TiKV 集群所消耗的时间
SST process duration	Split SST	将键值对流切分成若干 SST 文件所消耗的时间
SST process duration	SST upload	上传单个 SST 文件所消耗的时间
SST process duration	SST ingest	ingest 单个 SST 文件所消耗的时间
SST process duration	SST size	单个 SST 文件的大小

12.7.7.3 监控指标

本节将详细描述 `tikv-importer` 和 `tidb-lightning` 的监控指标。

12.7.7.3.1 `tikv-importer`

`tikv-importer` 的监控指标皆以 `tikv_import_*` 为前缀。

- `tikv_import_rpc_duration`（直方图）

  完成一次 RPC 用时直方图。标签：

  - request：所执行 RPC 请求的类型
    - `switch_mode` — 将一个 TiKV 节点切换为 import/normal 模式
    - `open_engine` — 打开引擎文件
    - `write_engine` — 接收数据并写入引擎文件
    - `close_engine` — 关闭一个引擎文件
    - `import_engine` — 导入一个引擎文件到 TiKV 集群中
    - `cleanup_engine` — 删除一个引擎文件
    - `compact_cluster` — 显式压缩 TiKV 集群
    - `upload` — 上传一个 SST 文件
    - `ingest` — Ingest 一个 SST 文件
    - `compact` — 显式压缩一个 TiKV 节点
  - result：RPC 请求的执行结果
    - `ok`
    - `error`

- `tikv_import_write_chunk_bytes`（直方图）

  从 TiDB Lightning 接收的键值对区块大小（未压缩）的直方图。

- `tikv_import_write_chunk_duration`（直方图）

  从 `tidb-lightning` 接收每个键值对区块所需时间的直方图。

- `tikv_import_upload_chunk_bytes`（直方图）

  上传到 TiKV 的每个 SST 文件区块大小（压缩）的直方图。

- `tikv_import_range_delivery_duration`（直方图）

  将一个 range 的键值对发送至 `dispatch-job` 任务所需时间的直方图。

- `tikv_import_split_sst_duration`（直方图）

  将 range 从引擎文件中分离到单个 SST 文件中所需时间的直方图。

- `tikv_import_sst_delivery_duration`（直方图）

  将 SST 文件从 `dispatch-job` 任务发送到 `ImportSSTJob` 任务所需时间的直方图

- `tikv_import_sst_recv_duration`（直方图）

  `ImportSSTJob` 任务接收从 `dispatch-job` 任务发送过来的 SST 文件所需时间的直方图。

- `tikv_import_sst_upload_duration`（直方图）

  从 `ImportSSTJob` 任务上传 SST 文件到 TiKV 节点所需时间的直方图。

- `tikv_import_sst_chunk_bytes`（直方图）

  上传到 TiKV 节点的 SST 文件（压缩）大小的直方图。

- `tikv_import_sst_ingest_duration`（直方图）

  将 SST 文件传入至 TiKV 所需时间的直方图。

- `tikv_import_each_phase`（测量仪）

  表示运行阶段。值为 1 时表示在阶段内运行，值为 0 时表示在阶段内运行。标签：

  - phase：prepare / import

- `tikv_import_wait_store_available_count`（计数器）

  计算出现 TiKV 节点没有充足空间上传 SST 文件现象的次数。标签：

  - store_id：TiKV 存储 ID。

- `tikv_import_upload_chunk_duration`（直方图）

  上传到 TiKV 的每个区块所需时间的直方图。

12.7.7.3.2 tidb-lightning

`tidb-lightning` 的监控指标皆以 `lightning_*` 为前缀。

- `lightning_importer_engine`（计数器）

  计算已开启及关闭的引擎文件数量。标签：

  - type:
    * open
    * closed

- `lightning_idle_workers`（计量表盘）

  计算闲置的 worker。标签：

  - name：
    * `table` — 未使用的 `table-concurrency` 的数量，通常为 0，直到进程结束

    * index — 未使用的 index-concurrency 的数量，通常为 0，直到进程结束
    * region — 未使用的 region-concurrency 的数量，通常为 0，直到进程结束
    * io — 未使用的 io-concurrency 的数量，通常接近配置值（默认为 5），接近 0 时表示磁盘运行太慢
    * closed-engine — 已关闭但未清理的引擎数量，通常接近 index-concurrency 与 table-concurrency 的和（默认为 8），接近 0 时表示 TiDB Lightning 比 TiKV Importer 快，导致 TiDB Lightning 延迟

- lightning_kv_encoder（计数器）

  计算已开启及关闭的 KV 编码器。KV 编码器是运行于内存的 TiDB 实例，用于将 SQL 的 INSERT 语句转换成键值对。此度量的净值（开启减掉关闭）在正常情况下不应持续增长。标签：

  - type:
    * open
    * closed

- lightning_tables（计数器）

  计算处理过的表及其状态。标签：

  - state：表的状态，表明当前应执行的操作
    * pending — 等待处理
    * written — 所有数据已编码和传输
    * closed — 所有对应的引擎文件已关闭
    * imported — 所有引擎文件已上传到目标集群
    * altered_auto_inc — 自增 ID 已改
    * checksum — 已计算校验和
    * analyzed — 已进行统计信息分析
    * completed — 表格已完全导入并通过验证
  - result：当前操作的执行结果
    * success — 成功
    * failure — 失败（未完成）

- lightning_engines（计数器）

  计算处理后引擎文件的数量以及其状态。标签：

  - state：引擎文件的状态，表明当前应执行的操作
    * pending — 等待处理
    * written — 所有数据已编码和传输
    * closed — 引擎文件已关闭
    * imported — 当前引擎文件已上传到目标集群
    * completed — 当前引擎文件已完全导入
  - result：当前操作的执行结果
    * success — 成功
    * failure — 失败（未完成）

- `lightning_chunks`（计数器）

  计算处理过的 Chunks 及其状态。标签：

  - state: 单个 Chunk 的状态，表明该 Chunk 当前所处的阶段
    * `estimated` — （非状态）当前任务中 Chunk 的数量
    * `pending` — 已载入但未执行
    * `running` — 正在编码和发送数据
    * `finished` — 该 Chunk 已处理完毕
    * `failed` — 处理过程中发生错误

- `lightning_import_seconds`（直方图）

  导入每个表所需时间的直方图。

- `lightning_row_read_bytes`（直方图）

  单行 SQL 数据大小的直方图。

- `lightning_row_encode_seconds`（直方图）

  解码单行 SQL 数据到键值对所需时间的直方图。

- `lightning_row_kv_deliver_seconds`（直方图）

  发送一组与单行 SQL 数据对应的键值对所需时间的直方图。

- `lightning_block_deliver_seconds`（直方图）

  每个键值对中的区块传送到 `tikv-importer` 所需时间的直方图。

- `lightning_block_deliver_bytes`（直方图）

  发送到 Importer 的键值对中区块（未压缩）的大小的直方图。

- `lightning_chunk_parser_read_block_seconds`（直方图）

  数据文件解析每个 SQL 区块所需时间的直方图。

- `lightning_checksum_seconds`（直方图）

  计算表中 Checksum 所需时间的直方图。

- `lightning_apply_worker_seconds`（直方图）

  获取闲置 worker 等待时间的直方图 (参见 `lightning_idle_workers` 计量表盘)。标签：

  - `name`:
    * `table`
    * `index`
    * `region`
    * `io`
    * `closed-engine`

12.7.8 TiDB Lightning 常见问题

本文列出了一些使用 TiDB Lightning 时可能会遇到的问题与解决办法。

12.7.8.1 TiDB Lightning 对 TiDB/TiKV/PD 的最低版本要求是多少？

TiDB Lightning 的版本应与集群相同。如果使用 Local-backend 模式，最低版本要求为 4.0.0。如果使用 Importer-backend 或 TiDB-backend 模式最低版本要求是 2.0.9，但建议使用最新的稳定版本 3.0。

12.7.8.2 TiDB Lightning 支持导入多个库吗？

支持。

12.7.8.3 TiDB Lightning 对下游数据库的账号权限要求是怎样的？

详细权限描述参考TiDB Lightning 使用前提。

12.7.8.4 TiDB Lightning 在导数据过程中某个表报错了，会影响其他表吗？进程会马上退出吗？

如果只是个别表报错，不会影响整体。报错的那个表会停止处理，继续处理其他的表。

12.7.8.5 如何正确重启 TiDB Lightning？

如果使用 Importer-backend，根据 `tikv-importer` 的状态，重启 TiDB Lightning 的基本顺序如下：

如果 `tikv-importer` 仍在运行：

1. 结束 `tidb-lightning` 进程。
2. 执行修改操作（如修复数据源、更改设置、更换硬件等）。
3. 如果上面的修改操作更改了任何表，你还需要清除对应的断点。
4. 重启 `tidb-lightning`。

如果 `tikv-importer` 需要重启：

1. 结束 `tidb-lightning` 进程。
2. 结束 `tikv-importer` 进程。
3. 执行修改操作（如修复数据源、更改设置、更换硬件等）。
4. 重启 `tikv-importer`。
5. 重启 `tidb-lightning` 并等待，直到程序因校验和错误（如果有的话）而失败。
    - 重启 `tikv-importer` 将清除所有仍在写入的引擎文件，但是 `tidb-lightning` 并不会感知到该操作。从 v3.0 开始，最简单的方法是让 `tidb-lightning` 继续，然后再重试。
6. 清除失败的表及断点。
7. 再次重启 `tidb-lightning`。

如果使用 Local-backend 和 TiDB-backend，操作和 Importer-backend 的 `tikv-importer` 仍在运行时相同。

12.7.8.6 如何校验导入的数据的正确性？

TiDB Lightning 默认会对导入数据计算校验和 (checksum)，如果校验和不一致就会停止导入该表。可以在日志看到相关的信息。

TiDB 也支持从 MySQL 命令行运行 `ADMIN CHECKSUM TABLE` 指令来计算校验和。

```
ADMIN CHECKSUM TABLE `schema`.`table`;
```

```
+---------+------------+--------------------+-----------+-------------+
| Db_name | Table_name | Checksum_crc64_xor | Total_kvs | Total_bytes |
+---------+------------+--------------------+-----------+-------------+
| schema  | table      | 5505282386844578743 |         3 |          96 |
+---------+------------+--------------------+-----------+-------------+
1 row in set (0.01 sec)
```

12.7.8.7 TiDB Lightning 支持哪些格式的数据源？

目前，TiDB Lightning 支持：

- 导入Dumpling、CSV 或Amazon Aurora Parquet 输出格式的数据源。
- 从本地盘或Amazon S3 云盘读取数据。

12.7.8.8 我已经在下游创建好库和表了，TiDB Lightning 可以忽略建库建表操作吗？

可以。在配置文档中的 `[mydumper]` 部分将 `no-schema` 设置为 `true` 即可。`no-schema=true` 会默认下游已经创建好所需的数据库和表，如果没有创建，会报错。

12.7.8.9 有些不合法的数据，能否通过关掉严格 SQL 模式 (Strict SQL Mode) 来导入？

可以。TiDB Lightning 默认的 `sql_mode` 为 `"STRICT_TRANS_TABLES,NO_ENGINE_SUBSTITUTION"`。

这个设置不允许一些非法的数值，例如 `1970-00-00` 这样的日期。可以修改配置文件 `[tidb]` 下的 `sql-mode` 值。

```
...
[tidb]
sql-mode = ""
...
```

12.7.8.10 可以启用一个 `tikv-importer`，同时有多个 `tidb-lightning` 进程导入数据吗？

只要每个 TiDB Lightning 操作的表互不相同就可以。

12.7.8.11 如何正确结束 `tikv-importer` 进程？

手动部署：如果 `tikv-importer` 正在前台运行，可直接按 Ctrl+C 退出。否则，可通过 `ps aux | grep tikv-importer` 获取进程 ID，然后通过 `kill «pid»` 结束进程。

12.7.8.12 如何正确结束 tidb-lightning 进程？

根据部署方式，选择相应操作结束进程

手动部署：如果 tidb-lightning 正在前台运行，可直接按 Ctrl+C 退出。否则，可通过 ps aux | grep tidb-lightning 获取进程 ID，然后通过 kill -2 «pid» 结束进程。

12.7.8.13 tidb-lightning 在服务器上运行，进程莫名其妙地退出了，是怎么回事呢？

这种情况可能是启动方式不正确，导致收到 SIGHUP 信号而退出。此时 tidb-lightning.log 通常有如下日志：

```
[2018/08/10 07:29:08.310 +08:00] [INFO] [main.go:41] ["got signal to exit"] [signal=hangup]
```

不推荐在命令行中直接使用 nohup 启动进程，推荐使用脚本启动 tidb-lightning。

另外，如果从 TiDB Lightning 的 log 的最后一条日志显示遇到的错误是 “Context canceled”，需要在日志中搜索第一条 “ERROR” 级别的日志。在这条日志之前，通常也会紧跟有一条 “got signal to exit”，表示 Lighting 是收到中断信号然后退出的。

12.7.8.14 为什么用过 TiDB Lightning 之后，TiDB 集群变得又慢又耗 CPU？

如果 tidb-lightning 异常退出，集群可能仍处于 “导入模式” (import mode)，该模式不适用于生产环境。此时可执行以下命令查看当前使用的模式：

```
tidb-lightning-ctl --config tidb-lightning.toml --fetch-mode
```

可执行以下命令强制切换回 “普通模式” (normal mode)：

```
tidb-lightning-ctl --config tidb-lightning.toml --switch-mode=normal
```

12.7.8.15 TiDB Lightning 可以使用千兆网卡吗？

使用 TiDB Lightning 建议配置万兆网卡。不推荐使用千兆网卡，尤其是在部署 tikv-importer 的机器上。

千兆网卡的总带宽只有 120 MB/s，而且需要与整个 TiKV 集群共享。在使用 TiDB Lightning 导入时，极易用尽所有带宽，继而因 PD 无法联络集群使集群断连。为了避免这种情况，你可以在tikv-importer 的配置文件中限制上传速度。

```
[import]
### Importer 上传至 TiKV 的最大速度（字节/秒）。
### 建议将该速度设为 100 MB/s 或更小。
upload-speed-limit = "100MB"
```

12.7.8.16 为什么 TiDB Lightning 需要在 TiKV 集群预留这么多空间？

当使用默认的 3 副本设置时，TiDB Lightning 需要 TiKV 集群预留数据源大小 6 倍的空间。多出来的 2 倍是算上下列没储存在数据源的因素的保守估计：

- 索引会占据额外的空间
- RocksDB 的空间放大效应

12.7.8.17 TiDB Lightning 使用过程中是否可以重启 TiKV Importer？

不能，TiKV Importer 会在内存中存储一些引擎文件，重启后，`tidb-lightning` 会因连接失败而停止。此时，你需要清除失败的断点，因为这些 TiKV Importer 特有的信息丢失了。你可以在之后重启 TiDB Lightning。

12.7.8.18 如何清除所有与 TiDB Lightning 相关的中间数据？

1. 删除断点文件。

```
tidb-lightning-ctl --config conf/tidb-lightning.toml --checkpoint-remove=all
```

   如果出于某些原因而无法运行该命令，你可以尝试手动删除 `/tmp/tidb_lightning_checkpoint.pb` 文件。

2. 如果使用 Local-backend，删除配置中 `sorted-kv-dir` 对应的目录；如果使用 Importer-backend，删除 `tikv-importer` 所在机器上的整个 `import` 文件目录。

3. 如果需要的话，删除 TiDB 集群上创建的所有表和库。

4. 清理残留的元信息。如果存在以下任意一种情况，需要手动清理元信息库：

   - 对于 v5.1.x 和 v5.2.x 版本的 TiDB Lightning, tidb-lightning-ctl 命令没有同时清理存储在目标集群的 metadata 库，需要手动清理。
   - 如果手动删除过断点文件，则需要手动清理下游的元信息库，否则可能影响后续导入的正确性。

   使用下面命令清理元信息：

```
DROP DATABASE IF EXISTS `lightning_metadata`;
```

12.7.8.19 TiDB Lightning 报错 `could not find first pair, this shouldn't happen`

报错原因是遍历本地排序的文件时出现异常，可能在 TiDB Lightning 打开的文件数量超过系统的上限时发生报错。在 Linux 系统中，可以使用 `ulimit -n` 命令确认此值是否过小。建议在导入期间将此设置调整为 1000000（即 `ulimit -n 1000000`）。

12.7.8.20 TiDB Lightning 导入速度太慢

TiDB Lightning 的正常速度为每条线程每 2 分钟导入一个 256 MB 的数据文件，如果速度远慢于这个数值就是有问题。导入的速度可以检查日志提及 `restore chunk ... takes` 的记录，或者观察 Grafana 的监控信息。

导入速度太慢一般有几个原因：

原因 1：`region-concurrency` 设定太高，线程间争用资源反而减低了效率。

1. 从日志的开头搜寻 `region-concurrency` 能知道 TiDB Lightning 读到的参数是多少。
2. 如果 TiDB Lightning 与其他服务（如 TiKV Importer）共用一台服务器，必需手动将 `region-concurrency` 设为该服务器 CPU 数量的 75%。
3. 如果 CPU 设有限额（例如从 Kubernetes 指定的上限），TiDB Lightning 可能无法自动判断出来，此时亦需要手动调整 `region-concurrency`。

原因 2：表结构太复杂。

每条索引都会额外增加键值对。如果有 N 条索引，实际导入的大小就差不多是 Dumpling 文件的 N+1 倍。如果索引不太重要，可以考虑先从 schema 去掉，待导入完成后再使用 CREATE INDEX 加回去。

原因 3: 单个文件过大。

把源数据分割为单个大小约为 256 MB 的多个文件时，TiDB Lightning 会并行处理数据，达到最佳效果。如果导入的单个文件过大，TiDB Lightning 可能无响应。

如果源数据是 CSV 格式文件，并且所有的 CSV 文件内都不存在包含字符换行符的字段 (U+000A 及 U+000D)，则可以启用 `strict-format`，TiDB Lightning 会自动分割大文件。

```
[mydumper]
strict-format = true
```

原因 4：TiDB Lightning 版本太旧。

试试最新的版本吧！可能会有改善。

12.7.8.21 checksum failed: checksum mismatched remote vs local

原因：本地数据源跟目标数据库某个表的校验和不一致。这通常有更深层的原因，可以通过检查日志中包含 `checksum mismatched` 的行进一步定位。

包含 `checksum mismatched` 的行中有 `total_kvs: x vs y` 的信息，x 表示导入集群在完成导入后计算出的键值对（KV pairs）数目，y 表示本地数据源产生的键值对数目。

- x 大，即导入集群键值对更多：
  - 可能这张表在导入前已有数据，因此影响了数据校验。这也包括 TiDB Lightning 之前失败停机过，但没有正确重启。
- y 大，即本地数据源键值对更多：
  - 如果目标数据库的校验和全是 0，表示没有发生任何导入，有可能是集群太忙无法接收任何数据。
  - 可能导出数据中包含重复数据，例如唯一键和主键 (UNIQUE and PRIMARY KEYs) 有重复的值、下游表结构为大小写不敏感而数据为大小写敏感。
- 其他情况
  - 如果数据源是由机器生成而不是从 Dumping 备份的，需确保数据符合表的限制，例如自增 (AUTO_INCREMENT) 的列需要为正数，不能为 0。

解决办法：

1. 使用 `tidb-lightning-ctl` 把出错的表删除，检查表结构与数据，重启 TiDB Lightning 重新导入此前出错的表。

   ```
   tidb-lightning-ctl --config conf/tidb-lightning.toml --checkpoint-error-destroy=all
   ```

2. 把断点存放在外部数据库（修改 `[checkpoint] dsn`），减轻目标集群压力。
3. 参考如何正确重启 TiDB Lightning中的解决办法。

12.7.8.22 Checkpoint for … has invalid status: (错误码)

原因：断点续传已启用。TiDB Lightning 或 TiKV Importer 之前发生了异常退出。为了防止数据意外损坏，TiDB Lightning 在错误解决以前不会启动。

错误码是小于 25 的整数，可能的取值是 0、3、6、9、12、14、15、17、18、20、21。整数越大，表示异常退出所发生的步骤在导入流程中越晚。

解决办法：

如果错误原因是非法数据源，使用 `tidb-lightning-ctl` 删除已导入数据，并重启 TiDB Lightning。

```
tidb-lightning-ctl --config conf/tidb-lightning.toml --checkpoint-error-destroy=all
```

其他解决方法请参考断点续传的控制。

12.7.8.23 ResourceTemporarilyUnavailable("Too many open engines …: …")

原因：并行打开的引擎文件 (engine files) 超出 `tikv-importer` 里的限制。这可能由配置错误引起。即使配置没问题，如果 `tidb-lightning` 曾经异常退出，也有可能令引擎文件残留在打开的状态，占据可用的数量。

解决办法：

1. 提高 `tikv-importer.toml` 内 `max-open-engines` 的值。这个设置主要由内存决定，计算公式为：

   最大内存使用量 ≈ `max-open-engines` × `write-buffer-size` × `max-write-buffer-number`

2. 降低 `table-concurrency` + `index-concurrency`，使之低于 `max-open-engines`。

3. 重启 `tikv-importer` 来强制移除所有引擎文件 (默认值为 `./data.import/`)。这样也会丢弃导入了一半的表，所以启动 TiDB Lightning 前必须清除过期的断点记录：

   ```
   tidb-lightning-ctl --config conf/tidb-lightning.toml --checkpoint-error-destroy=all
   ```

12.7.8.24 cannot guess encoding for input file, please convert to UTF-8 manually

原因：TiDB Lightning 只支持 UTF-8 和 GB-18030 编码的表架构。此错误代表数据源不是这里任一个编码。也有可能是文件中混合了不同的编码，例如，因为在不同的环境运行过 `ALTER TABLE`，使表架构同时出现 UTF-8 和 GB-18030 的字符。

解决办法：

1. 编辑数据源，保存为纯 UTF-8 或 GB-18030 的文件。
2. 手动在目标数量库创建所有的表，然后设置 `[mydumper] no-schema = true` 跳过创建表的步骤。
3. 设置 `[mydumper] character-set = "binary"` 跳过这个检查。但是这样可能使数据库出现乱码。

12.7.8.25 [sql2kv] sql encode error = [types:1292]invalid time format: '{1970 1 1 …}'

原因: 一个 `timestamp` 类型的时间戳记录了不存在的时间值。时间值不存在是由于夏时制切换或超出支持的范围 (1970 年 1 月 1 日至 2038 年 1 月 19 日)。

解决办法:

1. 确保 TiDB Lightning 与数据源时区一致。

    - 手动部署的话，通过设定 `$TZ` 环境变量强制时区设定。

      强制使用 Asia/Shanghai 时区：

      ```
      TZ='Asia/Shanghai' bin/tidb-lightning -config tidb-lightning.toml
      ```

2. 导出数据时，必须加上 `--skip-tz-utc` 选项。

3. 确保整个集群使用的是同一最新版本的 `tzdata` (2018i 或更高版本)。

    如果你使用的是 CentOS 机器，你可以运行 `yum info tzdata` 命令查看 `tzdata` 的版本及是否有更新。然后运行 `yum upgrade tzdata` 命令升级 `tzdata`。

12.7.8.26 `[Error 8025: entry too large, the max entry size is 6291456]`

原因： TiDB Lightning 生成的单行 KV 超过了 TiDB 的限制。

解决办法:

目前无法绕过 TiDB 的限制，只能忽略这张表，确保其它表顺利导入。

12.7.8.27 switch-mode 时遇到 `rpc error: code = Unimplemented ...`

原因： 集群中有不支持 switch-mode 的节点。目前已知的组件中，4.0.0-rc.2 之前的 TiFlash 不支持 switch-mode 操作。

解决办法：

- 如果集群中有 TiFlash 节点，可以将集群更新到 4.0.0-rc.2 或更新版本。
- 如果不方便升级，可以临时禁用 TiFlash。

12.7.8.28 `tidb lightning encountered error: TiDB version too old, expected '>=4.0.0', found '3.0.18'`

TiDB Lightning Local-backend 只支持导入到 v4.0.0 及以上版本的 TiDB 集群。如果尝试使用 Local-backend 导入到 v2.x 或 v3.x 的集群，就会报以上错误。此时可以修改配置使用 Importer-backend 或 TiDB-backend 进行导入。

部分 `nightly` 版本的 TiDB 集群的版本可能类似 4.0.0-beta.2。这种版本的 TiDB Lightning 实际支持 Local-backend，如果使用 `nightly` 版本遇到该报错，可以通过设置配置 `check-requirements = false` 跳过版本检查。在设置此参数之前，请确保 TiDB Lightning 的配置支持对应的版本，否则无法保证导入成功。

12.7.8.29 `restore table test.district failed: unknown columns in header [...]`

出现该错误通常是因为 CSV 格式的数据文件不包含 header（第一行也是数据），因此需要在 TiDB Lightning 的配置文件中增加如下配置项：

```
[mydumper.csv]
header = false
```

12.7.8.30 如何获取 TiDB Lightning 运行时的 goroutine 信息

1. 如果 TiDB Lightning 的配置文件中已经指定了`status-port`，可以跳过此步骤。否则，需要向 TiDB Lightning 发送 USR1 信号以开启 `status-port`。

   首先通过 `ps` 等命令获取 TiDB Lightning 的进程 PID，然后运行如下命令：

   ```
   kill -USR1 <lightning-pid>
   ```

   查看 TiDB Lightning 的日志，其中 `starting HTTP server` / `start HTTP server` / `started HTTP server` 的日志会显示新开启的 `status-port`。

2. 访问 `http://<lightning-ip>:<status-port>/debug/pprof/goroutine?debug=2` 可获取 goroutine 信息。

12.7.9 TiDB Lightning 术语表

本术语表提供了 TiDB Lightning 相关的术语和定义，这些术语会出现在 TiDB Lightning 的日志、监控指标、配置和文档中。

12.7.9.1 A

12.7.9.1.1 Analyze

统计信息分析。指重建 TiDB 表中的统计信息，即运行 `ANALYZE TABLE` 语句。

因为 TiDB Lightning 不通过 TiDB 导入数据，统计信息不会自动更新，所以 TiDB Lightning 在导入后显式地分析每个表。如果不需要该操作，可以将 `post-restore.analyze` 设置为 `false`。

12.7.9.1.2 AUTO_INCREMENT_ID

用于为自增列分配默认值的自增 ID 计数器。每张表都有一个相关联的 `AUTO_INCREMENT_ID` 计数器。在 TiDB 中，该计数器还用于分配行 ID。

因为 TiDB Lightning 不通过 TiDB 导入数据，`AUTO_INCREMENT_ID` 计数器不会自动更新，所以 TiDB Lightning 显式地将 `AUTO_INCREMENT_ID` 改为一个有效值。即使表中没有自增列，这步仍是会执行。

12.7.9.2 B

12.7.9.2.1 Backend

也称作 Back end（后端），用于接受 TiDB Lightning 解析结果。

详情参阅TiDB Lightning Backends。

12.7.9.3 C

12.7.9.3.1 Checkpoint

断点。用于保证 TiDB Lightning 在导入数据时不断地将进度保存到本地文件或远程数据库中。这样即使进程崩溃，TiDB Lightning 也能从中间状态恢复。

详情参见TiDB Lightning 断点续传。

12.7.9.3.2 Checksum

校验和。一种用于验证导入数据正确性的方法。

在 TiDB Lightning 中，表的校验和是由 3 个数字组成的集合，由该表中每个键值对的内容计算得出。这些数字分别是：

- 键值对的数量
- 所有键值对的总长度
- 每个键值对 CRC-64-ECMA 按位异或的结果

TiDB Lightning 通过比较每个表的本地校验和和远程校验和来验证导入数据的正确性。如果有任一对校验和不匹配，导入进程就会停止。如果你需要跳过校验和检查，可以将 `post-restore.checksum` 设置为 `false`。

遇到校验和不匹配的问题时，参考常见问题进行处理。

12.7.9.3.3 Chunk

一段连续的源数据，通常相当于数据源中的单个文件。

如果单个文件太大，TiDB Lightning 可以将单个文件拆分成多个文件块。

12.7.9.3.4 Compaction

压缩。指将多个小 SST 文件合并为一个大 SST 文件并清理已删除的条目。TiDB Lightning 导入数据时，TiKV 在后台会自动压缩数据。

> 注意：
>
> 出于遗留原因，你仍然可以将 TiDB Lightning 配置为在每次导入表时进行显式压缩，但是官方不推荐采用该操作，且该操作的相关设置默认是禁用的。
>
> 技术细节参阅 RocksDB 关于压缩的说明。

12.7.9.4 D

12.7.9.4.1 Data engine

数据引擎。用于对实际的行数据进行排序的引擎。

当一个表数据很多的时候，表的数据会被放置在多个数据引擎中以改善任务流水线并节省 TiKV Importer 的空间。默认条件下，每 100 GB 的 SQL 数据会打开一个新的数据引擎（可通过 `mydumper.batch-size` 配置项进行更改）。

TiDB Lightning 同时处理多个数据引擎（可通过 `lightning.table-concurrency` 配置项进行更改）。

12.7.9.5 E

12.7.9.5.1 Engine

引擎。在 TiKV Importer 中，一个引擎就是一个用于排序键值对的 RocksDB 实例。

TiDB Lightning 通过引擎将数据传送到 TiKV Importer 中。TiDB Lightning 先打开一个引擎，向其发送未排序的键值对，然后关闭引擎。随后，引擎会对收到的键值对进行排序操作。这些关闭的引擎可以进一步上传至 TiKV store 中为Ingest 做准备。

引擎使用 TiKV Importer 的 `import-dir` 作为临时存储，有时也会被称为引擎文件 (engine files)。

另见数据引擎和索引引擎。

12.7.9.6 F

12.7.9.6.1 Filter

配置列表，用于指定需要导入或不允许导入的表。

详情见表库过滤。

12.7.9.7 I

12.7.9.7.1 Import mode

导入模式。指通过降低读取速度和减少空间使用，来优化 TiKV 写入的配置模式。

导入过程中，TiDB Lightning 自动在导入模式和普通模式中来回切换。如果 TiKV 卡在导入模式，你可以使用 `tidb-lightning-ctl` 强制切换回普通模式。

12.7.9.7.2 Index engine

索引引擎。用于对索引进行排序的引擎。

不管表中有多少索引，每张表都只对应一个索引引擎。

TiDB Lightning 可同时处理多个索引引擎（可通过 `lightning.index-concurrency` 配置项进行更改）。由于每张表正好对应一个索引引擎，`lightning.index-concurrency` 配置项也限定了可同时处理的表的最大数量。

12.7.9.7.3 Ingest

指将SST 文件的全部内容插入到 RocksDB（TiKV）store 中的操作。

与逐个插入键值对相比，Ingest 的效率非常高。因此，该操作直接决定了 TiDB Lightning 的性能。

技术细节参阅 RocksDB 关于创建、Ingest SST 文件的 wiki 页面。

12.7.9.8 K

12.7.9.8.1 KV pair

即 key-value pair（键值对）。

12.7.9.8.2 KV encoder

用于将 SQL 或 CSV 行解析为键值对的例程。多个 KV encoder 会并行运行以加快处理速度。

12.7.9.9 L

12.7.9.9.1 Local checksum

本地校验和。在将键值对发送到 TiKV Importer 前，由 TiDB Lightning 计算的表的校验和。

12.7.9.10 N

12.7.9.10.1 Normal mode

普通模式。未启用导入模式时的模式。

12.7.9.11 P

12.7.9.11.1 Post-processing

指整个数据源被解析发送到 TiKV Importer 之后的一段时间。此时 TiDB Lightning 正在等待 TiKV Importer 上传、Ingest SST 文件。

12.7.9.12 R

12.7.9.12.1 Remote checksum

远程校验和。指导入 TiDB 后所计算的表的校验和。

12.7.9.13 S

12.7.9.13.1 Scattering

指随机再分配Region 中 leader 和 peer 的操作。Scattering 确保导入的数据在 TiKV store 中均匀分布，这样可以降低 PD 调度的压力。

12.7.9.13.2 Splitting

指 TiKV Importer 在上传之前会将单个引擎文件拆分为若干小SST 文件的操作。这是因为引擎文件通常很大（约为 100 GB），在 TiKV 中不适合视为单一的Region。拆分的文件大小可通过 `import.region-split-size` 配置项更改。

12.7.9.13.3 SST file

Sorted string table file（排序字符串表文件）。SST 文件是一种在 RocksDB 中（因而也是 TiKV 中）键值对集合在本地的存储形式。

TiKV Importer 从关闭的引擎中生成 SST 文件。这些 SST 文件接着被上传、 ingest 到 TiKV store 中。

12.8 TiDB Data Migration

12.8.1 Data Migration 简介

TiDB Data Migration (DM) 是一款便捷的数据迁移工具，支持从与 MySQL 协议兼容的数据库（MySQL、 MariaDB、Aurora MySQL）到 TiDB 的全量数据迁移和增量数据同步。使用 DM 工具有利于简化数据迁移过程，降低数据迁移运维成本。

使用 DM 进行数据迁移的时候，需要执行以下操作：

- 部署 DM 集群
- 创建上游数据源（source）对象，保存数据源访问信息
- 创建（多个）数据迁移任务从数据源迁移数据到 TiDB

数据迁移任务包含全量数据迁移、增量数据复制两个阶段：

- 全量数据迁移：从数据源迁移对应表的表结构到 TiDB，然后读取存量数据写入到 TiDB 集群；
- 增量数据复制：全量数据迁移完成后，从数据源读取对应的表变更然后写入到 TiDB 集群。

要快速了解 DM 的原理架构、适用场景，建议先观看下面的培训视频（时长 22 分钟）。注意本视频只作为学习参考，具体操作步骤和最新功能，请以文档内容为准。

12.8.1.1 版本说明

本文档中的说明基于 DM 的最新稳定版本 v5.4。

在 v5.4 之前，DM 工具的文档独立于 TiDB 文档。要访问这些早期版本的 DM 文档，请点击以下链接：

- DM v5.3 文档

- DM v2.0 文档
- DM v1.0 文档（较旧的版本，不推荐使用）

> 注意：
>
> - DM 的 GitHub 代码仓库已于 2021 年 10 月迁移至 pingcap/tiflow。如有任何关于 DM 的问题，请在 pingcap/tiflow 仓库提交，以获得后续反馈。
> - 在较早版本中（v1.0 和 v2.0），DM 采用独立于 TiDB 的版本号。从 DM v5.3 起，DM 采用与 TiDB 相同的版本号。DM v2.0 的下一个版本为 DM v5.3。DM v2.0 到 v5.3 无兼容性变更，升级过程与正常升级无差异。

12.8.1.2 基本功能

本节介绍 DM 工具的核心功能模块。

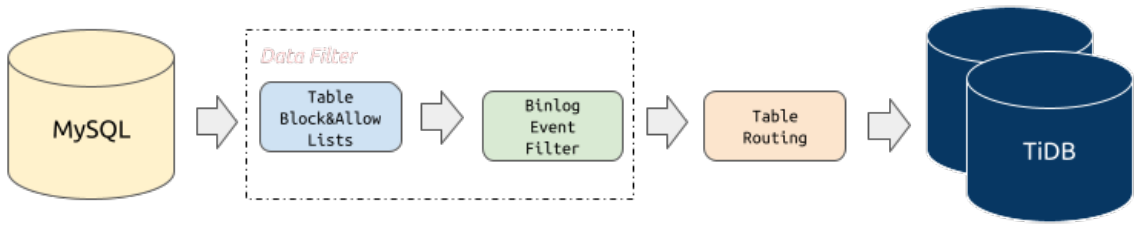


图 275: DM Core Features

12.8.1.2.1 Block & allow lists

Block & Allow Lists 的过滤规则类似于 MySQL `replication-rules-db/replication-rules-table`，用于过滤或指定只迁移某些数据库或某些表的所有操作。

12.8.1.2.2 Binlog event filter

Binlog Event Filter 用于过滤源数据库中特定表的特定类型操作，比如过滤掉表 `test.sbtest` 的 `INSERT` 操作或者过滤掉库 `test` 下所有表的 `TRUNCATE TABLE` 操作。

12.8.1.2.3 Table routing

Table Routing 是将源数据库的表迁移到下游指定表的路由功能，比如将源数据表 `test.sbtest1` 的表结构和数据迁移到 TiDB 的表 `test.sbtest2`。它也是分库分表合并迁移所需的一个核心功能。

12.8.1.3 高级功能

12.8.1.3.1 分库分表合并迁移

DM 支持对源数据的分库分表进行合并迁移，但有一些使用限制，详细信息请参考悲观模式分库分表合并迁移使用限制和乐观模式分库分表合并迁移使用限制。

12.8.1.3.2 对第三方 Online Schema Change 工具变更过程的同步优化

在 MySQL 生态中，gh-ost 与 pt-osc 等工具被广泛使用，DM 对其变更过程进行了特殊的优化，以避免对不必要的中间数据进行迁移。详细信息可参考online-ddl。

12.8.1.3.3 使用 SQL 表达式过滤某些行变更

在增量同步阶段，DM 支持配置 SQL 表达式过滤掉特定的行变更，以实现对同步数据的更精细控制。详细信息可参考使用 SQL 表达式过滤某些行变更。

12.8.1.4 使用限制

在使用 DM 工具之前，需了解以下限制：

- 数据库版本要求
  - MySQL 版本 > 5.5
  - MariaDB 版本 >= 10.1.2

> **注意：**
> 如果上游 MySQL/MariaDB servers 间构成主从复制结构，则需要 MySQL 版本高于 5.7.1 或者 MariaDB 版本等于或高于 10.1.3。

> **警告：**
> 使用 DM 从 MySQL v8.0 迁移数据到 TiDB 目前为实验特性（从 DM v2.0 引入），不建议在生产环境下使用。

- DDL 语法兼容性限制
  - 目前，TiDB 部分兼容 MySQL 支持的 DDL 语句。因为 DM 使用 TiDB parser 来解析处理 DDL 语句，所以目前仅支持 TiDB parser 支持的 DDL 语法。详见 TiDB DDL 语法支持。

  - DM 遇到不兼容的 DDL 语句时会报错。要解决此报错，需要使用 dmctl 手动处理，要么跳过该 DDL 语句，要么用指定的 DDL 语句来替换它。详见如何处理不兼容的 DDL 语句。

- 分库分表数据冲突合并

  - 如果业务分库分表之间存在数据冲突，可以参考自增主键冲突处理来解决；否则不推荐使用 DM 进行迁移，如果进行迁移则有冲突的数据会相互覆盖造成数据丢失。
  - 分库分表 DDL 同步限制，参见悲观模式下分库分表合并迁移使用限制以及乐观模式下分库分表合并迁移使用限制。

- 数据源 MySQL 实例切换

  - 当 DM-worker 通过虚拟 IP（VIP）连接到 MySQL 且要切换 VIP 指向的 MySQL 实例时，DM 内部不同的 connection 可能会同时连接到切换前后不同的 MySQL 实例，造成 DM 拉取的 binlog 与从上游获取到的其他状态不一致，从而导致难以预期的异常行为甚至数据损坏。如需切换 VIP 指向的 MySQL 实例，请参考虚拟 IP 环境下的上游主从切换对 DM 手动执行变更。

- GBK 字符集兼容性限制

  - DM 在 v5.4.0 之前不支持将 `charset=GBK` 的表迁移到 TiDB。

12.8.2 TiDB Data Migration 快速上手指南

本文介绍如何快速体验使用数据迁移工具 TiDB Data Migration (DM) 从 MySQL 迁移数据到 TiDB。

如需在生产环境中部署 DM，请参考以下文档：

- 使用 TiUP 部署 DM 集群
- 创建数据源
- 创建数据迁移任务

12.8.2.1 使用样例

在本地部署的 DM 集群组件和访问的 MySQL 和 TiDB 节点的信息如下：

实例	服务器地址	端口使用
DM-master	127.0.0.1	8261, 8291（内部端口）
DM-worker	127.0.0.1	8262
MySQL-3306	127.0.0.1	3306
TiDB	127.0.0.1	4000

12.8.2.2 使用 binary 包部署 DM

12.8.2.2.1 准备 DM binary 包

首先需要下载 DM 最新的 binary 或者手动编译。

第一种方式：下载最新 DM binary 包

```
wget http://download.pingcap.org/dm-nightly-linux-amd64.tar.gz
tar -xzvf dm-nightly-linux-amd64.tar.gz
cd dm-nightly-linux-amd64
```

第二种方式：编译最新 DM binary 包

```
git clone https://github.com/pingcap/dm.git
cd dm
make
```

12.8.2.2.2 部署 DM-master

执行如下命令启动 DM-master：

```
nohup bin/dm-master --master-addr='127.0.0.1:8261' --log-file=/tmp/dm-master.log --name="master1"
    ↪  >> /tmp/dm-master.log 2>&1 &
```

12.8.2.2.3 部署 DM-worker

执行如下命令启动 DM-worker：

```
nohup bin/dm-worker --worker-addr='127.0.0.1:8262' --log-file=/tmp/dm-worker.log --join
    ↪ ='127.0.0.1:8261' --name="worker1" >> /tmp/dm-worker.log 2>&1 &
```

12.8.2.2.4 检查 DM 集群部署是否正常

```
bin/dmctl --master-addr=127.0.0.1:8261 list-member
```

一个正常 DM 集群的范例返回结果如下所示：

```
{
    "result": true,
    "msg": "",
    "members": [
        {
            "leader": {
                "msg": "",
                "name": "master1",
                "addr": "127.0.0.1:8261"
            }
        },
        {
            "master": {
                "msg": "",
```

```
                "masters": [
                    {
                        "name": "master1",
                        "memberID": "11007177379717700053",
                        "alive": true,
                        "peerURLs": [
                            "http://127.0.0.1:8291"
                        ],
                        "clientURLs": [
                            "http://127.0.0.1:8261"
                        ]
                    }
                ]
            }
        },
        {
            "worker": {
                "msg": "",
                "workers": [
                    {
                        "name": "worker1",
                        "addr": "127.0.0.1:8262",
                        "stage": "free",
                        "source": ""
                    }
                ]
            }
        }
    ]
}
```

12.8.2.3 从 MySQL 同步数据到 TiDB

12.8.2.3.1 准备数据

使用 DM 之前，先准备好数据，向 MySQL-3306 写入示例数据。

```
drop database if exists `testdm`;
create database `testdm`;
use `testdm`;
create table t1 (id bigint, uid int, name varchar(80), info varchar(100), primary key (`id`),
    ↪ unique key(`uid`)) DEFAULT CHARSET=utf8mb4;
create table t2 (id bigint, uid int, name varchar(80), info varchar(100), primary key (`id`),
    ↪ unique key(`uid`)) DEFAULT CHARSET=utf8mb4;
```

```
insert into t1 (id, uid, name) values (1, 10001, 'Gabriel García Márquez'), (2, 10002, 'Cien años
    ↪  de soledad');
insert into t2 (id, uid, name) values (3, 20001, 'José Arcadio Buendía'), (4, 20002, 'Úrsula
    ↪ Iguarán'), (5, 20003, 'José Arcadio');
```

12.8.2.3.2 加载数据源 MySQL 配置

运行数据迁移任务前，需要加载数据源的配置，也就是 MySQL 的相关设置，到 DM。

对数据源 MySQL 访问密码进行加密

> 注意：
>
> - 如果数据源没有设置密码，则可以跳过该步骤。
> - 自 v2.0 起，DM 可以使用明文密码配置数据源的访问密码信息。

为了安全，可配置及使用加密后的 MySQL 访问密码，以密码为 “123456” 为例：

```
./bin/dmctl --encrypt "123456"
```

```
fCxfQ9XKCezSzuCD0Wf5dUD+LsKegSg=
```

记录该加密后的密码，用于下面新建 MySQL 数据源。

编写数据源 MySQL 配置

把以下配置文件内容写入到 `mysql-source-conf.yaml` 中。

MySQL1 的配置文件：

```
### MySQL Configuration.

source-id: "mysql-replica-01"

from:
  host: "127.0.0.1"
  user: "root"
  password: "fCxfQ9XKCezSzuCD0Wf5dUD+LsKegSg="
  port: 3306
```

加载数据源 MySQL 配置

在终端中执行下面的命令，使用 dmctl 将 MySQL 的数据源配置加载到 DM 集群中：

```
./bin/dmctl --master-addr=127.0.0.1:8261 operate-source create mysql-source-conf.yaml
```

结果如下：

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

这样就成功将 MySQL-3306 数据源添加到了 DM 集群。

12.8.2.3.3 创建数据迁移任务

在导入准备数据后，进行以下操作将 MySQL 的 `testdm.t1` 和 `testdm.t2` 两张表迁移到 TiDB。

1. 创建任务的配置文件 `testdm-task.yaml`：

```
---
name: testdm
task-mode: all

target-database:
  host: "127.0.0.1"
  port: 4000
  user: "root"
  password: "" # 如果密码不为空，则推荐使用经过 dmctl 加密的密文

mysql-instances:
  - source-id: "mysql-replica-01"
    block-allow-list:  "ba-rule1"

block-allow-list:
  ba-rule1:
    do-dbs: ["testdm"]
```

2. 使用 dmctl 创建任务：

```
./bin/dmctl --master-addr 127.0.0.1:8261 start-task testdm-task.yaml
```

结果如下：

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

这样就成功创建了一个将 MySQL-3306 数据迁移到 TiDB 的任务。

12.8.2.3.4 查看迁移任务状态

在创建迁移任务之后，可以用 `dmtcl query-status` 来查看任务的状态。

```
./bin/dmctl --master-addr 127.0.0.1:8261 query-status
```

结果如下：

```
{
    "result": true,
    "msg": "",
    "tasks": [
        {
            "taskName": "testdm",
            "taskStatus": "Running",
            "sources": [
                "mysql-replica-01"
            ]
        }
    ]
}
```

12.8.3 部署 DM 集群

12.8.3.1 DM 集群软硬件环境需求

DM 支持主流的 Linux 操作系统，具体版本要求见下表：

Linux 操作系统平台	版本
Red Hat Enterprise Linux	7.3 及以上

Linux 操作系统平台	版本
CentOS	7.3 及以上
Oracle Enterprise Linux	7.3 及以上
Ubuntu LTS	16.04 及以上

DM 可以在 Intel 架构服务器环境及主流虚拟化环境中部署和运行。

12.8.3.1.1 服务器建议配置

DM 支持部署和运行在 Intel x86-64 架构的 64 位通用硬件服务器平台。对于开发，测试，及生产环境的服务器硬件配置（不包含操作系统本身的占用）有以下要求和建议：

开发及测试环境

组件	CPU	内存	本地存储	网络	实例数量（最低要求）
DM-master	4 核 +	8 GB+	SAS，200 GB+	千兆网卡	1
DM-worker	8 核 +	16 GB+	SAS，200 GB+（大于迁移数据的大小）	千兆网卡	上游 MySQL 实例的数量

注意：

- 在功能验证的测试环境中的 DM-master 和 DM-worker 可以部署在同一台服务器上。
- 如进行性能相关的测试，避免采用低性能存储和网络硬件配置，防止对测试结果的正确性产生干扰。
- 如果仅验证功能，可以单机部署一个 DM-master，DM-worker 部署的数量至少是上游 MySQL 实例的数量。为了保证高可用性，建议部署更多的 DM-worker。
- DM-worker 在 dump 和 load 阶段需要存储全量数据，因此 DM-worker 的磁盘空间需要大于需要迁移数据的总量；如果迁移任务开启了 relay log，DM-worker 也需要一定的磁盘空间来存储上游的 binlog 数据。

生产环境

组件	CPU	内存	硬盘类型	网络	实例数量（最低要求）
DM-master	4 核 +	8 GB+	SAS，200 GB+	千兆网卡	3
DM-worker	16 核 +	32 GB+	SSD，200 GB+（大于迁移数据的大小）	万兆网卡	大于上游 MySQL 实例的数量
监控	8 核 +	16 GB+	SAS，200 GB+	千兆网卡	1

注意：

- 在生产环境中，不建议将 DM-master 和 DM-worker 部署和运行在同一个服务器上，以防

DM-worker 对磁盘的写入干扰 DM-master 高可用组件使用磁盘。

- 在遇到性能问题时可参照配置调优尝试修改任务配置。调优效果不明显时，可以尝试升级服务器配置。

12.8.3.2 使用 TiUP 部署 DM 集群

TiUP 是 TiDB 4.0 版本引入的集群运维工具，TiUP DM 是 TiUP 提供的使用 Golang 编写的集群管理组件，通过 TiUP DM 组件就可以进行日常的运维工作，包括部署、启动、关闭、销毁、扩缩容、升级 DM 集群以及管理 DM 集群参数。

目前 TiUP 可以支持部署 v2.0 及以上版本的 DM。本文将介绍不同集群拓扑的具体部署步骤。

注意：

如果部署机器的操作系统支持 SELinux，请确保 SELinux 处于关闭状态。

12.8.3.2.1 前提条件

当 DM 执行全量数据复制任务时，每个 DM-worker 只绑定一个上游数据库。DM-worker 首先在上游导出全部数据，然后将数据导入下游数据库。因此，DM-worker 的主机需要有足够的存储空间，具体存储路径在后续创建迁移任务时指定。

另外，部署 DM 集群需参照DM 集群软硬件环境需求，满足相应要求。

12.8.3.2.2 第 1 步：在中控机上安装 TiUP 组件

使用普通用户登录中控机，以 `tidb` 用户为例，后续安装 TiUP 及集群管理操作均通过该用户完成：

1. 执行如下命令安装 TiUP 工具：

```
curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
```

安装完成后，`~/.bashrc` 已将 TiUP 加入到路径中，你需要新开一个终端或重新声明全局变量 `source ~/.bashrc` 来使用 TiUP。

2. 安装 TiUP DM 组件：

```
tiup install dm dmctl
```

12.8.3.2.3 第 2 步：编辑初始化配置文件

请根据不同的集群拓扑，编辑 TiUP 所需的集群初始化配置文件。

请根据配置文件模板，新建一个配置文件 topology.yaml。如果有其他组合场景的需求，请根据多个模板自行调整。

可以使用 tiup dm template > topology.yaml 命令快速生成配置文件模板。

部署 3 个 DM-master、3 个 DM-worker 与 1 个监控组件的配置如下：

```
#全局变量适用于配置中的其他组件。如果组件实例中缺少一个特定值，则相应的全局变量将用作默认值。
global:
  user: "tidb"
  ssh_port: 22
  deploy_dir: "/dm-deploy"
  data_dir: "/dm-data"

server_configs:
  master:
    log-level: info
    # rpc-timeout: "30s"
    # rpc-rate-limit: 10.0
    # rpc-rate-burst: 40
  worker:
    log-level: info

master_servers:
  - host: 10.0.1.11
    name: master1
    ssh_port: 22
    port: 8261
    # peer_port: 8291
    # deploy_dir: "/dm-deploy/dm-master-8261"
    # data_dir: "/dm-data/dm-master-8261"
    # log_dir: "/dm-deploy/dm-master-8261/log"
    # numa_node: "0,1"
    # 下列配置项用于覆盖 `server_configs.master` 的值。
    config:
      log-level: info
      # rpc-timeout: "30s"
      # rpc-rate-limit: 10.0
      # rpc-rate-burst: 40
  - host: 10.0.1.18
    name: master2
    ssh_port: 22
    port: 8261
  - host: 10.0.1.19
```

```
    name: master3
    ssh_port: 22
    port: 8261
#### 如果不需要确保 DM 集群高可用，则可只部署 1 个 DM-master 节点，且部署的 DM-worker
    ↪ 节点数量不少于上游待迁移的 MySQL/MariaDB 实例数。
#### 如果需要确保 DM 集群高可用，则推荐部署 3 个 DM-master 节点，且部署的 DM-worker
    ↪ 节点数量大于上游待迁移的 MySQL/MariaDB 实例数（如 DM-worker 节点数量比上游实例数多 2 个
    ↪ ）。
worker_servers:
  - host: 10.0.1.12
    ssh_port: 22
    port: 8262
    # deploy_dir: "/dm-deploy/dm-worker-8262"
    # log_dir: "/dm-deploy/dm-worker-8262/log"
    # numa_node: "0,1"
    # 下列配置项用于覆盖 `server_configs.worker` 的值。
    config:
      log-level: info
  - host: 10.0.1.19
    ssh_port: 22
    port: 8262

monitoring_servers:
  - host: 10.0.1.13
    ssh_port: 22
    port: 9090
    # deploy_dir: "/tidb-deploy/prometheus-8249"
    # data_dir: "/tidb-data/prometheus-8249"
    # log_dir: "/tidb-deploy/prometheus-8249/log"

grafana_servers:
  - host: 10.0.1.14
    port: 3000
    # deploy_dir: /tidb-deploy/grafana-3000

alertmanager_servers:
  - host: 10.0.1.15
    ssh_port: 22
    web_port: 9093
    # cluster_port: 9094
    # deploy_dir: "/tidb-deploy/alertmanager-9093"
    # data_dir: "/tidb-data/alertmanager-9093"
    # log_dir: "/tidb-deploy/alertmanager-9093/log"
```

注意：

- 不建议在一台主机上运行太多 DM-worker。每个 DM-worker 至少应有 2 核 CPU 和 4 GiB 内存。
- 需要确保以下组件间端口可正常连通：
  - 各 DM-master 节点间的 peer_port（默认为 8291）可互相连通。
  - 各 DM-master 节点可连通所有 DM-worker 节点的 port（默认为 8262）。
  - 各 DM-worker 节点可连通所有 DM-master 节点的 port（默认为 8261）。
  - TiUP 节点可连通所有 DM-master 节点的 port（默认为 8261）。
  - TiUP 节点可连通所有 DM-worker 节点的 port（默认为 8262）。

更多 master_servers.host.config 参数说明，请参考 master parameter；更多 worker_servers.host.config 参数说明，请参考 worker parameter。

12.8.3.2.4 第 3 步：执行部署命令

注意：

通过 TiUP 进行集群部署可以使用密钥或者交互密码方式来进行安全认证：

- 如果是密钥方式，可以通过 -i 或者 --identity_file 来指定密钥的路径；
- 如果是密码方式，可以通过 -p 进入密码交互窗口；
- 如果已经配置免密登录目标机，则不需填写认证。

```
tiup dm deploy dm-test ${version} ./topology.yaml --user root [-p] [-i /home/root/.ssh/gcp_rsa]
```

以上部署命令中：

- 通过 TiUP DM 部署的集群名称为 dm-test。
- 部署版本为 ${version}，可以通过执行 tiup list dm-master 来查看 TiUP 支持的最新版本。
- 初始化配置文件为 topology.yaml。
- --user root：通过 root 用户登录到目标主机完成集群部署，该用户需要有 ssh 到目标机器的权限，并且在目标机器有 sudo 权限。也可以用其他有 ssh 和 sudo 权限的用户完成部署。
- -i 及 -p：非必选项，如果已经配置免密登录目标机，则不需填写，否则选择其一即可。-i 为可登录到目标机的 root 用户（或 --user 指定的其他用户）的私钥，也可使用 -p 交互式输入该用户的密码。
- TiUP DM 使用内置的 SSH 客户端，如需使用系统自带的 SSH 客户端，请参考 TiUP DM 文档中使用中控机系统自带的 SSH 客户端连接集群章节进行设置。

预期日志结尾输出会有 Deployed cluster `dm-test` successfully 关键词，表示部署成功。

12.8.3.2.5 第 4 步：查看 TiUP 管理的集群情况

```
tiup dm list
```

TiUP 支持管理多个 DM 集群，该命令会输出当前通过 TiUP DM 管理的所有集群信息，包括集群名称、部署用户、版本、密钥信息等：

```
Name  User  Version  Path                                       PrivateKey
----  ----  -------  ----                                       ----------
dm-test  tidb  ${version}  /root/.tiup/storage/dm/clusters/dm-test  /root/.tiup/storage/dm/
    ↪ clusters/dm-test/ssh/id_rsa
```

12.8.3.2.6 第 5 步：检查部署的 DM 集群情况

例如，执行如下命令检查 dm-test 集群情况：

```
tiup dm display dm-test
```

预期输出包括 dm-test 集群中实例 ID、角色、主机、监听端口和状态（由于还未启动，所以状态为 Down/inactive）、目录信息。

12.8.3.2.7 第 6 步：启动集群

```
tiup dm start dm-test
```

预期结果输出 Started cluster `dm-test` successfully 表示启动成功。

12.8.3.2.8 第 7 步：验证集群运行状态

通过以下 TiUP 命令检查集群状态：

```
tiup dm display dm-test
```

在输出结果中，如果 Status 状态信息为 Up，说明集群状态正常。

12.8.3.2.9 第 8 步：使用 dmctl 管理迁移任务

dmctl 是用来控制集群运行命令的工具，推荐通过 TiUP 获取该工具。

dmctl 支持命令模式与交互模式，具体请见使用 dmctl 运维集群。

12.8.3.3 使用 TiUP 离线镜像部署 DM 集群

本文介绍如何使用 TiUP 离线部署 DM 集群，具体的操作步骤如下。

12.8.3.3.1 第 1 步：准备 TiUP 离线组件包

- 在线环境中安装 TiUP 包管理器工具。

  1. 执行如下命令安装 TiUP 工具：

     ```
     curl --proto '=https' --tlsv1.2 -sSf https://tiup-mirrors.pingcap.com/install.sh | sh
     ```

  2. 重新声明全局环境变量：

     ```
     source .bash_profile
     ```

  3. 确认 TiUP 工具是否安装：

     ```
     which tiup
     ```

- 使用 TiUP 制作离线镜像。

  1. 在一台和外网相通的机器上拉取需要的组件：

     ```
     # 将 ${version} 修改成实际需要的版本
     tiup mirror clone tidb-dm-${version}-linux-amd64 --os=linux --arch=amd64 \
         --dm-master=${version} --dm-worker=${version} --dmctl=${version} \
         --alertmanager=v0.17.0 --grafana=v4.0.3 --prometheus=v4.0.3 \
         --tiup=v$(tiup --version|grep 'tiup'|awk -F ' ' '{print $1}') --dm=v$(tiup --
             ↪ version|grep 'tiup'|awk -F ' ' '{print $1}')
     ```

     该命令会在当前目录下创建一个名叫 `tidb-dm-${version}-linux-amd64` 的目录，里面包含 TiUP 管理的组件包。

  2. 通过 tar 命令将该组件包打包然后发送到隔离环境的中控机：

     ```
     tar czvf tidb-dm-${version}-linux-amd64.tar.gz tidb-dm-${version}-linux-amd64
     ```

     此时，`tidb-dm-${version}-linux-amd64.tar.gz` 就是一个独立的离线环境包。

12.8.3.3.2 第 2 步: 部署离线环境 TiUP 组件

将离线包发送到目标集群的中控机后，执行以下命令安装 TiUP 组件：

```
#### 将 ${version} 修改成实际需要的版本
tar xzvf tidb-dm-${version}-linux-amd64.tar.gz
sh tidb-dm-${version}-linux-amd64/local_install.sh
source /home/tidb/.bash_profile
```

`local_install.sh` 脚本会自动执行 `tiup mirror set tidb-dm-${version}-linux-amd64` 命令将当前镜像地址设置为 `tidb-dm-${version}-linux-amd64`。

若需将镜像切换到其他目录，可以通过手动执行 `tiup mirror set <mirror-dir>` 进行切换，切换后如果再想切换成官方镜像，可执行 `tiup mirror set https://tiup-mirrors.pingcap.com`。

12.8.3.3.3 第 3 步：编辑初始化配置文件

请根据不同的集群拓扑，编辑 TiUP 所需的集群初始化配置文件。

请根据配置文件模板，新建一个配置文件 topology.yaml。如果有其他组合场景的需求，请根据多个模板自行调整。

部署 3 个 DM-master、3 个 DM-worker 与 1 个监控组件的配置如下：

```
---
global:
  user: "tidb"
  ssh_port: 22
  deploy_dir: "/home/tidb/dm/deploy"
  data_dir: "/home/tidb/dm/data"
  # arch: "amd64"

master_servers:
  - host: 172.19.0.101
  - host: 172.19.0.102
  - host: 172.19.0.103

worker_servers:
  - host: 172.19.0.101
  - host: 172.19.0.102
  - host: 172.19.0.103

monitoring_servers:
  - host: 172.19.0.101

grafana_servers:
  - host: 172.19.0.101

alertmanager_servers:
  - host: 172.19.0.101
```

注意：

- 如果不需要确保 DM 集群高可用，则可只部署 1 个 DM-master 节点，且部署的 DM-worker 节点数量不少于上游待迁移的 MySQL/MariaDB 实例数。
- 如果需要确保 DM 集群高可用，则推荐部署 3 个 DM-master 节点，且部署的 DM-worker 节点数量大于上游待迁移的 MySQL/MariaDB 实例数（如 DM-worker 节点数量比上游实例数多 2 个）。
- 对于需要全局生效的参数，请在配置文件中 server_configs 的对应组件下配置。
- 对于需要某个节点生效的参数，请在具体节点的 config 中配置。

- 配置的层次结构使用 . 表示。如：log.slow-threshold。更多格式说明，请参考 TiUP 配置参数模版。
- 更多参数说明，请参考 master config.toml.example、worker config.toml.example。
- 需要确保以下组件间端口可正常连通：
  - 各 DM-master 节点间的 peer_port（默认为 8291）可互相连通。
  - 各 DM-master 节点可连通所有 DM-worker 节点的 port（默认为 8262）。
  - 各 DM-worker 节点可连通所有 DM-master 节点的 port（默认为 8261）。
  - TiUP 节点可连通所有 DM-master 节点的 port（默认为 8261）。
  - TiUP 节点可连通所有 DM-worker 节点的 port（默认为 8262）。

12.8.3.3.4 第 4 步：执行部署命令

注意：

通过 TiUP 进行集群部署可以使用密钥或者交互密码方式来进行安全认证：

- 如果是密钥方式，可以通过 -i 或者 --identity_file 来指定密钥的路径；
- 如果是密码方式，可以通过 -p 进入密码交互窗口；
- 如果已经配置免密登录目标机，则不需填写认证。

```
tiup dm deploy dm-test ${version} ./topology.yaml --user root [-p] [-i /home/root/.ssh/gcp_rsa]
```

以上部署命令中：

- 通过 TiUP DM 部署的集群名称为 dm-test。
- 部署版本为 ${version}，可以通过执行 tiup list dm-master 来查看 TiUP 支持的最新版本。
- 初始化配置文件为 topology.yaml。
- --user root：通过 root 用户登录到目标主机完成集群部署，该用户需要有 ssh 到目标机器的权限，并且在目标机器有 sudo 权限。也可以用其他有 ssh 和 sudo 权限的用户完成部署。
- -i 及 -p：非必选项，如果已经配置免密登录目标机，则不需填写，否则选择其一即可。-i 为可登录到目标机的 root 用户（或 --user 指定的其他用户）的私钥，也可使用 -p 交互式输入该用户的密码。
- TiUP DM 使用内置的 SSH 客户端，如需使用系统自带的 SSH 客户端，请参考 TiUP DM 文档中使用中控机系统自带的 SSH 客户端连接集群章节进行设置。

预期日志结尾输出会有 Deployed cluster `dm-test` successfully 关键词，表示部署成功。

12.8.3.3.5 第 5 步：查看 TiUP 管理的集群情况

```
tiup dm list
```

TiUP 支持管理多个 DM 集群，该命令会输出当前通过 TiUP DM 管理的所有集群信息，包括集群名称、部署用户、版本、密钥信息等：

```
Name  User  Version  Path                                       PrivateKey
----  ----  -------  ----                                       ----------
dm-test  tidb  ${version}  /root/.tiup/storage/dm/clusters/dm-test  /root/.tiup/storage/dm/
    ↪ clusters/dm-test/ssh/id_rsa
```

12.8.3.3.6 第 6 步：检查部署的 DM 集群情况

例如，执行如下命令检查 dm-test 集群情况：

```
tiup dm display dm-test
```

预期输出包括 dm-test 集群中实例 ID、角色、主机、监听端口和状态（由于还未启动，所以状态为 Down/inactive）、目录信息。

12.8.3.3.7 第 7 步：启动集群

```
tiup dm start dm-test
```

预期结果输出 Started cluster `dm-test` successfully 表示启动成功。

12.8.3.3.8 第 8 步：验证集群运行状态

通过以下 TiUP 命令检查集群状态：

```
tiup dm display dm-test
```

在输出结果中，如果 Status 状态信息为 Up，说明集群状态正常。

12.8.3.4 使用 DM binary 部署 DM 集群

本文将介绍如何使用 DM binary 快速部署 DM 集群。

> 注意：
>
> 对于生产环境，推荐使用 TiUP 部署 DM 集群及相关监控组件。

12.8.3.4.1 准备工作

使用下表中的链接下载官方 binary：

安装包	操作系统	架构	SHA256 校验和
https ↪ ://download ↪ .pingcap ↪ .org ↪ /dm ↪ -{version}- ↪ linux ↪ - ↪ amd64 ↪ .tar ↪ .gz	Linux	amd64	https ↪ ://download ↪ .pingcap ↪ .org ↪ /dm ↪ -{version}- ↪ linux ↪ - ↪ amd64 ↪ .sha256 ↪

> **注意：**
>
> 下载链接中的 {version} 为 DM 的版本号。例如，v1.0.1 版本的下载链接为 https://download.pingcap.org/dm-v1.0.1-linux-amd64.tar.gz。可以通过 DM Release 查看当前已发布版本。

下载的文件中包括子目录 bin 和 conf。bin 目录下包含 dm-master、dm-worker 以及 dmctl 的二进制文件。conf 目录下有相关的示例配置文件。

12.8.3.4.2 使用样例

假设在五台服务器上部署两个 DM-worker 实例和三个 DM-master 实例。各个节点的信息如下：

实例	服务器地址	端口
DM-master1	192.168.0.4	8261
DM-master2	192.168.0.5	8261
DM-master3	192.168.0.6	8261
DM-worker1	192.168.0.7	8262
DM-worker2	192.168.0.8	8262

下面以此为例，说明如何部署 DM。

> 注意：
>
> - 在单机部署多个 DM-master 或 DM-worker 时，需要确保每个实例的端口以及运行命令的当前目录各不相同。
> - 如果不需要确保 DM 集群高可用，则可只部署 1 个 DM-master 节点，且部署的 DM-worker 节点数量不少于上游待迁移的 MySQL/MariaDB 实例数。
> - 如果需要确保 DM 集群高可用，则推荐部署 3 个 DM-master 节点，且部署的 DM-worker 节点数量大于上游待迁移的 MySQL/MariaDB 实例数（如 DM-worker 节点数量比上游实例数多 2 个）。
> - 需要确保以下组件间端口可正常连通：
>   - 各 DM-master 节点间的 8291 端口可互相连通。
>   - 各 DM-master 节点可连通所有 DM-worker 节点的 8262 端口。
>   - 各 DM-worker 节点可连通所有 DM-master 节点的 8261 端口。

部署 DM-master

DM-master 提供命令行参数和配置文件两种配置方式。

使用命令行参数部署 DM-master

DM-master 的命令行参数说明：

```
./bin/dm-master --help
```

```
Usage of dm-master:
  -L string
        log level: debug, info, warn, error, fatal (default "info")
  -V    prints version and exit
  -advertise-addr string
        advertise address for client traffic (default "${master-addr}")
  -advertise-peer-urls string
        advertise URLs for peer traffic (default "${peer-urls}")
  -config string
        path to config file
  -data-dir string
        path to the data directory (default "default.${name}")
  -initial-cluster string
        initial cluster configuration for bootstrapping, e.g. dm-master=http://127.0.0.1:8291
  -join string
        join to an existing cluster (usage: cluster's "${master-addr}" list, e.g.
            ↪ "127.0.0.1:8261,127.0.0.1:18261"
  -log-file string
```

```
      log file path
  -master-addr string
      master API server and status addr
  -name string
      human-readable name for this DM-master member
  -peer-urls string
      URLs for peer traffic (default "http://127.0.0.1:8291")
  -print-sample-config
      print sample config file of dm-worker
```

> **注意：**
>
> 某些情况下，无法使用命令行参数来配置 DM-master，因为有的配置并未暴露给命令行。

使用配置文件部署 DM-master

推荐使用配置文件，把以下配置文件内容写入到 `conf/dm-master1.toml` 中。

DM-master 的配置文件：

```
#### Master Configuration.

name = "master1"

#### 日志配置
log-level = "info"
log-file = "dm-master.log"

#### DM-master 监听地址
master-addr = "192.168.0.4:8261"

#### DM-master 节点的对等 URL
peer-urls = "192.168.0.4:8291"

#### 初始集群中所有 DM-master 的 advertise-peer-urls 的值
initial-cluster = "master1=http://192.168.0.4:8291,master2=http://192.168.0.5:8291,master3=http
    ↪ ://192.168.0.6:8291"
```

在终端中使用下面的命令运行 DM-master：

> **注意：**
>
> 执行该命令后控制台不会输出日志，可以通过 `tail -f dm-master.log` 查看运行日志。

```
./bin/dm-master -config conf/dm-master1.toml
```

对于 DM-master2 和 DM-master3，修改配置文件中的 name 为 master2 和 master3，并将 peer-urls 的值改为 192.168.0.5:8291 和 192.168.0.6:8291 即可。

部署 DM-worker

DM-worker 提供命令行参数和配置文件两种配置方式。

使用命令行参数部署 DM-worker

查看 DM-worker 的命令行参数说明：

```
./bin/dm-worker --help
```

```
Usage of worker:
  -L string
        log level: debug, info, warn, error, fatal (default "info")
  -V    prints version and exit
  -advertise-addr string
        advertise address for client traffic (default "${worker-addr}")
  -config string
        path to config file
  -join string
        join to an existing cluster (usage: dm-master cluster's "${master-addr}")
  -keepalive-ttl int
        dm-worker's TTL for keepalive with etcd (in seconds) (default 10)
  -log-file string
        log file path
  -name string
        human-readable name for DM-worker member
  -print-sample-config
        print sample config file of dm-worker
  -worker-addr string
        listen address for client traffic
```

> 注意：
>
> 某些情况下，无法使用命令行参数的方法来配置 DM-worker，因为有的配置并未暴露给命令行。

使用配置文件部署 DM-worker

推荐使用配置文件来配置 DM-worker，把以下配置文件内容写入到 conf/dm-worker1.toml 中。

DM-worker 的配置文件：

```
#### Worker Configuration.

name = "worker1"

#### 日志配置
log-level = "info"
log-file = "dm-worker.log"

#### DM-worker 的地址
worker-addr = ":8262"

#### 对应集群中 DM-master 配置中的 master-addr
join = "192.168.0.4:8261,192.168.0.5:8261,192.168.0.6:8261"
```

在终端中使用下面的命令运行 DM-worker：

```
./bin/dm-worker -config conf/dm-worker1.toml
```

对于 DM-worker2，修改配置文件中的 name 为 worker2 即可。

这样，DM 集群就部署成功了。

12.8.3.5 在 Kubernetes 环境中部署

12.8.4 入门指南

12.8.4.0.1 管理上游数据源配置

本文介绍了如何使用dmctl 组件来管理数据源配置，包括如何加密数据库密码，数据源操作，查看数据源配置，改变数据源与 DM-worker 的绑定关系。

加密数据库密码

在 DM 相关配置文件中，推荐使用经 dmctl 加密后的密码。对于同一个原始密码，每次加密后密码不同。

```
./dmctl -encrypt 'abc!@#123'
```

```
MKxn0Qo3m3XOyjCnhEMtsUCm83EhGQDZ/T4=
```

数据源操作

operate-source 命令向 DM 集群加载、列出、移除数据源。

```
help operate-source
```

```
`create`/`update`/`stop`/`show` upstream MySQL/MariaDB source.

Usage:
```

```
  dmctl operate-source <operate-type> [config-file ...] [--print-sample-config] [flags]

Flags:
  -h, --help                  help for operate-source
  -p, --print-sample-config   print sample config file of source

Global Flags:
  -s, --source strings   MySQL Source ID
```

参数解释

- create：创建一个或多个上游的数据库源。创建多个数据源失败时，会尝试回滚到执行命令之前的状态
- update：更新一个上游的数据库源
- stop：停止一个或多个上游的数据库源。停止多个数据源失败时，可能有部分数据源已成功停止
- show：显示已添加的数据源以及对应的 DM-worker
- config-file：
  - 指定 source.yaml 的文件路径
  - 可传递多个文件路径
- --print-sample-config：打印示例配置文件。该参数会忽视其余参数

命令用法示例

使用 operate-source 命令创建数据源配置：

```
operate-source create ./source.yaml
```

其中 source.yaml 的配置参考上游数据库配置文件介绍。

结果如下：

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "dm-worker-1"
        }
    ]
}
```

查看数据源配置

> **注意：**
>
> `get-config` 命令仅在 DM v2.0.1 及其以后版本支持。

如果知道 source-id，可以通过 `dmctl --master-addr <master-addr> get-config source <source-id>` 命令直接查看数据源配置。

```
get-config source mysql-replica-01
```

```
{
  "result": true,
  "msg": "",
  "cfg": "enable-gtid: false
    flavor: mysql
    source-id: mysql-replica-01
    from:
      host: 127.0.0.1
      port: 8407
      user: root
      password: '******'
}
```

如果不知道 source-id，可以先通过 `dmctl --master-addr <master-addr> operate-source show` 查看源数据库列表。

```
operate-source show
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "source is added but there is no free worker to bound",
            "source": "mysql-replica-02",
            "worker": ""
        },
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "dm-worker-1"
```

```
        }
    ]
}
```

改变数据源与 DM-worker 的绑定关系

`transfer-source` 用于改变数据源与 DM-worker 的绑定关系。

```
help transfer-source
```

```
Transfers a upstream MySQL/MariaDB source to a free worker.

Usage:
  dmctl transfer-source <source-id> <worker-id> [flags]

Flags:
  -h, --help   help for transfer-source

Global Flags:
  -s, --source strings   MySQL Source ID.
```

在改变绑定关系前，DM 会检查待解绑的 worker 是否正在运行同步任务，如果正在运行则需要先暂停任务，并在改变绑定关系后恢复任务。

命令用法示例

如果不清楚 DM-worker 的绑定关系，可以通过 `dmctl --master-addr <master-addr> list-member --worker` 查看。

```
list-member --worker
```

```
{
    "result": true,
    "msg": "",
    "members": [
        {
            "worker": {
                "msg": "",
                "workers": [
                    {
                        "name": "dm-worker-1",
                        "addr": "127.0.0.1:8262",
                        "stage": "bound",
                        "source": "mysql-replica-01"
                    },
                    {
                        "name": "dm-worker-2",
                        "addr": "127.0.0.1:8263",
```

```
                    "stage": "free",
                    "source": ""
                }
            ]
        }
    }
]
}
```

在本示例中 mysql-replica-01 绑定到了 dm-worker-1 上。使用如下命令可以将该数据源绑定到 dm-worker-2 上

```
transfer-source mysql-replica-01 dm-worker-2
```

```
{
    "result": true,
    "msg": ""
}
```

再次通过 dmctl --master-addr <master-addr> list-member --worker 查看，检查命令已生效。

```
list-member --worker
```

```
{
    "result": true,
    "msg": "",
    "members": [
        {
            "worker": {
                "msg": "",
                "workers": [
                    {
                        "name": "dm-worker-1",
                        "addr": "127.0.0.1:8262",
                        "stage": "free",
                        "source": ""
                    },
                    {
                        "name": "dm-worker-2",
                        "addr": "127.0.0.1:8263",
                        "stage": "bound",
                        "source": "mysql-replica-01"
                    }
                ]
            }
        }
```

```
    ]
}
```

12.8.4.0.2 数据迁移任务配置向导

本文档介绍如何配置 Data Migration (DM) 的数据迁移任务。

配置需要迁移的数据源

配置需要迁移的数据源之前，首先应该确认已经在 DM 创建相应数据源：

- 查看数据源可以参考查看数据源配置
- 创建数据源可以参考在 DM 创建数据源
- 数据源配置可以参考数据源配置文件介绍

仿照下面的 `mysql-instances`: 示例定义数据迁移任务需要同步的单个或者多个数据源。

```
---

###### ********* 任务信息配置 *********
name: test              # 任务名称，需要全局唯一

###### ******** 数据源配置 **********
mysql-instances:
  - source-id: "mysql-replica-01"  # 从 source-id = mysql-replica-01 的数据源迁移数据
  - source-id: "mysql-replica-02"  # 从 source-id = mysql-replica-02 的数据源迁移数据
```

配置迁移的目标 TiDB 集群

仿照下面的 `target-database`: 示例定义迁移的目标 TiDB 集群。

```
---

###### ********* 任务信息配置 *********
name: test              # 任务名称，需要全局唯一

###### ******** 数据源配置 **********
mysql-instances:
  - source-id: "mysql-replica-01"  # 从 source-id = mysql-replica-01 的数据源迁移数据
  - source-id: "mysql-replica-02"  # 从 source-id = mysql-replica-02 的数据源迁移数据

###### ******** 目标 TiDB 配置 **********
target-database:        # 目标 TiDB 配置
  host: "127.0.0.1"
  port: 4000
  user: "root"
  password: ""          # 如果密码不为空，则推荐使用经过 dmctl 加密的密文
```

配置需要迁移的表

如果不需要过滤或迁移特定表，可以跳过该项配置。

配置从数据源迁移表的黑白名单，则需要添加两个定义，详细配置规则参考Block & Allow Lists：

1. 定义全局的黑白名单规则

```
block-allow-list:
  bw-rule-1:                            # 规则名称
    do-dbs: ["test.*", "user"]          # 迁移哪些库，支持通配符 "*" 和 "?"，do-dbs 和 ignore
        ↪ -dbs 只需要配置一个，如果两者同时配置只有 do-dbs 会生效
    # ignore-dbs: ["mysql", "account"] # 忽略哪些库，支持通配符 "*" 和 "?"
    do-tables:                          # 迁移哪些表，do-tables 和 ignore-tables
        ↪ 只需要配置一个，如果两者同时配置只有 do-tables 会生效
    - db-name: "test.*"
      tbl-name: "t.*"
    - db-name: "user"
      tbl-name: "information"
  bw-rule-2:                           # 规则名称
    ignore-tables:                     # 忽略哪些表
    - db-name: "user"
      tbl-name: "log"
```

2. 在数据源配置中引用黑白名单规则，过滤该数据源需要迁移的表

```
mysql-instances:
    - source-id: "mysql-replica-01"  # 从 source-id = mysql-replica-01 的数据源迁移数据
      block-allow-list:  "bw-rule-1" # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-beta.2
          ↪ 则使用 black-white-list
    - source-id: "mysql-replica-02"  # 从 source-id = mysql-replica-02 的数据源迁移数据
      block-allow-list:  "bw-rule-2" # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-beta.2
          ↪ 则使用 black-white-list
```

配置需要过滤的操作

如果不需要过滤特定库或者特定表的特定操作，可以跳过该项配置。

配置过滤特定操作，则需要添加两个定义，详细配置规则参考 Binlog Event Filter：

1. 定义全局的数据源操作过滤规则

```
filters:                                        # 定义过滤数据源特定操作的规则，
    ↪ 可以定义多个规则
  filter-rule-1:                                # 规则名称
    schema-pattern: "test_*"                    # 匹配数据源的库名，支持通配符 "*" 和 "?"
    table-pattern: "t_*"                        # 匹配数据源的表名，支持通配符 "*" 和 "?"
    events: ["truncate table", "drop table"]    # 匹配上 schema-pattern 和 table-pattern
        ↪ 的库或者表的操作类型
```

```
    action: Ignore                                  # 迁移 ( Do ) 还是忽略(Ignore)
  filter-rule-2:
    schema-pattern: "test"
    events: ["all dml"]
    action: Do
```

2. 在数据源配置中引用数据源操作过滤规则，过滤该数据源的指定库或表的指定操作

```
mysql-instances:
  - source-id: "mysql-replica-01"    # 从 source-id = mysql-replica-01 的数据源迁移数据
    block-allow-list:  "bw-rule-1"   # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-beta.2
        ↪ 则使用 black-white-list
    filter-rules: ["filter-rule-1"]  # 过滤数据源特定操作的规则，可以配置多个过滤规则
  - source-id: "mysql-replica-02"    # 从 source-id = mysql-replica-02 的数据源迁移数据
    block-allow-list:  "bw-rule-2"   # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-beta.2
        ↪ 则使用 black-white-list
    filter-rules: ["filter-rule-2"]  # 过滤数据源特定操作的规则，可以配置多个过滤规则
```

配置需要数据源表到目标 TiDB 表的映射

如果不需要将数据源表路由到不同名的目标 TiDB 表，可以跳过该项配置。分库分表合并迁移的场景必须配置该规则。

配置数据源表迁移到目标 TiDB 表的路由规则，则需要添加两个定义，详细配置规则参考 Table Routing：

1. 定义全局的路由规则

```
routes:                          # 定义数据源表迁移到目标 TiDB 表的路由规则，
    ↪ 可以定义多个规则
  route-rule-1:                  # 规则名称
    schema-pattern: "test_*"     # 匹配数据源的库名，支持通配符 "*" 和 "?"
    table-pattern: "t_*"         # 匹配数据源的表名，支持通配符 "*" 和 "?"
    target-schema: "test"        # 目标 TiDB 库名
    target-table: "t"            # 目标 TiDB 表名
  route-rule-2:
    schema-pattern: "test_*"
    target-schema: "test"
```

2. 在数据源配置中引用路由规则，过滤该数据源需要迁移的表

```
mysql-instances:
  - source-id: "mysql-replica-01"                  # 从 source-id = mysql-replica-01
      ↪ 的数据源迁移数据
    block-allow-list:  "bw-rule-1"                 # 黑白名单配置名称，如果 DM 版本早于
        ↪ v2.0.0-beta.2 则使用 black-white-list
    filter-rules: ["filter-rule-1"]                # 过滤数据源特定操作的规则，
        ↪ 可以配置多个过滤规则
```

```
    route-rules: ["route-rule-1", "route-rule-2"]     # 数据源表迁移到目标 TiDB 表的路由规则
        ↪ ，可以定义多个规则
  - source-id: "mysql-replica-02"                     # 从 source-id = mysql-replica-02
      ↪ 的数据源迁移数据
    block-allow-list:  "bw-rule-2"                    # 黑白名单配置名称，如果 DM 版本早于
        ↪ v2.0.0-beta.2 则使用 black-white-list
    filter-rules: ["filter-rule-2"]                   # 过滤数据源特定操作的规则，
        ↪ 可以配置多个过滤规则
```

配置是否进行分库分表合并

如果是分库分表合并的数据迁移场景，并且需要同步分库分表的 DDL，则必须显式配置 shard-mode，否则不要配置该选项。

分库分表 DDL 同步问题特别多，请确认了解 DM 同步分库分表 DDL 的原理和限制后，谨慎使用。

```
---

###### ********* 任务信息配置 *********
name: test                      # 任务名称，需要全局唯一
shard-mode: "pessimistic"       # 默认值为 "" 即无需协调。如果为分库分表合并任务，
    ↪ 请设置为悲观协调模式 "pessimistic"。在深入了解乐观协调模式的原理和使用限制后，
    ↪ 也可以设置为乐观协调模式 "optimistic"
```

其他配置

下面是本数据迁移任务配置向导的完整示例。完整的任务配置参见DM 任务完整配置文件介绍，其他各配置项的功能和配置也可参阅数据迁移功能。

```
---

###### ********* 任务信息配置 *********
name: test                      # 任务名称，需要全局唯一
shard-mode: "pessimistic"       # 默认值为 "" 即无需协调。如果为分库分表合并任务，
    ↪ 请设置为悲观协调模式 "pessimistic"。在深入了解乐观协调模式的原理和使用限制后，
    ↪ 也可以设置为乐观协调模式 "optimistic"
task-mode: all                  # 任务模式，可设为 "full" - "只进行全量数据迁移"、"incremental" -
    ↪  "Binlog 实时同步"、"all" - "全量 + Binlog 迁移"
##### timezone: "UTC"               # 指定数据迁移任务时 SQL Session 使用的时区。DM
    ↪ 默认使用目标库的全局时区配置进行数据迁移，并且自动确保同步数据的正确性。
    ↪ 使用自定义时区依然可以确保整个流程的正确性，但一般不需要手动指定。

###### ******** 数据源配置 **********
mysql-instances:
  - source-id: "mysql-replica-01"                 # 从 source-id = mysql-replica-01
      ↪ 的数据源迁移数据
    block-allow-list:  "bw-rule-1"                # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-
        ↪ beta.2 则使用 black-white-list
```

```
    filter-rules: ["filter-rule-1"]                # 过滤数据源特定操作的规则，
        ↪ 可以配置多个过滤规则
    route-rules: ["route-rule-1", "route-rule-2"]  # 数据源表迁移到目标 TiDB 表的路由规则，
        ↪ 可以定义多个规则
  - source-id: "mysql-replica-02"                  # 从 source-id = mysql-replica-02
      ↪ 的数据源迁移数据
    block-allow-list:  "bw-rule-2"                 # 黑白名单配置名称，如果 DM 版本早于 v2.0.0-
        ↪ beta.2 则使用 black-white-list
    filter-rules: ["filter-rule-2"]                # 过滤数据源特定操作的规则，
        ↪ 可以配置多个过滤规则
    route-rules: ["route-rule-2"]                  # 数据源表迁移到目标 TiDB 表的路由规则，
        ↪ 可以定义多个规则

###### ******** 目标 TiDB 配置 **********
target-database:       # 目标 TiDB 配置
  host: "127.0.0.1"
  port: 4000
  user: "root"
  password: ""         # 如果密码不为空，则推荐使用经过 dmctl 加密的密文

###### ******** 功能配置 **********
block-allow-list:                      # 定义数据源迁移表的过滤规则，可以定义多个规则。如果 DM
    ↪ 版本早于 v2.0.0-beta.2 则使用 black-white-list
  bw-rule-1:                           # 规则名称
    do-dbs: ["test.*", "user"]         # 迁移哪些库，支持通配符 "*" 和 "?"，do-dbs 和 ignore-dbs
        ↪ 只需要配置一个，如果两者同时配置只有 do-dbs 会生效
    # ignore-dbs: ["mysql", "account"] # 忽略哪些库，支持通配符 "*" 和 "?"
    do-tables:                         # 迁移哪些表，do-tables 和 ignore-tables 只需要配置一个，
        ↪ 如果两者同时配置只有 do-tables 会生效
    - db-name: "test.*"
      tbl-name: "t.*"
    - db-name: "user"
      tbl-name: "information"
  bw-rule-2:                         # 规则名称
    ignore-tables:                   # 忽略哪些表
    - db-name: "user"
      tbl-name: "log"

filters:                                       # 定义过滤数据源特定操作的规则，可以定义多个规则
  filter-rule-1:                               # 规则名称
    schema-pattern: "test_*"                   # 匹配数据源的库名，支持通配符 "*" 和 "?"
    table-pattern: "t_*"                       # 匹配数据源的表名，支持通配符 "*" 和 "?"
    events: ["truncate table", "drop table"]   # 匹配上 schema-pattern 和 table-pattern
        ↪ 的库或者表的操作类型
    action: Ignore                             # 迁移（Do）还是忽略(Ignore)
```

```
  filter-rule-2:
    schema-pattern: "test"
    events: ["all dml"]
    action: Do

routes:                             # 定义数据源表迁移到目标 TiDB 表的路由规则，可以定义多个规则
  route-rule-1:                     # 规则名称
    schema-pattern: "test_*"        # 匹配数据源的库名，支持通配符 "*" 和 "?"
    table-pattern: "t_*"            # 匹配数据源的表名，支持通配符 "*" 和 "?"
    target-schema: "test"           # 目标 TiDB 库名
    target-table: "t"               # 目标 TiDB 表名
  route-rule-2:
    schema-pattern: "test_*"
    target-schema: "test"
```

12.8.4.0.3 主要特性

本文档介绍 DM 提供的数据迁移功能以及相关的配置选项与使用示例。

Table Routing、 Block & Allow Lists、 Binlog Event Filter 在匹配库表名时，有以下版本差异：

- 对于 v1.0.5 版及后续版本，以上功能均支持通配符匹配。但注意所有版本中通配符匹配中的 * 符号 只能有一个且必须在末尾。
- 对于 v1.0.5 以前的版本，Table Routing 和 Binlog Event Filter 支持通配符，但不支持 `[...]` 与 `[!...]` 表达式。Block & Allow Lists 仅支持正则表达式。

在简单任务场景下推荐使用通配符匹配。

Table routing

Table routing 提供将上游 MySQL/MariaDB 实例的某些表迁移到下游指定表的功能。

> 注意：
>
> - 不支持对同一个表设置多个不同的路由规则。
> - Schema 的匹配规则需要单独设置，用来迁移 `CREATE/DROP SCHEMA xx`，例如下面参数配置中的 rule-2。

参数配置

```
routes:
  rule-1:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    target-schema: "test"
```

```
  target-table: "t"
rule-2:
  schema-pattern: "test_*"
  target-schema: "test"
```

参数解释

将根据 schema-pattern/table-pattern 匹配上该规则的上游 MySQL/MariaDB 实例的表迁移到下游的 target-schema/target-table。

使用示例

下面展示了三个不同场景下的配置示例。

分库分表合并

假设存在分库分表场景，需要将上游两个 MySQL 实例的表 test_{1,2,3...}.t_{1,2,3...} 迁移到下游 TiDB 的一张表 test.t。

为了迁移到下游实例的表 test.t，需要创建以下 table routing 规则：

- rule-1 用来迁移匹配上 schema-pattern: "test_*" 和 table-pattern: "t_*" 的表的 DML/DDL 语句到下游的 test.t。
- rule-2 用来迁移匹配上 schema-pattern: "test_*" 的库的 DDL 语句，例如 CREATE/DROP SCHEMA xx。

> 注意：
>
> - 如果下游 TiDB schema: test 已经存在，并且不会被删除，则可以省略 rule-2。
> - 如果下游 TiDB schema: test 不存在，只设置了 rule_1，则迁移会报错 schema test doesn't exist。

```
rule-1:
  schema-pattern: "test_*"
  table-pattern: "t_*"
  target-schema: "test"
  target-table: "t"
rule-2:
  schema-pattern: "test_*"
  target-schema: "test"
```

分库合并

假设存在分库场景，将上游两个 MySQL 实例 test_{1,2,3...}.t_{1,2,3...} 迁移到下游 TiDB 的 test.t_{1,2,3...}，创建一条路由规则即可：

```
rule-1:
  schema-pattern: "test_*"
  target-schema: "test"
```

错误的 table routing

假设存在下面两个路由规则，`test_1_bak.t_1_bak` 可以匹配上 `rule-1` 和 `rule-2`，违反 table 路由的限制而报错。

```
  rule-0:
    schema-pattern: "test_*"
    target-schema: "test"
  rule-1:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    target-schema: "test"
    target-table: "t"
  rule-2:
    schema-pattern: "test_1_bak"
    table-pattern: "t_1_bak"
    target-schema: "test"
    target-table: "t_bak"
```

Block & Allow Table Lists

上游数据库实例表的黑白名单过滤规则，可以用来过滤或者只迁移某些 `database/table` 的所有操作。

参数配置

```
block-allow-list:             # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  rule-1:
    do-dbs: ["test*"]         # 非 ~ 字符开头，表示规则是通配符；v1.0.5 及后续版本支持通配符规则
        ↪ 。
    do-tables:
    - db-name: "test[123]"    # 匹配 test1、test2、test3。
      tbl-name: "t[1-5]"      # 匹配 t1、t2、t3、t4、t5。
    - db-name: "test"
      tbl-name: "t"
  rule-2:
    do-dbs: ["~^test.*"]      # 以 ~ 字符开头，表示规则是正则表达式。
    ignore-dbs: ["mysql"]
    do-tables:
    - db-name: "~^test.*"
      tbl-name: "~^t.*"
    - db-name: "test"
      tbl-name: "t"
    ignore-tables:
    - db-name: "test"
      tbl-name: "log"
```

参数解释

- `do-dbs`：要迁移的库的白名单，类似于 MySQL 中的 `replicate-do-db`。

- `ignore-dbs`：要迁移的库的黑名单，类似于 MySQL 中的 `replicate-ignore-db`。
- `do-tables`：要迁移的表的白名单，类似于 MySQL 中的 `replicate-do-table`。必须同时指定 `db-name` 与 `tbl-name`。
- `ignore-tables`：要迁移的表的黑名单，类似于 MySQL 中的 `replicate-ignore-table`。必须同时指定 `db-name` 与 `tbl-name`。

以上参数值以 `~` 开头时均支持使用正则表达式来匹配库名、表名。

过滤规则

`do-dbs` 与 `ignore-dbs` 对应的过滤规则与 MySQL 中的 Evaluation of Database-Level Replication and Binary Logging Options 类似，`do-tables` 与 `ignore-tables` 对应的过滤规则与 MySQL 中的 Evaluation of Table-Level Replication Options 类似。

> 注意：
>
> DM 中黑白名单过滤规则与 MySQL 中相应规则存在以下区别：
>
> - MySQL 中存在 `replicate-wild-do-table` 与 `replicate-wild-ignore-table` 用于支持通配符，DM 中各配置参数直接支持以 `~` 字符开头的正则表达式。
> - DM 当前只支持 `ROW` 格式的 binlog，不支持 `STATEMENT/MIXED` 格式的 binlog，因此应与 MySQL 中 `ROW` 格式下的规则对应。
> - 对于 DDL，MySQL 仅依据默认的 database 名称（`USE` 语句显式指定的 database）进行判断，而 DM 优先依据 DDL 中的 database 名称部分进行判断，并当 DDL 中不包含 database 名称时再依据 `USE` 部分进行判断。假设需要判断的 SQL 为 `USE test_db_2; CREATE TABLE test_db_1.test_table (c1 INT PRIMARY KEY)`，且 MySQL 配置了 `replicate-do-db=test_db_1`、DM 配置了 `do-dbs: ["test_db_1"]`，则对于 MySQL 该规则不会生效，而对于 DM 该规则会生效。

判断 table `test.t` 是否应该被过滤的流程如下：

1. 首先进行 schema 过滤判断
    - 如果 `do-dbs` 不为空，判断 `do-dbs` 中是否存在一个匹配的 schema。
        - 如果存在，则进入 table 过滤判断。
        - 如果不存在，则过滤 `test.t`。
    - 如果 `do-dbs` 为空并且 `ignore-dbs` 不为空，判断 `ignore-dbs` 中是否存在一个匹配的 schema。
        - 如果存在，则过滤 `test.t`。
        - 如果不存在，则进入 table 过滤判断。
    - 如果 `do-dbs` 和 `ignore-dbs` 都为空，则进入 table 过滤判断。
2. 进行 table 过滤判断
    1. 如果 `do-tables` 不为空，判断 `do-tables` 中是否存在一个匹配的 table。
        - 如果存在，则迁移 `test.t`。

- 如果不存在，则过滤 test.t。

2. 如果 ignore-tables 不为空，判断 ignore-tables 中是否存在一个匹配的 table。
   - 如果存在，则过滤 test.t.
   - 如果不存在，则迁移 test.t。
3. 如果 do-tables 和 ignore-tables 都为空，则迁移 test.t。

> **注意：**
>
> 如果是判断 schema test 是否应该被过滤，则只进行 schema 过滤判断。

使用示例

假设上游 MySQL 实例包含以下表：

```
`logs`.`messages_2016`
`logs`.`messages_2017`
`logs`.`messages_2018`
`forum`.`users`
`forum`.`messages`
`forum_backup_2016`.`messages`
`forum_backup_2017`.`messages`
`forum_backup_2018`.`messages`
```

配置如下：

```
block-allow-list:  # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  bw-rule:
    do-dbs: ["forum_backup_2018", "forum"]
    ignore-dbs: ["~^forum_backup_"]
    do-tables:
    - db-name: "logs"
      tbl-name: "~_2018$"
    - db-name: "~^forum.*"
       tbl-name: "messages"
    ignore-tables:
    - db-name: "~.*"
       tbl-name: "^messages.*"
```

应用 bw-rule 规则后：

table	是否过滤	过滤的原因
logs ↪ .messages_2016 ↪	是	schema logs 没有匹配到 do-dbs 任意一项
logs ↪ .messages_2017 ↪	是	schema logs 没有匹配到 do-dbs 任意一项
logs ↪ .messages_2018 ↪	是	schema logs 没有匹配到 do-dbs 任意一项
forum_backup_2016 ↪ .messages ↪	是	schema forum_backup_2016 没有匹配到 do-dbs 任意一项
forum_backup_2017 ↪ .messages ↪	是	schema forum_backup_2017 没有匹配到 do-dbs 任意一项
forum ↪ .users ↪	是	1. schema forum 匹配到 do-dbs，进入 table 过滤判断 2. schema 和 table 没有匹配到 do-tables 和 ignore-tables 中任意一项，并且 do-tables 不为空，因此过滤
forum ↪ .messages ↪	否	1. schema forum 匹配到 do-dbs，进入 table 过滤判断 2. schema 和 table 匹配到 do-tables 的 db-name: "~^forum ↪ .*",tbl-name: ↪ "messages"

table	是否过滤	过滤的原因
`forum_backup_2018` ↪ `.messages` ↪	否	1. schema `forum_backup_2018` 匹配到 `do-dbs`，进入 table 过滤判断 2. schema 和 table 匹配到 `do-tables` 的 `db-name: "~^forum` ↪ `.*",tbl-name:` ↪ `"messages"`

Binlog event filter

Binlog event filter 是比迁移表黑白名单更加细粒度的过滤规则，可以指定只迁移或者过滤掉某些 `schema / table` 的指定类型 binlog，比如 `INSERT`、`TRUNCATE TABLE`。

> 注意：
>
> - 同一个表匹配上多个规则，将会顺序应用这些规则，并且黑名单的优先级高于白名单，即如果同时存在规则 `Ignore` 和 `Do` 应用在某个 table 上，那么 `Ignore` 生效。
> - 从 DM v2.0.2 开始，Binlog event filter 也可以在上游数据库配置文件中进行配置。见上游数据库配置文件介绍。

参数配置

```
filters:
  rule-1:
    schema-pattern: "test_*"
     table-pattern: "t_*"
     events: ["truncate table", "drop table"]
    sql-pattern: ["^DROP\\s+PROCEDURE", "^CREATE\\s+PROCEDURE"]
     action: Ignore
```

参数解释

- `schema-pattern/table-pattern`：对匹配上的上游 MySQL/MariaDB 实例的表的 binlog events 或者 DDL SQL 语句通过以下规则进行过滤。
- `events`：binlog events 数组，仅支持从以下 `Event` 中选择一项或多项。

Event	分类	解释
all		代表包含下面所有的 events

Event	分类	解释
all dml		代表包含下面所有 DML events
all ddl		代表包含下面所有 DDL events
none		代表不包含下面所有 events
none ddl		代表不包含下面所有 DDL events
none dml		代表不包含下面所有 DML events
insert	DML	insert DML event
update	DML	update DML event
delete	DML	delete DML event
create database	DDL	create database event
drop database	DDL	drop database event
create table	DDL	create table event
create index	DDL	create index event
drop table	DDL	drop table event
truncate table	DDL	truncate table event
rename table	DDL	rename table event
drop index	DDL	drop index event
alter table	DDL	alter table event

- `sql-pattern`：用于过滤指定的 DDL SQL 语句，支持正则表达式匹配，例如上面示例中的 `"^DROP\\s+PROCEDURE"`。
- `action`：string (`Do` / `Ignore`)；进行下面规则判断，满足其中之一则过滤，否则不过滤。
  - `Do`：白名单。binlog event 如果满足下面两个条件之一就会被过滤掉：
    * 不在该 rule 的 `events` 中。
    * 如果规则的 `sql-pattern` 不为空的话，对应的 SQL 没有匹配上 `sql-pattern` 中任意一项。
  - `Ignore`：黑名单。如果满足下面两个条件之一就会被过滤掉：
    * 在该 rule 的 `events` 中。
    * 如果规则的 `sql-pattern` 不为空的话，对应的 SQL 可以匹配上 `sql-pattern` 中任意一项。

使用示例

过滤分库分表的所有删除操作

需要设置下面两个 `Binlog event filter rule` 来过滤掉所有的删除操作：

- `filter-table-rule` 过滤掉所有匹配到 pattern `test_*.t_*` 的 table 的 `turncate table`、`drop table`、`delete statement` 操作。
- `filter-schema-rule` 过滤掉所有匹配到 pattern `test_*` 的 schema 的 `drop database` 操作。

```
filters:
  filter-table-rule:
    schema-pattern: "test_*"
    table-pattern: "t_*"
```

```
    events: ["truncate table", "drop table", "delete"]
    action: Ignore
  filter-schema-rule:
    schema-pattern: "test_*"
    events: ["drop database"]
    action: Ignore
```

只迁移分库分表的 DML 操作

需要设置下面两个 Binlog event filter rule 只迁移 DML 操作：

- do-table-rule 只迁移所有匹配到 pattern test_*.t_* 的 table 的 create table、insert、update、delete 操作。
- do-schema-rule 只迁移所有匹配到 pattern test_* 的 schema 的 create database 操作。

> 注意：
>
> 迁移 create database/table 的原因是创建库和表后才能迁移 DML。

```
filters:
  do-table-rule:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    events: ["create table", "all dml"]
    action: Do
  do-schema-rule:
    schema-pattern: "test_*"
    events: ["create database"]
    action: Do
```

过滤 TiDB 不支持的 SQL 语句

可设置如下规则过滤 TiDB 不支持的 PROCEDURE 语句：

```
filters:
  filter-procedure-rule:
    schema-pattern: "test_*"
    table-pattern: "t_*"
    sql-pattern: ["^DROP\\s+PROCEDURE", "^CREATE\\s+PROCEDURE"]
    action: Ignore
```

过滤 TiDB parser 不支持的 SQL 语句

对于 TiDB parser 不支持的 SQL 语句，DM 无法解析获得 schema/table 信息，因此需要使用全局过滤规则：schema-pattern: "*"。

> **注意：**
>
> 全局过滤规则的设置必须尽可能严格，以避免过滤掉需要迁移的数据。

可设置如下规则过滤某些版本的 TiDB parser 不支持的 PARTITION 语句：

```
filters:
  filter-partition-rule:
    schema-pattern: "*"
    sql-pattern: ["ALTER\\s+TABLE[\\s\\S]*ADD\\s+PARTITION", "ALTER\\s+TABLE[\\s\\S]*DROP\\s+
        ↪ PARTITION"]
    action: Ignore
```

online DDL 工具支持

在 MySQL 生态中，gh-ost 与 pt-osc 等工具较广泛地被使用，DM 对其提供了特殊的支持以避免对不必要的中间数据进行迁移。

有关 DM 对 online DDL 工具支持的原理、处理流程等，可参考online-ddl。

使用限制

- DM 仅针对 gh-ost 与 pt-osc 做了特殊支持。
- 在开启 `online-ddl` 时，增量复制对应的 checkpoint 应不处于 online DDL 执行过程中。如上游某次 online DDL 操作开始于 binlog `position-A`、结束于 `position-B`，则增量复制的起始点应早于 `position-A` 或晚于 `position-B`，否则可能出现迁移出错，具体可参考FAQ。

参数配置

在 v2.0.5 及之后的版本，请在 `task` 配置文件中使用 `online-ddl` 配置项。

如上游 MySQL/MariaDB（同时）使用 gh-ost 或 pt-osc 工具，则在 task 的配置文件中设置：

```
online-ddl: true
```

> **注意：**
>
> 自 v2.0.5 起，`online-ddl-scheme` 已被弃用，请使用 `online-ddl` 代替 `online-ddl-scheme`。如设置 `online-ddl: true` 会覆盖掉 `online-ddl-scheme`。如设置 `online-ddl-scheme: "pt"` 或 `online-ddl-scheme: "gh-ost"` 会被转换为 `online-ddl: true`。

在 v2.0.5 之前的版本（不含 v2.0.5），请在 `task` 配置文件中使用 `online-ddl-scheme` 配置项。

如上游 MySQL/MariaDB 使用的是 gh-ost 工具，则在 task 的配置文件中设置：

```
online-ddl-scheme: "gh-ost"
```

如上游 MySQL/MariaDB 使用的是 pt-osc 工具，则在 task 的配置文件中设置：

```
online-ddl-scheme: "pt"
```

分库分表合并

DM 支持将上游 MySQL/MariaDB 各分库分表中的 DML、 DDL 数据合并后迁移到下游 TiDB 的库表中。

使用限制

目前分库分表合并功能仅支持有限的场景，使用该功能前，请仔细阅读悲观模式分库分表合并迁移使用限制和乐观模式分库分表合并迁移使用限制。

参数配置

在 task 的配置文件中设置：

```
shard-mode: "pessimistic" # 默认值为 "" 即无需协调。如果为分库分表合并任务，请设置为悲观协调模式
    ↪ "pessimistic"。在深入了解乐观协调模式的原理和使用限制后，也可以设置为乐观协调模式 "
    ↪ optimistic"
```

手动处理 Sharding DDL Lock

如果分库分表合并迁移过程中发生了异常，对于部分场景，可尝试参考手动处理 Sharding DDL Lock进行处理。

12.8.4.0.4 通过 SQL 表达式过滤 DML

在进行增量数据迁移时，可以通过如何过滤 binlog 事件功能过滤某些类型的 binlog event，例如不向下游迁移 DELETE 事件以达到归档、审计等目的。但是 binlog event filter 无法以更细粒度判断某一行的 DELETE 事件是否要被过滤。

为了解决上述问题，从 v2.0.5 起，DM 支持在增量数据同步阶段使用`binlog value filter`过滤迁移数据。DM 支持的 ROW 格式的 binlog 中，binlog event 带有所有列的值。你可以基于这些值配置 SQL 表达式。如果该表达式对于某条行变更的计算结果是 TRUE，DM 就不会向下游迁移该条行变更。

具体操作步骤和实现细节，请参考如何通过 SQL 表达式过滤 DML。

12.8.4.0.5 迁移任务操作

上游 MySQL 实例配置前置检查

本文介绍了 DM 提供的前置检查功能，此功能用于在数据迁移任务启动时提前检测出上游 MySQL 实例配置中可能存在的一些错误。

使用命令

`check-task` 命令用于对上游 MySQL 实例配置是否满足 DM 要求进行前置检查。

检查内容

上下游数据库用户必须具备相应读写权限。当数据迁移任务启动时，DM 会自动检查下列权限和配置：

- 数据库版本
  - MySQL 版本 > 5.5
  - MariaDB 版本 >= 10.1.2

> **警告：**
>
> 支持从 MySQL v8.0 迁移数据是 DM v2.0 及以上版本的实验特性，不建议在生产环境下使用。

- 数据库配置
  - 是否设置 `server_id`
- MySQL binlog 配置
  - binlog 是否开启（DM 要求 binlog 必须开启）
  - 是否有 `binlog_format=ROW`（DM 只支持 ROW 格式的 binlog 迁移）
  - 是否有 `binlog_row_image=FULL`（DM 只支持 `binlog_row_image=FULL`）
- 上游 MySQL 实例用户的权限

  DM 配置中的 MySQL 用户至少需要具备以下权限：
  - REPLICATION SLAVE
  - REPLICATION CLIENT
  - RELOAD
  - SELECT
- 上游 MySQL 表结构的兼容性

  TiDB 和 MySQL 的兼容性存在以下一些区别：
  - TiDB 不支持外键
  - 字符集的兼容性不同，详见 TiDB 支持的字符集

  DM 还会检查上游表中是否存在主键或唯一键约束，在 v1.0.7 版本引入。
- 上游 MySQL 多实例分库分表的一致性
  - 所有分表的表结构是否一致，检查内容包括：
    * Column 数量
    * Column 名称
    * Column 位置
    * Column 类型
    * 主键
    * 唯一索引
  - 分表中自增主键冲突检查
    * 在两种情况下会造成检查失败：
      · 分表存在自增主键，且自增主键 column 类型不为 bigint

· 分表存在自增主键，自增主键 column 类型为 bigint，但没有为其配置列值转换

* 其他情况下检查将成功

关闭检查项

DM 会根据任务类型进行相应检查，用户可以在任务配置文件中使用 `ignore-checking-items` 配置关闭检查。`ignore-checking-items` 是一个列表，其中可能的取值包括：

取值	含义
all	关闭所有检查
dump_privilege	关闭检查上游 MySQL 实例用户的 dump 相关权限
replication_privilege	关闭检查上游 MySQL 实例用户的 replication 相关权限
version	关闭检查上游数据库版本
server_id	关闭检查上游数据库是否设置 server_id
binlog_enable	关闭检查上游数据库是否已启用 binlog
binlog_format	关闭检查上游数据库 binlog 格式是否为 ROW
binlog_row_image	关闭检查上游数据库 binlog_row_image 是否为 FULL
table_schema	关闭检查上游 MySQL 表结构的兼容性
schema_of_shard_tables	关闭检查上游 MySQL 多实例分库分表的表结构一致性
auto_increment_ID	关闭检查上游 MySQL 多实例分库分表的自增主键冲突

创建数据迁移任务

`start-task` 命令用于创建数据迁移任务。当数据迁移任务启动时，DM 将自动对相应权限和配置进行前置检查。

```
help start-task
```

```
Starts a task as defined in the configuration file
Usage:
  dmctl start-task [-s source ...] [--remove-meta] <config-file> [flags]
Flags:
  -h, --help          Help for start-task
      --remove-meta   Whether to remove task's metadata
Global Flags:
  -s, --source strings   MySQL Source ID
```

命令用法示例

```
start-task [ -s "mysql-replica-01"] ./task.yaml
```

参数解释

- `-s`:
  - 可选
  - 指定在特定的一个 MySQL 源上执行 `task.yaml`

  - 如果设置，则只启动指定任务在该 MySQL 源上的子任务

- `config-file`：
  - 必选
  - 指定 `task.yaml` 的文件路径

- `remove-meta`:
  - 可选
  - 如果设置，则在启动指定任务时会移除该任务之前存在的 metadata

返回结果示例

```
start-task task.yaml
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

TiDB Data Migration 查询状态

本文介绍 TiDB Data Migration (DM) `query-status` 命令的查询结果、任务状态与子任务状态。

查询结果

```
» query-status
```

```
{
    "result": true,     # 查询是否成功
    "msg": "",          # 查询失败原因描述
    "tasks": [          # 迁移 task 列表
        {
            "taskName": "test",         # 任务名称
            "taskStatus": "Running",    # 任务运行状态
            "sources": [                # 该任务的上游 MySQL 列表
                "mysql-replica-01",
                "mysql-replica-02"
            ]
        },
        {
```

```
            "taskName": "test2",
            "taskStatus": "Paused",
            "sources": [
                "mysql-replica-01",
                "mysql-replica-02"
            ]
        }
    ]
}
```

关于 tasks 下的 taskStatus 状态的详细定义，请参阅任务状态。

推荐的 `query-status` 使用方法是：

1. 首先使用 query-status 查看各个 task 的运行状态是否正常。
2. 如果发现其中某一 task 状态有问题，通过 `query-status <出错任务的 taskName>` 来得到更详细的错误信息。

任务状态

DM 的迁移任务状态取决于其分配到 DM-worker 上的子任务状态，定义见下表：

任务对应的所有子任务的状态	任务状态
任一子任务处于“Paused”状态且返回结果有错误信息	Error - Some error occurred in subtask

任务对应的所有子任务的状态	任务状态
任一处于 Sync 阶段的子任务处于“Running”状态但其 Relay 处理单元未运行 (处于 Error/-Paused/Stopped 状态)	Error - Relay status is Error/-Paused/Stopped
任一子任务处于“Paused”状态且返回结果没有错误信息	Paused

任务对应的所有子任务的状态	任务状态
所有子任务处于“New”状态	New
所有子任务处于“Finished”状态	Finished
所有子任务处于“Stopped”状态	Stopped
其他情况	Running

详情查询结果

```
» query-status test
```

```
{
    "result": true,     # 查询是否成功
    "msg": "",          # 查询失败原因描述
    "sources": [        # 上游 MySQL 列表
        {
            "result": true,
            "msg": "",
            "sourceStatus": {                   # 上游 MySQL 的信息
                "source": "mysql-replica-01",
                "worker": "worker1",
                "result": null,
                "relayStatus": null
```

```
},
"subTaskStatus": [              # 上游 MySQL 所有子任务的信息
    {
        "name": "test",         # 子任务名称
        "stage": "Running",     # 子任务运行状态，包括 “New”，“Running”，“Paused
            ↪ ”，“Stopped” 以及 “Finished”
        "unit": "Sync",         # DM 的处理单元，包括 “Check”，“Dump“，“Load”
            ↪ 以及 “Sync”
        "result": null,         # 子任务失败时显示错误信息
        "unresolvedDDLLockID": "test-`test`.`t_target`",    # sharding DDL lock ID,
            ↪ 可用于异常情况下手动处理 sharding DDL lock
        "sync": {                   # 当前 `Sync` 处理单元的迁移信息
            "totalEvents": "12",    # 该子任务中迁移的 binlog event 总数
            "totalTps": "1",        # 该子任务中每秒迁移的 binlog event 数量
            "recentTps": "1",       # 该子任务中最后一秒迁移的 binlog event 数量
            "masterBinlog": "(bin.000001, 3234)",                               #
                ↪ 上游数据库当前的 binlog position
            "masterBinlogGtid": "c0149e17-dff1-11e8-b6a8-0242ac110004:1-14",    #
                ↪ 上游数据库当前的 GTID 信息
            "syncerBinlog": "(bin.000001, 2525)",                               #
                ↪ 已被 `Sync` 处理单元迁移的 binlog position
            "syncerBinlogGtid": "",                                             #
                ↪ 使用 GTID 迁移的 binlog position
            "blockingDDLs": [       # 当前被阻塞的 DDL 列表。该项仅在当前 DM-worker
                ↪ 所有上游表都处于 “synced“ 状态时才有数值，
                ↪ 此时该列表包含的是待执行或待跳过的 sharding DDL 语句
                "USE `test`; ALTER TABLE `test`.`t_target` DROP COLUMN `age`;"
            ],
            "unresolvedGroups": [   # 没有被解决的 sharding group 信息
                {
                    "target": "`test`.`t_target`",                  # 待迁移的下游表
                    "DDLs": [
                        "USE `test`; ALTER TABLE `test`.`t_target` DROP COLUMN `age
                            ↪ `;"
                    ],
                    "firstPos": "(bin|000001.000001, 3130)",        # sharding DDL
                        ↪ 语句起始 binlog position
                    "synced": [                                     # `Sync`
                        ↪ 处理单元已经读到该 sharding DDL 的上游分表
                        "`test`.`t2`"
                        "`test`.`t3`"
                        "`test`.`t1`"
                    ],
                    "unsynced": [                                   # `Sync`
                        ↪ 处理单元未读到该 sharding DDL 的上游分表。
```

```
                        ↪ 如有上游分表未完成同步，`blockingDDLs` 为空
                    ]
                }
            ],
            "synced": false         # 增量复制是否已追上上游。由于后台 `Sync`
                ↪ 单元并不会实时刷新保存点，当前值为 “false“
                ↪ 并不一定代表发生了迁移延迟
        }
    }
]
},
{
    "result": true,
    "msg": "",
    "sourceStatus": {
        "source": "mysql-replica-02",
        "worker": "worker2",
        "result": null,
        "relayStatus": null
    },
    "subTaskStatus": [
        {
            "name": "test",
            "stage": "Running",
            "unit": "Load",
            "result": null,
            "unresolvedDDLLockID": "",
            "load": {                   # `Load` 处理单元的迁移信息
                "finishedBytes": "115", # 已全量导入字节数
                "totalBytes": "452",    # 总计需要导入的字节数
                "progress": "25.44 %"   # 全量导入进度
            }
        }
    ]
},
{
    "result": true,
    "sourceStatus": {
        "source": "mysql-replica-03",
        "worker": "worker3",
        "result": null,
        "relayStatus": null
    },
    "subTaskStatus": [
        {
```

```
                    "name": "test",
                    "stage": "Paused",
                    "unit": "Load",
                    "result": {                 # 错误示例
                        "isCanceled": false,
                        "errors": [
                            {
                                "Type": "ExecSQL",
                                "msg": "Error 1062: Duplicate entry '1155173304420532225' for key
                                    ↪  'PRIMARY'\n/home/jenkins/workspace/build_dm/go/src/
                                    ↪ github.com/pingcap/tidb-enterprise-tools/loader/db.go
                                    ↪ :160: \n/home/jenkins/workspace/build_dm/go/src/github.
                                    ↪ com/pingcap/tidb-enterprise-tools/loader/db.go:105: \n/
                                    ↪ home/jenkins/workspace/build_dm/go/src/github.com/pingcap
                                    ↪ /tidb-enterprise-tools/loader/loader.go:138: file test.t1
                                    ↪ .sql"
                            }
                        ],
                        "detail": null
                    },
                    "unresolvedDDLLockID": "",
                    "load": {
                        "finishedBytes": "0",
                        "totalBytes": "156",
                        "progress": "0.00 %"
                    }
                }
            ]
        }
    ]
}
```

关于 sources 下 subTaskStatus 中 stage 状态和状态转换关系的详细信息，请参阅子任务状态。

关于 sources 下 subTaskStatus 中 unresolvedDDLLockID的操作细节，请参阅手动处理 Sharding DDL Lock。

子任务状态

状态描述

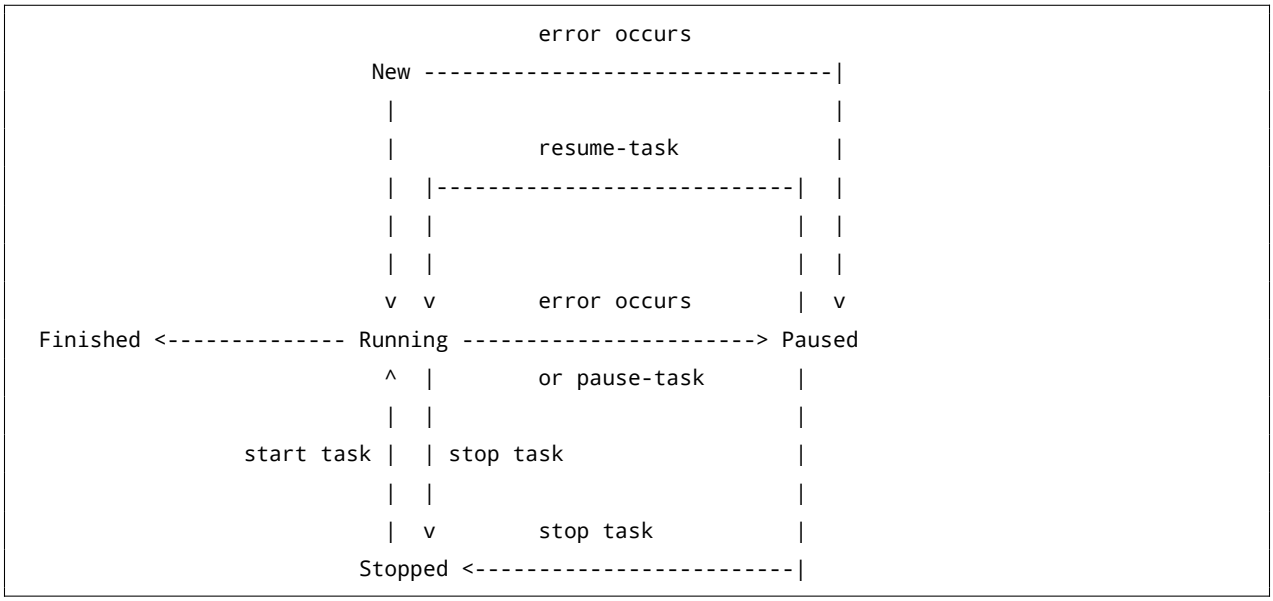
- New：
  - 初始状态。
  - 如果子任务没有发生错误，状态切换为 Running，其他情况则切换为 Paused。
- Running：正常运行状态。
- Paused：

- 暂停状态。
- 子任务发生错误，状态切换为 Paused。
- 如在子任务为 Running 状态下执行 pause-task 命令，任务状态会切换为 Paused。
- 如子任务处于该状态，可以使用 resume-task 命令恢复任务。

- Stopped：
  - 停止状态。
  - 如在子任务为 Running 或 Paused 状态下执行 stop-task 命令，任务状态会切换为 Stopped。
  - 如子任务处于该状态，不可使用 resume-task 命令恢复任务。

- Finished：
  - 任务完成状态。
  - 只有 task-mode 为 full 的任务正常完成后，任务才会切换为该状态。

状态转换图



暂停数据迁移任务

pause-task 命令用于暂停数据迁移任务。

> 注意：
>
> 有关 pause-task 与 stop-task 的区别如下：
>
> - 使用 pause-task 仅暂停迁移任务的执行，但仍然会在内存中保留任务的状态信息等，且可通过 query-status 进行查询；使用 stop-task 会停止迁移任务的执行，并移除内存中与该任务相关的信息，且不可再通过 query-status 进行查询，但不会移除已经写入到下游数据库中的数据以及其中的 checkpoint 等 dm_meta 信息。

- 使用 pause-task 暂停迁移任务期间，由于任务本身仍然存在，因此不能再启动同名的新任务，且会阻止对该任务所需 relay log 的清理；使用 stop-task 停止任务后，由于任务不再存在，因此可以再启动同名的新任务，且不会阻止对 relay log 的清理。
- pause-task 一般用于临时暂停迁移任务以排查问题等；stop-task 一般用于永久删除迁移任务或通过与 start-task 配合以更新配置信息。

```
help pause-task
```

```
pause a specified running task

Usage:
 dmctl pause-task [-s source ...] <task-name | task-file> [flags]

Flags:
 -h, --help   help for pause-task

Global Flags:
 -s, --source strings   MySQL Source ID
```

命令用法示例

```
pause-task [-s "mysql-replica-01"] task-name
```

参数解释

- -s：
  - 可选
  - 指定在特定的一个 MySQL 源上暂停数据迁移任务的子任务
  - 如果设置，则只暂停该任务在指定 MySQL 源上的子任务
- task-name| task-file：
  - 必选
  - 指定任务名称或任务文件路径

返回结果示例

```
pause-task test
```

```
{
    "op": "Pause",
    "result": true,
    "msg": "",
    "sources": [
```

```
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

恢复数据迁移任务

`resume-task` 命令用于恢复处于 Paused 状态的数据迁移任务，通常用于在人为处理完造成迁移任务暂停的故障后手动恢复迁移任务。

```
help resume-task
```

```
resume a specified paused task

Usage:
 dmctl resume-task [-s source ...] <task-name | task-file> [flags]

Flags:
 -h, --help   help for resume-task

Global Flags:
 -s, --source strings   MySQL Source ID
```

命令用法示例

```
resume-task [-s "mysql-replica-01"] task-name
```

参数解释

- `-s`:
    - 可选
    - 指定在特定的一个 MySQL 源上恢复数据迁移任务的子任务
    - 如果设置，则只恢复该任务在指定 MySQL 源上的子任务
- `task-name | task-file`:
    - 必选
    - 指定任务名称或任务文件路径

返回结果示例

```
resume-task test
```

```
{
    "op": "Resume",
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

停止数据迁移任务

stop-task 命令用于停止数据迁移任务。有关 stop-task 与 pause-task 的区别，请参考暂停数据迁移任务中的相关说明。

```
help stop-task
```

```
stop a specified task

Usage:
 dmctl stop-task [-s source ...] <task-name | task-file> [flags]

Flags:
 -h, --help   help for stop-task

Global Flags:
 -s, --source strings   MySQL Source ID
```

命令用法示例

```
stop-task [-s "mysql-replica-01"]  task-name
```

参数解释

- -s:
  - 可选
  - 指定在特定的一个 MySQL 源上停止数据迁移任务的子任务
  - 如果设置，则只停止该任务在指定 MySQL 源上的子任务
- task-name | task-file:
  - 必选
  - 指定任务名称或任务文件路径

返回结果示例

```
stop-task test
```

```
{
    "op": "Stop",
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

12.8.5 进阶教程

12.8.5.1 分库分表合并迁移

12.8.5.1.1 分库分表合并迁移

本文介绍了 DM 提供的分库分表的合并迁移功能，此功能可用于将上游 MySQL/MariaDB 实例中结构相同/不同的表迁移到下游 TiDB 的同一个表中。DM 不仅支持迁移上游的 DML 数据，也支持协调迁移多个上游分表的 DDL 表结构变更。

简介

DM 支持对上游多个分表的数据合并迁移到 TiDB 的一个表中，在迁移过程中需要协调各个分表的 DDL，以及该 DDL 前后的 DML。针对用户的使用场景，DM 支持悲观协调和乐观协调两种不同的模式。

> **注意：**
>
> - 要执行分库分表合并迁移任务，必须在任务配置文件中设置 `shard-mode`。
> - DM 对分库分表的合并默认使用悲观协调模式，在文档中如果没特殊说明则默认为悲观协调模式。
> - 在没有深入了解乐观模式的原理和使用限制的情况下不建议使用该模式，否则可能造成迁移中断甚至数据不一致的严重后果。

悲观协调模式

当上游一个分表执行某一 DDL 后，这个分表的迁移会暂停，并等待其他所有分表都执行了同样的 DDL 才在下游执行该 DDL 并继续数据迁移。“悲观协调”模式的优点是可以保证迁移到下游的数据不会出错，详细的介绍请参考悲观模式下分库分表合并迁移。

乐观协调模式

在一个分表上执行的 DDL，DM 会自动修改成兼容其他分表的语句，并立即迁移到下游，不会阻挡任何分表 DML 的迁移。“乐观协调”模式的优点是处理 DDL 时不会阻塞数据的迁移，但是使用不当会有迁移中断甚至数据不一致的风险，详细的介绍请参考乐观模式下分库分表合并迁移。

悲观协调模式与乐观协调模式的对比

悲观协调模式	乐观协调模式
发起 DDL 的分表会暂停 DML 迁移	发起 DDL 的分表会继续 DML 迁移
每个分表的 DDL 执行次序和语句必须相同	每个分表只需保持表结构互相兼容即可
DDL 在整个分表群达成一致后才迁移到下游	每个分表的 DDL 会即时影响下游
错误的 DDL 操作在侦测到后可以被拦截	错误的 DDL 操作也会被迁移到下游，可能在侦测到之前已使部分上下游数据不一

12.8.5.1.2 悲观模式下分库分表合并迁移

本文介绍了 DM 提供的悲观模式（默认模式）下分库分表合并迁移功能。此功能用于将上游 MySQL/MariaDB 实例中结构相同的表迁移到下游 TiDB 的同一个表中。

使用限制

DM 在悲观模式下进行分表 DDL 的迁移有以下几点使用限制：

- 对于一个逻辑 sharding group（需要合并迁移到下游同一个表的所有分表组成的 group），只能使用一个仅包含这些分表所在数据源的任务进行迁移。
- 在一个逻辑 sharding group 内，所有上游分表必须以相同的顺序执行相同的 DDL 语句（库名和表名可以不同），并且只有在所有分表执行完当前一条 DDL 语句后，下一条 DDL 语句才能执行。
    - 比如，如果在 table_1 表中先增加列 a 后再增加列 b，则在 table_2 表中就不能先增加列 b 后再增加列 a，因为 DM 不支持以不同的顺序来执行相同的 DDL 语句。
- 在一个逻辑 sharding group 内，所有上游分表都应该执行对应的 DDL 语句。
    - 比如，若 DM-worker-2 对应的一个或多个上游分表未执行 DDL 语句，则其他已执行 DDL 语句的 DM-worker 都会暂停迁移任务，直到等到 DM-worker-2 收到上游对应的 DDL 语句。
- sharding group 数据迁移任务不支持 `DROP DATABASE/TABLE` 语句。
    - DM-worker 中的 binlog 复制单元（sync）会自动忽略掉上游分表的 `DROP DATABASE` 和 `DROP TABLE` 语句。
- sharding group 数据迁移任务不支持 `TRUNCATE TABLE` 语句。
    - DM-worker 中的 binlog 复制单元（sync）会自动忽略掉上游分表的 `TRUNCATE TABLE` 语句。

- sharding group 数据迁移任务支持 RENAME TABLE 语句，但有如下限制（online DDL 中的 RENAME 有特殊方案进行支持）：
  - 只支持 RENAME TABLE 到一个不存在的表。
  - 一条 RENAME TABLE 语句只能包含一个 RENAME 操作。
- sharding group 数据迁移任务要求一个 DDL 语句仅包含对一张表的操作。
- 增量复制任务需要确认开始迁移的 binlog position 上各分表的表结构必须一致，才能确保来自不同分表的 DML 语句能够迁移到表结构确定的下游，并且后续各分表的 DDL 语句能够正确匹配与迁移。
- 如果需要变更 table routing 规则，必须先等所有 sharding DDL 语句迁移完成。
  - 在 sharding DDL 语句迁移过程中，使用 dmctl 尝试变更 router-rules 会报错。
- 如果需要创建新表加入到一个正在执行 DDL 语句的 sharding group 中，则必须保持新表结构和最新更改的表结构一致。
  - 比如，原 table_1, table_2 表初始时有 (a, b) 两列，sharding DDL 语句执行后有 (a, b, c) 三列，则迁移完成后新创建的表也应当有 (a, b, c) 三列。
- 由于已经收到 DDL 语句的 DM-worker 会暂停任务以等待其他 DM-worker 收到对应的 DDL 语句，因此数据迁移延迟会增加。

背景

目前，DM 使用 ROW 格式的 binlog 进行数据迁移，且 binlog 中不包含表结构信息。在 ROW 格式的 binlog 迁移过程中，如果不需要将多个上游表合并迁移到下游的同一个表，则只存在一个上游表的 DDL 语句会更新对应下游表结构。ROW 格式的 binlog 可以认为是具有 self-description 属性。

分库分表合并迁移过程中，可以根据 column values 及下游的表结构构造出相应的 DML 语句，但此时若上游的分表执行 DDL 语句进行了表结构变更，则必须对该 DDL 语句进行额外迁移处理，以避免因为表结构和 binlog 数据不一致而造成迁移出错的问题。

以下是一个简化后的例子：

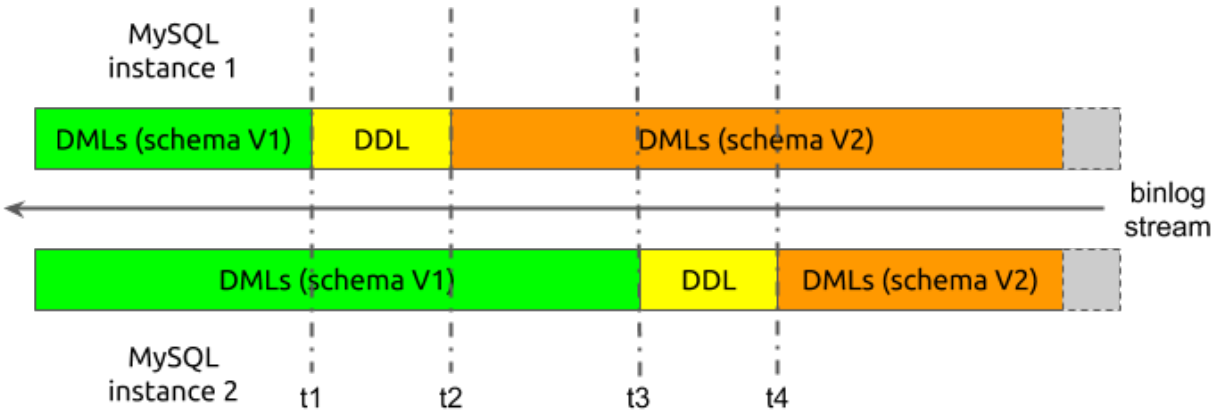


图 276: shard-ddl-example-1

在上图的例子中，分表的合库合表过程简化成了上游只有两个 MySQL 实例，每个实例内只有一个表。假设在数据迁移开始时，将两个分表的表结构版本记为 schema V1，将 DDL 语句执行完后的表结构版本记为 schema V2。

现在，假设数据迁移过程中，DM-worker 内的 binlog 复制单元（sync）从两个上游分表收到的 binlog 数据有如下时序：

1. 开始迁移时，sync 从两个分表收到的都是 schema V1 版本的 DML 语句。
2. 在 t1 时刻，sync 收到实例 1 上分表的 DDL 语句。
3. 从 t2 时刻开始，sync 从实例 1 收到的是 schema V2 版本的 DML 语句；但从实例 2 收到的仍是 schema V1 版本的 DML 语句。
4. 在 t3 时刻，sync 收到实例 2 上分表的 DDL 语句。
5. 从 t4 时刻开始，sync 从实例 2 收到的也是 schema V2 版本的 DML 语句。

假设在数据迁移过程中，不对分表的 DDL 语句进行额外处理。当实例 1 的 DDL 语句迁移到下游后，下游的表结构会变更成为 schema V2 版本。但在 t2 到 t3 这段时间内，sync 从实例 2 上收到的仍是 schema V1 版本的 DML 语句。当尝试把这些 schema V1 版本的 DML 语句迁移到下游时，就会由于 DML 语句与表结构的不一致而发生错误，从而无法正确迁移数据。

实现原理

基于上述例子，本部分介绍了 DM 在默认的悲观模式下合库合表过程中进行 DDL 迁移的实现原理。

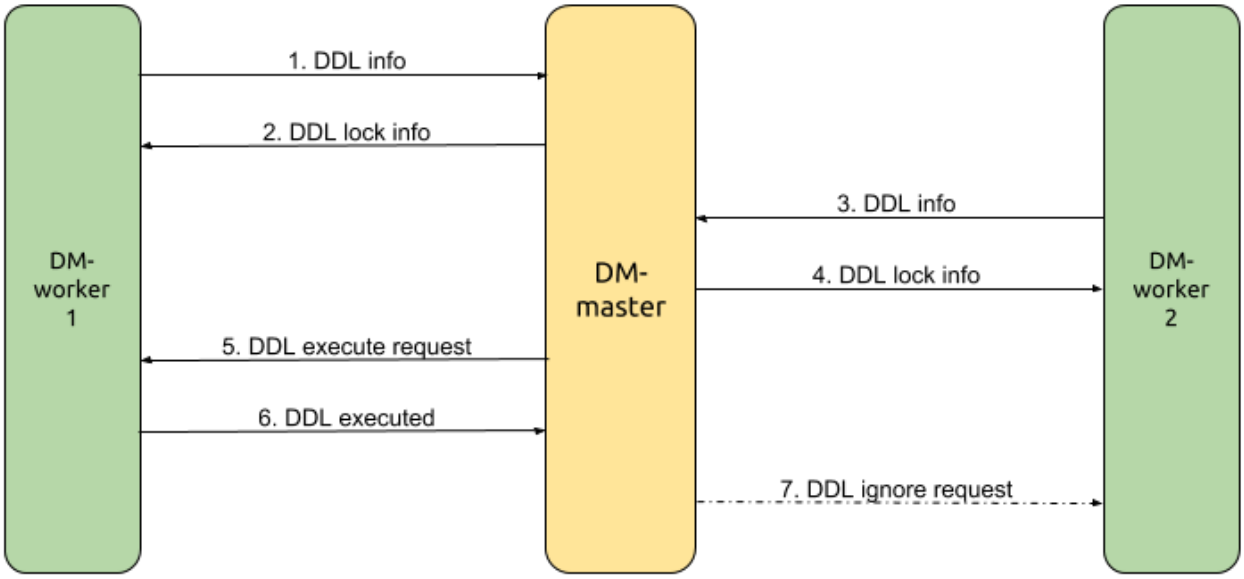


图 277: shard-ddl-flow

在这个例子中，DM-worker-1 负责迁移来自 MySQL 实例 1 的数据，DM-worker-2 负责迁移来自 MySQL 实例 2 的数据，DM-master 负责协调多个 DM-worker 间的 DDL 迁移。

从 DM-worker-1 收到 DDL 语句开始，简化后的 DDL 迁移流程为：

1. 在 t1 时刻，DM-worker-1 收到来自 MySQL 实例 1 的 DDL 语句，自身暂停该 DDL 语句对应任务的 DDL 及 DML 数据迁移，并将 DDL 相关信息发送给 DM-master。
2. DM-master 根据收到的 DDL 信息判断得知需要协调该 DDL 语句的迁移，于是为该 DDL 语句创建一个锁，并将 DDL 锁信息发回给 DM-worker-1，同时将 DM-worker-1 标记为这个锁的 owner。
3. DM-worker-2 继续进行 DML 语句的迁移，直到在 t3 时刻收到来自 MySQL 实例 2 的 DDL 语句，自身暂停该 DDL 语句对应任务的数据迁移，并将 DDL 相关信息发送给 DM-master。
4. DM-master 根据收到的 DDL 信息判断得知该 DDL 语句对应的锁信息已经存在，于是直接将对应锁信息发回给 DM-worker-2。
5. 根据任务启动时的配置信息、上游 MySQL 实例分表信息、部署拓扑信息等，DM-master 判断得知自身已经收到了来自待合表的所有上游分表的 DDL 语句，于是请求 DDL 锁的 owner（DM-worker-1）向下游迁移执行该 DDL。
6. DM-worker-1 根据第二步收到的 DDL 锁信息验证 DDL 语句执行请求；向下游执行 DDL，并将执行结果反馈给 DM-master；若 DDL 语句执行成功，则自身开始继续迁移后续的（从 t2 时刻对应的 binlog 开始的）DML 语句。
7. DM-master 收到来自 owner 执行 DDL 语句成功的响应，于是请求在等待该 DDL 锁的所有其他 DM-worker（DM-worker-2）忽略该 DDL 语句，直接继续迁移后续的（从 t4 时刻对应的 binlog 开始的）DML 语句。

根据上面的流程，可以归纳出 DM 协调多个 DM-worker 间 sharding DDL 迁移的特点：

- 根据任务配置与 DM 集群部署拓扑信息，DM-master 内部也会建立一个逻辑 sharding group 来协调 DDL 迁移，group 中的成员为负责处理该迁移任务拆解后的各子任务的 DM-worker。
- 各 DM-worker 从 binlog event 中收到 DDL 语句后，会将 DDL 信息发送给 DM-master。
- DM-master 根据来自 DM-worker 的 DDL 信息及 sharding group 信息创建或更新 DDL 锁。
- 如果 sharding group 的所有成员都收到了某一条相同的 DDL 语句，则表明上游分表在该 DDL 执行前的 DML 语句都已经迁移完成，此时可以执行该 DDL 语句，并继续后续的 DML 迁移。
- 上游所有分表的 DDL 在经过 table router 转换后需要保持一致，因此仅需 DDL 锁的 owner 执行一次该 DDL 语句即可，其他 DM-worker 可直接忽略对应的 DDL 语句。

在上面的示例中，每个 DM-worker 对应的上游 MySQL 实例中只有一个待合并的分表。但在实际场景下，一个 MySQL 实例可能有多个分库内的多个分表需要进行合并，这种情况下，sharding DDL 的协调迁移过程将更加复杂。

假设同一个 MySQL 实例中有 table_1 和 table_2 两个分表需要进行合并：

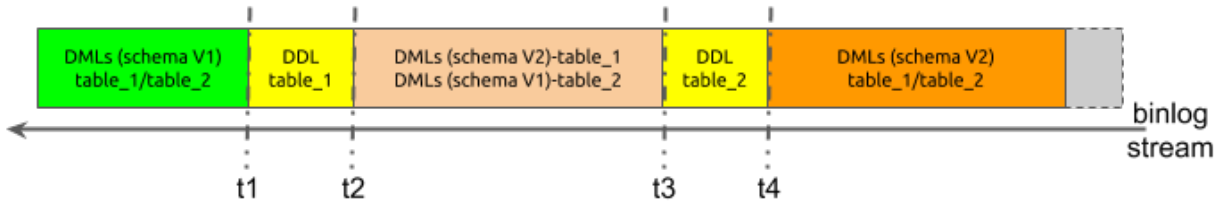


图 278: shard-ddl-example-2

在这个例子中，由于数据来自同一个 MySQL 实例，因此所有数据都是从同一个 binlog 流中获得，时序如下：

1. 开始迁移时，DM-worker 内的 sync 从两个分表收到的数据都是 schema V1 版本的 DML 语句。
2. 在 t1 时刻，sync 收到 table_1 分表的 DDL 语句。
3. 从 t2 到 t3 时刻，sync 收到的数据同时包含 table_1 的 DML 语句（schema V2 版本）及 table_2 的 DML 语句（schema V1 版本）。
4. 在 t3 时刻，sync 收到 table_2 分表的 DDL 语句。
5. 从 t4 时刻开始，sync 从两个分表收到的数据都是 schema V2 版本的 DML 语句。

假设在数据迁移过程中，不对分表的 DDL 语句进行额外处理。当 table_1 的 DDL 语句迁移到下游从而变更下游表结构后，table_2 的 DML 语句（schema V1 版本）将无法正常迁移。因此，在单个 DM-worker 内部，我们也构造了与 DM-master 内类似的逻辑 sharding group，但 group 的成员是同一个上游 MySQL 实例的不同分表。

DM-worker 内协调处理 sharding group 的迁移与 DM-master 处理 DM-worker 之间的迁移不完全一致，主要原因包括：

- 当 DM-worker 收到 table_1 分表的 DDL 语句时，迁移不能暂停，需要继续解析 binlog 才能获得后续 table_2 分表的 DDL 语句，即需要从 t2 时刻继续解析直到 t3 时刻。
- 在继续解析 t2 到 t3 时刻的 binlog 的过程中，table_1 分表的 DML 语句（schema V2 版本）不能向下游迁移；但当 sharding DDL 迁移并执行成功后，这些 DML 语句则需要迁移到下游。

DM-worker 内部 sharding DDL 迁移的简化流程为：

1. 在 t1 时刻，DM-worker 收到 table_1 的 DDL 语句，并记录 DDL 信息及此时的 binlog 位置点信息。
2. DM-worker 继续向前解析 t2 到 t3 时刻的 binlog。
3. 对于 table_1 的 DML 语句（schema V2 版本），忽略；对于 table_2 的 DML 语句（schema V1 版本），正常迁移到下游。
4. 在 t3 时刻，DM-worker 收到 table_2 的 DDL 语句，并记录 DDL 信息及此时的 binlog 位置点信息。
5. 根据迁移任务配置信息、上游库表信息等，DM-worker 判断得知该 MySQL 实例上所有分表的 DDL 语句都已收到；于是将该 DDL 语句迁移到下游执行并变更下游表结构。
6. DM-worker 设置 binlog 流的新解析起始位置点为第一步时保存的位置点。
7. DM-worker 重新开始解析从 t2 到 t3 时刻的 binlog。
8. 对于 table_1 的 DML 语句（schema V2 版本），正常迁移到下游；对于 table_2 的 DML 语句（schema V1 版本），忽略。
9. 解析到达第四步时保存的 binlog 位置点，可得知在第三步时被忽略的所有 DML 语句都已经重新迁移到下游。
10. DM-worker 继续从 t4 时刻对应的 binlog 位置点开始正常迁移。

综上可知，DM 在处理 sharding DDL 迁移时，主要通过两级 sharding group 来进行协调控制，简化的流程为：

1. 各 DM-worker 独立地协调对应上游 MySQL 实例内多个分表组成的 sharding group 的 DDL 迁移。
2. 当 DM-worker 收到所有分表的 DDL 语句时，向 DM-master 发送 DDL 相关信息。
3. DM-master 根据 DM-worker 发来的 DDL 信息，协调由各 DM-worker 组成的 sharing group 的 DDL 迁移。
4. 当 DM-master 收到所有 DM-worker 的 DDL 信息时，请求 DDL 锁的 owner（某个 DM-worker）执行该 DDL 语句。
5. DDL 锁的 owner 执行 DDL 语句，并将结果反馈给 DM-master；自身开始重新迁移在内部协调 DDL 迁移过程中被忽略的 DML 语句。
6. 当 DM-master 收到 owner 执行 DDL 成功的消息后，请求其他所有 DM-worker 继续开始迁移。
7. 其他所有 DM-worker 各自开始重新迁移在内部协调 DDL 迁移过程中被忽略的 DML 语句。
8. 在完成被忽略的 DML 语句的重新迁移后，所有 DM-worker 继续正常迁移。

12.8.5.1.3 乐观模式下分库分表合并迁移

本文介绍了 DM 提供的乐观模式下分库分表的合并迁移功能，此功能可用于将上游 MySQL/MariaDB 实例中结构相同/不同的表迁移到下游 TiDB 的同一个表中。

> **注意：**
>
> 在没有深入了解乐观模式的原理和使用限制的情况下不建议使用该模式，否则可能造成迁移中断甚至数据不一致的严重后果。

背景

DM 支持在线上执行分库分表的 DDL 语句（通称 Sharding DDL），默认使用“悲观模式”，即当上游一个分表执行某一 DDL 后，这个分表的迁移会暂停，等待其他所有分表都执行了同样的 DDL 才在下游执行该 DDL 并继续数据迁移。这种“悲观协调”模式的优点是可以保证迁移到下游的数据不会出错，缺点是会暂停数据迁移而不利于对上游进行灰度变更。有些用户可能会花较长时间在单一分表执行 DDL，验证一定时间后才会更改其他分表的结构。在悲观模式迁移的设定下，这些 DDL 会阻塞迁移，binlog 事件会大量积压。

因此，需要提供一种新的“乐观协调”模式，在一个分表上执行的 DDL，自动修改成兼容其他分表的语句后，立即迁移到下游，不会阻挡任何分表执行的 DML 的迁移。

乐观协调模式的配置

在任务的配置文件中指定 `shard-mode` 为 `optimistic` 则使用“乐观协调”模式，示例配置文件可以参考DM 任务完整配置文件介绍。

使用限制

使用“乐观协调”模式有一定的风险，需要严格遵照以下方针：

- 执行每个批次的 DDL 前和后，要确保每个分表的结构达成一致。

- 进行灰度 DDL 时，只集中在一个分表上测试。
- 灰度完成后，在其他分表上尽量以最简单直接的 DDL 迁移到最终的 schema，而不要重新执行灰度测试中对或错的每一步。
  - 例如：在分表执行过 `ADD COLUMN A INT; DROP COLUMN A; ADD COLUMN A FLOAT;`，在其他分表直接执行 `ADD COLUMN A FLOAT` 即可，不需要三条 DDL 都执行一遍。
- 执行 DDL 时要注意观察 DM 迁移状态。当迁移报错时，需要判断这个批次的 DDL 是否会造成数据不一致。

“乐观协调”模式暂不支持以下语句：

- `ALTER TABLE table_name ADD COLUMN column_name datatype NOT NULL`（添加无默认值的 not null 的列）。
- `ALTER TABLE table_name ADD COLUMN column_name datetime DEFAULT NOW()`（增加的列默认值不固定）。
- `ALTER TABLE table_name ADD COLUMN col1 INT, DROP COLUMN col2`（在一个 DDL 语句中同时包含 ADD COLUMN 与 DROP COLUMN）。
- `ALTER TABLE table_name RENAME COLUMN column_1 TO column_2;`（重命名列）。
- `ALTER TABLE table_name RENAME INDEX index_1 TO index_2;`（重命名索引）。

此外，不论是使用“乐观协调”或“悲观协调”，DM 仍是有以下限制：

- 增量复制任务需要确保开始迁移的 binlog position 对应的各分表的表结构必须一致。
- 进入 sharding group 的新表必须与其他成员的表结构一致（正在执行一个 DDL 批次时禁止 `CREATE/RENAME TABLE`）。
- 不支持 `DROP TABLE/DROP DATABASE`。
- 不支持 `TRUNCATE TABLE`。
- 单条 DDL 语句要求仅包含对一张表的操作。
- TiDB 不支持的 DDL 语句在 DM 也不支持。
- 新增列的默认值不能包含 `current_timestamp`、`rand()`、`uuid()` 等，否则会造成上下游数据不一致。

风险

使用乐观模式迁移时，由于 DDL 会即时迁移到下游，若使用不当，可能导致上下游数据不一致。

使数据不一致的操作

- 各分表的表结构不兼容，例：
  - 两个分表各自添加相同名称的列，但其类型不同。
  - 两个分表各自添加相同名称的列，但其默认值不同。
  - 两个分表各自添加相同名称的生成列，但其生成表达式不同。
  - 两个分表各自添加相同名称的索引，但其键组合不同。
  - 其他同名异构的情况。
- 在分表上执行对数据具有破坏性的 DDL，然后尝试回滚，例：
  - 删除一列 X，之后又把 X 加回来。

例子

例如以下三个分表合并迁移到 TiDB：

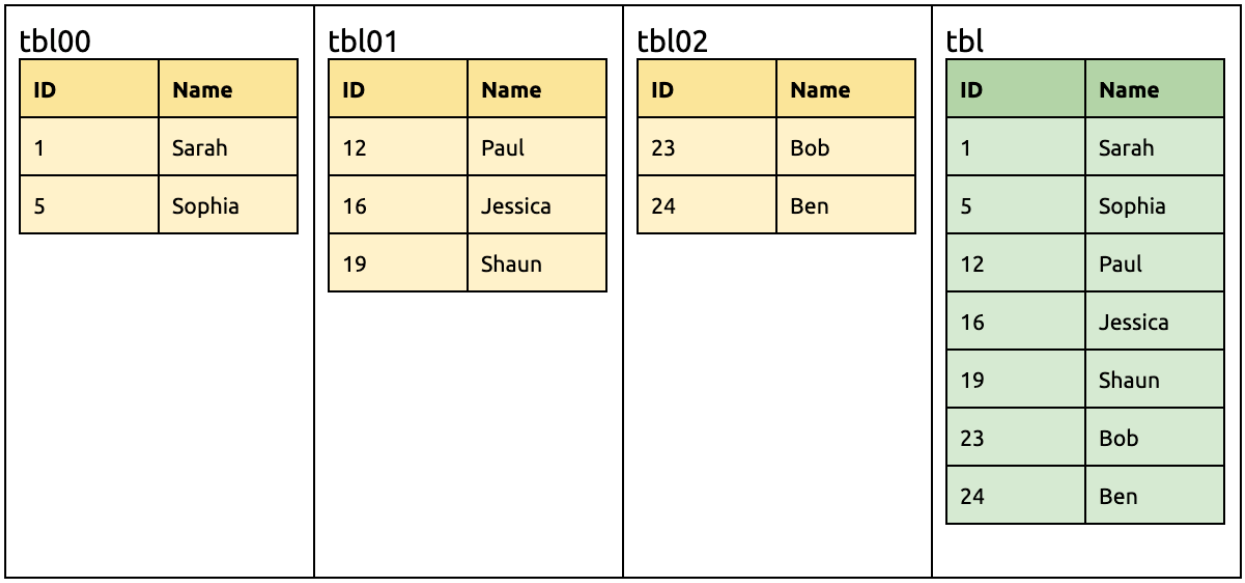


图 279: optimistic-ddl-fail-example-1

在 `tbl01` 新增一列 `Age`，默认值定为 `0`：

```
ALTER TABLE `tbl01` ADD COLUMN `Age` INT DEFAULT 0;
```

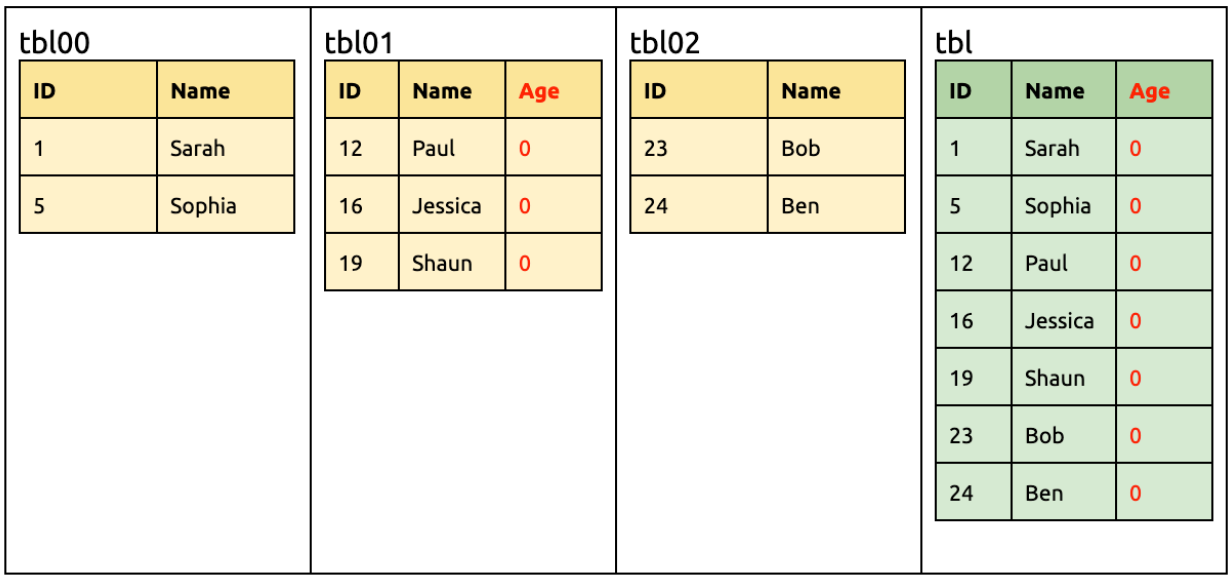


图 280: optimistic-ddl-fail-example-2

在 tbl00 新增一列 Age，但默认值定为 -1：

```
ALTER TABLE `tbl00` ADD COLUMN `Age` INT DEFAULT -1;
```

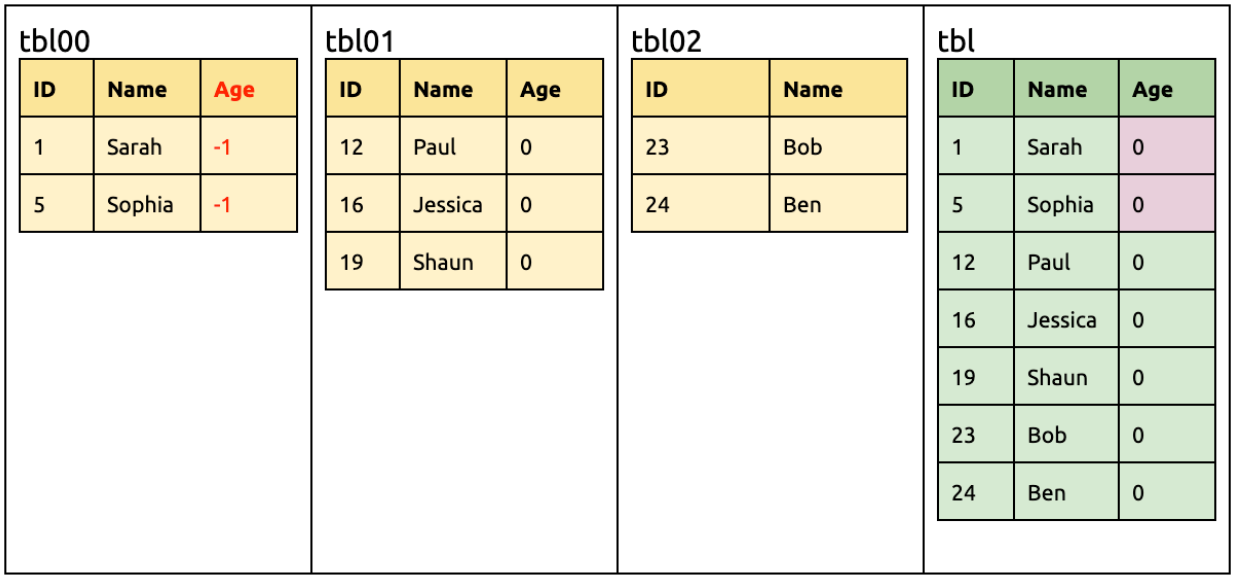


图 281: optimistic-ddl-fail-example-3

此时所有来自 tbl00 的 Age 都不一致了。这是由于 DEFAULT 0 和 DEFAULT -1 互不兼容。虽然 DM 遇到这种情况会报错，但上下游不一致的问题就需要手动去解决。

原理

在“乐观协调”模式下，DM-worker 接收到来自上游的 DDL 后，会把更新后的表结构转送给 DM-master。DM-worker 会追踪各分表当前的表结构，DM-master 合并成可兼容来自每个分表 DML 的合成结构，然后把与此对应的 DDL 迁移到下游；对于 DML 会直接迁移到下游。

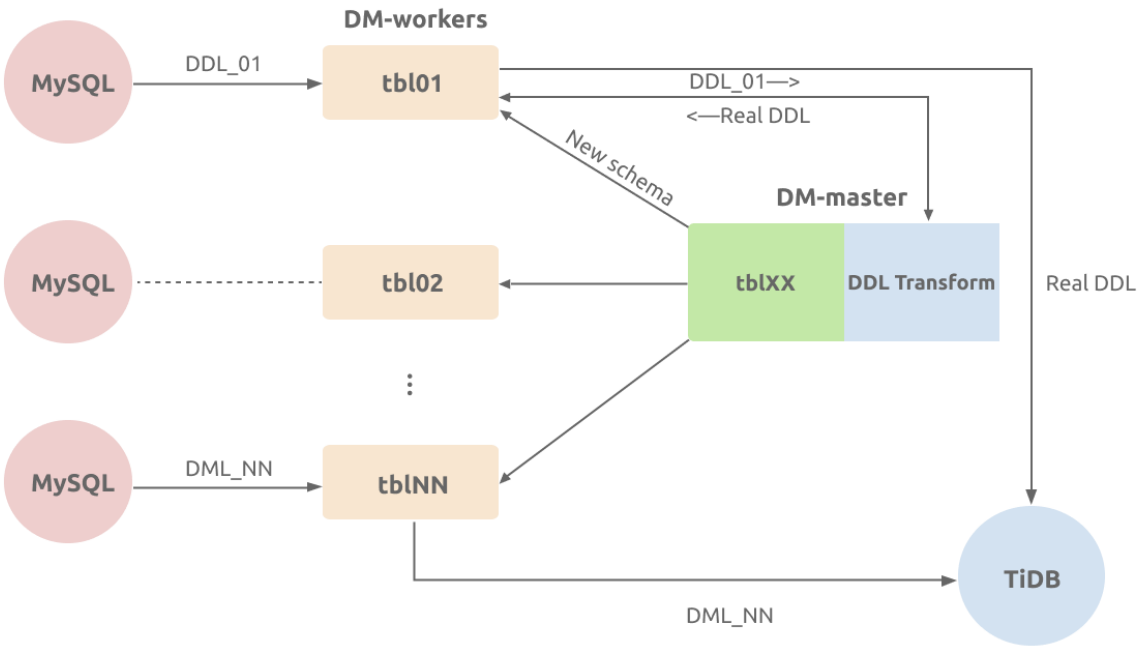


图 282: optimistic-ddl-flow

例子

例如上游 MySQL 有三个分表（`tbl00`, `tbl01` 以及 `tbl02`），使用 DM 迁移到下游 TiDB 的 `tbl` 表中，如下图所示：

tbl00

ID	Name
1	Sarah
5	Sophia

tbl01

ID	Name
12	Paul
16	Jessica
19	Shaun

tbl02

ID	Name
23	Bob
24	Ben

tbl

ID	Name
1	Sarah
5	Sophia
12	Paul
16	Jessica
19	Shaun
23	Bob
24	Ben

图 283: optimistic-ddl-example-1

在上游增加一列 Level：

```
ALTER TABLE `tbl00` ADD COLUMN `Level` INT;
```

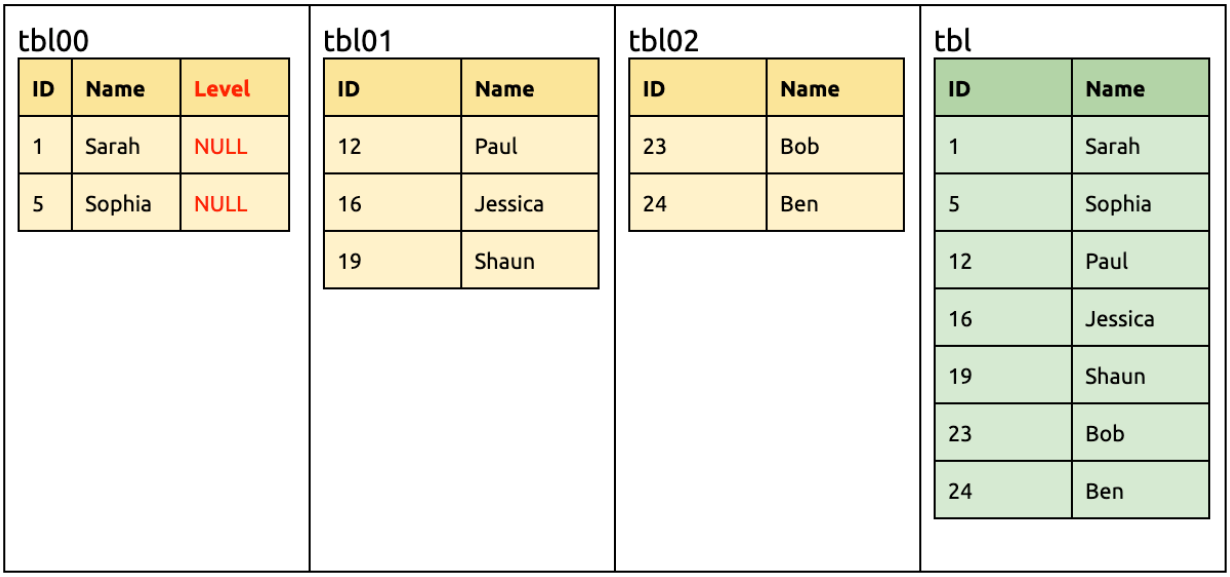


图 284: optimistic-ddl-example-2

此时下游 TiDB 要准备接受来自 tbl00 有 Level 的 DML、以及来自 tbl01 和 tbl02 没有 Level 的 DML。

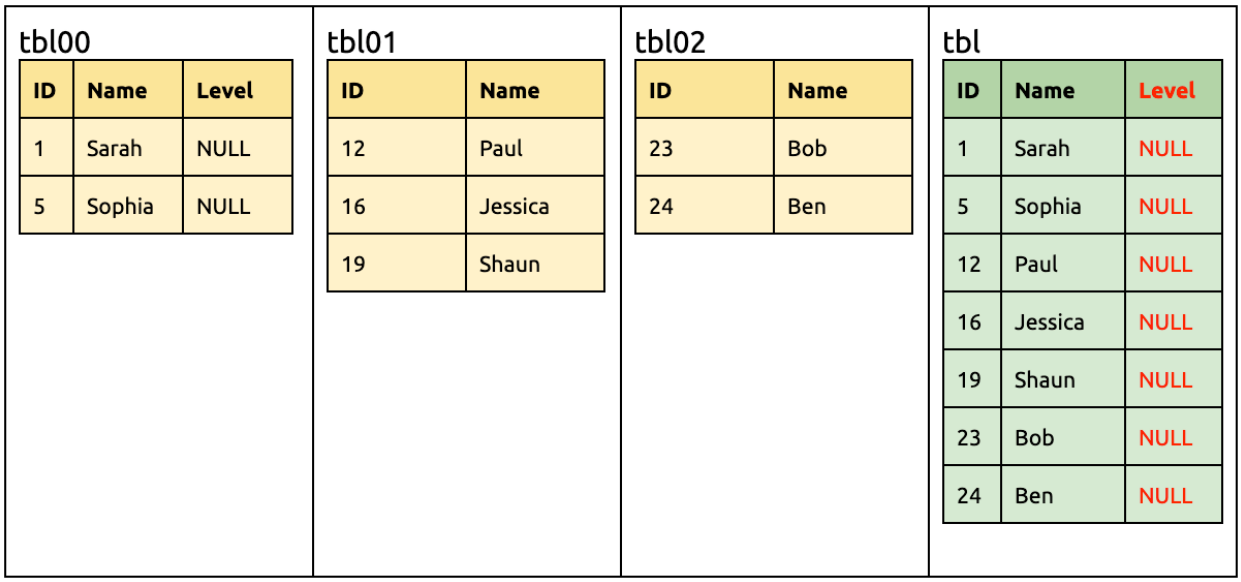


图 285: optimistic-ddl-example-3

这时候如下的 DML 无需修改就可以迁移到下游：

```
UPDATE `tbl00` SET `Level` = 9 WHERE `ID` = 1;
INSERT INTO `tbl02` (`ID`, `Name`) VALUES (27, 'Tony');
```

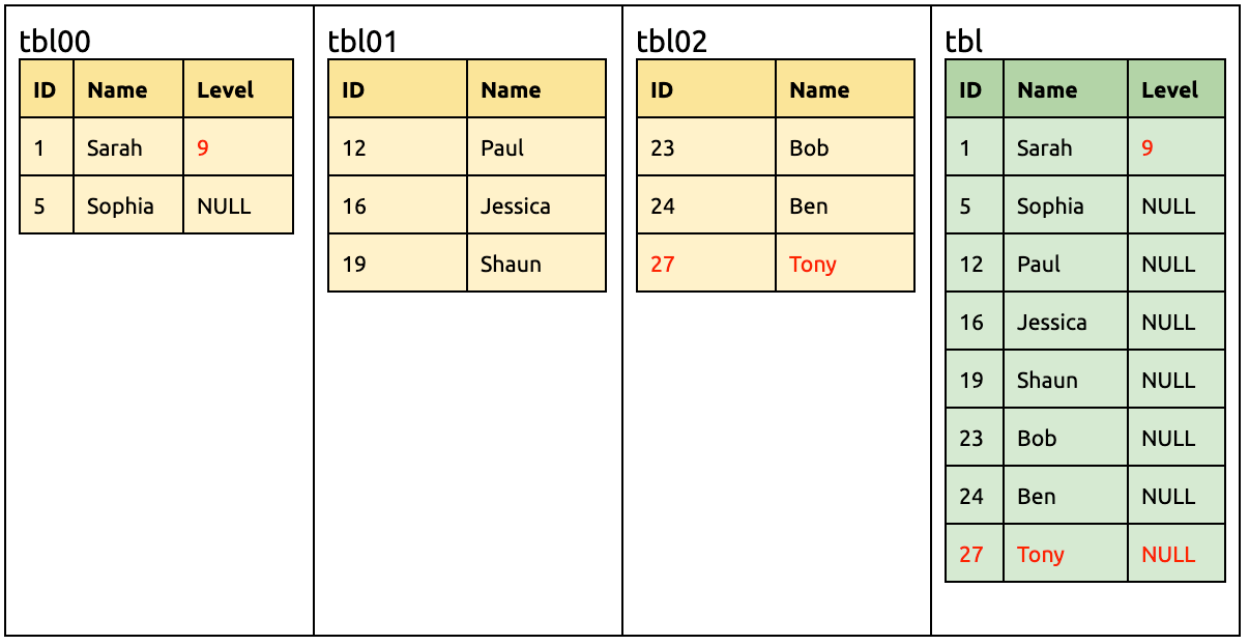


图 286: optimistic-ddl-example-4

在 `tbl01` 同样增加一列 `Level`：

```
ALTER TABLE `tbl01` ADD COLUMN `Level` INT;
```

tbl00

ID	Name	Level
1	Sarah	9
5	Sophia	NULL

tbl01

ID	Name	Level
12	Paul	NULL
16	Jessica	NULL
19	Shaun	NULL

tbl02

ID	Name
23	Bob
24	Ben
27	Tony

tbl

ID	Name	Level
1	Sarah	9
5	Sophia	NULL
12	Paul	NULL
16	Jessica	NULL
19	Shaun	NULL
23	Bob	NULL
24	Ben	NULL
27	Tony	NULL

图 287: optimistic-ddl-example-5

此时下游已经有相同的 Level 列了，所以 DM-master 比较表结构之后不做任何操作。

在 `tbl01` 删除一列 `Name`：

```
ALTER TABLE `tbl01` DROP COLUMN `Name`;
```

tbl00

ID	Name	Level
1	Sarah	9
5	Sophia	NULL

tbl01

ID	Level
12	NULL
16	NULL
19	NULL

tbl02

ID	Name
23	Bob
24	Ben
27	Tony

tbl

ID	Name	Level
1	Sarah	9
5	Sophia	NULL
12	Paul	NULL
16	Jessica	NULL
19	Shaun	NULL
23	Bob	NULL
24	Ben	NULL
27	Tony	NULL

图 288: optimistic-ddl-example-6

此时下游仍需要接收来自 `tbl00` 和 `tbl02` 含 `Name` 的 DML 语句，因此不会立刻删除该列。

同样，各种 DML 仍可直接迁移到下游：

```
INSERT INTO `tbl01` (`ID`, `Level`) VALUES (15, 7);
UPDATE `tbl00` SET `Level` = 5 WHERE `ID` = 5;
```

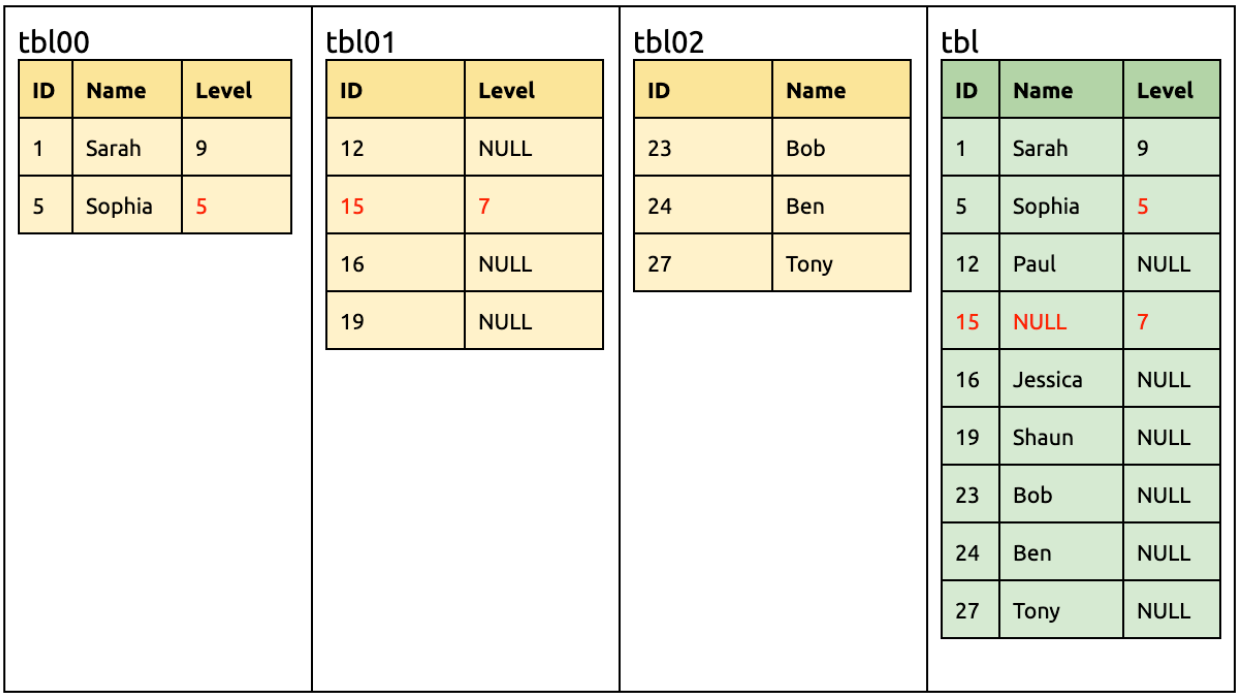


图 289: optimistic-ddl-example-7

在 `tbl02` 增加一列 `Level`：

```
ALTER TABLE `tbl02` ADD COLUMN `Level` INT;
```

tbl00

ID	Name	Level
1	Sarah	9
5	Sophia	5

tbl01

ID	Level
12	NULL
15	7
16	NULL
19	NULL

tbl02

ID	Name	Level
23	Bob	NULL
24	Ben	NULL
27	Tony	NULL

tbl

ID	Name	Level
1	Sarah	9
5	Sophia	5
12	Paul	NULL
15	NULL	7
16	Jessica	NULL
19	Shaun	NULL
23	Bob	NULL
24	Ben	NULL
27	Tony	NULL

图 290: optimistic-ddl-example-8

此时所有分表都已有 `Level` 列。

在 `tbl00` 和 `tbl02` 各删除一列 `Name`：

```
ALTER TABLE `tbl00` DROP COLUMN `Name`;
ALTER TABLE `tbl02` DROP COLUMN `Name`;
```

tbl00

ID	Level
1	9
5	5

tbl01

ID	Level
12	NULL
15	7
16	NULL
19	NULL

tbl02

ID	Level
23	NULL
24	NULL
27	NULL

tbl

ID	Name	Level
1	Sarah	9
5	Sophia	5
12	Paul	NULL
15	NULL	7
16	Jessica	NULL
19	Shaun	NULL
23	Bob	NULL
24	Ben	NULL
27	Tony	NULL

图 291: optimistic-ddl-example-9

到此步 Name 列也从所有分表消失了，所以可以安全从下游移除：

```
ALTER TABLE `tbl` DROP COLUMN `Name`;
```

tbl00

ID	Level
1	9
5	5

tbl01

ID	Level
12	NULL
15	7
16	NULL
19	NULL

tbl02

ID	Level
23	NULL
24	NULL
27	NULL

tbl

ID	Level
1	9
5	5
12	NULL
15	7
16	NULL
19	NULL
23	NULL
24	NULL
27	NULL

图 292: optimistic-ddl-example-10

12.8.5.1.4 手动处理 Sharding DDL Lock

DM (Data Migration) 使用 sharding DDL lock 来确保分库分表的 DDL 操作可以正确执行。绝大多数情况下，该锁定机制可自动完成；但在部分异常情况发生时，需要使用 `unlock-ddl-lock` 手动处理异常的 DDL lock。

> 注意：
>
> - 本文档只适用于悲观协调模式下 sharding DDL lock 的处理。
> - 本文档的命令在交互模式中进行，因此在以下命令示例中未添加转义字符。在命令行模式中，你需要添加转义字符，防止报错。
> - 不要轻易使用 `unlock-ddl-lock` 命令，除非完全明确当前场景下使用这些命令可能会造成的影响，并能接受这些影响。
> - 在手动处理异常的 DDL lock 前，请确保已经了解 DM 的分库分表合并迁移原理。

命令介绍

`show-ddl-locks`

该命令用于查询当前 DM-master 上存在的 DDL lock 信息。

命令示例

```
show-ddl-locks [--source=mysql-replica-01] [task-name | task-file]
```

参数解释

- source：
  - flag 参数，string，--source，可选
  - 不指定时，查询所有 MySQL source 相关的 lock 信息；指定时，仅查询与该 MySQL source 相关的 lock 信息，可重复多次指定
- task-name | task-file：
  - 非 flag 参数，string，可选
  - 不指定时，查询与所有任务相关的 lock 信息；指定时，仅查询特定任务相关的 lock 信息

返回结果示例

```
show-ddl-locks test
```

```
{
    "result": true,                                     # 查询 lock 操作本身是否成功
    "msg": "",                                          # 查询 lock
        ↪ 操作失败时的原因或其它描述信息（如不存在任务 lock）
    "locks": [                                          # 当前存在的 lock 信息列表
        {
            "ID": "test-`shard_db`.`shard_table`",      # lock 的 ID 标识，当前由任务名与 DDL
                ↪ 对应的 schema/table 信息组成
            "task": "test",                             # lock 所属的任务名
            "mode": "pessimistic"                       # shard DDL 协调模式，可为悲观模式 "
                ↪ pessimistic" 或乐观模式 "optimistic"
            "owner": "mysql-replica-01",                # lock 的 owner（
                ↪ 在悲观模式时为第一个遇到该 DDL 的 source ID），在乐观模式时总为空
            "DDLs": [                                   # 在悲观模式时为 lock 对应的 DDL 列表
                ↪ ，在乐观模式时总为空
                "USE `shard_db`; ALTER TABLE `shard_db`.`shard_table` DROP COLUMN `c2`;"
            ],
            "synced": [                                 # 已经收到对应 MySQL 实例内所有分表
                ↪ DDL 的 source 列表
                "mysql-replica-01"
            ],
            "unsynced": [                               # 尚未收到对应 MySQL 实例内所有分表
                ↪ DDL 的 source 列表
                "mysql-replica-02"
            ]
        }
    ]
}
```

unlock-ddl-lock

该命令用于主动请求 DM-master 解除指定的 DDL lock，包括的操作：请求 owner 执行 DDL 操作，请求其他非 owner 的 DM-worker 跳过 DDL 操作，移除 DM-master 上的 lock 信息。

> **注意：**
>
> unlock-ddl-lock 当前仅对悲观协调模式 (pessimistic) 下产生的 lock 有效。

命令示例

```
unlock-ddl-lock [--owner] [--force-remove] <lock-ID>
```

参数解释

- owner：
  - flag 参数，string，--owner，可选
  - 不指定时，请求默认的 owner（show-ddl-locks 返回结果中的 owner）执行 DDL 操作；指定时，请求该 MySQL source（替代默认的 owner）执行 DDL 操作
  - 除非原 owner 已经从集群中移除，否则不应该指定新的 owner
- force-remove：
  - flag 参数，boolean，--force-remove，可选
  - 不指定时，仅在 owner 执行 DDL 成功时移除 lock 信息；指定时，即使 owner 执行 DDL 失败也强制移除 lock 信息（此后将无法再次查询或操作该 lock）
- lock-ID：
  - 非 flag 参数，string，必选
  - 指定需要执行 unlock 操作的 DDL lock ID（即 show-ddl-locks 返回结果中的 ID）

返回结果示例

```
unlock-ddl-lock test-`shard_db`.`shard_table`
```

```
{
    "result": true,                                  # unlock lock 操作是否成功
    "msg": "",                                       # unlock lock 操作失败时的原因
}
```

支持场景

目前，使用 unlock-ddl-lock 命令仅支持处理以下两种 sharding DDL lock 异常情况。

场景一：部分 MySQL source 被移除

Lock 异常原因

在 DM-master 尝试自动 unlock sharding DDL lock 之前，需要等待所有 MySQL source 的 sharding DDL events 全部到达（详见分库分表合并迁移原理）。如果 sharding DDL 已经在迁移过程中，同时有部分 MySQL source 被移除，且不

再计划重新加载它们（按业务需求移除了这部分 MySQL source），则会由于永远无法等齐所有的 DDL 而造成 lock 无法自动 unlock。

手动处理示例

假设上游有 MySQL-1（mysql-replica-01）和 MySQL-2（mysql-replica-02）两个实例，其中 MySQL-1 中有 shard_db_1.shard_table_1 和 shard_db_1.shard_table_2 两个表，MySQL-2 中有 shard_db_2.shard_table_1 和 shard_db_2.shard_table_2 两个表。现在需要将这 4 个表合并后迁移到下游 TiDB 的 shard_db.shard_table 表中。

初始表结构如下：

```
SHOW CREATE TABLE shard_db_1.shard_table_1;
```

```
+---------------+----------------------------------------+
| Table         | Create Table                           |
+---------------+----------------------------------------+
| shard_table_1 | CREATE TABLE `shard_table_1` (
  `c1` int(11) NOT NULL,
  PRIMARY KEY (`c1`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 |
+---------------+----------------------------------------+
```

上游分表将执行以下 DDL 语句变更表结构：

```
ALTER TABLE shard_db_*.shard_table_* ADD COLUMN c2 INT;
```

MySQL 及 DM 操作与处理流程如下：

1. mysql-replica-01 对应的两个分表执行了对应的 DDL 操作进行表结构变更。

   ```
   ALTER TABLE shard_db_1.shard_table_1 ADD COLUMN c2 INT;
   ```

   ```
   ALTER TABLE shard_db_1.shard_table_2 ADD COLUMN c2 INT;
   ```

2. DM-worker 接受到 mysql-replica-01 两个分表的 DDL 之后，将对应的 DDL 信息发送给 DM-master，DM-master 创建相应的 DDL lock。

3. 使用 show-ddl-lock 查看当前的 DDL lock 信息。

   ```
   show-ddl-locks test
   ```

   ```
   {
       "result": true,
       "msg": "",
       "locks": [
           {
               "ID": "test-`shard_db`.`shard_table`",
               "task": "test",
   ```

```
            "mode": "pessimistic"
            "owner": "mysql-replica-01",
            "DDLs": [
                "USE `shard_db`; ALTER TABLE `shard_db`.`shard_table` ADD COLUMN `c2` int
                    ↪ (11);"
            ],
            "synced": [
                "mysql-replica-01"
            ],
            "unsynced": [
                "mysql-replica-02"
            ]
        }
    ]
}
```

4. 由于业务需要，`mysql-replica-02` 对应的数据不再需要迁移到下游 TiDB，对 `mysql-replica-02` 执行了移除操作。

5. DM-master 上 ID 为 `test-`shard_db`.`shard_table`` 的 lock 无法等到 `mysql-replica-02` 的 DDL 操作信息。`show-ddl-locks` 返回的 unsynced 中一直包含 `mysql-replica-02` 的信息。

6. 使用 `unlock-dll-lock` 来请求 DM-master 主动 unlock 该 DDL lock。

    - 如果 DDL lock 的 owner 也已经被移除，可以使用 `--owner` 参数指定其他 MySQL source 作为新 owner 来执行 DDL。
    - 当存在任意 MySQL source 报错时，`result` 将为 `false`，此时请仔细检查各 MySQL source 的错误是否是预期可接受的。

    ```
    unlock-ddl-lock test-`shard_db`.`shard_table`
    ```

    ```
    {
        "result": true,
        "msg": ""
    ```

7. 使用 `show-dd-locks` 确认 DDL lock 是否被成功 unlock。

```
show-ddl-locks test
```

```
{
    "result": true,
    "msg": "no DDL lock exists",
    "locks": [
    ]
}
```

8. 查看下游 TiDB 中的表结构是否变更成功。

```
SHOW CREATE TABLE shard_db.shard_table;
```

```
+-------------+-----------------------------------------------------+
| Table       | Create Table                                        |
+-------------+-----------------------------------------------------+
| shard_table | CREATE TABLE `shard_table` (
  `c1` int(11) NOT NULL,
  `c2` int(11) DEFAULT NULL,
  PRIMARY KEY (`c1`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin |
+-------------+-----------------------------------------------------+
```

9. 使用 `query-status` 确认迁移任务是否正常。

手动处理后的影响

使用 `unlock-ddl-lock` 手动执行 unlock 操作后，由于该任务的配置信息中仍然包含了已下线的 MySQL source，如果不进行处理，则当下次 sharding DDL 到达时，仍会出现 lock 无法自动完成迁移的情况。

因此，在手动解锁 DDL lock 后，需要再执行以下操作：

1. 使用 `stop-task` 停止运行中的任务。
2. 更新任务配置文件，将已下线 MySQL source 对应的信息从配置文件中移除。
3. 使用 `start-task` 及新任务配置文件重新启动任务。

> 注意：
>
> 在 `unlock-ddl-lock` 之后，如果已下线的 MySQL source 重新加载并尝试对其中的分表进行数据迁移，则会由于数据与下游的表结构不匹配而发生错误。

场景二：unlock 过程中部分 DM-worker 异常停止或网络中断

Lock 异常原因

在 DM-master 收到所有 DM-worker 的 DDL 信息后，执行自动 unlock DDL lock 的操作主要包括以下步骤：

1. 请求 lock owner 执行 DDL 操作，并更新对应分表的 checkpoint。
2. 在 owner 执行 DDL 操作成功后，移除 DM-master 上保存的 DDL lock 信息。
3. 在 owner 执行 DDL 操作成功后，请求其他所有非 owner 跳过 DDL 操作并更新对应分表的 checkpoint。
4. DM-master 在所有 owner/非 owner 操作成功后，移除对应的 DDL lock 信息。

上述 unlock DDL lock 的操作不是原子的。如果非 owner 跳过 DDL 操作成功后，所在的 DM-worker 异常停止或与下游 TiDB 发生网络异常，造成无法成功更新 checkpoint。

当非 owner 对应的 MySQL source 恢复数据迁移时，会尝试请求 DM-master 重新协调异常发生前已经协调过的 DDL、且永远无法等到其他 MySQL source 的对应 DDL，造成该 DDL 操作对应 lock 的自动 unlock。

手动处理示例

仍然假设是部分 MySQL source 被移除 示例中的上下游表结构及合表迁移需求。

当在 DM-master 自动执行 unlock 操作的过程中，owner（`mysql-replica-01`）成功执行了 DDL 操作且开始继续进行后续迁移，但在请求非 owner（`mysql-replica-02`）跳过 DDL 操作的过程中，由于对应的 DM-worker 发生了重启在跳过 DDL 后未能更新 checkpoint。

`mysql-replica-02` 对应的数据迁移子任务恢复后，将在 DM-master 上创建一个新的 lock，但其他 MySQL source 此时已经执行或跳过 DDL 操作并在进行后续迁移。

处理流程如下：

1. 使用 `show-ddl-locks` 确认 DM-master 上存在该 DDL 操作对应的 lock。

   应该仅有 `mysql-replica-02` 处于 `synced` 状态：

```
show-ddl-locks
```

```
{
    "result": true,
    "msg": "",
    "locks": [
        {
            "ID": "test-`shard_db`.`shard_table`",
            "task": "test",
            "mode": "pessimistic"
            "owner": "mysql-replica-02",
            "DDLs": [
                "USE `shard_db`; ALTER TABLE `shard_db`.`shard_table` ADD COLUMN `c2` int
                    ↪ (11);"
            ],
            "synced": [
                "mysql-replica-02"
            ],
            "unsynced": [
                "mysql-replica-01"
            ]
        }
    ]
}
```

2. 使用 `unlock-ddl-lock` 请求 DM-master unlock 该 lock。

   - Lock 过程中会尝试再次向下游执行该 DDL 操作（重启前的原 owner 已向下游执行过该 DDL 操作），需要确保该 DDL 操作可被多次执行。

```
unlock-ddl-lock test-`shard_db`.`shard_table`
```

```
{
    "result": true,
    "msg": "",
}
```

3. 使用 show-ddl-locks 确认 DDL lock 是否被成功 unlock。
4. 使用 query-status 确认迁移任务是否正常。

手动处理后的影响

手动 unlock sharding DDL lock 后，后续的 sharding DDL 将可以自动正常迁移。

12.8.5.2 迁移使用 GH-ost/PT-osc 的源数据库

在生产业务中执行 DDL 时，产生的锁表操作会一定程度阻塞数据库的读取或者写入。为了把对读写的影响降到最低，用户往往会选择 online DDL 工具执行 DDL。常见的 Online DDL 工具有 gh-ost 和 pt-osc。

在使用 DM 完成 MySQL 到 TiDB 的数据迁移时，可以开启online-ddl配置，实现 DM 工具与 gh-ost 或 pt-osc 的协同。关于如何开启 online-ddl配置及开启该配置后的工作流程，请参考上游使用 pt-osc/gh-ost 工具的持续同步场景。本文仅介绍 DM 与 online DDL 工具协作的细节。

12.8.5.2.1 DM 与 online DDL 工具协作细节

DM 与 online DDL 工具 gh-ost 和 pt-osc 在实现 online-schema-change 过程中的协作细节如下。

online-schema-change: gh-ost

gh-ost 在实现 online-schema-change 的过程会产生 3 种 table：

- gho：用于应用 DDL，待 gho 表中数据迁移到与 origin table 一致后，通过 rename 的方式替换 origin table。
- ghc：用于存放 online-schema-change 相关的信息。
- del：对 origin table 执行 rename 操作而生成。

DM 在迁移过程中会把上述 table 分成 3 类：

- ghostTable : _*_gho
- trashTable : _*_ghc、 _*_del
- realTable : 执行 online-ddl 的 origin table

gh-ost 涉及的主要 SQL 以及 DM 的处理：

1. 创建 _ghc 表：

```
Create /* gh-ost */ table `test`.`_test4_ghc` (
                        id bigint auto_increment,
                        last_update timestamp not null DEFAULT CURRENT_TIMESTAMP ON UPDATE
                            ↪ CURRENT_TIMESTAMP,
                        hint varchar(64) charset ascii not null,
                        value varchar(4096) charset ascii not null,
                        primary key(id),
                        unique key hint_uidx(hint)
                ) auto_increment=256 ;
```

DM：不执行 _test4_ghc 的创建操作。

2. 创建 _gho 表：

```
Create /* gh-ost */ table `test`.`_test4_gho` like `test`.`test4` ;
```

DM：不执行 _test4_gho 的创建操作，根据 ghost_schema、ghost_table 以及 dm_worker 的 server_id，删除下游 dm_meta.{task_name}_onlineddl 的记录，清理内存中的相关信息。

```
DELETE FROM dm_meta.{task_name}_onlineddl WHERE id = {server_id} and ghost_schema = {
    ↪ ghost_schema} and ghost_table = {ghost_table};
```

3. 在 _gho 表应用需要执行的 DDL：

```
Alter /* gh-ost */ table `test`.`_test4_gho` add column cl1 varchar(20) not null ;
```

DM：不执行 _test4_gho 的 DDL 操作，而是把该 DDL 记录到 dm_meta.{task_name}_onlineddl 以及内存中。

```
REPLACE INTO dm_meta.{task_name}_onlineddl (id, ghost_schema , ghost_table , ddls) VALUES
    ↪ (......);
```

4. 往 _ghc 表写入数据，以及往 _gho 表同步 origin table 的数据：

```
Insert /* gh-ost */ into `test`.`_test4_ghc` values (......);

Insert /* gh-ost `test`.`test4` */ ignore into `test`.`_test4_gho` (`id`, `date`, `
    ↪ account_id`, `conversion_price`, `ocpc_matched_conversions`, `ad_cost`, `cl2`)
  (select `id`, `date`, `account_id`, `conversion_price`, `ocpc_matched_conversions`, `
      ↪ ad_cost`, `cl2` from `test`.`test4` force index (`PRIMARY`)
    where (((`id` > _binary'1') or ((`id` = _binary'1'))) and ((`id` < _binary'2') or ((`id`
        ↪  = _binary'2')))) lock in share mode
  )   ;
```

DM：只要不是 realtable 的 DML 全部不执行。

5. 数据同步完成后 origin table 与 _gho 一起改名，完成 online DDL 操作：

```
Rename /* gh-ost */ table `test`.`test4` to `test`.`_test4_del`, `test`.`_test4_gho` to `
    ↪ test`.`test4`;
```

DM 执行以下两个操作:

- 把 rename 语句拆分成两个 SQL：

```
rename test.test4 to test._test4_del;
rename test._test4_gho to test.test4;
```

- 不执行 `rename to _test4_del`。当要执行 `rename ghost_table to origin table` 的时候，并不执行 rename 语句，而是把步骤 3 记录在内存中的 DDL 读取出来，然后把 ghost_table、ghost_schema 替换为 origin_table 以及对应的 schema，再执行替换后的 DDL。

```
alter table test._test4_gho add column c11 varchar(20) not null;
--替换为
alter table test.test4 add column c11 varchar(20) not null;
```

> 注意：
>
> 具体 gh-ost 的 SQL 会根据工具执行时所带的参数而变化。本文只列出主要的 SQL，具体可以参考 gh-ost 官方文档。

online-schema-change: pt

pt-osc 在实现 online-schema-change 的过程会产生 2 种 table：

- new：用于应用 DDL，待表中数据同步到与 origin table 一致后，再通过 rename 的方式替换 origin table。
- `old`：对 origin table 执行 rename 操作后生成。
- 3 种 trigger：`pt_osc\_\*\_ins`、`pt_osc\_\*\_upd`、`pt_osc\_\*\_del`，用于在 pt_osc 过程中，同步 origin table 新产生的数据到 new。

DM 在迁移过程中会把上述 table 分成 3 类：

- ghostTable : `\_\*\_new`
- trashTable : `\_\*\_old`
- realTable : 执行的 online-ddl 的 origin table

pt-osc 主要涉及的 SQL 以及 DM 的处理：

1. 创建 _new 表：

```
CREATE TABLE `test`.`_test4_new` (id int(11) NOT NULL AUTO_INCREMENT,
date date DEFAULT NULL, account_id bigint(20) DEFAULT NULL, conversion_price decimal(20,3)
    ↪ DEFAULT NULL,  ocpc_matched_conversions bigint(20) DEFAULT NULL, ad_cost decimal
    ↪ (20,3) DEFAULT NULL,cl2 varchar(20) COLLATE utf8mb4_bin NOT NULL,cl1 varchar(20)
    ↪ COLLATE utf8mb4_bin NOT NULL,PRIMARY KEY (id) ) ENGINE=InnoDB AUTO_INCREMENT=3
    ↪ DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_bin ;
```

DM: 不执行 _test4_new 的创建操作。根据 ghost_schema、ghost_table 以及 dm_worker 的 server_id，删除下游 dm_meta.{task_name}_onlineddl 的记录，清理内存中的相关信息。

```
DELETE FROM dm_meta.{task_name}_onlineddl WHERE id = {server_id} and ghost_schema = {
    ↪ ghost_schema} and ghost_table = {ghost_table};
```

2. 在 _new 表上执行 DDL：

```
ALTER TABLE `test`.`_test4_new` add column c3 int;
```

DM: 不执行 _test4_new 的 DDL 操作，而是把该 DDL 记录到 dm_meta.{task_name}_onlineddl 以及内存中。

```
REPLACE INTO dm_meta.{task_name}_onlineddl (id, ghost_schema , ghost_table , ddls) VALUES
    ↪ (......);
```

3. 创建用于同步数据的 3 个 Trigger：

```
CREATE TRIGGER `pt_osc_test_test4_del` AFTER DELETE ON `test`.`test4` ...... ;
CREATE TRIGGER `pt_osc_test_test4_upd` AFTER UPDATE ON `test`.`test4` ...... ;
CREATE TRIGGER `pt_osc_test_test4_ins` AFTER INSERT ON `test`.`test4` ...... ;
```

DM: 不执行 TiDB 不支持的相关 Trigger 操作。

4. 往 _new 表同步 origin table 的数据：

```
INSERT LOW_PRIORITY IGNORE INTO `test`.`_test4_new` (`id`, `date`, `account_id`, `
    ↪ conversion_price`, `ocpc_matched_conversions`, `ad_cost`, `cl2`, `cl1`) SELECT `id`,
    ↪  `date`, `account_id`, `conversion_price`, `ocpc_matched_conversions`, `ad_cost`, `
    ↪ cl2`, `cl1` FROM `test`.`test4` LOCK IN SHARE MODE /*pt-online-schema-change 3227
    ↪ copy table*/
```

DM: 只要不是 realTable 的 DML 全部不执行。

5. 数据同步完成后 origin table 与 _new 一起改名，完成 online DDL 操作：

```
RENAME TABLE `test`.`test4` TO `test`.`_test4_old`, `test`.`_test4_new` TO `test`.`test4`
```

DM 执行以下两个操作:

- 把 rename 语句拆分成两个 SQL。

  ```
  sql rename test.test4 to test._test4_old; rename test._test4_new to test.test4;
  ```

- 不执行 `rename to _test4_old`。当要执行 `rename ghost_table to origin table` 的时候，并不执行 rename，而是把步骤 2 记录在内存中的 DDL 读取出来，然后把 ghost_table、ghost_schema 替换为 origin_table 以及对应的 schema，再执行替换后的 DDL。

  ```
  sql  ALTER TABLE `test`.`_test4_new` add column c3 int; --替换为 ALTER TABLE `test`.`
  ↪ test4` add column c3 int;
  ```

6. 删除 `_old` 表以及 online DDL 的 3 个 Trigger：

```
DROP TABLE IF EXISTS `test`.`_test4_old`;
DROP TRIGGER IF EXISTS `pt_osc_test_test4_del` AFTER DELETE ON `test`.`test4` ...... ;
DROP TRIGGER IF EXISTS `pt_osc_test_test4_upd` AFTER UPDATE ON `test`.`test4` ...... ;
DROP TRIGGER IF EXISTS `pt_osc_test_test4_ins` AFTER INSERT ON `test`.`test4` ...... ;
```

DM: 不执行 `_test4_old` 以及 Trigger 的删除操作。

> 注意：
>
> 具体 pt-osc 的 SQL 会根据工具执行时所带的参数而变化。本文只列出主要的 SQL，具体可以参考 pt-osc 官方文档。

12.8.5.3 下游存在更多列的迁移场景

本文档介绍数据同步时，下游存在更多列的迁移场景需要的注意事项。具体迁移操作可参考已有数据迁移场景：

- 从小数据量 MySQL 迁移数据到 TiDB
- 从大数据量 MySQL 迁移数据到 TiDB
- 从小数据量分库分表 MySQL 合并迁移数据到 TiDB
- 从大数据量分库分表 MySQL 合并迁移数据到 TiDB

12.8.5.3.1 使用 DM 迁移至存在更多列的下游

DM 同步上游的 binlog 时，会尝试使用下游当前的表结构来解析 binlog 并生成相应的 DML 语句。如果上游的 binlog 里数据表的列数与下游表结构的列数不一致，则会产生如下错误：

```
"errors": [
    {
        "ErrCode": 36027,
        "ErrClass": "sync-unit",
        "ErrScope": "internal",
        "ErrLevel": "high",
        "Message": "startLocation: [position: (mysql-bin.000001, 2022), gtid-set:09bec856-ba95-11
            ↪ ea-850a-58f2b4af5188:1-9 ], endLocation: [position: (mysql-bin.000001, 2022),
            ↪ gtid-set: 09bec856-ba95-11ea-850a-58f2b4af5188:1-9]: gen insert sqls failed,
            ↪ schema: log, table: messages: Column count doesn't match value count: 3 (columns)
            ↪  vs 2 (values)",
```

```
        "RawCause": "",
        "Workaround": ""
    }
]
```

例如上游表结构为：

```
#### 上游表结构
CREATE TABLE `messages` (
  `id` int(11) NOT NULL,
  PRIMARY KEY (`id`)
)
```

例如下游表结构为：

```
#### 下游表结构
CREATE TABLE `messages` (
  `id` int(11) NOT NULL,
  `message` varchar(255) DEFAULT NULL, # 下游比上游多出的列。
  PRIMARY KEY (`id`)
)
```

当 DM 尝试使用下游表结构解析上游产生的 binlog event 时，DM 会报出上述 `Column count doesn't match` 错误。

此时，你可以使用 `operate-schema` 命令来为数据源中需要迁移的表指定表结构，表结构需要对应 DM 将要开始同步的 binlog event 的数据。如果你在进行分表合并的数据迁移，那么需要为每个分表按照如下步骤在 DM 中设置用于解析 binlog event 的表结构。具体操作为：

1. 在 DM 中，新建一个 `.sql` 文件，并将上游表结构对应的 `CREATE TABLE` 语句添加到该文件。例如，将以下表结构保存到 `log.messages.sql` 中。

    ```
    # 上游表结构
    CREATE TABLE `messages` (
    `id` int(11) NOT NULL,
    PRIMARY KEY (`id`)
    )
    ```

2. 使用 `operate-schema` 命令为数据源中需要迁移的表设置表结构（此时数据迁移任务应该由于上述 `Column count doesn't match` 错误而处于 Paused 状态）。

    ```
    tiup dmctl --master-addr ${advertise-addr} operate-schema set -s ${source-id} ${task-name} -
        ↪ d ${database-name} -t ${table-name} ${schema-file}
    ```

    该命令中的参数描述如下：

参数	描述

参数	描述
–master-addr	指定 dm-ctl 要连接的集群的任意 DM-master 节点的 $ ↪ { ↪ advertise ↪ - ↪ addr ↪ } ↪ 。 $ ↪ { ↪ advertise ↪ - ↪ addr ↪ } ↪ 表示 DM-master 向外界宣告的地址。

参数	描述
operate-schema set	手动设置 schema 信息

参数	描述
-s	指定 source。$ ↪ { ↪ source ↪ - ↪ id ↪ } ↪ 表示 MySQL 数据源 ID。$ ↪ { ↪ task ↪ - ↪ name ↪ } ↪ 表示数据同步任务配置文件 task ↪ . ↪ yaml ↪ 中定义的同步任务

参数	描述
-d	指定 database。$ ↪ { ↪ database ↪ - ↪ name ↪ } ↪ 表示上游数据库名。

参数	描述
-t	指定 table。 ${table-name} 表示上游数据表名。 ${schema-file} 表示将被设置的表结构文件。

例如：

```
tiup dmctl --master-addr 172.16.10.71:8261 operate-schema set -s mysql-01 task-test -d log -
    ↪ t message log.message.sql
```

3. 使用 `resume-task` 命令恢复处于 Paused 状态的同步任务。

```
tiup dmctl --master-addr ${advertise-addr} resume-task ${task-name}
```

4. 使用 `query-status` 命令确认数据迁移任务是否运行正常。

```
tiup dmctl --master-addr ${advertise-addr} query-status resume-task ${task-name}
```

12.8.6 运维管理

12.8.6.1 集群版本升级

12.8.6.1.1 使用 TiUP 运维 DM 集群

本文介绍如何使用 TiUP 的 DM 组件运维 DM 集群。

如果你还未部署 DM 集群，可参考使用 TiUP 部署 DM 集群。

> **注意：**
>
> - 需要确保以下组件间端口可正常连通：
>   - 各 DM-master 节点间的 `peer_port`（默认为 8291）可互相连通。
>   - 各 DM-master 节点可连通所有 DM-worker 节点的 `port`（默认为 8262）。
>   - 各 DM-worker 节点可连通所有 DM-master 节点的 `port`（默认为 8261）。
>   - TiUP 节点可连通所有 DM-master 节点的 `port`（默认为 8261）。
>   - TiUP 节点可连通所有 DM-worker 节点的 `port`（默认为 8262）。

TiUP DM 组件的帮助信息如下：

```
tiup dm --help
```

```
Deploy a DM cluster for production

Usage:
  tiup dm [flags]
  tiup dm [command]

Available Commands:
  deploy      Deploy a DM cluster for production
  start       Start a DM cluster
```

```
  stop        Stop a DM cluster
  restart     Restart a DM cluster
  list        List all clusters
  destroy     Destroy a specified DM cluster
  audit       Show audit log of cluster operation
  exec        Run shell command on host in the dm cluster
  edit-config Edit DM cluster config
  display     Display information of a DM cluster
  reload      Reload a DM cluster's config and restart if needed
  upgrade     Upgrade a specified DM cluster
  patch       Replace the remote package with a specified package and restart the service
  scale-out   Scale out a DM cluster
  scale-in    Scale in a DM cluster
  import      Import an exist DM 1.0 cluster from dm-ansible and re-deploy 2.0 version
  help        Help about any command

Flags:
  -h, --help               help for tiup-dm
      --native-ssh         Use the native SSH client installed on local system instead of the
          ↪ build-in one.
      --ssh-timeout int    Timeout in seconds to connect host via SSH, ignored for operations
          ↪ that don't need an SSH connection. (default 5)
  -v, --version            version for tiup-dm
      --wait-timeout int   Timeout in seconds to wait for an operation to complete, ignored for
          ↪ operations that don't fit. (default 60)
  -y, --yes                Skip all confirmations and assumes 'yes'
```

查看集群列表

集群部署成功后，可以通过 `tiup dm list` 命令在集群列表中查看该集群：

```
tiup dm list
```

```
Name  User  Version  Path                                       PrivateKey
----  ----  -------  ----                                       ----------
prod-cluster  tidb  ${version}  /root/.tiup/storage/dm/clusters/test  /root/.tiup/storage/dm/
    ↪ clusters/test/ssh/id_rsa
```

启动集群

集群部署成功后，可以执行以下命令启动该集群。如果忘记了部署的集群名字，可以使用 `tiup dm list` 命令查看。

```
tiup dm start prod-cluster
```

检查集群状态

如果想查看集群中每个组件的运行状态，逐一登录到各个机器上查看显然很低效。因此，TiUP 提供了 `tiup dm display` 命令，用法如下：

```
tiup dm display prod-cluster
```

```
dm Cluster: prod-cluster
dm Version: ${version}
ID                  Role          Host          Ports      OS/Arch       Status     Data Dir
    ↪                             Deploy Dir
--                  ----          ----          -----      -------       ------     --------
    ↪                             ----------
172.19.0.101:9093   alertmanager  172.19.0.101  9093/9094  linux/x86_64  Up         /home/tidb/
    ↪ data/alertmanager-9093  /home/tidb/deploy/alertmanager-9093
172.19.0.101:8261   dm-master     172.19.0.101  8261/8291  linux/x86_64  Healthy|L  /home/tidb/
    ↪ data/dm-master-8261     /home/tidb/deploy/dm-master-8261
172.19.0.102:8261   dm-master     172.19.0.102  8261/8291  linux/x86_64  Healthy    /home/tidb/
    ↪ data/dm-master-8261     /home/tidb/deploy/dm-master-8261
172.19.0.103:8261   dm-master     172.19.0.103  8261/8291  linux/x86_64  Healthy    /home/tidb/
    ↪ data/dm-master-8261     /home/tidb/deploy/dm-master-8261
172.19.0.101:8262   dm-worker     172.19.0.101  8262       linux/x86_64  Free       /home/tidb/
    ↪ data/dm-worker-8262     /home/tidb/deploy/dm-worker-8262
172.19.0.102:8262   dm-worker     172.19.0.102  8262       linux/x86_64  Free       /home/tidb/
    ↪ data/dm-worker-8262     /home/tidb/deploy/dm-worker-8262
172.19.0.103:8262   dm-worker     172.19.0.103  8262       linux/x86_64  Free       /home/tidb/
    ↪ data/dm-worker-8262     /home/tidb/deploy/dm-worker-8262
172.19.0.101:3000   grafana       172.19.0.101  3000       linux/x86_64  Up         -
    ↪                                   /home/tidb/deploy/grafana-3000
172.19.0.101:9090   prometheus    172.19.0.101  9090       linux/x86_64  Up         /home/tidb/
    ↪ data/prometheus-9090    /home/tidb/deploy/prometheus-9090
```

Status 列用 Up 或者 Down 表示该服务是否正常。对于 DM-master 组件，同时可能会带有 |L 表示该 DM-master 是 Leader, 对于 DM-worker 组件，Free 表示当前 DM-worker 没有与上游绑定。

缩容节点

缩容即下线服务，最终会将指定的节点从集群中移除，并删除遗留的相关数据文件。

缩容操作进行时，内部对 DM-master、 DM-worker 组件的操作流程为：

1. 停止组件进程
2. 调用 DM-master 删除 member 的 API
3. 清除节点的相关数据文件

缩容命令的基本用法：

```
tiup dm scale-in <cluster-name> -N <node-id>
```

它需要指定至少两个参数，一个是集群名字，另一个是节点 ID。节点 ID 可以参考上一节使用 `tiup dm display` 命令获取。

比如想缩容 172.16.5.140 上的 DM-worker 节点（DM-master 的缩容类似），可以执行：

```
tiup dm scale-in prod-cluster -N 172.16.5.140:8262
```

扩容节点

扩容的内部逻辑与部署类似，TiUP DM 组件会先保证节点的 SSH 连接，在目标节点上创建必要的目录，然后执行部署并且启动服务。

例如，在集群 `prod-cluster` 中扩容一个 DM-worker 节点（DM-master 的扩容类似）：

1. 新建 scale.yaml 文件，添加新增的 woker 节点信息：

> 注意：
> 需要新建一个拓扑文件，文件中只写入扩容节点的描述信息，不要包含已存在的节点。
> 其他更多配置项（如：部署目录等）请参考 TiUP 配置参数模版。

```
---

worker_servers:
  - host: 172.16.5.140
```

2. 执行扩容操作。TiUP DM 根据 scale.yaml 文件中声明的端口、目录等信息在集群中添加相应的节点：

```
tiup dm scale-out prod-cluster scale.yaml
```

执行完成之后可以通过 `tiup dm display prod-cluster` 命令检查扩容后的集群状态。

滚动升级

> 注意：
> 从 v2.0.5 版本开始，dmctl 支持导出和导入集群的数据源和任务配置。
> 升级前，可使用 `config export` 命令导出集群的配置文件，升级后如需降级回退到旧版本，可重建旧集群后，使用 `config import` 导入之前的配置。
> 对于 v2.0.5 之前版本的集群，可使用 v2.0.5 及之后版本的 dmctl 导出和导入集群配置。
> 对于 v2.0.2 之后的版本，导入集群配置时暂不支持自动恢复 relay worker 相关配置，可手动执行 `start-relay` 命令开启 relay log。

滚动升级过程中尽量保证对前端业务透明、无感知，其中对不同节点有不同的操作。

升级操作

可使用 `tiup dm upgrade` 命令来升级集群。例如，以下示例将集群升级到 `${version}`，执行命令前，将 `${version}` 替换为实际需要的版本：

```
tiup dm upgrade prod-cluster `${version}`
```

更新配置

如果想要动态更新组件的配置，TiUP DM 组件为每个集群保存了一份当前的配置，如果想要编辑这份配置，则执行 `tiup dm edit-config <cluster-name>` 命令。例如：

```
tiup dm edit-config prod-cluster
```

然后 TiUP DM 组件会使用 vi 打开配置文件供编辑（如果你想要使用其他编辑器，请使用 EDITOR 环境变量自定义编辑器，例如 `export EDITOR=nano`），编辑完之后保存即可。此时的配置并没有应用到集群，如果想要让它生效，还需要执行：

```
tiup dm reload prod-cluster
```

该操作会将配置发送到目标机器，滚动重启集群，使配置生效。

更新组件

常规的升级集群可以使用 upgrade 命令，但是在某些场景下（例如 Debug），可能需要用一个临时的包替换正在运行的组件，此时可以用 patch 命令：

```
tiup dm patch --help
```

```
Replace the remote package with a specified package and restart the service

Usage:
  tiup dm patch <cluster-name> <package-path> [flags]

Flags:
  -h, --help                    help for patch
  -N, --node strings            Specify the nodes
      --overwrite               Use this package in the future scale-out operations
  -R, --role strings            Specify the role
      --transfer-timeout int    Timeout in seconds when transferring dm-master leaders (default
          ↪ 300)

Global Flags:
      --native-ssh         Use the native SSH client installed on local system instead of the
          ↪ build-in one.
      --ssh-timeout int    Timeout in seconds to connect host via SSH, ignored for operations
          ↪ that don't need an SSH connection. (default 5)
      --wait-timeout int   Timeout in seconds to wait for an operation to complete, ignored for
          ↪ operations that don't fit. (default 60)
  -y, --yes                Skip all confirmations and assumes 'yes'
```

例如，有一个 DM-master 的 hotfix 包放在 `/tmp/dm-master-hotfix.tar.gz`，如果此时想要替换集群上的所有 DM-master，则可以执行：

```
tiup dm patch prod-cluster /tmp/dm-master-hotfix.tar.gz -R dm-master
```

或者只替换其中一个 DM-master：

```
tiup dm patch prod-cluster /tmp/dm--hotfix.tar.gz -N 172.16.4.5:8261
```

导入和升级由 DM-Ansible 部署的 DM 1.0 集群

> 注意：
>
> - TiUP 不支持在 DM 1.0 集群中导入 DM Portal 组件。
> - 导入 DM 集群前，终止运行待导入的集群。
> - 如果需要将迁移任务升级至 2.0，不可对该任务执行 `stop-run` 命令。
> - TiUP 仅支持导入数据到 v2.0.0-rc.2 或以上版本的 DM 集群。
> - `import` 命令用于将数据从 DM 1.0 集群导入到一个新的 DM 2.0 集群。如果你需要将 DM 迁移任务导入一个现有的 DM 2.0 集群，可以参考TiDB Data Migration 1.0.x 到 2.0+ 手动升级。
> - 导入集群某些组件的部署目录与原始集群的部署目录不同，可以执行 `display` 命令来查看相关信息。
> - 导入前，执行 `tiup update --self && tiup update dm`，以确保 DM 组件是最新版本。
> - 导入后，集群中只存在一个 DM-master 节点。请参考扩容节点 来扩展 DM-master。

引入 TiUP 前，DM-Ansible 用于部署 DM 集群。要使 TiUP 接管由 DM-Ansible 部署的 DM 1.0 集群，需要执行 `import` 命令：

例如，以下命令可导入使用 DM-Ansible 部署的集群：

```
tiup dm import --dir=/path/to/dm-ansible --cluster-version ${version}
```

可以执行 `tiup list dm-master` 来查看 TiUP 支持的最新集群版本。

`import` 命令的执行过程如下：

1. 根据之前使用 DM-Ansible 部署的 DM 集群，TiUP 生成一个拓扑文件 `topology.yml`。
2. 确认拓扑文件生成后，你可以用这个文件来部署 v2.0 或更高版本的 DM 集群。

部署完成后，执行 `tiup dm start` 命令来启动集群和开始 DM 内核升级流程。

查看操作日志

操作日志的查看可以借助 `audit` 命令，其用法如下：

```
Usage:
  tiup dm audit [audit-id] [flags]

Flags:
  -h, --help   help for audit
```

在不使用 [audit-id] 参数时，该命令会显示执行的命令列表，如下：

```
tiup dm audit
```

```
ID      Time                  Command
--      ----                  -------
4D5kQY  2020-08-13T05:38:19Z  tiup dm display test
4D5kNv  2020-08-13T05:36:13Z  tiup dm list
4D5kNr  2020-08-13T05:36:10Z  tiup dm deploy -p prod-cluster ${version} ./examples/dm/minimal.
    ↪ yaml
```

第一列为 audit-id，如果想看某个命令的执行日志，则传入这个 audit-id：

```
tiup dm audit 4D5kQY
```

在集群节点机器上执行命令

exec 命令可以很方便地到集群的机器上执行命令，使用方式如下：

```
Usage:
  tiup dm exec <cluster-name> [flags]

Flags:
      --command string   the command run on cluster host (default "ls")
  -h, --help             help for exec
  -N, --node strings     Only exec on host with specified nodes
  -R, --role strings     Only exec on host with specified roles
      --sudo             use root permissions (default false)
```

例如，如果要到所有的 DM 节点上执行 `ls /tmp`，则可以执行：

```
tiup dm exec prod-cluster --command='ls /tmp'
```

集群控制工具 (dmctl)

TiUP 集成了 DM 的控制工具 `dmctl`：

运行命令如下：

```
tiup dmctl [args]
```

指定 dmctl 版本。执行如下命令前，将 ${version} 修改为实际需要的版本。

```
tiup dmctl: ${version} [args]
```

例如，以前添加 source 命令为 `dmctl --master-addr master1:8261 operate-source create /tmp/source1.yml`，集成到 TiUP 中的命令为：

```
tiup dmctl --master-addr master1:8261 operate-source create /tmp/source1.yml
```

使用中控机系统自带的 SSH 客户端连接集群

在以上所有操作中，涉及到对集群机器的操作都是通过 TiUP 内置的 SSH 客户端连接集群执行命令，但是在某些场景下，需要使用系统自带的 SSH 客户端来对集群执行操作，比如：

- 使用 SSH 插件来做认证
- 使用定制的 SSH 客户端

此时可以通过命令行参数 `--native-ssh` 启用系统自带命令行：

- 部署集群: `tiup dm deploy <cluster-name> <version> <topo> --native-ssh`
- 启动集群: `tiup dm start <cluster-name> --native-ssh`
- 升级集群: `tiup dm upgrade ... --native-ssh`

所有涉及集群操作的步骤都可以加上 `--native-ssh` 来使用系统自带的客户端。

也可以使用环境变量 `TIUP_NATIVE_SSH` 来指定是否使用本地 SSH 客户端，避免每个命令都需要添加 `--native-ssh` 参数：

```
export TIUP_NATIVE_SSH=true
##### 或者
export TIUP_NATIVE_SSH=1
##### 或者
export TIUP_NATIVE_SSH=enable
```

若环境变量和 `--native-ssh` 同时指定，则以 `--native-ssh` 为准。

> 注意：
>
> 在部署集群的步骤中，若需要使用密码的方式连接 (-p)，或者密钥文件设置了 passphrase，则需要保证中控机上安装了 sshpass，否则连接时会报错。

12.8.6.1.2 TiDB Data Migration 1.0.x 到 2.0+ 手动升级

本文档主要介绍如何手动从 DM v1.0.x 升级到 v2.0+，主要思路为利用 v1.0.x 时的全局 checkpoint 信息在 v2.0+ 集群中启动一个新的增量数据复制任务。

> 注意：
>
> - DM 当前不支持在数据迁移任务处于全量导出或全量导入过程中从 v1.0.x 升级到 v2.0+。
> - 由于 DM 各组件间用于交互的 gRPC 协议进行了较大变更，因此需确保升级前后 DM 集群各组件（包括 dmctl）使用相同的版本。

- 由于 DM 集群的元数据存储（如 checkpoint、shard DDL lock 状态及 online DDL 元信息等）发生了较大变更，升级到 v2.0+ 后无法自动复用 v1.0.x 的元数据，因此在执行升级操作前需要确保：
  - 所有数据迁移任务不处于 shard DDL 协调过程中。
  - 所有数据迁移任务不处于 online DDL 协调过程中。

下面是手动升级的具体步骤。

第 1 步：准备 v2.0+ 的配置文件

准备的 v2.0+ 的配置文件包括上游数据库的配置文件以及数据迁移任务的配置文件。

上游数据库配置文件

在 v2.0+ 中将上游数据库 source 相关的配置从 DM-worker 的进程配置中独立了出来，因此需要根据v1.0.x 的 DM-worker 配置拆分得到 source 配置。

> 注意：
>
> 当前从 v1.0.x 升级到 v2.0+ 时，如在 source 配置中启用了 `enable-gtid`，则后续需要通过解析 binlog 或 relay log 文件获取 binlog position 对应的 GTID sets。

从 DM-Ansible 部署的 v1.0.x 升级

如果 v1.0.x 是使用 DM-Ansible 部署的，且假设在 `inventory.ini` 中有如下 `dm_worker_servers` 配置：

```
[dm_master_servers]
dm_worker1 ansible_host=172.16.10.72 server_id=101 source_id="mysql-replica-01" mysql_host
    ↪ =172.16.10.81 mysql_user=root mysql_password='VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU='
    ↪ mysql_port=3306
dm_worker2 ansible_host=172.16.10.73 server_id=102 source_id="mysql-replica-02" mysql_host
    ↪ =172.16.10.82 mysql_user=root mysql_password='VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU='
    ↪ mysql_port=3306
```

则可以转换得到如下两个 source 配置文件：

```
##### 原 dm_worker1 对应的 source 配置，如命名为 source1.yaml
server-id: 101                                  # 对应原 `server_id`
source-id: "mysql-replica-01"                   # 对应原 `source_id`
from:
  host: "172.16.10.81"                          # 对应原 `mysql_host`
  port: 3306                                    # 对应原 `mysql_port`
  user: "root"                                  # 对应原 `mysql_user`
  password: "VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU="  # 对应原 `mysql_password`
```

```
##### 原 dm_worker2 对应的 source 配置，如命名为 source2.yaml
server-id: 102                                  # 对应原 `server_id`
source-id: "mysql-replica-02"                   # 对应原 `source_id`
from:
  host: "172.16.10.82"                          # 对应原 `mysql_host`
  port: 3306                                    # 对应原 `mysql_port`
  user: "root"                                  # 对应原 `mysql_user`
  password: "VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU="   # 对应原 `mysql_password`
```

从 Binary 部署的 v1.0.x 升级

如果 v1.0.x 是使用 Binary 部署的，且对应的 DM-worker 配置如下：

```
log-level = "info"
log-file = "dm-worker.log"
worker-addr = ":8262"

server-id = 101
source-id = "mysql-replica-01"
flavor = "mysql"

[from]
host = "172.16.10.81"
user = "root"
password = "VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU="
port = 3306
```

则可转换得到如下的一个 source 配置文件：

```
server-id: 101                                  # 对应原 `server-id`
source-id: "mysql-replica-01"                   # 对应原 `source-id`
flavor: "mysql"                                 # 对应原 `flavor`
from:
  host: "172.16.10.81"                          # 对应原 `from.host`
  port: 3306                                    # 对应原 `from.port`
  user: "root"                                  # 对应原 `from.user`
  password: "VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU="   # 对应原 `from.password`
```

数据迁移任务配置文件

对于数据迁移任务配置向导，v2.0+ 基本与 v1.0.x 保持兼容，可直接复制 v1.0.x 的配置。

第 2 步：部署 v2.0+ 集群

> **注意：**
>
> 如果已有其他可用的 v2.0+ 集群，可跳过此步。

使用 TiUP 按所需要节点数部署新的 v2.0+ 集群。

第 3 步：下线 v1.0.x 集群

如果原 v1.0.x 集群是使用 DM-Ansible 部署的，则使用 DM-Ansible 下线 v1.0.x 集群。

如果原 v1.0.x 集群是使用 Binary 部署，则直接停止 DM-worker 与 DM-master 进程。

第 4 步：升级数据迁移任务

1. 使用operate-source 命令将准备 v2.0+ 的配置文件 中得到的上游数据库 source 配置加载到 v2.0+ 集群中。
2. 在下游 TiDB 中，从 v1.0.x 的数据复制任务对应的增量 checkpoint 表中获取对应的全局 checkpoint 信息。
    - 假设 v1.0.x 的数据迁移配置中未额外指定 `meta-schema`（或指定其值为默认的`dm_meta`），且对应的任务名为 `task_v1`，则对应的 checkpoint 信息在下游 TiDB 的 `` `dm_meta`.`task_v1_syncer_checkpoint` `` 表中。
    - 使用以下 SQL 语句分别获取该数据迁移任务对应的所有上游数据库 source 的全局 checkpoint 信息。

```
> SELECT `id`, `binlog_name`, `binlog_pos` FROM `dm_meta`.`task_v1_syncer_checkpoint`
    ↪ WHERE `is_global`=1;
+------------------+--------------------------+------------+
| id               | binlog_name              | binlog_pos |
+------------------+--------------------------+------------+
| mysql-replica-01 | mysql-bin|000001.000123 | 15847      |
| mysql-replica-02 | mysql-bin|000001.000456 | 10485      |
+------------------+--------------------------+------------+
```

3. 更新 v1.0.x 的数据迁移任务配置文件以启动新的 v2.0+ 数据迁移任务。
    - 如 v1.0.x 的数据迁移任务配置文件为 `task_v1.yaml`，则将其复制一份为 `task_v2.yaml`。
    - 对 `task_v2.yaml` 进行以下修改：
        - 将 name 修改为一个新的、不存在的名称，如 `task_v2`
        - 将 `task-mode` 修改为 `incremental`
        - 根据 step.2 中获取的全局 checkpoint 信息，为各 source 设置增量复制的起始点，如：

```
mysql-instances:
  - source-id: "mysql-replica-01"        # 对应 checkpoint 信息所属的 `id`
    meta:
      binlog-name: "mysql-bin.000123"    # 对应 checkpoint 信息中的 `binlog_name`，
          ↪ 但不包含 `|000001` 部分
      binlog-pos: 15847                  # 对应 checkpoint 信息中的 `binlog_pos`

  - source-id: "mysql-replica-02"
    meta:
      binlog-name: "mysql-bin.000456"
      binlog-pos: 10485
```

> 注意：
> 如在 source 配置中启动了 `enable-gtid`，当前需要通过解析 binlog 或 relay log 文件获取 binlog position 对应的 GTID sets 并在 `meta` 中设置为 `binlog-gtid`。

4. 使用`start-task` 命令以 v2.0+ 的数据迁移任务配置文件启动升级后的数据迁移任务。
5. 使用`query-status` 命令确认数据迁移任务是否运行正常。

如果数据迁移任务运行正常，则表明 DM 升级到 v2.0+ 的操作成功。

12.8.6.2 集群运维工具

12.8.6.2.1 使用 dmctl 运维集群

> 注意：
> 对于用 TiUP 部署的 DM 集群，推荐直接使用`tiup dmctl` 命令。

dmctl 是用来运维 DM 集群的命令行工具，支持交互模式和命令模式。

dmctl 交互模式

进入交互模式，与 DM-master 进行交互：

> 注意：
> 交互模式下不具有 bash 的特性，比如不需要通过引号传递字符串参数而应当直接传递。

```
./dmctl --master-addr 172.16.30.14:8261
```

```
Welcome to dmctl
Release Version: ${version}
Git Commit Hash: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Git Branch: release-x.x
UTC Build Time: yyyy-mm-dd hh:mm:ss
Go Version: go version gox.xx linux/amd64

» help
DM control
```

```
Usage:
  dmctl [command]

Available Commands:
  check-task      Checks the configuration file of the task.
  config          Commands to import/export config.
  get-config      Gets the configuration.
  handle-error    `skip`/`replace`/`revert` the current error event or a specific binlog position
      ↪  (binlog-pos) event.
  help            Help about any command
  list-member     Lists member information.
  offline-member  Offlines member which has been closed.
  operate-leader  `evict`/`cancel-evict` the leader.
  operate-schema  `get`/`set`/`remove` the schema for an upstream table.
  operate-source  `create`/`update`/`stop`/`show` upstream MySQL/MariaDB source.
  pause-relay     Pauses DM-worker's relay unit.
  pause-task      Pauses a specified running task.
  purge-relay     Purges relay log files of the DM-worker according to the specified filename.
  query-status    Queries task status.
  resume-relay    Resumes DM-worker's relay unit.
  resume-task     Resumes a specified paused task.
  show-ddl-locks  Shows un-resolved DDL locks.
  start-task      Starts a task as defined in the configuration file.
  stop-task       Stops a specified task.
  unlock-ddl-lock Unlocks DDL lock forcefully.

Flags:
  -h, --help             help for dmctl
  -s, --source strings   MySQL Source ID.

Use "dmctl [command] --help" for more information about a command.
```

dmctl 命令模式

命令模式跟交互模式的区别是，执行命令时只需要在 dmctl 命令后紧接着执行任务操作，任务操作同交互模式的参数一致。

> 注意：
>
> - 一条 dmctl 命令只能跟一个任务操作
> - 从 v2.0.4 版本开始，支持从环境变量 (DM_MASTER_ADDR) 里读取 `-master-addr` 参数

```
./dmctl --master-addr 172.16.30.14:8261 start-task task.yaml
```

```
./dmctl --master-addr 172.16.30.14:8261 stop-task task
./dmctl --master-addr 172.16.30.14:8261 query-status

export DM_MASTER_ADDR="172.16.30.14:8261"
./dmctl query-status
```

```
Available Commands:
  check-task            check-task <config-file> [--error count] [--warn count]
  config                commands to import/export config
  get-config            get-config <task | master | worker | source> <name> [--file filename]
  handle-error          handle-error <task-name | task-file> [-s source ...] [-b binlog-pos] <
      ↪ skip/replace/revert> [replace-sql1;replace-sql2;]
  list-member           list-member [--leader] [--master] [--worker] [--name master-name/worker-
      ↪ name ...]
  offline-member        offline-member <--master/--worker> <--name master-name/worker-name>
  operate-leader        operate-leader <operate-type>
  operate-schema        operate-schema <operate-type> <-s source ...> <task-name | task-file> <-d
      ↪  database> <-t table> [schema-file]
  operate-source        operate-source <operate-type> [config-file ...] [--print-sample-config]
  pause-relay           pause-relay <-s source ...>
  pause-task            pause-task [-s source ...] <task-name | task-file>
  purge-relay           purge-relay <-s source> <-f filename> [--sub-dir directory]
  query-status          query-status [-s source ...] [task-name | task-file] [--more]
  resume-relay          resume-relay <-s source ...>
  resume-task           resume-task [-s source ...] <task-name | task-file>
  show-ddl-locks        show-ddl-locks [-s source ...] [task-name | task-file]
  start-task            start-task [-s source ...] [--remove-meta] <config-file>
  stop-task             stop-task [-s source ...] <task-name | task-file>
  unlock-ddl-lock       unlock-ddl-lock [-s source ...] <lock-ID>

Special Commands:
  --encrypt Encrypts plaintext to ciphertext.
  --decrypt Decrypts ciphertext to plaintext.

Global Options:
  --V Prints version and exit.
  --config Path to configuration file.
  --master-addr Master API server address.
  --rpc-timeout RPC timeout, default is 10m.
  --ssl-ca Path of file that contains list of trusted SSL CAs for connection.
  --ssl-cert Path of file that contains X509 certificate in PEM format for connection.
  --ssl-key Path of file that contains X509 key in PEM format for connection.
```

12.8.6.3 性能调优

12.8.6.3.1 DM 5.4.0 性能测试报告

TiDB Data Migration (DM) 5.4.0 的性能测试数据与 DM 5.3.0 的测试数据持平，因此不单独提供测试报告。若要了解 DM 5.3.0 版本的性能测试目的、环境、场景和结果，请参阅 DM 5.3.0 性能测试报告。

12.8.6.3.2 DM 配置优化

本文档介绍如何对迁移任务的配置进行优化，从而提高 DM 的数据迁移性能。

全量导出

全量导出相关的配置项为 `mydumpers`，下面介绍和性能相关的参数如何配置。

`rows`

设置 `rows` 选项可以开启单表多线程并发导出，值为导出的每个 chunk 包含的最大行数。开启后，DM 会在 MySQL 的单表并发导出时，优先选出一列做拆分基准，选择的优先级为主键 > 唯一索引 > 普通索引，选出目标列后需保证该列为整数类型（如 `INT`、`MEDIUMINT`、`BIGINT` 等）。

`rows` 的值可以设置为 10000，具体设置的值可以根据表中包含数据的总行数以及数据库的性能做调整。另外也需要设置 `threads` 来控制并发线程数量，默认值为 4，可以适当做些调整。

`chunk-filesize`

DM 全量备份时会根据 `chunk-filesize` 参数的值把每个表的数据划分成多个 chunk，每个 chunk 保存到一个文件中，大小约为 `chunk-filesize`。根据这个参数把数据切分到多个文件中，这样就可以利用 DM Load 处理单元的并行处理逻辑提高导入速度。该参数默认值为 64（单位为 MB），正常情况下不需要设置，也可以根据全量数据的大小做适当的调整。

> 注意：
>
> - `mydumpers` 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
> - `mydumpers.threads` 可以使用配置项 `mydumper-thread` 替代来简化配置。
> - 如果设置了 `rows`，DM 会忽略 `chunk-filesize` 的值。

全量导入

全量导入相关的配置项为 `loaders`，下面介绍和性能相关的参数如何配置。

`pool-size`

`pool-size` 为 DM Load 阶段线程数量的设置，默认值为 16，正常情况下不需要设置，也可以根据全量数据的大小以及数据库的性能做适当的调整。

> 注意：

- loaders 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
- loaders.pool-size 可以使用配置项 loader-thread 替代来简化配置。

增量复制

增量复制相关的配置为 syncers，下面介绍和性能相关的参数如何配置。

worker-count

worker-count 为 DM Sync 阶段并发迁移 DML 的线程数量设置，默认值为 16，如果对迁移速度有较高的要求，可以适当调高改参数的值。

batch

batch 为 DM Sync 阶段迁移数据到下游数据库时，每个事务包含的 DML 的数量，默认值为 100，正常情况下不需要调整。

注意：

- syncers 的参数值不支持在迁移任务创建后更新，所以需要在创建任务前确定好各个参数的值。如果需要更新，则需要使用 dmctl stop 任务后更新配置文件，然后再重新创建任务。
- syncers.worker-count 可以使用配置项 syncer-thread 替代来简化配置。
- worker-count 和 batch 的设置需要根据实际的场景进行调整，例如：DM 到下游数据库的网络延迟较高，可以适当调高 worker-count，调低 batch。

12.8.6.3.3 DM 集群性能测试

本文档介绍如何构建测试场景对 DM 集群进行性能测试，包括数据迁移速度、延迟等。

迁移数据流

可以使用简单的迁移数据流来测试 DM 集群的数据迁移性能，即单个 MySQL 实例到 TiDB 的数据迁移：MySQL -> DM -> TiDB。

部署测试环境

- 使用 TiUP 部署 TiDB 测试集群，所有配置使用 TiUP 提供的默认配置。
- 部署 MySQL 服务，开启 ROW 模式 binlog，其他配置项使用默认配置。
- 部署 DM 集群，部署一个 DM-worker 和一个 DM-master 即可。

性能测试

迁移数据表结构

使用如下结构的表进行性能测试：

```
CREATE TABLE `sbtest` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `k` int(11) NOT NULL DEFAULT '0',
  `c` char(120) CHARSET utf8mb4 COLLATE utf8mb4_bin NOT NULL DEFAULT '',
  `pad` char(60) CHARSET utf8mb4 COLLATE utf8mb4_bin NOT NULL DEFAULT '',
  PRIMARY KEY (`id`),
  KEY `k_1` (`k`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_bin
```

全量导入性能测试用例

生成测试数据

使用 sysbench 在上游创建测试表，并生成全量导入的测试数据。sysbench 生成数据的命令如下所示：

```
sysbench --test=oltp_insert --tables=4 --mysql-host=172.16.4.40 --mysql-port=3306 --mysql-user=
    ↪ root --mysql-db=dm_benchmark --db-driver=mysql --table-size=50000000 prepare
```

创建数据迁移任务

1. 创建上游 MySQL 的 source，将 source-id 配置为 source-1。详细操作方法参考：加载数据源配置。
2. 创建 full 模式的 DM 迁移任务，示例任务配置文件如下：

“ ‘yaml

name: test-full

task-mode: full

使用实际测试环境中 TiDB 的信息配置 target-database: host: “192.168.0.1” port: 4000 user: “root” password: “”

mysql-instances: - source-id: “source-1” block-allow-list: “instance” # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list mydumper-config-name: “global” loader-thread: 16

配置 sysbench 生成数据所在的库的名称 block-allow-list: # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list instance: do-dbs: [“dm_benchmark”]

mydumpers: global: rows: 32000 threads: 32 “ ‘

创建数据迁移任务的详细操作参考创建数据迁移任务。

> 注意：
>
> - 在 mydumpers 配置项中使用 rows 选项，可以开启单表多线程并发导出，加快数据导出速度。
> - mysql-instances 配置中的 loader-thread 以及 mydumpers 配置项中的 rows 和 threads 可以做适当调整，测试在不同配置下对性能的影响。

获取测试结果

观察 DM-worker 日志，当出现 `all data files have been finished` 时，表示全量数据导入完成，此时可以看到消耗时间。示例日志如下：

```
[INFO] [loader.go:604] ["all data files have been finished"] [task=test] [unit=load] ["cost time
    ↪ "=52.439796ms]
```

根据测试数据的数据量和导入消耗时间，可以算出全量数据的迁移速度。

增量复制性能测试用例

初始化表

使用 sysbench 在上游创建测试表。

创建数据迁移任务

1. 创建上游 MySQL 的 source, source-id 配置为 `source-1`（如果在全量迁移性能测试中已经创建，则不需要再次创建）。详细操作方法参考：加载数据源配置。
2. 创建 `all` 模式的 DM 迁移任务，示例任务配置文件如下：

“‘yaml

name: test-all

task-mode: all

使用实际测试环境中 TiDB 的信息配置 target-database: host: “192.168.0.1” port: 4000 user: “root” password: “”

mysql-instances: - source-id: “source-1” block-allow-list: “instance” # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list syncer-config-name: “global”

配置 sysbench 生成数据所在的库的名称 block-allow-list: # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list instance: do-dbs: [“dm_benchmark”]

syncers: global: worker-count: 16 batch: 100 “‘

创建数据迁移任务的详细操作参考创建数据迁移任务。

> **注意：**
>
> `syncers` 配置项中的 `worker-count` 和 `batch` 可以做适当调整，测试在不同配置下性能的差异。

生成增量数据

执行 sysbench 命令在上游持续生成增量数据：

```
sysbench --test=oltp_insert --tables=4 --num-threads=32 --mysql-host=172.17.4.40 --mysql-port
    ↪ =3306 --mysql-user=root --mysql-db=dm_benchmark --db-driver=mysql --report-interval=10 --
    ↪ time=1800 run
```

> **注意：**
>
> 可以通过调整 sysbench 的语句类型，测试在不同业务场景下 DM 的数据迁移性能。

获取测试结果

通过 `query-status` 命令观测 DM 的迁移状态，通过 Grafana 观测 DM 的监控指标。主要包括单位时间内完成的 job 数量 `finished sqls jobs` 等，详细的监控指标说明参考Binlog replication 监控指标。

12.8.6.3.4 性能问题及处理方法

本文档介绍 DM 中可能存在的、常见的性能问题及其处理方法。

在诊断与处理性能问题时，请确保已经正确配置并安装 DM 的监控组件，并能在 Grafana 监控面板查看DM 的监控指标。

在诊断性能问题时，请先确保对应组件正在正常运行，否则可能出现监控指标异常的情况，对性能问题的诊断造成干扰。

在诊断问题前，也可以先了解 DM 的性能测试报告。

当数据迁移过程存在较大延迟时，若需快速定位瓶颈是在 DM 组件内部还是在 TiDB 集群，可先排查写入 SQL 到下游 部分的 `DML queue remain length`。

relay log 模块的性能问题及处理方法

在relay log 的监控部分，可以主要通过 `binlog file gap between master and relay` 监控项确认是否存在性能问题。如果该指标长时间大于 1，通常表明存在性能问题；如果该指标基本为 0，一般表明没有性能问题。

如果 `binlog file gap between master and relay` 基本为 0，但仍怀疑存在性能问题，则可以继续查看 `binlog pos`，如果该指标中 `master` 远大于 `relay`，则表明可能存在性能问题。

如果存在性能问题，则继续根据 relay log 模块的主要处理流程分别进行诊断与处理。

读取 binlog 数据

与 relay log 模块从上游读取 binlog 数据相关的主要性能指标是 `read binlog event duration`，该指标表示从上游 MySQL/MariaDB 读取到单个 binlog event 所需要的时间，理想情况下应接近于 DM-worker 与 MySQL/MariaDB 实例间的网络延迟。

对于同机房的数据迁移，这部分一般不会成为性能瓶颈；如果该值过大，请排查 DM-worker 与 MySQL/MariaDB 间的网络连通情况。

对于跨机房的数据迁移，可尝试将 DM-worker 与 MySQL/MariaDB 部署在同一机房，而仍将 TiDB 集群部署在目标机房。

从上游读取 binlog 数据这一流程细分后包括以下三个子流程：

- 上游 MySQL/MariaDB 从本地读取 binlog 数据并通过网络进行发送。上游 MySQL/MariaDB 负载无异常时，该子流程通常不会成为瓶颈。

- binlog 数据通过网络从 MySQL/MariaDB 所在机器传输到 DM-worker 所在机器。该子流程主要由 DM-worker 与上游 MySQL/MariaDB 的网络连通情况决定。
- DM-worker 从网络数据流中读取 binlog 数据，并构造成 binlog event。当 DM-worker 负载无异常时，该子流程通常不会成为瓶颈。

> 注意：
>
> 如果 `read binlog event duration` 的值较大，另一个可能的原因是上游 MySQL/MariaDB 负载较低，一段时间内暂时没有需要发送给 DM 的 binlog event，relay log 模块处于等待状态，导致该值包含了额外的等待时间。

binlog 数据解码与验证

将 binlog event 读取到 DM 内存后，会进行必要的数据解码与验证，这部分通常不会存在性能瓶颈，因此监控面板上默认无对应性能指标。如果需要查看相应指标，可手动为 Prometheus 中的 `dm_relay_read_transform_duration` 添加相应的监控。

写入 relay log 文件

在将 binlog event 写入 relay log 文件时，相关的主要性能指标是 `write relay log duration`，该指标在 `binlog event size` 不是特别大时，值应在微秒级别。如果该值过大，需排查磁盘写入性能，如尽量优先为 DM-worker 使用本地 SSD 等。

Load 模块的性能问题及处理方法

Load 模块主要操作为从本地读取 SQL 文件数据并写入到下游，对应的主要性能指标是 `transaction execution latency`，如果该值过大，则通常需要根据下游数据库的监控对下游性能进行排查。

另外，也可以查看是否 DM 到下游数据库间的网络存在较大的延迟。

Binlog replication 模块的性能问题及处理方法

在Binlog replication 的监控部分，可以主要通过 `binlog file gap between master and syncer` 监控项确认是否存在性能问题，如果该指标长时间大于 1，则通常表明存在性能问题；如果该指标基本为 0，则一般表明没有性能问题。

如果 `binlog file gap between master and syncer` 长时间大于 1，则可以再通过 `binlog file gap between relay and syncer` 判断延迟主要存在于哪个模块，如果该值基本为 0，则延迟可能存在于 relay log 模块，请先参考relay log 模块的性能问题及处理方法 进行处理；否则继续对 Binlog replication 进行排查。

读取 binlog 数据

Binlog replication 模块会根据配置选择从上游 MySQL/MariaDB 或 relay log 文件中读取 binlog event，对应的主要性能指标是 `read binlog event duration`，该值的范围一般是几微秒至几十微秒。

- 如果是从上游 MySQL/MariaDB 读取 binlog event，则可参考 relay log 模块下的读取 binlog 数据进行排查与处理。
- 如果是从 relay log 文件中读取，则在 `binlog event size` 不是特别大时，`read binlog event duration` 的值应在微秒级别。如果 `read binlog event duration` 过大，则需排查磁盘读取性能，如尽量优先为 DM-worker 使用本地 SSD 等。

binlog event 转换

Binlog replication 模块从 binlog event 数据中尝试构造 DML、解析 DDL 以及进行 table router 转换等，主要的性能指标是 `transform binlog event duration`。

这部分的耗时受上游写入的业务特点影响较大，如对于 `INSERT INTO` 语句，转换单个 `VALUES` 的时间和转换大量 `VALUES` 的时间差距很多，其波动范围可能从几十微秒至上百微秒，但一般不会成为系统的瓶颈。

写入 SQL 到下游

Binlog replication 模块将转换后的 SQL 写入到下游时，涉及到的性能指标主要包括 `DML queue remain length` 与 `transaction execution latency`。

DM 在从 binlog event 构造出 SQL 后，会使用 `worker-count` 个队列尝试并发写入到下游。但为了避免监控条目过多，会将并发队列编号按 8 取模，即所有并发队列在监控上会对应到 `q_0` 到 `q_7` 的某一项。

`DML queue remain length` 用于表示并发处理队列中尚未取出并开始用于向下游写入的 DML 语句数，理想情况下，各 `q_*` 对应的曲线应该基本一致，如果极不一致则表明并发的负载极不均衡。

如果负载不均衡，请确认需要迁移的所有表结构中都有主键或唯一键，如没有主键或唯一键则请尝试为其添加主键或唯一键；如果存在主键或唯一键时仍存在该问题，可尝试升级 DM 到 v1.0.5 及以上的版本。

当整个数据迁移链路无明显延迟时，`DML queue remain length` 对应曲线应基本为 0，且最大通常应不超过任务配置文件中的 `batch` 值。

如果确认数据迁移链路存在明显延迟，且 `DML queue remain length` 中各 `q_*` 对应的曲线基本一致且基本为 0，则表明 DM 未能及时地从上游读取数据、进行转换或进行并行分发（如瓶颈存在于 relay log 模块等），请参考本文档前述各节进行排查。

如果 `DML queue remain length` 对应曲线不为 0（最大一般不超过 1024），则通常表明向下游写入 SQL 时存在瓶颈，可通过 `transaction execution latency` 查看向下游执行单个事务的耗时情况。

`transaction execution latency` 一般应在几十毫秒。如果该值过高，则通常需要根据下游数据库的监控对下游性能进行排查，另外也可以关注是否 DM 到下游数据库间的网络存在较大的延迟。

此外，也可通过 `statement execution latency` 查看向下游写入 `BEGIN`、`INSERT/UPDATE/DELETE`、`COMMIT` 等单条语句的耗时情况。

12.8.6.4 数据源管理

12.8.6.4.1 切换 DM-worker 与上游 MySQL 实例的连接

当需要对 DM-worker 所连接的上游 MySQL 实例进行停机维护或该实例意外宕机时，需要将 DM-worker 的连接切换到同一个主从复制集群内的另一个 MySQL 实例上。本文介绍如何将 DM-worker 的连接从一个 MySQL 实例切换到另一个 MySQL 实例上。

> 注意：
>
> - 仅支持在同一个主从复制集内的 MySQL 实例间进行切换。
> - 将要切换到的 MySQL 实例必须拥有 DM-worker 所需的 binlog。

- DM-worker 必须以 GTID sets 模式运行，即对应 source 配置文件中指定 `enable-gtid: true`。
- DM 仅支持以下两种切换场景，且必须严格按照各场景的步骤执行操作，否则可能需要根据切换后的 MySQL 实例重新搭建 DM 集群并完整重做数据迁移任务。

有关 GTID sets 的概念解释，请参考 MySQL 文档。

虚拟 IP 环境下切换 DM-worker 与 MySQL 实例的连接

如果 DM-worker 通过虚拟 IP (VIP) 连接上游的 MySQL 实例，更改 VIP 所指向的 MySQL 实例，即是在 DM-worker 对应上游连接地址不变的情况下切换 DM-worker 实际所连接的 MySQL 实例。

> 注意：
>
> 如果不对 DM 执行必要变更，当切换 VIP 所指向的 MySQL 实例时，DM 内部不同的 connection 可能会同时连接到切换前后不同的 MySQL 实例，造成 DM 拉取的 binlog 与从上游获取到的其他状态不一致，从而导致难以预期的异常行为甚至数据损坏。

如果 DM-worker 连接的 VIP 需要指向新的 MySQL 实例，需要按以下步骤进行操作：

1. 使用 `query-status` 命令获取当前 binlog replication 处理单元已复制到下游的 binlog 对应的 GTID sets (`syncerBinlogGtid`)，记为 `gtid-S`。
2. 在将要切换到的 MySQL 实例上使用 `SELECT @@GLOBAL.gtid_purged;` 获取已经被 purged 的 binlog 对应的 GTID sets，记为 `gtid-P`。
3. 在将要切换到的 MySQL 实例上使用 `SELECT @@GLOBAL.gtid_executed;` 获取所有已经执行成功的事务对应的 GTID sets，记为 `gtid-E`。
4. 确保满足以下关系，否则不支持将 DM-worker 连接切换到相应的 MySQL 实例：
    - `gtid-S` 包含 `gtid-P`（`gtid-P` 可以为空）
    - `gtid-E` 包含 `gtid-S`
5. 使用 `pause-task` 命令暂停所有运行中的数据迁移任务。
6. 变更 VIP 以指向新的 MySQL 实例。
7. 使用 `resume-task` 命令恢复之前的数据迁移任务。

变更 DM-worker 连接的上游 MySQL 实例地址

若要变更 DM-worker 的配置信息来使 DM-worker 连接到新的上游 MySQL 实例，需要按以下步骤进行操作：

1. 使用 `query-status` 命令获取当前 binlog replication 处理单元已复制到下游的 binlog 对应的 GTID sets (`syncerBinlogGtid`)，记为 `gtid-S`。
2. 在将要切换到的 MySQL 实例上使用 `SELECT @@GLOBAL.gtid_purged;` 获取已经被 purged 的 binlog 对应的 GTID sets，记为 `gtid-P`。
3. 在将要切换到的 MySQL 实例上使用 `SELECT @@GLOBAL.gtid_executed;` 获取所有已经执行成功的事务对应的 GTID sets，记为 `gtid-E`。

4. 确保满足以下关系，否则不支持将 DM-worker 连接切换到相应的 MySQL 实例：
    - `gtid-S` 包含 `gtid-P`（`gtid-P` 可以为空）
    - `gtid-E` 包含 `gtid-S`
5. 使用 `stop-task` 命令停止所有运行中的数据迁移任务。
6. 使用 `operate-source stop` 命令从 DM 集群中移除原 MySQL 实例地址对应的 source 配置。
7. 更新 source 配置文件中的 MySQL 实例地址并使用 `operate-source create` 命令将新的 source 配置重新加载到 DM 集群中。
8. 使用 `start-task` 命令重新启动数据迁移任务。

12.8.6.5 任务管理

12.8.6.5.1 处理出错的 DDL 语句

本文介绍了如何使用 DM 来处理出错的 DDL 语句。

目前，TiDB 并不完全兼容所有的 MySQL 语法（详见 DDL 的限制）。当使用 DM 从 MySQL 迁移数据到 TiDB 时，如果 TiDB 不支持对应的 DDL 语句，可能会造成错误并中断迁移任务。在这种情况下，DM 提供 `handle-error` 命令来恢复迁移。

使用限制

如果业务不能接受下游 TiDB 跳过异常的 DDL 语句，也不接受使用其他 DDL 语句作为替代，则不适合使用此方式进行处理。

比如：`DROP PRIMARY KEY`，这种情况下，只能在下游重建一个（DDL 执行完后的）新表结构对应的表，并将原表的全部数据重新导入该新表。

支持场景

迁移过程中，上游执行了 TiDB 不支持的 DDL 语句并迁移到了 DM，造成迁移任务中断。

- 如果业务能接受下游 TiDB 不执行该 DDL 语句，则使用 `handle-error <task-name> skip` 跳过对该 DDL 语句的迁移以恢复迁移任务。
- 如果业务能接受下游 TiDB 执行其他 DDL 语句来作为替代，则使用 `handle-error <task-name> replace` 替代该 DDL 的迁移以恢复迁移任务。

命令介绍

使用 dmctl 手动处理出错的 DDL 语句时，主要使用的命令包括 `query-status`、`handle-error`。

query-status

`query-status` 命令用于查询当前 MySQL 实例内子任务及 relay 单元等的状态和错误信息，详见查询状态。

handle-error

`handle-error` 命令用于处理错误的 DDL 语句。

命令用法

```
» handle-error -h
```

```
Usage:
  dmctl handle-error <task-name | task-file> [-s source ...] [-b binlog-pos] <skip/replace/revert
      ↪ > [replace-sql1;replace-sql2;] [flags]

Flags:
  -b, --binlog-pos string   position used to match binlog event if matched the handler-error
      ↪ operation will be applied. The format like "mysql-bin|000001.000003:3270"
  -h, --help                help for handle-error

Global Flags:
  -s, --source strings   MySQL Source ID
```

参数解释

- `task-name`：
  - 非 flag 参数，string，必选；
  - 指定预设的操作将生效的任务。
- `source`：
  - flag 参数，string，`--source`；
  - `source` 指定预设操作将生效的 MySQL 实例。
- `skip`：跳过该错误
- `replace`：替代错误的 DDL 语句
- `revert`：重置该错误先前的 skip/replace 操作，仅在先前的操作没有最终生效前执行
- `binlog-pos`：
  - flag 参数，string，`--binlog-pos`；
  - 若不指定，DM 会自动处理当前出错的 DDL 语句
  - 在指定时表示操作将在 `binlog-pos` 与 binlog event 的 position 匹配时生效，格式为 `binlog-filename:binlog-pos`，如 `mysql-bin|000001.000003:3270`。
  - 在迁移执行出错后，binlog position 可直接从 `query-status` 返回的 `startLocation` 中的 `position` 获得；在迁移执行出错前，binlog position 可在上游 MySQL 中使用 `SHOW BINLOG EVENTS` 获得。

使用示例

迁移中断执行跳过操作

非合库合表场景

假设现在需要将上游的 `db1.tbl1` 表迁移到下游 TiDB，初始时表结构为：

```
SHOW CREATE TABLE db1.tbl1;
```

```
+-------+-------------------------------------------------+
| Table | Create Table                                    |
+-------+-------------------------------------------------+
| tbl1  | CREATE TABLE `tbl1` (
  `c1` int(11) NOT NULL,
  `c2` decimal(11,3) DEFAULT NULL,
  PRIMARY KEY (`c1`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 |
+-------+-------------------------------------------------+
```

此时，上游执行以下 DDL 操作修改表结构（将列的 DECIMAL(11, 3) 修改为 DECIMAL(10, 3)）：

```
ALTER TABLE db1.tbl1 CHANGE c2 c2 DECIMAL (10, 3);
```

则会由于 TiDB 不支持该 DDL 语句而导致 DM 迁移任务中断，使用 `query-status <task-name>` 命令可看到如下错误：

```
ERROR 8200 (HY000): Unsupported modify column: can't change decimal column precision
```

假设业务上可以接受下游 TiDB 不执行此 DDL 语句（即继续保持原有的表结构），则可以通过使用 `handle-error <task-name> skip` 命令跳过该 DDL 语句以恢复迁移任务。操作步骤如下：

1. 使用 `handle-error <task-name> skip` 跳过当前错误的 DDL 语句

```
» handle-error test skip
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

2. 使用 `query-status <task-name>` 查看任务状态

```
» query-status test
```

执行结果

```
{
    "result": true,
```

```
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "sourceStatus": {
                "source": "mysql-replica-01",
                "worker": "worker1",
                "result": null,
                "relayStatus": null
            },
            "subTaskStatus": [
                {
                    "name": "test",
                    "stage": "Running",
                    "unit": "Sync",
                    "result": null,
                    "unresolvedDDLLockID": "",
                    "sync": {
                        "totalEvents": "4",
                        "totalTps": "0",
                        "recentTps": "0",
                        "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                        "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
                        "blockingDDLs": [
                        ],
                        "unresolvedGroups": [
                        ],
                        "synced": true,
                        "binlogType": "remote"
                    }
                }
            ]
        }
    ]
}
```


可以看到任务运行正常，错误的 DDL 被跳过。

合库合表场景

假设现在存在如下四个上游表需要合并迁移到下游的同一个表 `shard_db`.`shard_table`，任务模式为悲观协调模式：

- MySQL 实例 1 内有 shard_db_1 库，包括 shard_table_1 和 shard_table_2 两个表。
- MySQL 实例 2 内有 shard_db_2 库，包括 shard_table_1 和 shard_table_2 两个表。

初始时表结构为：

```
SHOW CREATE TABLE shard_db.shard_table;
```

```
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
| Table | Create Table
    ↪
    ↪ |
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
| tb    | CREATE TABLE `shard_table` (
  `id` int(11) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin |
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
```

此时，在上游所有分表上都执行以下 DDL 操作修改表字符集

```
ALTER TABLE `shard_db_*`.`shard_table_*` CHARACTER SET LATIN1 COLLATE LATIN1_DANISH_CI;
```

则会由于 TiDB 不支持该 DDL 语句而导致 DM 迁移任务中断，使用 query-status 命令可以看到 MySQL 实例 1 的 shard_db_1.shard_table_1 表和 MySQL 实例 2 的 shard_db_2.shard_table_1 表报错：

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_1`.`shard_table_1` CHARACTER SET UTF8
        ↪ COLLATE UTF8_UNICODE_CI",
    "RawCause": "[ddl:8200]Unsupported modify charset from latin1 to utf8"
}
```

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_2`.`shard_table_1` CHARACTER SET UTF8
        ↪ COLLATE UTF8_UNICODE_CI",
    "RawCause": "[ddl:8200]Unsupported modify charset from latin1 to utf8"
}
```

假设业务上可以接受下游 TiDB 不执行此 DDL 语句（即继续保持原有的表结构），则可以通过使用 handle-
↪ error <task-name> skip 命令跳过该 DDL 语句以恢复迁移任务。操作步骤如下：

1. 使用 handle-error <task-name> skip 跳过 MySQL 实例 1 和实例 2 当前错误的 DDL 语句

```
» handle-error test skip
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        },
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-02",
            "worker": "worker2"
        }
    ]
}
```

2. 使用 `query-status <task-name>` 查看任务状态，可以看到 MySQL 实例 1 的 `shard_db_1.shard_table_2` 表和 MySQL 实例 2 的 `shard_db_2.shard_table_2` 表报错：

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_1`.`shard_table_2` CHARACTER SET
        ↪ UTF8 COLLATE UTF8_UNICODE_CI",
    "RawCause": "[ddl:8200]Unsupported modify charset from latin1 to utf8"
}
```

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_2`.`shard_table_2` CHARACTER SET
        ↪ UTF8 COLLATE UTF8_UNICODE_CI",
    "RawCause": "[ddl:8200]Unsupported modify charset from latin1 to utf8"
}
```

3. 继续使用 `handle-error <task-name> skip` 跳过 MySQL 实例 1 和实例 2 当前错误的 DDL 语句

```
» handle-error test skip
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
```

```
            "source": "mysql-replica-01",
            "worker": "worker1"
        },
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-02",
            "worker": "worker2"
        }
    ]
}
```

4. 使用 `query-status <task-name>` 查看任务状态

```
» query-status test
```

执行结果

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "sourceStatus": {
                "source": "mysql-replica-01",
                "worker": "worker1",
                "result": null,
                "relayStatus": null
            },
            "subTaskStatus": [
                {
                    "name": "test",
                    "stage": "Running",
                    "unit": "Sync",
                    "result": null,
                    "unresolvedDDLLockID": "",
                    "sync": {
                        "totalEvents": "4",
                        "totalTps": "0",
                        "recentTps": "0",
                        "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                        "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
```

```
                    "blockingDDLs": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "synced": true,
                    "binlogType": "remote"
                }
            }
        ]
    },
    {
        "result": true,
        "msg": "",
        "sourceStatus": {
            "source": "mysql-replica-02",
            "worker": "worker2",
            "result": null,
            "relayStatus": null
        },
        "subTaskStatus": [
            {
                "name": "test",
                "stage": "Running",
                "unit": "Sync",
                "result": null,
                "unresolvedDDLLockID": "",
                "sync": {
                    "totalEvents": "4",
                    "totalTps": "0",
                    "recentTps": "0",
                    "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                    "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                    "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                    "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
                    "blockingDDLs": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "synced": true,
                    "binlogType": "remote"
                }
            }
        ]
    }
]
```

```
}
```

可以看到任务运行正常，无错误信息。四条 DDL 全部被跳过。

迁移中断执行替代操作

非合库合表场景

假设现在需要将上游的 db1.tbl1 表迁移到下游 TiDB，初始时表结构为：

```
SHOW CREATE TABLE db1.tbl1;
```

```
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
| Table | Create Table
    ↪
    ↪ |
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
| tb    | CREATE TABLE `tbl1` (
  `id` int(11) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin |
+-------+-----------------------------------------------------------------------------------------------------------------
    ↪
```

此时，上游执行以下 DDL 操作增加新列，并添加 UNIQUE 约束

```
ALTER TABLE `db1`.`tbl1` ADD COLUMN new_col INT UNIQUE;
```

则会由于 TiDB 不支持该 DDL 语句而导致 DM 迁移任务中断，使用 query-status 命令可看到如下错误：

```
{
    "Message": "cannot track DDL: ALTER TABLE `db1`.`tbl1` ADD COLUMN `new_col` INT UNIQUE KEY",
    "RawCause": "[ddl:8200]unsupported add column 'new_col' constraint UNIQUE KEY when altering '
        ↪ db1.tbl1'",
}
```

我们将该 DDL 替换成两条等价的 DDL。操作步骤如下：

1. 使用如下命令替换错误的 DDL 语句

```
» handle-error test replace "ALTER TABLE `db1`.`tbl1` ADD COLUMN `new_col` INT;ALTER TABLE `
    ↪ db1`.`tbl1` ADD UNIQUE(`new_col`)";
```

```
{
    "result": true,
    "msg": "",
```

```
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

2. 使用 query-status <task-name> 查看任务状态

```
» query-status test
```

执行结果

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "sourceStatus": {
                "source": "mysql-replica-01",
                "worker": "worker1",
                "result": null,
                "relayStatus": null
            },
            "subTaskStatus": [
                {
                    "name": "test",
                    "stage": "Running",
                    "unit": "Sync",
                    "result": null,
                    "unresolvedDDLLockID": "",
                    "sync": {
                        "totalEvents": "4",
                        "totalTps": "0",
                        "recentTps": "0",
                        "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                        "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
                        "blockingDDLs": [
                        ],
```

```
                        "unresolvedGroups": [
                        ],
                        "synced": true,
                        "binlogType": "remote"
                    }
                }
            ]
        }
    ]
}
```

可以看到任务运行正常，错误的 DDL 已被替换且执行成功。

合库合表场景

假设现在存在如下四个上游表需要合并迁移到下游的同一个表 `shard_db`.`shard_table`，任务模式为悲观协调模式：

- MySQL 实例 1 内有 shard_db_1 库，包括 shard_table_1 和 shard_table_2 两个表。
- MySQL 实例 2 内有 shard_db_2 库，包括 shard_table_1 和 shard_table_2 两个表。

初始时表结构为：

```
SHOW CREATE TABLE shard_db.shard_table;
```

```
+-------+-------------------------------------------------------------------------------------------------------------
    ↪
| Table | Create Table
    ↪
    ↪ |
+-------+-------------------------------------------------------------------------------------------------------------
    ↪
| tb    | CREATE TABLE `shard_table` (
  `id` int(11) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin |
+-------+-------------------------------------------------------------------------------------------------------------
    ↪
```

此时，在上游所有分表上都执行以下 DDL 操作增加新列，并添加 UNIQUE 约束：

```
ALTER TABLE `shard_db_*`.`shard_table_*` ADD COLUMN new_col INT UNIQUE;
```

则会由于 TiDB 不支持该 DDL 语句而导致 DM 迁移任务中断，使用 query-status 命令可以看到 MySQL 实例 1 的 shard_db_1.shard_table_1 表和 MySQL 实例 2 的 shard_db_2.shard_table_1 表报错：

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_1`.`shard_table_1` ADD COLUMN `new_col`
        ↪ INT UNIQUE KEY",
    "RawCause": "[ddl:8200]unsupported add column 'new_col' constraint UNIQUE KEY when altering '
        ↪ shard_db_1.shard_table_1'",
}
```

```
{
    "Message": "cannot track DDL: ALTER TABLE `shard_db_2`.`shard_table_1` ADD COLUMN `new_col`
        ↪ INT UNIQUE KEY",
    "RawCause": "[ddl:8200]unsupported add column 'new_col' constraint UNIQUE KEY when altering '
        ↪ shard_db_2.shard_table_1'",
}
```

我们将该 DDL 替换成两条等价的 DDL。操作步骤如下：

1. 使用如下命令分别替换 MySQL 实例 1 和实例 2 中错误的 DDL 语句

```
» handle-error test -s mysql-replica-01 replace "ALTER TABLE `shard_db_1`.`shard_table_1`
    ↪ ADD COLUMN `new_col` INT;ALTER TABLE `shard_db_1`.`shard_table_1` ADD UNIQUE(`
    ↪ new_col`)";
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

```
» handle-error test -s mysql-replica-02 replace "ALTER TABLE `shard_db_2`.`shard_table_1`
    ↪ ADD COLUMN `new_col` INT;ALTER TABLE `shard_db_2`.`shard_table_1` ADD UNIQUE(`
    ↪ new_col`)";
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
```

```
            "msg": "",
            "source": "mysql-replica-02",
            "worker": "worker2"
        }
    ]
}
```

2. 使用 query-status <task-name> 查看任务状态，可以看到 MySQL 实例 1 的 shard_db_1.shard_table_2 表和 MySQL 实例 2 的 shard_db_2.shard_table_2 表报错：

```
{
    "Message": "detect inconsistent DDL sequence from source ... ddls: [ALTER TABLE `
        ↪ shard_db`.`tb` ADD COLUMN `new_col` INT UNIQUE KEY] source: `shard_db_1`.`
        ↪ shard_table_2`], right DDL sequence should be ..."
}
```

```
{
    "Message": "detect inconsistent DDL sequence from source ... ddls: [ALTER TABLE `
        ↪ shard_db`.`tb` ADD COLUMN `new_col` INT UNIQUE KEY] source: `shard_db_2`.`
        ↪ shard_table_2`], right DDL sequence should be ..."
}
```

3. 使用如下命令继续分别替换 MySQL 实例 1 和实例 2 中错误的 DDL 语句

```
» handle-error test -s mysql-replica-01 replace "ALTER TABLE `shard_db_1`.`shard_table_2`
    ↪ ADD COLUMN `new_col` INT;ALTER TABLE `shard_db_1`.`shard_table_2` ADD UNIQUE(`
    ↪ new_col`)";
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "worker1"
        }
    ]
}
```

```
» handle-error test -s mysql-replica-02 replace "ALTER TABLE `shard_db_2`.`shard_table_2`
    ↪ ADD COLUMN `new_col` INT;ALTER TABLE `shard_db_2`.`shard_table_2` ADD UNIQUE(`
    ↪ new_col`)";
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-02",
            "worker": "worker2"
        }
    ]
}
```

4. 使用 `query-status <task-name>` 查看任务状态

```
» query-status test
```

执行结果

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "sourceStatus": {
                "source": "mysql-replica-01",
                "worker": "worker1",
                "result": null,
                "relayStatus": null
            },
            "subTaskStatus": [
                {
                    "name": "test",
                    "stage": "Running",
                    "unit": "Sync",
                    "result": null,
                    "unresolvedDDLLockID": "",
                    "sync": {
                        "totalEvents": "4",
                        "totalTps": "0",
                        "recentTps": "0",
                        "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                        "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                        "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
```

```
                    "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
                    "blockingDDLs": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "synced": true,
                    "binlogType": "remote"
                }
            }
        ]
    },
    {
        "result": true,
        "msg": "",
        "sourceStatus": {
            "source": "mysql-replica-02",
            "worker": "worker2",
            "result": null,
            "relayStatus": null
        },
        "subTaskStatus": [
            {
                "name": "test",
                "stage": "Running",
                "unit": "Sync",
                "result": null,
                "unresolvedDDLLockID": "",
                "sync": {
                    "totalEvents": "4",
                    "totalTps": "0",
                    "recentTps": "0",
                    "masterBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                    "masterBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-10",
                    "syncerBinlog": "(DESKTOP-T561TSO-bin.000001, 2388)",
                    "syncerBinlogGtid": "143bdef3-dd4a-11ea-8b00-00155de45f57:1-4",
                    "blockingDDLs": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "unresolvedGroups": [
                    ],
                    "synced": try,
                    "binlogType": "remote"
```

```
                    }
                }
            ]
        }
    ]
}
```

可以看到任务运行正常，无错误信息。四条 DDL 全部被替换。

12.8.6.5.2 管理迁移表的表结构

本文介绍如何使用dmctl 组件来管理通过 DM 迁移的表在 DM 内部的表结构。

原理介绍

在使用 DM 迁移数据表时，DM 对于表结构主要包含以下相关处理。

对于全量导出与导入，DM 直接导出当前时刻上游的表结构到 SQL 格式的文件中，并将该表结构直接应用到下游。

对于增量复制，在整个数据链路中则包含以下几类可能相同或不同的表结构。

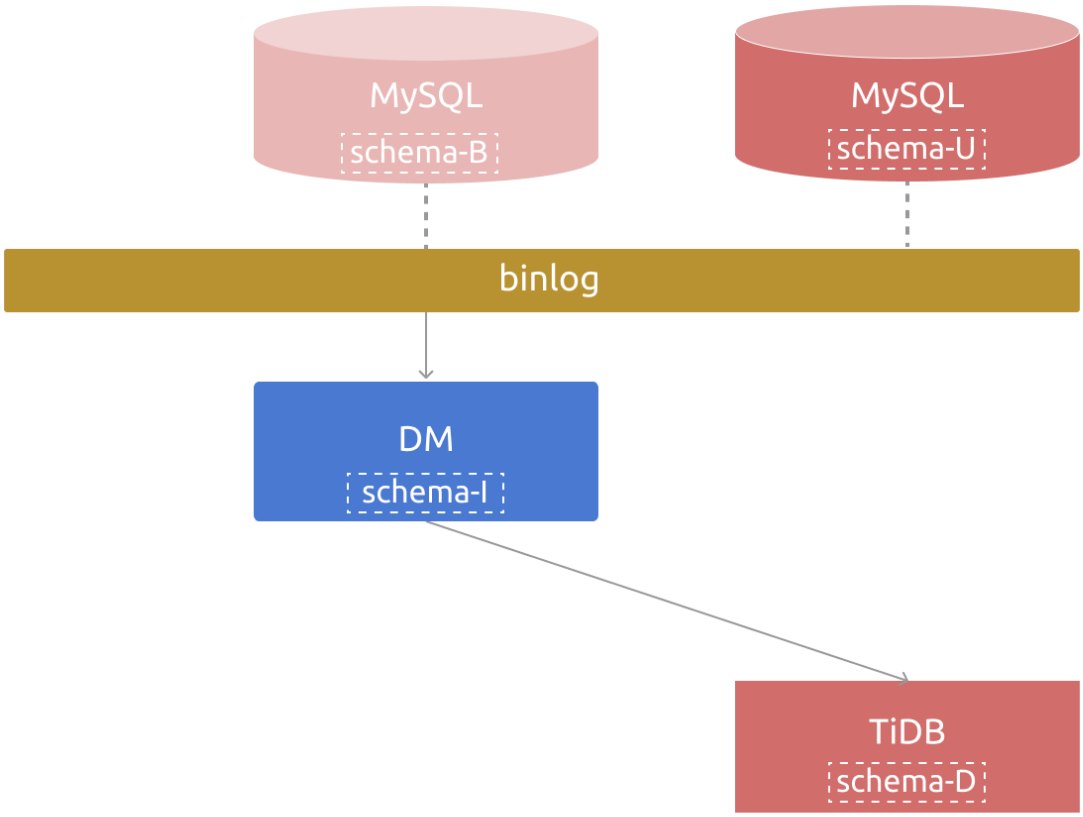


图 293: 表结构

- 上游当前时刻的表结构（记为 schema-U）。
- 当前 DM 正在消费的 binlog event 的表结构（记为 schema-B，其对应于上游某个历史时刻的表结构）。
- DM 内部（schema tracker 组件）当前维护的表结构（记为 schema-I）。
- 下游 TiDB 集群中的表结构（记为 schema-D）。

在大多数情况下，以上 4 类表结构一致。

当上游执行 DDL 变更表结构后，schema-U 即会发生变更；DM 通过将该 DDL 应用于内部的 schema tracker 组件及下游 TiDB，会先后更新 schema-I、schema-D 以与 schema-U 保持一致，因而随后能正常消费 binlog 中在 DDL 之后对应表结构为 schema-B 的 binlog event。即当 DDL 被复制成功后，仍能保持 schema-U、schema-B、schema-I 及 schema-D 的一致。

但在开启乐观 shard DDL 支持 的数据迁移过程中，下游合并表的 schema-D 可能与部分分表对应的 schema-B 及 schema-I 并不一致，但 DM 仍保持 schema-I 与 schema-B 的一致以确保能正常解析 DML 对应的 binlog event。

此外，在其他一些场景下（如：下游比上游多部分列），schema-D 也可能会与 schema-B 及 schema-I 并不一致。

为了支持以上的特殊场景及处理其他可能的由于 schema 不匹配导致的迁移中断等问题，DM 提供了 operate-schema 命令来获取、修改、删除 DM 内部维护的表结构 schema-I。

命令介绍

```
help operate-schema
```

```
`get`/`set`/`remove` the schema for an upstream table.

Usage:
  dmctl operate-schema <operate-type> <-s source ...> <task-name | task-file> <-d database> <-t
      ↪ table> [schema-file] [--flush] [--sync] [flags]

Flags:
  -d, --database string   database name of the table
      --flush             flush the table info and checkpoint immediately
  -h, --help              help for operate-schema
      --sync              sync the table info to master to resolve shard ddl lock, only for
          ↪ optimistic mode now
  -t, --table string      table name

Global Flags:
  -s, --source strings   MySQL Source ID.
```

> **注意：**
>
> - 由于表结构在数据迁移过程中可能会发生变更，为获取确定性的表结构，当前 `operate-schema` 命令仅能在数据迁移任务处于 `Paused` 状态时可用。
> - 强烈建议在修改表结构前，首先获取并备份表结构，以免误操作导致数据丢失。

参数解释

- `operate-type`:
  - 必选
  - 指定对于 schema 的操作类型，可以为 `get`、`set` 或 `remove`
- `-s`:
  - 必选
  - 指定操作将应用到的 MySQL 源
- `task-name | task-file`:
  - 必选
  - 指定任务名称或任务文件路径
- `-d`:
  - 必选

  - 指定数据表所属的上游数据库名

- -t:
  - 必选
  - 指定数据表对应的上游表名

- schema-file:
  - set 操作时必选，其他操作不需指定
  - 将被设置的表结构文件，文件内容应为一个合法的 CREATE TABLE 语句

- --flush:
  - 可选
  - 同时将表结构写入 checkpoint，从而在任务重启后加载到该表结构
  - 默认值为 true

- --sync:
  - 可选
  - 仅在乐观协调 DDL 出错时使用。使用该表结构更新乐观协调元数据中的表结构

使用示例

获取表结构

假设要获取 db_single 任务对应于 mysql-replica-01 MySQL 源的 `db_single`.`t1` 表的表结构，则执行如下命令：

```
operate-schema get -s mysql-replica-01 task_single -d db_single -t t1
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "CREATE TABLE `t1` ( `c1` int(11) NOT NULL, `c2` int(11) DEFAULT NULL, PRIMARY
                ↪  KEY (`c1`)) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin",
            "source": "mysql-replica-01",
            "worker": "127.0.0.1:8262"
        }
    ]
}
```

设置表结构

假设要设置 db_single 任务对应于 mysql-replica-01 MySQL 源的 `db_single`.`t1` 表的表结构为

```
CREATE TABLE `t1` (
    `c1` int(11) NOT NULL,
    `c2` bigint(11) DEFAULT NULL,
```

```
    PRIMARY KEY (`c1`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COLLATE=latin1_bin
```

则将上述 CREATE TABLE 语句保存为文件（如 db_single.t1-schema.sql）后执行如下命令：

```
operate-schema set -s mysql-replica-01 task_single -d db_single -t t1 db_single.t1-schema.sql
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "127.0.0.1:8262"
        }
    ]
}
```

删除表结构

> 注意：
>
> 删除 DM 内部维护的表结构后，如果后续有该表的 DDL/DML 需要复制到下游，则 DM 会依次尝试从 checkpoint 表里 table_info 字段、乐观 shard DDL 协调中的元信息以及下游 TiDB 中对应的该表获取表结构。

假设要删除 db_single 任务对应于 mysql-replica-01 MySQL 源的 `db_single`.`t1` 表的表结构，则执行如下命令：

```
operate-schema remove -s mysql-replica-01 task_single -d db_single -t t1
```

```
{
    "result": true,
    "msg": "",
    "sources": [
        {
            "result": true,
            "msg": "",
            "source": "mysql-replica-01",
            "worker": "127.0.0.1:8262"
        }
    ]
}
```

12.8.6.6 导出和导入集群的数据源和任务配置

config 命令用于导出和导入集群的数据源和任务配置。

> **注意：**
>
> 对于 v2.0.5 版本之前的集群，可使用 v2.0.5 版本及之后的 dmctl 导出和导入集群的数据源和任务配置文件。

```
» help config
Commands to import/export config

Usage:
  dmctl config [command]

Available Commands:
  export      Export the configurations of sources and tasks.
  import      Import the configurations of sources and tasks.

Flags:
  -h, --help   help for config

Global Flags:
  -s, --source strings   MySQL Source ID.

Use "dmctl config [command] --help" for more information about a command.
```

12.8.6.6.1 导出集群的数据源和任务配置

使用 export 子命令导出集群的数据源和任务配置到指定文件夹中。

```
config export [--dir directory]
```

参数解释

- dir：
  - 可选
  - 指定导出文件夹路径
  - 默认值为 ./configs

返回结果示例

```
config export -d /tmp/configs
```

```
export configs to directory `/tmp/configs` succeed
```

12.8.6.6.2 导入集群的数据源和任务配置

使用 `import` 子命令从指定文件夹中导入集群的数据源和任务配置。

```
config import [--dir directory]
```

> **注意：**
>
> 对于 v2.0.2 版本之后的集群，暂不支持自动导入 relay worker 的相关配置，可以手动使用 `start-relay` 命令开启 relay log。

参数解释

- `dir`：
  - 可选
  - 指定导入文件夹路径
  - 默认值为 `./configs`

返回结果示例

```
config import -d /tmp/configs
```

```
start creating sources
start creating tasks
import configs from directory `/tmp/configs` succeed
```

12.8.6.7 处理告警

本文档介绍 DM 中各主要告警信息的处理方法。

12.8.6.7.1 高可用告警

`DM_master_all_down`

当全部 DM-master 离线时触发该告警。发生该错误时，需要检查集群环境，并通过各节点日志排查错误。

`DM_worker_offline`

存在离线的 DM-worker 超过一小时会触发该告警。在高可用架构下，该告警可能不会直接中断任务，但是会提升任务中断的风险。处理告警可以查看对应 DM-worker 节点的工作状态，检查是否连通，并通过日志排查错误。

`DM_DDL_error`

处理 shard DDL 时出现错误，此时需要参考DM 故障诊断进行处理。

`DM_pending_DDL`

存在未完成的 shard DDL 并超过一小时会触发该告警。在某些应用场景下，存在未完成的 shard DDL 可能是用户所期望的。在用户预期以外的场景下，可以通过手动处理 Sharding DDL Lock解决。

12.8.6.7.2 任务状态告警

`DM_task_state`

当 DM-worker 内有子任务处于 Paused 状态超过 20 分钟时会触发该告警，此时需要参考DM 故障诊断进行处理。

12.8.6.7.3 relay log 告警

`DM_relay_process_exits_with_error`

当 relay log 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

`DM_remain_storage_of_relay_log`

当 relay log 所在磁盘的剩余可用容量小于 10G 时会触发该告警，对应的处理方法包括：

- 手动清理该磁盘上其他无用数据以增加可用容量。
- 尝试调整 relay log 的自动清理策略或执行手动清理。
- 使用 `pause-relay` 命令暂停 relay log 的拉取，并在磁盘空间合适之后使用 `resume-relay` 命令恢复。需要注意上游数据源不要清理尚未拉取的 binlog。

`DM_relay_log_data_corruption`

当 relay log 处理单元在校验从上游读取到的 binlog event 且发现 checksum 信息异常时会转为 Paused 状态并立即触发告警，此时需要参考DM 故障诊断进行处理。

`DM_fail_to_read_binlog_from_master`

当 relay log 处理单元在尝试从上游读取 binlog event 发生错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

`DM_fail_to_write_relay_log`

当 relay log 处理单元在尝试将 binlog event 写入 relay log 文件发生错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

`DM_binlog_file_gap_between_master_relay`

当 relay log 处理单元已拉取到的最新的 binlog 文件个数落后于当前上游 MySQL/MariaDB 超过 1 个（不含 1 个）且持续 10 分钟时会触发该告警，此时需要参考性能问题及处理方法对 relay log 处理单元相关的性能问题进行排查与处理。

12.8.6.7.4 Dump/Load 告警

`DM_dump_process_exists_with_error`

当 Dump 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

`DM_load_process_exists_with_error`

当 Load 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

12.8.6.7.5 Binlog replication 告警

`DM_sync_process_exists_with_error`

当 Binlog replication 处理单元遇到错误时，会转为 Paused 状态并立即触发该告警，此时需要参考DM 故障诊断进行处理。

`DM_binlog_file_gap_between_master_syncer`

当 Binlog replication 处理单元已处理到的最新的 binlog 文件个数落后于当前上游 MySQL/MariaDB 超过 1 个（不含 1 个）且持续 10 分钟时 DM 会触发该告警，此时需要参考性能问题及处理方法对 Binlog replication 处理单元相关的性能问题进行排查与处理。

`DM_binlog_file_gap_between_relay_syncer`

当 Binlog replication 处理单元已处理到的最新的 binlog 文件个数落后于当前 relay log 处理单元超过 1 个（不含 1 个）且持续 10 分钟时 DM 会触发该告警，此时需要参考性能问题及处理方法对 Binlog replication 处理单元相关的性能问题进行排查与处理。

12.8.7 参考手册

12.8.7.0.1 架构

Data Migration 架构

DM 主要包括三个组件：DM-master，DM-worker 和 dmctl。

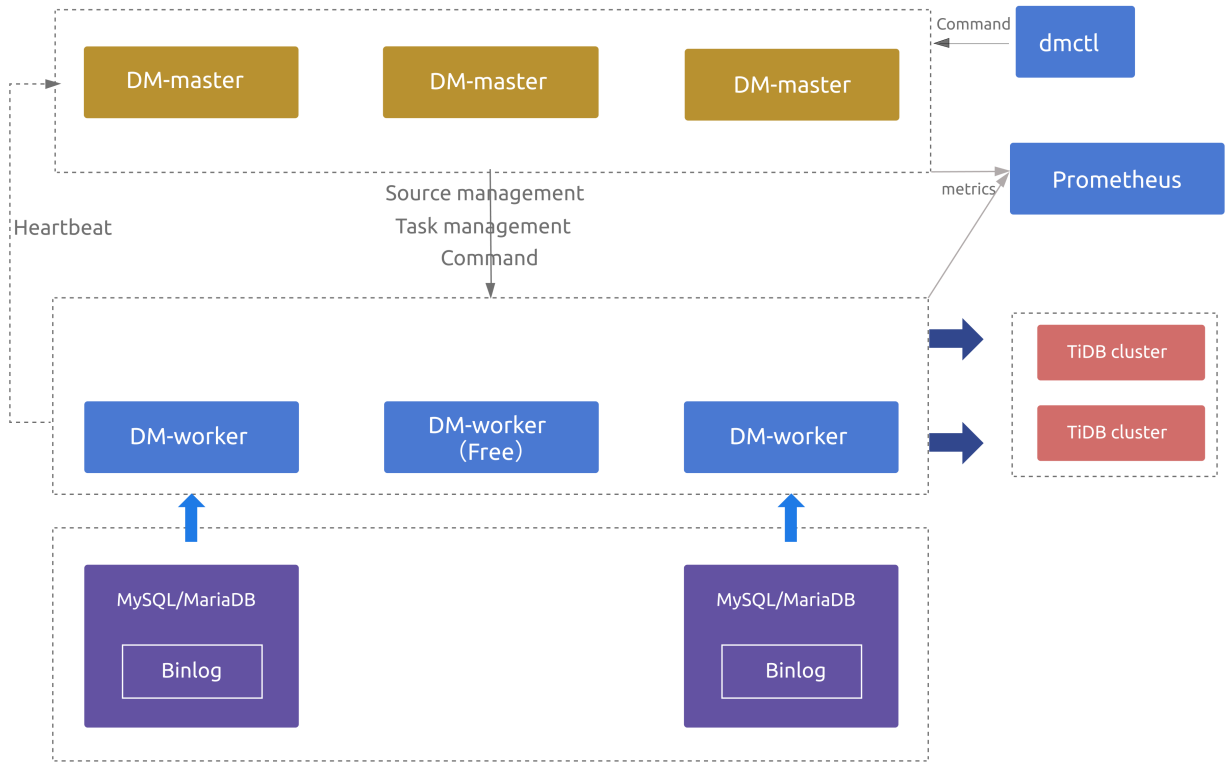


图 294: Data Migration architecture

架构组件

DM-master

DM-master 负责管理和调度数据迁移任务的各项操作。

- 保存 DM 集群的拓扑信息
- 监控 DM-worker 进程的运行状态
- 监控数据迁移任务的运行状态
- 提供数据迁移任务管理的统一入口
- 协调分库分表场景下各个实例分表的 DDL 迁移

DM-worker

DM-worker 负责执行具体的数据迁移任务。

- 将 binlog 数据持久化保存在本地
- 保存数据迁移子任务的配置信息
- 编排数据迁移子任务的运行
- 监控数据迁移子任务的运行状态

有关于 DM-worker 的更多介绍，详见DM-worker 简介。

dmctl

dmctl 是用来控制 DM 集群的命令行工具。

- 创建、更新或删除数据迁移任务
- 查看数据迁移任务状态
- 处理数据迁移任务错误
- 校验数据迁移任务配置的正确性

有关于 dmctl 的使用介绍，详见dmctl 使用。

架构特性

高可用

当部署多个 DM-master 节点时，所有 DM-master 节点将使用内部嵌入的 etcd 组成集群。该 DM-master 集群用于存储集群节点信息、任务配置等元数据，同时通过 etcd 选举出 leader 节点。该 leader 节点用于提供集群管理、数据迁移任务管理相关的各类服务。因此，若可用的 DM-master 节点数超过部署节点的半数，即可正常提供服务。

当部署的 DM-worker 节点数超过上游 MySQL/MariaDB 节点数时，超出上游节点数的相关 DM-worker 节点默认将处于空闲状态。若某个 DM-worker 节点下线或与 DM-master leader 发生网络隔离，DM-master 能自动将与原 DM-worker 节点相关的数据迁移任务调度到其他空闲的 DM-worker 节点上（若原 DM-worker 节点为网络隔离状态，则其会自动停止相关的数据迁移任务）；若无空闲的 DM-worker 节点可供调度，则原 DM-worker 相关的数据迁移任务将暂时挂起，直到有空闲 DM-worker 节点后自动恢复。

注意：

当数据迁移任务处于全量导出或导入阶段时，该迁移任务暂不支持高可用，主要原因为：

- 对于全量导出，MySQL 暂不支持指定从特定快照点导出，也就是说数据迁移任务被重新调度或重启后，无法继续从前一次中断时刻继续导出。
- 对于全量导入，DM-worker 暂不支持跨节点读取全量导出数据，也就是说数据迁移任务被调度到的新 DM-worker 节点无法读取调度发生前原 DM-worker 节点上的全量导出数据。

DM-worker 简介

DM-worker 是 DM (Data Migration) 的一个组件，负责执行具体的数据迁移任务。

其主要功能如下：

- 注册为一台 MySQL 或 MariaDB 服务器的 slave。
- 读取 MySQL 或 MariaDB 的 binlog event，并将这些 event 持久化保存在本地 (relay log)。
- 单个 DM-worker 支持迁移一个 MySQL 或 MariaDB 实例的数据到下游的多个 TiDB 实例。
- 多个 DM-Worker 支持迁移多个 MySQL 或 MariaDB 实例的数据到下游的一个 TiDB 实例。

DM-worker 处理单元

DM-worker 任务包含如下多个逻辑处理单元。

Relay log

Relay log 持久化保存从上游 MySQL 或 MariaDB 读取的 binlog，并对 binlog replication 处理单元提供读取 binlog event 的功能。

其原理和功能与 MySQL relay log 类似，详见 MySQL Relay Log。

dump 处理单元

dump 处理单元从上游 MySQL 或 MariaDB 导出全量数据到本地磁盘。

load 处理单元

load 处理单元读取 dump 处理单元导出的数据文件，然后加载到下游 TiDB。

Binlog replication/sync 处理单元

Binlog replication/sync 处理单元读取上游 MySQL/MariaDB 的 binlog event 或 relay log 处理单元的 binlog event，将这些 event 转化为 SQL 语句，再将这些 SQL 语句应用到下游 TiDB。

DM-worker 所需权限

本小节主要介绍使用 DM-worker 时所需的上下游数据库用户权限以及各处理单元所需的用户权限。

上游数据库用户权限

上游数据库 (MySQL/MariaDB) 用户必须拥有以下权限：

权限	作用域
SELECT	Tables
RELOAD	Global
REPLICATION SLAVE	Global
REPLICATION CLIENT	Global

如果要迁移 db1 的数据到 TiDB，可执行如下的 GRANT 语句：

```
GRANT RELOAD,REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'your_user'@'your_wildcard_of_host'
GRANT SELECT ON db1.* TO 'your_user'@'your_wildcard_of_host';
```

如果还要迁移其他数据库的数据到 TiDB, 请确保已赋予这些库跟 db1 一样的权限。

下游数据库用户权限

下游数据库 (TiDB) 用户必须拥有以下权限：

权限	作用域
SELECT	Tables
INSERT	Tables
UPDATE	Tables
DELETE	Tables
CREATE	Databases，tables
DROP	Databases，tables
ALTER	Tables
INDEX	Tables

对要执行迁移操作的数据库或表执行下面的 GRANT 语句：

```
GRANT SELECT,INSERT,UPDATE,DELETE,CREATE,DROP,ALTER,INDEX  ON db.table TO 'your_user'@'
    ↪ your_wildcard_of_host';
```

处理单元所需的最小权限

处理单元	最小上游 (MySQL/MariaDB) 权限	最小下游 (TiDB) 权限	最小系统权限
Relay log	REPLICATION SLAVE (读取 binlog) REPLICATION CLIENT (show master status, show slave status)	无	本地读/写磁盘
Dump	SELECTRELOAD (获取读锁将表数据刷到磁盘，进行一些操作后，再释放读锁对表进行解锁)	无	本地写磁盘

处理单元	最小上游 (MySQL/MariaDB) 权限	最小下游 (TiDB) 权限	最小系统权限
Load	无	SELECT（查询 checkpoint 历史） CREATE（创建数据库或表） DELETE（删除 checkpoint） INSERT（插入 dump 数据）	读/写本地文件
Binlog replication	REPLICATION SLAVE（读 binlog） REPLICATION CLIENT (show master status, show slave status)	SELECT（显示索引和列）INSERT (DML)UPDATE (DML)DELETE (DML)CREATE（创建数据库或表） DROP（删除数据库或表）ALTER（修改表）INDEX（创建或删除索引）	本地读/写磁盘

> **注意：**
>
> 这些权限并非一成不变。随着需求改变，这些权限也可能会改变。

12.8.7.0.2 命令行

DM 命令行参数

本文档介绍 DM 中各组件的主要命令行参数。

DM-master

`--advertise-addr`

- DM-master 用于接收客户端请求的外部地址
- 默认值为 `"{master-addr}"`
- 可选参数，可以为 `"域名:port"` 的形式

`--advertise-peer-urls`

- DM-master 节点间通信的外部连接地址
- 默认值为 `"{peer-urls}"`

- 可选参数，可以为 "http(s)://域名:port" 的形式

--config

- DM-master 配置文件路径
- 默认值为 ""
- 可选参数

--data-dir

- DM-master 用于存储自身数据的目录
- 默认值为 "default.{name}"
- 可选参数

--initial-cluster

- 用于 bootstrap DM-master 集群的 "{节点名}={外部地址}" 列表
- 默认值为 "{name}={advertise-peer-urls}"
- 在未指定 join 参数时需要指定该参数。一个 3 节点集群的配置示例为 "dm-master-1=http
  ↪ ://172.16.15.11:8291,dm-master-2=http://172.16.15.12:8291,dm-master-3=http://172.16.15.13:8291"
  ↪

--join

- DM-master 节点加入到已有集群时，已有集群的 advertise-addr 地址列表
- 默认值为 ""
- 未指定 initial-cluster 参数时需要指定该参数。一个新节点加入到一个已有 2 个节点的集群的示例为 "172.16.15.11:8261,172.16.15.12:8261"

--log-file

- log 输出文件名
- 默认值为 ""
- 可选参数

-L

- log 级别
- 默认值为 "info"
- 可选参数

--master-addr

- DM-master 监听客户端请求的地址

- 默认值为 ""
- 必选参数

--name

- DM-master 节点名称
- 默认值为 "dm-master-{hostname}"
- 必选参数

--peer-urls

- DM-master 节点间通信的监听地址
- 默认值为 "http://127.0.0.1:8291"
- 必选参数

DM-worker

--advertise-addr

- DM-worker 用于接受客户端请求的外部地址
- 默认值为 "{worker-addr}"
- 可选参数，可以为 "域名:port" 的形式

--config

- DM-worker 配置文件路径
- 默认值为 ""
- 可选参数

--join

- DM-worker 注册到集群时，相应集群的 DM-master 节点的 {advertise-addr} 列表
- 默认值为 ""
- 必选参数，一个 3 DM-master 节点的集群配置示例为 "172.16.15.11:8261,172.16.15.12:8261,172.16.15.13:8261"
  ↪

--log-file

- log 输出文件名
- 默认值为 ""
- 可选参数

-L

- log 级别

- 默认值为 "info"
- 可选参数

--name

- DM-worker 节点名称
- 默认值为 "{advertise-addr}"
- 必选参数

--worker-addr

- DM-worker 监听客户端请求的地址
- 默认值为 ""
- 必选参数

dmctl

--config

- dmctl 配置文件路径
- 默认值为 ""
- 可选参数

--master-addr

- dmctl 要连接的集群的任意 DM-master 节点的 {advertise-addr}
- 默认值为 ""
- 需要与 DM-master 交互时为必选参数

--encrypt

- 将明文数据库密码加密成密文
- 默认值为 ""
- 指定该参数时，仅用于加密明文而不会与 DM-master 交互

--decrypt

- 将使用 dmctl 加密过的密文解密为明文
- 默认值为 ""
- 指定该参数时，仅用于解密密文而不会与 DM-master 交互

12.8.7.0.3 配置文件

DM 配置简介

本文档简要介绍 DM (Data Migration) 的配置文件和数据迁移任务的配置。

配置文件

- `dm-master.toml`：DM-master 进程的配置文件，包括 DM-master 的拓扑信息、日志等各项配置。配置说明详见DM-master 配置文件介绍。
- `dm-worker.toml`：DM-worker 进程的配置文件，包括 DM-worker 的拓扑信息、日志等各项配置。配置说明详见DM-worker 配置文件介绍。
- `source.yaml`：上游数据库 MySQL/MariaDB 相关配置。配置说明详见上游数据库配置文件介绍。

迁移任务配置

创建数据迁移任务

具体步骤如下：

1. 使用 dmctl 将数据源配置加载到 DM 集群；
2. 参考数据任务配置向导来创建 `your_task.yaml`；
3. 使用 dmctl 创建数据迁移任务。

关键概念

DM 配置的关键概念如下：

概念	解释	配置文件
source-id	唯一确定一个 MySQL 或 MariaDB 实例，或者一个具有主从结构的复制组，字符串长度不大于 32	`source.yaml` 的 `source-id`；`task.yaml` 的 `source-id`
DM-master ID	唯一确定一个 DM-master（取值于 `dm-master.toml` 的 `master-addr` 参数）	`dm-master.toml` 的 `master-addr`
DM-worker ID	唯一确定一个 DM-worker（取值于 `dm-worker.toml` 的 `worker-addr` 参数）	`dm-worker.toml` 的 `worker-addr`

上游数据库配置文件介绍

本文介绍上游数据库的配置文件，包括配置文件示例与配置项说明。

配置文件示例

上游数据库的示例配置文件如下所示：

```
source-id: "mysql-replica-01"

###### 是否开启 GTID
enable-gtid: false

###### 是否开启 relay log
enable-relay: false       # 该配置项从 DM v2.0.2 版本起弃用，使用 `start-relay` 命令开启 relay
    ↪ log
relay-binlog-name: ""     # 拉取上游 binlog 的起始文件名
relay-binlog-gtid: ""     # 拉取上游 binlog 的起始 GTID
relay-dir: "relay-dir"    # 存储 relay log 的目录，默认值为 "relay-dir"

from:
  host: "127.0.0.1"
  port: 3306
  user: "root"
  password: "ZqMLjZ2j5khNelDEfDoUhkD5aV5fIJOe0fiog9w=" # 推荐使用 dmctl
      ↪ 对上游数据库的用户密码加密之后的密码
  security:                       # 上游数据库 TLS 相关配置
    ssl-ca: "/path/to/ca.pem"
    ssl-cert: "/path/to/cert.pem"
    ssl-key: "/path/to/key.pem"

###### purge:
######   interval: 3600
######   expires: 0
######   remain-space: 15

###### checker:
######   check-enable: true
######   backoff-rollback: 5m0s
######   backoff-max: 5m0s       # backoff 的最大值，不能小于 1s

###### 从 DM v2.0.2 开始，Binlog event filter 也可以在上游数据库配置文件中进行配置
###### case-sensitive: false
###### filters:
###### - schema-pattern: dmctl
######   table-pattern: t_1
######   events: []
######   sql-pattern:
```

```
######   - alter table .* add column `aaa` int
######   action: Ignore
```

> **注意：**
>
> 在 DM v2.0.1 版本中，请勿同时配置 `enable-gtid` 与 `enable-relay` 为 `true`，否则可能引发增量数据丢失问题。

配置项说明

Global 配置

配置项	说明
`source-id`	标识一个 MySQL 实例。
`enable-gtid`	是否使用 GTID 方式从上游拉取 binlog，默认值为 false。一般情况下不需要手动配置，如果上游数据库启用了 GTID 支持，且需要做主从切换，则将该配置项设置为 true。
`enable-relay`	是否开启 relay log，默认值为 false。自 DM v2.0.2 版本起，该配置项弃用，需使用 `start-relay` 命令开启 relay log。
`relay-binlog-name`	拉取上游 binlog 的起始文件名，例如“mysql-bin.000002”，该配置在 `enable-gtid` 为 false 的情况下生效。如果不配置该项，DM-worker 将从最新时间点的 binlog 文件开始拉取 binlog，一般情况下不需要手动配置。
`relay-binlog-gtid`	拉取上游 binlog 的起始 GTID，例如“e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849”，该配置在 `enable-gtid` 为 true 的情况下生效。如果不配置该项，DM-worker 将从最新时间点的 binlog GTID 开始拉取 binlog，一般情况下不需要手动配置。
`relay-dir`	存储 relay log 的目录，默认值为“./relay_log”。
`host`	上游数据库的 host。
`port`	上游数据库的端口。
`user`	上游数据库使用的用户名。
`password`	上游数据库的用户密码。注意：推荐使用 dmctl 加密后的密码。
`security`	上游数据库 TLS 相关配置。配置的证书文件路径需能被所有节点访问。若配置为本地路径，则集群所有节点需要将证书文件拷贝一份放在各节点机器相同的路径位置上。

relay log 清理策略配置（purge 配置项）

一般情况下不需要手动配置，如果 relay log 数据量较大，磁盘空间不足，则可以通过设置该配置项来避免 relay log 写满磁盘。

配置项	说明
`interval`	定期检查 relay log 是否过期的间隔时间，默认值：3600，单位：秒。
`expires`	relay log 的过期时间，默认值为 0，单位：小时。未由 relay 处理单元进行写入、或已有数据迁移任务当前或未来不需要读取的 relay log 在超过过期时间后会被 DM 删除。如果不设置则 DM 不会自动清理过期的 relay log。
`remain-space`	设置最小的可用磁盘空间。当磁盘可用空间小于这个值时，DM-worker 会尝试删除 relay log，默认值：15，单位：GB。

> 注意：
>
> 仅在 `interval` 不为 0 且 `expires` 和 `remain-space` 两个配置项中至少有一个不为 0 的情况下 DM 的自动清理策略才会生效。

任务状态检查配置（checker 配置项）

DM 会定期检查当前任务状态以及错误信息，判断恢复任务能否消除错误，并自动尝试恢复任务进行重试。DM 会使用指数回退策略调整检查间隔。这些行为可以通过如下配置进行调整：

配置项	说明
`check-enable`	启用自动重试功能。
`backoff-rollback`	如果指数回退策略的间隔大于该值，且任务处于正常状态，尝试减小间隔。
`backoff-max`	指数回退策略的间隔的最大值，该值必须大于 1 秒。

Binlog event filter

从 DM v2.0.2 开始，Binlog event filter 也可以在上游数据库配置文件中进行配置。

配置项	说明
`case-sensitive`	Binlog event filter 标识符是否大小写敏感。默认值：false。
`filters`	配置 Binlog event filter，含义见 Binlog event filter 参数解释。

DM 任务完整配置文件介绍

本文档主要介绍 Data Migration (DM) 的任务完整的配置文件，包含全局配置 和实例配置 两部分。

关于各配置项的功能和配置，请参阅数据迁移功能。

关键概念

关于包括 source-id 和 DM-worker ID 在内的关键概念的介绍，请参阅关键概念。

完整配置文件示例

下面是一个完整的配置文件示例，通过该示例可以完成复杂的数据迁移功能。

```
---

###### ----------- 全局配置 -----------
####### ********* 基本信息配置 *********
name: test                        # 任务名称，需要全局唯一
task-mode: all                    # 任务模式，可设为 "full" - "只进行全量数据迁移"、"incremental" -
    ↪  "Binlog 实时同步"、"all" - "全量 + Binlog 实时同步"
is-sharding: true                 # 该配置项从 DM v2.0.0 起弃用，其功能被 `shard-mode` 取代，
    ↪ 建议使用 `shard-mode` 代替 `is-sharding`
shard-mode: "pessimistic"         # 任务协调模式，可选的模式有 ""、"pessimistic、"optimistic"。
    ↪ 默认值为 "" 即无需协调。如果是分库分表合并任务，请设置为悲观协调模式 "pessimistic"。
                                  # 在 v2.0.6 版本后乐观模式逐渐成熟，
                                      ↪ 深入了解乐观协调模式的原理和使用限制后，
                                      ↪ 也可以设置为乐观协调模式 "optimistic"
meta-schema: "dm_meta"            # 下游储存 `meta` 信息的数据库
###### timezone: "Asia/Shanghai"      # 指定数据迁移任务时 SQL Session 使用的时区。DM
    ↪ 默认使用目标库的全局时区配置进行数据迁移，并且自动确保同步数据的正确性。
    ↪ 使用自定义时区依然可以确保整个流程的正确性，但一般不需要手动指定。

case-sensitive: false             # schema/table 是否大小写敏感
online-ddl: true                  # 支持上游 "gh-ost" 、"pt" 的自动处理
online-ddl-scheme: "gh-ost"       # `online-ddl-scheme` 在未来将会被弃用，建议使用 `online-ddl`
    ↪ 代替 `online-ddl-scheme`
clean-dump-file: true             # 是否清理 dump 阶段产生的文件，包括 metadata 文件、建库建表 SQL
    ↪ 文件以及数据导入 SQL 文件
collation_compatible: "loose"     # 同步 CREATE 语句中缺省 Collation 的方式，可选 "loose" 和 "
    ↪ strict"，默认为 "loose"。"loose" 模式不会显式补充上游缺省的 Collation，"strict"
    ↪ 会显式补充上游缺省的 Collation。当使用 "strict" 模式，但下游不支持上游缺省的 Collation 时
    ↪ ，下游可能会报错。

target-database:                  # 下游数据库实例配置
  host: "192.168.0.1"
  port: 4000
  user: "root"
  password: "/Q7B9DizNLLTTfiZHv9WoEAKamfpIUs="  # 推荐使用经 `dmctl encrypt` 加密后的密码
  max-allowed-packet: 67108864                 # 设置 DM 内部连接 TiDB 服务器时，TiDB 客户端的 "
      ↪ max_allowed_packet" 限制（即接受的最大数据包限制），单位为字节，默认 67108864 (64 MB)
                                               # 该配置项从 DM v2.0.0 起弃用，DM 会自动获取连接
```

```
                                                    ↪ TiDB 的 "max_allowed_packet"
  session:                                        # 设置 TiDB 的 session 变量，在 v1.0.6 版本引入。
      ↪ 更多变量及解释参见 `https://docs.pingcap.com/zh/tidb/stable/system-variables`
    sql_mode: "ANSI_QUOTES,NO_ZERO_IN_DATE,NO_ZERO_DATE" # 从 DM v2.0.0 起，
        ↪ 如果配置文件中没有出现该项，DM 会自动从下游 TiDB 中获得适合用于 "sql_mode" 的值。
        ↪ 手动配置该项具有更高优先级
    tidb_skip_utf8_check: 1                                 # 从 DM v2.0.0 起，
        ↪ 如果配置文件中没有出现该项，DM 会自动从下游 TiDB 中获得适合用于 "tidb_skip_utf8_check
        ↪ " 的值。手动配置该项具有更高优先级
    tidb_constraint_check_in_place: 0
  security:                         # 下游 TiDB TLS 相关配置
    ssl-ca: "/path/to/ca.pem"
    ssl-cert: "/path/to/cert.pem"
    ssl-key: "/path/to/key.pem"

####### ******** 功能配置集 **********

routes:                             # 上游和下游表之间的路由 table routing 规则集
  route-rule-1:                     # 配置名称
    schema-pattern: "test_*"        # 库名匹配规则，支持通配符 "*" 和 "?"
    table-pattern: "t_*"            # 表名匹配规则，支持通配符 "*" 和 "?"
    target-schema: "test"           # 目标库名称
    target-table: "t"               # 目标表名称
  route-rule-2:
    schema-pattern: "test_*"
    target-schema: "test"

filters:                                        # 上游数据库实例匹配的表的 binlog event filter
    ↪ 规则集
  filter-rule-1:                                # 配置名称
    schema-pattern: "test_*"                    # 库名匹配规则，支持通配符 "*" 和 "?"
    table-pattern: "t_*"                        # 表名匹配规则，支持通配符 "*" 和 "?"
    events: ["truncate table", "drop table"]    # 匹配哪些 event 类型
    action: Ignore                              # 对与符合匹配规则的 binlog 迁移（Do）还是忽略(
        ↪ Ignore)
  filter-rule-2:
    schema-pattern: "test_*"
    events: ["all dml"]
    action: Do

expression-filter:                  # 定义数据源迁移行变更的过滤规则，可以定义多个规则
  # 过滤 `expr_filter`.`tbl` 的 c 为偶数的插入
  even_c:                           # 规则名称
    schema: "expr_filter"           # 要匹配的上游数据库库名，不支持通配符匹配或正则匹配
    table: "tbl"                    # 要匹配的上游表名，不支持通配符匹配或正则匹配
```

```
    insert-value-expr: "c % 2 = 0"

block-allow-list:                     # 定义数据源迁移表的过滤规则，可以定义多个规则。如果 DM
    ↪ 版本早于 v2.0.0-beta.2 则使用 black-white-list
  bw-rule-1:                          # 规则名称
    do-dbs: ["~^test.*", "user"]      # 迁移哪些库
    ignore-dbs: ["mysql", "account"] # 忽略哪些库
    do-tables:                        # 迁移哪些表
    - db-name: "~^test.*"
      tbl-name: "~^t.*"
    - db-name: "user"
      tbl-name: "information"
  bw-rule-2:                          # 规则名称
    ignore-tables:                    # 忽略哪些表
    - db-name: "user"
      tbl-name: "log"

mydumpers:                            # dump 处理单元的运行配置参数
  global:                             # 配置名称
    threads: 4                        # dump 处理单元从上游数据库实例导出数据的线程数量，默认值为
        ↪ 4
    chunk-filesize: 64                # dump 处理单元生成的数据文件大小，默认值为 64，单位为 MB
    extra-args: "--consistency none" # dump 处理单元的其他参数，不需要在 extra-args 中配置 table-
        ↪ list，DM 会自动生成

loaders:                              # load 处理单元的运行配置参数
  global:                             # 配置名称
    pool-size: 16                     # load 处理单元并发执行 dump 处理单元的 SQL 文件的线程数量，
        ↪ 默认值为 16，当有多个实例同时向 TiDB 迁移数据时可根据负载情况适当调小该值
    dir: "./dumped_data"              # dump 处理单元输出 SQL 文件的目录，同时也是 load
        ↪ 处理单元读取文件的目录。该配置项的默认值为 "./dumped_data"。
        ↪ 同实例对应的不同任务必须配置不同的目录


syncers:                              # sync 处理单元的运行配置参数
  global:                             # 配置名称
    worker-count: 16                  # 应用已传输到本地的 binlog 的并发线程数量，默认值为 16。
        ↪ 调整此参数不会影响上游拉取日志的并发，但会对下游产生显著压力。
    batch: 100                        # sync 迁移到下游数据库的一个事务批次 SQL 语句数，默认值为
        ↪ 100，建议一般不超过 500。
    enable-ansi-quotes: true          # 若 `session` 中设置 `sql-mode: "ANSI_QUOTES"`，
        ↪ 则需开启此项

    # 设置为 true，则将来自上游的 `INSERT` 改写为 `REPLACE`，将 `UPDATE` 改写为 `DELETE` 与 `
        ↪ REPLACE`，保证在表结构中存在主键或唯一索引的条件下迁移数据时可以重复导入 DML。
```

```
    safe-mode: false
    # 设置为 true，DM 会在不增加延迟的情况下，尽可能地将上游对同一条数据的多次操作压缩成一次操作
        ↪ 。
    # 如 INSERT INTO tb(a,b) VALUES(1,1); UPDATE tb SET b=11 WHERE a=1; 会被压缩成 INSERT INTO tb
        ↪ (a,b) VALUES(1,11); 其中 a 为主键
    # 如 UPDATE tb SET b=1 WHERE a=1; UPDATE tb(a,b) SET b=2 WHERE a=1; 会被压缩成 UPDATE tb(a,b)
        ↪  SET b=2 WHERE a=1; 其中 a 为主键
    # 如 DELETE FROM tb WHERE a=1; INSERT INTO tb(a,b) VALUES(1,1); 会被压缩成 REPLACE INTO tb(a,
        ↪ b) VALUES(1,1); 其中 a 为主键
    compact: false
    # 设置为 true，DM 会尽可能地将多条同类型的语句合并到一条语句中，生成一条带多行数据的 SQL 语句
        ↪ 。
    # 如 INSERT INTO tb(a,b) VALUES(1,1); INSERT INTO tb(a,b) VALUES(2,2); 会变成 INSERT INTO tb(
        ↪ a,b) VALUES(1,1),(2,2);
    # 如 UPDATE tb SET b=11 WHERE a=1; UPDATE tb(a,b) set b=22 WHERE a=2; 会变成 INSERT INTO tb(a
        ↪ ,b) VALUES(1,11),(2,22) ON DUPLICATE KEY UPDATE a=VALUES(a), b=VALUES(b); 其中 a
        ↪ 为主键
    # 如 DELETE FROM tb WHERE a=1; DELETE FROM tb WHERE a=2 会变成 DELETE FROM tb WHERE (a) IN
        ↪ (1),(2)；其中 a 为主键
    multiple-rows: false

###### ----------- 实例配置 -----------
mysql-instances:
  -
    source-id: "mysql-replica-01"           # 对应 source.toml 中的 `source-id`
    meta:                                   # `task-mode` 为 `incremental` 且下游数据库的 `
        ↪ checkpoint` 不存在时 binlog 迁移开始的位置；如果 checkpoint 存在，则以 `checkpoint`
        ↪ 为准
      binlog-name: binlog.000001
      binlog-pos: 4
      binlog-gtid: "03fc0263-28c7-11e7-a653-6c0b84d59f30:1-7041423,05474d3c-28c7-11e7-8352-203
          ↪ db246dd3d:1-170"  # 对于 source 中指定了 `enable-gtid: true` 的增量任务，
          ↪ 需要指定该值

    route-rules: ["route-rule-1", "route-rule-2"]    # 该上游数据库实例匹配的表到下游数据库的
        ↪ table routing 规则名称
    filter-rules: ["filter-rule-1", "filter-rule-2"] # 该上游数据库实例匹配的表的 binlog event
        ↪ filter 规则名称
    block-allow-list:  "bw-rule-1"                   # 该上游数据库实例匹配的表的 block-allow-
        ↪ list 过滤规则名称，如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list
    expression-filters: ["even_c"]                   # 使用名为 even_c 的表达式过滤规则

    mydumper-config-name: "global"          # mydumpers 配置的名称
    loader-config-name: "global"            # loaders 配置的名称
    syncer-config-name: "global"            # syncers 配置的名称
```

```
-
  source-id: "mysql-replica-02"  # 对应 source.toml 中的 `source-id`
  mydumper-thread: 4             # dump 处理单元用于导出数据的线程数量，等同于 mydumpers
      ↪ 配置中的 `threads`，当同时指定它们时 `mydumper-thread` 优先级更高
  loader-thread: 16              # load 处理单元用于导入数据的线程数量，等同于 loaders 配置中的
      ↪  `pool-size`，当同时指定它们时 `loader-thread` 优先级更高。当有多个实例同时向 TiDB
      ↪ 迁移数据时可根据负载情况适当调小该值
  syncer-thread: 16              # sync 处理单元用于复制增量数据的线程数量，等同于 syncers
      ↪ 配置中的 `worker-count`，当同时指定它们时 `syncer-thread` 优先级更高。
      ↪ 当有多个实例同时向 TiDB 迁移数据时可根据负载情况适当调小该值
```

配置顺序

通过上面的配置文件示例，可以看出配置文件总共分为两个部分：全局配置和实例配置，其中全局配置又分为基本信息配置和实例配置，配置顺序如下：

1. 编辑全局配置。
2. 根据全局配置编辑实例配置。

全局配置

任务基本信息配置

配置任务的基本信息，配置项的说明参见以上示例配置文件中的注释。其中 `task-mode` 需要特殊说明：

`task-mode`

- 描述：任务模式，可以通过任务模式来指定需要执行的数据迁移工作。
- 值为字符串（`full`，`incremental` 或 `all`）。
  - `full`：只全量备份上游数据库，然后将数据全量导入到下游数据库。
  - `incremental`：只通过 binlog 把上游数据库的增量修改复制到下游数据库，可以设置实例配置的 `meta` 配置项来指定增量复制开始的位置。
  - `all`：`full` + `incremental`。先全量备份上游数据库，将数据全量导入到下游数据库，然后从全量数据备份时导出的位置信息 (binlog position) 开始通过 binlog 增量复制数据到下游数据库。

功能配置集

全局配置主要包含下列功能配置集：

配置项	说明
`routes`	上游和下游表之间的路由 table routing 规则集。如果上游与下游的库名、表名一致，则不需要配置该项。使用场景及示例配置参见 Table Routing
`filters`	上游数据库实例匹配的表的 binlog event filter 规则集。如果不需要对 binlog 进行过滤，则不需要配置该项。使用场景及示例配置参见 Binlog Event Filter

配置项	说明
block-allow-list	该上游数据库实例匹配的表的 block & allow lists 过滤规则集。建议通过该项指定需要迁移的库和表，否则会迁移所有的库和表。使用场景及示例配置参见Block & Allow Lists
mydumpers	dump 处理单元的运行配置参数。如果默认配置可以满足需求，则不需要配置该项，也可以只使用 mydumper-thread 对 thread 配置项单独进行配置。
loaders	load 处理单元的运行配置参数。如果默认配置可以满足需求，则不需要配置该项，也可以只使用 loader-thread 对 pool-size 配置项单独进行配置。
syncers	sync 处理单元的运行配置参数。如果默认配置可以满足需求，则不需要配置该项，也可以只使用 syncer-thread 对 worker-count 配置项单独进行配置。

各个功能配置集的参数及解释参见完整配置文件示例中的注释说明。

实例配置

本小节定义具体的数据迁移子任务，DM 支持从单个或者多个上游 MySQL 实例迁移数据到同一个下游数据库实例。

在该项配置中设置数据迁移子任务中各个功能对应的配置集中的配置名称，关于这些配置项的更多配置细节，参见功能配置集的相关配置项，对应关系如下：

配置项	相关配置项
route-rules	routes
filter-rules	filters
block-allow-list	block-allow-list
mydumper-config-name	mydumpers
loader-config-name	loaders
syncer-config-name	syncers

DM-master 配置文件介绍

本文介绍 DM-master 的配置文件，包括配置文件示例与配置项说明。

配置文件示例

DM-master 的示例配置文件如下所示：

```
name = "dm-master"

###### log configuration
log-level = "info"
log-file = "dm-master.log"
```

```
###### DM-master listening address
master-addr = ":8261"
advertise-addr = "127.0.0.1:8261"

###### URLs for peer traffic
peer-urls = "http://127.0.0.1:8291"
advertise-peer-urls = "http://127.0.0.1:8291"

###### cluster configuration
initial-cluster = "master1=http://127.0.0.1:8291,master2=http://127.0.0.1:8292,master3=http
    ↪ ://127.0.0.1:8293"
join = ""

ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/cert.pem"
ssl-key = "/path/to/key.pem"
cert-allowed-cn = ["dm"]
```

配置项说明

Global 配置

配置项	说明
`name`	标识一个 DM-master。
`log-level`	日志级别：debug、info、warn、error、fatal。默认为 info。
`log-file`	日志文件，如果不配置，日志会输出到标准输出中。
`master-addr`	DM-master 服务的地址，可以省略 IP 信息，例如：“:8261”。
`advertise-addr`	DM-master 向外界宣告的地址。
`peer-urls`	DM-master 节点的对等 URL。
`advertise-peer-urls`	DM-master 向外界宣告的对等 URL。默认为 `peer-urls` 的值。
`initial-cluster`	初始集群中所有 DM-master 的 `advertise-peer-urls` 的值。
`join`	集群里已有的 DM-master 的 `advertise-peer-urls` 的值。如果是新加入的 DM-master 节点，使用 `join` 替代 `initial-cluster`。
`ssl-ca`	DM-master 组件用于与其它组件连接的 SSL CA 证书所在的路径
`ssl-cert`	DM-master 组件用于与其它组件连接的 PEM 格式的 X509 证书所在的路径
`ssl-key`	DM-master 组件用于与其它组件连接的 PEM 格式的 X509 密钥所在的路径
`cert-allowed-cn`	证书检查 Common Name 列表

配置项	说明

DM-worker 配置文件介绍

本文介绍 DM-worker 的配置文件，包括配置文件示例与配置项说明。

配置文件示例

```
###### Worker Configuration.

name = "worker1"

###### Log configuration.
log-level = "info"
log-file = "dm-worker.log"

###### DM-worker listen address.
worker-addr = ":8262"
advertise-addr = "127.0.0.1:8262"
join = "http://127.0.0.1:8261,http://127.0.0.1:8361,http://127.0.0.1:8461"

keepalive-ttl = 60
relay-keepalive-ttl = 1800 # 版本 2.0.2 新增
###### relay-dir = "relay_log" # 版本 5.4.0 新增。使用相对路径时注意结合部署、
    ↪ 启动方式确认路径位置。

ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/cert.pem"
ssl-key = "/path/to/key.pem"
cert-allowed-cn = ["dm"]
```

配置项说明

Global 配置

配置项	说明
name	标识一个 DM-worker。
log-level	日志级别：debug、info、warn、error、fatal。默认为 info。
log-file	日志文件，如果不配置日志会输出到标准输出中。
worker-addr	DM-worker 服务的地址，可以省略 IP 信息，例如：“:8262”。
advertise-addr	DM-worker 向外界宣告的地址。
join	对应一个或多个 DM-master 配置中的master-addr。
keepalive-ttl	当绑定的上游数据源没有启用 relay log 时，DM-worker 向 DM-master 保持存活的周期，单位为秒。默认是 60 秒。

配置项	说明
relay-keepalive ↪ -ttl	当绑定的上游数据源启用 relay log 时，DM-worker 向 DM-master 保持存活的周期，单位为秒。默认是 1800 秒。在版本 2.0.2 新增。
relay-dir	当绑定的上游数据源启用 relay log 时，DM-worker 将 relay log 保存在该路径下。该配置优先级比上游数据源配置更高。在版本 5.4.0 新增。
ssl-ca	DM-worker 组件用于与其它组件连接的 SSL CA 证书所在的路径
ssl-cert	DM-worker 组件用于与其它组件连接的 PEM 格式的 X509 证书所在的路径
ssl-key	DM-worker 组件用于与其它组件连接的 PEM 格式的 X509 密钥所在的路径
cert-allowed-cn	证书检查 Common Name 列表

12.8.7.0.4 使用 OpenAPI 运维集群

DM 提供 OpenAPI 功能，您可以通过 OpenAPI 方便地对 DM 集群进行查询和运维操作。OpenAPI 的功能范围和dmctl工具相当。如需开启 OpenAPI，请在 DM-master 的配置文件中增加如下配置项：

```
openapi = true
```

> 注意：
>
> - DM 提供符合 OpenAPI 3.0.0 标准的 Spec 文档，其中包含了所有 API 的请求参数和返回体，你可自行复制到如 Swagger Editor 等工具中在线预览文档。
> - 部署 DM-master 后，你可访问 `http://{master-addr}/api/v1/docs` 在线预览文档。

你可以通过 OpenAPI 完成 DM 集群的如下运维操作：

集群相关 API

- 获取 DM-master 节点信息
- 下线 DM-master 节点
- 获取 DM-worker 节点信息
- 下线 DM-worker 节点

数据源相关 API

- 创建数据源
- 获取数据源列表
- 删除数据源

- 获取数据源状态
- 对数据源开启 relay-log 功能
- 对数据源停止 relay-log 功能
- 对数据源暂停 relay-log 功能
- 对数据源恢复 relay-log 功能
- 更改数据源和 DM-worker 的绑定关系
- 获取数据源的数据库名列表
- 获取数据源的指定数据库的表名列表

同步任务相关 API

- 创建同步任务
- 获取同步任务列表
- 停止同步任务
- 获取同步任务状态
- 暂停同步任务
- 恢复同步任务
- 获取同步任务关联数据源的数据库名列表
- 获取同步任务关联数据源的数据表名列表
- 获取同步任务关联数据源的数据表的创建语句
- 更新同步任务关联数据源的数据表的创建语句
- 删除同步任务关联数据源的数据表

本文档以下部分描述当前提供的 API 的具体使用方法。

API 统一错误格式

对 API 发起的请求后，如发生错误，返回错误信息的格式如下所示：

```
{
    "error_msg": "",
    "error_code": ""
}
```

如上所示，`error_msg` 描述错误信息，`error_code` 则是对应的错误码。

获取 DM-master 节点信息

该接口是一个同步接口，请求成功会返回对应节点的状态信息。

请求 URI

`GET /api/v1/cluster/masters`

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/cluster/masters' \
  -H 'accept: application/json'
```

```
{
  "total": 1,
  "data": [
    {
      "name": "master1",
      "alive": true,
      "leader": true,
      "addr": "127.0.0.1:8261"
    }
  ]
}
```

下线 DM-master 节点

该接口是一个同步接口，请求成功后返回体的 Status Code 是 204。

请求 URI

```
DELETE /api/v1/cluster/masters/{master-name}
```

使用样例

```
curl -X 'DELETE' \
  'http://127.0.0.1:8261/api/v1/cluster/masters/master1' \
  -H 'accept: */*'
```

获取 DM-worker 节点信息

该接口是一个同步接口，请求成功会返回对应节点的状态信息。

请求 URI

```
GET /api/v1/cluster/workers
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/cluster/workers' \
  -H 'accept: application/json'
```

```
{
  "total": 1,
  "data": [
    {
      "name": "worker1",
      "addr": "127.0.0.1:8261",
      "bound_stage": "bound",
      "bound_source_name": "mysql-01"
    }
  ]
}
```

下线 DM-worker 节点

该接口是一个同步接口，请求成功后返回体的 Status Code 是 204。

请求 URI

```
DELETE /api/v1/cluster/workers/{worker-name}
```

使用样例

```
curl -X 'DELETE' \
  'http://127.0.0.1:8261/api/v1/cluster/workers/worker1' \
  -H 'accept: */*'
```

创建数据源

该接口是一个同步接口，请求成功会返回对应数据源信息。

请求 URI

```
POST /api/v1/sources
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "source_name": "mysql-01",
  "host": "127.0.0.1",
  "port": 3306,
  "user": "root",
  "password": "123456",
  "enable_gtid": false,
  "security": {
    "ssl_ca_content": "",
    "ssl_cert_content": "",
    "ssl_key_content": "",
    "cert_allowed_cn": [
      "string"
    ]
  },
  "purge": {
    "interval": 3600,
    "expires": 0,
    "remain_space": 15
  }
}'
```

```
{
  "source_name": "mysql-01",
  "host": "127.0.0.1",
  "port": 3306,
  "user": "root",
  "password": "123456",
  "enable_gtid": false,
  "security": {
    "ssl_ca_content": "",
    "ssl_cert_content": "",
    "ssl_key_content": "",
    "cert_allowed_cn": [
      "string"
    ]
  },
  "purge": {
    "interval": 3600,
    "expires": 0,
    "remain_space": 15
  },
  "status_list": [
    {
      "source_name": "mysql-replica-01",
      "worker_name": "worker-1",
      "relay_status": {
        "master_binlog": "(mysql-bin.000001, 1979)",
        "master_binlog_gtid": "e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849",
        "relay_dir": "./sub_dir",
        "relay_binlog_gtid": "e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849",
        "relay_catch_up_master": true,
        "stage": "Running"
      },
      "error_msg": "string"
    }
  ]
}
```


获取数据源列表

该接口是一个同步接口，请求成功会返回数据源列表信息。

请求 URI

```
GET /api/v1/sources
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/sources?with_status=true' \
  -H 'accept: application/json'
```

```
{
  "data": [
    {
      "enable_gtid": false,
      "host": "127.0.0.1",
      "password": "******",
      "port": 3306,
      "purge": {
        "expires": 0,
        "interval": 3600,
        "remain_space": 15
      },
      "security": null,
      "source_name": "mysql-01",
      "user": "root"
    },
    {
      "enable_gtid": false,
      "host": "127.0.0.1",
      "password": "******",
      "port": 3307,
      "purge": {
        "expires": 0,
        "interval": 3600,
        "remain_space": 15
      },
      "security": null,
      "source_name": "mysql-02",
      "user": "root"
    }
  ],
  "total": 2
}
```

删除数据源

该接口是一个同步接口，请求成功后返回体的 Status Code 是 204。

请求 URI

```
DELETE /api/v1/sources/{source-name}
```

使用样例

```
curl -X 'DELETE' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01?force=true' \
  -H 'accept: application/json'
```

获取数据源状态

该接口是一个同步接口，请求成功会返回对应节点的状态信息。

请求 URI

```
GET /api/v1/sources/{source-name}/status
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-replica-01/status' \
  -H 'accept: application/json'
```

```
{
  "total": 1,
  "data": [
    {
      "source_name": "mysql-replica-01",
      "worker_name": "worker-1",
      "relay_status": {
        "master_binlog": "(mysql-bin.000001, 1979)",
        "master_binlog_gtid": "e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849",
        "relay_dir": "./sub_dir",
        "relay_binlog_gtid": "e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849",
        "relay_catch_up_master": true,
        "stage": "Running"
      },
      "error_msg": "string"
    }
  ]
}
```

对数据源开启 relay-log 功能

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取数据源状态接口获取最新的状态。

请求 URI

```
POST /api/v1/sources/{source-name}/start-relay
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01/start-relay' \
  -H 'accept: */*' \
```

```
  -H 'Content-Type: application/json' \
  -d '{
  "worker_name_list": [
    "worker-1"
  ],
  "relay_binlog_name": "mysql-bin.000002",
  "relay_binlog_gtid": "e9a1fc22-ec08-11e9-b2ac-0242ac110003:1-7849",
  "relay_dir": "./relay_log"
}'
```

对数据源停止 relay-log 功能

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取数据源状态接口获取最新的状态。

请求 URI

```
POST /api/v1/sources/{source-name}/stop-relay
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01/stop-relay' \
  -H 'accept: */*' \
  -H 'Content-Type: application/json' \
  -d '{
  "worker_name_list": [
    "worker-1"
  ]
}'
```

对数据源暂停 relay-log 功能

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取数据源状态接口获取最新的状态。

请求 URI

```
POST /api/v1/sources/{source-name}/pause-relay
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01/pause-relay' \
  -H 'accept: */*'
```

对数据源恢复 relay-log 功能

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取数据源状态接口获取最新的状态。

请求 URI

```
POST /api/v1/sources/{source-name}/resume-relay
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01/resume-relay' \
  -H 'accept: */*'
```

更改数据源和 DM-worker 的绑定关系

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取 DM-worker 节点信息接口获取最新的状态。

请求 URI

```
POST /api/v1/sources/{source-name}/transfer
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/sources/mysql-01/transfer' \
  -H 'accept: */*' \
  -H 'Content-Type: application/json' \
  -d '{
  "worker_name": "worker-1"
}'
```

获取数据源的数据库名列表

该接口是一个同步接口，请求成功会返回对应的列表。

请求 URI

```
GET /api/v1/sources/{source-name}/schemas
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/sources/source-1/schemas' \
  -H 'accept: application/json'
```

```
[
  "db1"
]
```

获取数据源的指定数据库的表名列表

该接口是一个同步接口，请求成功会返回对应的列表。

请求 URI

```
GET /api/v1/sources/{source-name}/schemas/{schema-name}
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/sources/source-1/schemas/db1' \
  -H 'accept: application/json'
```

```
[
  "table1"
]
```

创建同步任务

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取同步任务状态接口获取最新的状态。

请求 URI

```
POST /api/v1/tasks
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/tasks' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "remove_meta": false,
  "task": {
    "name": "task-1",
    "task_mode": "all",
    "shard_mode": "pessimistic",
    "meta_schema": "dm-meta",
    "enhance_online_schema_change": true,
    "on_duplicate": "overwrite",
    "target_config": {
      "host": "127.0.0.1",
      "port": 3306,
      "user": "root",
      "password": "123456",
      "security": {
        "ssl_ca_content": "",
        "ssl_cert_content": "",
        "ssl_key_content": "",
        "cert_allowed_cn": [
          "string"
        ]
      }
    },
    "binlog_filter_rule": {
      "rule-1": {
        "ignore_event": [
          "all dml"
        ],
        "ignore_sql": [
```

```
        "^Drop"
      ]
    },
    "rule-2": {
      "ignore_event": [
        "all dml"
      ],
      "ignore_sql": [
        "^Drop"
      ]
    },
    "rule-3": {
      "ignore_event": [
        "all dml"
      ],
      "ignore_sql": [
        "^Drop"
      ]
    }
  },
  "table_migrate_rule": [
    {
      "source": {
        "source_name": "source-name",
        "schema": "db-*",
        "table": "tb-*"
      },
      "target": {
        "schema": "db1",
        "table": "tb1"
      },
      "binlog_filter_rule": [
        "rule-1",
        "rule-2",
        "rule-3",
      ]
    }
  ],
  "source_config": {
    "full_migrate_conf": {
      "export_threads": 4,
      "import_threads": 16,
      "data_dir": "./exported_data",
      "consistency": "auto"
    },
```

```
      "incr_migrate_conf": {
        "repl_threads": 16,
        "repl_batch": 100
      },
      "source_conf": [
        {
          "source_name": "mysql-replica-01",
          "binlog_name": "binlog.000001",
          "binlog_pos": 4,
          "binlog_gtid": "03fc0263-28c7-11e7-a653-6c0b84d59f30:1-7041423,05474d3c-28c7-11e7
              ↪ -8352-203db246dd3d:1-170"
        }
      ]
    }
  },
  "source_name_list": [
    "source-1"
  ]
}'
```

```
{
  "name": "task-1",
  "task_mode": "all",
  "shard_mode": "pessimistic",
  "meta_schema": "dm-meta",
  "enhance_online_schema_change": true,
  "on_duplicate": "overwrite",
  "target_config": {
    "host": "127.0.0.1",
    "port": 3306,
    "user": "root",
    "password": "123456",
    "security": {
      "ssl_ca_content": "",
      "ssl_cert_content": "",
      "ssl_key_content": "",
      "cert_allowed_cn": [
        "string"
      ]
    }
  },
  "binlog_filter_rule": {
    "rule-1": {
      "ignore_event": [
        "all dml"
```

```
    ],
    "ignore_sql": [
      "^Drop"
    ]
  },
  "rule-2": {
    "ignore_event": [
      "all dml"
    ],
    "ignore_sql": [
      "^Drop"
    ]
  },
  "rule-3": {
    "ignore_event": [
      "all dml"
    ],
    "ignore_sql": [
      "^Drop"
    ]
  }
},
"table_migrate_rule": [
  {
    "source": {
      "source_name": "source-name",
      "schema": "db-*",
      "table": "tb-*"
    },
    "target": {
      "schema": "db1",
      "table": "tb1"
    },
    "binlog_filter_rule": [
      "rule-1",
      "rule-2",
      "rule-3",
    ]
  }
],
"source_config": {
  "full_migrate_conf": {
    "export_threads": 4,
    "import_threads": 16,
    "data_dir": "./exported_data",
```

```
      "consistency": "auto"
    },
    "incr_migrate_conf": {
      "repl_threads": 16,
      "repl_batch": 100
    },
    "source_conf": [
      {
        "source_name": "mysql-replica-01",
        "binlog_name": "binlog.000001",
        "binlog_pos": 4,
        "binlog_gtid": "03fc0263-28c7-11e7-a653-6c0b84d59f30:1-7041423,05474d3c-28c7-11e7
            ↪ -8352-203db246dd3d:1-170"
      }
    ]
  }
}
```

获取同步任务列表

该接口是一个同步接口，请求成功会返回对应的同步任务信息。

请求 URI

```
GET /api/v1/tasks
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/tasks' \
  -H 'accept: application/json'
```

```
{
  "total": 2,
  "data": [
    {
      "name": "task-1",
      "task_mode": "all",
      "shard_mode": "pessimistic",
      "meta_schema": "dm-meta",
      "enhance_online_schema_change": true,
      "on_duplicate": "overwrite",
      "target_config": {
        "host": "127.0.0.1",
        "port": 3306,
        "user": "root",
        "password": "123456",
        "security": {
```

```
    "ssl_ca_content": "",
    "ssl_cert_content": "",
    "ssl_key_content": "",
    "cert_allowed_cn": [
      "string"
    ]
  }
},
"binlog_filter_rule": {
  "rule-1": {
    "ignore_event": [
      "all dml"
    ],
    "ignore_sql": [
      "^Drop"
    ]
  },
  "rule-2": {
    "ignore_event": [
      "all dml"
    ],
    "ignore_sql": [
      "^Drop"
    ]
  },
  "rule-3": {
    "ignore_event": [
      "all dml"
    ],
    "ignore_sql": [
      "^Drop"
    ]
  }
},
"table_migrate_rule": [
  {
    "source": {
      "source_name": "source-name",
      "schema": "db-*",
      "table": "tb-*"
    },
    "target": {
      "schema": "db1",
      "table": "tb1"
    },
```

```
          "binlog_filter_rule": [
            "rule-1",
            "rule-2",
            "rule-3",
          ]
        }
      ],
      "source_config": {
        "full_migrate_conf": {
          "export_threads": 4,
          "import_threads": 16,
          "data_dir": "./exported_data",
          "consistency": "auto"
        },
        "incr_migrate_conf": {
          "repl_threads": 16,
          "repl_batch": 100
        },
        "source_conf": [
          {
            "source_name": "mysql-replica-01",
            "binlog_name": "binlog.000001",
            "binlog_pos": 4,
            "binlog_gtid": "03fc0263-28c7-11e7-a653-6c0b84d59f30:1-7041423,05474d3c-28c7-11e7
                ↪ -8352-203db246dd3d:1-170"
          }
        ]
      }
    }
  ]
}
```

停止同步任务

这是一个异步接口，请求成功的 Status Code 是 204，可通过获取同步任务状态接口获取最新的状态。

请求 URI

```
DELETE /api/v1/tasks/{task-name}
```

使用样例

```
curl -X 'DELETE' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1' \
  -H 'accept: */*'
```

获取同步任务状态

该接口是一个同步接口，请求成功会返回对应节点的状态信息。

请求 URI

```
GET /api/v1/tasks/task-1/status
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/status?stage=running' \
  -H 'accept: application/json'
```

```
{
  "total": 1,
  "data": [
    {
      "name": "string",
      "source_name": "string",
      "worker_name": "string",
      "stage": "runing",
      "unit": "sync",
      "unresolved_ddl_lock_id": "string",
      "load_status": {
        "finished_bytes": 0,
        "total_bytes": 0,
        "progress": "string",
        "meta_binlog": "string",
        "meta_binlog_gtid": "string"
      },
      "sync_status": {
        "total_events": 0,
        "total_tps": 0,
        "recent_tps": 0,
        "master_binlog": "string",
        "master_binlog_gtid": "string",
        "syncer_binlog": "string",
        "syncer_binlog_gtid": "string",
        "blocking_ddls": [
          "string"
        ],
        "unresolved_groups": [
          {
            "target": "string",
            "ddl_list": [
              "string"
            ],
            "first_location": "string",
            "synced": [
              "string"
```

```
            ],
            "unsynced": [
              "string"
            ]
          }
        ],
        "synced": true,
        "binlog_type": "string",
        "seconds_behind_master": 0
      }
    }
  ]
}
```

暂停同步任务

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取同步任务状态接口获取最新的状态。

请求 URI

```
POST /api/v1/tasks/task-1/pause
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/pause' \
  -H 'accept: */*' \
  -H 'Content-Type: application/json' \
  -d '[
  "source-1"
]'
```

恢复同步任务

这是一个异步接口，请求成功的 Status Code 是 200，可通过获取同步任务状态接口获取最新的状态。

请求 URI

```
POST /api/v1/tasks/task-1/resume
```

使用样例

```
curl -X 'POST' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/resume' \
  -H 'accept: */*' \
  -H 'Content-Type: application/json' \
  -d '[
  "source-1"
]'
```

获取同步任务关联数据源的数据库名列表

该接口是一个同步接口，请求成功会返回对应的列表。

请求 URI

```
GET /api/v1/tasks/{task-name}/sources/{source-name}/schemas
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/sources/source-1/schemas' \
  -H 'accept: application/json'
```

```
[
  "db1"
]
```

获取同步任务关联数据源的数据表名列表

该接口是一个同步接口，请求成功会返回对应的列表。

请求 URI

```
GET /api/v1/tasks/{task-name}/sources/{source-name}/schemas/{schema-name}
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/sources/source-1/schemas/db1' \
  -H 'accept: application/json'
```

```
[
  "table1"
]
```

获取同步任务关联数据源的数据表的创建语句

该接口是一个同步接口，请求成功会返回对应的创建语句。

请求 URI

```
GET /api/v1/tasks/{task-name}/sources/{source-name}/schemas/{schema-name}/{table-name}
```

使用样例

```
curl -X 'GET' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/sources/source-1/schemas/db1/table1' \
  -H 'accept: application/json'
```

```
{
  "schema_name": "db1",
  "table_name": "table1",
  "schema_create_sql": "CREATE TABLE `t1` (`id` int(11) NOT NULL AUTO_INCREMENT,PRIMARY KEY (`id
      ↪ `) /*T![clustered_index] CLUSTERED */) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=
      ↪ utf8mb4_bin"
```

```
}
```

更新同步任务关联数据源的数据表的创建语句

该接口是一个同步接口，返回体的 Status Code 是 200。

请求 URI

```
POST /api/v1/tasks/{task-name}/sources/{source-name}/schemas/{schema-name}/{table-name}
```

使用样例

```
curl -X 'PUT' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/sources/task-1/schemas/db1/table1' \
  -H 'accept: */*' \
  -H 'Content-Type: application/json' \
  -d '{
  "sql_content": "CREATE TABLE `t1` ( `c1` int(11) DEFAULT NULL, `c2` int(11) DEFAULT NULL, `c3`
      ↪ int(11) DEFAULT NULL) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_bin;",
  "flush": true,
  "sync": true
}'
```

删除同步任务关联数据源的数据表

该接口是一个同步接口，返回体的 Status Code 是 200。

请求 URI

```
DELETE /api/v1/tasks/{task-name}/sources/{source-name}/schemas/{schema-name}/{table-name}
```

使用样例

```
curl -X 'DELETE' \
  'http://127.0.0.1:8261/api/v1/tasks/task-1/sources/source-1/schemas/db1/table1' \
  -H 'accept: */*'
```

12.8.7.0.5 安全

为 DM 的连接开启加密传输

本文介绍如何为 DM 的连接开启加密传输，包括 DM-master，DM-worker，dmctl 组件之间的连接以及 DM 组件与上下游数据库之间的连接。

为 DM-master，DM-worker，dmctl 组件之间的连接开启加密传输

本节介绍如何为 DM-master，DM-worker，dmctl 组件之间的连接开启加密传输。

配置开启加密传输

1. 准备证书。

   推荐为 DM-master、DM-worker 分别准备一个 Server 证书，并保证可以相互验证，而 dmctl 工具则可选择共用 Client 证书。

有多种工具可以生成自签名证书，如 openssl，cfssl 及 easy-rsa 等基于 openssl 的工具。

这里提供一个使用 openssl 生成证书的示例：生成自签名证书。

2. 配置证书。

> **注意：**
>
> DM-master、 DM-worker 与 dmctl 三个组件可使用同一套证书。

- DM-master

  在 DM-master 配置文件或命令行参数中设置：

```
ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/master-cert.pem"
ssl-key = "/path/to/master-key.pem"
```

- DM-worker

  在 DM-worker 配置文件或命令行参数中设置：

```
ssl-ca = "/path/to/ca.pem"
ssl-cert = "/path/to/worker-cert.pem"
ssl-key = "/path/to/worker-key.pem"
```

- dmctl

  若 DM 集群各个组件间开启加密传输后，在使用 dmctl 工具连接集群时，需要指定 Client 证书，示例如下：

```
./dmctl --master-addr=127.0.0.1:8261 --ssl-ca /path/to/ca.pem --ssl-cert /path/to/
    ↪ client-cert.pem --ssl-key /path/to/client-key.pem
```

认证组件调用者身份

通常被调用者除了校验调用者提供的密钥、证书和 CA 有效性外，还需要校验调用者身份以防止拥有有效证书的非法访问者进行访问（例如：DM-worker 只能被 DM-master 访问，需阻止拥有合法证书但非 DM-master 的其他访问者访问 DM-worker）。

如希望进行组件调用者身份认证，需要在生成证书时通过 Common Name (CN) 标识证书使用者身份，并在被调用者配置检查证书 Common Name 列表时检查调用者身份。

- DM-master

  在 config 文件或命令行参数中设置：

```
cert-allowed-cn = ["dm"]
```

- DM-worker

  在 config 文件或命令行参数中设置：

```
cert-allowed-cn = ["dm"]
```

证书重加载

DM-master、 DM-worker 和 dmctl 都会在每次新建相互通讯的连接时重新读取当前的证书和密钥文件内容，实现证书和密钥的重加载。

当 `ssl-ca`、 `ssl-cert` 或 `ssl-key` 的文件内容更新后，可通过重启 DM 组件使其重新加载证书与密钥内容并重新建立连接。

DM 组件与上下游数据库之间的连接开启加密传输

本节介绍如何为 DM 组件与上下游数据库之间的连接开启加密传输。

为上游数据库连接开启加密传输

1. 配置上游数据库，启用加密连接支持并设置 Server 证书，具体可参考 Using encrypted connections
2. 在 source 配置文件中设置 MySQL Client 证书：

> **注意：**
> 请确保所有 DM-master 与 DM-worker 组件能通过指定路径读取到证书与密钥文件的内容。

```
from:
    security:
        ssl-ca: "/path/to/mysql-ca.pem"
        ssl-cert: "/path/to/mysql-client-cert.pem"
        ssl-key: "/path/to/mysql-client-key.pem"
```

为下游 TiDB 连接开启加密传输

1. 配置下游 TiDB 启用加密连接支持，具体可参考配置 TiDB 启用加密连接支持
2. 在 task 配置文件中设置 TiDB Client 证书：

> **注意：**
> 请确保所有 DM-master 与 DM-worker 组件能通过指定路径读取到证书与密钥文件的内容。

```
target-database:
    security:
        ssl-ca: "/path/to/tidb-ca.pem"
        ssl-cert: "/path/to/tidb-client-cert.pem"
        ssl-key: "/path/to/tidb-client-key.pem"
```

生成自签名证书

本文档提供使用 `openssl` 生成自签名证书的一个示例，用户也可以根据自己的需求生成符合要求的证书和密钥。

假设实例集群拓扑如下：

Name	Host IP	Services
node1	172.16.10.11	DM-master1
node2	172.16.10.12	DM-master2
node3	172.16.10.13	DM-master3
node4	172.16.10.14	DM-worker1
node5	172.16.10.15	DM-worker2
node6	172.16.10.16	DM-worker3

安装 OpenSSL

对于 Debian 或 Ubuntu 操作系统：

```
apt install openssl
```

对于 RedHat 或 CentOS 操作系统：

```
yum install openssl
```

也可以参考 OpenSSL 官方的下载文档进行安装。

生成 CA 证书

CA 的作用是签发证书。实际情况中，请联系你的管理员签发证书或者使用信任的 CA 机构。CA 会管理多个证书对，这里只需生成原始的一对证书，步骤如下：

1. 生成 CA 密钥：

    ```
    openssl genrsa -out ca-key.pem 4096
    ```

2. 生成 CA 证书：

    ```
    openssl req -new -x509 -days 1000 -key ca-key.pem -out ca.pem
    ```

3. 验证 CA 证书：

    ```
    openssl x509 -text -in ca.pem -noout
    ```

签发各个组件的证书

集群中可能使用到的证书

- master certificate 由 DM-master 使用，为其他组件验证 DM-master 身份。

- worker certificate 由 DM-worker 使用，为其他组件验证 DM-worker 身份。
- client certificate 由 dmctl 使用，用于 DM-master、 DM-worker 验证客户端。

为 DM-master 签发证书

给 DM-master 实例签发证书的步骤如下：

1. 生成该证书对应的私钥：

```
openssl genrsa -out master-key.pem 2048
```

2. 拷贝一份 OpenSSL 的配置模板文件。

   模板文件可能存在多个位置，请以实际位置为准：

```
cp /usr/lib/ssl/openssl.cnf .
```

   如果不知道实际位置，请在根目录下查找：

```
find / -name openssl.cnf
```

3. 编辑 openssl.cnf，在 [req] 字段下加入 req_extensions = v3_req，然后在 [v3_req] 字段下加入 subjectAltName = @alt_names。最后新建一个字段，根据前述的集群拓扑编辑 Subject Alternative Name (SAN) 的信息：

```
[ alt_names ]
IP.1 = 127.0.0.1
IP.2 = 172.16.10.11
IP.3 = 172.16.10.12
IP.4 = 172.16.10.13
```

   目前支持以下 SAN 检查项：

   - IP
   - DNS
   - URI

> 注意：
> 如果要使用 0.0.0.0 等特殊 IP 用于连接通讯，也需要将其加入到 alt_names 中。

4. 保存 openssl.cnf 文件后，生成证书请求文件（在这一步中提供 Common Name (e.g. server FQDN or YOUR name) []: 输入时，可以为该证书指定 Common Name (CN)，如 dm。其作用是让服务端验证接入的客户端的身份，各个组件默认不会开启验证，需要在配置文件中启用该功能才生效）：

```
openssl req -new -key master-key.pem -out master-cert.pem -config openssl.cnf
```

5. 签发生成证书：

```
openssl x509 -req -days 365 -CA ca.pem -CAkey ca-key.pem -CAcreateserial -in master-cert.pem
    ↪  -out master-cert.pem -extensions v3_req -extfile openssl.cnf
```

6. 验证证书携带 SAN 字段信息（可选）：

```
openssl x509 -text -in master-cert.pem -noout
```

7. 确认在当前目录下得到如下文件：

```
ca.pem
master-cert.pem
master-key.pem
```

> 注意：
>
> 为 DM-worker 组件签发证书的过程类似，此文档不再赘述。

为 dmctl 签发证书

为客户端签发证书的步骤如下。

1. 生成该证书对应的私钥：

```
openssl genrsa -out client-key.pem 2048
```

2. 生成证书请求文件（在这一步也可以为该证书指定 Common Name，其作用是让服务端验证接入的客户端的身份，默认不会开启对各个组件的验证，需要在配置文件中启用该功能才生效）

```
openssl req -new -key client-key.pem -out client-cert.pem
```

3. 签发生成证书：

```
openssl x509 -req -days 365 -CA ca.pem -CAkey ca-key.pem -CAcreateserial -in client-cert.pem
    ↪  -out client-cert.pem
```

12.8.7.0.6 监控告警

DM 监控指标

使用 TiUP 部署 DM 集群的时候，会默认部署一套监控系统。

Task

在 Grafana dashboard 中，DM 默认名称为 DM-task。

Overview

overview 下包含运行当前选定 task 的所有 DM-worker/master instance/source 的部分监控指标。当前默认告警规则只针对于单个 DM-worker/master instance/source。

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	N/A	N/A
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	N/A	N/A
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
binlog file gap between master and syncer	与上游 master 相比 binlog replication unit 落后的 binlog file 个数	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A

Operate error

metric 名称	说明	告警说明	告警级别
before any operate error	在进行操作之前出错的次数	N/A	N/A
source bound error	数据源绑定操作出错次数	N/A	N/A
start error	子任务启动的出错次数	N/A	N/A
pause error	子任务暂停的出错次数	N/A	N/A
resume error	子任务恢复的出错次数	N/A	N/A
auto-resume error	子任务自动恢复的出错次数	N/A	N/A
update error	子任务更新的出错次数	N/A	N/A
stop error	子任务停止的出错次数	N/A	N/A

HA 高可用

metric 名称	说明	告警说明	告警级别
number of dm-masters start leader components per minute	每分钟内 DM-master 尝试启用 leader 相关组件次数	N/A	N/A
number of workers in different state	不同状态下有多少个 DM-worker	存在离线的 DM-worker 超过一小时	critical
workers' state	DM-worker 的状态	N/A	N/A
number of worker event error	不同类型的 DM-worker 错误出现次数	N/A	N/A
shard ddl error per minute	每分钟内不同类型的 shard DDL 错误次数	发生 shard DDL 错误	critical
number of pending shard ddl	未完成的 shard DDL 数目	存在未完成的 shard DDL 数目超过一小时	critical

Task 状态

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	当子任务状态处于 Paused 超过 20 分钟时	critical

Dump/Load unit

下面 metrics 仅在 `task-mode` 为 `full` 或者 `all` 模式下会有值。

metric 名称	说明	告警说明	告警级别
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
data file size	load unit 导入的全量数据中数据文件（内含 `INSERT INTO` 语句）的总大小	N/A	N/A
dump process exits with error	dump unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
load process exits with error	load unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
table count	load unit 导入的全量数据中 table 的数量总和	N/A	N/A
data file count	load unit 导入的全量数据中数据文件（内含 `INSERT INTO` 语句）的数量总和	N/A	N/A
transaction execution latency	load unit 在执行事务的时延，单位：秒	N/A	N/A
statement execution latency	load unit 执行语句的耗时，单位：秒	N/A	N/A
remaining time	load unit 完成同步的剩余时间，单位：秒	N/A	N/A

Binlog replication

下面 metrics 仅在 `task-mode` 为 `incremental` 或者 `all` 模式下会有值。

metric 名称	说明	告警说明	告警级别
remaining time to sync	预计 Syncer 还需要多少分钟可以和上游 master 完全同步，单位：分钟	N/A	N/A
replicate lag gauge	上游 master 到下游的 binlog 复制延迟时间，单位：秒	N/A	N/A
replicate lag histogram	上游 master 到下游的 binlog 复制延迟分布，单位：秒。注意由于统计机制不同，数据会有误差	N/A	N/A
process exist with error	binlog replication unit 在 DM-worker 内部遇到错误并且退出了	立即告警	critical

metric 名称	说明	告警说明	告警级别
binlog file gap between master and syncer	与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个（不含 1 个）且持续 10 分钟时	critical
binlog file gap between relay and syncer	与 relay 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个（不含 1 个）且持续 10 分钟时	critical
binlog event QPS	单位时间内接收到的 binlog event 数量 (不包含需要跳过的 event)	N/A	N/A
skipped binlog event QPS	单位时间内接收到的需要跳过的 binlog event 数量	N/A	N/A
read binlog event duration	binlog replication unit 从 relay log 或上游 MySQL 读取 binlog 的耗时，单位：秒	N/A	N/A
transform binlog event duration	binlog replication unit 解析 binlog 并将 binlog 转换成 SQL 语句的耗时，单位：秒	N/A	N/A
dispatch binlog event duration	binlog replication unit 调度一条 binlog event 的耗时，单位：秒	N/A	N/A

metric 名称	说明	告警说明	告警级别
transaction execution latency	binlog replication unit 执行事务到下游的耗时，单位：秒	N/A	N/A
binlog event size	binlog replication unit 从 relay log 或上游 MySQL 读取的单条 binlog event 的大小	N/A	N/A
DML queue remain length	剩余 DML job 队列的长度	N/A	N/A
total sqls jobs	单位时间内新增的 job 数量	N/A	N/A
finished sqls jobs	单位时间内完成的 job 数量	N/A	N/A
statement execution latency	binlog replication unit 执行语句到下游的耗时，单位：秒	N/A	N/A
add job duration	binlog replication unit 增加一条 job 到队列的耗时，单位：秒	N/A	N/A
DML conflict detect duration	binlog replication unit 检测 DML 间冲突的耗时，单位：秒	N/A	N/A
skipped event duration	binlog replication unit 跳过 binlog event 的耗时，单位：秒	N/A	N/A
unsynced tables	当前子任务内还未收到 shard DDL 的分表数量	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A

metric 名称	说明	告警说明	告警级别
ideal QPS	在 DM 运行耗时为 0 时可以达到的最高 QPS	N/A	N/A
binlog event row	一个 binlog 事件中的行数	N/A	N/A
finished transaction total	执行完毕的事务数量	N/A	N/A
replication transaction batch	执行到下游的事务里中 sql 行数	N/A	N/A
flush checkpoints time interval	检查点刷新时间间隔，单位：秒	N/A	N/A

Relay log

metric 名称	说明	告警说明	告警级别
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	小于 10G 的时候需要告警	critical
process exits with error	relay log 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
relay log data corruption	relay log 文件损坏的个数	立即告警	emergency
fail to read binlog from master	relay 从上游的 MySQL 读取 binlog 时遇到的错误数	立即告警	critical

metric 名称	说明	告警说明	告警级别
fail to write relay log	relay 写 binlog 到磁盘时遇到的错误数	立即告警	critical
binlog file index	relay log 最大的文件序列号。如 value = 1 表示 relay-log.000001	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个 (不含 1 个) 且持续 10 分钟时	critical
binlog pos	relay log 最新文件的写入 offset	N/A	N/A
read binlog event duration	relay log 从上游的 MySQL 读取 binlog 的时延，单位：秒	N/A	N/A
write relay log duration	relay log 每次写 binlog 到磁盘的时延，单位：秒	N/A	N/A
binlog event size	relay log 写到磁盘的单条 binlog 的大小	N/A	N/A

Instance

在 Grafana dashboard 中，instance 的默认名称为 `DM-instance`。

Relay log

metric 名称	说明	告警说明	告警级别
storage capacity	relay log 占有的磁盘的总容量	N/A	N/A
storage remain	relay log 占有的磁盘的剩余可用容量	小于 10G 的时候需要告警	critical

metric 名称	说明	告警说明	告警级别
process exits with error	relay log 在 DM-worker 内部遇到错误并且退出了	立即告警	critical
relay log data corruption	relay log 文件损坏的个数	立即告警	emergency
fail to read binlog from master	relay 从上游的 MySQL 读取 binlog 时遇到的错误数	立即告警	critical
fail to write relay log	relay 写 binlog 到磁盘时遇到的错误数	立即告警	critical
binlog file index	relay log 最大的文件序列号。如 value = 1 表示 relay-log.000001	N/A	N/A
binlog file gap between master and relay	relay 与上游 master 相比落后的 binlog file 个数	落后 binlog file 个数超过 1 个 (不含 1 个) 且持续 10 分钟时	critical
binlog pos	relay log 最新文件的写入 offset	N/A	N/A
read binlog duration	relay log 从上游的 MySQL 读取 binlog 的时延，单位：秒	N/A	N/A
write relay log duration	relay log 每次写 binlog 到磁盘的时延，单位：秒	N/A	N/A
binlog size	relay log 写到磁盘的单条 binlog 的大小	N/A	N/A

task

metric 名称	说明	告警说明	告警级别
task state	迁移子任务的状态	当子任务状态处于 paused 超过 10 分钟时	critical
load progress	load unit 导入过程的进度百分比，值变化范围为：0% - 100%	N/A	N/A
binlog file gap between master and syncer	与上游 master 相比 binlog replication unit 落后的 binlog file 个数	N/A	N/A
shard lock resolving	当前子任务是否正在等待 shard DDL 迁移，大于 0 表示正在等待迁移	N/A	N/A

DM 告警信息

使用 TiUP 部署 DM 集群的时候，会默认部署一套告警系统。

DM 的告警规则及其对应的处理方法可参考告警处理。

DM 的告警信息与监控指标均基于 Prometheus，告警规则与监控指标的对应关系可参考DM 监控指标。

12.8.7.0.7 故障及处理方法

本文档介绍 DM 的错误系统及常见故障的处理方法。

DM 错误系统

在 DM 的错误系统中，对于一条特定的错误，通常主要包含以下信息：

- code：错误码。

  同一种错误都使用相同的错误码。错误码不随 DM 版本改变。

  在 DM 迭代过程中，部分错误可能会被移除，但错误码不会。新增的错误会使用新的错误码，不会复用已有的错误码。

- class：发生错误的类别。

  用于标记出现错误的系统子模块。

下表展示所有的错误类别、错误对应的系统子模块、错误样例：

|

错误类别

| 错误对应的系统子模块 | 错误样例
↪ |

:———— | :———————— | :————————————————————— |
database | 执行数据库操作出现错误 | [code=10003:class=database:scope=downstream:level=medium]
↪ database driver: invalid connection |
functional | 系统底层的基础函数错误 | [code=11005:class=functional:scope=internal:level=high]
↪ not allowed operation: alter multiple tables in one statement |
config | 配置错误 | [code=20005:class=config:scope=internal:level=medium] empty source-id not
↪ valid |
binlog-op | binlog 操作出现错误 | [code=22001:class=binlog-op:scope=internal:level=high] empty
↪ UUIDs not valid |
checkpoint | checkpoint 相关操作出现错误 | [code=24002:class=checkpoint:scope=internal:level=
↪ high] save point bin.1234 is older than current pos bin.1371 |
task-check | 进行任务检查时出现错误 | [code=26003:class=task-check:scope=internal:level=medium
↪] new table router error |
relay-event-lib| 执行 relay 模块基础功能时出现错误 | [code=28001:class=relay-event-lib:scope=
↪ internal:level=high] parse server-uuid.index |
relay-unit | relay 处理单元内出现错误 | [code=30015:class=relay-unit:scope=upstream:level=high]
↪ TCPReader get event: ERROR 1236 (HY000): Could not open log file |
dump-unit | dump 处理单元内出现错误 | [code=32001:class=dump-unit:scope=internal:level=high]
↪ mydumper runs with error: CRITICAL **: 15:12:17.559: Error connecting to database: Access
↪ denied for user 'root'@'172.17.0.1' (using password: NO) |
load-unit | load 处理单元内出现错误 | [code=34002:class=load-unit:scope=internal:level=high]
↪ corresponding ending of sql: ')' not found |
sync-unit | sync 处理单元内出现错误 | [code=36027:class=sync-unit:scope=internal:level=high]
↪ Column count doesn't match value count: 9 (columns)vs 10 (values) |
dm-master | DM-master 服务内部出现错误 | [code=38008:class=dm-master:scope=internal:level
↪ =high] grpc request error: rpc error: code = Unavailable desc = all SubConns are in
↪ TransientFailure, latest connection error: connection error: desc = "transport: Error
↪ while dialing dial tcp 172.17.0.2:8262: connect: connection refused" |
dm-worker | DM-worker 服务内部出现错误 | [code=40066:class=dm-worker:scope=internal:level=high]
↪ ExecuteDDL timeout, try use query-status to query whether the DDL is still blocking |
dm-tracer | DM-tracer 服务内部出现错误 | [code=42004:class=dm-tracer:scope=internal:level=medium
↪] trace event test.1 not found |
schema-tracker | 增量复制时记录 schema 变更出现错误 | [code=44006:class=schema-tracker:scope=
↪ internal:level=high],"cannot track DDL: ALTER TABLE test DROP COLUMN col1" |
scheduler | 数据迁移任务调度相关操作出错操作 | [code=46001:class=scheduler:scope=internal:
↪ level=high],"the scheduler has not started" |
dmctl | dmctl 内部或与其他组件交互出现错误 | [code=48001:class=dmctl:scope=internal:level=high
↪],"can not create grpc connection" |

- `scope`：错误作用域。

  用于标识错误发生时 DM 作用对象的范围和来源，包括未设置 (`not-set`)、上游数据库 (`upstream`)、下游数据库 (`downstream`)、内部 (`internal`) 四种类型。

  如果错误发生的逻辑直接涉及到上下游数据库请求，作用域会设置为 `upstream` 或 `downstream`，其他出错场景目前都设置为 `internal`。

- `level`：错误级别。

  错误的严重级别，包括低级别 (`low`)、中级别 (`medium`)、高级别 (`high`) 三种。

  低级别通常是用户操作、输入错误，不影响正常迁移任务；中级别通常是用户配置等错误，会影响部分新启动服务，不影响已有系统迁移状态；高级别通常是用户需要关注的一些错误，可能存在迁移任务中断等风险，需要用户进行处理。

- `message`：错误描述。

  错误的详细描述信息。对于错误调用链上每一层额外增加的错误 message，采用 errors.Wrap 的模式来叠加和保存错误 message。wrap 最外层的 message 是 DM 内部对该错误的描述，wrap 最内层的 message 是该错误最底层出错位置的错误描述。

- workaround: 错误处理方法（可选）。

  对该错误的处理方法。对于部分明确的错误（如：配置信息错误等），DM 会在 workaround 中给出相应的人为处理方法。

- 错误堆栈信息（可选）。

  DM 根据错误的严重程度和必要性来选择是否输出错误堆栈。错误堆栈记录了错误发生时完整的堆栈调用信息。如果用户通过错误基本信息和错误 message 描述不能完全诊断出错误发生的原因，可以通过错误堆栈进一步跟进出错时代码的运行路径。

可在 DM 代码仓库 已发布的错误码 中查询完整的错误码列表。

DM 故障诊断

如果在运行 DM 工具时出现了错误，请尝试以下解决方案：

1. 执行 `query-status` 命令查看任务运行状态以及相关错误输出。
2. 查看与该错误相关的日志文件。日志文件位于 DM-master、DM-worker 部署节点上，通过DM 错误系统 获取错误的关键信息，然后查看常见故障处理方法以寻找相应的解决方案。
3. 如果该错误还没有相应的解决方案，并且你无法通过查询日志或监控指标自行解决此问题，你可以联系相关技术支持人员。
4. 一般情况下，错误处理完成后，只需使用 dmctl 重启任务即可。

```
resume-task ${task name}
```

但在某些情况下，你还需要重置数据迁移任务。有关何时需要重置以及如何重置，详见重置数据迁移任务。

常见故障处理方法

|

错误码

错误说明	解决方法
:———	:———————————————	:———————————————
code=10001	数据库操作异常	进一步分析错误信息和错误堆栈	
code=10002	数据库底层的 bad connection 错误，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启等）且当前请求的数据暂时未能发送到 TiDB。	DM 提供针对此类错误的自动恢复。如果长时间未恢复，需要用户检查网络或 TiDB 状态。	
code=10003	数据库底层 invalid connection 错误，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启、TiKV busy 等）且当前请求已有部分数据发送到了 TiDB。	DM 提供针对此类错误的自动恢复。如果未能正常恢复，需要用户进一步检查错误信息并根据具体场景进行分析。	
code=10005	数据库查询类语句出错		
code=10006	数据库 EXECUTE 类型语句出错，包括 DDL 和 INSERT/UPDATE/DELETE 类型的 DML。更详细的错误信息可通过错误 message 获取。错误 message 中通常包含操作数据库所返回的错误码和错误信息。		
code=11006	DM 内置的 parser 解析不兼容的 DDL 时出错	可参考Data Migration 故障诊断 -处理不兼容的 DDL 语句 提供的解决方案	
code=20010	处理任务配置时，解密数据库的密码出错	检查任务配置中提供的下游数据库密码是否有使用 dmctl 正确加密	
code=26002	任务检查创建数据库连接失败。更详细的错误信息可通过错误 message 获取。错误 message 中包含操作数据库所返回的错误码和错误信息。	检查 DM-master 所在的机器是否有权限访问上游	
code=32001	dump 处理单元异常	如果报错 msg 包含 mydumper: argument list too long.，则需要用户根据 block-allow-list，在 task.yaml 的 dump 处理单元的 extra-args 参数中手动加上 --regex 正则表达式设置要导出的库表。例如，如果要导出所有库中表名字为 hello 的表，可加上 --regex '.*\\.hello$'，如果要导出所有表，可加上 --regex '.*'。	
code=38008	DM 组件间的 gRPC 通信出错	检查 class，定位错误发生在哪些组件的交互环节，根据错误 message 判断是哪类通信错误。如果是 gRPC 建立连接出错，可检查通信服务端是否运行正常。	

迁移任务中断并包含 invalid connection 错误

原因

发生 invalid connection 错误时，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启、TiKV busy 等）且当前请求已有部分数据发送到了 TiDB。

解决方案

由于 DM 中存在迁移任务并发向下游复制数据的特性，因此在任务中断时可能同时包含多个错误（可通过 query-status 查询当前错误）。

- 如果错误中仅包含 invalid connection 类型的错误且当前处于增量复制阶段，则 DM 会自动进行重试。
- 如果 DM 由于版本问题等未自动进行重试或自动重试未能成功，则可尝试先使用 stop-task 停止任务，然后再使用 start-task 重启任务。

迁移任务中断并包含 driver: bad connection 错误

原因

发生 driver: bad connection 错误时，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启等）且当前请求的数据暂时未能发送到 TiDB。

解决方案

当前版本 DM 会自动进行重试，如果由于版本问题等未自动重试，可先使用 `stop-task` 停止任务后再使用 `start-task` 重启任务。

relay 处理单元报错 `event from * in * diff from passed-in event *` 或迁移任务中断并包含 `get binlog error ERROR 1236 (HY000)`、`binlog checksum mismatch, data may be corrupted` 等 binlog 获取或解析失败错误

原因

在 DM 进行 relay log 拉取与增量复制过程中，如果遇到了上游超过 4GB 的 binlog 文件，就可能出现这两个错误。

原因是 DM 在写 relay log 时需要依据 binlog position 及文件大小对 event 进行验证，且需要保存迁移的 binlog position 信息作为 checkpoint。但是 MySQL binlog position 官方定义使用 uint32 存储，所以超过 4G 部分的 binlog position 的 offset 值会溢出，进而出现上面的错误。

解决方案

对于 relay 处理单元，可通过以下步骤手动恢复：

1. 在上游确认出错时对应的 binlog 文件的大小超出了 4GB。
2. 停止 DM-worker。
3. 将上游对应的 binlog 文件复制到 relay log 目录作为 relay log 文件。
4. 更新 relay log 目录内对应的 `relay.meta` 文件以从下一个 binlog 开始拉取。如果 DM worker 已开启 `enable_gtid`，那么在修改 `relay.meta` 文件时，同样需要修改下一个 binlog 对应的 GTID。如果未开启 `enable_gtid` 则无需修改 GTID。

   例如：报错时有 `binlog-name = "mysql-bin.004451"` 与 `binlog-pos = 2453`，则将其分别更新为 `binlog-name = "mysql-bin.004452"` 和 `binlog-pos = 4`，同时更新 `binlog-gtid = "f0e914ef-54cf-11e7-813d-6c92bf2fa791:1-138218058"`。
5. 重启 DM-worker。

对于 binlog replication 处理单元，可通过以下步骤手动恢复：

1. 在上游确认出错时对应的 binlog 文件的大小超出了 4GB。
2. 通过 `stop-task` 停止迁移任务。
3. 将下游 `dm_meta` 数据库中 global checkpoint 与每个 table 的 checkpoint 中的 `binlog_name` 更新为出错的 binlog 文件，将 `binlog_pos` 更新为已迁移过的一个合法的 position 值，比如 4。

   例如：出错任务名为 `dm_test`，对应的 `source-id` 为 `replica-1`，出错时对应的 binlog 文件为 `mysql-bin|000001.004451`，则执行 `UPDATE dm_test_syncer_checkpoint SET binlog_name='mysql-bin|000001.004451', binlog_pos = 4 WHERE id='replica-1';`。
4. 在迁移任务配置中为 `syncers` 部分设置 `safe-mode: true` 以保证可重入执行。
5. 通过 `start-task` 启动迁移任务。

6. 通过 query-status 观察迁移任务状态，当原造成出错的 relay log 文件迁移完成后，即可还原 safe-mode 为原始值并重启迁移任务。

执行 query-status 或查看日志时出现 Access denied for user 'root'@'172.31.43.27' (using password: YES)

在所有 DM 配置文件中，数据库相关的密码都推荐使用经 dmctl 加密后的密文（若数据库密码为空，则无需加密）。有关如何使用 dmctl 加密明文密码，参见使用 dmctl 加密数据库密码。

此外，在 DM 运行过程中，上下游数据库的用户必须具备相应的读写权限。在启动迁移任务过程中，DM 会自动进行相应权限的前置检查，详见上游 MySQL 实例配置前置检查。

load 处理单元报错 packet for query is too large. Try adjusting the 'max_allowed_packet' variable

原因

- MySQL client 和 MySQL/TiDB Server 都有 max_allowed_packet 配额的限制，如果在使用过程中违反其中任何一个 max_allowed_packet 配额，客户端程序就会收到对应的报错。目前最新版本的 DM 和 TiDB Server 的默认 max_allowed_packet 配额都为 64M。
- DM 的全量数据导入处理模块不支持对 dump 处理模块导出的 SQL 文件进行切分。

解决方案

- 推荐在 DM 的 dump 处理单元提供的配置 extra-args 中设置 statement-size:

  依据默认的 --statement-size 设置，DM 的 dump 处理单元默认生成的 Insert Statement 大小一般会在 1M 左右，使用默认值就可以确保绝大部分情况 load 处理单元不会报错 packet for query is too large ↪ . Try adjusting the 'max_allowed_packet' variable。

  有时候在 dump 过程中会出现下面的 WARN log。这个 WARN log 不影响 dump 的过程，只是说明 dump 的表可能是宽表。

```
Row bigger than statement_size for xxx
```

- 如果宽表的单行超过了 64M，需要修改以下两项配置，并且确保其生效。
  - 在 TiDB Server 执行 set @@global.max_allowed_packet=134217728（134217728 = 128M）
  - 根据实际情况为 DM 的任务配置文件中的 target-database 增加配置 max-allowed-packet: ↪ 134217728 (128M)，执行 stop-task 后再重新 start-task。

12.8.7.0.8 TiDB Data Migration 术语表

本文档介绍 TiDB Data Migration (TiDB DM) 相关术语。

B

Binlog

在 TiDB DM 中，Binlog 通常指 MySQL/MariaDB 生成的 binary log 文件，具体请参考 MySQL Binary Log 与 MariaDB Binary Log。

Binlog event

MySQL/MariaDB 生成的 Binlog 文件中的数据变更信息，具体请参考 MySQL Binlog Event 与 MariaDB Binlog Event。

Binlog event filter

比 Block & allow table list 更加细粒度的过滤功能，具体可参考Binlog Event Filter。

Binlog position

特定 Binlog event 在 Binlog 文件中的位置偏移信息，具体请参考 MySQL `SHOW BINLOG EVENTS` 与 MariaDB `SHOW BINLOG EVENTS`。

Binlog replication 处理单元/ sync 处理单元

DM-worker 内部用于读取上游 Binlog 或本地 Relay log 并迁移到下游的处理单元，每个 Subtask 对应一个 Binlog replication 处理单元。在当前文档中，有时也称作 Sync 处理单元。

Block & allow table list

针对上游数据库实例表的黑白名单过滤功能，具体可参考Block & Allow Table Lists。该功能与 MySQL Replication Filtering 及 MariaDB Replication Filters 类似。

C

Checkpoint

TiDB DM 在全量导入与增量复制过程中的断点信息，用于在重新启动或恢复任务时从之前已经处理过的位置继续执行。

- 对于全量导入，Checkpoint 信息对应于每个数据文件已经被成功导入的数据对应的文件内偏移量等信息，其在每个导入数据的事务中迁移更新；
- 对于增量复制，Checkpoint 信息对应于已经成功解析并导入到下游的Binlog event 对应的Binlog position 等信息，其在 DDL 导入成功后或距上次更新时间超过 30 秒等条件下更新。

另外，Relay 处理单元 对应的 `relay.meta` 内记录的信息也相当于 Checkpoint，其对应于 Relay 处理单元已经成功从上游拉取并写入到Relay log 的Binlog event 对应的Binlog position 或GTID 信息。

D

Dump 处理单元

DM-worker 内部用于从上游导出全量数据的处理单元，每个 Subtask 对应一个 Dump 处理单元。

F

复制/增量复制

使用 TiDB Data Migration 工具将上游数据库的增量数据复制到下游数据库的过程。

本用户手册中，在明确提到是 “增量” 的情况下，将使用 “复制” 或 “增量复制” 进行文档描述。

G

GTID

MySQL/MariaDB 的全局事务 ID，当启用该功能后会在 Binlog 文件中记录 GTID 相关信息，多个 GTID 即组成为 GTID Set，具体请参考 MySQL GTID Format and Storage 与 MariaDB Global Transaction ID。

L

Load 处理单元

DM-worker 内部用于将全量导出数据导入到下游的处理单元，每个 Subtask 对应一个 Load 处理单元。在当前文档中，有时也称作 Import 处理单元。

Q

迁移/全量迁移

使用 TiDB Data Migration 工具将上游数据库的全量数据迁移到下游数据库的过程。

本用户手册中，在明确提到是“全量”的情况下，将使用“迁移”或“全量迁移”进行文档描述；在明确提到是“全量 + 增量”的情况下，也将统一使用“迁移”进行文档描述。

R

Relay log

DM-worker 从上游 MySQL/MariaDB 拉取 Binlog 后存储在本地的文件，当前其格式为标准的 Binlog 格式，可使用版本兼容的 mysqlbinlog 等工具进行解析。其作用与 MySQL Relay Log 及 MariaDB Relay Log 相近。

有关 TiDB DM 内 Relay log 的目录结构、初始迁移规则、数据清理等内容，可参考 TiDB DM Relay Log。

Relay 处理单元

DM-worker 内部用于从上游拉取 Binlog 并写入数据到 Relay log 的处理单元，每个 DM-worker 实例内部仅存在一个该处理单元。

S

Safe mode

指增量复制过程中，用于支持在表结构中存在主键或唯一索引的条件下可重复导入 DML 的模式。该模式的主要特点是：将来自上游的 INSERT 改写为 REPLACE，将 UPDATE 改写为 DELETE 与 REPLACE 后再向下游执行。

该模式会在满足如下任一条件时启用：

- 任务配置文件中设置 `safe-mode: true` 时会始终启用
- 合库合表模式下，DDL 尚未在所有分表完成同步时保持启用
- 在全量迁移任务中的 dump 处理单元配置 `--consistency none` 后，不能确定导出开始时的 binlog 变动是否影响了导出数据。Safe mode 会在增量复制这部分 binlog 时保持启用
- 任务出错停止并恢复后，对有些数据的操作可能会被执行两次时保持启用

Shard DDL

指合库合表迁移过程中，在上游各分表 (shard) 上执行的需要 TiDB DM 进行协调迁移的 DDL。在当前文档中，有时也称作 Sharding DDL。

Shard DDL lock

用于协调 Shard DDL 迁移的锁机制，具体原理可查看悲观模式下分库分表合并迁移实现原理。在当前文档中，有时也称作 Sharding DDL lock。

Shard group

指合库合表迁移过程中，需要合并迁移到下游同一张表的所有上游分表 (shard)，TiDB DM 内部具体实现时使用了两级抽象的 Shard group，具体可查看悲观模式下分库分表合并迁移实现原理。在当前文档中，有时也称作 Sharding group。

Subtask

数据迁移子任务，即数据迁移任务运行在单个 DM-worker 实例上的部分。根据任务配置的不同，单个数据迁移任务可能只有一个子任务，也可能有多个子任务。

Subtask status

数据迁移子任务所处的状态，目前包括 New、Running、Paused、Stopped 及 Finished 5 种状态。有关数据迁移任务、子任务状态的更多信息可参考任务状态。

T

Table routing

用于支持将上游 MySQL/MariaDB 实例的某些表迁移到下游指定表的路由功能，可以用于分库分表的合并迁移，具体可参考 Table routing。

Task

数据迁移任务，执行 `start-task` 命令成功后即启动一个数据迁移任务。根据任务配置的不同，单个数据迁移任务既可能只在单个 DM-worker 实例上运行，也可能同时在多个 DM-worker 实例上运行。

Task status

数据迁移子任务所处的状态，由Subtask status 整合而来，具体信息可查看任务状态。

12.8.7.1 使用示例

12.8.7.1.1 使用 DM 迁移数据

本文介绍如何使用 DM 工具迁移数据。

第 1 步：部署 DM 集群

推荐使用 TiUP 部署 DM 集群；也可以使用 binary 部署 DM 集群用于体验或测试。

> 注意：
>
> - 在 DM 所有的配置文件中，对于数据库密码推荐使用 dmctl 加密后的密文。如果数据库密码为空，则不需要加密。关于如何使用 dmctl 加密明文密码，参考使用 dmctl 加密数据库密码。
> - 上下游数据库用户必须拥有相应的读写权限。

第 2 步：检查集群信息

使用 TiUP 部署 DM 集群后，相关配置信息如下：

- DM 集群相关组件配置信息

组件	主机	端口
dm_worker1	172.16.10.72	8262
dm_worker2	172.16.10.73	8262
dm_master	172.16.10.71	8261

- 上下游数据库实例相关信息

数据库实例	主机	端口	用户名	加密密码
上游 MySQL-1	172.16.10.81	3306	root	VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU=
上游 MySQL-2	172.16.10.82	3306	root	VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU=
下游 TiDB	172.16.10.83	4000	root	

上游 MySQL 数据库实例用户所需权限参见上游 MySQL 实例配置前置检查介绍。

第 3 步：创建数据源

1. 将 MySQL-1 的相关信息写入到 `conf/source1.yaml` 中：

```
# MySQL1 Configuration.

source-id: "mysql-replica-01"

# DM-worker 是否使用全局事务标识符（GTID）拉取 binlog。使用前提是在上游 MySQL 已开启 GTID
    ↪ 模式。
enable-gtid: false

from:
  host: "172.16.10.81"
  user: "root"
  password: "VjX8cEeTX+qcvZ3bPaO4h0C80pe/1aU="
  port: 3306
```

2. 在终端中执行下面的命令，使用 `tiup dmctl` 将 MySQL-1 的数据源配置加载到 DM 集群中：

```
tiup dmctl --master-addr 172.16.10.71:8261 operate-source create conf/source1.yaml
```

3. 对于 MySQL-2，修改配置文件中的相关信息，并执行相同的 dmctl 命令。

第 4 步：配置任务

假设需要将 MySQL-1 和 MySQL-2 实例的 `test_db` 库的 `test_table` 表以全量 + 增量的模式迁移到下游 TiDB 的 `test_db` 库的 `test_table` 表。

编辑任务配置文件 `task.yaml`：

```
##### 任务名，多个同时运行的任务不能重名。
name: "test"
##### 全量+增量 (all) 迁移模式。
task-mode: "all"
##### 下游 TiDB 配置信息。
target-database:
  host: "172.16.10.83"
  port: 4000
  user: "root"
  password: ""

##### 当前数据迁移任务需要的全部上游 MySQL 实例配置。
mysql-instances:
-
  # 上游实例或者复制组 ID，参考 `inventory.ini` 的 `source_id` 或者 `dm-master.toml` 的 `source-
      ↪ id 配置`。
  source-id: "mysql-replica-01"
  # 需要迁移的库名或表名的黑白名单的配置项名称，用于引用全局的黑白名单配置，全局配置见下面的 `
      ↪ block-allow-list` 的配置。
  block-allow-list: "global"          # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  # dump 处理单元的配置项名称，用于引用全局的 dump 处理单元配置。
  mydumper-config-name: "global"

-
  source-id: "mysql-replica-02"
  block-allow-list: "global"          # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  mydumper-config-name: "global"

##### 黑白名单全局配置，各实例通过配置项名引用。
block-allow-list:                     # 如果 DM 版本早于 v2.0.0-beta.2 则使用 black-white-list。
  global:
    do-tables:                        # 需要迁移的上游表的白名单。
    - db-name: "test_db"              # 需要迁移的表的库名。
      tbl-name: "test_table"          # 需要迁移的表的名称。

##### dump 处理单元全局配置，各实例通过配置项名引用。
mydumpers:
  global:
    extra-args: ""
```

第 5 步：启动任务

为了提前发现数据迁移任务的一些配置错误，DM 中增加了前置检查功能：

- 启动数据迁移任务时，DM 自动检查相应的权限和配置。

- 也可使用 `check-task` 命令手动前置检查上游的 MySQL 实例配置是否符合 DM 的配置要求。

> **注意：**
>
> 第一次启动数据迁移任务时，必须确保上游数据库已配置。否则，启动任务时会报错。

使用 `tiup dmctl` 执行以下命令启动数据迁移任务。其中，`task.yaml` 是之前编辑的配置文件。

```
tiup dmctl --master-addr 172.16.10.71:8261 start-task ./task.yaml
```

- 如果执行该命令后返回的结果如下，则表明任务已成功启动。

```
{
    "result": true,
    "msg": "",
    "workers": [
        {
            "result": true,
            "worker": "172.16.10.72:8262",
            "msg": ""
        },
        {
            "result": true,
            "worker": "172.16.10.73:8262",
            "msg": ""
        }
    ]
}
```

- 如果任务启动失败，可根据返回结果的提示进行配置变更后执行 `start-task task.yaml` 命令重新启动任务。

第 6 步：查询任务

如需了解 DM 集群中是否存在正在运行的迁移任务及任务状态等信息，可使用 `tiup dmctl` 执行以下命令进行查询：

```
tiup dmctl --master-addr 172.16.10.71:8261 query-status
```

第 7 步：停止任务

如果不再需要进行数据迁移，可以使用 `tiup dmctl` 执行以下命令停止迁移任务：

```
tiup dmctl --master-addr 172.16.10.71:8261 stop-task test
```

其中的 `test` 是 `task.yaml` 配置文件中 name 配置项设置的任务名。

第 8 步：监控任务与查看日志

如果使用 TiUP 部署 DM 集群时，正确部署了 Prometheus、Alertmanager 与 Grafana，且其地址均为 `172.16.10.71`。可在浏览器中打开 http://172.16.10.71:9093 进入 Alertmanager 查看 DM 告警信息；可在浏览器中打开 http://172.16.10.71:3000 进入 Grafana，选择 DM 的 dashboard 查看 DM 相关监控项。

DM 在运行过程中，DM-worker, DM-master 及 dmctl 都会通过日志输出相关信息。各组件的日志目录如下：

- DM-master 日志目录：通过 DM-master 进程参数 `--log-file` 设置。如果使用 TiUP 部署 DM，则日志目录位于 `{log_dir}`。
- DM-worker 日志目录：通过 DM-worker 进程参数 `--log-file` 设置。如果使用 TiUP 部署 DM，则日志目录位于 `{log_dir}`。

12.8.7.2 异常解决

12.8.7.2.1 Data Migration 常见问题

DM 是否支持迁移阿里 RDS 以及其他云数据库的数据？

DM 仅支持解析标准版本的 MySQL/MariaDB 的 binlog，对于阿里云 RDS 以及其他云数据库没有进行过测试，如果确认其 binlog 为标准格式，则可以支持。

已知问题的兼容情况：

- 阿里云 RDS
  - 即使上游表没有主键，阿里云 RDS 的 binlog 中也会包含隐藏的主键列，与上游表结构不一致。
- 华为云 RDS
  - 不支持，详见：华为云数据库 RDS 是否支持直接读取 Binlog 备份文件。

task 配置中的黑白名单的正则表达式是否支持非获取匹配 (?!)?

目前不支持，DM 仅支持 golang 标准库的正则，可以通过 re2-syntax 了解 golang 支持的正则表达式。

如果在上游执行的一个 statement 包含多个 DDL 操作，DM 是否支持迁移？

DM 会尝试将包含多个 DDL 变更操作的单条语句拆分成只包含一个 DDL 操作的多条语句，但是可能没有覆盖所有的场景。建议在上游执行的一条 statement 中只包含一个 DDL 操作，或者在测试环境中验证一下，如果不支持，可以给 DM 提 issue。

如何处理不兼容的 DDL 语句？

你需要使用 dmctl 手动处理 TiDB 不兼容的 DDL 语句（包括手动跳过该 DDL 语句或使用用户指定的 DDL 语句替换原 DDL 语句，详见处理出错的 DDL 语句 ）。

> 注意：
>
> TiDB 目前并不兼容 MySQL 支持的所有 DDL 语句。

如何重置数据迁移任务？

当数据迁移过程中发生异常且无法恢复时，需要重置数据迁移任务，对数据重新进行迁移：

1. 使用 `stop-task` 停止异常的数据迁移任务。
2. 清理下游已迁移的数据。
3. 从下面两种方式中选择其中一种重启数据迁移任务：
    - 修改任务配置文件以指定新的任务名，然后使用 `start-task {task-config-file}` 重启迁移任务。
    - 使用 `start-task --remove-meta {task-config-file}` 重启数据迁移任务。

设置了 `online-ddl-scheme: "gh-ost"`，gh-ost 表相关的 DDL 报错该如何处理？

```
[unit=Sync] ["error information"="{\"msg\":\"[code=36046:class=sync-unit:scope=internal:level=
    ↪ high] online ddls on ghost table `xxx`.`_xxxx_gho`\\ngithub.com/pingcap/dm/pkg/terror.(*
    ↪ Error).Generate ......
```

出现上述错误可能有以下原因：

DM 在最后 `rename ghost_table to origin table` 的步骤会把内存的 DDL 信息读出，并且还原为 origin table 的 DDL。而内存中的 DDL 信息是在 `alter ghost_table` 的时候进行处理，记录 ghost_table DDL 的信息；或者是在重启 dm-worker 启动 task 的时候，从 `dm_meta.{task_name}_onlineddl` 中读取出来。

因此，如果在增量复制过程中，指定的 Pos 跳过了 `alter ghost_table` 的 DDL，但是该 Pos 仍在 gh-ost 的 online-ddl 的过程中，就会因为 ghost_table 没有正确写入到内存以及 `dm_meta.{task_name}_onlineddl`，而导致该问题。

可以通过以下方式绕过这个问题：

1. 取消 task 的 `online-ddl-schema` 的配置。
2. 把 `_{table_name}_gho`、`_{table_name}_ghc`、`_{table_name}_del` 配置到 `block-allow-list.ignore-tables` 中。
3. 手工在下游的 TiDB 执行上游的 DDL。
4. 待 Pos 复制到 gh-ost 整体流程后的位置，再重新启用 `online-ddl-schema` 以及注释掉 `block-allow-list.ignore-tables`。

如何为已有迁移任务增加需要迁移的表？

假如已有数据迁移任务正在运行，但又有其他的表需要添加到该迁移任务中，可根据当前数据迁移任务所处的阶段按下列方式分别进行处理。

> 注意：
>
> 向已有数据迁移任务中增加需要迁移的表操作较复杂，请仅在确有强烈需求时进行。

迁移任务当前处于 Dump 阶段

由于 MySQL 不支持指定 snapshot 来进行导出，因此在导出过程中不支持更新迁移任务并重启以通过断点继续导出，故无法支持在该阶段动态增加需要迁移的表。

如果确实需要增加其他的表用于迁移，建议直接使用新的配置文件重新启动迁移任务。

迁移任务当前处于 Load 阶段

多个不同的数据迁移任务在导出时，通常对应于不同的 binlog position，如将它们在 Load 阶段合并导入，则无法就 binlog position 达成一致，因此不建议在 Load 阶段向数据迁移任务中增加需要迁移的表。

迁移任务当前处于 Sync 阶段

当数据迁移任务已经处于 Sync 阶段时，在配置文件中增加额外的表并重启任务，DM 并不会为新增的表重新执行全量导出与导入，而是会继续从之前的断点进行增量复制。

因此，如果需要新增的表对应的全量数据尚未导入到下游，则需要先使用单独的数据迁移任务将其全量数据导出并导入到下游。

将已有迁移任务对应的全局 checkpoint （`is_global=1`）中的 position 信息记为 `checkpoint-T`，如 `(mysql-bin.000100, 1234)`。将需要增加到迁移任务的表在全量导出的 `metedata`（或另一个处于 Sync 阶段的数据迁移任务的 checkpoint）的 position 信息记为 `checkpoint-S`，如 `(mysql-bin.000099, 5678)`。则可通过以下步骤将表增加到迁移任务中：

1. 使用 `stop-task` 停止已有迁移任务。如果需要增加的表属于另一个运行中的迁移任务，则也将其停止。
2. 使用 MySQL 客户连接到下游 TiDB 数据库，手动更新已有迁移任务对应的 checkpoint 表中的信息为 `checkpoint-T` 与 `checkpoint-S` 中的较小值（在本例中，为 `(mysql-bin.000099, 5678)`）。
    - 需要更新的 checkpoint 表为 `{dm_meta}` 库中的 `{task-name}_syncer_checkpoint`。
    - 需要更新的 checkpoint 行为 `id={source-id}` 且 `is_global=1`。
    - 需要更新的 checkpoint 列为 `binlog_name` 与 `binlog_pos`。
3. 在迁移任务配置中为 `syncers` 部分设置 `safe-mode: true` 以保证可重入执行。
4. 通过 `start-task` 启动迁移任务。
5. 通过 `query-status` 观察迁移任务状态，当 `syncerBinlog` 超过 `checkpoint-T` 与 `checkpoint-S` 中的较大值后（在本例中，为 `(mysql-bin.000100, 1234)`），即可还原 `safe-mode` 为原始值并重启迁移任务。

全量导入过程中遇到报错 `packet for query is too large. Try adjusting the 'max_allowed_packet' variable`

尝试将

- TiDB Server 的全局变量 `max_allowed_packet`
- 任务配置文件中的配置项 `target-database.max-allowed-packet`（详情参见DM 任务完整配置文件介绍）

设置为比默认 67108864 (64M) 更大的值。

2.0+ 集群运行 1.0 已有数据迁移任务时报错 `Error 1054: Unknown column 'binlog_gtid' in 'field list'`

在 DM 2.0 之后，为 checkpoint 等元信息表引入了更多的字段。如果通过 `start-task` 直接使用 1.0 集群的任务配置文件从增量复制阶段继续运行，则会出现 `Error 1054: Unknown column 'binlog_gtid' in 'field list'` 错误。

对于此错误，可手动将 DM 1.0 的数据迁移任务导入到 2.0+ 集群。

TiUP 无法部署 DM 的某个版本（如 v2.0.0-hotfix）

你可以通过 `tiup list dm-master` 命令查看 TiUP 支持部署的 DM 版本。该命令未展示的版本不能由 TiUP 管理。

DM 同步报错 `parse mydumper metadata error: EOF`

该错误需要查看报错信息以及日志进一步分析。报错原因可能是 dump 单元由于缺少权限没有产生正确的 metadata 文件。

DM 分库分表同步中没有明显报错，但是下游数据丢失

需要检查配置项 `block-allow-list` 和 `table-route`：

- `block-allow-list` 填写的是上游数据库表，可以在 `do-tables` 前通过加 “~” 来进行正则匹配。
- `table-route` 不支持正则，采用的是通配符模式，所以 `table_parttern_[0-63]` 只会匹配 table_parttern_0 到 table_pattern_6 这 7 张表。

DM 上游无写入，replicate lag 监控无数据

在 DM v1.0 中，需要开启 `enable-heartbeat` 才会产生该监控数据。v2.0 及以后版本中，尚未启用该功能，replicate lag 监控无数据是预期行为。

DM v2.0.0 启动任务时出现 `fail to initial unit Sync of subtask`，报错信息的 RawCause 显示 `context deadline exceeded`

该问题是 DM v2.0.0 的已知问题，在同步任务的表数目较多时触发，将在 v2.0.1 修复。使用 TiUP 部署的用户可以升级到开发版 nightly 解决该问题，或者访问 GitHub 上 DM 仓库的 release 页面下载 v2.0.0-hotfix 版本手动替换可执行文件。

DM 同步中报错 `duplicate entry`

用户需要首先确认任务中没有配置 `disable-detect`（v2.0.7 及之前版本），没有其他同步程序或手动插入数据，表中没有配置相关的 DML 过滤器。

为了便于排查问题，用户收集到下游 TiDB 相关 general log 后可以在 AskTUG 社区 联系专家进行排查。收集 general log 的方式如下：

```
##### 开启 general log
curl -X POST -d "tidb_general_log=1" http://{TiDBIP}:10080/settings
##### 关闭 general log
curl -X POST -d "tidb_general_log=0" http://{TiDBIP}:10080/settings
```

在发生 `duplicate entry` 报错时，确认日志中包含冲突数据的记录。

监控中部分面板显示 `No data point`

请参照DM 监控指标查看各面板含义，部分面板没有数据是正常现象。例如没有发生错误、不存在 DDL lock、没有启用 relay 功能等情况，均可能使得对应面板没有数据。

DM v1.0 在任务出错时使用 `sql-skip` 命令无法跳过某些语句

首先需要检查执行 `sql-skip` 之后 binlog 位置是否在推进，如果是的话表示 `sql-skip` 已经生效。重复出错的原因是上游发送了多个不支持的 DDL，可以通过 `sql-skip -s <sql-pattern>` 进行模式匹配。

对于类似下面这种报错（报错中包含 `parse statement`）：

```
if the DDL is not needed, you can use a filter rule with \"*\" schema-pattern to ignore it.\n\t :
    ↪  parse statement: line 1 column 11 near \"EVENT `event_del_big_table` \r\nDISABLE\" %!(
    ↪ MISSING)(EXTRA string=ALTER EVENT `event_del_big_table` \r\nDISABLE
```

出现报错的原因是 TiDB parser 无法解析上游的 DDL，例如 `ALTER EVENT`，所以 `sql-skip` 不会按预期生效。可以在任务配置文件中添加 Binlog 过滤规则进行过滤，并设置 `schema-pattern: "*"`。从 DM 2.0.1 版本开始，已预设过滤了 EVENT 相关语句。

在 DM v2.0 版本之后 `sql-skip` 已经被 `handle-error` 替代，`handle-error` 可以跳过该类错误。

DM 同步时下游长时间出现 REPLACE 语句

请检查是否符合safe mode 触发条件。如果任务发生错误并自动恢复，或者发生高可用调度，会满足“启动或恢复任务的前 1 分钟”这一条件，因此启用 safe mode。

可以检查 DM-worker 日志，在其中搜索包含 `change count` 的行，该行的 `new count` 非零时会启用 safe mode。检查 safe mode 启用时间以及启用前是否有报错，以定位启用原因。

使用 DM v2.0 同步数据时重启 DM 进程，出现全量数据导入失败错误

在 DM v2.0.1 及更早版本中，如果全量导入操作未完成时发生重启，重启后的上游数据源与 DM worker 的绑定关系可能会发生变化。例如，可能出现 dump 单元的中间数据在 DM worker A 机器上，但却由 DM worker B 进行 load 单元的情况，进而导致操作失败。

该情况有两种解决方案：

- 如果数据量较小（TB 级以下）或任务有合库合表：清空下游数据库的已导入数据，同时清空导出数据目录，使用 dmctl 删除并 `start-task --remove-meta` 重建任务。后续尽量保证全量导出导入阶段 DM 没有冗余 worker 以及避免在该时段内重启或升级 DM 集群。
- 如果数据量较大（数 TB 或更多）：清空下游数据库的已导入数据，将 lightning 部署到数据所在的 DM worker 节点，使用lightning local backend 模式 导入 DM dump 单元导出的数据。全量导入完成后，修改任务的 `task-mode` 为 `incremental`，修改 `mysql-instance.meta.pos` 为 dump 单元导出数据 `metadata` 中记录的位置，启动一个增量任务。

使用 DM 同步数据时重启 DM 进程，增量任务出现 `ERROR 1236 (HY000): The slave is connecting using CHANGE MASTER TO MASTER_AUTO_POSITION = 1, but the master has purged binary logs containing GTIDs that the slave requires.` 错误

该错误表明全量迁移期间，dump 单元记录 metadata 中的 binlog 位置已经被上游清理。

解决方案：出现该问题时只能清空下游数据库已同步数据，并在停止任务后加上 `--remove-meta` 参数重建任务。

如要提前避免该问题，需要进行以下配置：

1. 在 DM 全量迁移未完成时调大上游 MySQL 的 `expire_logs_days` 变量，保证全量进行结束时 metadata 中的 binlog 位置到当前时间的 binlog 都还没有被清理掉。如果数据量较大，应该使用 dumpling + lightning 的方式加快全量迁移速度。
2. DM 任务开启 relay log 选项，保证 binlog 被清理后 DM 仍有 relay log 可读取。

使用 TiUP v1.3.0, v1.3.1 部署 DM 集群，DM 集群的 grafana 监控报错显示 `failed to fetch dashboard`

该问题为 TiUP 已知 bug，在 TiUP v1.3.2 中已进行修复。可采取以下任一方法解决：

- 方法一：使用 `tiup update --self && tiup update dm` 升级 TiUP 到更新版本，随后先缩容再扩容集群中的 grafana 节点，重建 grafana 服务。
- 方法二：

    1. 备份 `deploy/grafana-$port/bin/public` 文件夹。
    2. 下载 TiUP DM 离线镜像包，并进行解压，将其中的 grafana-v4.0.3-**.tar.gz 文件解压后，用解压出的 public/ 文件夹替换前面所描述的文件夹，运行 `tiup dm restart $cluster_name -R grafana` 重启 grafana 服务监控。

在 DM v2.0 中，同时开启 relay 与 gtid 同步 MySQL 时，运行 `query-status` 发现 syncer checkpoint 中 GTID 不连续

该问题为 DM 已知 bug，在完全满足以下两个条件时将会触发，DM 将在 v2.0.2 修复该问题：

1. DM 配置的 source 同时设置了 `enable-relay` 与 `enable-gtid` 为 `true`
2. DM 同步上游为 MySQL 从库，并且该从库通过 `show binlog events in '<newest-binlog>' limit 2` 查询出的 `previous_gtids` 区间不连续，例如：

```
mysql> show binlog events in 'mysql-bin.000005' limit 2;
+------------------+------+---------------+-----------+-------------+-----------------------------------------
    ↪
| Log_name         | Pos  | Event_type    | Server_id | End_log_pos | Info
    ↪                                                              |
+------------------+------+---------------+-----------+-------------+-----------------------------------------
    ↪
| mysql-bin.000005 |    4 | Format_desc   |    123452 |         123 | Server ver: 5.7.32-35-log,
    ↪  Binlog ver: 4                          |
| mysql-bin.000005 |  123 | Previous_gtids |    123452 |         194 | d3618e68-6052-11eb-a68b
    ↪ -0242ac110002:6-7                          |
+------------------+------+---------------+-----------+-------------+-----------------------------------------
    ↪
```

使用 dmctl 的 `query-status <task>` 指令查询任务信息，如果已经出现 `subTaskStatus.sync.syncerBinlogGtid` 不连续但 `subTaskStatus.sync.masterBinlogGtid` 连续的情况，例如下述例子：

```
query-status test
{
    ...
    "sources": [
```

```
{
    ...
    "sourceStatus": {
        "source": "mysql1",
        ...
        "relayStatus": {
            "masterBinlog": "(mysql-bin.000006, 744)",
            "masterBinlogGtid": "f8004e25-6067-11eb-9fa3-0242ac110003:1-50",
            ...
        }
    },
    "subTaskStatus": [
        {
            ...
            "sync": {
                ...
                "masterBinlog": "(mysql-bin.000006, 744)",
                "masterBinlogGtid": "f8004e25-6067-11eb-9fa3-0242ac110003:1-50",
                "syncerBinlog": "(mysql-bin|000001.000006, 738)",
                "syncerBinlogGtid": "f8004e25-6067-11eb-9fa3-0242ac110003:1-20:40-49",
                ...
                "synced": false,
                "binlogType": "local"
            }
        }
    ]
},
{
    ...
    "sourceStatus": {
        "source": "mysql2",
        ...
        "relayStatus": {
            "masterBinlog": "(mysql-bin.000007, 1979)",
            "masterBinlogGtid": "ddb8974e-6064-11eb-8357-0242ac110002:1-25",
            ...
        }
    },
    "subTaskStatus": [
        {
            ...
            "sync": {
                "masterBinlog": "(mysql-bin.000007, 1979)",
                "masterBinlogGtid": "ddb8974e-6064-11eb-8357-0242ac110002:1-25",
                "syncerBinlog": "(mysql-bin|000001.000008, 1979)",
```

```
                    "syncerBinlogGtid": "ddb8974e-6064-11eb-8357-0242ac110002:1-25",
                    ...
                    "synced": true,
                    "binlogType": "local"
                }
            }
        ]
    }
  ]
}
```

其中 mysql1 的 syncerBinlogGtid 不连续，已有数据丢失需要按下述方案之一处理：

- 如果全量导出任务 metadata 中的 position 到当前时间的上游数据库的 binlog 仍未被清理：
    1. 停止当前任务并删除所有 GTID 不连续的 source
    2. 设置所有 source 的 enable-relay 为 false
    3. 针对 GTID 不连续的 source（上例 mysql1），在对应的任务配置文件 task.yaml 中，把 task-mode 修改为 incremental 并配置增量任务起始点 mysql-instances.meta 为各个全量导出任务 metadata 的 binlog name，position 和 gtid 信息
    4. 配置 task.yaml 中的 syncers.safe-mode 为 true 并重启任务
    5. 待增量同步追上后，停止任务并在任务配置文件中设置 safe-mode 为 false
    6. 再次重启任务
- 如果上游数据库 binlog 已被清理但是本地 relay log 仍未被清理：
    1. 停止当前任务
    2. 针对 GTID 不连续的 source（上例 mysql1），在对应的任务配置文件 task.yaml 中，把 task-mode 修改为 incremental 并配置增量任务起始点 mysql-instances.meta 为各个全量导出任务 metadata 的 binlog name，position 和 gtid 信息
    3. 修改其中的 GTID 信息的 1-y 为 previous_gtids 的前段值，例如，上述例子需要改为 6-y
    4. 配置 task.yaml 中的 syncers.safe-mode 为 true 并重启任务
    5. 待增量同步追上后，停止任务并在任务配置文件中设置 safe-mode 为 false
    6. 再次重启任务
    7. 重启 source 并关闭 gtid 或 relay
- 如果上述条件均不满足或任务同步数据量较小：
    1. 清空下游数据库中数据
    2. 重启 source 并关闭 gtid 或 relay
    3. 重建任务并通过 start-task task.yaml --remove-meta 重新同步

上述处理方案中，针对正常同步的 source（如上例 mysql2），重设增量任务时起始点需设置 mysql-instances.
↪ meta 为 subTaskStatus.sync 的 syncerBinlog 与 syncerBinlogGtid。

在 DM 2.0 中开启 heartbeat，虚拟 IP 环境下切换 DM-worker 与 MySQL 实例的连接，遇到 “heartbeat config is different from previous used: serverID not equal” 错误

heartbeat 功能在 DM v2.0 及之后版本已经默认关闭，如果用户在同步任务配置文件中开启会干扰高可用特性，在配置文件中关闭该项（通过设置 enable-heartbeat: false，然后更新任务配置）即可解决。DM 将会在后续版本强制关闭该功能。

DM-master 在重启后无法加入集群，报错信息为 “fail to start embed etcd, RawCause: member xxx has already been bootstrapped”

DM-master 会在启动时将 etcd 信息记录在当前目录。如果重启后当前目录发生变化，会导致 DM 缺失 etcd 信息，从而启动失败。

推荐使用 TiUP 运维 DM 避免这一问题。在需要使用二进制部署的场合，需要在 DM-master 配置文件中使用绝对路径配置 data-dir 项，或者注意运行命令的当前目录。

使用 dmctl 执行命令时无法连接 DM-master

在使用 dmctl 执行相关命令时，发现连接 DM-master 失败（即使已在命令中指定 --master-addr 的参数值），报错内容类似 RawCause: context deadline exceeded, Workaround: please check your network connection.，但使用 telnet <master-addr> 之类的命令检查网络却没有发现异常。

这种情况可以检查下环境变量 https_proxy（注意，这里是 https ）。如果配置了该环境变量，dmctl 会自动去连接 https_proxy 指定的主机及端口，而如果该主机没有相应的 proxy 转发服务，则会导致连接失败。

解决方案：确认 https_proxy 是否必须要配置，如果不是必须的，取消该设置即可。如果环境必须，那么在原命令前加环境变量设置 https_proxy="" ./dmctl --master-addr "x.x.x.x:8261" 即可。

> 注意：
>
> 关于 proxy 的环境变量有 http_proxy，https_proxy，no_proxy 等。如果依据上述解决方案处理后仍无法连接，可以考虑检查 http_proxy 和 no_proxy 的参数配置是否有影响。

v2.0.2 - v2.0.6 版本执行 start-relay 命令报错该如何处理？

```
flush local meta, Rawcause: open relay-dir/xxx.000001/relay.metayyyy: no such file or directory
```

上述报错在以下情况下有可能会被触发：

- DM 从 v2.0.1 及之前的版本升级到 v2.0.2 - v2.0.6 版本，且升级之前曾开启过 relay log，升级完后重新开启。
- 使用 stop-relay 命令暂停 relay log 后重新开启 relay log。

可以通过以下方式解决该问题：

- 重启 relay log:

```
» stop-relay -s sourceID workerName
» start-relay -s sourceID workerName
```

- 升级 DM 至 v2.0.7 或之后版本。

12.8.7.3 TiDB Data Migration 版本发布历史

从 DM v5.4.0 起，TiDB Data Migration 的 Release Notes 合并入相同版本号的 TiDB Release Notes。

- 如需阅读 v5.4.0 及之后版本的 DM Release Notes，请查看对应版本的 TiDB Release Notes 中 DM 相关的内容。
- 如需阅读 v5.3.0 及更早版本的 DM Release Notes，请参考以下链接：

12.8.7.3.1 5.3

- 5.3.0

12.8.7.3.2 2.0

- 2.0.7
- 2.0.6
- 2.0.5
- 2.0.4
- 2.0.3
- 2.0.2
- 2.0.1
- 2.0 GA
- 2.0.0-rc.2
- 2.0.0-rc

12.8.7.3.3 1.0

- 1.0.7
- 1.0.6
- 1.0.5
- 1.0.4
- 1.0.3
- 1.0.2

12.9 Backup & Restore (BR)

12.9.1 备份与恢复工具 BR 简介

BR 全称为 Backup & Restore，是 TiDB 分布式备份恢复的命令行工具，用于对 TiDB 集群进行数据备份和恢复。

相比 `dumpling`，BR 更适合大数据量的场景。

BR 除了可以用来进行常规备份恢复外，也可以在保证兼容性前提下用来做大规模的数据迁移。

本文介绍了 BR 的工作原理、推荐部署配置、使用限制以及几种使用方式。

12.9.1.1 工作原理

BR 将备份或恢复操作命令下发到各个 TiKV 节点。TiKV 收到命令后执行相应的备份或恢复操作。

在一次备份或恢复中，各个 TiKV 节点都会有一个对应的备份路径，TiKV 备份时产生的备份文件将会保存在该路径下，恢复时也会从该路径读取相应的备份文件。

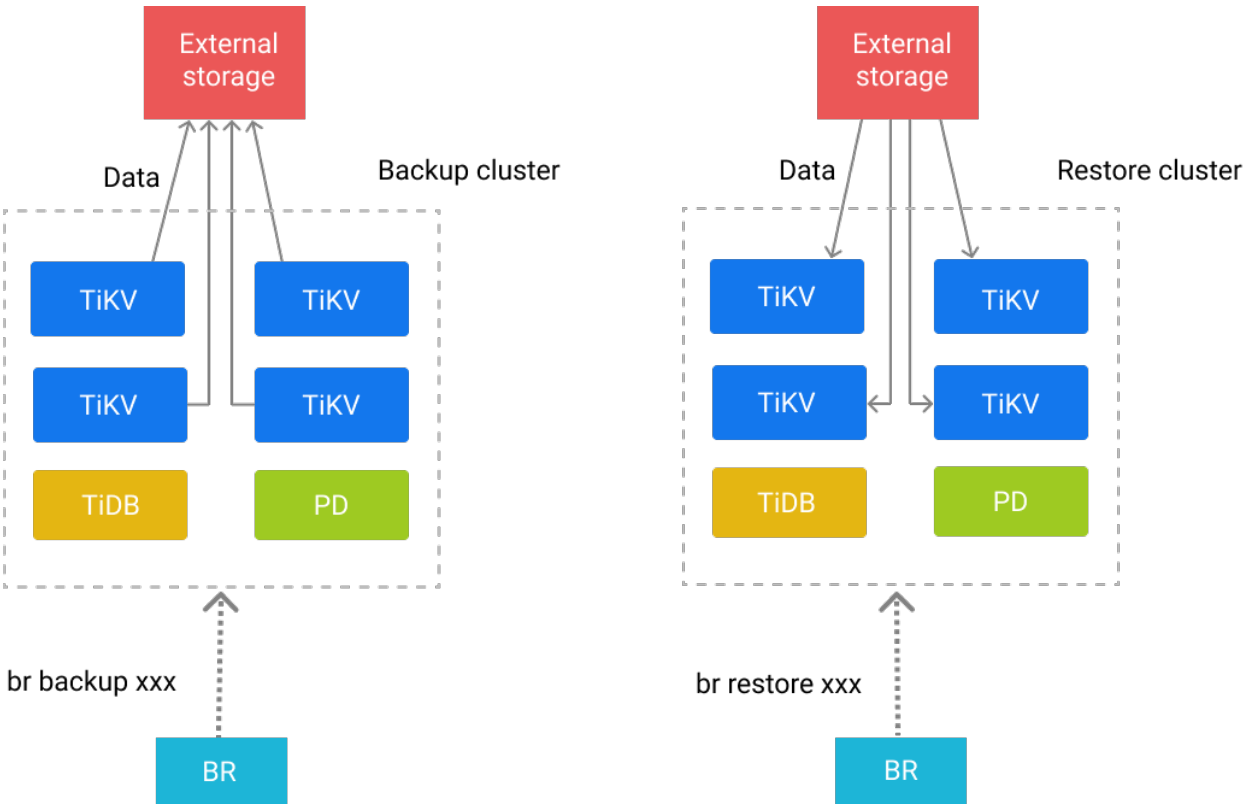


图 295: br-arch

更多信息请参阅备份恢复设计方案。

12.9.1.1.1 备份文件类型

备份路径下会生成以下两种类型文件：

- SST 文件：存储 TiKV 备份下来的数据信息
- `backupmeta` 文件：存储本次备份的元信息，包括备份文件数、备份文件的 Key 区间、备份文件大小和备份文件 Hash (sha256) 值
- `backup.lock` 文件：用于防止多次备份到同一目录

12.9.1.1.2 SST 文件命名格式

SST 文件以 `storeID_regionID_regionEpoch_keyHash_cf` 的格式命名。格式名的解释如下：

- storeID：TiKV 节点编号

- regionID：Region 编号
- regionEpoch：Region 版本号
- keyHash：Range startKey 的 Hash (sha256) 值，确保唯一性
- cf：RocksDB 的 ColumnFamily（默认为 `default` 或 `write`）

12.9.1.2 部署使用 BR 工具

12.9.1.2.1 推荐部署配置

- 推荐 BR 部署在 PD 节点上。
- 推荐使用一块高性能 SSD 网盘，挂载到 BR 节点和所有 TiKV 节点上，网盘推荐万兆网卡，否则带宽有可能成为备份恢复时的性能瓶颈。

> 注意：
>
> - 如果没有挂载网盘或者使用其他共享存储，那么 BR 备份的数据会生成在各个 TiKV 节点上。由于 BR 只备份 leader 副本，所以各个节点预留的空间需要根据 leader size 来预估。
> - 同时由于 TiDB 默认使用 leader count 进行平衡，所以会出现 leader size 差别大的问题，导致各个节点备份数据不均衡。

12.9.1.2.2 使用限制

下面是使用 BR 进行备份恢复的几条限制：

- BR 恢复到 TiCDC / Drainer 的上游集群时，恢复数据无法由 TiCDC / Drainer 同步到下游。
- BR 只支持在 `new_collations_enabled_on_first_bootstrap` 开关值相同的集群之间进行操作。这是因为 BR 仅备份 KV 数据。如果备份集群和恢复集群采用不同的排序规则，数据校验会不通过。所以恢复集群时，你需要确保 `select VARIABLE_VALUE from mysql.tidb where VARIABLE_NAME='new_collation_enabled';` 语句的开关值查询结果与备份时的查询结果相一致，才可以进行恢复。

12.9.1.2.3 兼容性

BR 和 TiDB 集群的兼容性问题分为以下两方面：

- BR 部分版本和 TiDB 集群的接口不兼容

BR 在 v5.4.0 之前不支持恢复 `charset=GBK` 的表。并且，任何版本的 BR 都不支持恢复 `charset=GBK` 的表到 5.4.0 之前的 TiDB 集群。

- 某些功能在开启或关闭状态下，会导致 KV 格式发生变化，因此备份和恢复期间如果没有统一开启或关闭，就会带来不兼容的问题

下表整理了会导致 KV 格式发生变化的功能。

功能	相关 issue	解决方式
聚簇索引	#565	确保备份时 `tidb_enable_clustered_index` 全局变量和恢复时一致，否则会导致数据不一致的问题，例如 `default not found` 和数据索引不一致。
New collation	#352	确保恢复时集群的 `new_collations_enabled_on_first_bootstrap` 变量值和备份时的一致，否则会导致数据索引不一致和 checksum 通不过。

功能	相关 issue	解决方式
恢复集群开启 TiCDC 同步	#364	TiKV 暂不能将 BR ingest 的 SST 文件下推到 TiCDC，因此使用 BR 恢复时候需要关闭 TiCDC。
全局临时表		确保使用 BR v5.3.0 及以上版本进行备份和恢复，否则会导致全局临时表的表定义错误。

在上述功能确保备份恢复一致的前提下，BR 和 TiKV/TiDB/PD 还可能因为版本内部协议不一致/接口不一致出现不兼容的问题，因此 BR 内置了版本检查。

版本检查

BR 内置版本会在执行备份和恢复操作前，对 TiDB 集群版本和自身版本进行对比检查。如果大版本不匹配（比如 BR v4.x 和 TiDB v5.x 上），BR 会提示退出。如要跳过版本检查，可以通过设置 `--check-requirements=false` 强行跳过版本检查。

需要注意的是，跳过检查可能会遇到版本不兼容的问题。BR 和 TiDB 各版本兼容情况如下表所示：

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR nightly 备份 TiDB nightly	✓	✓	✓（如果恢复了使用非整数类型聚簇主键的表到 v4.0 的 TiDB 集群，BR 会无任何警告地导致数据错误）

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR v5.0 备份 TiDB v5.0	✓	✓	✓（如果恢复了使用非整数类型聚簇主键的表到 v4.0 的 TiDB 集群，BR 会无任何警告地导致数据错误）

恢复版本（横向） 备份版本（纵向）	用 BR nightly 恢复 TiDB nightly	用 BR v5.0 恢复 TiDB v5.0	用 BR v4.0 恢复 TiDB v4.0
用 BR v4.0 备份 TiDB v4.0	□	□	□（如果 TiKV >= v4.0.0-rc.1，BR 包含 #233 Bug 修复，且 TiKV 不包含 #7241 Bug 修复，那么 BR 会导致 TiKV 节点重启）
用 BR nightly 或 v5.0 备份 TiDB v4.0	□（当 TiDB 版本小于 v4.0.9 时会出现 #609 问题）	□（当 TiDB 版本小于 v4.0.9 会出现 #609 问题）	□（当 TiDB 版本小于 v4.0.9 会出现 #609 问题）

12.9.1.2.4 备份和恢复 mysql 系统库下的表数据（实验特性）

> **警告：**
>
> 该功能目前为实验性功能，尚未经过充分测试。不建议在生产环境中使用该功能。

在 v5.1.0 之前，BR 备份时会过滤掉系统库 `mysql.*` 的表数据。自 v5.1.0 起，BR 默认备份集群内的全部数据，包括系统库 `mysql.*` 中的数据。但由于恢复 `mysql.*` 中系统表数据的技术实现尚不完善，因此 BR 默认不恢复系统库 `mysql` 中的表数据。

如果你希望将系统表（如 `mysql.usertable1`）的数据恢复到系统库 `mysql` 中，可以设置`filter` 参数 来过滤表名 (`-f "mysql.usertable1"`)。完成设置后，系统库下的表数据会恢复到临时库中，然后通过对临时库表进行重命名的方式恢复到系统库。

由于技术原因，以下系统表不能被正确恢复。即使你指定了 `-f "mysql.*"`，以下表也不会被恢复：

- 统计信息相关的表：“stats_buckets”，“stats_extended”，“stats_feedback”，“stats_fm_sketch”，“stats_histograms”，“stats_meta”，“stats_top_n”
- 权限或系统相关的表：“tidb”，“global_variables”，“columns_priv”，“db”，“default_roles”，“global_grants”，“global_priv”，“role_edges”，“tables_priv”，“user”，“gc_delete_range”，“gc_delete_range_done”，“schema_index_usage”

12.9.1.2.5 运行 BR 的最低机型配置要求

运行 BR 的最低机型配置要求如下：

CPU	内存	硬盘类型	网络
1 核	4 GB	HDD	千兆网卡

一般场景下（备份恢复的表少于 1000 张），BR 在运行期间的 CPU 消耗不会超过 200%，内存消耗不会超过 4 GB。但在备份和恢复大量数据表时，BR 的内存消耗可能会上升到 4 GB 以上。在实际测试中，备份 24000 张表大概需要消耗 2.7 GB 内存，CPU 消耗维持在 100% 以下。

12.9.1.2.6 最佳实践

下面是使用 BR 进行备份恢复的几种推荐操作：

- 推荐在业务低峰时执行备份操作，这样能最大程度地减少对业务的影响。
- BR 支持在不同拓扑的集群上执行恢复，但恢复期间对在线业务影响很大，建议低峰期或者限速 (`rate-limit`) 执行恢复。
- BR 备份最好串行执行。不同备份任务并行会导致备份性能降低，同时也会影响在线业务。
- BR 恢复最好串行执行。不同恢复任务并行会导致 Region 冲突增多，恢复的性能降低。
- 推荐在 `-s` 指定的备份路径上挂载一个共享存储，例如 NFS。这样能方便收集和管理备份文件。
- 在使用共享存储时，推荐使用高吞吐的存储硬件，因为存储的吞吐会限制备份或恢复的速度。
- BR 默认会分别在备份、恢复完成后，进行一轮数据校验，将文本数据同集群数据比较，来保证正确性。在迁移场景下，为了减少迁移耗时，推荐备份时关闭校验 (`--checksum=false`)，只在恢复时开启校验。

12.9.1.2.7 使用方式

目前支持以下几种方式来运行 BR 工具，分别是通过 SQL 语句、命令行工具或在 Kubernetes 环境下进行备份恢复。

通过 SQL 语句

TiDB 支持使用 SQL 语句BACKUP 和RESTORE 进行备份恢复。如果要查看备份恢复的进度，你可以使用 SHOW BACKUPS|RESTORES 语句。

通过命令行工具

TiDB 支持使用 BR 命令行工具进行备份恢复（需手动下载）。关于 BR 命令行工具的具体使用方法，请参阅使用备份与恢复工具 BR。

在 Kubernetes 环境下

目前支持使用 BR 工具备份 TiDB 集群数据到兼容 S3 的存储、Google Cloud Storage 以及持久卷，并作恢复：

> 注意：
>
> Amazon S3 和 Google Cloud Storage (GCS) 参数描述见外部存储文档。

- 备份 TiDB 集群数据到兼容 S3 的存储
- 恢复 S3 兼容存储上的备份数据
- 备份 TiDB 集群到 Google Cloud Storage
- 恢复 Google Cloud Storage 上的备份数据
- 备份 TiDB 集群到持久卷
- 恢复持久卷上的备份数据

12.9.1.3 BR 相关文档

- 使用 BR 命令行备份恢复
- BR 备份与恢复场景示例
- BR 常见问题
- 外部存储

12.9.2 使用 BR 命令行进行备份恢复

本文介绍如何 BR 命令行进行 TiDB 集群数据的备份和恢复。

在阅读本文前，请确保你已通读备份与恢复工具 BR 简介，尤其是使用限制和最佳实践这两节。

12.9.2.1 BR 命令行描述

一条 br 命令是由子命令、选项和参数组成的。子命令即不带 - 或者 -- 的字符。选项即以 - 或者 -- 开头的字符。参数即子命令或选项字符后紧跟的、并传递给命令和选项的字符。

以下是一条完整的 br 命令行：

```
br backup full --pd "${PDIP}:2379" -s "local:///tmp/backup"
```

命令行各部分的解释如下：

- `backup`：br 的子命令

- `full`：backup 的子命令
- `-s` 或 `--storage`：备份保存的路径
- `"local:///tmp/backup"`：`-s` 的参数，保存的路径为各个 TiKV 节点本地磁盘的 `/tmp/backup`
- `--pd`：PD 服务地址
- `"${PDIP}:2379"`：`--pd` 的参数

> 注意：
>
> 在使用 `local` storage 的时候，备份数据会分散在各个节点的本地文件系统中。
>
> 不建议在生产环境中备份到本地磁盘，因为在日后恢复的时候，必须手动聚集这些数据才能完成恢复工作（见恢复集群数据）。
>
> 聚集这些备份数据可能会造成数据冗余和运维上的麻烦，而且在不聚集这些数据便直接恢复的时候会遇到颇为迷惑的 `SST file not found` 报错。
>
> 建议在各个节点挂载 NFS 网盘，或者直接备份到 S3 对象存储中。

12.9.2.1.1 命令和子命令

BR 由多层命令组成。目前，BR 包含 `backup`、`restore` 和 `version` 三个子命令：

- `br backup` 用于备份 TiDB 集群
- `br restore` 用于恢复 TiDB 集群

以上三个子命令可能还包含这些子命令：

- `full`：可用于备份或恢复全部数据。
- `db`：可用于备份或恢复集群中的指定数据库。
- `table`：可用于备份或恢复集群指定数据库中的单张表。

12.9.2.1.2 常用选项

- `--pd`：用于连接的选项，表示 PD 服务地址，例如 `"${PDIP}:2379"`。
- `-h/--help`：获取所有命令和子命令的使用帮助。例如 `br backup --help`。
- `-V`（或 `--version`）：检查 BR 版本。
- `--ca`：指定 PEM 格式的受信任 CA 的证书文件路径。
- `--cert`：指定 PEM 格式的 SSL 证书文件路径。
- `--key`：指定 PEM 格式的 SSL 证书密钥文件路径。
- `--status-addr`：BR 向 Prometheus 提供统计数据的监听地址。

12.9.2.2 使用 BR 命令行备份集群数据

使用 `br backup` 命令来备份集群数据。可选择添加 `full` 或 `table` 子命令来指定备份的范围：全部集群数据或单张表的数据。

12.9.2.2.1 备份全部集群数据

要备份全部集群数据，可使用 `br backup full` 命令。该命令的使用帮助可以通过 `br backup full -h` 或 `br backup full --help` 来获取。

用例：将所有集群数据备份到各个 TiKV 节点的 `/tmp/backup` 路径，同时也会将备份的元信息文件 `backupmeta` 写到该路径下。

> 注意：
>
> - 经测试，在全速备份的情况下，如果备份盘和服务盘不同，在线备份会让只读线上服务的 QPS 下降 15%~25% 左右。如果希望降低影响，请参考 `--ratelimit` 进行限速。
> - 假如备份盘和服务盘相同，备份将会和服务争夺 I/O 资源，这可能会让只读线上服务的 QPS 骤降一半以上。请尽量禁止将在线服务的数据备份到 TiKV 的数据盘。

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
```

以上命令中，`--ratelimit` 选项限制了每个 TiKV 执行备份任务的速度上限（单位 MiB/s）。`--log-file` 选项指定把 BR 的 log 写到 `backupfull.log` 文件中。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。进度条效果如下：

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
Full Backup <---------/.....................................................> 17.12%.
```

12.9.2.2.2 备份单个数据库的数据

要备份集群中指定单个数据库的数据，可使用 `br backup db` 命令。同样可通过 `br backup db -h` 或 `br backup db --help` 来获取子命令 db 的使用帮助。

用例：将数据库 `test` 备份到各个 TiKV 节点的 `/tmp/backup` 路径，同时也会将备份的元信息文件 `backupmeta` 写到该路径下。

```
br backup db \
    --pd "${PDIP}:2379" \
    --db test \
```

```
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backuptable.log
```

db 子命令的选项为 --db，用来指定数据库名。其他选项的含义与备份全部集群数据相同。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。

12.9.2.2.3 备份单张表的数据

要备份集群中指定单张表的数据，可使用 br backup table 命令。同样可通过 br backup table -h 或 br backup table --help 来获取子命令 table 的使用帮助。

用例：将表 test.usertable 备份到各个 TiKV 节点的 /tmp/backup 路径，同时也会将备份的元信息文件 backupmeta 写到该路径下。

```
br backup table \
    --pd "${PDIP}:2379" \
    --db test \
    --table usertable \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backuptable.log
```

table 子命令有 --db 和 --table 两个选项，分别用来指定数据库名和表名。其他选项的含义与备份全部集群数据相同。

备份期间有进度条在终端中显示。当进度条前进到 100% 时，说明备份已完成。在完成备份后，BR 为了确保数据安全性，还会校验备份数据。

12.9.2.2.4 使用表库过滤功能备份多张表的数据

如果你需要以更复杂的过滤条件来备份多个表，执行 br backup full 命令，并使用 --filter 或 -f 来指定表库过滤规则。

用例：以下命令将所有 db*.tbl* 形式的表格数据备份到每个 TiKV 节点上的 /tmp/backup 路径，并将 backupmeta 文件写入该路径。

```
br backup full \
    --pd "${PDIP}:2379" \
    --filter 'db*.tbl*' \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file backupfull.log
```

12.9.2.2.5 备份数据到 Amazon S3 后端存储

如果备份的存储并不是在本地，而是在 Amazon 的 S3 后端存储，那么需要在 storage 子命令中指定 S3 的存储路径，并且赋予 BR 节点和 TiKV 节点访问 Amazon S3 的权限。

这里可以参照 AWS 官方文档在指定的 Region 区域中创建一个 S3 桶 Bucket，如果有需要，还可以参照 AWS 官方文档在 Bucket 中创建一个文件夹 Folder。

> 注意：
>
> 要完成一次备份，通常 TiKV 和 BR 需要的最小权限为 s3:ListBucket，s3:PutObject 和 s3:AbortMultipartUpload。

将有权限访问该 S3 后端存储的账号的 SecretKey 和 AccessKey 作为环境变量传入 BR 节点，并且通过 BR 将权限传给 TiKV 节点。

```
export AWS_ACCESS_KEY_ID=${AccessKey}
export AWS_SECRET_ACCESS_KEY=${SecretKey}
```

在进行 BR 备份时，显示指定参数 --s3.region 和 --send-credentials-to-tikv, --s3.region 表示 S3 存储所在的区域，--send-credentials-to-tikv 表示将 S3 的访问权限传递给 TiKV 节点。

```
br backup full \
    --pd "${PDIP}:2379" \
    --storage "s3://${Bucket}/${Folder}" \
    --s3.region "${region}" \
    --send-credentials-to-tikv=true \
    --ratelimit 128 \
    --log-file backuptable.log
```

12.9.2.2.6 增量备份

如果想要备份增量，只需要在备份的时候指定上一次的备份时间戳 --lastbackupts 即可。

注意增量备份有以下限制：

- 增量备份需要与前一次全量备份在不同的路径下
- GC safepoint 必须在 lastbackupts 之前

```
br backup full\
    --pd ${PDIP}:2379 \
    --ratelimit 128 \
    -s local:///home/tidb/backupdata/incr \
    --lastbackupts ${LAST_BACKUP_TS}
```

以上命令会备份 (LAST_BACKUP_TS, current PD timestamp] 之间的增量数据。

你可以使用 validate 指令获取上一次备份的时间戳，示例如下：

```
LAST_BACKUP_TS=`br validate decode --field="end-version" -s local:///home/tidb/backupdata | tail
    ↪ -n1`
```

示例备份的增量数据记录 (LAST_BACKUP_TS, current PD timestamp] 之间的数据变更，以及这段时间内的 DDL。在恢复的时候，BR 会先把所有 DDL 恢复，而后才会恢复数据。

12.9.2.2.7 加密备份数据（实验性功能）

自 TiDB v5.3.0 起，TiDB 开始支持备份加密功能，你可配置下列参数在备份过程中到达加密数据的效果：

- --crypter.method：加密算法，支持 aes128-ctr/aes192-ctr/aes256-ctr 三种算法，缺省值为 plaintext，表示不加密
- --crypter.key：加密密钥，十六进制字符串格式，aes128-ctr 对应 128 位（16 字节）密钥长度，aes192-ctr 为 24 字节，aes256-ctr 为 32 字节
- --crypter.key-file：密钥文件，可直接将存放密钥的文件路径作为参数传入，此时 crypter.key 不需要传入

> 警告：
>
> - 当前该功能为实验特性，不建议在生产环境中使用。
> - 密钥丢失，备份的数据将无法恢复到集群中。
> - 加密功能需在 br 工具和 TiDB 集群都不低于 v5.3.0 的版本上使用，且加密备份得到的数据无法在低于 v5.3.0 版本的集群上恢复。

备份加密的示例如下：

```
br backup full\
    --pd ${PDIP}:2379 \
    -s local:///home/tidb/backupdata/incr \
    --crypter.method aes128-ctr \
    --crypter.key 0123456789abcdef0123456789abcdef
```

12.9.2.2.8 Raw KV 备份（实验性功能）

> 警告：
>
> Raw KV 备份功能还在实验中，没有经过完备的测试。暂时请避免在生产环境中使用该功能。

在某些使用场景下，TiKV 可能会独立于 TiDB 运行。考虑到这点，BR 也提供跳过 TiDB 层，直接备份 TiKV 中数据的功能：

```
br backup raw --pd $PD_ADDR \
    -s "local://$BACKUP_DIR" \
    --start 31 \
    --ratelimit 128 \
    --end 3130303030303030 \
    --format hex \
    --cf default
```

以上命令会备份 default CF 上 [0x31, 0x3130303030303030) 之间的所有键到 $BACKUP_DIR 去。

这里，--start 和 --end 的参数会先依照 --format 指定的方式解码，再被送到 TiKV 上去，目前支持以下解码方式：

- “raw”：不进行任何操作，将输入的字符串直接编码为二进制格式的键。
- “hex”：将输入的字符串视作十六进制数字。这是默认的编码方式。
- “escape”：对输入的字符串进行转义之后，再编码为二进制格式。

12.9.2.3 使用 BR 命令行恢复集群数据

使用 br restore 命令来恢复备份数据。可选择添加 full、db 或 table 子命令来指定恢复操作的范围：全部集群数据、某个数据库或某张数据表。

> 注意：
>
> 如果使用本地存储，在恢复前必须将所有备份的 SST 文件复制到各个 TiKV 节点上 --storage 指定的目录下。
>
> 即使每个 TiKV 节点最后只需要读取部分 SST 文件，这些节点也需要有所有 SST 文件的完全访问权限。原因如下：
>
> - 数据被复制到了多个 Peer 中。在读取 SST 文件时，这些文件必须要存在于所有 Peer 中。这与数据的备份不同，在备份时，只需从单个节点读取。
> - 在数据恢复的时候，每个 Peer 分布的位置是随机的，事先并不知道哪个节点将读取哪个文件。
>
> 使用共享存储可以避免这些情况。例如，在本地路径上安装 NFS，或使用 S3。利用这些网络存储，各个节点都可以自动读取每个 SST 文件，此时上述注意事项不再适用。
>
> 同时，请注意同一时间对同一个集群只能运行一个恢复任务，否则可能会出现非预期的行为，详见FAQ。

12.9.2.3.1 恢复全部备份数据

要将全部备份数据恢复到集群中来，可使用 `br restore full` 命令。该命令的使用帮助可以通过 `br restore full -h` 或 `br restore full --help` 来获取。

用例：将 `/tmp/backup` 路径中的全部备份数据恢复到集群中。

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file restorefull.log
```

以上命令中，`--ratelimit` 选项限制了每个 TiKV 执行恢复任务的速度上限（单位 MiB/s）。`--log-file` 选项指定把 BR 的 log 写到 `restorefull.log` 文件中。

恢复期间还有进度条会在终端中显示，当进度条前进到 100% 时，说明恢复已完成。在完成恢复后，BR 为了确保数据安全性，还会校验恢复数据。进度条效果如下：

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "local:///tmp/backup" \
    --ratelimit 128 \
    --log-file restorefull.log
Full Restore <---------/.....................................................> 17.12%.
```

12.9.2.3.2 恢复单个数据库的数据

要将备份数据中的某个数据库恢复到集群中，可以使用 `br restore db` 命令。该命令的使用帮助可以通过 `br restore db -h` 或 `br restore db --help` 来获取。

用例：将 `/tmp/backup` 路径中备份数据中的某个数据库恢复到集群中。

```
br restore db \
    --pd "${PDIP}:2379" \
    --db "test" \
    --ratelimit 128 \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

以上命令中 `--db` 选项指定了需要恢复的数据库名字。其余选项的含义与恢复全部备份数据相同。

> 注意：
>
> 恢复备份数据的时候，`--db` 选项指定的数据库名必须与执行备份时候 `--db`选项指定的数据库名相同，否则无法恢复成功。由于备份数据的元文件`backupmeta` 记录了该数据库名，因此只能将数据恢复到同名的数据库。推荐做法是把备份文件恢复到另一个集群的同名数据库中。

12.9.2.3.3 恢复单张表的数据

要将备份数据中的某张数据表恢复到集群中，可以使用 `br restore table` 命令。该命令的使用帮助可通过 `br restore table -h` 或 `br restore table --help` 来获取。

用例：将 `/tmp/backup` 路径下的备份数据中的某个数据表恢复到集群中。

```
br restore table \
    --pd "${PDIP}:2379" \
    --db "test" \
    --table "usertable" \
    --ratelimit 128 \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

12.9.2.3.4 使用表库功能过滤恢复数据

如果你需要用复杂的过滤条件来恢复多个表，执行 `br restore full` 命令，并用 `--filter` 或 `-f` 指定使用表库过滤。

用例：以下命令将备份在 `/tmp/backup` 路径的表的子集恢复到集群中。

```
br restore full \
    --pd "${PDIP}:2379" \
    --filter 'db*.tbl*' \
    --storage "local:///tmp/backup" \
    --log-file restorefull.log
```

12.9.2.3.5 从 Amazon S3 后端存储恢复数据

如果需要恢复的数据并不是存储在本地，而是在 Amazon 的 S3 后端，那么需要在 `storage` 子命令中指定 S3 的存储路径，并且赋予 BR 节点和 TiKV 节点访问 Amazon S3 的权限。

> **注意：**
>
> 要完成一次恢复，通常 TiKV 和 BR 需要的最小权限为 `s3:ListBucket` 和 `s3:GetObject`。

将有权限访问该 S3 后端存储的账号的 `SecretKey` 和 `AccessKey` 作为环境变量传入 BR 节点，并且通过 BR 将权限传给 TiKV 节点。

```
export AWS_ACCESS_KEY_ID=${AccessKey}
export AWS_SECRET_ACCESS_KEY=${SecretKey}
```

在进行 BR 恢复时，显示指定参数 `--s3.region` 和 `--send-credentials-to-tikv`, `--s3.region` 表示 S3 存储所在的区域，`--send-credentials-to-tikv` 表示将 S3 的访问权限传递给 TiKV 节点。`--storage`参数中的 `Bucket` 和 `Folder` 分别代表需要恢复的数据所在的 S3 存储桶和文件夹。

```
br restore full \
    --pd "${PDIP}:2379" \
    --storage "s3://${Bucket}/${Folder}" \
    --s3.region "${region}" \
    --ratelimit 128 \
    --send-credentials-to-tikv=true \
    --log-file restorefull.log
```

以上命令中 `--table` 选项指定了需要恢复的表名。其余选项的含义与恢复单个数据库相同。

12.9.2.3.6 增量恢复

增量恢复的方法和使用 BR 进行全量恢复的方法并无差别。需要注意，恢复增量数据的时候，需要保证备份时指定的 `last backup ts` 之前备份的数据已经全部恢复到目标集群。

12.9.2.3.7 恢复创建在 mysql 数据库下的表（实验性功能）

BR 可以并且会默认备份 `mysql` 数据库下的表。

在执行恢复时，`mysql` 下的表默认不会被恢复。如果需要恢复 `mysql` 下的用户创建的表，可以通过table filter 来显式地包含目标表。以下示例中要恢复目标用户表为 `mysql.usertable`；该命令会在执行正常的恢复的同时恢复 `mysql.usertable`。

```
br restore full -f '*.*' -f '!mysql.*' -f 'mysql.usertable' -s $external_storage_url --ratelimit
    ↪ 128
```

在如上的命令中，`-f '*.*'` 用于覆盖掉默认的规则，`-f '!mysql.*'` 指示 BR 不要恢复 `mysql` 中的表，除非另有指定。`-f 'mysql.usertable'` 则指定需要恢复 `mysql.usertable`。具体原理请参考table filter 的文档。

如果只需要恢复 `mysql.usertable`，而无需恢复其他表，可以使用以下命令：

```
br restore full -f 'mysql.usertable' -s $external_storage_url --ratelimit 128
```

> **警告：**
>
> 系统表（例如 `mysql.tidb`）可以通过 BR 进行备份。但恢复系统表存在限制。即便是使用了 `-filter` 设置，也不能通过 BR 恢复以下系统表：
>
> - 统计信息表（`mysql.stat_*`）
> - 系统变量表（`mysql.tidb`，`mysql.global_variables`）
> - 用户信息表（`mysql.user`，`mysql.columns_priv`，等等）
> - 其他系统表
>
> 恢复系统表可能还存在更多兼容性问题。为了防止意外发生，请避免在生产环境中恢复系统表。

12.9.2.3.8 解密恢复数据（实验性功能）

> **警告：**
>
> - 当前该功能为实验特性，不建议在生产环境中使用。

在对数据做加密备份后，恢复操作需传入相应的解密参数，解密算法或密钥不正确则无法恢复，解密参数和加密参数一致即可。解密恢复的示例如下：

```
br restore full\
    --pd ${PDIP}:2379 \
    -s local:///home/tidb/backupdata/incr \
    --crypter.method aes128-ctr \
    --crypter.key 0123456789abcdef0123456789abcdef
```

12.9.2.3.9 Raw KV 恢复（实验性功能）

> **警告：**
>
> Raw KV 恢复功能还在实验中，没有经过完备的测试。暂时请避免在生产环境中使用该功能。

和Raw KV 备份相似地，恢复 Raw KV 的命令如下：

```
br restore raw --pd $PD_ADDR \
    -s "local://$BACKUP_DIR" \
    --start 31 \
    --end 3130303030303030 \
    --ratelimit 128 \
    --format hex \
    --cf default
```

以上命令会将范围在 `[0x31, 0x3130303030303030)` 的已备份键恢复到 TiKV 集群中。这里键的编码方式和备份时相同。

12.9.2.3.10 在线恢复（实验性功能）

> **警告：**
>
> 在线恢复功能还在实验中，没有经过完备的测试，同时还依赖 PD 的不稳定特性 Placement Rules。暂时请避免在生产环境中使用该功能。

在恢复的时候，写入过多的数据会影响在线集群的性能。为了尽量避免影响线上业务，BR 支持通过Placement rules 隔离资源。让下载、导入 SST 的工作仅仅在指定的几个节点（下称“恢复节点”）上进行，具体操作如下：

1. 配置 PD，启动 Placement rules：

```
echo "config set enable-placement-rules true" | pd-ctl
```

2. 编辑恢复节点 TiKV 的配置文件，在 server 一项中指定：

```
[server]
labels = { exclusive = "restore" }
```

3. 启动恢复节点的 TiKV，使用 BR 恢复备份的文件，和非在线恢复相比，这里只需要加上 --online 标志即可：

```
br restore full \
    -s "local://$BACKUP_DIR" \
    --ratelimit 128 \
    --pd $PD_ADDR \
    --online
```

12.9.3 BR 备份与恢复场景示例

BR 是一款分布式的快速备份和恢复工具。

本文展示了以下几种备份和恢复场景下的 BR 操作过程：

- 将单表数据备份到网络盘（推荐生产环境使用）
- 从网络磁盘恢复备份数据（推荐生产环境使用）
- 将单表数据备份到本地磁盘（推荐测试环境试用）
- 从本地磁盘恢复备份数据（推荐测试环境试用）

以帮助读者达到以下目标：

- 正确使用网络盘或本地盘进行备份或恢复
- 通过相关监控指标了解备份或恢复的状态
- 了解在备份或恢复时如何调优性能
- 处理备份时可能发生的异常

12.9.3.1 目标读者

你需要对 TiDB 和 TiKV 有一定的了解。

在阅读本文前，请确保你已通读备份与恢复工具 BR 简介，尤其是使用限制和最佳实践这两节。

12.9.3.2 环境准备

本节介绍 TiDB 的推荐部署方式、BR 使用示例中的集群版本、TiKV 集群硬件信息和集群配置。

你可以根据自己的硬件和配置来预估备份恢复的性能。

12.9.3.2.1 部署方式

推荐使用TiUP 部署 TiDB 集群，再下载TiDB Toolkit 获取 BR 工具。

12.9.3.2.2 集群版本

- TiDB: v5.0.0
- TiKV: v5.0.0
- PD: v5.0.0
- BR: v5.0.0

> **注意：**
>
> v5.0.0 为编写本文档时的最新版本。推荐使用最新版本 TiDB/TiKV/PD/BR，同时需要确保 BR 版本和 TiDB 相同。

12.9.3.2.3 TiKV 集群硬件信息

- 操作系统：CentOS Linux release 7.6.1810 (Core)
- CPU：16-Core Common KVM processor
- RAM：32GB
- 硬盘：500G SSD * 2
- 网卡：万兆网卡

12.9.3.2.4 配置

BR 可以直接将命令下发到 TiKV 集群来执行备份和恢复，不依赖 TiDB server 组件，因此无需对 TiDB server 进行配置。

- TiKV: 默认配置
- PD : 默认配置

12.9.3.3 使用场景

本节描述以下几种使用场景：

- 将单表数据备份到网络盘（推荐生产环境使用）
- 从网络磁盘恢复备份数据（推荐生产环境使用）

- 将单表数据备份到本地磁盘（推荐测试环境试用）
- 从本地磁盘恢复备份数据（推荐测试环境试用）

推荐使用网络盘来进行备份和恢复操作，这样可以省去收集备份数据文件的繁琐步骤。尤其在 TiKV 集群规模较大的情况下，使用网络盘可以大幅提升操作效率。

在使用 BR 进行备份或恢复操作前，需要先进行如下准备工作。

12.9.3.3.1 备份前的准备工作

BR 工具已支持自适应 GC，会自动将 backupTS（默认是最新的 PD timestamp）注册到 PD 的 `safePoint`，保证 TiDB 的 GC Safe Point 在备份期间不会向前移动，即可避免手动设置 GC。

关于 `br backup` 命令的具体使用方法，参见使用备份与恢复工具 BR。

运行 `br backup` 命令前，请确保以下条件：

1. TiDB 集群中没有正在运行中的 DDL。
2. 用于创建备份的存储设备有足够的空间。

12.9.3.3.2 恢复前的准备工作

使用 BR 进行恢复前的准备工作如下：

运行`br restore` 命令前，需要检查新集群，确保集群内没有同名的表。

12.9.3.3.3 将单表数据备份到网络盘（推荐生产环境使用）

使用 `br backup` 命令，将单表数据 `--db batchmark --table order_line` 备份到指定的网络盘路径 `local:///br_data` 下。

前置要求

- 备份前的准备工作。
- 配置一台高性能 SSD 硬盘主机为 NFS server 存储数据。其他所有 BR 节点，TiKV 节点和 TiFlash 节点为 NFS client，挂载相同的路径（例如 `/br_data`）到 NFS server 上以访问 NFS server。
- NFS server 和 NFS client 间的数据传输速率至少要达到备份集群的 `TiKV 实例数 * 150MB/s`。否则网络 I/O 有可能成为性能瓶颈。

> 注意：
>
> - 因为备份时候只备份单副本 (leader) 数据，所以即使集群中存在 TiFlash 副本，无需挂载 TiFlash 节点 BR 也能完成备份。
> - BR 在恢复数据时，会恢复全部副本的数据。因此在恢复时，TiFlash 节点需要有备份数据的访问权限 BR 才能完成恢复，此时也必须将 TiFlash 节点挂载到 NFS server 上。

部署拓扑

部署拓扑如下图所示：

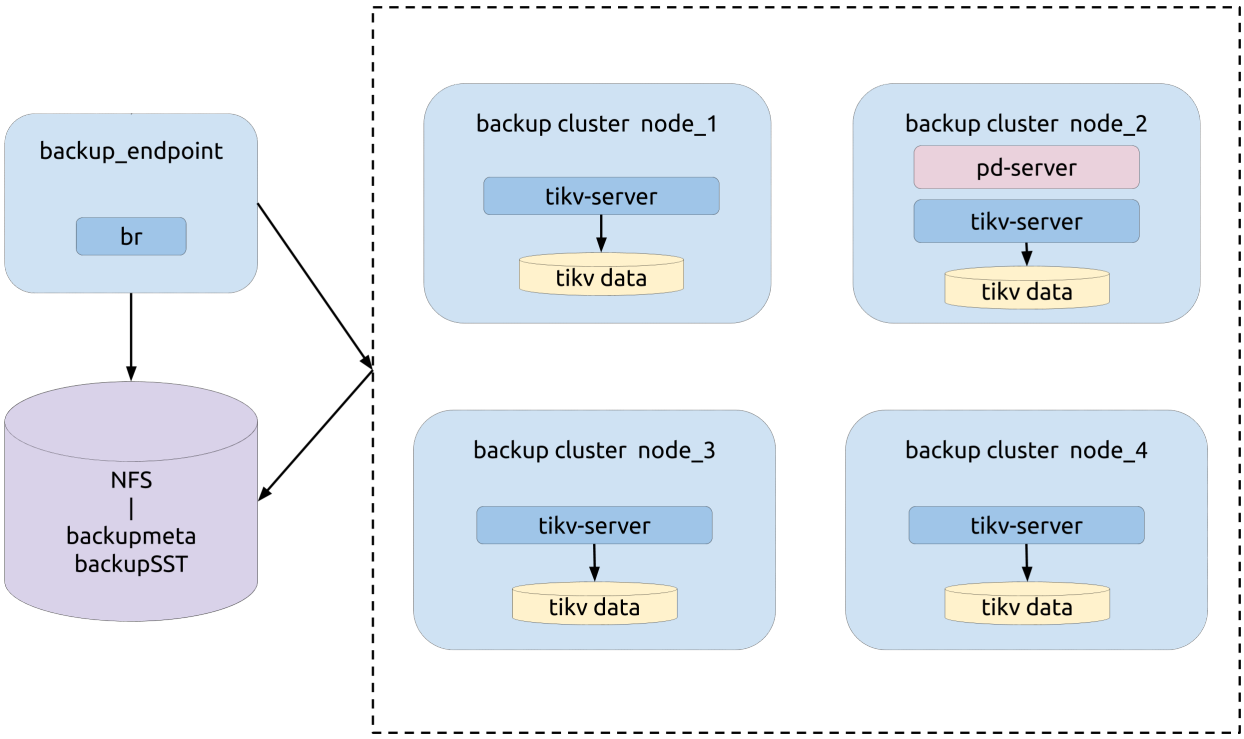


图 296: img

运行备份

备份操作前，在 TiDB 中使用 `admin checksum table order_line` 命令获得备份目标表 `--db batchmark --table order_line` 的统计信息。统计信息示例如下：

```
+-----------+------------+---------------------+------------+--------------+
| Db_name   | Table_name | Checksum_crc64_xor  | Total_kvs  | Total_bytes  |
+-----------+------------+---------------------+------------+--------------+
| batchmark | order_line | 10912722838344822475 | 5659888624 | 370385538778 |
+-----------+------------+---------------------+------------+--------------+
1 row in set (5 min 47.59 sec)
```

图 297: img

运行 BR 备份命令：

```
bin/br backup table \
    --db batchmark \
    --table order_line \
```

```
    -s local:///br_data \
    --pd ${PD_ADDR}:2379 \
    --log-file backup-nfs.log
```

备份过程中的运行指标

在 BR 备份过程中，需关注以下监控面版中的运行指标来了解备份的状态。

Backup CPU Utilization：参与备份的 TiKV 节点（例如 backup-worker 和 backup-endpoint）的 CPU 使用率。

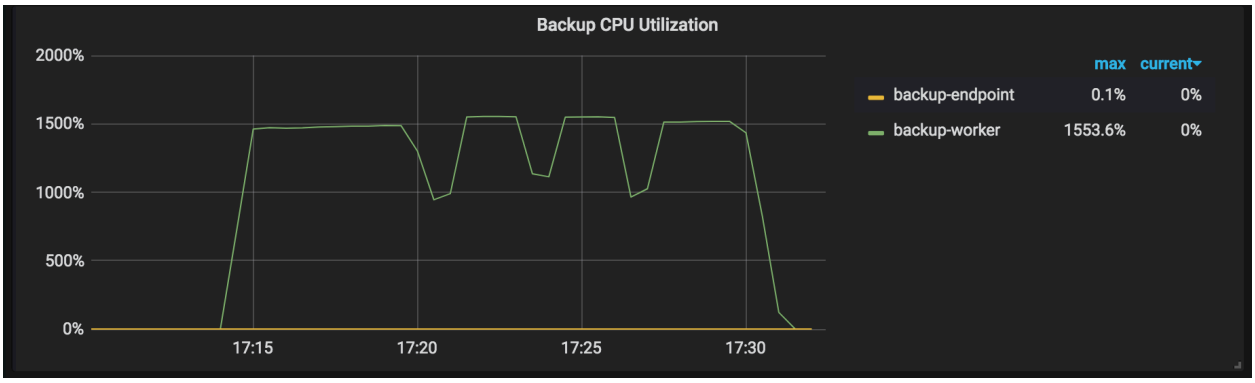


图 298: img

IO Utilization：参与备份的 TiKV 节点的 I/O 使用率。

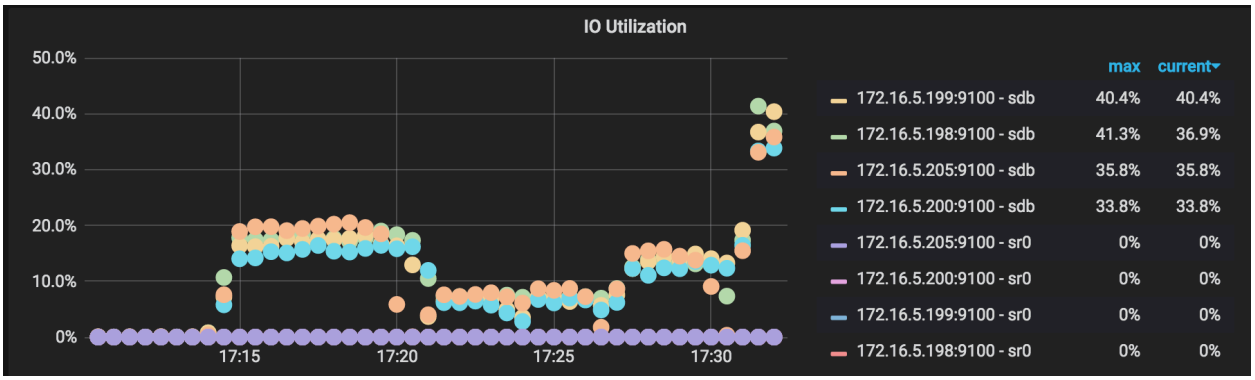


图 299: img

BackupSST Generation Throughput：参与备份的 TiKV 节点生成 backupSST 文件的吞吐。正常时单个 TiKV 节点的吞吐在 150MB/s 左右。

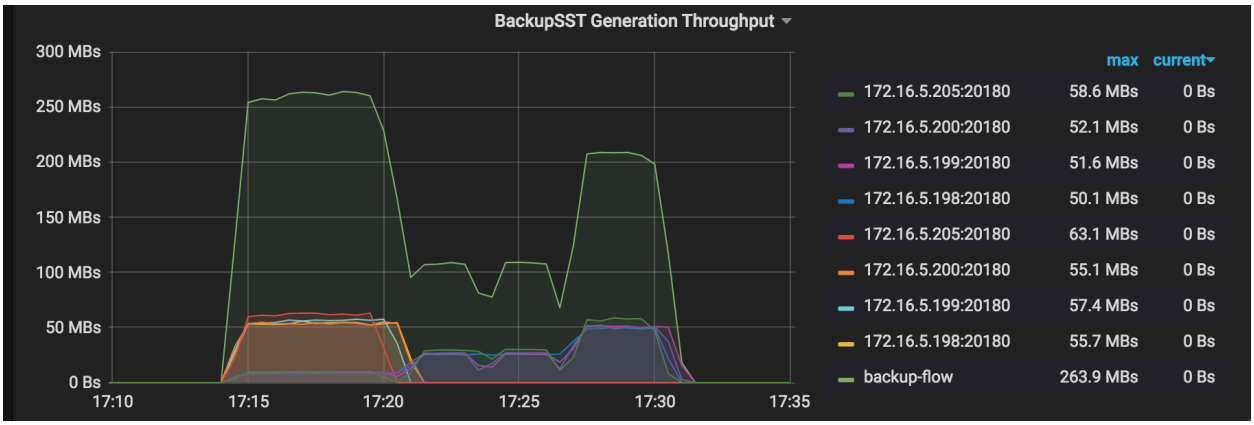


图 300: img

One Backup Range Duration：备份一个 range 的操作耗时，包括扫描耗时 (scan KV) 和保存耗时（保存为 backupSST 文件）。

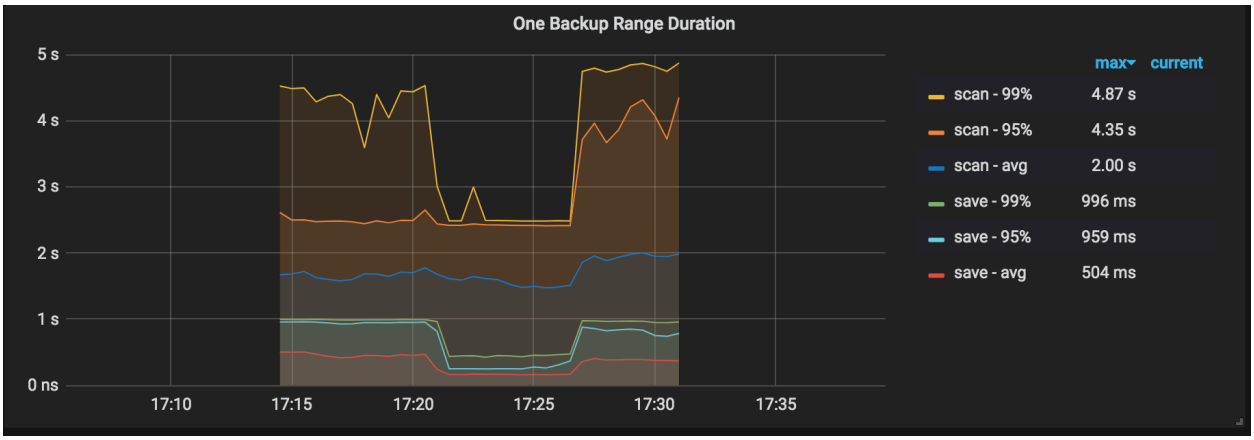


图 301: img

One Backup Subtask Duration：一次备份任务会被拆分成多个子任务。该监控项显示子任务的耗时。

注意：

- 虽然本次任务是备份单表，但因为表中有 3 个索引，所以正常会拆分成 4 个子任务。
- 下图中有 13 个点，说明有 9 次 (13-4) 重试。备份过程中可能发生 Region 调度行为，少量重试是正常的。

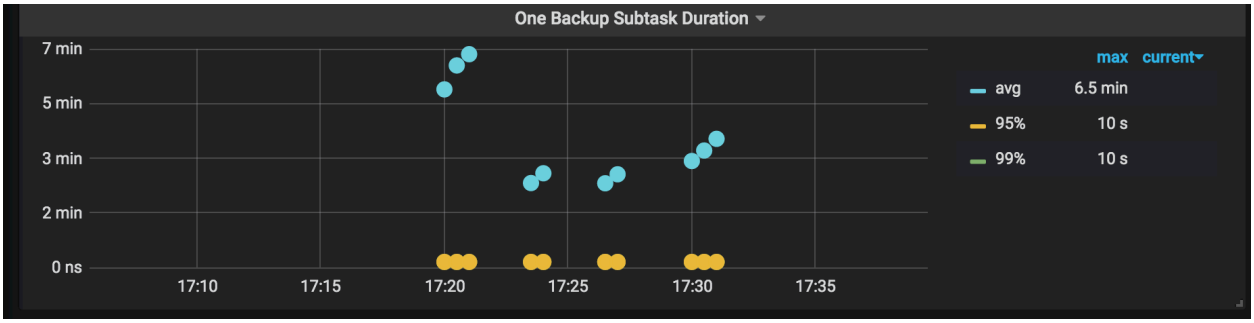


图 302: img

Backup Errors：备份过程中的错误。正常时无错误。即使出现少量错误，备份操作也有重试机制，可能会导致备份时间增加，但不会影响备份的正确性。

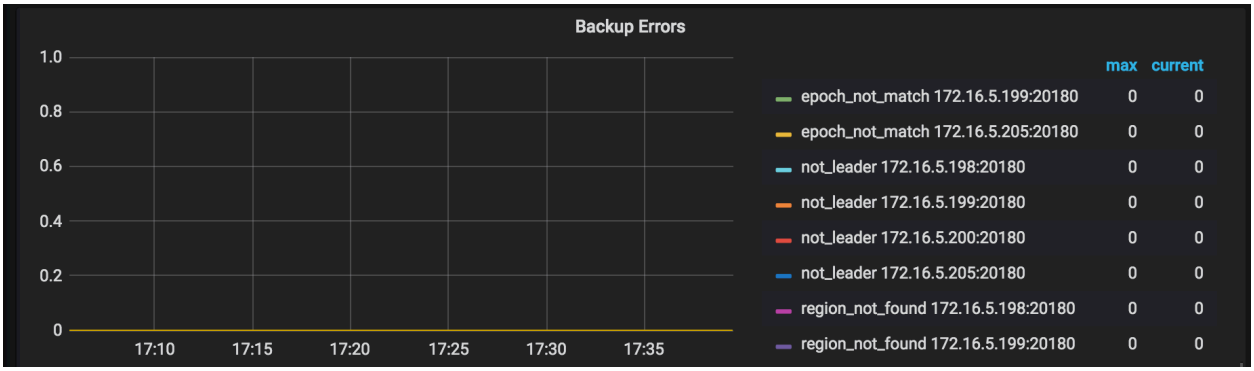


图 303: img

Checksum Request Duration：对备份集群执行 admin checksum 的耗时统计。

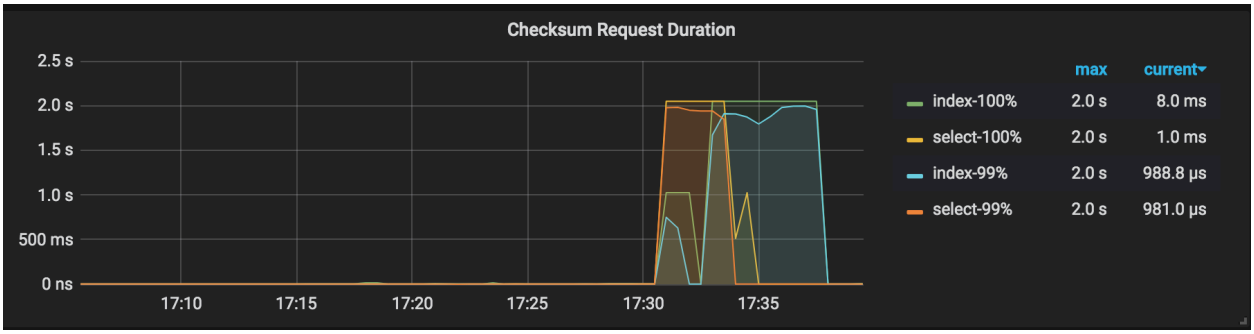


图 304: img

结果解读

BR 会在备份结束时输出备份总结到控制台。

同时使用 BR 前已设置日志的存放路径。从路径下存放的日志中可以获取此次备份的相关统计信息。在日志中搜关键字“summary”，可以看到以下信息：

```
["Full backup Success summary:
    total backup ranges: 2,
    total success: 2,
    total failed: 0,
    total take(Full backup time): 31.802912166s,
    total take(real time): 49.799662427s,
    total size(MB): 5997.49,
    avg speed(MB/s): 188.58,
    total kv: 120000000"]
    ["backup checksum"=17.907153678s]
    ["backup fast checksum"=349.333µs]
    ["backup total regions"=43]
    [BackupTS=422618409346269185]
    [Size=826765915]
```

以上日志信息中包含以下内容：

- 备份耗时：`total take(Full backup time): 31.802912166s`
- 程序运行总耗时：`total take(real time): 49.799662427s`
- 备份数据大小：`total size(MB): 5997.49`
- 备份吞吐：`avg speed(MB/s): 188.58`
- 备份 KV 对数：`total kv: 120000000`
- 校验耗时：`["backup checksum"=17.907153678s]`
- 计算各表 checksum、KV 和 bytes 信息总和的耗时：`["backup fast checksum"=349.333µs]`
- 备份 Region 总数：`["backup total regions"=43]`
- 备份存档经压缩后在磁盘中的实际大小：`[Size=826765915]`
- 备份存档的快照时间戳：`[BackupTS=422618409346269185]`

通过以上数据可以计算得到单个 TiKV 实例的吞吐为：`avg speed(MB/s)/tikv_count = 62.86`。

性能调优

如果 TiKV 的资源使用没有出现明显的瓶颈（例如备份过程中的运行指标中的 Backup CPU Utilization 最高为 1500% 左右，IO Utilization 普遍低于 30%），可以尝试调大 `--concurrency`（默认是 4）参数以进行性能调优。该方法不适用于存在许多小表的场景。

示例如下：

```
bin/br backup table \
    --db batchmark \
    --table order_line \
    -s local:///br_data/ \
    --pd ${PD_ADDR}:2379 \
    --log-file backup-nfs.log \
    --concurrency 16
```

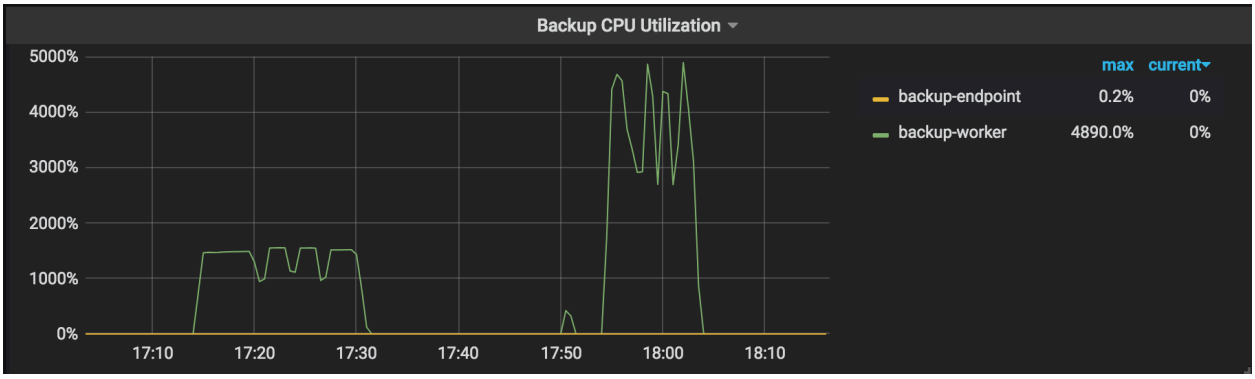


图 305: img

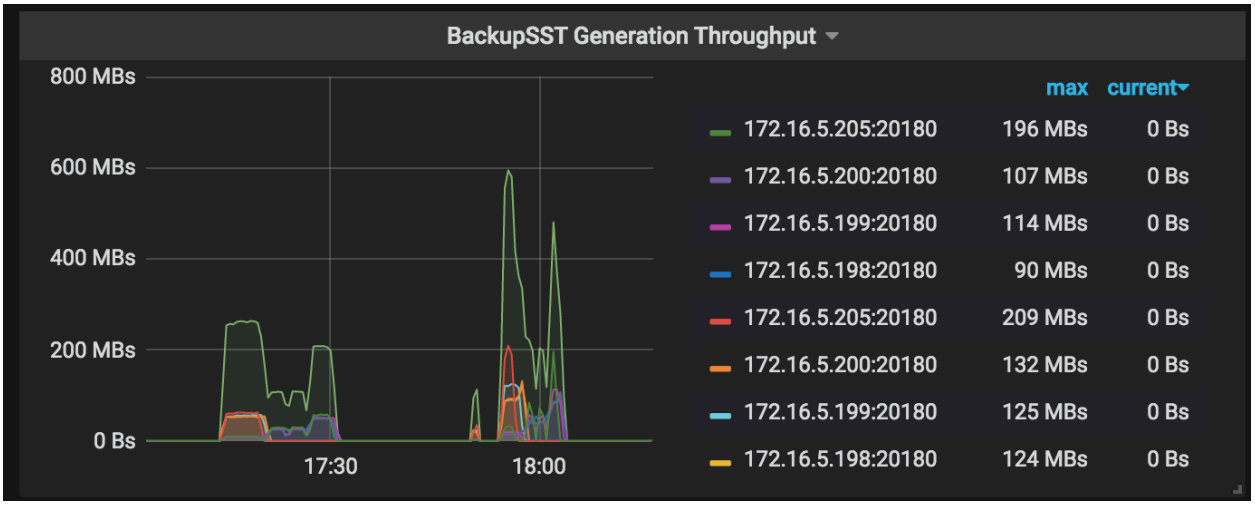


图 306: img

性能调优后的结果如下所示（保持数据大小不变）：

- 备份耗时：`total take(s)` 从 `986.43` 减少到 `535.53`
- 数据大小：`total size(MB): 353227.18`
- 备份吞吐：`avg speed(MB/s)` 从 `358.09` 提升到 `659.59`
- 单个 TiKV 实例的吞吐：`avg speed(MB/s)/tikv_count` 从 `89` 提升到 `164.89`

12.9.3.3.4 从网络磁盘恢复备份数据（推荐生产环境使用）

使用 `br restore` 命令，将一份完整的备份数据恢复到一个离线集群。暂不支持恢复到在线集群。

前置要求

- 恢复前的准备工作。

部署拓扑

部署拓扑如下图所示：

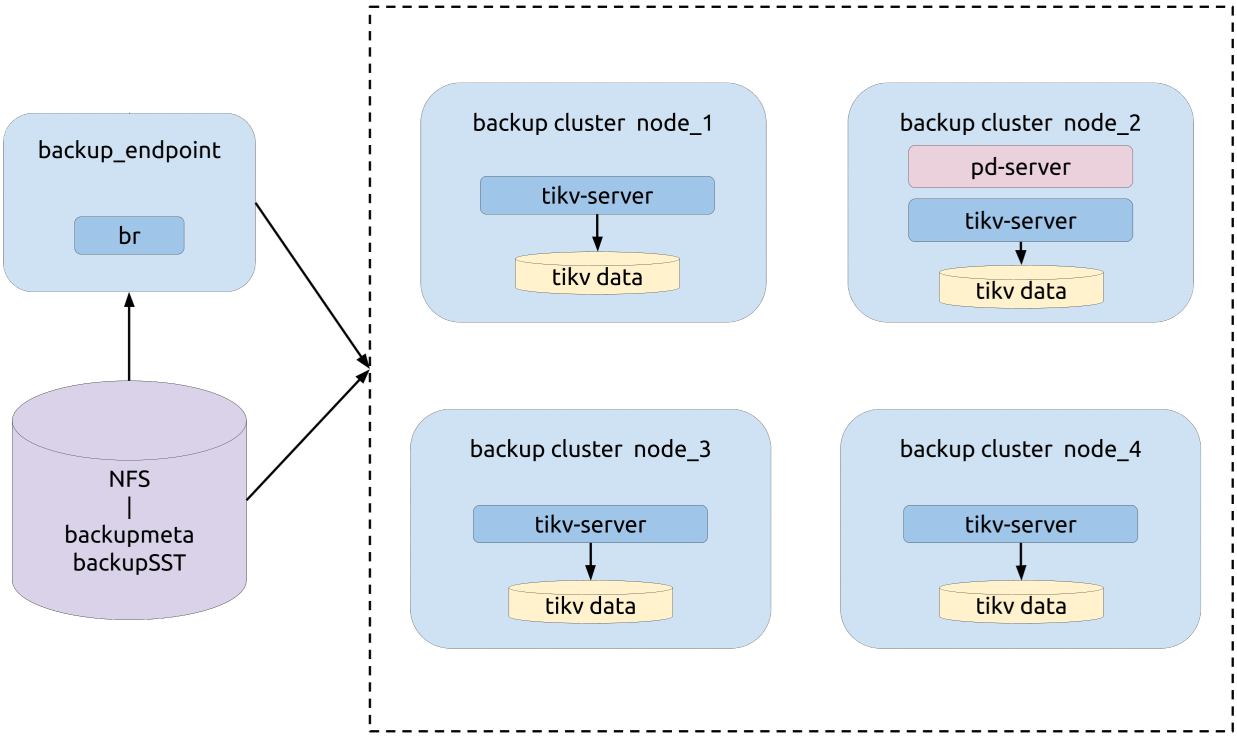


图 307: img

运行恢复

运行 `br restore` 命令：

```
bin/br restore table --db batchmark --table order_line -s local:///br_data --pd 172.16.5.198:2379
    ↪  --log-file restore-nfs.log
```

恢复过程中的运行指标

在 BR 恢复过程中，需关注以下监控面版中的运行指标来了解恢复的状态。

CPU Utilization：参与恢复的 TiKV 节点 CPU 使用率。

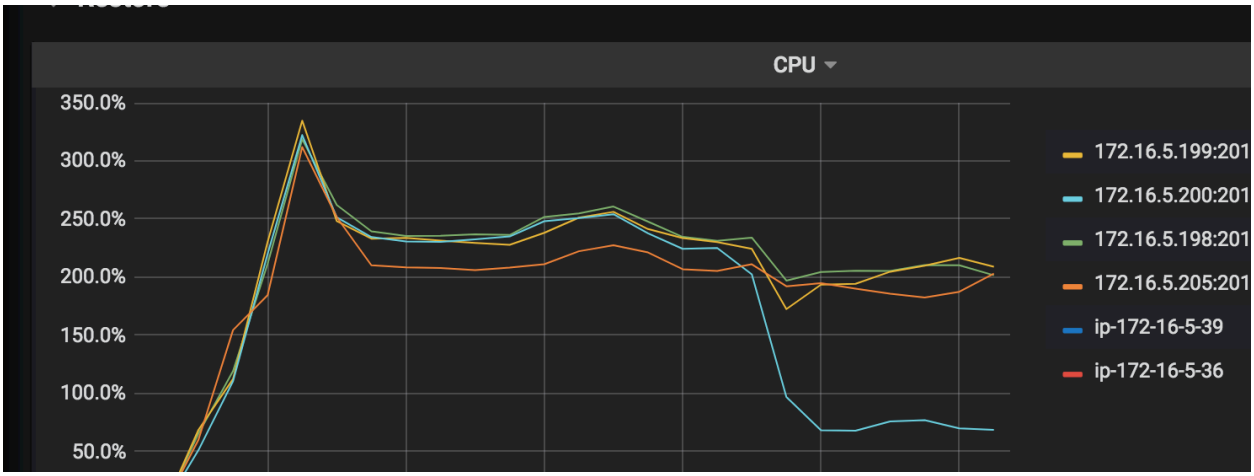


图 308: img

IO Utilization：参与恢复的 TiKV 节点的 I/O 使用率。

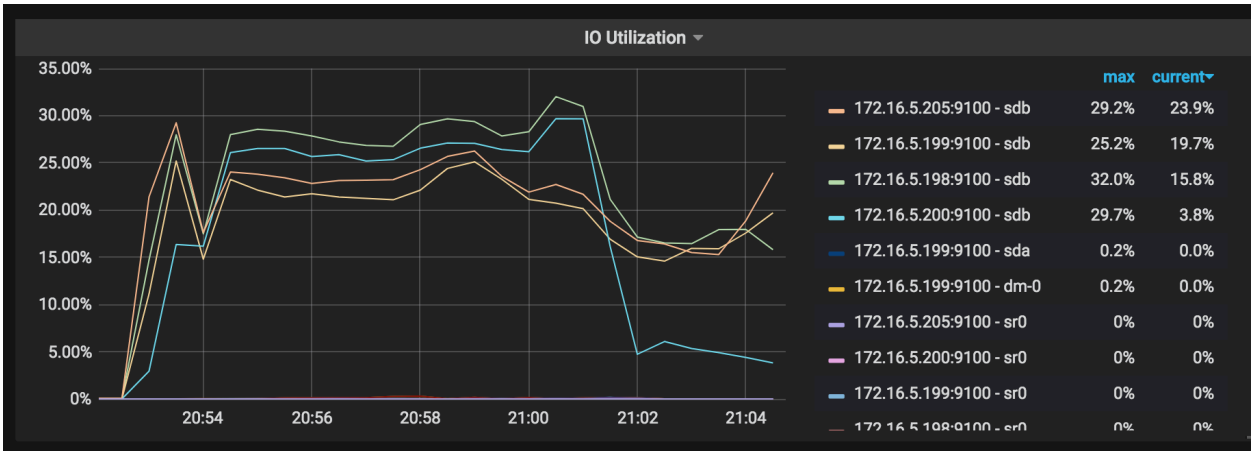


图 309: img

Region 分布：Region 分布越均匀，说明恢复资源利用越充分。

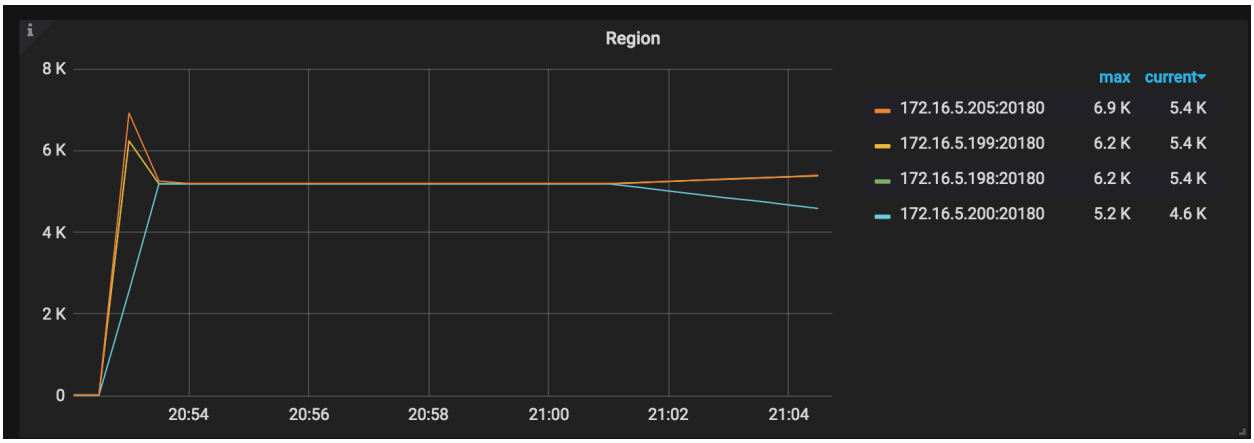


图 310: img

Process SST Duration：处理 SST 文件的延迟。恢复一张表时，如果 `tableID` 发生了变化，需要对 `tableID` 进行 `rewrite`，否则会进行 `rename`。通常 `rewrite` 延迟要高于 `rename` 的延迟。

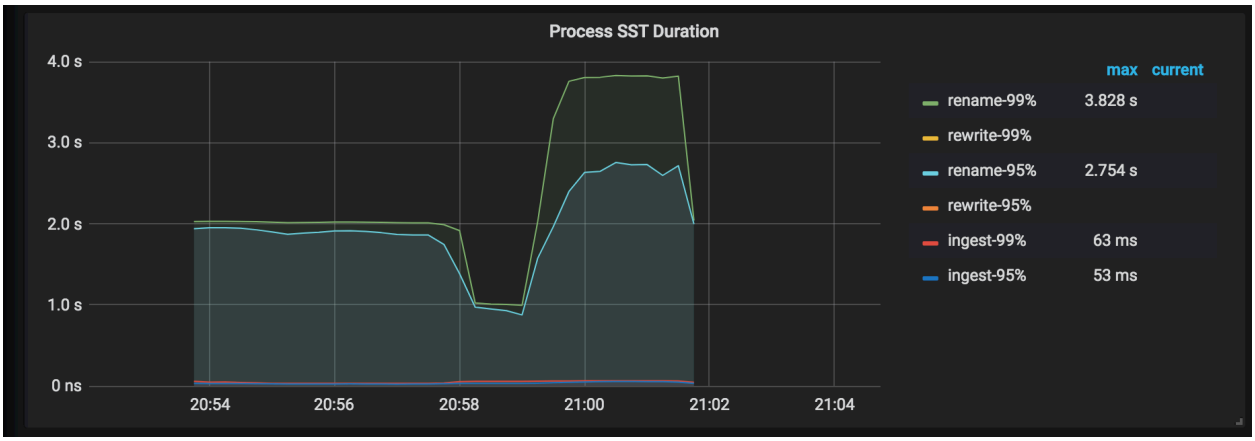


图 311: img

DownLoad SST Throughput：从 External Storage 下载 SST 文件的吞吐。

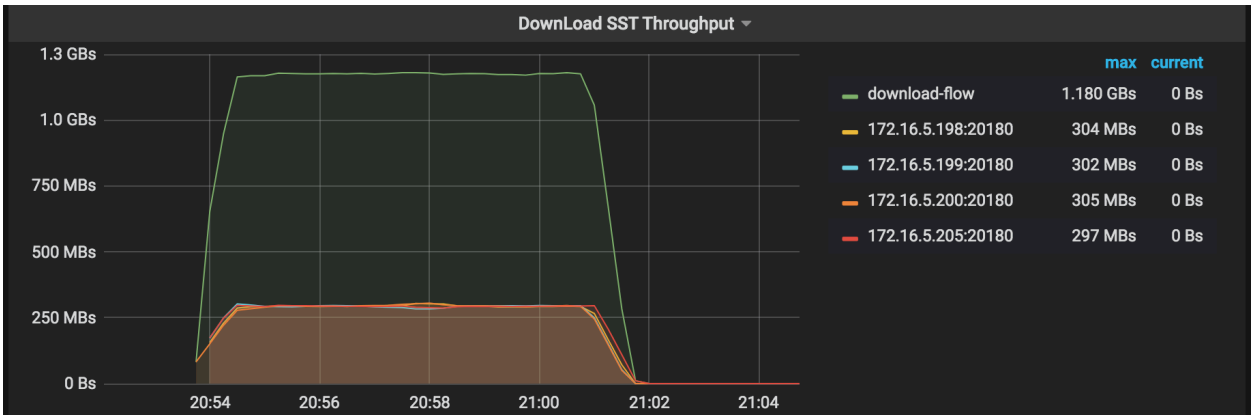


图 312: img

Restore Errors：恢复过程中的错误。

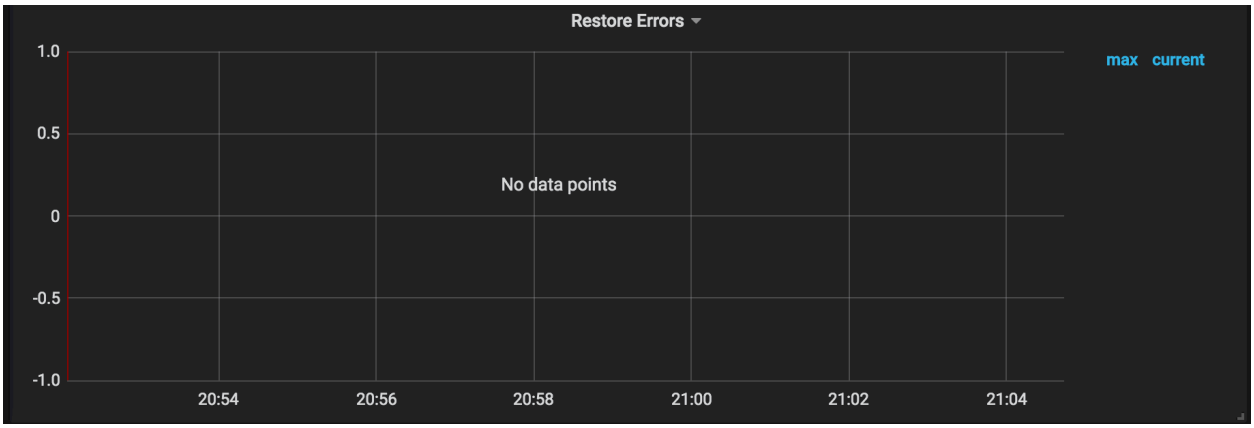


图 313: img

Checksum Request duration：对恢复集群执行 admin checksum 的耗时统计，会比备份时的 checksum 延迟高。

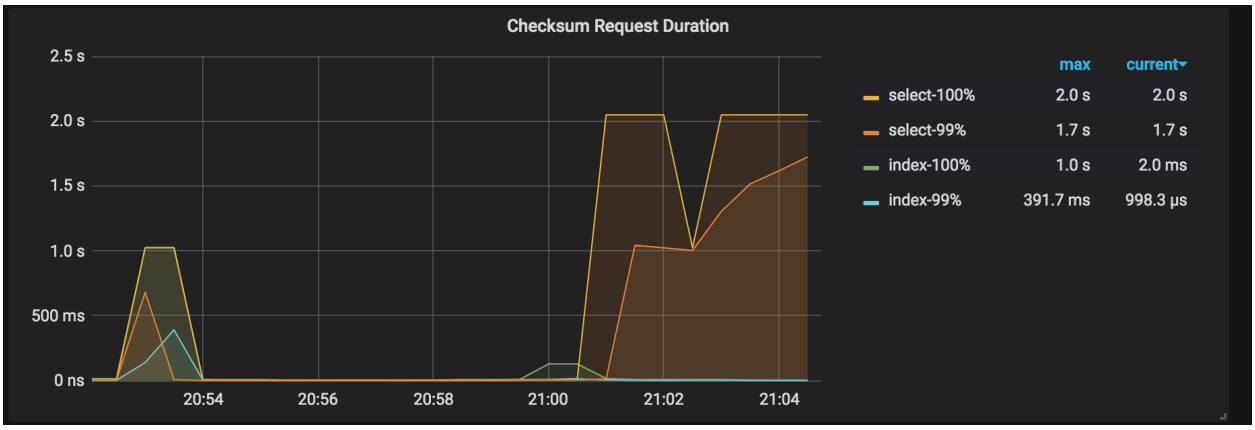


图 314: img

结果解读

使用 BR 前已设置日志的存放路径。从路径下存放的日志中可以获取此次恢复的相关统计信息。在日志中搜关键字“summary”，可以看到以下信息：

```
["Table Restore summary:
    total restore tables: 1,
    total success: 1,
    total failed: 0,
    total take(Full restore time): 17m1.001611365s,
    total take(real time): 16m1.371611365s,
    total kv: 5659888624,
    total size(MB): 353227.18,
    avg speed(MB/s): 367.42"]
    ["restore files"=9263]
    ["restore ranges"=6888]
    ["split region"=49.049182743s]
    ["restore checksum"=6m34.879439498s]
    [Size=48693068713]
```

以上日志信息中包含以下内容：

- 恢复耗时：`total take(Full restore time): 17m1.001611365s`
- 程序运行总耗时：`total take(real time): 16m1.371611365s`
- 恢复数据大小：`total size(MB): 353227.18`
- 恢复 KV 对数：`total kv: 5659888624`
- 恢复吞吐：`avg speed(MB/s): 367.42`
- Region Split 耗时：`take=49.049182743s`
- 校验耗时：`restore checksum=6m34.879439498s`
- 恢复存档在磁盘中的实际大小：`[Size=48693068713]`

根据上表数据可以计算得到：

- 单个 TiKV 吞吐：avg speed(MB/s)/tikv_count = 91.8
- 单个 TiKV 平均恢复速度：total size(MB)/(split time + restore time)/tikv_count = 87.4

性能调优

如果 TiKV 资源使用没有明显的瓶颈，可以尝试调大 --concurrency 参数（默认为 128），示例如下：

```
bin/br restore table --db batchmark --table order_line -s local:///br_data/ --pd
    ↪ 172.16.5.198:2379 --log-file restore-concurrency.log --concurrency 1024
```

性能调优后的结果如下表所示（保持数据大小不变）：

- 恢复耗时：total take(s) 从 961.37 减少到 443.49
- 恢复吞吐：avg speed(MB/s) 从 367.42 提升到 796.47
- 单个 TiKV 实例的吞吐：avg speed(MB/s)/tikv_count 从 91.8 提升到 199.1
- 单个 TiKV 实例的平均恢复速度：total size(MB)/(split time + restore time)/tikv_count 从 87.4 提升到 162.3

12.9.3.3.5 将单表数据备份到本地磁盘（推荐测试环境试用）

使用 br backup 命令，将单表数据 --db batchmark --table order_line 备份到指定的本地磁盘路径 local ↪ :///home/tidb/backup_local 下。

前置要求

- 备份前的准备工作。
- 各个 TiKV 节点有单独的磁盘用来存放 backupSST 数据。
- backup_endpoint 节点有单独的磁盘用来存放备份的 backupmeta 文件。
- TiKV 和 backup_endpoint 节点需要有相同的备份目录，例如 /home/tidb/backup_local。

部署拓扑

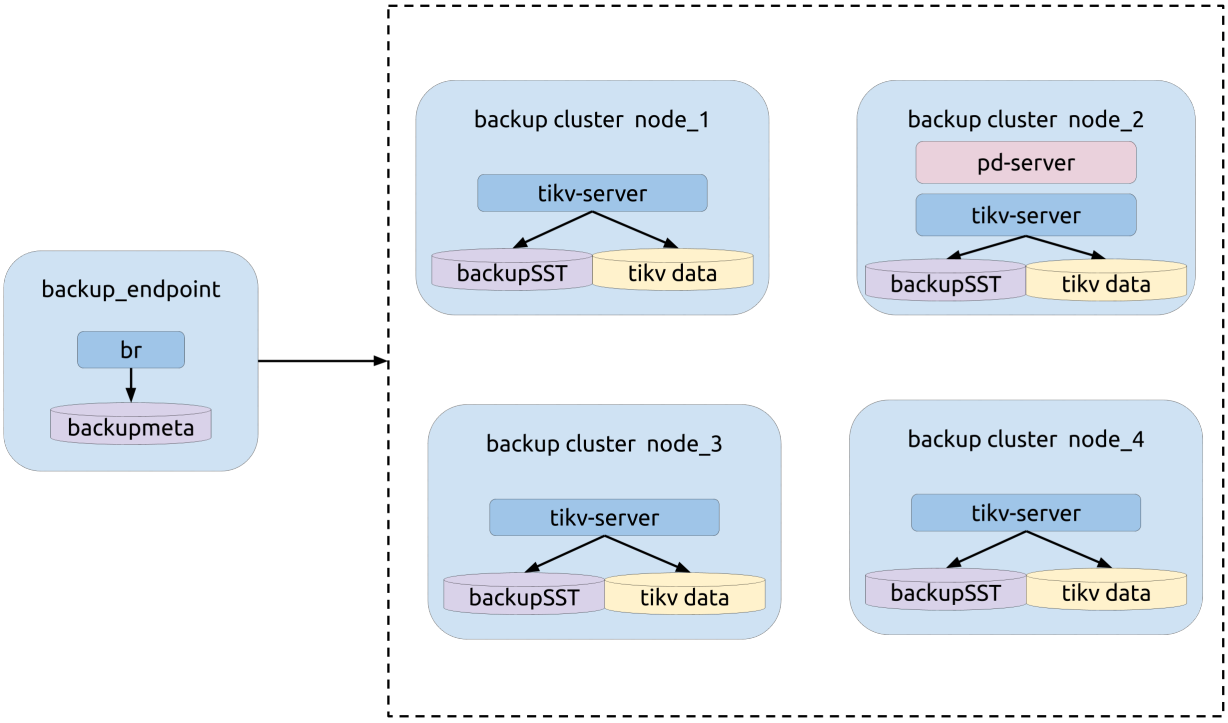


图 315: img

运行备份

备份前在 TiDB 里通过 `admin checksum table order_line` 获得备份的目标表 `--db batchmark --table order_line` 的统计信息。统计信息示例如下：

```
+----------+------------+--------------------+-----------+-------------+
| Db_name  | Table_name | Checksum_crc64_xor | Total_kvs | Total_bytes |
+----------+------------+--------------------+-----------+-------------+
| batchmark | order_line | 10912722838344822475 | 5659888624 | 370385538778 |
+----------+------------+--------------------+-----------+-------------+
1 row in set (5 min 47.59 sec)
```

图 316: img

运行 `br backup` 命令：

```
bin/br backup table \
    --db batchmark \
    --table order_line \
    -s local:///home/tidb/backup_local/ \
    --pd ${PD_ADDR}:2379 \
    --log-file backup_local.log
```

运行备份时，参考备份过程中的运行指标对相关指标进行监控，以了解备份状态。

结果解读

使用 BR 前已设置日志的存放路径。从该路径下存放的日志获取此次备份的相关统计信息。在日志中搜关键字 “summary”，可以看到以下信息：

```
["Table backup summary: total backup ranges: 4, total success: 4, total failed: 0, total take(s):
    ↪  551.31, total kv: 5659888624, total size(MB): 353227.18, avg speed(MB/s): 640.71"] ["
    ↪ backup total regions"=6795] ["backup checksum"=6m33.962719217s] ["backup fast checksum
    ↪ "=22.995552ms]
```

以上日志信息中包含以下内容：

- 备份耗时：`total take(s): 551.31`
- 数据大小：`total size(MB): 353227.18`
- 备份吞吐：`avg speed(MB/s): 640.71`
- 校验耗时：`take=6m33.962719217s`

根据上表数据可以计算得到单个 TiKV 实例的吞吐：`avg speed(MB/s)/tikv_count` = 160。

12.9.3.3.6 从本地磁盘恢复备份数据（推荐测试环境试用）

使用 `br restore` 命令，将一份完整的备份数据恢复到一个离线集群。暂不支持恢复到在线集群。

前置要求

- 恢复前的准备工作。
- 集群中没有与备份数据相同的库表。目前 BR 不支持 table route。
- 集群中各个 TiKV 节点有单独的磁盘用来存放要恢复的 backupSST 数据。
- `restore_endpoint` 节点有单独的磁盘用来存放要恢复的 `backupmeta` 文件。
- 集群中 TiKV 和 `restore_endpoint` 节点需要有相同的备份目录，例如 `/home/tidb/backup_local/`。

如果备份数据存放在本地磁盘，那么需要执行以下的步骤：

1. 汇总所有 backupSST 文件到一个统一的目录下。
2. 将汇总后的 backupSST 文件到复制到集群的所有 TiKV 节点下。
3. 将 `backupmeta` 文件复制到到 `restore endpoint` 节点下。

部署拓扑

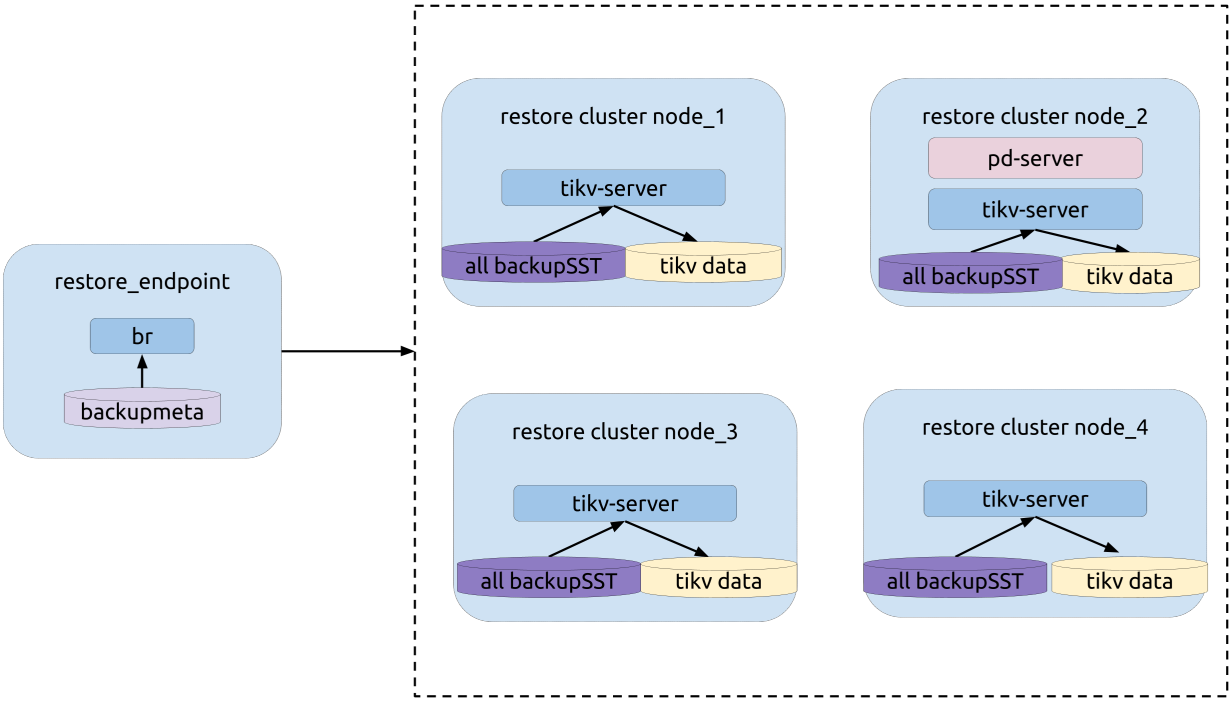


图 317: img

运行恢复

运行 `br restore` 命令：

```
bin/br restore table --db batchmark --table order_line -s local:///home/tidb/backup_local/ --pd
    ↪ 172.16.5.198:2379 --log-file restore_local.log
```

运行恢复时，参考恢复过程中的运行指标对相关指标进行监控，以了解恢复状态。

结果解读

使用 BR 前已设置日志的存放路径。从该日志中可以获取此次恢复的相关统计信息。在日志中搜关键字“summary”，可以看到以下信息：

```
["Table Restore summary: total restore tables: 1, total success: 1, total failed: 0, total take(s
    ↪ ): 908.42, total kv: 5659888624, total size(MB): 353227.18, avg speed(MB/s): 388.84"] ["
    ↪ restore files"=9263] ["restore ranges"=6888] ["split region"=58.7885518s] ["restore
    ↪ checksum"=6m19.349067937s]
```

以上日志信息中包含以下内容：

- 恢复耗时：`total take(s): 908.42`
- 数据大小：`total size(MB): 353227.18`
- 恢复吞吐：`avg speed(MB/s): 388.84`
- Region Split 耗时：`take=58.7885518s`

- 校验耗时：take=6m19.349067937s

根据上表数据可以计算得到：

- 单个 TiKV 实例的吞吐：avg speed(MB/s)/tikv_count = 97.2
- 单个 TiKV 实例的平均恢复速度：total size(MB)/(split time + restore time)/tikv_count = 92.4

12.9.3.4 备份过程中的异常处理

本节介绍如何处理备份过程中出现的常见错误。

12.9.3.4.1 备份日志中出现 key locked Error

日志中的错误消息：log - ["backup occur kv error"][error="{\"KvError\":{\"locked\":

如果在备份过程中遇到 key 被锁住，目前 BR 会尝试清锁。少量报错不会影响备份的正确性。

12.9.3.4.2 备份失败

日志中的错误消息：log - Error: msg:"Io(Custom { kind: AlreadyExists, error: \"[5_5359_42_123_default
↪ .sst] is already exists in /dir/backup_local/\" })"

若备份失败并出现以上错误消息，采取以下其中一种操作后再重新备份：

- 更换备份数据目录。例如将 /dir/backup-2020-01-01/ 改为 /dir/backup_local/。
- 删除所有 TiKV 和 BR 节点的备份目录。

12.9.4 BR 特性

12.9.4.1 BR 自动调节从 v5.4 版本开始引入

在 TiDB v5.4.0 之前，默认情况下，使用 BR 工具进行备份任务时使用的线程数量占总逻辑 CPU 数量的 75%。在没有限速的前提下，备份会消耗大量的集群资源，这会对在线集群的性能造成相当大的影响。虽然你可以通过调节线程池的大小的方式来减少备份对集群性能的影响，但观察负载、手动调节线程池大小也是一件繁琐的事情。

为了减少备份任务对在线集群的影响，从 TiDB v5.4.0 起，BR 引入了自动调节功能，此功能会默认开启。在集群资源占用率较高的情况下，BR 可以通过该功能自动限制备份使用的资源，从而减少对集群的影响。

12.9.4.1.1 使用场景

如果你希望减少 BR 备份对集群的影响，那么，你可以开启自动调节功能。开启该功能后，BR 会在不过度影响集群的前提下，以最快的速度进行备份。

或者，你也可以使用 TiKV 配置项backup.num-threads 或参数 --ratelimit 进行备份限速。

12.9.4.1.2 使用方法

自动调节功能默认打开，无需额外配置。

> **注意：**
>
> v5.3.x 版本的集群，在升级到 v5.4.0 及以上版本后，自动调节功能默认关闭，需手动开启。

如需开启 BR 自动调节功能，可以通过把 TiKV 配置项`backup.enable-auto-tune` 设置为 `true` 的方式来完成。

TiKV 支持动态配置自动调节功能，因此，在开启或关闭该功能时，无需重启集群。你可以运行以下命令动态启动或停止 BR 自动调节功能：

```
tikv-ctl modify-tikv-config -n backup.enable-auto-tune -v <true|false>
```

在离线备份场景中，你也可以使用 `tikv-ctl` 把 `backup.num-threads` 修改为更大的数字，从而提升备份速度。

12.9.4.1.3 使用限制

自动调节是一个粗粒度的限流方案，它的优势在无需手动调节。但是，由于调节的粒度不够精确，该功能有可能无法彻底移除备份对集群的影响。

该功能的已知问题及其解决方案如下：

- 问题 1：对于以写负载为主的集群，自动调节可能会让工作负载和备份进入一种“正反馈循环”：备份会占用较多资源，导致工作负载使用的资源变少。此时，自动调节会误以为资源使用率下降，从而让 BR 运行得更加激进。在这种情况下，自动调节实际上失效。
  - 解决方法：手动调节 `backup.num-threads`，限制处理备份的工作线程数量。具体原理如下：
    目前，备份过程会涉及大量的 SST 解码、编码、压缩、解压，而此过程会需要消耗大量的 CPU 资源。另外，以往的测试证明，备份过程中，用于备份的线程池的 CPU 利用率接近 100%。也就是说，备份任务会占用大量 CPU 资源。通过调整备份任务使用的线程数量，TiKV 可以控制备份任务使用的 CPU 核心数，从而减少其任务对集群性能的影响。
- 问题 2：对于存在热点的集群，产生热点的 TiKV 节点可能会被过度限流，从而拉慢备份的整体进度。
  - 解决方法：消除热点节点，或者在热点节点上关闭自动调节（关闭此功能可能会导致集群性能降低）。
- 问题 3：对于流量抖动非常大的场景，由于自动调节每隔`auto-tune-refresh-interval`（默认为一分钟）才会计算出新的限流，所以可能无法很好地应对流量抖动非常厉害的场景。
  - 解决方法：关闭自动调节。

12.9.4.1.4 实现原理

自动调节会通过调节 BR 备份时使用的工作线程池的大小，保证集群的 CPU 总体使用率不超过某个特定的值。

这个特性还有两个配置项未在 TiKV 文档中列出，仅在内部调试使用，正常备份时无需配置这两个参数。

- backup.auto-tune-remain-threads：
  - 通过控制备份任务占用的资源，自动调节会保证该节点中至少有该数量的核心会保持空闲的状态。
  - 默认值：round(0.2 * vCPU)
- backup.auto-tune-refresh-interval：
  - 每隔该值的时间段，自动调节会刷新统计信息并重新计算备份任务使用的 CPU 核心数的上限。
  - 默认值：1m

以下是一个使用自动调节功能的示例，其中 * 代表集群中被备份任务占用的 CPU，^ 代表其它任务占用的 CPU，- 代表空闲 CPU。

```
|--------| 系统总共有 8 颗逻辑 CPU。
|****----| 默认配置 `backup.num-threads` 为 `4`。请注意，在任何时候自动调节都不会让线程池大小大于
    ↪ `backup.num-threads`。
|^^****--| 默认配置 `auto-tune-remain-threads` = round(8 * 0.2) = 2。
    ↪ 自动调节会将备份任务的线程池大小调节至 `4`。
|^^^^**--| 由于集群的工作负载加重，自动调节将备份任务的线程池大小调节至 `2`。调节后，集群中仍有 2
    ↪ 个 CPU 核心数保持空闲。
```

在监控面板的 “Backup CPU Utilization” 中，可以看到自动限流目前选择的线程池的大小：

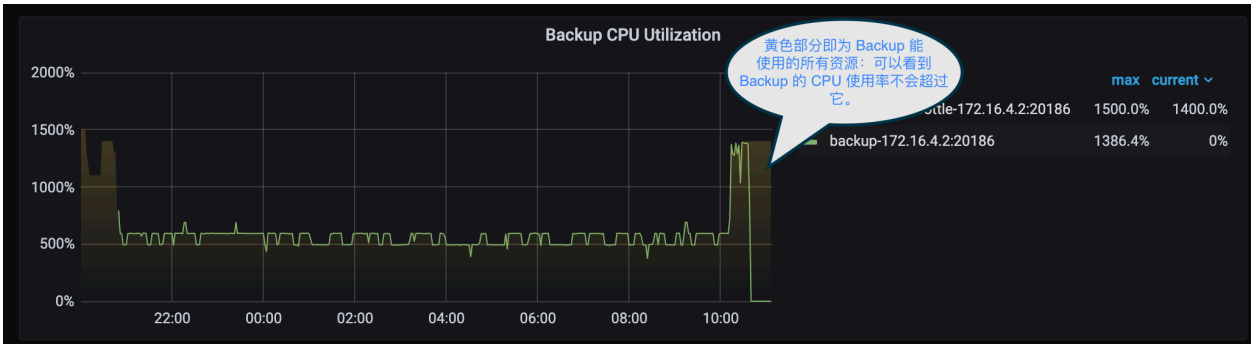


图 318: Grafana dashboard example of backup auto-tune metrics

图片中，黄色半透明的填充部分为开启自动调节后备份任务可用的线程，即备份任务能使用的所有资源。从中可以看到备份任务的 CPU 使用率不会超过黄色部分。

12.9.5 Backup & Restore 常见问题

本文列出了在使用 Backup & Restore (BR) 时，可能会遇到的问题及相应的解决方法。

如果遇到未包含在此文档且无法解决的问题，可以在 AskTUG 社区中提问。

12.9.5.1 在 TiDB v5.4.0 及后续版本中，当在有负载的集群进行备份时，备份速度为什么会变得很慢？

从 TiDB v5.4.0 起，TiKV 的备份新增了自动调节功能。对于 v5.4.0 及以上版本，该功能会默认开启。当集群负载较高时，该功能会自动限制备份任务使用的资源，从而减少备份对在线集群的性能造成的影响。如需了解关于自动调节功能的更多信息，请参见自动调节。

TiKV 支持动态配置自动调节功能。因此，在开启或关闭该功能时，你不需要重启集群。以下为该功能的具体使用方法：

- 关闭功能：把 TiKV 配置项backup.enable-auto-tune 设置为 false。
- 开启功能：把 backup.enable-auto-tune 设置为 true。对于 v5.3.x 版本的集群，当 TiDB 升级到 v5.4.0 及以上版本后，自动调节功能会默认关闭。这时，你可以通过该方式手动开启此功能。

如需了解通过 tikv-ctl 在线修改自动调节功能的命令行，请参阅自动调节功能的使用方法。

另外，该功能也降低了进行备份任务时默认使用的工作线程数量（详见backup.num-threads）。因此，你通过 Grafana 监控面板看到的备份速度、CPU 使用率、I/O 资源利用率都会小于 v5.4 之前的版本。在 v5.4 之前的版本中，backup.num-threads 的默认值为 CPU * 0.75，即处理备份任务的工作线程数量占了 75% 的逻辑 CPU，最大值为 32；在 v5.4 及之后的版本中，该配置项的默认值为 CPU * 0.5，最大值为 8。

在离线备份场景中，你也可以使用 tikv-ctl 把 backup.num-threads 修改为更大的数字，从而提升备份任务的速度。

12.9.5.2 恢复的时候，报错 could not read local://...:download sst failed，该如何处理？

在恢复的时候，每个节点都必须能够访问到所有的备份文件 (SST files)，默认情况下，假如使用 local storage，备份文件会分散在各个节点中，此时是无法直接恢复的，必须将每个 TiKV 节点的备份文件拷贝到其它所有 TiKV 节点才能恢复。

建议在备份的时候挂载一块 NFS 网盘作为备份盘，详见将单表数据备份到网络盘。

12.9.5.3 BR 备份时，对集群影响多大？

对于 TiDB v5.4.0 及以后的版本，BR 不仅调低了备份任务会使用的默认 CPU 使用率，还引入了自动调节功能。开启该功能后，当在集群高负载的场景下进行备份时，BR 能够自动限制备份任务使用的资源，从而限制 BR 备份的速度。因此，在 v5.4.0 高负载集群中使用默认配置进行备份时，备份任务对其集群性能的影响会显著小于 v5.4.0 之前的版本。有关自动调节功能的具体内容，请参阅BR 自动调节。

以下为针对内部的一个单节点进行的内部测试。通过测试结果，可以确定在全速备份场景下使用 v5.4.0 版本及其前期版本的默认配置时，BR 备份对集群性能的影响有较大区别。具体测试结果如下：

- 使用 v5.3.0 的默认配置时，纯写负载的 QPS 降低了 75%；
- 使用 v5.4.0 的默认配置时，相同负载的 QPS 降低了 25%。不过，使用该配置时，BR 备份任务的耗时也相应得拉慢，其所需时间为 v5.3.0 配置下耗时的 1.7 倍。

如果需要手动控制 BR 备份对集群性能带来的影响，可以通过以下方案实现。这两种方案在减少 BR 备份对集群的影响的同时，也会降低备份任务的速度。

- 使用 --ratelimit 参数对备份任务进行限速。请注意，这个参数限制的是把备份文件存储到外部存储的速度。计算备份文件的总大小的时候，请以备份日志中的 backup data size(after compressed) 为基准。

- 调节 TiKV 配置项backup.num-threads，限制备份任务使用的资源。该配置项决定处理备份任务的工作线程数量。过往的测试数据表明，当 BR 备份的线程数量不大于 8、集群总 CPU 利用率不超过 60% 时，BR 备份任务对集群（无论读写负载）几乎没有影响。

12.9.5.4 BR 会备份系统表吗？在数据恢复的时候，这些系统表会冲突吗？

在 v5.1.0 之前，BR 备份时会过滤掉系统库 `mysql.*` 的表数据。自 v5.1.0 起，BR 默认备份集群内的全部数据，包括系统库 `mysql.*` 中的数据。

恢复数据时，系统表不会冲突。由于恢复 `mysql.*` 中系统表数据的技术实现尚不完善，因此 BR 默认不恢复系统库 `mysql` 中的表数据，也就不会产生任何冲突。详情参阅备份和恢复 `mysql` 系统库下的表数据（实验特性）。

12.9.5.5 BR 遇到 Permission denied 或者 No such file or directory 错误，即使用 root 运行 BR 也无法解决，该如何处理？

需要确认 TiKV 是否有访问备份目录的权限。如果是备份，确认是否有写权限；如果是恢复，确认是否有读权限。

在进行备份操作时，如果使用本地磁盘或 NFS 作为存储介质，请确保执行 BR 的用户和启动 TiKV 的用户相同（如果 BR 和 TiKV 位于不同的机器，则需要用户的 UID 相同），否则备份可能会出现该问题。。

使用 root 运行 BR 仍旧有可能会因为磁盘权限而失败，因为备份文件 (SST) 的保存是由 TiKV 执行的。

> **注意：**
>
> 在恢复的时候也可能遇到同样的问题。
>
> 使用 BR 进行数据的恢复时，检验读权限的时机是在第一次读取 SST 文件时，考虑到执行 DDL 的耗时，这个时刻可能会离开始运行 BR 的时间很远。这样可能会出现等了很长时间之后遇到 Permission denied 错误失败的情况。
>
> 因此，最好在恢复前提前检查权限。

你可以按照如下步骤进行权限检查：

1. 执行 Linux 原生的进程查询命令

```
ps aux | grep tikv-server
```

命令输出示例如下:

```
tidb_ouo  9235 10.9  3.8 2019248 622776 ?      Ssl  08:28   1:12 bin/tikv-server --addr
    ↪ 0.0.0.0:20162 --advertise-addr 172.16.6.118:20162 --status-addr 0.0.0.0:20188 --
    ↪ advertise-status-addr 172.16.6.118:20188 --pd 172.16.6.118:2379 --data-dir /home/
    ↪ user1/tidb-data/tikv-20162 --config conf/tikv.toml --log-file /home/user1/tidb-
    ↪ deploy/tikv-20162/log/tikv.log
```

```
tidb_ouo  9236  9.8  3.8 2048940 631136 ?      Ssl  08:28   1:05 bin/tikv-server --addr
    ↪ 0.0.0.0:20161 --advertise-addr 172.16.6.118:20161 --status-addr 0.0.0.0:20189 --
    ↪ advertise-status-addr 172.16.6.118:20189 --pd 172.16.6.118:2379 --data-dir /home/
    ↪ user1/tidb-data/tikv-20161 --config conf/tikv.toml --log-file /home/user1/tidb-
    ↪ deploy/tikv-20161/log/tikv.log
```

或者执行以下命令：

```
ps aux | grep tikv-server | awk '{print $1}'
```

命令输出示例如下:

```
tidb_ouo
tidb_ouo
```

2. 使用 TiUP 命令查询集群的启动信息

```
tiup cluster list
```

命令输出示例如下：

```
[root@Copy-of-VM-EE-CentOS76-v1 br]# tiup cluster list
Starting component `cluster`: /root/.tiup/components/cluster/v1.5.2/tiup-cluster list
Name          User      Version  Path
    ↪ PrivateKey
----          ----      -------  ----
    ↪ ----------
tidb_cluster  tidb_ouo  v5.0.2   /root/.tiup/storage/cluster/clusters/tidb_cluster  /root/.
    ↪ tiup/storage/cluster/clusters/tidb_cluster/ssh/id_rsa
```

3. 检查备份目录的权限，例如 backup 目录是备份数据存储目录。命令示例如下：

```
ls -al backup
```

命令输出示例如下：

```
[root@Copy-of-VM-EE-CentOS76-v1 user1]# ls -al backup
total 0
drwxr-xr-x  2 root root   6 Jun 28 17:48 .
drwxr-xr-x 11 root root 310 Jul  4 10:35 ..
```

由以上命令输出结果可知，tikv-server 实例由用户 tidb_ouo 启动，但用户账号 tidb_ouo 没有 backup 目录的写入权限，所以备份失败。

12.9.5.6 BR 遇到错误信息 Io(Os...)，该如何处理？

这类问题几乎都是 TiKV 在写盘的时候遇到的系统调用错误。例如遇到 Io(Os { code: 13, kind: ↪ PermissionDenied...}) 或者 Io(Os { code: 2, kind: NotFound...}) 这类错误信息，首先检查备份目录的挂载方式和文件系统，试试看备份到其它文件夹或者其它硬盘。

目前已知备份到 samba 搭建的网盘时可能会遇到 Code: 22(invalid argument) 错误。

12.9.5.7 BR 遇到错误信息 `rpc error: code = Unavailable desc =...`，该如何处理？

该问题一般是因为使用 BR 恢复数据的时候，恢复集群的性能不足导致的。可以从恢复集群的监控或者 TiKV 日志来辅助确认。

要解决这类问题，可以尝试扩大集群资源，以及调小恢复时的并发度 (concurrency)，打开限速 (ratelimit) 设置。

12.9.5.8 使用 local storage 的时候，BR 备份的文件会存在哪里？

在使用 local storage 的时候，会在运行 BR 的节点生成 `backupmeta`，在各个 Region 的 Leader 节点生成备份文件。

12.9.5.9 备份数据有多大，备份会有副本吗？

备份的时候仅仅在每个 Region 的 Leader 处生成该 Region 的备份文件。因此备份的大小等于数据大小，不会有多余的副本数据。所以最终的总大小大约是 TiKV 数据总量除以副本数。

但是假如想要从本地恢复数据，因为每个 TiKV 都必须要能访问到所有备份文件，在最终恢复的时候会有等同于恢复时 TiKV 节点数量的副本。

12.9.5.10 BR 恢复到 TiCDC / Drainer 的上游集群时，要注意些什么？

- BR 恢复的数据无法被同步到下游，因为 BR 直接导入 SST 文件，而下游集群目前没有办法获得上游的 SST 文件。
- 在 4.0.3 版本之前，BR 恢复时产生的 DDL jobs 还可能会让 TiCDC / Drainer 执行异常的 DDL。所以，如果一定要在 TiCDC / Drainer 的上游集群执行恢复，请将 BR 恢复的所有表加入 TiCDC / Drainer 的阻止名单。

TiCDC 可以通过配置项中的 `filter.rules` 项完成，Drainer 则可以通过`syncer.ignore-table` 完成。

12.9.5.11 BR 会备份表的 SHARD_ROW_ID_BITS 和 PRE_SPLIT_REGIONS 信息吗？恢复出来的表会有多个 Region 吗？

会的，BR 会备份表的`SHARD_ROW_ID_BITS 和 PRE_SPLIT_REGIONS` 信息，并恢复成多个 Region。

12.9.5.12 使用 BR 恢复备份数据后，SQL 查询报错 `region is unavailable`

如果 BR 备份的集群有 TiFlash，恢复时会将 TiFlash 信息存进 `TableInfo`。此时如果恢复的集群没有 TiFlash，则会报该错误。计划在未来版本中修复该错误。

12.9.5.13 BR 是否支持就地 (in-place) 全量恢复某个历史备份？

不支持。

12.9.5.14 在 Kubernetes 环境中如何使用 BR 进行增量备份？

可以使用 kubectl 执行 `kubectl -n ${namespace} get bk ${name}` 以获得上次 BR 备份 `commitTs` 字段，该字段的内容可作为 `--lastbackupts` 使用。

12.9.5.15 BR backupTS 如何转化成 Unix 时间？

BR backupTS 默认是在备份开始前，从 PD 获取到的最新时间戳。可以使用 `pd-ctl tso timestamp` 来解析该时间戳，以获得精确值，也可以通过 `backupTS >> 18` 来快速获取估计值。

12.9.5.16 BR 恢复存档后是否需要对表执行 ANALYZE 以更新 TiDB 在表和索引上留下的统计信息？

BR 不会备份统计信息（v4.0.9 除外）。所以在恢复存档后需要手动执行 `ANALYZE TABLE` 或等待 TiDB 自动进行 ANALYZE。

BR v4.0.9 备份统计信息使 BR 消耗过多内存，为保证备份过程正常，从 v4.0.10 开始默认关闭备份统计信息的功能。

如果不对表执行 ANALYZE，TiDB 会因统计信息不准确而选不中最优化的执行计划。如果查询性能不是重点关注项，可以忽略 ANALYZE。

12.9.5.17 是否可以同时使用多个 BR 进程对单个集群进行恢复？

强烈不建议在单个集群中同时使用多个 BR 进程进行恢复，原因如下：

- BR 在恢复数据时，会修改 PD 的一些全局配置。如果同时使用多个 BR 进程进行恢复，这些配置可能会被错误地覆写，导致集群状态异常。
- 因为 BR 在恢复数据的时候会占用大量集群资源，事实上并行恢复能获得的速度提升也非常有限。
- 多个 BR 并行恢复的场景没有经过测试，无法保证成功。

12.10 TiDB Binlog

12.10.1 TiDB Binlog 简介

TiDB Binlog 是一个用于收集 TiDB 的 binlog，并提供准实时备份和同步功能的商业工具。

TiDB Binlog 支持以下功能场景：

- 数据同步：同步 TiDB 集群数据到其他数据库
- 实时备份和恢复：备份 TiDB 集群数据，同时可以用于 TiDB 集群故障时恢复

> **注意：**
>
> TiDB Binlog 与 TiDB v5.0 版本开始引入的一些特性不兼容，无法一起使用，详情参照注意事项。
> 建议使用TiCDC 替代 TiDB Binlog。

要快速了解 Binlog 的基本原理和使用方法，建议先观看下面的培训视频（时长 28 分钟）。注意本视频只为学习参考，具体操作步骤和最新功能，请以文档内容为准。

12.10.1.1 TiDB Binlog 整体架构

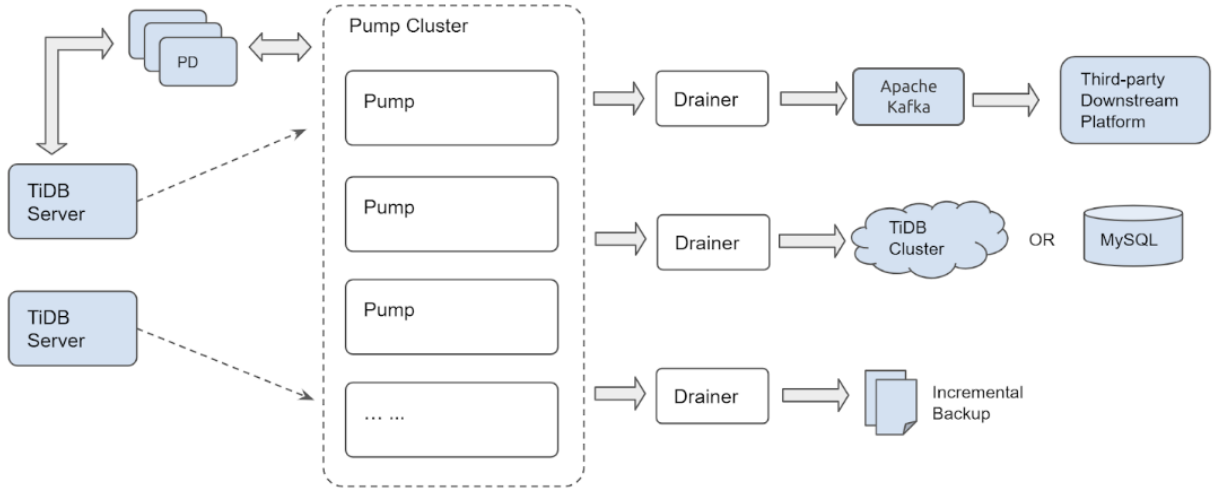


图 319: TiDB Binlog 架构

TiDB Binlog 集群主要分为 Pump 和 Drainer 两个组件，以及 binlogctl 工具：

12.10.1.1.1 Pump

Pump 用于实时记录 TiDB 产生的 Binlog，并将 Binlog 按照事务的提交时间进行排序，再提供给 Drainer 进行消费。

12.10.1.1.2 Drainer

Drainer 从各个 Pump 中收集 Binlog 进行归并，再将 Binlog 转化成 SQL 或者指定格式的数据，最终同步到下游。

12.10.1.1.3 binlogctl 工具

`binlogctl` 是一个 TiDB Binlog 配套的运维工具，具有如下功能：

- 获取 TiDB 集群当前的 TSO
- 查看 Pump/Drainer 状态
- 修改 Pump/Drainer 状态
- 暂停/下线 Pump/Drainer

12.10.1.2 主要特性

- 多个 Pump 形成一个集群，可以水平扩容。
- TiDB 通过内置的 Pump Client 将 Binlog 分发到各个 Pump。
- Pump 负责存储 Binlog，并将 Binlog 按顺序提供给 Drainer。
- Drainer 负责读取各个 Pump 的 Binlog，归并排序后发送到下游。
- Drainer 支持relay log 功能，通过 relay log 保证下游集群的一致性状态。

12.10.1.3 注意事项

- TiDB Binlog 和 TiDB 在 v5.1 版本中解决了 v5.0 版本中引入的聚簇索引与 TiDB Binlog 不兼容问题。在升级 TiDB Binlog 和 TiDB Server 到 v5.1 版本后：开启 TiDB Binlog 后，TiDB 支持创建聚簇索引表；聚簇索引表的数据插入、删除和更新操作支持通过 TiDB Binlog 同步到下游。对于同步聚簇索引表时需注意：
  - 如果从 v5.0 版本手动控制组件升级顺序进行升级，请确保先将 TiDB Binlog 升级至 v5.1 版本后再将 TiDB Server 升级至 v5.1 版本。
  - 推荐将上下游的 TiDB 系统变量`tidb_enable_clustered_index`配置为一致的值来保证上下游 TiDB 聚簇索引表结构一致。
- TiDB Binlog 与 TiDB v5.0 版本开始引入的以下特性不兼容，无法一起使用：
  - TiDB 聚簇索引特性：开启 TiDB Binlog 后 TiDB 不允许创建非单个整数列作为主键的聚簇索引；已创建的聚簇索引表的数据插入、删除和更新动作不会通过 TiDB Binlog 同步到下游。如需同步聚簇索引表，请升级至 v5.1 版本或使用TiCDC；
  - TiDB 系统变量 tidb_enable_async_commit：启用 TiDB Binlog 后，开启该选项无法获得性能提升。要获得性能提升，建议使用TiCDC 替代 TiDB Binlog。
  - TiDB 系统变量 tidb_enable_1pc：启用 TiDB Binlog 后，开启该选项无法获得性能提升。要获得性能提升，建议使用TiCDC 替代 TiDB Binlog。
- TiDB Binlog 与 TiDB v4.0.7 版本开始引入的以下特性不兼容，无法一起使用：
  - TiDB 系统变量 tidb_enable_amend_pessimistic_txn：两个功能存在兼容性问题，一起使用会造成 TiDB Binlog 复制数据不一致的正确性问题。
- Drainer 支持将 Binlog 同步到 MySQL、TiDB、Kafka 或者本地文件。如果需要将 Binlog 同步到其他 Drainer 不支持的类型的系统中，可以设置 Drainer 将 Binlog 同步到 Kafka，然后根据 binlog consumer protocol 进行定制处理，参考Binlog Consumer Client 用户文档。
- 如果 TiDB Binlog 用于增量恢复，可以设置配置项 `db-type="file"`，Drainer 会将 binlog 转化为指定的 proto buffer 格式的数据，再写入到本地文件中。这样就可以使用Reparo 恢复增量数据。

  关于 `db-type` 的取值，应注意：
  - 如果 TiDB 版本 < 2.1.9，则 `db-type="pb"`。
  - 如果 TiDB 版本 > = 2.1.9，则 `db-type="file"` 或 `db-type="pb"`。
- 如果下游为 MySQL/TiDB，数据同步后可以使用sync-diff-inspector 进行数据校验。

12.10.2 TiDB Binlog 教程

本文档主要介绍如何使用 TiDB Binlog 将数据从 TiDB 推送到 MariaDB 实例。文中的 TiDB Binlog 集群包含 Pump 和 Drainer 的单个节点，TiDB 集群包含 TiDB、TiKV 和 Placement Driver (PD) 各组件的单个节点。

希望上手实践 TiDB Binlog 工具的用户需要对TiDB 架构有一定的了解，最好有创建过 TiDB 集群的经验。该文档也有助于简单快速了解 TiDB Binlog 架构以及相关概念。

> **警告：**
>
> 该文档中部署 TiDB 的操作指导不适用于在生产或研发环境中部署 TiDB 的情况。

该文档假设用户使用的是现代 Linux 发行版本中的 x86-64。示例中使用的是 VMware 中运行的 CentOS 7 最小化安装。建议在一开始就进行清洁安装，以避免受现有环境中未知情况的影响。如果不想使用本地虚拟环境，也可以使用云服务启动 CentOS 7 VM。

12.10.2.1 TiDB Binlog 简介

TiDB Binlog 用于收集 TiDB 中二进制日志数据、提供实时数据备份和同步以及将 TiDB 集群的数据增量同步到下游。

TiDB Binlog 支持以下功能场景：

- 增量备份，将 TiDB 集群中的数据增量同步到另一个集群，或通过 Kafka 增量同步到选择的下游。
- 当使用 TiDB DM (Data Migration) 将数据从上游 MySQL 或者 MariaDB 迁移到 TiDB 集群时，可使用 TiDB Binlog 保持 TiDB 集群与其一个独立下游 MySQL 或 MariaDB 实例或集群同步。当 TiDB 集群上游数据迁移过程中出现问题，下游数据同步过程中可使用 TiDB Binlog 恢复数据到原先的状态。

更多信息参考TiDB Binlog Cluster 版本用户文档。

12.10.2.2 架构

TiDB Binlog 集群由 Pump 和 Drainer 两个组件组成。一个 Pump 集群中有若干个 Pump 节点。TiDB 实例连接到各个 Pump 节点并发送 binlog 数据到 Pump 节点。Pump 集群连接到 Drainer 节点，Drainer 将接收到的更新数据转换到某个特定下游（例如 Kafka、另一个 TiDB 集群或 MySQL 或 MariaDB Server）指定的正确格式。

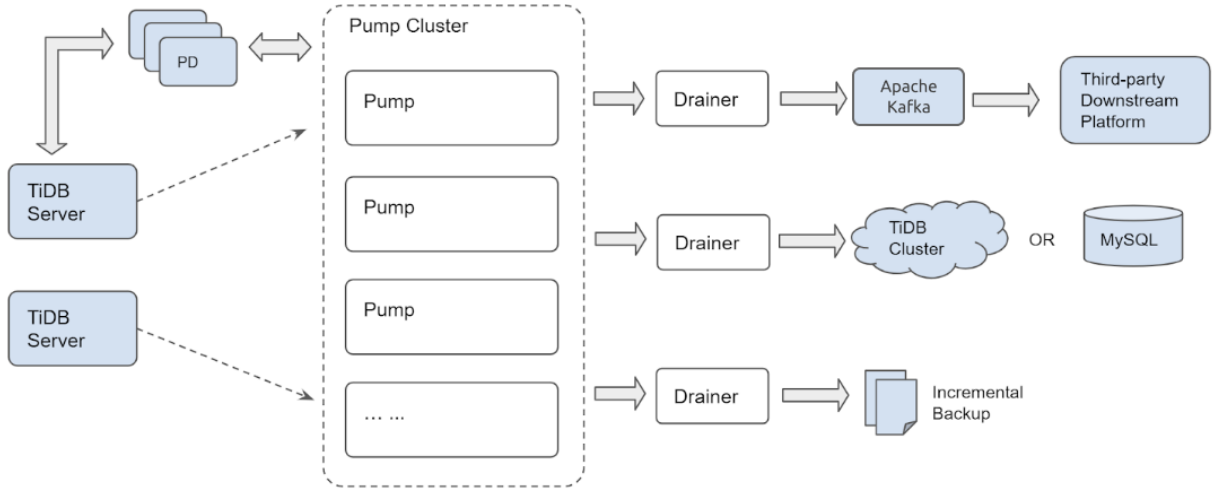


图 320: TiDB Binlog architecture

Pump 的集群架构能确保 TiDB 或 Pump 集群中有新的实例加入或退出时更新数据不会丢失。

12.10.2.3 安装

由于 RHEL/CentOS 7 的默认包装库中包括 MariaDB Server，本示例选择的是 MariaDB Server。后续除了安装服务器，也需要安装客户端。安装指令如下：

```
sudo yum install -y mariadb-server
```

```
curl -LO https://download.pingcap.org/tidb-latest-linux-amd64.tar.gz | tar xzf - &&
cd tidb-latest-linux-amd64
```

预期输出：

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  368M  100  368M    0     0  8394k      0  0:00:44  0:00:44 --:--:-- 11.1M
```

12.10.2.4 配置

通过执行以下步骤配置一个 TiDB 集群，该集群包括 pd-server、tikv-server 和 tidb-server 各组件的单个实例。

1. 填充配置文件：

```
printf > pd.toml %s\\n 'log-file="pd.log"' 'data-dir="pd.data"' &&
printf > tikv.toml %s\\n 'log-file="tikv.log"' '[storage]' 'data-dir="tikv.data"' '[pd]' '
    ↪ endpoints=["127.0.0.1:2379"]' '[rocksdb]' max-open-files=1024 '[raftdb]' max-open-
    ↪ files=1024 &&
printf > pump.toml %s\\n 'log-file="pump.log"' 'data-dir="pump.data"' 'addr="127.0.0.1:8250"
    ↪ ' 'advertise-addr="127.0.0.1:8250"' 'pd-urls="http://127.0.0.1:2379"' &&
printf > tidb.toml %s\\n 'store="tikv"' 'path="127.0.0.1:2379"' '[log.file]' 'filename="tidb
    ↪ .log"' '[binlog]' 'enable=true' &&
printf > drainer.toml %s\\n 'log-file="drainer.log"' '[syncer]' 'db-type="mysql"' '[syncer.
    ↪ to]' 'host="127.0.0.1"' 'user="root"' 'password=""' 'port=3306'
```

2. 查看配置细节：

```
for f in *.toml; do echo "$f:"; cat "$f"; echo; done
```

预期输出：

```
drainer.toml:
log-file="drainer.log"
[syncer]
db-type="mysql"
[syncer.to]
```

```
host="127.0.0.1"
user="root"
password=""
port=3306

pd.toml:
log-file="pd.log"
data-dir="pd.data"

pump.toml:
log-file="pump.log"
data-dir="pump.data"
addr="127.0.0.1:8250"
advertise-addr="127.0.0.1:8250"
pd-urls="http://127.0.0.1:2379"

tidb.toml:
store="tikv"
path="127.0.0.1:2379"
[log.file]
filename="tidb.log"
[binlog]
enable=true

tikv.toml:
log-file="tikv.log"
[storage]
data-dir="tikv.data"
[pd]
endpoints=["127.0.0.1:2379"]
[rocksdb]
max-open-files=1024
[raftdb]
max-open-files=1024
```

12.10.2.5 启动程序

现在可启动各个组件。推荐启动顺序依次为 Placement Driver (PD)、 TiKV、 Pump（ TiDB 发送 binlog 日志必须连接 Pump 服务 ）、 TiDB。

1. 启动所有服务：

```
./bin/pd-server --config=pd.toml &>pd.out &
```

```
[1] 20935
```

```
./bin/tikv-server --config=tikv.toml &>tikv.out &
```

```
[2] 20944
```

```
./bin/pump --config=pump.toml &>pump.out &
```

```
[3] 21050
```

```
sleep 3 &&
./bin/tidb-server --config=tidb.toml &>tidb.out &
```

```
[4] 21058
```

2. 如果执行 jobs，可以看到后台正在运行的程序，列表如下：

```
jobs
```

```
[1]   Running                 ./bin/pd-server --config=pd.toml &>pd.out &
[2]   Running                 ./bin/tikv-server --config=tikv.toml &>tikv.out &
[3]-  Running                 ./bin/pump --config=pump.toml &>pump.out &
[4]+  Running                 ./bin/tidb-server --config=tidb.toml &>tidb.out &
```

如果有服务启动失败（例如出现“Exit 1”而不是“Running”），尝试重启单个组件。

12.10.2.6 连接

按以上步骤操作后，TiDB 的 4 个组件开始运行。接下来可以使用以下 MariaDB 或 MySQL 命令行客户端，通过 4000 端口连接到 TiDB 服务：

```
mysql -h 127.0.0.1 -P 4000 -u root -e 'select tidb_version();'
```

预期输出：

```
*************************** 1. row ***************************
tidb_version(): Release Version: v3.0.0-beta.1-154-gd5afff70c
Git Commit Hash: d5afff70cdd825d5fab125c8e52e686cc5fb9a6e
Git Branch: master
UTC Build Time: 2019-04-24 03:10:00
GoVersion: go version go1.12 linux/amd64
Race Enabled: false
TiKV Min Version: 2.1.0-alpha.1-ff3dd160846b7d1aed9079c389fc188f7f5ea13e
Check Table Before Drop: false
```

连接后 TiDB 集群已开始运行，pump 读取集群中的 binlog 数据，并在其数据目录中将 binlog 数据存储为 relay log。下一步是启动一个可供 drainer 写入的 MariaDB Server。

1. 启动 drainer：

```
sudo systemctl start mariadb &&
./bin/drainer --config=drainer.toml &>drainer.out &
```

如果你的操作系统更易于安装 MySQL，只需保证监听 3306 端口。另外，可使用密码为空的“root”用户连接到 MySQL，或调整 drainer.toml 连接到 MySQL。

```
mysql -h 127.0.0.1 -P 3306 -u root
```

预期输出：

```
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MariaDB connection id is 20
Server version: 5.5.60-MariaDB MariaDB Server

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]>
```

```
show databases;
```

预期输出：

```
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| test               |
| tidb_binlog        |
+--------------------+
5 rows in set (0.01 sec)
```

如下表格是包含 checkpoint 表格的 tidb_binlog 数据库。drainer 使用 checkpoint 表格，记录 TiDB 集群中的 binlog 已经更新到了哪个位置。

```
use tidb_binlog;
```

```
Database changed
```

```
select * from checkpoint;
```

```
+--------------------+--------------------------------------------+
| clusterID          | checkPoint                                 |
+--------------------+--------------------------------------------+
| 6678715361817107733 | {"commitTS":407637466476445697,"ts-map":{}} |
+--------------------+--------------------------------------------+
1 row in set (0.00 sec)
```

打开另一个连接到 TiDB 的客户端，创建一个表格并插入几行数据。建议在 GNU Screen 软件中操作，从而同时打开多个客户端。

```
mysql -h 127.0.0.1 -P 4000 --prompt='TiDB [\d]> ' -u root
```

```
create database tidbtest;
```

```
Query OK, 0 rows affected (0.12 sec)
```

```
use tidbtest;
```

```
Database changed
```

```
create table t1 (id int unsigned not null AUTO_INCREMENT primary key);
```

```
Query OK, 0 rows affected (0.11 sec)
```

```
insert into t1 () values (),(),(),(),();
```

```
Query OK, 5 rows affected (0.01 sec)
Records: 5  Duplicates: 0  Warnings: 0
```

```
select * from t1;
```

```
+----+
| id |
+----+
|  1 |
|  2 |
|  3 |
|  4 |
|  5 |
+----+
5 rows in set (0.00 sec)
```

切换回 MariaDB 客户端可看到新的数据库、新的表格和最近插入的行数据。

```
use tidbtest;
```

```
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
```

```
show tables;
```

```
+--------------------+
| Tables_in_tidbtest |
+--------------------+
| t1                 |
+--------------------+
1 row in set (0.00 sec)
```

```
select * from t1;
```

```
+----+
| id |
+----+
|  1 |
|  2 |
|  3 |
|  4 |
|  5 |
+----+
5 rows in set (0.00 sec)
```

可看到查询 MariaDB 时插入到 TiDB 相同的行数据，表明 TiDB Binlog 安装成功。

12.10.2.7 binlogctl

加入到集群的 Pump 和 Drainer 的数据存储在 Placement Driver (PD) 中。binlogctl 可用于查询和修改状态信息。更多信息请参考binlogctl guide。

使用 `binlogctl` 查看集群中 Pump 和 Drainer 的当前状态：

```
./bin/binlogctl -cmd drainers
```

```
[2019/04/11 17:44:10.861 -04:00] [INFO] [nodes.go:47] ["query node"] [type=drainer] [node="{
    ↪ NodeID: localhost.localdomain:8249, Addr: 192.168.236.128:8249, State: online,
    ↪ MaxCommitTS: 407638907719778305, UpdateTime: 2019-04-11 17:44:10 -0400 EDT}"]
```

```
./bin/binlogctl -cmd pumps
```

```
[2019/04/11 17:44:13.904 -04:00] [INFO] [nodes.go:47] ["query node"] [type=pump] [node="{NodeID:
    ↪ localhost.localdomain:8250, Addr: 192.168.236.128:8250, State: online, MaxCommitTS:
    ↪ 407638914024079361, UpdateTime: 2019-04-11 17:44:13 -0400 EDT}"]
```

如果结束 Drainer 进程，集群会改进程设置 “已暂停（即集群等待 Drainer 重新加入）” 的状态。

```
pkill drainer &&
./bin/binlogctl -cmd drainers
```

预期输出：

```
[2019/04/11 17:44:22.640 -04:00] [INFO] [nodes.go:47] ["query node"] [type=drainer] [node="{
    ↪ NodeID: localhost.localdomain:8249, Addr: 192.168.236.128:8249, State: paused,
    ↪ MaxCommitTS: 407638915597467649, UpdateTime: 2019-04-11 17:44:18 -0400 EDT}"]
```

使用 binlogctl 的 “NodeIDs” 可控制单个对应节点。在该情况下，Drainer 的节点 ID 是 “localhost.localdomain:8249”，Pump 的节点 ID 是 “localhost.localdomain:8250”。

本文档中的 binlogctl 主要用于集群重启。如果在 TiDB 集群中终止并尝试重启所有的进程，由于 Pump 无法连接 Drainer 且认为 Drainer 依旧 “在线”，Pump 会拒绝启动。这里的进程并不包括下游的 MySQL 或 MariaDB 或 Drainer。

以下有三个方案可解决上述问题：

- 使用 binlogctl 停止 Drainer，而不是结束进程：

  ```
  ./bin/binlogctl --pd-urls=http://127.0.0.1:2379 --cmd=drainers &&
  ./bin/binlogctl --pd-urls=http://127.0.0.1:2379 --cmd=pause-drainer --node-id=localhost.
      ↪ localdomain:8249
  ```

- 在启动 Pump 之前先启动 Drainer。

- 在启动 PD 之后但在启动 Drainer 和 Pump 之前，使用 binlogctl 更新已暂定 Drainer 的状态。

  ```
  ./bin/binlogctl --pd-urls=http://127.0.0.1:2379 --cmd=update-drainer --node-id=localhost.
      ↪ localdomain:8249 --state=paused
  ```

12.10.2.8 清理

在 shell 终端里可启动创建集群的所有进程（`pd-server`、`tikv-server`、`pump`、`tidb-server`、`drainer`）。可通过在 shell 终端中执行 `pkill -P $$` 停止 TiDB 集群服务和 TiDB Binlog 进程。按一定的顺序停止这些进程有利于留出足够的时间彻底关闭每个组件。

```
for p in tidb-server drainer pump tikv-server pd-server; do pkill "$p"; sleep 1; done
```

预期输出：

```
[4]-  Done                    ./bin/tidb-server --config=tidb.toml &>tidb.out
[5]+  Done                    ./bin/drainer --config=drainer.toml &>drainer.out
[3]+  Done                    ./bin/pump --config=pump.toml &>pump.out
[2]+  Done                    ./bin/tikv-server --config=tikv.toml &>tikv.out
[1]+  Done                    ./bin/pd-server --config=pd.toml &>pd.out
```

如果需要所有服务退出后重启集群，可以使用一开始启动服务的命令。如以上binlogctl 部分所述，需要先启动 Drainer 再启动 Pump，最后启动 tidb-server。

```
./bin/pd-server --config=pd.toml &>pd.out &
./bin/tikv-server --config=tikv.toml &>tikv.out &
./bin/drainer --config=drainer.toml &>drainer.out &
sleep 3
./bin/pump --config=pump.toml &>pump.out &
sleep 3
./bin/tidb-server --config=tidb.toml &>tidb.out &
```

如果有组件启动失败，请尝试单独重启该组件。

12.10.2.9 总结

本文档介绍了如何通过设置 TiDB Binlog，使用单个 Pump 和 Drainer 组成的集群同步 TiDB 集群数据到下游的 MariaDB。可以发现，TiDB Binlog 是用于获取处理 TiDB 集群中更新数据的综合性平台工具。

在更稳健的开发、测试或生产部署环境中，可以使用多个 TiDB 服务以实现高可用性和扩展性。使用多个 Pump 实例可以避免 Pump 集群中的问题影响发送到 TiDB 实例的应用流量。或者可以使用增加的 Drainer 实例同步数据到不同的下游或实现数据增量备份。

12.10.3 TiDB Binlog 集群部署

12.10.3.1 服务器要求

Pump 和 Drainer 均可部署和运行在 Intel x86-64 架构的 64 位通用硬件服务器平台上。在开发、测试和生产环境下，对服务器硬件配置的要求和建议如下：

服务	部署数量	CPU	磁盘	内存
Pump	3	8 核 +	SSD, 200 GB+	16G
Drainer	1	8 核 +	SAS, 100 GB+（如果输出 binlog 为本地文件，磁盘大小视保留数据天数而定）	16G

12.10.3.2 使用 TiUP 部署 TiDB Binlog

推荐使用 TiUP 部署 TiDB Binlog。即在使用 TiUP 部署 TiDB 时，在拓扑文件中添加 TiDB Binlog 的 drainer 和 pump 节点信息后，再随 TiDB 一起部署。详细部署方式参考TiUP 部署 TiDB 集群。

12.10.3.3 使用 Binary 部署 TiDB Binlog

12.10.3.3.1 下载官方 Binary

```
wget https://download.pingcap.org/tidb-{version}-linux-amd64.tar.gz &&
wget https://download.pingcap.org/tidb-{version}-linux-amd64.sha256
```

检查文件完整性，返回 ok 则正确：

```
sha256sum -c tidb-{version}-linux-amd64.sha256
```

对于 v2.1.0 GA 及以上版本，Pump 和 Drainer 已经包含在 TiDB 的下载包中，其他版本需要单独下载 Pump 和 Drainer:

```
wget https://download.pingcap.org/tidb-binlog-latest-linux-amd64.tar.gz &&
wget https://download.pingcap.org/tidb-binlog-latest-linux-amd64.sha256
```

检查文件完整性，返回 ok 则正确：

```
sha256sum -c tidb-binlog-latest-linux-amd64.sha256
```

12.10.3.3.2 使用样例

假设有三个 PD，一个 TiDB，另外有两台机器用于部署 Pump，一台机器用于部署 Drainer。各个节点信息如下：

```
TiDB="192.168.0.10"
PD1="192.168.0.16"
PD2="192.168.0.15"
PD3="192.168.0.14"
Pump="192.168.0.11"
Pump="192.168.0.12"
Drainer="192.168.0.13"
```

下面以此为例，说明 Pump/Drainer 的使用。

1. 使用 binary 部署 Pump

    - Pump 命令行参数说明（以在 “192.168.0.11” 上部署为例）

```
Usage of Pump:
-L string
    日志输出信息等级设置：debug，info，warn，error，fatal（默认 "info"）
-V
    打印版本信息
-addr string
    Pump 提供服务的 RPC 地址(-addr="192.168.0.11:8250")
-advertise-addr string
    Pump 对外提供服务的 RPC 地址(-advertise-addr="192.168.0.11:8250")
-config string
    配置文件路径，如果你指定了配置文件，Pump 会首先读取配置文件的配置；
```

```
    如果对应的配置在命令行参数里面也存在，Pump
        ↪ 就会使用命令行参数的配置来覆盖配置文件里的配置。
-data-dir string
    Pump 数据存储位置路径
-gc int
    Pump 只保留多少天以内的数据（默认 7）
-heartbeat-interval int
    Pump 向 PD 发送心跳间隔（单位 秒）
-log-file string
    log 文件路径
-log-rotate string
    log 文件切换频率，hour/day
-metrics-addr string
    Prometheus Pushgateway 地址，不设置则禁止上报监控信息
-metrics-interval int
    监控信息上报频率（默认 15，单位 秒）
-node-id string
    Pump 节点的唯一识别 ID，如果不指定，程序会根据主机名和监听端口自动生成
-pd-urls string
    PD 集群节点的地址（-pd-urls="http://192.168.0.16:2379,http://192.168.0.15:2379,http
        ↪ ://192.168.0.14:2379"）
-fake-binlog-interval int
    Pump 节点生成 fake binlog 的频率（默认 3，单位 秒）
```

- Pump 配置文件（以在“192.168.0.11”上部署为例）

```
# Pump Configuration

# Pump 绑定的地址
addr = "192.168.0.11:8250"

# Pump 对外提供服务的地址
advertise-addr = "192.168.0.11:8250"

# Pump 只保留多少天以内的数据（默认 7）
gc = 7

# Pump 数据存储位置路径
data-dir = "data.pump"

# Pump 向 PD 发送心跳的间隔（单位 秒）
heartbeat-interval = 2

# PD 集群节点的地址（英文逗号分割，中间不加空格）
pd-urls = "http://192.168.0.16:2379,http://192.168.0.15:2379,http://192.168.0.14:2379"
```

```
# [security]
# 如无特殊安全设置需要，该部分一般都注解掉
# 包含与集群连接的受信任 SSL CA 列表的文件路径
# ssl-ca = "/path/to/ca.pem"
# 包含与集群连接的 PEM 形式的 X509 certificate 的路径
# ssl-cert = "/path/to/drainer.pem"
# 包含与集群链接的 PEM 形式的 X509 key 的路径
# ssl-key = "/path/to/drainer-key.pem"

# [storage]
# 设置为 true（默认值）来保证可靠性，确保 binlog 数据刷新到磁盘
# sync-log = true

# 当可用磁盘容量小于该设置值时，Pump 将停止写入数据
# 42 MB -> 42000000, 42 mib -> 44040192
# default: 10 gib
# stop-write-at-available-space = "10 gib"

# Pump 内嵌的 LSM DB 设置，除非对该部分很了解，否则一般注解掉
# [storage.kv]
# block-cache-capacity = 8388608
# block-restart-interval = 16
# block-size = 4096
# compaction-L0-trigger = 8
# compaction-table-size = 67108864
# compaction-total-size = 536870912
# compaction-total-size-multiplier = 8.0
# write-buffer = 67108864
# write-L0-pause-trigger = 24
# write-L0-slowdown-trigger = 17
```

- 启动示例

```
./bin/pump -config pump.toml
```

如果命令行参数与配置文件中的参数重合，则使用命令行设置的参数的值。

2. 使用 binary 部署 Drainer

- Drainer 命令行参数说明（以在 “192.168.0.13” 上部署为例）

```
Usage of Drainer
-L string
    日志输出信息等级设置：debug，info，warn，error，fatal（默认 "info"）
-V
    打印版本信息
-addr string
```

```
    Drainer 提供服务的地址(-addr="192.168.0.13:8249")
-c int
    同步下游的并发数，该值设置越高同步的吞吐性能越好 (default 1)
-cache-binlog-count int
    缓存中的 binlog 数目限制 ( 默认 8 )
    如果上游的单个 binlog 较大导致 Drainer 出现 OOM 时，可尝试调小该值减少内存使用
-config string
    配置文件路径，Drainer 会首先读取配置文件的配置；
    如果对应的配置在命令行参数里面也存在，Drainer
        ↪ 就会使用命令行参数的配置来覆盖配置文件里面的配置
-data-dir string
    Drainer 数据存储位置路径 (默认 "data.drainer")
-dest-db-type string
    Drainer 下游服务类型 (默认为 mysql，支持 tidb、kafka、file)
-detect-interval int
    向 PD 查询在线 Pump 的时间间隔 (默认 10，单位 秒)
-disable-detect
    是否禁用冲突监测
-disable-dispatch
    是否禁用拆分单个 binlog 的 SQL 的功能，如果设置为 true，则每个 binlog
    按顺序依次还原成单个事务进行同步 ( 下游服务类型为 MySQL，该项设置为 False )
-ignore-schemas string
    db 过滤列表 (默认 "INFORMATION_SCHEMA,PERFORMANCE_SCHEMA,mysql,test"),
    不支持对 ignore schemas 的 table 进行 rename DDL 操作
-initial-commit-ts ( 默认为 `-1` )
    如果 Drainer 没有相关的断点信息，可以通过该项来设置相关的断点信息
    该参数值为 `-1` 时，Drainer 会自动从 PD 获取一个最新的时间戳
-log-file string
    log 文件路径
-log-rotate string
    log 文件切换频率，hour/day
-metrics-addr string
    Prometheus Pushgateway 地址，不设置则禁止上报监控信息
-metrics-interval int
    监控信息上报频率 ( 默认 15，单位：秒 )
-node-id string
    drainer 节点的唯一识别 ID，如果不指定，程序会根据主机名和监听端口自动生成
-pd-urls string
    PD 集群节点的地址 (-pd-urls="http://192.168.0.16:2379,http://192.168.0.15:2379,http
        ↪ ://192.168.0.14:2379")
-safe-mode
    是否开启安全模式使得下游 MySQL/TiDB 可被重复写入
    即将 insert 语句换为 replace 语句，将 update 语句拆分为 delete + replace 语句
-txn-batch int
    输出到下游数据库一个事务的 SQL 数量 ( 默认 1 )
```

- Drainer 配置文件（以在 “192.168.0.13” 上部署为例）

```
# Drainer Configuration.

# Drainer 提供服务的地址("192.168.0.13:8249")
addr = "192.168.0.13:8249"

# Drainer 对外提供服务的地址
advertise-addr = "192.168.0.13:8249"

# 向 PD 查询在线 Pump 的时间间隔（默认 10，单位 秒）
detect-interval = 10

# Drainer 数据存储位置路径（默认 "data.drainer"）
data-dir = "data.drainer"

# PD 集群节点的地址（英文逗号分割，中间不加空格）
pd-urls = "http://192.168.0.16:2379,http://192.168.0.15:2379,http://192.168.0.14:2379"

# log 文件路径
log-file = "drainer.log"

# Drainer 从 Pump 获取 binlog 时对数据进行压缩，值可以为 "gzip"，如果不配置则不进行压缩
# compressor = "gzip"

# [security]
# 如无特殊安全设置需要，该部分一般都注解掉
# 包含与集群连接的受信任 SSL CA 列表的文件路径
# ssl-ca = "/path/to/ca.pem"
# 包含与集群连接的 PEM 形式的 X509 certificate 的路径
# ssl-cert = "/path/to/pump.pem"
# 包含与集群链接的 PEM 形式的 X509 key 的路径
# ssl-key = "/path/to/pump-key.pem"

# Syncer Configuration
[syncer]
# 如果设置了该项，会使用该 sql-mode 解析 DDL 语句，此时如果下游是 MySQL 或 TiDB 则
# 下游的 sql-mode 也会被设置为该值
# sql-mode = "STRICT_TRANS_TABLES,NO_ENGINE_SUBSTITUTION"

# 输出到下游数据库一个事务的 SQL 语句数量（默认 20）
txn-batch = 20

# 同步下游的并发数，该值设置越高同步的吞吐性能越好（默认 16）
worker-count = 16
```

```
# 是否禁用拆分单个 binlog 的 SQL 的功能，如果设置为 true，则按照每个 binlog
# 顺序依次还原成单个事务进行同步（下游服务类型为 MySQL，该项设置为 False）
disable-dispatch = false

# safe mode 会使写下游 MySQL/TiDB 可被重复写入
# 会用 replace 替换 insert 语句，用 delete + replace 替换 update 语句
safe-mode = false

# Drainer 下游服务类型（默认为 mysql）
# 参数有效值为 "mysql"，"tidb"，"file"，"kafka"
db-type = "mysql"

# 事务的 commit ts 若在该列表中，则该事务将被过滤，不会同步至下游
ignore-txn-commit-ts = []

# db 过滤列表（默认 "INFORMATION_SCHEMA,PERFORMANCE_SCHEMA,mysql,test"），
# 不支持对 ignore schemas 的 table 进行 rename DDL 操作
ignore-schemas = "INFORMATION_SCHEMA,PERFORMANCE_SCHEMA,mysql"

# replicate-do-db 配置的优先级高于 replicate-do-table。如果配置了相同的库名，
    ↪ 支持使用正则表达式进行配置。
# 以 '~' 开始声明使用正则表达式

# replicate-do-db = ["~^b.*","s1"]

# [syncer.relay]
# 保存 relay log 的目录，空值表示不开启。
# 只有下游是 TiDB 或 MySQL 时该配置才生效。
# log-dir = ""
# 每个文件的大小上限
# max-file-size = 10485760

# [[syncer.replicate-do-table]]
# db-name ="test"
# tbl-name = "log"

# [[syncer.replicate-do-table]]
# db-name ="test"
# tbl-name = "~^a.*"

# 忽略同步某些表
# [[syncer.ignore-table]]
# db-name = "test"
# tbl-name = "log"
```

```
# db-type 设置为 mysql 时，下游数据库服务器参数
[syncer.to]
host = "192.168.0.13"
user = "root"
# 如果你不想在配置文件中写明文密码，则可以使用 `./binlogctl -cmd encrypt -text string`
    ↪ 生成加密的密码
# 如果配置了 encrypted_password 且非空，那么配置的 password 不生效。encrypted_password
    ↪ 和 password 无法同时生效。
password = ""
encrypted_password = ""
port = 3306

[syncer.to.checkpoint]
# 当 checkpoint type 是 mysql 或 tidb 时可以开启该选项，以改变保存 checkpoint 的数据库
# schema = "tidb_binlog"
# 目前只支持 mysql 或者 tidb 类型。可以去掉注释来控制 checkpoint 保存的位置。
# db-type 默认的 checkpoint 保存方式是:
# mysql/tidb -> 对应的下游 mysql/tidb
# file/kafka -> file in `data-dir`
# type = "mysql"
# host = "127.0.0.1"
# user = "root"
# password = ""
# 使用 `./binlogctl -cmd encrypt -text string` 加密的密码
# encrypted_password 非空时 password 会被忽略
# encrypted_password = ""
# port = 3306

# db-type 设置为 file 时，存放 binlog 文件的目录
# [syncer.to]
# dir = "data.drainer"

# db-type 设置为 kafka 时，Kafka 相关配置
# [syncer.to]
# kafka-addrs 和 zookeeper-addrs 只需要一个，两者都有时程序会优先用 zookeeper 中的
    ↪ kafka 地址
# zookeeper-addrs = "127.0.0.1:2181"
# kafka-addrs = "127.0.0.1:9092"
# kafka-version = "0.8.2.0"
# 配置单条 broker request 中的最大 message 数（即 binlog 数），不配置或配置小于等于 0
    ↪ 时会使用默认值 1024
# kafka-max-messages = 1024
# 配置单条 broker request 的最大 size（单位为 Byte），默认为 1 GiB，最大可配置为 2 GiB
# kafka-max-message-size = 1073741824
```

```
# 保存 binlog 数据的 Kafka 集群的 topic 名称，默认值为 <cluster-id>_obinlog
# 如果运行多个 Drainer 同步数据到同一个 Kafka 集群，每个 Drainer 的 topic-name
    ↪ 需要设置不同的名称
# topic-name = ""
```

- 启动示例

> 注意：
> 如果下游为 MySQL/TiDB，为了保证数据的完整性，在 Drainer 初次启动前需要获取 `initial-commit-ts` 的值，并进行全量数据的备份与恢复。

初次启动时使用参数 `initial-commit-ts`，命令如下：

```
./bin/drainer -config drainer.toml -initial-commit-ts {initial-commit-ts}
```

如果命令行参数与配置文件中的参数重合，则使用命令行设置的参数的值。

注意：

- 在运行 TiDB 时，需要保证至少一个 Pump 正常运行。
- 通过给 TiDB 增加启动参数 `enable-binlog` 来开启 binlog 服务。尽量保证同一集群的所有 TiDB 都开启了 binlog 服务，否则在同步数据时可能会导致上下游数据不一致。如果要临时运行一个不开启 binlog 服务的 TiDB 实例，需要在 TiDB 的配置文件中设置 `run_ddl= false`。
- Drainer 不支持对 ignore schemas（在过滤列表中的 schemas）的 table 进行 rename DDL 操作。
- 在已有的 TiDB 集群中启动 Drainer，一般需要全量备份并且获取快照时间戳，然后导入全量备份，最后启动 Drainer 从对应的快照时间戳开始同步增量数据。
- 下游使用 MySQL 或 TiDB 时应当保证上下游数据库的 sql_mode 具有一致性，即下游数据库同步每条 SQL 语句时的 sql_mode 应当与上游数据库执行该条 SQL 语句时的 sql_mode 保持一致。可以在上下游分别执行 `select @@sql_mode;` 进行查询和比对。
- 如果存在上游 TiDB 能运行但下游 MySQL 不支持的 DDL 语句时（例如下游 MySQL 使用 InnoDB 引擎时同步语句 `CREATE TABLE t1(a INT)ROW_FORMAT=FIXED;`），Drainer 也会同步失败，此时可以在 Drainer 配置中跳过该事务，同时在下游手动执行兼容的语句，详见跳过事务。

12.10.4 TiDB Binlog 集群运维

本文首先介绍 Pump 和 Drainer 的状态及启动、退出的内部处理流程，然后说明如何通过 binlogctl 工具或者直接在 TiDB 执行 SQL 操作来管理 binlog 集群，最后的 FAQ 部分会介绍一些常见问题以及处理方法。

12.10.4.1 Pump/Drainer 的状态

Pump/Drainer 中状态的定义：

- online：正常运行中
- pausing：暂停中
- paused：已暂停
- closing：下线中
- offline：已下线

这些状态由 Pump/Drainer 服务自身进行维护，并定时将状态信息更新到 PD 中。

12.10.4.2 Pump/Drainer 的启动、退出流程

12.10.4.2.1 Pump

- 启动：Pump 启动时会通知所有 online 状态的 Drainer，如果通知成功，则 Pump 将状态设置为 online，否则 Pump 将报错，然后将状态设置为 paused 并退出进程。
- 退出：Pump 进程正常退出前要选择进入暂停或者下线状态；非正常退出（kill -9、进程 panic、宕机）都依然保持 online 状态。
  - 暂停：使用 kill（非 kill -9）、Ctrl+C 或者使用 binlogctl 的 pause-pump 命令都可以暂停 Pump。接收到暂停指令后，Pump 会变更状态为 pausing，并停止接受 binlog 的写请求，也停止向 Drainer 提供 binlog 数据。安全退出所有线程后，更新状态为 paused 然后退出进程。
  - 下线：仅在使用 binlogctl 的 offline-pump 命令的情况下才会下线 Pump。接收到下线指令后，Pump 会变更状态为 closing，并停止接受 binlog 的写请求。Pump 继续向 Drainer 提供 binlog，等待所有 binlog 数据都被 Drainer 消费后再将状态设置为 offline 并退出进程。

12.10.4.2.2 Drainer

- 启动：Drainer 启动时将状态设置为 online，并尝试从所有非 offline 状态的 Pump 获取 binlog，如果获取 binlog 失败，会不断尝试重新获取。
- 退出：Drainer 进程正常退出前要选择进入暂停或者下线状态；非正常退出（kill -9 、进程 panic、宕机）都依然保持 online 状态。
  - 暂停：使用 kill（非 kill -9）、Ctrl+C 或者使用 binlogctl 的 pause-drainer 命令都可以暂停 Drainer。接收到指令后，Drainer 会变更状态为 pausing，并停止从 Pump 获取 binlog。安全退出所有线程后，更新状态为 paused 然后退出进程。
  - 下线：仅在使用 binlogctl 的 offline-drainer 命令的情况下才会下线 Drainer。接收到下线指令后，Drainer 变更状态为 closing，并停止从 Pump 获取 binlog。安全退出所有线程后，更新状态为 offline 然后退出进程。

关于 Pump/Drainer 暂停、下线、状态查询、状态修改等具体的操作方法，参考如下 binlogctl 工具的使用方法介绍。

12.10.4.3 使用 binlogctl 工具管理 Pump/Drainer

binlogctl 支持如下这些功能：

- 查看 Pump/Drainer 状态
- 暂停/下线 Pump/Drainer
- Pump/Drainer 异常状态处理

详细的介绍和使用方法请参考binlogctl 工具。

12.10.4.4 使用 TiDB SQL 管理 Pump/Drainer

要查看和管理 binlog 相关的状态，可在 TiDB 中执行相应的 SQL 语句。

- 查看 TiDB 是否开启 binlog，0 代表关闭，1 代表开启

```
show variables like "log_bin";
```

```
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| log_bin       |  0    |
+---------------+-------+
```

- 查看 Pump/Drainer 状态

```
show pump status;
```

```
+--------|----------------|--------|--------------------|---------------------|
| NodeID |     Address    | State  |   Max_Commit_Ts    |    Update_Time      |
+--------|----------------|--------|--------------------|---------------------|
| pump1  | 127.0.0.1:8250 | Online | 408553768673342237 | 2019-05-01 00:00:01 |
+--------|----------------|--------|--------------------|---------------------|
| pump2  | 127.0.0.2:8250 | Online | 408553768673342335 | 2019-05-01 00:00:02 |
+--------|----------------|--------|--------------------|---------------------|
```

```
show drainer status;
```

```
+----------|----------------|--------|--------------------|---------------------|
|  NodeID  |     Address    | State  |   Max_Commit_Ts    |    Update_Time      |
+----------|----------------|--------|--------------------|---------------------|
| drainer1 | 127.0.0.3:8249 | Online | 408553768673342532 | 2019-05-01 00:00:03 |
+----------|----------------|--------|--------------------|---------------------|
| drainer2 | 127.0.0.4:8249 | Online | 408553768673345531 | 2019-05-01 00:00:04 |
+----------|----------------|--------|--------------------|---------------------|
```

- 异常情况下修改 Pump/Drainer 状态

```
change pump to node_state ='paused' for node_id 'pump1';
```

```
Query OK, 0 rows affected (0.01 sec)
```

```
change drainer to node_state ='paused' for node_id 'drainer1';
```

```
Query OK, 0 rows affected (0.01 sec)
```

该 SQL 的功能和 binlogctl 中的 update-pump 和 update-drainer 命令的功能一样，因此也只有在 Pump/Drainer 异常的情况下使用。

注意：

1. 查看 binlog 开启状态以及 Pump/Drainer 状态的功能在 TiDB v2.1.7 及以上版本中支持。
2. 修改 Pump/Drainer 状态的功能在 TiDB v3.0.0-rc.1 及以上版本中支持。该功能只修改 PD 中存储的 Pump/Drainer 状态，如果需要暂停/下线节点，仍然需要使用 `binlogctl`。

12.10.5 TiDB Binlog 配置说明

本文档介绍 TiDB Binlog 的各项配置说明。

12.10.5.1 Pump

本节介绍 Pump 的配置项。可以在 Pump Configuration 中查看完整的 Pump 配置文件示例。

12.10.5.1.1 addr

- HTTP API 的监听地址，格式为 `host:port`。
- 默认：`"127.0.0.1:8250"`

12.10.5.1.2 advertise-addr

- 对外可访问的 HTTP API 地址。这个地址会被注册到 PD，格式为 `host:port`。
- 默认：与 `addr` 的配置相同。

12.10.5.1.3 socket

- HTTP API 监听的 Unix socket 地址。
- 默认：`""`

12.10.5.1.4 pd-urls

- 由逗号分隔的 PD URL 列表。如果指定了多个地址，PD 客户端在连接一个地址时出错时会自动尝试连接另一个地址。
- 默认：`"http://127.0.0.1:2379"`

12.10.5.1.5 data-dir

- 本地存放 binlog 及其索引的目录。
- 默认：`"data.pump"`

12.10.5.1.6 heartbeat-interval

- 心跳间隔，即每隔指定秒数向 PD 汇报最新的状态。
- 默认：2

12.10.5.1.7 gen-binlog-interval

- 指定写入 fake binlog 的间隔秒数。
- 默认：3

12.10.5.1.8 gc

- 指定 binlog 可在本地存储的天数（整型）。超过指定天数的 binlog 会被自动删除。
- 默认：7

12.10.5.1.9 log-file

- 保存日志文件的路径。如果为空，日志不会被保存。
- 默认：`""`

12.10.5.1.10 log-level

- Log 等级。
- 默认：`"info"`

12.10.5.1.11 node-id

- Pump 节点的 ID，用于在集群中识别这个进程。
- 默认：主机名:端口号，例如 node-1:8250。

12.10.5.1.12 security

以下是与安全相关的配置项。

ssl-ca

- 包含可信 SSL 证书或 CA 证书列表的文件路径，例如 `/path/to/ca.pem`。
- 默认：`""`

ssl-cert

- 包含 PEM 格式编码的 X509 证书文件路径，例如 `/path/to/pump.pem`。
- 默认：`""`

ssl-key

- 包含 PEM 格式编码的 X509 Key 文件路径，例如 `/path/to/pump-key.pem`。
- 默认：`""`

12.10.5.1.13 storage

以下是与存储相关的配置项。

sync-log

- 指定是否在每次批量写入 binlog 后使用 `fsync` 以确保数据安全。
- 默认：`true`

kv_chan_cap

- 在 Pump 接收写入请求时会先将请求放入一个缓冲区，该项指定缓冲区能存放的请求数量。
- 默认：1048576（即 2 的 20 次方）

slow_write_threshold

- 写入单个 binlog 的耗时超过该项设定的秒数就被认为是慢写入，并输出一条包含 `"take a long time to write binlog"` 的日志。
- 默认：1

stop-write-at-available-space

- 可用存储空间低于指定值时不再接收 binlog 写入请求。可以用例如 `900 MB`、`5 GB`、`12 GiB` 的格式指定空间大小。如果集群中 Pump 节点多于一个，那么在某个 Pump 节点因为空间不足而拒绝写入时，TiDB 端会自动写入到其他 Pump 节点。
- 默认：`10 GiB`

kv

目前 Pump 的存储是基于 GoLevelDB 实现的。`storage` 下还有一个 kv 子分组可以用于调整 GoLevelDB 的配置，支持的配置项包括：

- block-cache-capacity
- block-restart-interval
- block-size
- compaction-L0-trigger
- compaction-table-size
- compaction-total-size
- compaction-total-size-multiplier
- write-buffer
- write-L0-pause-trigger
- write-L0-slowdown-trigger

配置具体含义可在 GoLevelDB 文档中查看。

12.10.5.2 Drainer

本节介绍 Drainer 的配置项。可以在 Drainer Configuration 中查看完整的配置文件示例。

12.10.5.2.1 addr

- HTTP API 的监听的地址，格式为 `host:port`。
- 默认：`"127.0.0.1:8249"`

12.10.5.2.2 advertise-addr

- 对外可访问的 HTTP API 地址，这个地址会被注册到 PD，格式为 `host:port`。
- 默认：设定成与 `addr` 相同的配置

12.10.5.2.3 log-file

- 日志的文件保存路径。如果为空，日志不会被保存。
- 默认：`""`

12.10.5.2.4 log-level

- Log 等级。
- 默认：`"info"`

12.10.5.2.5 node-id

- Drainer 节点 ID，用于在集群中识别这个进程。
- 默认：主机名:端口号，例如 `node-1:8249`。

12.10.5.2.6 data-dir

- 用于存放 Drainer 运行中需要保存的文件的目录。
- 默认：`"data.drainer"`

12.10.5.2.7 detect-interval

- 每隔指定秒数从 PD 更新一次 Pumps 信息，以获取节点加入或离开等事件。
- 默认：5

12.10.5.2.8 pd-urls

- 由逗号分隔的 PD URL 列表。如果指定了多个地址，PD 客户端在连接一个地址出错时会自动尝试连接另一个地址。
- 默认：`"http://127.0.0.1:2379"`

12.10.5.2.9 initial-commit-ts

- 指定从哪个事务提交时间点（事务的 commit ts）之后开始同步。这个配置仅适用于初次开始同步的 Drainer 节点。如果下游已经有 checkpoint 存在，则会根据 checkpoint 里记录的时间进行同步。
- commit ts（即 commit timestamp）是 TiDB 事务的提交时间点。该时间点是从 PD 获取的全局唯一递增的时间戳，作为当前事务的唯一 ID。典型的 `initial-commit-ts` 配置可以通过以下方式获得：
  - BR 备份的元信息（即 backupmeta）中记录的 backup TS
  - Dumpling 备份的元信息（即 metadata）中记录的 Pos
  - PD Control 中 `tso` 命令返回的结果
- 默认：-1。Drainer 会从 PD 得到一个最新的 timestamp 作为初始时间。即从当前的时间点开始同步。

12.10.5.2.10 synced-check-time

- 通过 HTTP API 访问 `/status` 路径可以查询 Drainer 同步的状态。`synced-check-time` 指定距离上次成功同步的时间超过多少分钟可以认为是 `synced`，即同步完成。
- 默认：5

12.10.5.2.11 compressor

- 指定 Pump 与 Drainer 间的数据传输所用的压缩算法。目前仅支持一种算法，即 `gzip`。
- 默认：`""`，表示不压缩。

12.10.5.2.12 security

以下是与 Drainer 安全相关的配置。

ssl-ca

- 包含可信 SSL 证书或 CA 证书列表的文件路径，例如 `/path/to/ca.pem`。
- 默认：`""`

ssl-cert

- 包含 PEM 格式编码的 X509 证书文件路径，例如 `/path/to/drainer.pem`。
- 默认：`""`

ssl-key

- 包含 PEM 格式编码的 X509 Key 文件路径，例如 `/path/to/drainer-key.pem`。
- 默认：`""`

12.10.5.2.13 syncer

syncer 分组包含一些与同步下游相关的配置项。

db-type

下游类型，目前支持的选项有：

- `mysql`
- `tidb`
- `kafka`
- `file`

默认：`mysql`

sql-mode

- 当下游为 `mysql/tidb` 类型时，可以指定 SQL mode，如果超过一个 mode，则用逗号隔开。
- 默认：`""`

ignore-txn-commit-ts

- 同步时，该项所指定的 commit timestamp 的 binlog 会被忽略，例如 `[416815754209656834, 421349811963822081]`。
- 默认：`[]`

ignore-schemas

- 同步时忽略指定数据库的变更。如果超过一个需忽略的数据库，则用逗号隔开。如果一个 binlog 中的变更全部被过滤掉，则忽略整个 binlog。
- 默认：`"INFORMATION_SCHEMA,PERFORMANCE_SCHEMA,mysql"`

ignore-table

同步时忽略指定表格的变更。在 `toml` 中可以用以下方式指定多个需忽略的表格：

```
[[syncer.ignore-table]]
db-name = "test"
tbl-name = "log"

[[syncer.ignore-table]]
db-name = "test"
tbl-name = "audit"
```

如果一个 binlog 中的变更全部被过滤掉，则忽略整个 binlog。

默认：`[]`

replicate-do-db

- 指定要同步的数据库，例如 `[db1, db2]`。
- 默认：`[]`

replicate-do-table

指定要同步的表格，示例如下：

```
[[syncer.replicate-do-table]]
db-name ="test"
tbl-name = "log"

[[syncer.replicate-do-table]]
db-name ="test"
tbl-name = "~^a.*"
```

默认：`[]`

txn-batch

- 当下游为 `mysql/tidb` 类型时，会将 DML 分批执行。这个配置可以用于设置每个事务中包含多少个 DML。
- 默认：`20`

worker-count

- 当下游为 `mysql/tidb` 类型时，会并发执行 DML，`worker-count` 可以指定并发数。
- 默认：`16`

disable-dispatch

- 关掉并发，强制将 worker-count 置为 1。
- 默认：false

safe-mode

如果打开 Safe mode，Drainer 会对同步的变更作以下修改，使其变成可重入的操作：

- Insert 变为 Replace Into
- Update 变为 Delete 和 Replace Into

默认：false

12.10.5.2.14 syncer.to

不同类型的下游配置都在 syncer.to 分组。以下按配置类型进行介绍。

mysql/tidb

用于连接下游数据库的配置项：

- host：如果没有设置，会尝试检查环境变量 MYSQL_HOST，默认值为 "localhost"。
- port：如果没有设置，会尝试检查环境变量 MYSQL_PORT，默认值为 3306。
- user：如果没有设置，会尝试检查环境变量 MYSQL_USER，默认值为 "root"。
- password：如果没有设置，会尝试检查环境变量 MYSQL_PSWD，默认值为 ""。

file

- dir：指定用于保存 binlog 的目录。如果不指定该项，则使用 data-dir。

kafka

当下游为 kafka 时，有效的配置包括：

- zookeeper-addrs
- kafka-addrs
- kafka-version
- kafka-max-messages
- kafka-max-message-size
- topic-name

12.10.5.2.15 syncer.to.checkpoint

- type：指定用哪种方式保存同步进度，目前支持的选项为 mysql、tidb 和 file。

  该配置选项默认与下游类型相同。例如 file 类型的下游 checkpoint 进度保存在本地文件 <data-dir>/
  ↪ savepoint 中，mysql 类型的下游进度保存在下游数据库。当明确指定要使用 mysql 或 tidb 保存同步进度时，需要指定以下配置项：

- schema：默认为 "tidb_binlog"。

> 注意：
> 在同个 TiDB 集群中部署多个 Drainer 时，需要为每个 Drainer 节点指定不同的 checkpoint schema，否则两个实例的同步进度会互相覆盖。

- host
- user
- password
- port

12.10.5.3 TiDB Binlog 配置说明

本文档介绍 TiDB Binlog 的各项配置说明。

12.10.5.3.1 Pump

本节介绍 Pump 的配置项。可以在 Pump Configuration 中查看完整的 Pump 配置文件示例。

addr

- HTTP API 的监听地址，格式为 host:port。
- 默认："127.0.0.1:8250"

advertise-addr

- 对外可访问的 HTTP API 地址。这个地址会被注册到 PD，格式为 host:port。
- 默认：与 addr 的配置相同。

socket

- HTTP API 监听的 Unix socket 地址。
- 默认：""

pd-urls

- 由逗号分隔的 PD URL 列表。如果指定了多个地址，PD 客户端在连接一个地址时出错时会自动尝试连接另一个地址。
- 默认："http://127.0.0.1:2379"

data-dir

- 本地存放 binlog 及其索引的目录。

- 默认：`"data.pump"`

heartbeat-interval

- 心跳间隔，即每隔指定秒数向 PD 汇报最新的状态。
- 默认：`2`

gen-binlog-interval

- 指定写入 fake binlog 的间隔秒数。
- 默认：`3`

gc

- 指定 binlog 可在本地存储的天数（整型）。超过指定天数的 binlog 会被自动删除。
- 默认：`7`

log-file

- 保存日志文件的路径。如果为空，日志不会被保存。
- 默认：`""`

log-level

- Log 等级。
- 默认：`"info"`

node-id

- Pump 节点的 ID，用于在集群中识别这个进程。
- 默认：主机名:端口号，例如 `node-1:8250`。

security

以下是与安全相关的配置项。

ssl-ca

- 包含可信 SSL 证书或 CA 证书列表的文件路径，例如 `/path/to/ca.pem`。
- 默认：`""`

ssl-cert

- 包含 PEM 格式编码的 X509 证书文件路径，例如 `/path/to/pump.pem`。
- 默认：`""`

ssl-key

- 包含 PEM 格式编码的 X509 Key 文件路径，例如 `/path/to/pump-key.pem`。
- 默认：`""`

storage

以下是与存储相关的配置项。

sync-log

- 指定是否在每次批量写入 binlog 后使用 `fsync` 以确保数据安全。
- 默认：`true`

kv_chan_cap

- 在 Pump 接收写入请求时会先将请求放入一个缓冲区，该项指定缓冲区能存放的请求数量。
- 默认：1048576（即 2 的 20 次方）

slow_write_threshold

- 写入单个 binlog 的耗时超过该项设定的秒数就被认为是慢写入，并输出一条包含 `"take a long time to write binlog"` 的日志。
- 默认：1

stop-write-at-available-space

- 可用存储空间低于指定值时不再接收 binlog 写入请求。可以用例如 `900 MB`、`5 GB`、`12 GiB` 的格式指定空间大小。如果集群中 Pump 节点多于一个，那么在某个 Pump 节点因为空间不足而拒绝写入时，TiDB 端会自动写入到其他 Pump 节点。
- 默认：`10 GiB`

kv

目前 Pump 的存储是基于 GoLevelDB 实现的。`storage` 下还有一个 kv 子分组可以用于调整 GoLevelDB 的配置，支持的配置项包括：

- block-cache-capacity
- block-restart-interval
- block-size
- compaction-L0-trigger
- compaction-table-size
- compaction-total-size
- compaction-total-size-multiplier
- write-buffer
- write-L0-pause-trigger
- write-L0-slowdown-trigger

配置具体含义可在 GoLevelDB 文档中查看。

12.10.5.3.2 Drainer

本节介绍 Drainer 的配置项。可以在 Drainer Configuration 中查看完整的配置文件示例。

addr

- HTTP API 的监听的地址，格式为 `host:port`。
- 默认：`"127.0.0.1:8249"`

advertise-addr

- 对外可访问的 HTTP API 地址，这个地址会被注册到 PD，格式为 `host:port`。
- 默认：设定成与 `addr` 相同的配置

log-file

- 日志的文件保存路径。如果为空，日志不会被保存。
- 默认：`""`

log-level

- Log 等级。
- 默认：`"info"`

node-id

- Drainer 节点 ID，用于在集群中识别这个进程。
- 默认：主机名:端口号，例如 `node-1:8249`。

data-dir

- 用于存放 Drainer 运行中需要保存的文件的目录。
- 默认：`"data.drainer"`

detect-interval

- 每隔指定秒数从 PD 更新一次 Pumps 信息，以获取节点加入或离开等事件。
- 默认：5

pd-urls

- 由逗号分隔的 PD URL 列表。如果指定了多个地址，PD 客户端在连接一个地址出错时会自动尝试连接另一个地址。
- 默认：`"http://127.0.0.1:2379"`

initial-commit-ts

- 指定从哪个事务提交时间点（事务的 commit ts）之后开始同步。这个配置仅适用于初次开始同步的 Drainer 节点。如果下游已经有 checkpoint 存在，则会根据 checkpoint 里记录的时间进行同步。
- commit ts（即 commit timestamp）是 TiDB 事务的提交时间点。该时间点是从 PD 获取的全局唯一递增的时间戳，作为当前事务的唯一 ID。典型的 `initial-commit-ts` 配置可以通过以下方式获得：
  - BR 备份的元信息（即 backupmeta）中记录的 backup TS
  - Dumpling 备份的元信息（即 metadata）中记录的 Pos
  - PD Control 中 `tso` 命令返回的结果
- 默认：-1。Drainer 会从 PD 得到一个最新的 timestamp 作为初始时间。即从当前的时间点开始同步。

synced-check-time

- 通过 HTTP API 访问 `/status` 路径可以查询 Drainer 同步的状态。`synced-check-time` 指定距离上次成功同步的时间超过多少分钟可以认为是 `synced`，即同步完成。
- 默认：5

compressor

- 指定 Pump 与 Drainer 间的数据传输所用的压缩算法。目前仅支持一种算法，即 `gzip`。
- 默认：`""`，表示不压缩。

security

以下是与 Drainer 安全相关的配置。

ssl-ca

- 包含可信 SSL 证书或 CA 证书列表的文件路径，例如 `/path/to/ca.pem`。
- 默认：`""`

ssl-cert

- 包含 PEM 格式编码的 X509 证书文件路径，例如 `/path/to/drainer.pem`。
- 默认：`""`

ssl-key

- 包含 PEM 格式编码的 X509 Key 文件路径，例如 `/path/to/drainer-key.pem`。
- 默认：`""`

syncer

syncer 分组包含一些与同步下游相关的配置项。

db-type

下游类型，目前支持的选项有：

- `mysql`
- `tidb`
- `kafka`
- `file`

默认：`mysql`

sql-mode

- 当下游为 `mysql/tidb` 类型时，可以指定 SQL mode，如果超过一个 mode，则用逗号隔开。
- 默认：`""`

ignore-txn-commit-ts

- 同步时，该项所指定的 commit timestamp 的 binlog 会被忽略，例如 `[416815754209656834, 421349811963822081]`。
- 默认：`[]`

ignore-schemas

- 同步时忽略指定数据库的变更。如果超过一个需忽略的数据库，则用逗号隔开。如果一个 binlog 中的变更全部被过滤掉，则忽略整个 binlog。
- 默认：`"INFORMATION_SCHEMA,PERFORMANCE_SCHEMA,mysql"`

ignore-table

同步时忽略指定表格的变更。在 `toml` 中可以用以下方式指定多个需忽略的表格：

```
[[syncer.ignore-table]]
db-name = "test"
tbl-name = "log"

[[syncer.ignore-table]]
db-name = "test"
tbl-name = "audit"
```

如果一个 binlog 中的变更全部被过滤掉，则忽略整个 binlog。

默认：`[]`

replicate-do-db

- 指定要同步的数据库，例如 `[db1, db2]`。
- 默认：`[]`

replicate-do-table

指定要同步的表格，示例如下：

```
[[syncer.replicate-do-table]]
db-name ="test"
tbl-name = "log"

[[syncer.replicate-do-table]]
db-name ="test"
tbl-name = "~^a.*"
```

默认：[]

txn-batch

- 当下游为 mysql/tidb 类型时，会将 DML 分批执行。这个配置可以用于设置每个事务中包含多少个 DML。
- 默认：20

worker-count

- 当下游为 mysql/tidb 类型时，会并发执行 DML，worker-count 可以指定并发数。
- 默认：16

disable-dispatch

- 关掉并发，强制将 worker-count 置为 1。
- 默认：false

safe-mode

如果打开 Safe mode，Drainer 会对同步的变更作以下修改，使其变成可重入的操作：

- Insert 变为 Replace Into
- Update 变为 Delete 和 Replace Into

默认：false

syncer.to

不同类型的下游配置都在 syncer.to 分组。以下按配置类型进行介绍。

mysql/tidb

用于连接下游数据库的配置项：

- host：如果没有设置，会尝试检查环境变量 MYSQL_HOST，默认值为 "localhost"。
- port：如果没有设置，会尝试检查环境变量 MYSQL_PORT，默认值为 3306。
- user：如果没有设置，会尝试检查环境变量 MYSQL_USER，默认值为 "root"。
- password：如果没有设置，会尝试检查环境变量 MYSQL_PSWD，默认值为 ""。

file

- `dir`：指定用于保存 binlog 的目录。如果不指定该项，则使用 `data-dir`。

kafka

当下游为 `kafka` 时，有效的配置包括：

- `zookeeper-addrs`
- `kafka-addrs`
- `kafka-version`
- `kafka-max-messages`
- `kafka-max-message-size`
- `topic-name`

syncer.to.checkpoint

- `type`：指定用哪种方式保存同步进度，目前支持的选项为 `mysql`、`tidb` 和 `file`。

  该配置选项默认与下游类型相同。例如 `file` 类型的下游 checkpoint 进度保存在本地文件 `<data-dir>/savepoint` 中，`mysql` 类型的下游进度保存在下游数据库。当明确指定要使用 `mysql` 或 `tidb` 保存同步进度时，需要指定以下配置项：
- `schema`：默认为 `"tidb_binlog"`。

> 注意：
> 在同个 TiDB 集群中部署多个 Drainer 时，需要为每个 Drainer 节点指定不同的 checkpoint schema，否则两个实例的同步进度会互相覆盖。

- `host`
- `user`
- `password`
- `port`

12.10.6 TiDB Binlog 版本升级方法

如未特别指明，文中出现的 TiDB Binlog 均指最新的Cluster 版本。

本文介绍通过手动部署的 TiDB Binlog 的版本升级方法，另外有一小节介绍如何从更早的不兼容版本（Kafka/Local 版本）升级到最新版本。

12.10.6.1 手动部署

12.10.6.1.1 升级 Pump

对集群里的每个 Pump 逐一升级，确保集群中总有 Pump 可以接收 TiDB 发来的 Binlog。

1. 用新版本的 pump 替换原来的文件
2. 重启 Pump 进程

12.10.6.1.2 升级 Drainer

1. 用新版本的 `drainer` 替换原来的文件
2. 重启 Drainer 进程

12.10.6.2 从 Kafka/Local 版本升级到 Cluster 版本

新版本的 TiDB（v2.0.8-binlog、v2.1.0-rc.5 及以上版本）不兼容 Kafka 版本以及 Local 版本的 TiDB Binlog，集群升级到新版本后只能使用 Cluster 版本的 TiDB Binlog。如果在升级前已经使用了 Kafka / Local 版本的 TiDB Binlog，必须将其升级到 Cluster 版本。

TiDB Binlog 版本与 TiDB 版本的对应关系如下：

TiDB Binlog 版本	TiDB 版本	说明
Local	TiDB 1.0 及更低版本	
Kafka	TiDB 1.0 ~ TiDB 2.1 RC5	TiDB 1.0 支持 local 版本和 Kafka 版本的 TiDB Bin-log。

TiDB Binlog 版本	TiDB 版本	说明
Cluster	TiDB v2.0.8-binlog，TiDB 2.1 RC5 及更高版本	TiDB v2.0.8-binlog 是一个支持 Cluster 版本 TiDB Binlog 的 2.0 特殊版本。

12.10.6.2.1 升级流程

> **注意：**
>
> 如果能接受重新导全量数据，则可以直接废弃老版本，按TiDB Binlog 集群部署中的步骤重新部署。

如果想从原来的 checkpoint 继续同步，使用以下升级流程：

1. 部署新版本 Pump。
2. 暂停 TiDB 集群业务。
3. 更新 TiDB 以及配置，写 Binlog 到新的 Pump Cluster。
4. TiDB 集群重新接入业务。
5. 确认老版本的 Drainer 已经将老版本的 Pump 的数据完全同步到下游。

   查询 Drainer 的 `status` 接口，示例命令如下：

```
curl 'http://172.16.10.49:8249/status'
```

```
{"PumpPos":{"172.16.10.49:8250":{"offset":32686}},"Synced": true ,"DepositWindow":{"Upper
    ↪ ":398907800202772481,"Lower":398907799455662081}}
```

如果返回的 Synced 为 true，则可以认为 Binlog 数据已经全部同步到了下游。

6. 启动新版本 Drainer；
7. 下线老版本的 Pump、 Drainer 以及依赖的 Kafka 和 ZooKeeper。

12.10.7 TiDB Binlog 集群监控

成功部署 TiDB Binlog 集群后，可以进入 Grafana Web 界面（默认地址: http://grafana_ip:3000，默认账号：admin，密码：admin）查看 Pump 和 Drainer 的运行状态。

12.10.7.1 监控指标

12.10.7.1.1 Pump

metric 名称	说明
Storage Size	记录磁盘的总空间大小 (capacity)，以及可用磁盘空间大小 (available)
Metadata	记录每个 Pump 的可删除 binlog 的最大 tso (gc_tso)，以及保存的 binlog 的最大的 commit tso (max_commit_tso)。
Write Binlog QPS by Instance	每个 Pump 接收到的写 binlog 请求的 QPS
Write Binlog Latency	记录每个 Pump 写 binlog 的延迟时间
Storage Write Binlog Size	Pump 写 binlog 数据的大小
Storage Write Binlog Latency	Pump 中的 storage 模块写 binlog 数据的延迟

metric 名称	说明
Pump Storage Error By Type	Pump 遇到的 error 数量，按照 error 的类型进行统计
Query TiKV	Pump 通过 TiKV 查询事务状态的次数

12.10.7.1.2 Drainer

metric 名称	说明
Checkpoint TSO	Drainer 已经同步到下游的 binlog 的最大 TSO 对应的时间。可以通过该指标估算同步延迟时间
Pump Handle TSO	记录 Drainer 从各个 Pump 获取到的 binlog 的最大 TSO 对应的时间
Pull Binlog QPS by Pump NodeID	Drainer 从每个 Pump 获取 binlog 的 QPS
95% Binlog Reach Duration By Pump	记录 binlog 从写入 Pump 到被 Drainer 获取到这个过程的延迟时间
Error By Type	Drainer 遇到的 error 数量，按照 error 的类型进行统计
SQL Query Time	Drainer 在下游执行 SQL 的耗时
Drainer Event	各种类型 event 的数量，event 包括 ddl、insert、delete、update、flush、savepoint
Execute Time	写入 binlog 到同步下游模块所消耗的时间
95% Binlog Size	Drainer 从各个 Pump 获取到 binlog 数据的大小
DDL Job Count	Drainer 处理的 DDL 的数量
Queue Size	Drainer 内部工作队列大小

12.10.7.2 监控报警规则

本节介绍了 TiDB Binlog 组件的报警项及相应的报警规则。根据严重级别，报警项可分为三类，按照严重程度由高到低依次为：紧急级别 (Emergency)、重要级别 (Critical)、警告级别 (Warning)。

12.10.7.2.1 紧急级别报警项

紧急级别的报警通常由于服务停止或节点故障导致，此时需要马上进行人工干预操作。

`binlog_pump_storage_error_count`

- 报警规则：

  ```
  changes(binlog_pump_storage_error_count[1m])> 0
  ```

- 规则描述：

  Pump 写 binlog 到本地存储时失败。

- 处理方法：

  确认 `pump_storage_error` 监控是否存在错误，查看 Pump 日志确认原因。

12.10.7.2.2 重要级别报警项

对于重要级别的报警，需要密切关注异常指标。

`binlog_drainer_checkpoint_high_delay`

- 报警规则：

  ```
  (time()- binlog_drainer_checkpoint_tso / 1000)> 3600
  ```

- 规则描述：

  Drainer 同步落后延迟超过 1 个小时。

- 处理方法：

  - 判断从 Pump 获取数据是否太慢：

    监控 Pump handle tso 可以看每个 Pump 最近一条消息的时间，是不是有延迟特别大的 Pump，确认对应 Pump 正常运行。

  - 根据 Drainer event 和 Drainer execute latency 来判断是否下游同步太慢：

    * 如果 Drainer execute time 过大，则检查到目标库网络带宽和延迟，以及目标库状态。
    * 如果 Drainer execute time 不大，Drainer event 过小，则增加 work count 和 batch 进行重试。

  - 如果上面都不满足或者操作后没有改观，则报备开发人员 support@pingcap.com 进行处理。

12.10.7.2.3 警告级别报警项

警告级别的报警是对某一问题或错误的提醒。

`binlog_pump_write_binlog_rpc_duration_seconds_bucket`

- 报警规则：

  ```
  histogram_quantile(0.9, rate(binlog_pump_rpc_duration_seconds_bucket{method="WriteBinlog"}[5
  ↪ m]))> 1
  ```

- 规则描述：

  Pump 处理 TiDB 写 Binlog 请求耗时过大。

- 处理方法：

  - 确认磁盘性能压力，通过 `node exported` 查看 disk performance 监控。
  - 如果 `disk latency` 和 `util` 都很低，那么报备开发人员 support@pingcap.com 进行处理。

`binlog_pump_storage_write_binlog_duration_time_bucket`

- 报警规则：

  ```
  histogram_quantile(0.9, rate(binlog_pump_storage_write_binlog_duration_time_bucket{type="
  ↪ batch"}[5m]))> 1
  ```

- 规则描述：

  Pump 写本地 binlog 到本地盘的耗时。

- 处理方法：

  确认 Pump 本地盘情况，进行修复。

`binlog_pump_storage_available_size_less_than_20G`

- 报警规则：

  `binlog_pump_storage_storage_size_bytes{type="available"} < 20 * 1024 * 1024 * 1024`

- 规则描述：

  Pump 剩余可用磁盘空间不足 20 G。

- 处理方法：

  监控确认 Pump 的 `gc_tso` 是否正常。如果不正常，调整 Pump 的 GC 时间配置或者下线对应 Pump。

`binlog_drainer_checkpoint_tso_no_change_for_1m`

- 报警规则：

  `changes(binlog_drainer_checkpoint_tso[1m])< 1`

- 规则描述：

  Drainer 的 `checkpoint` 在 1 分钟内没有更新。

- 处理方法：

  确认所有非下线的 Pump 是否正常运行。

`binlog_drainer_execute_duration_time_more_than_10s`

- 报警规则：

  `histogram_quantile(0.9, rate(binlog_drainer_execute_duration_time_bucket[1m]))> 10`

- 规则描述：

  Drainer 同步到 TiDB 的事务耗时。如果这个值过大，会影响 Drainer 同步。

- 处理方法：

  - 查看 TiDB 集群的状态。
  - 查看 Drainer 日志或监控，如果是 DDL 操作导致了该问题，则忽略。

12.10.8 Reparo 使用文档

Reparo 是 TiDB Binlog 的一个配套工具，用于增量的恢复。使用 TiDB Binlog 中的 Drainer 将 binlog 按照 protobuf 格式输出到文件，通过这种方式来备份增量数据。当需要恢复增量数据时，使用 Reparo 解析文件中的 binlog，并将其应用到 TiDB / MySQL 中。

下载链接：tidb-binlog-cluster-latest-linux-amd64.tar.gz

12.10.8.1 Reparo 使用

12.10.8.1.1 命令行参数说明

```
Usage of Reparo:
-L string
    日志输出信息等级设置：debug, info, warn, error, fatal（默认值：info）。
-V 打印版本信息。
-c int
    同步下游的并发数，该值设置越高同步的吞吐性能越好（默认 16）。
-config string
    配置文件路径，如果指定了配置文件，Reparo 会首先读取配置文件的配置；
        ↪ 如果对应的配置在命令行参数里面也存在，Reparo
        ↪ 就会使用命令行参数的配置来覆盖配置文件里面的。
-data-dir string
    Drainer 输出的 protobuf 格式 binlog 文件的存储路径（默认值： data.drainer）。
-dest-type string
    下游服务类型。 取值为 print, mysql（默认值：print）。当值为 print 时，只做解析打印到标准输出
        ↪ ，不执行 SQL；如果为 mysql，则需要在配置文件内配置 host、port、user、password 等信息
        ↪ 。
-log-file string
    log 文件路径。
-log-rotate string
    log 文件切换频率，取值为 hour、day。
-start-datetime string
    用于指定开始恢复的时间点，格式为 “2006-01-02 15:04:05”。如果不设置该参数则从最早的 binlog
        ↪ 文件开始恢复。
-stop-datetime string
    用于指定结束恢复的时间点，格式同上。如果不设置该参数则恢复到最后一个 binlog 文件。
-safe-mode bool
    指定是否开启安全模式，开启后可支持反复同步。
-txn-batch int
    输出到下游数据库一个事务的 SQL 语句数量（默认 20）。
```

12.10.8.1.2 配置文件说明

```
### Drainer 输出的 protobuf 格式 binlog 文件的存储路径。
data-dir = "./data.drainer"

### 日志输出信息等级设置：debug, info, warn, error, fatal（默认值：info）。
log-level = "info"

### 使用 start-datetime 和 stop-datetime 来选择恢复指定时间范围内的 binlog，格式为 “2006-01-02
    ↪ 15:04:05”。
### start-datetime = ""
```

```
### stop-datetime = ""

### start-tso、stop-tso 分别对应 start-datetime 和 stop-datetime，也是用于恢复指定时间范围内的
    ↪ binlog，用 tso 的值来设置。如果已经设置了 start-datetime 和 stop-datetime，就不需要再设置
    ↪  start-tso 和 stop-tso。
### 在从全量或者上次增量位置继续同步时，start-tso 应当指定为全量 tso + 1 或者上次增量的 stop-tso
    ↪ + 1
### start-tso = 0
### stop-tso = 0

### 下游服务类型。 取值为 print, mysql（默认值：print）。当值为 print 时，只做解析打印到标准输出
    ↪ ，不执行 SQL；如果为 mysql，则需要在 [dest-db] 中配置 host、port、user、password 等信息。
dest-type = "mysql"

### 输出到下游数据库一个事务的 SQL 语句数量（默认 20）。
txn-batch = 20

### 同步下游的并发数，该值设置越高同步的吞吐性能越好（默认 16）。
worker-count = 16

### 安全模式配置。取值为 true 或 false（默认值：false）。当值为 true 时，Reparo 会将 update
    ↪ 语句拆分为 delete + replace 语句。
safe-mode = false

### replicate-do-db 和 replicate-do-table 用于指定恢复的库和表，replicate-do-db 的优先级高于
    ↪ replicate-do-table。支持使用正则表达式来配置，需要以 '~' 开始声明使用正则表达式。
### 注：replicate-do-db 和 replicate-do-table 使用方式与 Drainer 的使用方式一致。
### replicate-do-db = ["~^b.*","s1"]
### [[replicate-do-table]]
### db-name ="test"
### tbl-name = "log"
### [[replicate-do-table]]
### db-name ="test"
### tbl-name = "~^a.*"

### 如果 dest-type 设置为 mysql，需要配置 dest-db。
[dest-db]
host = "127.0.0.1"
port = 3309
user = "root"
password = ""
```

12.10.8.1.3 启动示例

```
./bin/reparo -config reparo.toml
```

注意：

- data-dir 用于指定 Drainer 输出的 binlog 文件目录。
- start-datatime 和 start-tso 效果一样，只是时间格式上的区别，用于指定开始恢复的时间点；如果不指定，则默认在第一个 binlog 文件开始恢复。
- stop-datetime 和 stop-tso 效果一样，只是时间格式上的区别，用于指定结束恢复的时间点；如果不指定，则恢复到最后一个 binlog 文件的结尾。
- dest-type 指定目标类型，取值为 `mysql`、`print`。当值为 `mysql` 时，可以恢复到 MySQL/TiDB 等使用或兼容 MySQL 协议的数据库，需要在配置下面的 [dest-db] 填写数据库信息；当取值为 `print` 的时候，只是打印 binlog 信息，通常用于 debug，以及查看 binlog 的内容，此时不需要填写 `[dest-db]`。
- replicate-do-db 用于指定恢复的库，不指定的话，则全部都恢复。
- replicate-do-table 用于指定要恢复的表，不指定的话，则全部都恢复。

12.10.9 binlogctl 工具

Binlog Control（以下简称 binlogctl）是 TiDB Binlog 的命令行工具，用于管理 TiDB Binlog 集群。

binlogctl 支持如下这些功能：

- 查看 Pump/Drainer 状态
- 暂停/下线 Pump/Drainer
- Pump/Drainer 异常状态处理

使用 binlogctl 的场景：

- 同步出现故障/检查运行情况，需要查看 Pump/Drainer 的状态
- 维护集群，需要暂停/下线 Pump/Drainer
- Pump/Drainer 异常退出，状态没有更新，或者状态不符合预期，对业务造成影响

12.10.9.1 binlogctl 下载

注意：

建议使用的 Control 工具版本与集群版本保持一致。

binlogctl 下载链接：

```
wget https://download.pingcap.org/tidb-{version}-linux-amd64.tar.gz &&
wget https://download.pingcap.org/tidb-{version}-linux-amd64.sha256
```

检查文件完整性，返回 ok 则正确：

```
sha256sum -c tidb-{version}-linux-amd64.sha256
```

检查文件完整性，返回 ok 则正确：

```
sha256sum -c tidb-enterprise-tools-latest-linux-amd64.sha256
```

12.10.9.2 binlogctl 使用说明

命令行参数：

```
Usage of binlogctl:
-V
输出 binlogctl 的版本信息
-cmd string
    命令模式，包括 "generate_meta"（已废弃）, "pumps", "drainers", "update-pump" ,"update-drainer
        ↪ ", "pause-pump", "pause-drainer", "offline-pump", "offline-drainer"
-data-dir string
    保存 Drainer 的 checkpoint 的文件的路径（默认 "binlog_position"）（已废弃）
-node-id string
    Pump/Drainer 的 ID
-pd-urls string
    PD 的地址，如果有多个，则用"," 连接（默认 "http://127.0.0.1:2379"）
-ssl-ca string
    SSL CAs 文件的路径
-ssl-cert string
        PEM 格式的 X509 认证文件的路径
-ssl-key string
        PEM 格式的 X509 key 文件的路径
-time-zone string
    如果设置时区，在 "generate_meta" 模式下会打印出获取到的 tso 对应的时间。例如 "Asia/Shanghai"
        ↪ 为 CST 时区，"Local" 为本地时区
-show-offline-nodes
    在用 `-cmd pumps` 或 `-cmd drainers` 命令时使用，这两个命令默认不显示 offline 的节点，
        ↪ 仅当明确指定 `-show-offline-nodes` 时会显示
```

命令示例：

- 查询所有的 Pump/Drainer 的状态：

  设置 cmd 为 pumps 或者 drainers 来查看所有 Pump 或者 Drainer 的状态。例如：

  ```
  bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd pumps
  ```

```
[2019/04/28 09:29:59.016 +00:00] [INFO] [nodes.go:48] ["query node"] [type=pump] [node="{
    ↪ NodeID: 1.1.1.1:8250, Addr: pump:8250, State: online, MaxCommitTS:
    ↪ 408012403141509121, UpdateTime: 2019-04-28 09:29:57 +0000 UTC}"]
```

```
bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd drainers
```

```
[2019/04/28 09:29:59.016 +00:00] [INFO] [nodes.go:48] ["query node"] [type=drainer] [node="{
    ↪ NodeID: 1.1.1.1:8249, Addr: 1.1.1.1:8249, State: online, MaxCommitTS:
    ↪ 408012403141509121, UpdateTime: 2019-04-28 09:29:57 +0000 UTC}"]
```

- 暂停/下线 Pump/Drainer

  binlogctl 提供以下命令暂停/下线服务：

cmd	说明	示例
pause-pump	暂停 Pump	`bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd pause-pump -node-id ip-127-0-0-1:8250`
pause-drainer	暂停 Drainer	`bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd pause-drainer -node-id ip-127-0-0-1:8249`
offline-pump	下线 Pump	`bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd offline-pump -node-id ip-127-0-0-1:8250`
offline-drainer	下线 Drainer	`bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd offline-drainer -node-id ip-127-0-0-1:8249`

  binlogctl 会发送 HTTP 请求给 Pump/Drainer，Pump/Drainer 收到命令后会主动执行对应的退出流程。

- 异常情况下修改 Pump/Drainer 的状态

  在服务正常运行以及符合流程的暂停、下线过程中，Pump/Drainer 的状态都是可以正确的。但是在一些异常情况下 Pump/Drainer 无法正确维护自己的状态，可能会影响数据同步任务，在这种情况下需要使用 binlogctl 修复状态信息。

  设置 cmd 为 `update-pump` 或者 `update-drainer` 来更新 Pump 或者 Drainer 的状态。Pump 和 Drainer 的状态可以为 paused 或者 offline。例如：

```
bin/binlogctl -pd-urls=http://127.0.0.1:2379 -cmd update-pump -node-id ip-127-0-0-1:8250 -
    ↪ state paused
```

> **注意：**
>
> Pump/Drainer 在正常运行过程中会定期在 PD 中更新自己的状态，而这条命令是直接去修改 Pump/Drainer 保存在 PD 中的状态，所以在 Pump/Drainer 服务正常的情况下使用这些命令是没有意义的。仅在 Pump/Drainer 服务异常的情况下使用，具体哪些场景下使用这条命令可以参考 FAQ。

12.10.10 Binlog Consumer Client 用户文档

目前 Drainer 提供了多种输出方式，包括 MySQL、TiDB、file 等。但是用户往往有一些自定义的需求，比如输出到 Elasticsearch、Hive 等，这些需求 Drainer 现在还没有实现，因此 Drainer 增加了输出到 Kafka 的功能，将 binlog 数据解析后按一定的格式再输出到 Kafka 中，用户编写代码从 Kafka 中读出数据再进行处理。

12.10.10.1 配置 Kafka Drainer

修改 Drainer 的配置文件，设置输出为 Kafka，相关配置如下：

```
[syncer]
db-type = "kafka"

[syncer.to]
### Kafka 地址
kafka-addrs = "127.0.0.1:9092"
### Kafka 版本号
kafka-version = "0.8.2.0"
```

12.10.10.2 自定义开发

12.10.10.2.1 数据格式

首先需要了解 Drainer 写入到 Kafka 中的数据格式：

```
// Column 保存列的数据，针对数据的类型，保存在对应的变量中
message Column {
  // 数据是否为 null
  optional bool is_null = 1 [ default = false ];
  // 保存 int 类型的数据
  optional int64 int64_value = 2;
  // 保存 uint、enum, set 类型的数据
  optional uint64 uint64_value = 3;
  // 保存 float、double 类型的数据
  optional double double_value = 4;
  // 保存 bit、blob、binary、json 类型的数据
  optional bytes bytes_value = 5;
  // 保存 date、time、decimal、text、char 类型的数据
  optional string string_value = 6;
}

// ColumnInfo 保存列的信息，包括列名、类型、是否为主键
message ColumnInfo {
  optional string name = 1 [ (gogoproto.nullable) = false ];
  // MySQL 中小写的列字段类型
  // https://dev.mysql.com/doc/refman/8.0/en/data-types.html
```

```
  // numeric 类型：int bigint smallint tinyint float double decimal bit
  // string 类型：text longtext mediumtext char tinytext varchar
  // blob longblob mediumblob binary tinyblob varbinary
  // enum set
  // json 类型：json
  optional string mysql_type = 2 [ (gogoproto.nullable) = false ];
  optional bool is_primary_key = 3 [ (gogoproto.nullable) = false ];
}

// Row 保存一行的具体数据
message Row { repeated Column columns = 1; }

// MutationType 表示 DML 的类型
enum MutationType {
  Insert = 0;
  Update = 1;
  Delete = 2;
}

// Table 包含一个表的数据变更
message Table {
  optional string schema_name = 1;
  optional string table_name = 2;
  repeated ColumnInfo column_info = 3;
  repeated TableMutation mutations = 4;
}

// TableMutation 保存一行数据的变更
message TableMutation {
  required MutationType type = 1;
  // 修改后的数据
  required Row row = 2;
  // 修改前的数据，只对 Update MutationType 有效
  optional Row change_row = 3;
}

// DMLData 保存一个事务所有的 DML 造成的数据变更
message DMLData {
  // `tables` 包含事务中所有表的数据变更
  repeated Table tables = 1;
}

// DDLData 保存 DDL 的信息
message DDLData {
  // 当前使用的数据库
```

```
  optional string schema_name = 1;
  // 相关表
  optional string table_name = 2;
  // `ddl_query` 是原始的 DDL 语句 query
  optional bytes ddl_query = 3;
}

// BinlogType 为 Binlog 的类型，分为 DML 和 DDL
enum BinlogType {
  DML = 0; //  Has `dml_data`
  DDL = 1; //  Has `ddl_query`
}

// Binlog 保存一个事务所有的变更，Kafka 中保存的数据为该结构数据序列化后的结果
message Binlog {
  optional BinlogType type = 1 [ (gogoproto.nullable) = false ];
  optional int64 commit_ts = 2 [ (gogoproto.nullable) = false ];
  optional DMLData dml_data = 3;
  optional DDLData ddl_data = 4;
}
```

查看数据格式的具体定义，参见 binlog.proto。

12.10.10.2.2 Driver

TiDB-Tools 项目提供了用于读取 Kafka 中 binlog 数据的 Driver，具有如下功能：

- 读取 Kafka 的数据
- 根据 commit ts 查找 binlog 在 kafka 中的储存位置

使用该 Driver 时，用户需要配置如下信息：

- KafkaAddr：Kafka 集群的地址
- CommitTS：从哪个 commit ts 开始读取 binlog
- Offset：从 Kafka 哪个 offset 开始读取，如果设置了 CommitTS 就不用配置该参数
- ClusterID：TiDB 集群的 cluster ID
- Topic: Kafka Topic 名称，如果 Topic 名称为空，将会使用 drainer `<ClusterID>_obinlog` 中的默认名称

用户以包的形式引用 Driver 的代码即可使用，可以参考 Driver 中提供的示例代码来学习如何使用 Driver 以及 binlog 数据的解析，目前提供了两个例子：

- 使用该 Driver 将数据同步到 MySQL，该示例包含将 binlog 转化为 SQL 的具体方法
- 使用该 Driver 将数据打印出来

Driver 项目地址：Binlog Slave Driver。

注意：

- 示例代码仅仅用于示范如何使用 Driver，如果需要用于生产环境需要优化代码。
- 目前仅提供了 golang 版本的 Driver 以及示例代码。如果需要使用其他语言，用户需要根据 binlog 的 proto 文件生成相应语言的代码文件，并自行开发程序读取 Kafka 中的 binlog 数据、解析数据、输出到下游。也欢迎用户优化 example 代码，以及提交其他语言的示例代码到 TiDB-Tools。

12.10.11 TiDB Binlog Relay Log

Drainer 同步 binlog 时会拆分上游的事务，并将拆分的事务并发同步到下游。在极端情况下，上游集群不可用并且 Drainer 异常退出后，下游集群（MySQL 或 TiDB）可能处于数据不一致的中间状态。在此场景下，Drainer 借助 relay log 可以确保将下游集群同步到一个一致的状态。

12.10.11.1 Drainer 同步时的一致性状态

下游集群达到一致的状态是指：下游集群的数据等同于上游设置了 `tidb_snapshot = ts` 的快照。

checkpoint 状态一致性是指：Drainer checkpoint 通过 `consistent` 保存了同步的一致性状态。Drainer 运行时 `consistent` 为 `false`，Drainer 正常退出后 `consistent` 更新为 `true`。

查询下游 checkpoint 表的示例如下：

```
mysql> select  * from tidb_binlog.checkpoint;
+--------------------+--------------------------------------------------------------------+
| clusterID          | checkPoint                                                         |
+--------------------+--------------------------------------------------------------------+
| 6791641053252586769 | {"consistent":false,"commitTS":414529105591271429,"ts-map":{}} |
+--------------------+--------------------------------------------------------------------+
```

12.10.11.2 工作原理

Drainer 开启 relay log 后会先将 binlog event 写到磁盘上，然后再同步给下游集群。如果上游集群不可用，Drainer 可以通过读取 relay log 把下游集群恢复到一个一致的状态。

注意：

若同时丢失 relay log 数据，该方法将不可用，不过这是概率极小的事件。此外可以使用 NFS 等网络文件系统来保证 relay log 的数据安全。

12.10.11.2.1 Drainer 从 relay log 消费 binlog 的触发场景

如果 Drainer 启动时无法连接到上游集群的 PD，并且探测到 checkpoint 的 `consistent = false`，此时会尝试读取 relay log，并将下游集群恢复到一致的状态。然后 Drainer 进程将 checkpoint 的 `consistent` 设置为 `true` 后主动退出。

12.10.11.2.2 Relay log 的清理（GC）机制

Drainer 在将数据同步到下游之前，会先将数据写入到 relay log 文件中。当一个 relay log 文件大小达到 10MB（默认）并且当前事务的 binlog 数据被写入完成后，Drainer 就会开始将数据写入到下一个 relay log 文件中。当 Drainer 将数据成功同步到下游后，就会自动清除当前正在写入的 relay log 文件以外其他已完成同步的 relay log 文件。

12.10.11.3 配置

在 Drainer 中添加以下配置来开启 relay log 功能：

```
[syncer.relay]
### 保存 relay log 的目录，空值表示不开启。
### 只有下游是 TiDB 或 MySQL 时该配置才有生效。
log-dir = "/dir/to/save/log"
### 单个 relay log 文件大小限制（单位：字节）。
### 超出该值后会将 binlog 数据写入到下一个 relay log 文件。
max-file-size = 10485760
```

12.10.12 集群间双向同步

> **警告：**
>
> 目前双向同步属于实验特性，尚未经过完备的测试，不建议在生产环境中使用该功能。

本文档介绍如何将一个 TiDB 集群的数据双向同步到另一个 TiDB 集群、双向同步的实现原理、如何开启双向同步、以及如何同步 DDL 操作。

12.10.12.1 使用场景

当用户需要在两个 TiDB 集群之间双向同步数据时，可使用 TiDB Binlog 进行操作。例如要将集群 A 的数据同步到集群 B，而且要将集群 B 的数据同步到集群 A。

> **注意：**
>
> 集群间双向同步的前提条件是，写入两个集群的数据必须保证无冲突，即在两个集群中，不会同时修改同一张表的同一主键和具有唯一索引的行。

使用场景示例图如下：

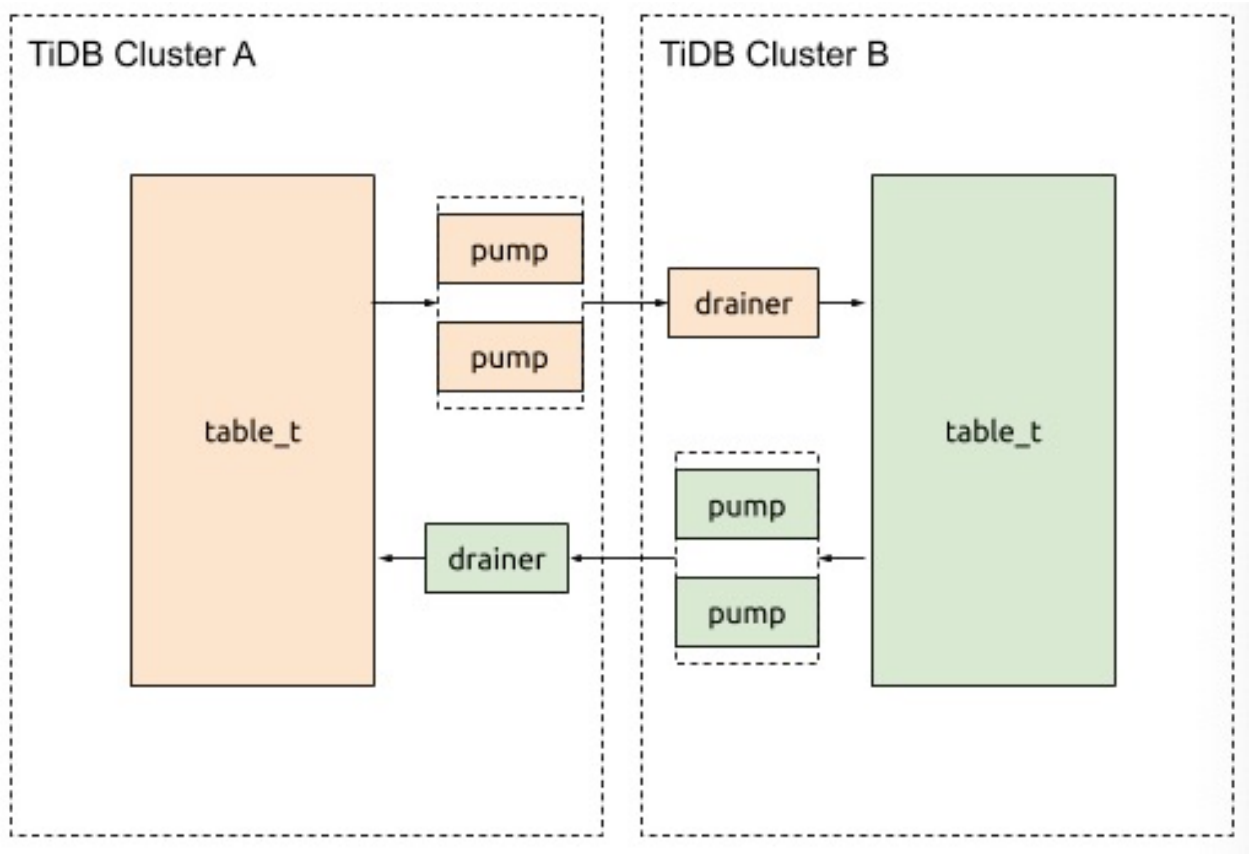


图 321: 使用场景示例图

12.10.12.2 实现原理

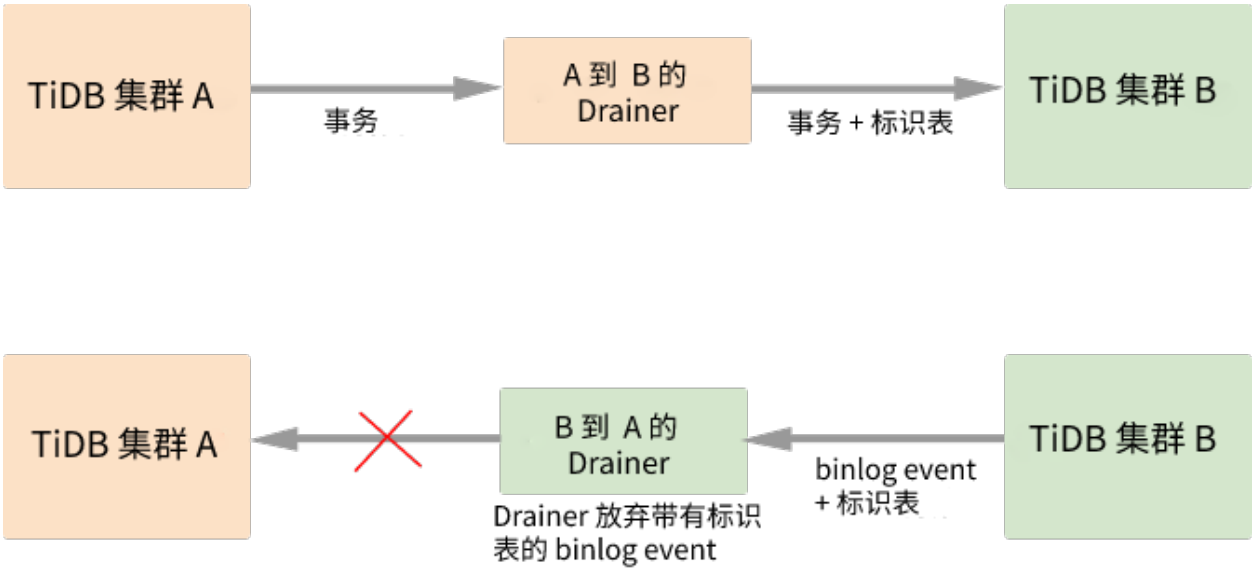


图 322: 原理示例图

在 A 和 B 两个集群间开启双向同步，则写入集群 A 的数据会同步到集群 B 中，然后这部分数据又会继续同步到集群 A，这样就会出现无限循环同步的情况。如上图所示，在同步数据的过程中 Drainer 对 binlog 加上标记，通过过滤掉有标记的 binlog 来避免循环同步。详细的实现流程如下：

1. 为两个集群分别启动 TiDB Binlog 同步程序。
2. 待同步的事务经过 A 的 Drainer 时，Drainer 为事务加入 `_drainer_repl_mark` 标识表，并在表中写入本次 DML event 更新，将事务同步至集群 B。
3. 集群 B 向集群 A 返回带有 `_drainer_repl_mark` 标识表的 binlog event。集群 B 的 Drainer 在解析该 binlog event 时发现带有 DML event 的标识表，放弃同步该 binlog event 到集群 A。

将集群 B 中的数据同步到集群 A 的流程与以上流程相同，两个集群可以互为上下游。

> 注意：
>
> - 更新 `_drainer_repl_mark` 标识表时，一定要有数据改动才会产生 binlog。
> - DDL 操作没有事务概念，因此采取单向同步的方案，见同步 DDL

Drainer 与下游的每个连接可以使用一个 ID 以避免冲突。`channel_id` 用来表示进行双向同步的一个通道。A 和 B 两个集群进行双向同步的配置应使用相同的值。

当有添加或者删除列时，要同步到下游的数据可能会出现多列或者少列的情况。Drainer 通过添加配置来允许这种情况，会忽略多了的列值或者给少了的列写入默认值。

12.10.12.3 标识表

`_drainer_repl_mark` 标识表的结构如下：

```
CREATE TABLE `_drainer_repl_mark` (
  `id` bigint(20) NOT NULL,
  `channel_id` bigint(20) NOT NULL DEFAULT '0',
  `val` bigint(20) DEFAULT '0',
  `channel_info` varchar(64) DEFAULT NULL,
  PRIMARY KEY (`id`,`channel_id`)
);
```

Drainer 使用如下 SQL 语句更新 _drainer_repl_mark 可保证数据改动，从而保证产生 binlog：

```
update drainer_repl_mark set val = val + 1 where id = ? && channel_id = ?;
```

12.10.12.4 同步 DDL

因为 Drainer 无法为 DDL 操作加入标识表，所以采用单向同步的方式来同步 DDL 操作。

比如集群 A 到集群 B 开启了 DDL 同步，则集群 B 到集群 A 会关闭 DDL 同步。即 DDL 操作全部在 A 上执行。

> 注意：
>
> DDL 操作无法在两个集群上同时执行。执行 DDL 时，若同时有 DML 操作或者 DML binlog 没同步完，会可能出现 DML 同步的上下游表结构不一致的情况。

12.10.12.5 配置并开启双向同步

若要在集群 A 和集群 B 间进行双向同步，假设统一在集群 A 上执行 DDL。在集群 A 到集群 B 的同步路径上，向 Drainer 添加以下配置：

```
[syncer]
loopback-control = true
channel-id = 1 # 互相同步的两个集群配置相同的 ID。
sync-ddl = true # 需要同步 DDL 操作

[syncer.to]
### 1 表示 SyncFullColumn，2 表示 SyncPartialColumn。
### 若设为 SyncPartialColumn，Drainer 会允许下游表结构比当前要同步的数据多或者少列
### 并且去掉 SQL mode 的 STRICT_TRANS_TABLES，来允许少列的情况，并插入零值到下游。
sync-mode = 2

### 忽略 checkpoint 表。
[[syncer.ignore-table]]
db-name = "tidb_binlog"
tbl-name = "checkpoint"
```

在集群 B 到集群 A 的同步路径上，向 Drainer 添加以下配置：

```
[syncer]
loopback-control = true
channel-id = 1 # 互相同步的两个集群配置相同的 ID。
sync-ddl = false  # 不需要同步 DDL 操作。

[syncer.to]
### 1 表示 SyncFullColumn，2 表示 SyncPartialColumn。
### 若设为 SyncPartialColumn，Drainer 会允许下游表结构比当前要同步的数据多或者少列
### 并且去掉 SQL mode 的 STRICT_TRANS_TABLES，来允许少列的情况，并插入零值到下游。
sync-mode = 2

### 忽略 checkpoint 表。
[[syncer.ignore-table]]
db-name = "tidb_binlog"
tbl-name = "checkpoint"
```

12.10.13 TiDB Binlog 术语表

本文档介绍 TiDB Binlog 相关术语。

12.10.13.1 Binlog

在 TiDB Binlog 中，Binlog 通常指 TiDB 写的二进制日志数据，也指 Drainer 写到 Kafka 或者文件的二进制日志数据，前者与后者的格式不同。此外，TiDB 的 binlog 格式与 MySQL 的 binlog 格式也不同。

12.10.13.2 Binlog event

TiDB 写的 DML Binlog 有 3 种 event，分别为：INSERT、UPDATE 和 DELETE。在 Drainer 监控面板上可以看到同步数据对应的不同 event 的个数。

12.10.13.3 Checkpoint

Checkpoint 指保存的断点信息，记录了 Drainer 同步到下游的 commit-ts。Drainer 重启时可以读取 checkpoint，并从对应的 commit-ts 开始同步数据到下游。

12.10.13.4 Safe mode

Safe mode 指增量复制过程中，当表结构中存在主键或唯一索引时，用于支持幂等导入 DML 的模式。

该模式将来自上游的 INSERT 改写为 REPLACE，将 UPDATE 改写为 DELETE 与 REPLACE，然后再向下游执行。在 Drainer 启动的前 5 分钟，会自动启动 Safe mode；另外也可以在配置文件中通过 safe-mode 参数手动开启，但该配置仅在下游是 MySQL 或 TiDB 时有效。

12.10.14 故障诊断

12.10.14.1 TiDB Binlog 故障诊断

本文总结了在 TiDB Binlog 的使用过程中遇到问题的诊断流程，并指引用户通过监控、状态、日志等信息查找相应的解决方案。

如果你在使用 TiDB Binlog 时出现了异常，请尝试以下方式排查问题：

1. 查看各个监控指标是否异常，参见TiDB Binlog 集群监控。
2. 使用binlogctl 工具查看各个 Pump、 Drainer 的状态是否有异常。
3. 查看 Pump、 Drainer 日志中是否有 ERROR、 WARN，并根据详细的日志信息初步判断问题原因。

通过以上方式定位到问题后，在FAQ 以及常见错误及修复 中查找解决方案，如果没有查找到解决方案或者提供的解决方案无法解决问题，请提交 issue 或者联系相关技术支持人员。

12.10.14.2 TiDB Binlog 常见错误修复

本文档介绍 TiDB Binlog 中常见的错误以及修复方法。

12.10.14.2.1 Drainer 同步数据到 Kafka 时报错 “kafka server: Message was too large, server rejected it to avoid allocation error”

报错原因：如果在 TiDB 中执行了大事务，则生成的 binlog 数据会比较大，可能超过了 Kafka 的消息大小限制。

解决方法：需要调整 Kafka 的配置参数，如下所示。

```
message.max.bytes=1073741824
replica.fetch.max.bytes=1073741824
fetch.message.max.bytes=1073741824
```

12.10.14.2.2 Pump 报错 “no space left on device”

报错原因：本地磁盘空间不足，Pump 无法正常写 binlog 数据。

解决方法：需要清理磁盘空间，然后重启 Pump。

12.10.14.2.3 Pump 启动时报错 “fail to notify all living drainer”

报错原因：Pump 启动时需要通知所有 Online 状态的 Drainer，如果通知失败则会打印该错误日志。

解决方法：可以使用binlogctl 工具查看所有 Drainer 的状态是否有异常，保证 Online 状态的 Drainer 都在正常工作。如果某个 Drainer 的状态和实际运行情况不一致，则使用 binlogctl 修改状态，然后再重启 Pump。

12.10.14.2.4 TiDB Binlog 同步中发现数据丢失

需要确认所有 TiDB 实例均开启了 TiDB Binlog，并且运行状态正常。如果集群版本大于 v3.0，可以使用 `curl {TiDB_IP}:{STATUS_PORT}/info/all` 命令确认所有 TiDB 实例上的 TiDB Binlog 状态。

12.10.14.2.5 当上游事务较大时，Pump 报错 `rpc error: code = ResourceExhausted desc = trying to send message larger than max (2191430008 vs. 2147483647)`

出现该错误的原因是 TiDB 发送给 Pump 的 gRPC message 超过限值。可以在启动 Pump 时通过指定 `max-message-size` 来调整 Pump 可接受 gRPC message 的最大大小。

12.10.14.2.6 Drainer 输出 file 格式的增量数据，数据有什么清理机制吗？数据会被删除吗？

- 在 v3.0.x 版本的 Drainer 中，file 格式的增量数据没有任何清理机制。
- 在 v4.0.x 版本中，有基于时间的数据清理机制，详见 Drainer 的 `retention-time` 配置项。

12.10.15 TiDB Binlog 常见问题

本文介绍 TiDB Binlog 使用过程中的常见问题及解决方案。

12.10.15.1 开启 binog 对 TiDB 的性能有何影响？

- 对于查询无影响。
- 对于有写入或更新数据的事务有一点性能影响。延迟上，在 Prewrite 阶段要并发写一条 p-binlog 成功后才可以提交事务，一般写 binlog 比 KV Prewrite 快，所以不会增加延迟。可以在 Pump 的监控面板看到写 binlog 的响应时间。

12.10.15.2 TiDB Binlog 的同步延迟一般为多少？

TiDB Binlog 的同步延迟为秒级别，在非业务高峰时延迟一般为 3 秒左右。

12.10.15.3 Drainer 同步下游 TiDB/MySQL 的帐号需要哪些权限？

Drainer 同步帐号需要有如下权限：

- Insert
- Update
- Delete
- Create
- Drop
- Alter
- Execute
- Index
- Select

12.10.15.4 Pump 磁盘快满了怎么办？

确认 GC 正常：

- 确认 pump 监控面板 gc_tso 时间是否与配置一致。

如 gc 正常以下调整可以降低单个 pump 需要的空间大小：

- 调整 pump GC 参数减少保留数据天数。
- 添加 pump 结点。

12.10.15.5 Drainer 同步中断怎么办？

使用以下 binlogctl 命令查看 Pump 状态是否正常，以及是否全部非 `offline` 状态的 Pump 都在正常运行。

```
binlogctl -cmd pumps
```

查看 Drainer 监控与日志是否有对应报错，根据具体问题进行处理。

12.10.15.6 Drainer 同步下游 TiDB/MySQL 慢怎么办？

特别关注以下监控项：

- 通过 Drainer 监控 drainer event，可以看到 Drainer 当前每秒同步 Insert/Update/Delete 事件到下游的速度。
- 通过 Drainer 监控 sql query time，可以看到 Drainer 在下游执行 SQL 的响应时间。

同步慢的可能原因与解决方案：

- 同步的数据库包含没有主键或者唯一索引的表，需要给表加上主键。
- Drainer 与下游之间延迟大，可以调大 Drainer `worker-count` 参数（跨机房同步建议将 Drainer 部署在下游）。
- 下游负载不高，可以尝试调大 Drainer `worker-count` 参数。

12.10.15.7 假如有一个 Pump crash 了会怎样？

Drainer 会因为获取不到这个 Pump 的数据没法同步数据到下游。如果这个 Pump 能恢复，Drainer 就能恢复同步。

如果 Pump 没法恢复，可采用以下方式进行处理：

1. 使用binlogctl 将该 Pump 状态修改为 `offline`（丢失这个 Pump 的数据）
2. Drainer 获取到的数据会丢失这个 Pump 上的数据，下游跟上游数据会不一致，需要重新做全量 + 增量同步。具体步骤如下：
    1. 停止当前 Drainer。
    2. 上游做全量备份。
    3. 清理掉下游数据，包括 checkpoint 表 `tidb_binlog.checkpoint`。
    4. 使用上游的全量备份恢复下游数据。
    5. 部署 Drainer，使用 `initialCommitTs=`{从全量备份获取快照的时间戳}。

12.10.15.8 什么是 checkpoint？

Checkpoint 记录了 Drainer 同步到下游的 commit-ts，Drainer 重启时可以读取 checkpoint 接着从对应 commit-ts 同步数据到下游。Drainer 日志 `["write save point"] [ts=411222863322546177]` 表示保存对应时间戳的 checkpoint。

下游类型不同，checkpoint 的保存方式也不同：

- 下游 MySQL/TiDB 保存在 `tidb_binlog.checkpoint` 表。
- 下游 kafka/file 保存在对应配置目录里的文件。

因为 kafka/file 的数据内容包含了 commit-ts，所以如果 checkpoint 丢失，可以消费下游最新的一条数据看写到下游数据的最新 commit-ts。

Drainer 启动的时候会去读取 checkpoint，如果读取不到，就会使用配置的 `initial-commit-ts` 做为初次启动开始的同步时间点。

12.10.15.9 Drainer 机器发生故障，下游数据还在，如何在新机器上重新部署 Drainer？

如果下游数据还在，只要保证能从对应 checkpoint 接着同步即可。

假如 checkpoint 还在，可采用以下方式进行处理：

1. 部署新的 Drainer 并启动即可（参考 checkpoint 介绍，Drainer 可以读取 checkpoint 接着同步）。
2. 使用binlogctl 将老的 Drainer 状态修改成 `offline`。

假如 checkpoint 不在，可以如下处理：

1. 获取之前 Drainer 的 checkpoint `commit-ts`，做为新部署 Drainer 的 `initial-commit-ts` 配置来部署新的 Drainer。
2. 使用binlogctl 将老的 Drainer 状态修改成 `offline`。

12.10.15.10 如何用全量 + binlog 备份文件来恢复一个集群？

1. 清理集群数据并使用全部备份恢复数据。
2. 使用 reparo 设置 `start-tso` = {全量备份文件快照时间戳 +1}，`end-ts` = 0（或者指定时间点），恢复到备份文件最新的数据。

12.10.15.11 主从同步开启 `ignore-error` 触发 critical error 后如何重新部署？

TiDB 配置开启 `ignore-error` 写 binlog 失败后触发 critical error 告警，后续都不会再写 binlog，所以会有 binlog 数据丢失。如果要恢复同步，需要如下处理：

1. 停止当前 Drainer。
2. 重启触发 critical error 的 `tidb-server` 实例重新开始写 binlog（触发 critical error 后不会再写 binlog 到 pump）。
3. 上游做全量备份。
4. 清理掉下游数据包括 checkpoint 表 `tidb_binlog.checkpoint`。
5. 使用上游的全量备份恢复下游。
6. 部署 Drainer，使用 `initialCommitTs=`{从全量备份获取快照的时间戳}。

12.10.15.12 同步时出现上游数据库支持但是下游数据库执行会出错的 DDL，应该怎么办？

1. 查看 drainer.log 日志，查找 `exec failed` 找到 Drainer 退出前最后一条执行失败的 DDL。
2. 将 DDL 改为下游兼容支持的版本，在下游数据库中手动执行。
3. 查看 drainer.log 日志，查找执行失败的 DDL 语句，可以查询到该 DDL 的 commit-ts。例如：

```
[2020/05/21 09:51:58.019 +08:00] [INFO] [syncer.go:398] ["add ddl item to syncer, you can
    ↪ add this commit ts to `ignore-txn-commit-ts` to skip this ddl if needed"] [sql="
    ↪ ALTER TABLE `test` ADD INDEX (`index1`)"] ["commit ts"=416815754209656834]。
```

4. 编辑 `drainer.toml` 配置文件，在 `ignore-txn-commit-ts` 项中添加该 commit-ts，重启 Drainer。

在绝大部分情况下，TiDB 和 MySQL 的语句都是兼容的。用户需要注意的是上下游的 `sql_mode` 应当保持一致。

12.10.15.13 在什么情况下暂停和下线 Pump/Drainer？

首先需要通过以下内容来了解 Pump/Drainer 的状态定义和启动、退出的流程。

暂停主要针对临时需要停止服务的场景，例如：

- 版本升级：停止进程后使用新的 binary 启动服务。
- 服务器维护：需要对服务器进行停机维护。退出进程，等维护完成后重启服务。

下线主要针对永久（或长时间）不再使用该服务的场景，例如：

- Pump 缩容：不再需要那么多 Pump 服务了，所以下线部分服务。
- 同步任务取消：不再需要将数据同步到某个下游，所以下线对应的 Drainer。
- 服务器迁移：需要将服务迁移到其他服务器。下线服务，在新的服务器上重新部署。

12.10.15.14 可以通过哪些方式暂停 Pump/Drainer？

- 直接 kill 进程。

> 注意：
> 不能使用 `kill -9` 命令，否则 Pump/Drainer 无法对信号进行处理。

- 如果 Pump/Drainer 运行在前台，则可以通过按下 Ctrl+C 来暂停。
- 使用 binlogctl 的 `pause-pump` 或 `pause-drainer` 命令。

12.10.15.15 可以使用 binlogctl 的 update-pump/update-drainer 命令来下线 Pump/Drainer 服务吗？

不可以。使用 update-pump/update-drainer 命令会直接修改 PD 中保存的状态信息，并且不会通知 Pump/Drainer 做相应的操作。使用不当时，可能会干扰数据同步，某些情况下还可能会造成数据不一致的严重后果。例如：

- 当 Pump 正常运行或者处于暂停状态时，如果使用 update-pump 将该 Pump 设置为 offline，那么 Drainer 会放弃获取处于 offline 状态的 Pump 的 binlog 数据，导致该 Pump 最新的 binlog 数据没有同步到 Drainer，造成上下游数据不一致。
- 当 Drainer 正常运行时，使用 update-drainer 命令将该 Drainer 设置为 offline。如果这时启动一个 Pump 节点，Pump 只会通知 online 状态的 Drainer，导致该 Drainer 没有及时获取到该 Pump 的 binlog 数据，造成上下游数据不一致。

12.10.15.16 可以使用 binlogctl 的 update-pump/update-drainer 命令来暂停 Pump/Drainer 服务吗？

不可以。update-pump/update-drainer 命令直接修改 PD 中保存的状态信息。执行这个命令并不会通知 Pump/Drainer 做相应的操作，而且使用不当会使数据同步中断，甚至造成数据丢失。

12.10.15.17 什么情况下使用 binlogctl 的 update-pump 命令设置 Pump 状态为 paused？

在某些异常情况下，Pump 没有正确维护自己的状态，实际上状态应该为 paused。这时可以使用 update-pump 对状态进行修正，例如：

- Pump 异常退出（可能由 panic 或者误操作执行 kill -9 命令直接 kill 掉进程导致），Pump 保存在 PD 中的状态仍然为 online。如果暂时不需要重启 Pump 恢复服务，可以使用 update-pump 把该 Pump 状态设置为 paused，避免对 TiDB 写 binlog 和 Drainer 获取 binlog 造成干扰。

12.10.15.18 什么情况下使用 binlogctl 的 update-drainer 命令设置 Drainer 状态为 paused？

在某些异常情况下，Drainer 没有正确维护自己的状态，，对数据同步造成了影响，实际上状态应该为 paused。这时可以使用 update-drainer 对状态进行修正，例如：

- Drainer 异常退出（出现 panic 直接退出进程，或者误操作执行 kill -9 命令直接 kill 掉进程），Drainer 保存在 PD 中的状态仍然为 online。当 Pump 启动时无法正常通知该 Drainer（报错 notify drainer ...），导致 Pump 无法正常运行。这个时候可以使用 update-drainer 将 Drainer 状态更新为 paused，再启动 Pump。

12.10.15.19 可以通过哪些方式下线 Pump/Drainer？

目前只可以使用 binlogctl 的 offline-pump 和 offline-drainer 命令来下线 Pump 和 Drainer。

12.10.15.20 什么情况下使用 binlogctl 的 update-pump 命令设置 Pump 状态为 offline?

> 警告：
>
> 仅在可以容忍 binlog 数据丢失、上下游数据不一致或者确认不再需要使用该 Pump 存储的 binlog 数据的情况下，才能使用 update-pump 修改 Pump 状态为 offline。

可以使用 update-pump 修改 Pump 状态为 offline 的情况有：

- 在某些情况下，Pump 异常退出进程，且无法恢复服务，同步就会中断。如果希望恢复同步且可以容忍部分 binlog 数据丢失，可以使用 update-pump 命令将该 Pump 状态设置为 offline，则 Drainer 会放弃拉取该 Pump 的 binlog 然后继续同步数据。
- 有从历史任务遗留下来且不再使用的 Pump 且进程已经退出（例如测试使用的服务），之后不再需要使用该服务，使用 update-pump 将该 Pump 设置为 offline。

在其他情况下一定要使用 offline-pump 命令让 Pump 走正常的下线处理流程。

12.10.15.21 Pump 进程已经退出，且状态为 paused，现在不想使用这个 Pump 节点了，能否用 binlogctl 的 update-pump 命令设置节点状态为 offline？

Pump 以 paused 状态退出进程时，不保证所有 binlog 数据被下游 Drainer 消费。所以这样做会有上下游数据不一致的风险。正确的做法是重新启动 Pump，然后使用 offline-pump 下线该 Pump。

12.10.15.22 什么情况下使用 binlogctl 的 update-drainer 命令设置 Drainer 状态为 offline？

- 有从历史任务遗留下来且不再使用的 Drainer 且进程已经退出（例如测试使用的服务），之后不再需要使用该服务，使用 update-drainer 将该 Drainer 设置为 offline。

12.10.15.23 可以使用 change pump、change drainer 等 SQL 操作来暂停或者下线 Pump/Drainer 服务吗？

目前还不支持。这种 SQL 操作会直接修改 PD 中保存的状态，在功能上等同于使用 binlogctl 的 update-pump、update-drainer 命令。如果需要暂停或者下线，仍然要使用 binlogctl。

12.10.15.24 TiDB 写入 binlog 失败导致 TiDB 卡住，日志中出现 listener stopped, waiting for manual stop

在 TiDB v3.0.12 以及之前，binlog 写入失败会导致 TiDB 报 fatal error。但是 TiDB 不会自动退出只是停止服务，看起来像服务卡住。TiDB 日志中可看到 listener stopped, waiting for manual stop。

遇到该问题需要根据具体情况判断是什么原因导致 binlog 写入失败。如果是 binlog 写入下游缓慢导致的，可以考虑扩容 Pump 或增加写 binlog 的超时时间。

TiDB 在 v3.0.13 版本中已优化此逻辑，写入 binlog 失败将使事务执行失败返回报错，而不会导致 TiDB 卡住。

12.10.15.25 TiDB 向 Pump 写入了重复的 binlog？

TiDB 在写入 binlog 失败或者超时的情况下，会重试将 binlog 写入到下一个可用的 Pump 节点直到写入成功。所以如果写入到某个 Pump 节点较慢，导致 TiDB 超时（默认 15s），此时 TiDB 判定写入失败并尝试写入下一个 Pump 节点。如果超时的 Pump 节点实际也写入成功，则会出现同一条 binlog 被写入到多个 Pump 节点。Drainer 在处理 binlog 的时候，会自动去重 TSO 相同的 binlog，所以这种重复的写入对下游无感知，不会对同步逻辑产生影响。

12.10.15.26 在使用全量 + 增量方式恢复的过程中，Reparo 中断了，可以使用日志里面最后一个 TSO 恢复同步吗？

可以。Reparo 不会在启动时自动开启 safe-mode 模式，需要手动操作：

1. Reparo 中断后，记录日志中最后一个 TSO，记为 `checkpoint-tso`。
2. 修改 Reparo 配置文件，将配置项 `start-tso` 设为 `checkpoint-tso + 1`，将 `stop-tso` 设为 `checkpoint-tso + 80,000,000,000`（大概是 `checkpoint-tso` 延后 5 分钟），将 `safe-mode` 设置为 `true`。启动 Reparo，Reparo 会将数据同步到 `stop-tso` 后自动停止。
3. Reparo 自动停止后，将 `start-tso` 设置为 `checkpoint tso + 80,000,000,001`，将 `stop-tso` 设置为 0，将 `safe-mode` 设为 `false`。启动 Reparo 继续同步。

12.11 TiCDC

12.11.1 TiCDC 简介

TiCDC 是一款通过拉取 TiKV 变更日志实现的 TiDB 增量数据同步工具，具有将数据还原到与上游任意 TSO 一致状态的能力，同时提供开放数据协议 (TiCDC Open Protocol)，支持其他系统订阅数据变更。

要快速了解 TiCDC 的基本原理和使用方法，建议先观看下面的培训视频（时长 28 分钟）。注意本视频只为学习参考，具体操作步骤和最新功能，请以文档内容为准。

12.11.1.1 TiCDC 架构

TiCDC 运行时是一种无状态节点，通过 PD 内部的 etcd 实现高可用。TiCDC 集群支持创建多个同步任务，向多个不同的下游进行数据同步。

TiCDC 的系统架构如下图所示：

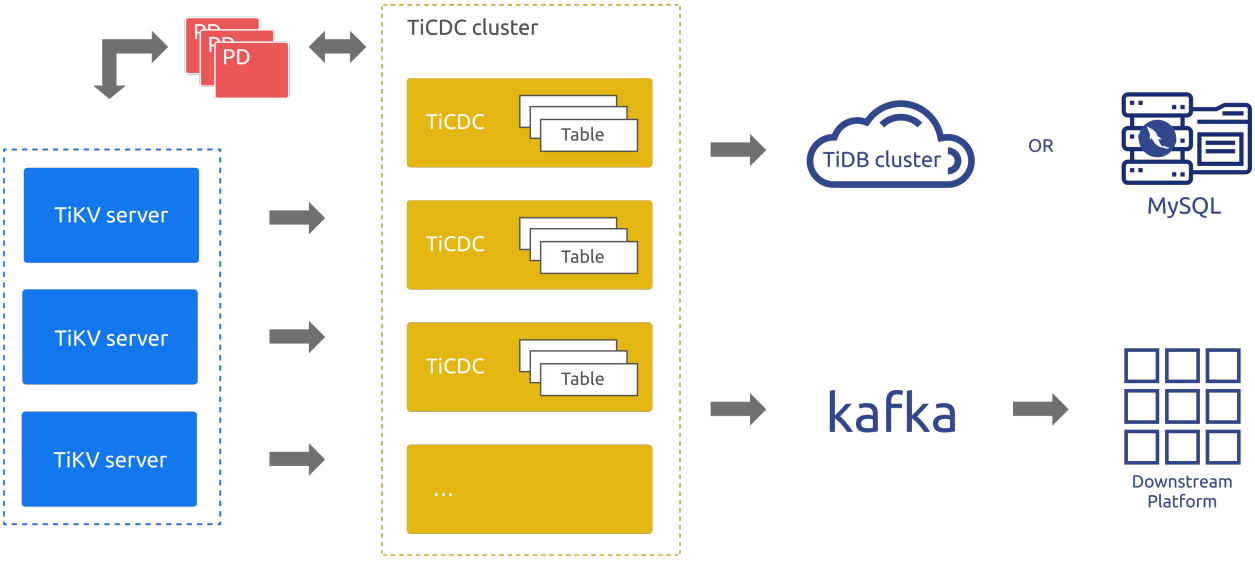


图 323: TiCDC architecture

12.11.1.1.1 系统角色

- TiKV CDC 组件：只输出 key-value (KV) change log。
  - 内部逻辑拼装 KV change log。
  - 提供输出 KV change log 的接口，发送数据包括实时 change log 和增量扫的 change log。
- `capture`：TiCDC 运行进程，多个 `capture` 组成一个 TiCDC 集群，负责 KV change log 的同步。
  - 每个 `capture` 负责拉取一部分 KV change log。
  - 对拉取的一个或多个 KV change log 进行排序。
  - 向下游还原事务或按照 TiCDC Open Protocol 进行输出。

12.11.1.2 同步功能介绍

本部分介绍 TiCDC 的同步功能。

12.11.1.2.1 sink 支持

目前 TiCDC sink 模块支持同步数据到以下下游：

- MySQL 协议兼容的数据库，提供最终一致性支持。
- 以 TiCDC Open Protocol 输出到 Kafka，可实现行级别有序、最终一致性或严格事务一致性三种一致性保证。

12.11.1.2.2 同步顺序保证和一致性保证

数据同步顺序

- TiCDC 对于所有的 DDL/DML 都能对外输出至少一次。
- TiCDC 在 TiKV/TiCDC 集群故障期间可能会重复发相同的 DDL/DML。对于重复的 DDL/DML：
  - MySQL sink 可以重复执行 DDL，对于在下游可重入的 DDL（譬如 truncate table）直接执行成功；对于在下游不可重入的 DDL（譬如 create table），执行失败，TiCDC 会忽略错误继续同步。
  - Kafka sink 会发送重复的消息，但重复消息不会破坏 Resolved Ts 的约束，用户可以在 Kafka 消费端进行过滤。

数据同步一致性

- MySQL sink
  - TiCDC 不拆分单表事务，保证单表事务的原子性。
  - TiCDC 不保证下游事务的执行顺序和上游完全一致。
  - TiCDC 以表为单位拆分跨表事务，不保证跨表事务的原子性。
  - TiCDC 保证单行的更新与上游更新顺序一致。
- Kafka sink
  - TiCDC 提供不同的数据分发策略，可以按照表、主键或 ts 等策略分发数据到不同 Kafka partition。
  - 不同分发策略下 consumer 的不同实现方式，可以实现不同级别的一致性，包括行级别有序、最终一致性或跨表事务一致性。
  - TiCDC 没有提供 Kafka 消费端实现，只提供了TiCDC 开放数据协议，用户可以依据该协议实现 Kafka 数据的消费端。

12.11.1.3 同步限制

使用 TiCDC 进行同步的时候，请注意以下相关限制要求以及暂不支持的场景。

12.11.1.3.1 有效索引的相关要求

TiCDC 只能同步至少存在一个有效索引的表，有效索引的定义如下：

- 主键 (PRIMARY KEY) 为有效索引。
- 同时满足下列条件的唯一索引 (UNIQUE INDEX) 为有效索引：
  - 索引中每一列在表结构中明确定义非空 (NOT NULL)。
  - 索引中不存在虚拟生成列 (VIRTUAL GENERATED COLUMNS)。

TiCDC 从 4.0.8 版本开始，可通过修改任务配置来同步没有有效索引的表，但在数据一致性的保证上有所减弱。具体使用方法和注意事项参考同步没有有效索引的表。

12.11.1.3.2 暂不支持的场景

目前 TiCDC 暂不支持的场景如下：

- 暂不支持单独使用 RawKV 的 TiKV 集群。
- 暂不支持在 TiDB 中创建 SEQUENCE 的 DDL 操作和 SEQUENCE 函数。在上游 TiDB 使用 SEQUENCE 时，TiCDC 将会忽略掉上游执行的 SEQUENCE DDL 操作/函数，但是使用 SEQUENCE 函数的 DML 操作可以正确地同步。
- 对上游存在较大事务的场景提供部分支持，详见TiCDC 是否支持同步大事务？有什么风险吗？

> 注意：
>
> 从 v5.3.0 版本起，TiCDC 不再支持环形同步功能。

12.11.1.4 TiCDC 安装和部署

要安装 TiCDC，可以选择随新集群一起部署，也可以对现有 TiDB 集群新增 TiCDC 组件。详请参阅TiCDC 安装部署。

12.11.1.5 TiCDC 集群管理和同步任务管理

目前支持使用 cdc cli 工具或 HTTP 接口来管理 TiCDC 集群状态和数据同步任务。详细操作见：

- 使用 cdc cli 工具来管理集群状态和数据同步
- 使用 OpenAPI 接口管理集群状态和数据同步

12.11.1.6 TiCDC 开放数据协议

TiCDC Open Protocol 是一种行级别的数据变更通知协议，为监控、缓存、全文索引、分析引擎、异构数据库的主从复制等提供数据源。TiCDC 遵循 TiCDC Open Protocol，向 MQ (Message Queue) 等第三方数据媒介复制 TiDB 的数据变更。详细信息参考TiCDC 开放数据协议。

12.11.1.7 兼容性问题

本文介绍了兼容性相关的问题。

12.11.1.7.1 使用 TiCDC v5.0.0-rc 版本的 cdc cli 工具操作 v4.0.x 集群导致不兼容问题

使用 TiCDC v5.0.0-rc 版本的 cdc cli 工具操作 v4.0.x 版本的 TiCDC 集群时，可能会遇到如下异常情况：

- 若 TiCDC 集群版本为 v4.0.8 或以下，使用 v5.0.0-rc 版本的 cdc cli 创建同步任务 changefeed 时，可能导致 TiCDC 集群陷入异常状态，导致同步卡住。
- 若 TiCDC 集群版本为 v4.0.9 或以上，使用 v5.0.0-rc 版本的 cdc cli 创建同步任务 changefeed，会导致 Old Value 和 Unified Sorter 特性被非预期地默认开启。

处理方案：

使用和 TiCDC 集群版本对应的 cdc 可执行文件进行如下操作：

1. 删除使用 v5.0.0-rc 版本创建的 changefeed，例如：`tiup cdc:v4.0.9 cli changefeed remove -c xxxx --pd=xxxxx --force`。
2. 如果 TiCDC 同步已经卡住，重启 TiCDC 集群，例如：`tiup cluster restart <cluster_name> -R cdc`。
3. 重新创建 changefeed，例如：`tiup cdc:v4.0.9 cli changefeed create --sink-uri=xxxx --pd=xxx`。

> 注意：
>
> 上述问题仅在 cdc cli 的版本是 v5.0.0-rc 时存在。未来其他 v5.0.x 版本的 cdc cli 可以兼容 v4.0.x 版本的集群。

12.11.1.7.2 sort-dir 及 data-dir 配置项的兼容性说明

sort-dir 配置项用于给 TiCDC 内部的排序器指定临时文件目录，其作用在各版本有过如下兼容性更改：

版本	sort-engine 的使用	说明	使用建议
版本	sort-engine 的使用	说明	使用建议
v4.0.11 及之前的 v4.0 版本，v5.0.0-rc	作为 changefeed 配置项，给 sorter 和 Sorter 指定临时文件目录	在这些版本中，sorter 和 sorter 均不是正式功能 (GA)，不推荐在生产环境中使用。如果有多个 changefeed 被配	不推荐在生产环境中使用 Unified Sorter

版本	sort-engine 的使用	说明	使用建议
v4.0.12，v4.0.13，v5.0.0 及 v5.0.1	作为 change-feed 配置项或 cdc server 配置项	在默认情况下 change-feed 的 sort-dir 配置不会生效，而 cdc server 的 sort-dir 配置默认为 /tmp/cdc_sort。建议生产环境下仅	需要通过 cdc server 命令行参数（或 TiUP）配置 sort-dir

版本	sort-engine 的使用	说明	使用建议
v4.0.14 及之后的 v4.0 版本，v5.0.3 及之后的 v5.0 版本，更新的版本	sort-dir 被弃用，建议配置 data-dir	data-dir 可以通过最新版本的 TiUP 进行配置。这些版本中 unified sorter 是默认开启的，升级时请确保 data-dir 已经被正	需要通过 cdc server 命令行参数（或 TiUP）配置 data-dir
1304			

	sort- ↪ engine ↪ 的	说	使 用 建
版本	使用	明	议

12.11.1.7.3 全局临时表兼容性说明

TiCDC 从 v5.3.0 开始支持全局临时表。

你需要使用 TiCDC v5.3.0 及以上版本同步全局临时表到下游。低于该版本，会导致表定义错误。

如果 TiCDC 的上游集群包含全局临时表，下游集群也必须是 TiDB 5.3.0 及以上版本，否则同步报错。

12.11.1.8 TiCDC 常见问题

在使用 TiCDC 过程中经常遇到的问题以及相对应的解决方案请参考TiCDC 常见问题。

12.11.2 TiCDC 安装部署

本文档介绍 TiCDC 集群的软硬件环境要求，以及如何安装部署 TiCDC 集群。你可以选择随新集群一起部署 TiCDC，也可以对原有 TiDB 集群新增 TiCDC 组件。通常推荐使用 TiUP 完成部署，如有特殊情况也可以用 binary 部署。

12.11.2.1 软件和硬件环境推荐配置

在生产环境中，TiCDC 的软件和硬件配置推荐如下：

Linux 操作系统平台	版本
Red Hat Enterprise Linux	7.3 及以上
CentOS	7.3 及以上

CPU	内存	硬盘类型	网络	TiCDC 集群实例数量（生产环境最低要求）
16 核 +	64 GB+	SSD	万兆网卡（2 块最佳）	2

更多信息参见TiDB 软件和硬件环境建议配置。

12.11.2.2 使用 TiUP 部署包含 TiCDC 组件的全新 TiDB 集群

在使用 TiUP 部署全新 TiDB 集群时，支持同时部署 TiCDC 组件。只需在 TiUP 启动 TiDB 集群时的配置文件中加入 `TiCDC` 部分即可，详细操作参考编辑初始化配置文件，具体可配置字段参考通过 TiUP 配置 `cdc_servers`。

12.11.2.3 使用 TiUP 在原有 TiDB 集群上新增 TiCDC 组件

目前也支持在原有 TiDB 集群上使用 TiUP 新增 TiCDC 组件，操作步骤如下：

1. 首先确认当前 TiDB 的版本支持 TiCDC，否则需要先升级 TiDB 集群至 4.0.0 rc.1 或更新版本。TiCDC 在 4.0.6 版本已经 GA，建议使用 4.0.6 及以后的版本。
2. 参考扩容 TiDB/TiKV/PD/TiCDC 节点章节对 TiCDC 进行部署。

12.11.2.4 使用 binary 在原有 TiDB 集群上新增 TiCDC 组件（不推荐）

假设 PD 集群有一个可以提供服务的 PD 节点（client URL 为 10.0.10.25:2379）。若要部署三个 TiCDC 节点，可以按照以下命令启动集群。只需要指定相同的 PD 地址，新启动的节点就可以自动加入 TiCDC 集群。

```
cdc server --pd=http://10.0.10.25:2379 --log-file=ticdc_1.log --addr=0.0.0.0:8301 --advertise-
    ↪ addr=127.0.0.1:8301
cdc server --pd=http://10.0.10.25:2379 --log-file=ticdc_2.log --addr=0.0.0.0:8302 --advertise-
    ↪ addr=127.0.0.1:8302
cdc server --pd=http://10.0.10.25:2379 --log-file=ticdc_3.log --addr=0.0.0.0:8303 --advertise-
    ↪ addr=127.0.0.1:8303
```

12.11.2.5 TiCDC cdc server 命令行参数说明

对于 cdc server 命令中可用选项解释如下：

- gc-ttl：TiCDC 在 PD 设置的服务级别 GC safepoint 的 TTL (Time To Live) 时长，和 TiCDC 同步任务所能够停滞的时长。单位为秒，默认值为 86400，即 24 小时。注意：TiCDC 同步任务的停滞会影响 TiCDC GC safepoint 的推进，即会影响上游 TiDB GC 的推进，详情可以参考TiCDC GC safepoint 的完整行为 。
- pd：PD client 的 URL。
- addr：TiCDC 的监听地址，提供服务的 HTTP API 查询地址和 Prometheus 查询地址。
- advertise-addr：TiCDC 对外访问地址。
- tz：TiCDC 服务使用的时区。TiCDC 在内部转换 timestamp 等时间数据类型和向下游同步数据时使用该时区，默认为进程运行本地时区。（注意如果同时指定 tz 参数和 sink-uri 中的 time-zone 参数，TiCDC 进程内部使用 tz 指定的时区，sink 向下游执行时使用 time-zone 指定的时区）
- log-file：TiCDC 进程运行日志的地址，默认为 cdc.log。
- log-level：TiCDC 进程运行时默认的日志级别，默认为 info。
- ca：TiCDC 使用的 CA 证书文件路径，PEM 格式，可选。
- cert：TiCDC 使用的证书文件路径，PEM 格式，可选。
- key：TiCDC 使用的证书密钥文件路径，PEM 格式，可选。
- config：可选项，表示 TiCDC 使用的配置文件地址。TiCDC 从 v5.0.0 开始支持该选项，TiUP 从 v1.4.0 开始支持在部署 TiCDC 时使用该配置。
- data-dir：指定 TiCDC 需要使用磁盘储存文件时使用的目录。目前 Unified Sorter 会使用该目录储存临时文件，请确保该目录所在设备可用空间充足。对于使用 TiUP 的用户，本选项可以通过配置 cdc_servers 小节中的 data_dir 来指定或默认使用 global 中 data_dir 路径。
- sort-dir：指定排序引擎使用的临时文件目录。自 TiDB v4.0.13, v5.0.3 和 v5.1.0 起已经无效，请不要使用。

12.11.3 TiCDC 运维操作及任务管理

本文档介绍如何通过 TiCDC 提供的命令行工具 cdc cli 管理 TiCDC 集群和同步任务，并介绍了如何使用 TiUP 来升级和修改 TiCDC 集群的配置。你也可以通过 HTTP 接口，即 TiCDC OpenAPI 来管理 TiCDC 集群和同步任务，详见TiCDC OpenAPI。

12.11.3.1 使用 TiUP 升级 TiCDC

本部分介绍如何使用 TiUP 来升级 TiCDC 集群。在以下例子中，假设需要将 TiCDC 组件和整个 TiDB 集群升级到 v5.4.0。

```
tiup update --self && \
tiup update --all && \
tiup cluster upgrade <cluster-name> v5.4.0
```

12.11.3.1.1 升级的注意事项

- TiCDC v4.0.2 对 changefeed 的配置做了调整，请参阅配置文件兼容注意事项。
- 升级期间遇到的问题及其解决办法，请参阅使用 TiUP 升级 TiDB。

12.11.3.2 使用 TiUP 修改 TiCDC 配置

本节介绍如何使用 TiUP 的`tiup cluster edit-config` 命令来修改 TiCDC 的配置。在以下例子中，假设需要把 TiCDC 的 `gc-ttl` 从默认值 86400 修改为 3600，即 1 小时。

首先执行以下命令。将 `<cluster-name>` 替换成实际的集群名。

```
tiup cluster edit-config <cluster-name>
```

执行以上命令之后，进入到 vi 编辑器页面，修改`server-configs` 下的 cdc 配置，如下所示：

```
server_configs:
  tidb: {}
  tikv: {}
  pd: {}
  tiflash: {}
  tiflash-learner: {}
  pump: {}
  drainer: {}
  cdc:
    gc-ttl: 3600
```

修改完毕后执行 `tiup cluster reload -R cdc` 命令重新加载配置。

12.11.3.3 使用加密传输 (TLS) 功能

请参阅为 TiDB 组件间通信开启加密传输。

12.11.3.4 使用 `cdc cli` 工具来管理集群状态和数据同步

本部分介绍如何使用 `cdc cli` 工具来管理集群状态和数据同步。`cdc cli` 是指通过 cdc binary 执行 `cli` 子命令。在以下描述中，通过 cdc binary 直接执行 `cli` 命令，PD 的监听 IP 地址为 `10.0.10.25`，端口为 2379。

> **注意：**
>
> PD 监听的 IP 和端口对应为 pd-server 启动时指定的 advertise-client-urls 参数。多个 pd-server 会包含多个该参数，用户可以指定其中任意一个或多个参数。例如 --pd=http://10.0.10.25:2379 或 --pd=http://10.0.10.25:2379,http://10.0.10.26:2379,http://10.0.10.27:2379。

如果你使用的 TiCDC 是用 TiUP 部署的，需要将以下命令中的 cdc cli 替换为 tiup ctl cdc。

12.11.3.4.1 管理 TiCDC 服务进程 (capture)

- 查询 capture 列表：

```
cdc cli capture list --pd=http://10.0.10.25:2379
```

```
[
  {
    "id": "806e3a1b-0e31-477f-9dd6-f3f2c570abdd",
    "is-owner": true,
    "address": "127.0.0.1:8300"
  },
  {
    "id": "ea2a4203-56fe-43a6-b442-7b295f458ebc",
    "is-owner": false,
    "address": "127.0.0.1:8301"
  }
]
```

  - id：服务进程的 ID。
  - is-owner：表示该服务进程是否为 owner 节点。
  - address：该服务进程对外提供接口的地址。

12.11.3.4.2 管理同步任务 (changefeed)

同步任务状态流转

同步任务状态标识了同步任务的运行情况。在 TiCDC 运行过程中，同步任务可能会运行出错、手动暂停、恢复，或达到指定的 TargetTs，这些行为都可以导致同步任务状态发生变化。本节描述 TiCDC 同步任务的各状态以及状态之间的流转关系。

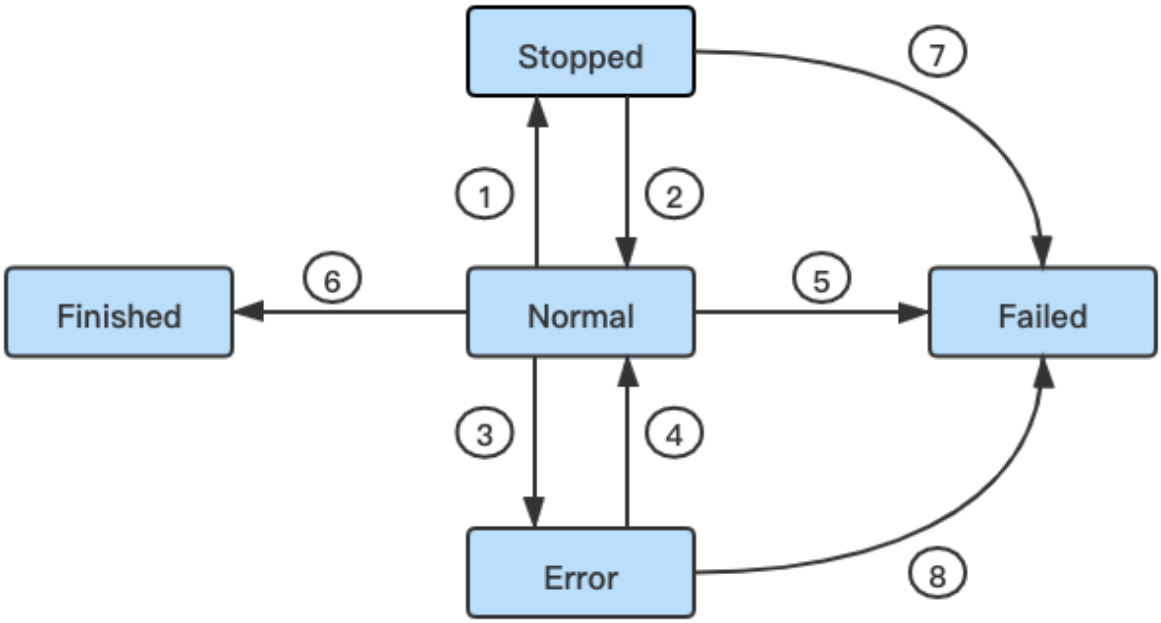


图 324: TiCDC state transfer

以上状态流转图中的状态说明如下：

- Normal：同步任务正常进行，checkpoint-ts 正常推进。
- Stopped：同步任务停止，由于用户手动暂停 (pause) changefeed。处于这个状态的 changefeed 会阻挡 GC 推进。
- Error：同步任务报错，由于某些可恢复的错误导致同步无法继续进行，处于这个状态的 changefeed 会不断尝试继续推进，直到状态转为 Normal。处于这个状态的 changefeed 会阻挡 GC 推进。
- Finished：同步任务完成，同步任务进度已经达到预设的 TargetTs。处于这个状态的 changefeed 不会阻挡 GC 推进。
- Failed：同步任务失败。由于发生了某些不可恢复的错误，导致同步无法继续进行，并且无法恢复。处于这个状态的 changefeed 不会阻挡 GC 推进。

以上状态流转图中的编号说明如下：

- ① 执行 `changefeed pause` 命令。
- ② 执行 `changefeed resume` 恢复同步任务。
- ③ changefeed 运行过程中发生可恢复的错误，自动进行恢复。
- ④ 执行 `changefeed resume` 恢复同步任务。
- ⑤ changefeed 运行过程中发生不可恢复的错误。
- ⑥ changefeed 已经进行到预设的 TargetTs，同步自动停止。
- ⑦ changefeed 停滞时间超过 `gc-ttl` 所指定的时长，不可被恢复。
- ⑧ changefeed 尝试自动恢复过程中发生不可恢复的错误。

创建同步任务

使用以下命令来创建同步任务：

```
cdc cli changefeed create --pd=http://10.0.10.25:2379 --sink-uri="mysql://root:123456@127
    ↪ .0.0.1:3306/" --changefeed-id="simple-replication-task" --sort-engine="unified"
```

```
Create changefeed successfully!
ID: simple-replication-task
Info: {"sink-uri":"mysql://root:123456@127.0.0.1:3306/","opts":{},"create-time":"2020-03-12T22
    ↪ :04:08.103600025+08:00","start-ts":415241823337054209,"target-ts":0,"admin-job-type":0,"
    ↪ sort-engine":"unified","sort-dir":".","config":{"case-sensitive":true,"filter":{"rules
    ↪ ":["*.*"],"ignore-txn-start-ts":null,"ddl-allow-list":null},"mounter":{"worker-num":16},"
    ↪ sink":{"dispatchers":null},"scheduler":{"type":"table-number","polling-time":-1}},"state
    ↪ ":"normal","history":null,"error":null}
```

- `--changefeed-id`：同步任务的 ID，格式需要符合正则表达式 `^[a-zA-Z0-9]+(\-[a-zA-Z0-9]+)*$`。如果不指定该 ID，TiCDC 会自动生成一个 UUID（version 4 格式）作为 ID。
- `--sink-uri`：同步任务下游的地址，需要按照以下格式进行配置，目前 scheme 支持 `mysql/tidb/kafka` ↪ `/pulsar`。

  ```
  [scheme]://[userinfo@][host]:[port][/path]?[query_parameters]
  ```

  URI 中包含特殊字符时，需要以 URL 编码对特殊字符进行处理。
- `--start-ts`：指定 changefeed 的开始 TSO。TiCDC 集群将从这个 TSO 开始拉取数据。默认为当前时间。
- `--target-ts`：指定 changefeed 的目标 TSO。TiCDC 集群拉取数据直到这个 TSO 停止。默认为空，即 TiCDC 不会自动停止。
- `--sort-engine`：指定 changefeed 使用的排序引擎。因 TiDB 和 TiKV 使用分布式架构，TiCDC 需要对数据变更记录进行排序后才能输出。该项支持 `unified`（默认）`/memory/file`：
  - `unified`：优先使用内存排序，内存不足时则自动使用硬盘暂存数据。该选项默认开启。
  - `memory`：在内存中进行排序。不建议使用，同步大量数据时易引发 OOM。
  - `file`：完全使用磁盘暂存数据。已经弃用，不建议在任何情况使用。
- `--config`：指定 changefeed 配置文件。
- `--sort-dir`: 用于指定排序器使用的临时文件目录。自 TiDB v4.0.13, v5.0.3 和 v5.1.0 起已经无效，请不要使用。

Sink URI 配置 `mysql/tidb`

配置样例如下所示：

```
--sink-uri="mysql://root:123456@127.0.0.1:3306/?worker-count=16&max-txn-row=5000"
```

URI 中可配置的参数如下：

参数	解析
root	下游数据库的用户名
123456	下游数据库密码
127.0.0.1	下游数据库的 IP
3306	下游数据的连接端口
worker-count	向下游执行 SQL 的并发度（可选，默认值为 16）
max-txn-row	向下游执行 SQL 的 batch 大小（可选，默认值为 256）
ssl-ca	连接下游 MySQL 实例所需的 CA 证书文件路径（可选）
ssl-cert	连接下游 MySQL 实例所需的证书文件路径（可选）
ssl-key	连接下游 MySQL 实例所需的证书密钥文件路径（可选）
time-zone	连接下游 MySQL 实例时使用的时区名称，从 v4.0.8 开始生效。（可选。如果不指定该参数，使用 TiCDC 服务进程的时区；如果指定该参数但使用空值，则表示连接 MySQL 时不指定时区，使用下游默认时区）

Sink URI 配置 kafka

配置样例如下所示：

```
--sink-uri="kafka://127.0.0.1:9092/topic-name?protocol=canal-json&kafka-version=2.4.0&partition-
    ↪ num=6&max-message-bytes=67108864&replication-factor=1"
```

URI 中可配置的的参数如下：

参数	解析
127.0.0.1	下游 Kafka 对外提供服务的 IP
9092	下游 Kafka 的连接端口
topic-name	变量，使用的 Kafka topic 名字
kafka-version	下游 Kafka 版本号（可选，默认值 2.4.0，目前支持的最低版本为 0.11.0.2，最高版本为 2.7.0。该值需要与下游 Kafka 的实际版本保持一致）
kafka-client-id	指定同步任务的 Kafka 客户端的 ID（可选，默认值为 TiCDC_sarama_producer_同步任务的 ID）
partition-num	下游 Kafka partition 数量（可选，不能大于实际 partition 数量，否则创建同步任务会失败，默认值 3）
max-message-bytes	每次向 Kafka broker 发送消息的最大数据量（可选，默认值 10MB）。从 v5.0.6 和 v4.0.6 开始，默认值分别从 64MB 和 256MB 调整至 10MB。
replication-factor	Kafka 消息保存副本数（可选，默认值 1）
protocol	输出到 Kafka 的消息协议，可选值有 canal-json、open-protocol、canal、avro、maxwell
auto-create-topic	当传入的 topic-name 在 Kafka 集群不存在时，TiCDC 是否要自动创建该 topic（可选，默认值 true）
enable-tidb-extension	当输出协议为 canal-json 时，如果该值为 true，TiCDC 会发送 Resolved 事件，并在 Kafka 消息中添加 TiDB 扩展字段（可选，默认值 false）
max-batch-size	从 v4.0.9 开始引入。当消息协议支持把多条变更记录输出至一条 Kafka 消息时，该参数用于指定这一条 Kafka 消息中变更记录的最多数量。目前，仅当 Kafka 消息的 protocol 为 open-protocol 时有效（可选，默认值 16）

参数	解析
ca	连接下游 Kafka 实例所需的 CA 证书文件路径（可选）
cert	连接下游 Kafka 实例所需的证书文件路径（可选）
key	连接下游 Kafka 实例所需的证书密钥文件路径（可选）
sasl-user	连接下游 Kafka 实例所需的 SASL/PLAIN 或 SASL/SCRAM 验证的用户名（authcid）（可选）
sasl-password	连接下游 Kafka 实例所需的 SASL/PLAIN 或 SASL/SCRAM 验证的密码（可选）
sasl-mechanism	连接下游 Kafka 实例所需的 SASL/PLAIN 或 SASL/SCRAM 验证的名称（可选）

最佳实践：

- TiCDC 推荐用户自行创建 Kafka Topic，你至少需要设置该 Topic 每次向 Kafka broker 发送消息的最大数据量和下游 Kafka partition 的数量。在创建 changefeed 的时候，这两项设置分别对应 `max-message-bytes` 和 `partition-num` 参数。
- 如果你在创建 changefeed 时，使用了尚未存在的 Topic，那么 TiCDC 会尝试使用 `partition-num` 和 `replication-factor` 参数自行创建 Topic。建议明确指定这两个参数。
- 在大多数情况下，建议使用 `canal-json` 协议。

> 注意：
>
> 当 `protocol` 为 `open-protocol` 时，TiCDC 会尽量避免产生长度超过 `max-message-bytes` 的消息。但如果单条数据变更记录需要超过 `max-message-bytes` 个字节来表示，为了避免静默失败，TiCDC 会试图输出这条消息并在日志中输出 Warning。

TiCDC 集成 Kafka Connect (Confluent Platform)

> 警告：
>
> 当前该功能为实验特性，不建议在生产环境中使用。

配置样例如下所示：

```
--sink-uri="kafka://127.0.0.1:9092/topic-name?protocol=canal-json&kafka-version=2.4.0&protocol=
    ↪ avro&partition-num=6&max-message-bytes=67108864&replication-factor=1"
--opts registry="http://127.0.0.1:8081"
```

如要使用 Confluent 提供的 data connectors 向关系型或非关系型数据库传输数据，应当选择 `avro` 协议，并在 `opts` 中提供 Confluent Schema Registry 的 URL。请注意，`avro` 协议和 Confluent 集成目前均为实验特性。

集成具体步骤详见TiDB 集成 Confluent Platform 快速上手指南。

Sink URI 配置 `pulsar`

配置样例如下所示：

```
--sink-uri="pulsar://127.0.0.1:6650/topic-name?connectionTimeout=2s"
```

URI 中可配置的的参数如下：

参数	解析
connectionTimeout	连接下游 Pulsar 的超时时间。可选参数，默认值为 30s。
operationTimeout	对下游 Pulsar 进行操作的超时时间（例如创建 topic）。可选参数，默认值为 30s。
tlsTrustCertsFilePath	连接下游 Pulsar 实例所需的 CA 证书文件路径（可选）
tlsAllowInsecureConnection	在开启 TLS 之后是否允许非加密连接（可选）
tlsValidateHostname	是否校验下游 Pulsar 证书中的 host name（可选）
maxConnectionsPerBroker	下游单个 Pulsar broker 最多允许的连接数（可选，默认值为 1）
auth.tls	使用 TLS 模式认证下游 Pulsar（可选，示例 auth=tls&auth.tlsCertFile=/path/to/cert&auth.tlsKeyFile=/path/to/key）
auth.token	使用 token 模式认证下游（可选，示例 auth=token&auth.token=secret-token 或者 auth=token&auth.file=path/to/secret-token-file）
name	TiCDC 中 Pulsar producer 名字（可选）
maxPendingMessages	Pending 消息队列的最大大小，例如，等待接收来自 Pulsar 的确认的消息（可选，默认值为 1000）
disableBatching	禁止自动批量发送消息（可选）
batchingMaxPublishDelay	设置发送消息的批处理时间（默认值为 10ms）
compressionType	设置发送消息时使用的压缩算法（可选 NONE，LZ4，ZLIB 和 ZSTD，默认值为 NONE）
hashingScheme	用于选择发送分区的哈希算法（可选 JavaStringHash 和 Murmur3，默认值为 JavaStringHash）
properties.*	在 TiCDC 中 Pulsar producer 上添加用户定义的属性（可选，示例 properties.location=Hangzhou）

更多关于 Pulsar 的参数解释，参见“pulsar-client-go ClientOptions 文档”和“pulsar-client-go ProducerOptions 文档”。

使用同步任务配置文件

如需设置更多同步任务的配置，比如指定同步单个数据表，请参阅同步任务配置文件描述。

使用配置文件创建同步任务的方法如下：

```
cdc cli changefeed create --pd=http://10.0.10.25:2379 --sink-uri="mysql://root:123456@127
    ↪ .0.0.1:3306/" --config changefeed.toml
```

其中 changefeed.toml 为同步任务的配置文件。

查询同步任务列表

使用以下命令来查询同步任务列表：

```
cdc cli changefeed list --pd=http://10.0.10.25:2379
```

```
[{
    "id": "simple-replication-task",
    "summary": {
      "state": "normal",
      "tso": 417886179132964865,
      "checkpoint": "2020-07-07 16:07:44.881",
      "error": null
    }
}]
```

- checkpoint 即为 TiCDC 已经将该时间点前的数据同步到了下游。
- state 为该同步任务的状态：
  - normal: 正常同步
  - stopped: 停止同步（手动暂停）
  - error: 停止同步（出错）
  - removed: 已删除任务（只在指定 --all 选项时才会显示该状态的任务。未指定时，可通过 query 查询该状态的任务）
  - finished: 任务已经同步到指定 target-ts，处于已完成状态（只在指定 --all 选项时才会显示该状态的任务。未指定时，可通过 query 查询该状态的任务）。

查询特定同步任务

使用 changefeed query 命令可以查询特定同步任务（对应某个同步任务的信息和状态），指定 --simple 或 -s 参数会简化输出，提供最基本的同步状态和 checkpoint 信息。不指定该参数会输出详细的任务配置、同步状态和同步表信息。

```
cdc cli changefeed query -s --pd=http://10.0.10.25:2379 --changefeed-id=simple-replication-task
```

```
{
 "state": "normal",
 "tso": 419035700154597378,
 "checkpoint": "2020-08-27 10:12:19.579",
 "error": null
}
```

以上命令中：

- state 代表当前 changefeed 的同步状态，各个状态必须和 changefeed list 中的状态相同。
- tso 代表当前 changefeed 中已经成功写入下游的最大事务 TSO。
- checkpoint 代表当前 changefeed 中已经成功写入下游的最大事务 TSO 对应的时间。
- error 记录当前 changefeed 是否有错误发生。

```
cdc cli changefeed query --pd=http://10.0.10.25:2379 --changefeed-id=simple-replication-task
```


```
{
  "info": {
    "sink-uri": "mysql://127.0.0.1:3306/?max-txn-row=20\u0026worker-number=4",
    "opts": {},
    "create-time": "2020-08-27T10:33:41.687983832+08:00",
    "start-ts": 419036036249681921,
    "target-ts": 0,
    "admin-job-type": 0,
    "sort-engine": "unified",
    "sort-dir": ".",
    "config": {
      "case-sensitive": true,
      "enable-old-value": false,
      "filter": {
        "rules": [
          "*.*"
        ],
        "ignore-txn-start-ts": null,
        "ddl-allow-list": null
      },
      "mounter": {
        "worker-num": 16
      },
      "sink": {
        "dispatchers": null,
      },
      "scheduler": {
        "type": "table-number",
        "polling-time": -1
      }
    },
    "state": "normal",
    "history": null,
    "error": null
  },
  "status": {
    "resolved-ts": 419036036249681921,
    "checkpoint-ts": 419036036249681921,
    "admin-job-type": 0
  },
  "count": 0,
  "task-status": [
```

```
    {
      "capture-id": "97173367-75dc-490c-ae2d-4e990f90da0f",
      "status": {
        "tables": {
          "47": {
            "start-ts": 419036036249681921
          }
        },
        "operation": null,
        "admin-job-type": 0
      }
    }
  ]
}
```

以上命令中：

- `info` 代表查询 changefeed 的同步配置。
- `status` 代表查询 changefeed 的同步状态信息。
  - `resolved-ts` 代表当前 changefeed 中已经成功从 TiKV 发送到 TiCDC 的最大事务 TS。
  - `checkpoint-ts` 代表当前 changefeed 中已经成功写入下游的最大事务 TS。
  - `admin-job-type` 代表一个 changefeed 的状态：
    * 0: 状态正常。
    * 1: 任务暂停，停止任务后所有同步 `processor` 会结束退出，同步任务的配置和同步状态都会保留，可以从 `checkpoint-ts` 恢复任务。
    * 2: 任务恢复，同步任务从 `checkpoint-ts` 继续同步。
    * 3: 任务已删除，接口请求后会结束所有同步 `processor`，并清理同步任务配置信息。同步状态保留，只提供查询，没有其他实际功能。
- `task-status` 代表查询 changefeed 所分配的各个同步子任务的状态信息。

12.11.3.4.3 停止同步任务

使用以下命令来停止同步任务：

```
cdc cli changefeed pause --pd=http://10.0.10.25:2379 --changefeed-id simple-replication-task
```

以上命令中：

- `--changefeed-id=uuid` 为需要操作的 changefeed ID。

12.11.3.4.4 恢复同步任务

使用以下命令恢复同步任务：

```
cdc cli changefeed resume --pd=http://10.0.10.25:2379 --changefeed-id simple-replication-task
```

以上命令中：

- `--changefeed-id=uuid` 为需要操作的 changefeed ID。

12.11.3.4.5 删除同步任务

使用以下命令删除同步任务：

```
cdc cli changefeed remove --pd=http://10.0.10.25:2379 --changefeed-id simple-replication-task
```

- `--changefeed-id=uuid` 为需要操作的 changefeed ID。

12.11.3.4.6 更新同步任务配置

TiCDC 从 4.0.4 开始支持非动态修改同步任务配置，修改 changefeed 配置需要按照 暂停任务 -> 修改配置 -> 恢复任务 的流程。

```
cdc cli changefeed pause -c test-cf --pd=http://10.0.10.25:2379
cdc cli changefeed update -c test-cf --pd=http://10.0.10.25:2379 --sink-uri="mysql
    ↪ ://127.0.0.1:3306/?max-txn-row=20&worker-number=8" --config=changefeed.toml
cdc cli changefeed resume -c test-cf --pd=http://10.0.10.25:2379
```

当前支持修改的配置包括：

- changefeed 的 `sink-uri`
- changefeed 配置文件及文件内所有配置
- changefeed 是否使用文件排序和排序目录
- changefeed 的 `target-ts`

12.11.3.4.7 管理同步子任务处理单元 (processor)

- 查询 `processor` 列表：

```
cdc cli processor list --pd=http://10.0.10.25:2379
```

```
[
        {
                "id": "9f84ff74-abf9-407f-a6e2-56aa35b33888",
                "capture-id": "b293999a-4168-4988-a4f4-35d9589b226b",
                "changefeed-id": "simple-replication-task"
        }
]
```

- 查询特定 `processor`，对应于某个节点处理的同步子任务信息和状态：

```
cdc cli processor query --pd=http://10.0.10.25:2379 --changefeed-id=simple-replication-task
    ↪ --capture-id=b293999a-4168-4988-a4f4-35d9589b226b
```

```
{
  "status": {
    "tables": {
      "56": {    # 56 表示同步表 id，对应 TiDB 中表的 tidb_table_id
        "start-ts": 417474117955485702
      }
    },
    "operation": null,
    "admin-job-type": 0
  },
  "position": {
    "checkpoint-ts": 417474143881789441,
    "resolved-ts": 417474143881789441,
    "count": 0
  }
}
```

以上命令中：

- status.tables 中每一个作为 key 的数字代表同步表的 id，对应 TiDB 中表的 tidb_table_id。
- resolved-ts 代表当前 processor 中已经排序数据的最大 TSO。
- checkpoint-ts 代表当前 processor 已经成功写入下游的事务的最大 TSO。

12.11.3.5 同步任务配置文件描述

本部分详细介绍了同步任务的配置。

```
### 指定配置文件中涉及的库名、表名是否为大小写敏感
### 该配置会同时影响 filter 和 sink 相关配置，默认为 true
case-sensitive = true

### 是否输出 old value，从 v4.0.5 开始支持，从 v5.0 开始默认为 true
enable-old-value = true

[filter]
### 忽略指定 start_ts 的事务
ignore-txn-start-ts = [1, 2]

### 过滤器规则
### 过滤规则语法：https://docs.pingcap.com/zh/tidb/stable/table-filter#表库过滤语法
rules = ['*.*', '!test.*']

[mounter]
### mounter 线程数，用于解码 TiKV 输出的数据
worker-num = 16
```

```
[sink]
### 对于 MQ 类的 Sink，可以通过 dispatchers 配置 event 分发器
### 支持 default、ts、rowid、table 四种分发器，分发规则如下：
### - default：有多个唯一索引（包括主键）时按照 table 模式分发；只有一个唯一索引（或主键）按照
    ↪ rowid 模式分发；如果开启了 old value 特性，按照 table 分发
### - ts：以行变更的 commitTs 做 Hash 计算并进行 event 分发
### - rowid：以表的主键或者唯一索引列名和列值做 Hash 计算并进行 event 分发
### - table：以表的 schema 名和 table 名做 Hash 计算并进行 event 分发
### matcher 的匹配语法和过滤器规则语法相同
dispatchers = [
    {matcher = ['test1.*', 'test2.*'], dispatcher = "ts"},
    {matcher = ['test3.*', 'test4.*'], dispatcher = "rowid"},
]
### 对于 MQ 类的 Sink，可以指定消息的协议格式
### 目前支持 canal-json、open-protocol、canal、avro 和 maxwell 五种协议。
protocol = "canal-json"
```

12.11.3.5.1 配置文件兼容性的注意事项

- TiCDC v4.0.0 中移除了 `ignore-txn-commit-ts`，添加了 `ignore-txn-start-ts`，使用 start_ts 过滤事务。
- TiCDC v4.0.2 中移除了 `db-dbs/db-tables/ignore-dbs/ignore-tables`，添加了 `rules`，使用新版的数据库和数据表过滤规则，详细语法参考表库过滤。

12.11.3.6 输出行变更的历史值从 v4.0.5 版本开始引入

> **警告：**
>
> 目前输出行变更历史值属于实验特性，尚未经过完备的测试，不建议在生产环境中使用该功能。

在默认配置下同步任务输出的 TiCDC Open Protocol 行变更数据只包含变更后的值，不包含变更前行的值，因此该输出数据不支持 TiDB 4.0 新的 Collation 框架，也不满足 TiCDC Open Protocol 的消费端使用行变更历史值的需求。

从 v4.0.5 开始，TiCDC 支持输出行变更数据的历史值。若要开启该特性，需要在 changefeed 的配置文件的根级别指定以下配置：

```
enable-old-value = true
```

开启该特性后，TiCDC Open Protocol 的输出格式参考TiCDC 开放数据协议 - Row Changed Event，使用 MySQL sink 时也会自动支持的 TiDB 4.0 新 Collation 特性。

12.11.3.7 同步没有有效索引的表

从 v4.0.8 开始，TiCDC 支持通过修改任务配置来同步没有有效索引的表。若要开启该特性，需要在 changefeed 配置文件的根级别进行如下指定：

```
enable-old-value = true
force-replicate = true
```

> **警告：**
>
> 对于没有有效索引的表，INSERT 和 REPLACE 等操作不具备可重入性，因此会有数据冗余的风险。TiCDC 在同步过程中只保证数据至少分发一次，因此开启该特性同步没有有效索引的表，一定会导致数据冗余出现。如果不能接受数据冗余，建议增加有效索引，譬如增加具有 AUTO RANDOM 属性的主键列。

12.11.3.8 Unified Sorter 功能

Unified Sorter 是 TiCDC 中的排序引擎功能，用于缓解以下场景造成的内存溢出问题：

- 如果 TiCDC 数据订阅任务的暂停中断时间长，其间积累了大量的增量更新数据需要同步。
- 从较早的时间点启动数据订阅任务，业务写入量大，积累了大量的更新数据需要同步。

对 v4.0.13 版本之后的 cdc cli 创建的 changefeed，默认开启 Unified Sorter。对 v4.0.13 版本前已经存在的 changefeed，则使用之前的配置。

要确定一个 changefeed 上是否开启了 Unified Sorter 功能，可执行以下示例命令查看（假设 PD 实例的 IP 地址为 http://10.0.10.25:2379）：

```
cdc cli --pd="http://10.0.10.25:2379" changefeed query --changefeed-id=simple-replication-task |
    ↪ grep 'sort-engine'
```

以上命令的返回结果中，如果 sort-engine 的值为 “unified”，则说明 Unified Sorter 已在该 changefeed 上开启。

> **注意：**
>
> - 如果服务器使用机械硬盘或其他有延迟或吞吐有瓶颈的存储设备，请谨慎开启 Unified Sorter。
> - Unified Sorter 默认使用 data_dir 储存临时文件。建议保证硬盘的空闲容量大于等于 500 GiB。对于生产环境，建议保证每个节点上的磁盘可用空间大于（业务允许的最大）checkpoint-ts 延迟 * 业务高峰上游写入流量。此外，如果在 changefeed 创建后预期需要同步大量历史数据，请确保每个节点的空闲容量大于等于要追赶的同步数据。
> - Unified Sorter 默认开启，如果您的服务器不符合以上条件，并希望关闭 Unified Sorter，请手动将 changefeed 的 sort-engine 设为 memory。
> - 如需在已使用 memory 排序的 changefeed 上开启 Unified Sorter，参见同步任务中断，尝试再次启动后 TiCDC 发生 OOM，如何处理回答中提供的方法。

12.11.3.9 灾难场景的最终一致性复制

从 v5.3.0 版本开始，TiCDC 开始提供灾难场景下的最终一致性复制能力。具体解决的场景是，当生产集群（即 TiCDC 同步的上游集群）发生灾难、且短时间内无法恢复对外提供服务，TiCDC 需要具备保证从集群数据一致性的能力，并允许业务快速的将流量切换至从集群，避免数据库长时间不可用而对业务造成影响。

该功能支持 TiCDC 将 TiDB 集群的增量数据复制到备用关系型数据库 TiDB/Aurora/MySQL/MariaDB，在 TiCDC 正常同步没有延迟的情况下，上游发生灾难后，可以在 5 分钟内将下游集群恢复到上游的某个 snapshot 状态，并且允许丢失的数据小于 30 分钟。即 RTO <= 5min，RPO <= 30min。

12.11.3.9.1 使用前提

- 准备好具有高可用的 S3 存储或 NFS 系统，用于存储 TiCDC 的实时增量数据备份文件，在上游发生灾难情况下该文件存储可以访问。
- TiCDC 对需要具备灾难场景最终一致性的 changefeed 开启该功能，开启方式是在 changefeed 配置文件中增加以下配置：

```
[consistent]
### 一致性级别，选项有：
### - none： 默认值，非灾难场景，只有在任务指定 finished-ts 情况下保证最终一致性。
### - eventual： 使用 redo log，提供上游灾难情况下的最终一致性。
level = "eventual"

### 单个 redo log 文件大小，单位 MiB，默认值 64，建议该值不超过 128。
max-log-size = 64

### 刷新或上传 redo log 至 S3 的间隔，单位毫秒，默认 1000，建议范围 500-2000。
flush-interval = 1000

### 存储 redo log 的形式，包括 nfs（NFS 目录），S3（上传至S3）
storage = "s3://logbucket/test-changefeed?endpoint=http://$S3_ENDPOINT/"
```

12.11.3.9.2 灾难恢复

当上游发生灾难后，需要通过 cdc redo 命令在下游手动恢复。恢复流程如下：

1. 确保 TiCDC 进程已经退出，防止在数据恢复过程中上游恢复服务，TiCDC 重新开始同步数据。
2. 使用 cdc binary 进行数据恢复，具体命令如下：

```
cdc redo apply --tmp-dir="/tmp/cdc/redo/apply" \
    --storage="s3://logbucket/test-changefeed?endpoint=http://10.0.10.25:24927/" \
    --sink-uri="mysql://normal:123456@10.0.10.55:3306/"
```

以上命令中：

- tmp-dir：指定用于下载 TiCDC 增量数据备份文件的临时目录。
- storage：指定存储 TiCDC 增量数据备份文件的地址，为 S3 或者 NFS 目录。
- sink-uri：恢复数据到的下游地址。scheme 仅支持 mysql。

12.11.4 TiCDC 常见问题和故障处理

本文档总结了在使用 TiCDC 过程中经常遇到的问题，给出合适的运维方法。本文档还总结了常见的运行故障，并给出相对应的解决方案。

> **注意：**
>
> 本文档 `cdc cli` 命令中指定 PD 地址为 `--pd=http://10.0.10.25:2379`，用户在使用时需根据实际地址进行替换。

12.11.4.1 TiCDC 创建任务时如何选择 start-ts？

首先需要理解同步任务的 `start-ts` 对应于上游 TiDB 集群的一个 TSO，同步任务会从这个 TSO 开始请求数据。所以同步任务的 `start-ts` 需要满足以下两个条件：

- `start-ts` 的值需要大于 TiDB 集群当前的 `tikv_gc_safe_point`，否则创建任务时会报错。
- 启动任务时，需要保证下游已经具有 `start-ts` 之前的所有数据。对于同步到消息队列等场景，如果不需要保证上下游数据的一致，可根据业务场景放宽此要求。

如果不指定 `start-ts` 或者指定 `start-ts=0`，在启动任务的时候会去 PD 获取一个当前 TSO，并从该 TSO 开始同步。

12.11.4.2 为什么 TiCDC 创建任务时提示部分表不能同步？

在使用 `cdc cli changefeed create` 创建同步任务时会检查上游表是否符合同步限制。如果存在表不满足同步限制，会提示 `some tables are not eligible to replicate` 并列出这些不满足的表。用户选择 `Y` 或 `y` 则会继续创建同步任务，并且同步过程中自动忽略这些表的所有更新。用户选择其他输入，则不会创建同步任务。

12.11.4.3 如何查看 TiCDC 同步任务的状态？

可以使用 `cdc cli` 查询同步任务的状态。例如：

```
cdc cli changefeed list --pd=http://10.0.10.25:2379
```

上述命令输出如下：

```
[{
    "id": "4e24dde6-53c1-40b6-badf-63620e4940dc",
    "summary": {
      "state": "normal",
      "tso": 417886179132964865,
      "checkpoint": "2020-07-07 16:07:44.881",
      "error": null
    }
}]
```

- `checkpoint`：即为 TiCDC 已经将该时间点前的数据同步到了下游。
- `state` 为该同步任务的状态：
  - `normal`：正常同步。
  - `stopped`：停止同步（手动暂停或出错）。
  - `removed`：已删除任务。

> 注意：
>
> 该功能在 TiCDC 4.0.3 版本引入。

12.11.4.4 TiCDC 同步任务出现中断

12.11.4.4.1 如何判断 TiCDC 同步任务出现中断？

- 通过 Grafana 检查同步任务的 `changefeed checkpoint` 监控项。注意选择正确的 `changefeed id`。如果该值不发生变化或者查看 `checkpoint lag` 是否不断增大，可能同步任务出现中断。
- 通过 Grafana 检查 `exit error count` 监控项，该监控项大于 0 代表同步任务出现错误。
- 通过 `cdc cli changefeed list` 和 `cdc cli changefeed query` 命令查看同步任务的状态信息。任务状态为 `stopped` 代表同步中断，`error` 项会包含具体的错误信息。任务出错后可以在 TiCDC server 日志中搜索 `error on running processor` 查看错误堆栈，帮助进一步排查问题。
- 部分极端异常情况下 TiCDC 出现服务重启，可以在 TiCDC server 日志中搜索 FATAL 级别的日志排查问题。

12.11.4.4.2 如何查看 TiCDC 同步任务是否被人为终止？

可以使用 `cdc cli` 查询同步任务是否被人为终止。例如：

```
cdc cli changefeed query --pd=http://10.0.10.25:2379 --changefeed-id 28c43ffc-2316-4f4f-a70b-
    ↪ d1a7c59ba79f
```

上述命令的输出中 `admin-job-type` 标志这个同步的任务的状态：

- 0: 任务进行中，没有被人为停止。
- 1: 任务暂停，停止任务后所有同步 `processor` 会结束退出，同步任务的配置和同步状态都会保留，可以从 `checkpoint-ts` 恢复任务。
- 2: 任务恢复，同步任务从 `checkpoint-ts` 继续同步。
- 3: 任务已删除，接口请求后会结束所有同步 `processor`，并清理同步任务配置信息。同步状态保留，只提供查询，没有其他实际功能。

12.11.4.4.3 如何处理 TiCDC 同步任务的中断？

目前已知可能发生的同步中断包括以下场景：

- 下游持续异常，TiCDC 多次重试后仍然失败。

  - 该场景下 TiCDC 会保存任务信息，由于 TiCDC 已经在 PD 中设置的 service GC safepoint，在 `gc-ttl` 的有效期内，同步任务 checkpoint 之后的数据不会被 TiKV GC 清理掉。
  - 处理方法：用户可以在下游恢复正常后，通过 HTTP 接口恢复同步任务。

- 因下游存在不兼容的 SQL 语句，导致同步不能继续。

  - 该场景下 TiCDC 会保存任务信息，由于 TiCDC 已经在 PD 中设置的 service GC safepoint，在 `gc-ttl` 的有效期内，同步任务 checkpoint 之后的数据不会被 TiKV GC 清理掉。
  - 处理方法：

    1. 用户需先通过 `cdc cli changefeed query` 查询同步任务状态信息，记录 `checkpoint-ts` 值。
    2. 使用新的任务配置文件，增加`ignore-txn-start-ts` 参数跳过指定 `start-ts` 对应的事务。
    3. 通过 HTTP API 停止旧的同步任务，使用 `cdc cli changefeed create` ，指定新的任务配置文件，指定 `start-ts` 为刚才记录的 `checkpoint-ts`，启动新的同步任务恢复同步。

- 在 v4.0.13 及之前的版本中 TiCDC 同步分区表存在问题，导致同步停止。

  - 该场景下 TiCDC 会保存任务信息，由于 TiCDC 已经在 PD 中设置的 service GC safepoint，在 `gc-ttl` 的有效期内，同步任务 checkpoint 之后的数据不会被 TiKV GC 清理掉。
  - 处理方法：

    1. 通过 `cdc cli changefeed pause -c <changefeed-id>` 暂停同步。
    2. 等待约一分钟后，通过 `cdc cli changefeed resume -c <changefeed-id>` 恢复同步。

12.11.4.4.4 同步任务中断，尝试再次启动后 TiCDC 发生 OOM，应该如何处理？

升级 TiDB 集群和 TiCDC 集群到最新版本。该 OOM 问题在 v4.0.14 及之后的 v4.0 版本，v5.0.2 及之后的 v5.0 版本，更新的版本上已得到缓解。

在这些版本上，可以开启 Unified Sorter 排序功能，该功能会在系统内存不足时使用磁盘进行排序。启用的方式是创建同步任务时在 `cdc cli` 内传入 `--sort-engine=unified`，使用示例如下：

```
cdc cli changefeed update -c <changefeed-id> --sort-engine="unified" --pd=http://10.0.10.25:2379
```

如果无法升级到上述版本，需要在之前的版本上开启 Unified Sorter，可以在创建同步任务时在 `cdc cli` 内传入 `--sort-engine=unified` 和 `--sort-dir=/path/to/sort_dir`，使用示例如下：

```
cdc cli changefeed update -c <changefeed-id> --sort-engine="unified" --sort-dir="/data/cdc/sort"
    ↪ --pd=http://10.0.10.25:2379
```

> 注意：
>
> - TiCDC 从 4.0.9 版本起支持 Unified Sorter 排序引擎。
> - TiCDC（4.0 发布版本）还不支持动态修改排序引擎设置前，请务必确保 changefeed 已经停止 (stopped)。
> - `sort-dir` 在不同版本之间有不同的行为，请参考sort-dir 及 data-dir 配置项的兼容性说明，谨慎配置。

- 目前 Unified Sorter 排序引擎为实验特性，在数据表较多 (>= 100) 时可能出现性能问题，影响同步速度，故不建议在生产环境中使用。开启 Unified Sorter 前请保证各 TiCDC 节点机器上有足够硬盘空间。如果积攒的数据总量有可能超过 1 TB，则不建议使用 TiCDC 进行同步。

12.11.4.5 TiCDC 的 gc-ttl 是什么？

从 TiDB v4.0.0-rc.1 版本起，PD 支持外部服务设置服务级别 GC safepoint。任何一个服务可以注册更新自己服务的 GC safepoint。PD 会保证任何晚于该 GC safepoint 的 KV 数据不会在 TiKV 中被 GC 清理掉。

在 TiCDC 中启用了这一功能，用来保证 TiCDC 在不可用、或同步任务中断情况下，可以在 TiKV 内保留 TiCDC 需要消费的数据不被 GC 清理掉。

启动 TiCDC server 时可以通过 `gc-ttl` 指定 GC safepoint 的 TTL，也可以通过 TiUP 修改 TiCDC 的 `gc-ttl`，默认值为 24 小时。在 TiCDC 中这个值有如下两重含义：

- 当 TiCDC 服务全部停止后，由 TiCDC 在 PD 所设置的 GC safepoint 保存的最长时间。
- TiCDC 中某个同步任务中断或者被手动停止时所能停滞的最长时间，若同步任务停滞时间超过 `gc-ttl` 所设置的值，那么该同步任务就会进入 `failed` 状态，无法被恢复，并且不会继续影响 GC safepoint 的推进。

以上第二种行为是在 TiCDC v4.0.13 版本及之后版本中新增的。目的是为了防止 TiCDC 中某个同步任务停滞时间过长，导致上游 TiKV 集群的 GC safepoint 长时间不推进，保留的旧数据版本过多，进而影响上游集群性能。

注意在某些应用场景中，比如使用 Dumpling/BR 全量同步后使用 TiCDC 接增量同步时，默认的 `gc-ttl` 为 24 小时可能无法满足需求。此时应该根据实际情况，在启动 TiCDC server 时指定 `gc-ttl` 的值。

12.11.4.6 TiCDC GC safepoint 的完整行为是什么

TiCDC 服务启动后，如果有任务开始同步，TiCDC owner 会根据所有同步任务最小的 checkpoint-ts 更新到 PD service GC safepoint，service GC safepoint 可以保证该时间点及之后的数据不被 GC 清理掉。如果 TiCDC 中某个同步任务中断、或者被用户主动停止，则该任务的 checkpoint-ts 不会再改变，PD 对应的 service GC safepoint 最终会停滞在该任务的 checkpoint-ts 处不再更新。

如果该同步任务停滞的时间超过了 `gc-ttl` 指定的时长，那么该同步任务就会进入 `failed` 状态，并且无法被恢复，PD 对应的 service GC safepoint 就会继续推进。

TiCDC 为 service GC safepoint 设置的存活有效期为 24 小时，即 TiCDC 服务中断 24 小时内恢复能保证数据不因 GC 而丢失。

12.11.4.7 如何处理 TiCDC 创建同步任务或同步到 MySQL 时遇到 `Error 1298: Unknown or incorrect time zone: 'UTC'` 错误？

这是因为下游 MySQL 没有加载时区，可以通过 mysql_tzinfo_to_sql 命令加载时区，加载后就可以正常创建任务或同步任务。

```
mysql_tzinfo_to_sql /usr/share/zoneinfo | mysql -u root mysql -p
```

显示类似于下面的输出则表示导入已经成功：

```
Enter password:
Warning: Unable to load '/usr/share/zoneinfo/iso3166.tab' as time zone. Skipping it.
Warning: Unable to load '/usr/share/zoneinfo/leap-seconds.list' as time zone. Skipping it.
Warning: Unable to load '/usr/share/zoneinfo/zone.tab' as time zone. Skipping it.
Warning: Unable to load '/usr/share/zoneinfo/zone1970.tab' as time zone. Skipping it.
```

如果下游是特殊的 MySQL 环境（某种公有云 RDS 或某些 MySQL 衍生版本等），使用上述方式导入时区失败，就需要通过 sink-uri 中的 `time-zone` 参数指定下游的 MySQL 时区。可以首先在 MySQL 中查询其使用的时区：

```
show variables like '%time_zone%';
```

```
+------------------+--------+
| Variable_name    | Value  |
+------------------+--------+
| system_time_zone | CST    |
| time_zone        | SYSTEM |
+------------------+--------+
```

然后在创建同步任务和创建 TiCDC 服务时使用该时区：

```
cdc cli changefeed create --sink-uri="mysql://root@127.0.0.1:3306/?time-zone=CST" --pd=http
    ↪ ://10.0.10.25:2379
```

> 注意：
>
> CST 可能是以下四个不同时区的缩写：
>
> - 美国中部时间：Central Standard Time (USA) UT-6:00
> - 澳大利亚中部时间：Central Standard Time (Australia) UT+9:30
> - 中国标准时间：China Standard Time UT+8:00
> - 古巴标准时间：Cuba Standard Time UT-4:00
>
> 在中国，CST 通常表示中国标准时间，使用时请注意甄别。

12.11.4.8 如何理解 TiCDC 时区和上下游数据库系统时区之间的关系？

	上游时区	TiCDC 时区	下游时区
配置方式	见时区支持	启动 ticdc server 时的 -- ↪ tz ↪ 参数	sink-uri 中的 time ↪ -zone ↪ 参数

	上游时区	TiCDC 时区	下游时区
说明	上游 TiDB 的时区，影响 timestamp 类型的 DML 操作和与 timestamp 类型列相关的 DDL 操作。	TiCDC 会将假设上游 TiDB 的时区和 TiCDC 时区配置相同，对 timestamp 类型的列进行相关处理。	下游 MySQL 将按照下游的时区设置对 DML 和 DDL 操作中包含的 timestamp 进行处理。

注意：

请谨慎设置 TiCDC server 的时区，因为该时区会用于时间类型的转换。上游时区、TiCDC 时区和下游时区应该保持一致。TiCDC server 时区使用的优先级如下：

- 最优先使用 `--tz` 传入的时区。

- 没有 `--tz` 参数，会尝试读取 TZ 环境变量设置的时区。
- 如果还没有 TZ 环境变量，会从 TiCDC server 运行机器的默认时区。

12.11.4.9 创建同步任务时，如果不指定 `--config` 配置文件，TiCDC 的默认的行为是什么？

在使用 `cdc cli changefeed create` 命令时如果不指定 `--config` 参数，TiCDC 会按照以下默认行为创建同步任务：

- 同步所有的非系统表
- 开启 old value 功能
- 不同步不包含有效索引的表

12.11.4.10 如何处理升级 TiCDC 后配置文件不兼容的问题？

请参阅配置文件兼容注意事项。

12.11.4.11 TiCDC 是否支持输出 Canal 格式的变更数据？

支持。要开启 Canal 格式输出，只需在 `--sink-uri` 中指定 protocol 为 `canal` 即可，例如：

```
cdc cli changefeed create --pd=http://10.0.10.25:2379 --sink-uri="kafka://127.0.0.1:9092/cdc-test
    ↪ ?kafka-version=2.4.0&protocol=canal" --config changefeed.toml
```

注意：

- 该功能在 TiCDC 4.0.2 版本引入。
- 目前 TiCDC 仅支持将 Canal 格式的变更数据输出到 MQ 类的 Sink（例如：Kafka，Pulsar）。

更多信息请参考创建同步任务。

12.11.4.12 为什么 TiCDC 到 Kafka 的同步任务延时越来越大？

- 请参考如何查看 TiCDC 同步任务的状态？检查下同步任务的状态是否正常。
- 请适当调整 Kafka 的以下参数：
  - `message.max.bytes`，将 Kafka 的 `server.properties` 中该参数调大到 1073741824 (1 GB)。
  - `replica.fetch.max.bytes`，将 Kafka 的 `server.properties` 中该参数调大到 1073741824 (1 GB)。
  - `fetch.message.max.bytes`，适当调大 `consumer.properties` 中该参数，确保大于 `message.max.bytes`。

12.11.4.13 TiCDC 把数据同步到 Kafka 时，是把一个事务内的所有变更都写到一个消息中吗？如果不是，是根据什么划分的？

不是，根据配置的分发策略不同，有不同的划分方式，包括 `default`、`row id`、`table`、`ts`。更多请参考同步任务配置文件描述。

12.11.4.14 TiCDC 把数据同步到 Kafka 时，能在 TiDB 中控制单条消息大小的上限吗？

可以通过 `max-message-bytes` 控制每次向 Kafka broker 发送消息的最大数据量（可选，默认值 10MB）；通过 `max-batch-size` 参数指定每条 kafka 消息中变更记录的最大数量，目前仅对 Kafka 的 `protocol` 为 `open-protocol` 时有效（可选，默认值 16）。

12.11.4.15 TiCDC 把数据同步到 Kafka 时，一条消息中会不会包含多种数据变更？

会，一条消息中可能出现多个 `update` 或 `delete`，`update` 和 `delete` 也有可能同时存在。

12.11.4.16 TiCDC 把数据同步到 Kafka 时，如何查看 TiCDC Open protocol 输出变更数据中的时间戳、表名和库名？

这些信息包含在 Kafka 消息的 Key 中，比如：

```
{
    "ts":<TS>,
    "scm":<Schema Name>,
    "tbl":<Table Name>,
    "t":1
}
```

更多信息请参考Open protocol Event 格式定义

12.11.4.17 TiCDC 把数据同步到 Kafka 时，如何确定一条消息中包含的数据变更发生在哪个时间点？

把 Kafka 消息的 Key 中的 `ts` 右移 18 位即得 unix timestamp。

12.11.4.18 TiCDC Open protocol 如何标示 null 值？

Open protocol 的输出中 type = 6 即为 null，比如：

类型	Code	输出示例	说明
Null	6	`{"t":6,"v":null}`	

更多信息请参考Open protocol Event 格式定义。

12.11.4.19 TiCDC 启动任务的 start-ts 时间戳与当前时间差距较大，任务执行过程中同步中断，出现错误 `[CDC:ErrBufferReachLimit]`，怎么办？

自 v4.0.9 起可以尝试开启 unified sorter 特性进行同步；或者使用 BR 工具进行一次增量备份和恢复，然后从新

的时间点开启 TiCDC 同步任务。TiCDC 将会在后续版本中对该问题进行优化。

12.11.4.20 如何区分 TiCDC Open Protocol 中的 Row Changed Event 是 INSERT 事件还是 UPDATE 事件？

如果没有开启 Old Value 功能，用户无法区分 TiCDC Open Protocol 中的 Row Changed Event 是 INSERT 事件还是 UPDATE 事件。如果开启了 Old Value 功能，则可以通过事件中的字段判断事件类型：

- 如果同时存在 "p" 和 "u" 字段为 UPDATE 事件
- 如果只存在 "u" 字段则为 INSERT 事件
- 如果只存在 "d" 字段则为 DELETE 事件

更多信息请参考Open protocol Row Changed Event 格式定义。

12.11.4.21 TiCDC 占用多少 PD 的存储空间

TiCDC 使用 PD 内部的 etcd 来存储元数据并定期更新。因为 etcd 的多版本并发控制 (MVCC) 以及 PD 默认的 compaction 间隔是 1 小时，TiCDC 占用的 PD 存储空间与 1 小时内元数据的版本数量成正比。在 v4.0.5、v4.0.6、v4.0.7 三个版本中 TiCDC 存在元数据写入频繁的问题，如果 1 小时内有 1000 张表创建或调度，就会用尽 etcd 的存储空间，出现 `etcdserver: mvcc: database space exceeded` 错误。出现这种错误后需要清理 etcd 存储空间，参考 etcd maintaince space-quota。推荐使用这三个版本的用户升级到 v4.0.9 及以后版本。

12.11.4.22 TiCDC 支持同步大事务吗？有什么风险吗？

TiCDC 对大事务（大小超过 5 GB）提供部分支持，根据场景不同可能存在以下风险：

- 当 TiCDC 内部处理能力不足时，可能出现同步任务报错 `ErrBufferReachLimit`。
- 当 TiCDC 内部处理能力不足或 TiCDC 下游吞吐能力不足时，可能出现内存溢出 (OOM)。

当遇到上述错误时，建议将包含大事务部分的增量数据通过 BR 进行增量恢复，具体操作如下：

1. 记录因为大事务而终止的 changefeed 的 `checkpoint-ts`，将这个 TSO 作为 BR 增量备份的 `--lastbackupts`，并执行增量备份。
2. 增量备份结束后，可以在 BR 日志输出中找到类似 `["Full backup Failed summary : total backup ranges: 0, total success: 0, total failed: 0"] [BackupTS=421758868510212097]` 的日志，记录其中的 BackupTS。
3. 执行增量恢复。
4. 建立一个新的 changefeed，从 BackupTS 开始同步任务。
5. 删除旧的 changefeed。

12.11.4.23 当 changefeed 的下游为类 MySQL 数据库时，TiCDC 执行了一个耗时较长的 DDL 语句，阻塞了所有其他 changefeed，应该怎样处理？

1. 首先暂停执行耗时较长的 DDL 的 changefeed。此时可以观察到，这个 changefeed 暂停后，其他的 changefeed 不再阻塞了。
2. 在 TiCDC log 中搜寻 `apply job` 字段，确认耗时较长的 DDL 的 `start-ts`。
3. 手动在下游执行该 DDL 语句，执行完毕后进行下面的操作。
4. 修改 changefeed 配置，将上述 `start-ts` 添加到 `ignore-txn-start-ts` 配置项中。
5. 恢复被暂停的 changefeed。

12.11.4.24 TiCDC 集群升级到 v4.0.8 之后，changefeed 报错 [CDC:ErrKafkaInvalidConfig]Canal requires old value to be enabled，为什么？

自 v4.0.8 起，如果 changefeed 使用 canal-json、canal 或者 maxwell 协议输出，TiCDC 会自动开启 Old Value 功能。但是，当 TiCDC 是从较旧版本升级到 v4.0.8 或以上版本时，在 changefeed 使用 canal-json、canal 或 maxwell 协议的同时 TiCDC 的 Old Value 功能会被禁用。此时，会出现该报错。可以按照以下步骤解决该报错：

1. 将 changefeed 配置文件中 enable-old-value 的值设为 true。
2. 使用 cdc cli changefeed pause 暂停同步任务。

```
cdc cli changefeed pause -c test-cf --pd=http://10.0.10.25:2379
```

3. 使用 cdc cli changefeed update 更新原有 changefeed 的配置。

```
cdc cli changefeed update -c test-cf --pd=http://10.0.10.25:2379 --sink-uri="mysql
    ↪ ://127.0.0.1:3306/?max-txn-row=20&worker-number=8" --config=changefeed.toml
```

4. 使用 cdc cli changefeed resume 恢复同步任务。

```
cdc cli changefeed resume -c test-cf --pd=http://10.0.10.25:2379
```

12.11.4.25 使用 TiCDC 创建 changefeed 时报错 [tikv:9006]GC life time is shorter than transaction duration, transaction starts at xx, GC safe point is yy

解决方案：需要执行 pd-ctl service-gc-safepoint --pd <pd-addrs> 命令查询当前的 GC safepoint 与 service GC safepoint。如果 GC safepoint 小于 TiCDC changefeed 同步任务的开始时间戳 start-ts，则用户可以直接在 cdc cli create changefeed 命令后加上 --disable-gc-check 参数创建 changefeed。

如果 pd-ctl service-gc-safepoint --pd <pd-addrs> 的结果中没有 gc_worker service_id：

- 如果 PD 的版本 <= v4.0.8，详见 PD issue #3128。
- 如果 PD 是由 v4.0.8 或更低版本滚动升级到新版，详见 PD issue #3366。
- 对于其他情况，请将上述命令执行结果反馈到 AskTUG 论坛。

12.11.4.26 使用 TiCDC 创建同步任务时将 enable-old-value 设置为 true 后，为什么上游的 INSERT/UPDATE 语句经 TiCDC 同步到下游后变为了 REPLACE INTO？

TiCDC 创建 changefeed 时会默认指定 safe-mode 为 true，从而为上游的 INSERT/UPDATE 语句生成 REPLACE INTO 的执行语句。

目前用户暂时无法修改 safe-mode 设置，因此该问题暂无解决办法。

12.11.4.27 使用 TiCDC 同步消息到 Kafka 时 Kafka 报错 Message was too large

v4.0.8 或更低版本的 TiCDC，仅在 Sink URI 中为 Kafka 配置 max-message-bytes 参数不能有效控制输出到 Kafka 的消息大小，需要在 Kafka server 配置中加入如下配置以增加 Kafka 接收消息的字节数限制。

```
### broker 能接收消息的最大字节数
message.max.bytes=2147483648
### broker 可复制的消息的最大字节数
replica.fetch.max.bytes=2147483648
### 消费者端的可读取的最大消息字节数
fetch.message.max.bytes=2147483648
```

12.11.4.28 TiCDC 同步时，在下游执行 DDL 语句失败会有什么表现，如何恢复？

如果某条 DDL 语句执行失败，同步任务 (changefeed) 会自动停止，checkpoint-ts 断点时间戳为该条出错 DDL 语句的结束时间戳 (finish-ts) 减去一。如果希望让 TiCDC 在下游重试执行这条 DDL 语句，可以使用 `cdc cli changefeed resume` 恢复同步任务。例如：

```
cdc cli changefeed resume -c test-cf --pd=http://10.0.10.25:2379
```

如果希望跳过这条出错的 DDL 语句，可以将 changefeed 的 start-ts 设为报错时的 checkpoint-ts 加上一，然后通过 `cdc cli changefeed resume` 恢复同步任务。假设报错时的 checkpoint-ts 为 415241823337054209，可以进行如下操作来跳过该 DDL 语句：

```
cdc cli changefeed update -c test-cf --pd=http://10.0.10.25:2379 --start-ts 415241823337054210

cdc cli changefeed resume -c test-cf --pd=http://10.0.10.25:2379
```

12.11.4.29 同步 DDL 到下游 MySQL 5.7 时间类型字段默认值不一致

比如上游 TiDB 的建表语句为 `create table test (id int primary key, ts timestamp)`，TiCDC 同步该语句到下游 MySQL 5.7，MySQL 使用默认配置，同步得到的表结构如下所示，timestamp 字段默认值会变成 CURRENT_TIMESTAMP：

```
mysql root@127.0.0.1:test> show create table test;
+-------+-----------------------------------------------------------------------------------------+
| Table | Create Table                                                                            |
+-------+-----------------------------------------------------------------------------------------+
| test  | CREATE TABLE `test` (                                                                   |
|       |   `id` int(11) NOT NULL,                                                                |
|       |   `ts` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP,        |
|       |   PRIMARY KEY (`id`)                                                                    |
|       | ) ENGINE=InnoDB DEFAULT CHARSET=latin1                                                  |
+-------+-----------------------------------------------------------------------------------------+
1 row in set
```

产生表结构不一致的原因是 `explicit_defaults_for_timestamp` 的默认值在 TiDB 和 MySQL 5.7 不同。从 TiCDC v5.0.1/v4.0.13 版本开始，同步到 MySQL 会自动设置 session 变量 `explicit_defaults_for_timestamp = ON`，保证同步时间类型时上下游行为一致。对于 v5.0.1/v4.0.13 以前的版本，同步时间类型时需要注意 `explicit_defaults_for_timestamp` 默认值不同带来的兼容性问题。

12.11.4.30 数据同步下游的 Sink 为 TiDB 或 MySQL 时，下游数据库的用户需要哪些权限？

Sink 为 TiDB 或 MySQL 时，下游数据库的用户需要以下权限：

- Select
- Index
- Insert
- Update
- Delete
- Create
- Drop
- Alter
- Create View

如果要同步 `recover table` 到下游 TiDB，需要有 Super 权限。

12.11.5 TiCDC 集群监控报警规则

本文介绍了 TiCDC 组件的报警项及相应的报警规则。根据严重级别，报警项按照严重程度由高到低依次为：重要级别 (Critical)、警告级别 (Warning)。

12.11.5.1 重要级别报警项

对于重要级别的报警，需要密切关注异常指标。

12.11.5.1.1 cdc_checkpoint_high_delay

- 报警规则：

  (time() - ticdc_processor_checkpoint_ts / 1000) > 600

- 规则描述：

  TiCDC 某个同步任务延迟超过 10 分钟。

- 处理方法：

  参考TiCDC 同步任务出现中断 的处理方法。

12.11.5.1.2 cdc_resolvedts_high_delay

- 报警规则：

  (time() - ticdc_processor_resolved_ts / 1000) > 300

- 规则描述：

  TiCDC 某个同步任务的 resolved ts 延迟超过 10 分钟。

- 处理方法：

  该告警与同步任务中断类似，可参考TiCDC 同步任务出现中断 的处理方法。

12.11.5.1.3 ticdc_processor_exit_with_error_count

- 报警规则：

  ```
  changes(ticdc_processor_exit_with_error_count[1m])> 0
  ```

- 规则描述：

  TiCDC 某个同步任务报错退出。

- 处理方法：

  参考TiCDC 同步任务出现中断 的处理方法。

12.11.5.2 警告级别报警项

警告级别的报警是对某一问题或错误的提醒。

12.11.5.2.1 cdc_multiple_owners

- 报警规则：

  ```
  sum(rate(ticdc_owner_ownership_counter[30s]))>= 2
  ```

- 规则描述：

  TiCDC 集群有多个 owner。

- 处理方法：

  收集 TiCDC 日志，定位原因。

12.11.5.2.2 ticdc_mounter_unmarshal_and_mount_time_more_than_1s

- 报警规则：

  ```
  histogram_quantile(0.9, rate(ticdc_mounter_unmarshal_and_mount_bucket[1m]))* 1000 > 1000
  ```

- 规则描述：

  TiCDC 某一同步任务解码变更数据的耗时超过 1 秒。

- 处理方法：

  收集 TiCDC 日志，定位原因。

12.11.5.2.3 cdc_sink_execute_duration_time_more_than_10s

- 报警规则：

  ```
  histogram_quantile(0.9, rate(ticdc_sink_txn_exec_duration_bucket[1m]))> 10
  ```

- 规则描述：

  TiCDC 某一同步任务写下游执行时间超过 10 秒。

- 处理方法：

  检查下游是否出现问题。

12.11.5.2.4 cdc_processor_checkpoint_tso_no_change_for_1m

- 报警规则：

```
changes(ticdc_processor_checkpoint_ts[1m])< 1
```

- 规则描述：

TiCDC 某一个同步任务进度超过 1 分钟没有推进。

- 处理方法：

参考TiCDC 同步任务出现中断 的处理方法。

12.11.5.2.5 ticdc_puller_entry_sorter_sort_bucket

- 报警规则：

```
histogram_quantile(0.9, rate(ticdc_puller_entry_sorter_sort_bucket{}[1m]))> 1
```

- 规则描述：

TiCDC puller entry sorter 排序延迟太高。

- 处理方法：

收集 TiCDC 日志，定位原因。

12.11.5.2.6 ticdc_puller_entry_sorter_merge_bucket

- 报警规则：

```
histogram_quantile(0.9, rate(ticdc_puller_entry_sorter_merge_bucket{}[1m]))> 1
```

- 规则描述：

TiCDC puller entry sorter merge 延迟太高。

- 处理方法：

收集 TiCDC 日志，定位原因。

12.11.5.2.7 tikv_cdc_min_resolved_ts_no_change_for_1m

- 报警规则：

```
changes(tikv_cdc_min_resolved_ts[1m])< 1 and ON (instance)tikv_cdc_region_resolve_status{
↪ status="resolved"} > 0
```

- 规则描述：

TiKV CDC 模块最小的 resolved ts 1 分钟没推进。

- 处理方法：

收集 TiKV 日志，定位原因。

12.11.5.2.8 tikv_cdc_scan_duration_seconds_more_than_10min

- 报警规则：

  ```
  histogram_quantile(0.9, rate(tikv_cdc_scan_duration_seconds_bucket{}[1m]))> 600
  ```

- 规则描述：

  TiKV CDC 模块的增量扫描耗时超过 10 分钟。

- 处理方法：

  收集 TiCDC 监控和 TiKV 日志，定位原因。

12.11.5.2.9 ticdc_sink_mysql_execution_error

- 报警规则：

  ```
  changes(ticdc_sink_mysql_execution_error[1m])> 0
  ```

- 规则描述：

  TiCDC 某一同步任务写下游 MySQL 时遇到错误。

- 处理方法：

  MySQL 报错的情况较多，参考TiCDC 常见问题和故障处理。

12.11.5.2.10 ticdc_memory_abnormal

- 报警规则：

  ```
  go_memstats_heap_alloc_bytes{job="ticdc"} > 1e+10
  ```

- 规则描述：

  TiCDC 堆内存使用量超过 10 GiB。

- 处理方法：

  收集 TiCDC 监控和 TiCDC 日志，定位原因。

12.11.6 TiCDC OpenAPI

TiCDC 提供 OpenAPI 功能，用户可通过 OpenAPI 对 TiCDC 集群进行查询和运维操作。OpenAPI 的总体功能和cdc cli 工具类似。

你可以通过 OpenAPI 完成 TiCDC 集群的如下运维操作：

- 获取 TiCDC 节点状态信息
- 检查 TiCDC 集群的健康状态
- 创建同步任务
- 删除同步任务
- 更新同步任务配置

- 查询同步任务列表
- 查询特定同步任务
- 暂停同步任务
- 恢复同步任务
- 查询同步子任务列表
- 查询特定同步子任务
- 查询 TiCDC 服务进程列表
- 驱逐 owner 节点
- 手动触发表的负载均衡
- 手动调度表到其他节点
- 动态调整 TiCDC Server 日志级别

所有 API 的请求体与返回值统一使用 JSON 格式数据。本文档以下部分描述当前提供的 API 的具体使用方法。

在下文的示例描述中，假设 TiCDC server 的监听 IP 地址为 `127.0.0.1`，端口为 8300（在启动 TiCDC server 时可以通过 `--addr=ip:port` 指定绑定的 IP 和端口）。

12.11.6.1 API 统一错误格式

对 API 发起的请求后，如发生错误，返回错误信息的格式如下所示：

```
{
    "error_msg": "",
    "error_code": ""
}
```

如上所示，`error_msg` 描述错误信息，`error_code` 则是对应的错误码。

12.11.6.2 获取 TiCDC 节点状态信息

该接口是一个同步接口，请求成功会返回对应节点的状态信息。

12.11.6.2.1 请求 URI

```
GET /api/v1/status
```

12.11.6.2.2 使用样例

以下请求会获取 IP 地址为 `127.0.0.1`，端口号为 8300 的 TiCDC 节点的状态信息。

```
curl -X GET http://127.0.0.1:8300/api/v1/status
```

```
{
    "version": "v5.2.0-master-dirty",
    "git_hash": "f191cd00c53fdf7a2b1c9308a355092f9bf8824e",
    "id": "c6a43c16-0717-45af-afd6-8b3e01e44f5d",
    "pid": 25432,
```

```
    "is_owner": true
}
```

以上返回信息的字段解释如下：

- version：当前 TiCDC 版本号。
- git_hash：Git 哈希值。
- id：该节点的 capture ID。
- pid：该节点 capture 进程的 PID。
- is_owner：表示该节点是否是 owner。

12.11.6.3 检查 TiCDC 集群的健康状态

该接口是一个同步接口，在集群健康的时候会返回 200 OK。

12.11.6.3.1 请求 URI

```
GET /api/v1/health
```

12.11.6.3.2 使用样例

```
curl -X GET http://127.0.0.1:8300/api/v1/health
```

12.11.6.4 创建同步任务

该接口是一个异步接口，请求成功会返回 202 Accepted。该返回结果只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.4.1 请求 URI

```
POST /api/v1/changefeeds
```

12.11.6.4.2 参数说明

使用 API 创建同步任务可选的参数不如使用 cli 命令创建同步任务的参数完备，以下是该 API 支持的参数。

请求体参数

参数名	说明
changefeed_id	STRING 类型，同步任务的 ID。（非必选）
start_ts	UINT64 类型，指定 changefeed 的开始 TSO。（非必选）
target_ts	UINT64 类型，指定 changefeed 的目标 TSO。（非必选）
sink_uri	STRING 类型，同步任务下游的地址。（必选）
force_replicate	BOOLEAN 类型，是否强制同步没有唯一索引的表。（非必选）
ignore_ineligible_table	BOOLEAN 类型，是否忽略无法进行同步的表。（非必选）
filter_rules	STRING 类型数组，表库过滤的规则。（非必选）

参数名	说明
ignore_txn_start_ts	UINT64 类型数组，忽略指定 start_ts 的事务。（非必选）
mounter_worker_num	INT 类型，mounter 线程数。（非必选）
sink_config	sink 的配置参数。（非必选）

changefeed_id、start_ts、target_ts、sink_uri 的含义和格式与使用 cli 创建同步任务中所作的解释相同，具体解释请参见该文档。需要注意，当在 sink_uri 中指定证书的路径时，须确保已将对应证书上传到对应的 TiCDC server 上。

下面会对一些需要补充说明的参数进行进一步阐述。

force_replicate：该值默认为 false，当指定为 true 时，同步任务会尝试强制同步没有唯一索引的表。

ignore_ineligible_table：该值默认为 false，当指定为 true 时，同步任务会忽略无法进行同步的表。

filter_rules：表库过滤的规则，如 filter_rules = ['foo*.*', 'bar*.*'] 详情参考表库过滤。

ignore_txn_start_ts：指定之后会忽略指定 start_ts 的事务，如 ignore-txn-start-ts = [1, 2]。

mounter_worker_num：mounter 线程数，mounter 用于解码 TiKV 输出的数据，默认值为 16。

sink_config：sink 的配置参数，如下

```
{
  "dispatchers":[
    {"matcher":["test1.*", "test2.*"], "dispatcher":"ts"},
    {"matcher":["test3.*", "test4.*"], "dispatcher":"rowid"}
  ],
  "protocol":"canal-json"
}
```

dispatchers：对于 MQ 类的 Sink，可以通过 dispatchers 配置 event 分发器，支持 default、ts、rowid、table 四种分发器，分发规则如下：

- default：有多个唯一索引（包括主键）时按照 table 模式分发；只有一个唯一索引（或主键）按照 rowid 模式分发；如果开启了 old value 特性，按照 table 分发。
- ts：以行变更的 commitTs 做 Hash 计算并进行 event 分发。
- rowid：以所选的 HandleKey 列名和列值做 Hash 计算并进行 event 分发。
- table：以表的 schema 名和 table 名做 Hash 计算并进行 event 分发。

matcher：匹配语法和过滤器规则语法相同。

protocol：对于 MQ 类的 Sink，可以指定消息的协议格式。目前支持 canal-json、open-protocol、canal、avro 和 maxwell 五种协议。

12.11.6.4.3 使用样例

以下请求会创建一个 ID 为 test5，sink_uri 为 blackhome://的同步任务。

```
curl -X POST -H "'Content-type':'application/json'" http://127.0.0.1:8300/api/v1/changefeeds -d
    ↪ '{"changefeed_id":"test5","sink_uri":"blackhole://"}'
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.5 删除同步任务

该接口是一个异步接口，请求成功会返回 202 Accepted，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.5.1 请求 URI

```
DELETE /api/v1/changefeeds/{changefeed_id}
```

12.11.6.5.2 参数说明

路径参数

参数名	说明
changefeed_id	需要删除的同步任务 (changefeed) 的 ID

12.11.6.5.3 使用样例

以下请求会删除 ID 为 test1 的同步任务。

```
curl -X DELETE http://127.0.0.1:8300/api/v1/changefeeds/test1
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.6 更新同步任务配置

该接口是一个异步接口，请求成功会返回 202 Accepted，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

修改 changefeed 配置需要按照暂停任务 -> 修改配置 -> 恢复任务的流程。

12.11.6.6.1 请求 URI

```
PUT /api/v1/changefeeds/{changefeed_id}
```

12.11.6.6.2 参数说明

路径参数

参数名	说明
changefeed_id	需要更新的同步任务 (changefeed) 的 ID

请求体参数

目前仅支持通过 API 修改同步任务的如下配置。

参数名	说明
target_ts	UINT64 类型，指定 changefeed 的目标 TSO。(非必选)
sink_uri	STRING 类型，同步任务下游的地址。(非必选)
filter_rules	STRING 类型数组，表库过滤的规则。(非必选)
ignore_txn_start_ts	UINT64 类型数组，忽略指定 start_ts 的事务。 (非必选)
mounter_worker_num	INT 类型，mounter 线程数。(非必选)
sink_config	sink 的配置参数。(非必选)

以上参数含义与创建同步任务中的参数相同，此处不再赘述。

12.11.6.6.3 使用样例

以下请求会更新 ID 为 test1 的同步任务的 mounter_worker_num 为 32。

```
curl -X PUT -H "'Content-type':'application/json'" http://127.0.0.1:8300/api/v1/changefeeds/
    ↪ test1 -d '{"mounter_worker_num":32}'
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.7 查询同步任务列表

该接口是一个同步接口，请求成功会返回 TiCDC 集群中所有同步任务 (changefeed) 的基本信息。

12.11.6.7.1 请求 URI

```
GET /api/v1/changefeeds
```

12.11.6.7.2 参数说明

查询参数

参数名	说明
state	非必选，指定后将会只返回该状态的同步任务的信息

state 可选值为 all、 normal、 stopped、 error、 failed、 finished。

若不指定该参数，则默认返回处于 normal、 stopped、 failed 状态的同步任务基本信息。

12.11.6.7.3 使用样例

以下请求查询所有状态 (state) 为 normal 的同步任务的基本信息。

```
curl -X GET http://127.0.0.1:8300/api/v1/changefeeds?state=normal
```

```
[
    {
        "id": "test1",
        "state": "normal",
        "checkpoint_tso": 426921294362574849,
        "checkpoint_time": "2021-08-10 14:04:54.242",
        "error": null
    },
    {
        "id": "test2",
        "state": "normal",
        "checkpoint_tso": 426921294362574849,
        "checkpoint_time": "2021-08-10 14:04:54.242",
        "error": null
    }
]
```

此处对以上返回的信息做进一步阐述：

- id：同步任务的 ID
- state：同步任务当前所处的状态。
- checkpoint_tso：同步任务当前 checkpoint 的 TSO 表示。
- checkpoint_tso：同步任务当前 checkpoint 的格式化时间表示。
- error：同步任务的错误信息。

12.11.6.8 查询特定同步任务

该接口是一个同步接口，请求成功会返回指定同步任务 (changefeed) 的详细信息。

12.11.6.8.1 请求 URI

```
GET /api/v1/changefeeds/{changefeed_id}
```

12.11.6.8.2 参数说明

路径参数

参数名	说明
changefeed_id	需要查询的同步任务 (changefeed) 的 ID

12.11.6.8.3 使用样例

以下请求会查询 ID 为 test1 的同步任务的详细信息。

```
curl -X GET http://127.0.0.1:8300/api/v1/changefeeds/test1
```

```
{
    "id": "test1",
    "sink_uri": "blackhole://",
    "create_time": "2021-08-10 11:41:30.642",
    "start_ts": 426919038970232833,
    "target_ts": 0,
    "checkpoint_tso": 426921014615867393,
    "checkpoint_time": "2021-08-10 13:47:07.093",
    "sort_engine": "unified",
    "state": "normal",
    "error": null,
    "error_history": null,
    "creator_version": "",
    "task_status": [
        {
            "capture_id": "d8924259-f52f-4dfb-97a9-c48d26395945",
            "table_ids": [
                63,
                65
            ],
            "table_operations": {}
        }
    ]
}
```


12.11.6.9 暂停同步任务

该接口是一个异步接口，请求成功会返回 `202 Accepted`，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.9.1 请求 URI

```
POST /api/v1/changefeeds/{changefeed_id}/pause
```

12.11.6.9.2 参数说明

路径参数

参数名	说明
`changefeed_id`	需要暂停的同步任务 (changefeed) 的 ID

12.11.6.9.3 使用样例

以下请求会暂停 ID 为 `test1` 的同步任务。

```
curl -X POST http://127.0.0.1:8300/api/v1/changefeeds/test1/pause
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.10 恢复同步任务

该接口是一个异步接口，请求成功会返回 202 Accepted，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.10.1 请求 URI

```
POST /api/v1/changefeeds/{changefeed_id}/resume
```

12.11.6.10.2 参数说明

路径参数

参数名	说明
changefeed_id	需要恢复的同步任务 (changefeed) 的 ID

12.11.6.10.3 使用样例

以下请求会恢复 ID 为 test1 的同步任务。

```
curl -X POST http://127.0.0.1:8300/api/v1/changefeeds/test1/resume
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.11 查询同步子任务列表

该接口是一个同步接口，请求成功会返回当前 TiCDC 集群中的所有同步子任务 (processor) 的基本信息。

12.11.6.11.1 请求 URI

```
GET /api/v1/processors
```

12.11.6.11.2 使用样例

```
curl -X GET http://127.0.0.1:8300/api/v1/processors
```

```
[
    {
        "changefeed_id": "test1",
        "capture_id": "561c3784-77f0-4863-ad52-65a3436db6af"
    }
]
```

12.11.6.12 查询特定同步子任务

该接口是一个同步接口，请求成功会返回指定同步子任务 (processor) 的详细信息。

12.11.6.12.1 请求 URI

```
GET /api/v1/processors/{changefeed_id}/{capture_id}
```

12.11.6.12.2 参数说明

路径参数

参数名	说明
changefeed_id	需要查询的子任务的 Changefeed ID
capture_id	需要查询的子任务的 Capture ID

12.11.6.12.3 使用样例

以下请求查询 changefeed_id 为 test、capture_id 为 561c3784-77f0-4863-ad52-65a3436db6af 的同步子任务。一个同步子任务通过 changefeed_id 和 capture_id 来标识。

```
curl -X GET http://127.0.0.1:8300/api/v1/processors/test1/561c3784-77f0-4863-ad52-65a3436db6af
```

```
{
    "checkpoint_ts": 426919123303006208,
    "resolved_ts": 426919123369066496,
    "table_ids": [
        63,
        65
    ],
    "error": null
}
```

12.11.6.13 查询 TiCDC 服务进程列表

该接口是一个同步接口，请求成功会返回当前 TiCDC 集群中的所有服务进程 (capture) 的基本信息。

12.11.6.13.1 请求 URI

```
GET /api/v1/captures
```

12.11.6.13.2 使用样例

```
curl -X GET http://127.0.0.1:8300/api/v1/captures
```

```
[
    {
        "id": "561c3784-77f0-4863-ad52-65a3436db6af",
        "is_owner": true,
        "address": "127.0.0.1:8300"
    }
]
```

12.11.6.14 驱逐 owner 节点

该接口是一个异步的请求，请求成功会返回 `202 Accepted`，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.14.1 请求 URI

```
POST /api/v1/owner/resign
```

12.11.6.14.2 使用样例

以下请求会驱逐 TiCDC 当前的 owner 节点，并会触发新一轮的选举，产生新的 owner 节点。

```
curl -X POST http://127.0.0.1:8300/api/v1/owner/resign
```

若是请求成功，则返回 `202 Accepted`，若请求失败，则返回错误信息和错误码。

12.11.6.15 手动触发表的负载均衡

该接口是一个异步的请求，请求成功会返回 `202 Accepted`它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.15.1 请求 URI

```
POST /api/v1/changefeeds/{changefeed_id}/tables/rebalance_table
```

12.11.6.15.2 参数说明

路径参数

参数名	说明
changefeed_id	进行调度的 Changefeed ID

12.11.6.15.3 使用样例

以下请求会触发 ID 为 `test1` 的 changefeed 表的负载均衡。

```
curl -X POST http://127.0.0.1:8300/api/v1/changefeeds/test1/tables/rebalance_table
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.16 手动调度表到其他节点

该接口是一个异步的请求，请求成功会返回 202 Accepted，它只代表服务器答应执行该命令，不保证命令会被成功的执行。

12.11.6.16.1 请求 URI

```
POST /api/v1/changefeeds/{changefeed_id}/tables/move_table
```

12.11.6.16.2 参数说明

路径参数

参数名	说明
changefeed_id	进行调度的 Changefeed ID

请求体参数

参数名	说明
target_capture_id	目标 Capture ID
table_id	需要调度的 Table ID

12.11.6.16.3 使用样例

以下请求会将 ID 为 test1 的 changefeed 中 ID 为 49 的 table 调度到 ID 为 6f19a6d9-0f8c-4dc9-b299-3ba7c0f216f5 的 capture 上去。

```
curl -X POST -H "'Content-type':'application/json'" http://127.0.0.1:8300/api/v1/changefeeds/
    ↪ changefeed-test1/tables/move_table -d '{"capture_id":"6f19a6d9-0f8c-4dc9-b299-3
    ↪ ba7c0f216f5","table_id":49}'
```

若是请求成功，则返回 202 Accepted，若请求失败，则返回错误信息和错误码。

12.11.6.17 动态调整 TiCDC Server 日志级别

该接口是一个同步接口，请求成功会返回 200 OK。

12.11.6.17.1 请求 URI

```
POST /api/v1/log
```

12.11.6.17.2 请求参数

请求体参数

参数名	说明
log_level	想要设置的日志等级

log_level 支持 zap 提供的日志级别：“debug”、“info”、“warn”、“error”、“dpanic”、“panic”、“fatal”。

12.11.6.17.3 使用样例

```
curl -X POST -H "'Content-type':'application/json'" http://127.0.0.1:8300/api/v1/log -d '{"
    ↪ log_level":"debug"}'
```

若是请求成功，则返回 200 OK，若请求失败，则返回错误信息和错误码。

12.11.7 TiCDC Open Protocol

12.11.7.1 概述

TiCDC Open Protocol 是一种行级别的数据变更通知协议，为监控、缓存、全文索引、分析引擎、异构数据库的主从复制等提供数据源。TiCDC 遵循 TiCDC Open Protocol，向 MQ(Message Queue) 等第三方数据媒介复制 TiDB 的数据变更。

TiCDC Open Protocol 以 Event 为基本单位向下游复制数据变更事件，Event 分为三类：

- Row Changed Event：代表一行的数据变化，在行发生变更时该 Event 被发出，包含变更后该行的相关信息。
- DDL Event：代表 DDL 变更，在上游成功执行 DDL 后发出，DDL Event 会广播到每一个 MQ Partition 中。
- Resolved Event：代表一个特殊的时间点，表示在这个时间点前的收到的 Event 是完整的。

12.11.7.2 协议约束

- 在绝大多数情况下，一个版本的 Row Changed Event 只会发出一次，但是特殊情况（节点故障、网络分区等）下，同一版本的 Row Changed Event 可能会多次发送。
- 同一张表中的每一个版本第一次发出的 Row Changed Event 在 Event 流中一定是按 TS (timestamp) 顺序递增的。
- Resolved Event 会被周期性的广播到各个 MQ Partition，Resolved Event 意味着任何 TS 小于 Resolved Event TS 的 Event 已经发送给下游。
- DDL Event 将被广播到各个 MQ Partition。
- 一行数据的多个 Row Changed Event 一定会被发送到同一个 MQ Partition 中。

12.11.7.3 Message 格式定义

一个 Message 中包含一个或多个 Event，按照以下格式排列：

Key:

Offset(Byte)	0~7	8~15	16~(15+ 长度 1)	…	…
参数	协议版本号	长度 1	Event Key1	长度 N	Event KeyN

Value:

Offset(Byte)	0~7	8~(7+ 长度 1)	…	…
参数	长度 1	Event Value1	长度 N	Event ValueN

- 长度N代表第 N 个 Key/Value 的长度
- 长度及协议版本号均为大端序 int64 类型
- 当前协议版本号为 1

12.11.7.4 Event 格式定义

本部分介绍 Row Changed Event、 DDL Event 和 Resolved Event 的格式定义。

12.11.7.4.1 Row Changed Event

- Key:

```
{
    "ts":<TS>,
    "scm":<Schema Name>,
    "tbl":<Table Name>,
    "t":1
}
```

参数	类型	说明
TS	Number	造成 Row 变更的事务的 TS
Schema Name	String	Row 所在的 Schema 的名字
Table Name	String	Row 所在的 Table 的名字

- Value:

  Insert 事件，输出新增的行数据。

```
{
    "u":{
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
```

```
        },
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        }
    }
}
```

Update 事件，输出新增的行数据（“u”）以及修改前的行数据（“p”），仅当 Old Value 特性开启时，才会输出修改前的行数据。

```
{
    "u":{
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        },
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        }
    },
    "p":{
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        },
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        }
    }
}
```

Delete 事件，输出被删除的行数据。当 Old Value 特性开启时，Delete 事件中包含被删除的行数据中的所有列；当 Old Value 特性关闭时，Delete 事件中仅包含HandleKey 列。

```
{
    "d":{
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        },
        <Column Name>:{
            "t":<Column Type>,
            "h":<Where Handle>,
            "f":<Flag>,
            "v":<Column Value>
        }
    }
}
```

参数	类型	说明
Column Name	String	列名
Column Type	Number	列类型，详见：Column 的类型码
Where Handle	Bool	表示该列是否可以作为 Where 筛选条件，当该列在表内具有唯一性时，Where Handle 为 true。
Flag	Number	列标志位，详见：列标志位
Column Value	Any	列值

12.11.7.4.2 DDL Event

- Key:

```
{
    "ts":<TS>,
    "scm":<Schema Name>,
    "tbl":<Table Name>,
    "t":2
}
```

参数	类型	说明
TS	Number	进行 DDL 变更的事务的 TS
Schema Name	String	DDL 变更的 Schema 的名字，可能为空字符串
Table Name	String	DDL 变更的 Table 的名字，可能为空字符串

- Value:

```
{
    "q":<DDL Query>,
    "t":<DDL Type>
}
```

参数	类型	说明
DDL Query	String	DDL Query SQL
DDL Type	String	DDL 类型，详见：DDL 的类型码

12.11.7.4.3 Resolved Event

- Key:

```
{
    "ts":<TS>,
    "t":3
}
```

参数	类型	说明
TS	Number	Resolved TS，任意小于该 TS 的 Event 已经发送完毕

- Value: None

12.11.7.5 Event 流的输出示例

本部分展示并描述 Event 流的输出日志。

假设在上游执行以下 SQL 语句，MQ Partition 数量为 2：

```
CREATE TABLE test.t1(id int primary key, val varchar(16));
```

如以下执行日志中的 Log 1、Log 3 所示，DDL Event 将被广播到所有 MQ Partition，Resolved Event 会被周期性地广播到各个 MQ Partition：

```
1. [partition=0] [key="{\"ts\":415508856908021766,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":2}"] [
    ↪ value="{\"q\":\"CREATE TABLE test.t1(id int primary key, val varchar(16))\",\"t\":3}"]
2. [partition=0] [key="{\"ts\":415508856908021766,\"t\":3}"] [value=]
3. [partition=1] [key="{\"ts\":415508856908021766,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":2}"] [
    ↪ value="{\"q\":\"CREATE TABLE test.t1(id int primary key, val varchar(16))\",\"t\":3}"]
4. [partition=1] [key="{\"ts\":415508856908021766,\"t\":3}"] [value=]
```

在上游执行以下 SQL 语句：

```
BEGIN;
INSERT INTO test.t1(id, val) VALUES (1, 'aa');
INSERT INTO test.t1(id, val) VALUES (2, 'aa');
UPDATE test.t1 SET val = 'bb' WHERE id = 2;
INSERT INTO test.t1(id, val) VALUES (3, 'cc');
COMMIT;
```

- 如以下执行日志中的 Log 5 和 Log 6 所示，同一张表内的 Row Changed Event 可能会根据主键被分派到不同的 Partition，但同一行的变更一定会分派到同一个 Partition，方便下游并发处理。
- 如 Log 6 所示，在一个事务内对同一行进行多次修改，只会发出一个 Row Changed Event。
- Log 8 是 Log 7 的重复 Event。 Row Changed Event 可能重复，但每个版本的 Event 第一次发出的次序一定是有序的。

```
5. [partition=0] [key="{\"ts\":415508878783938562,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":1},\"val\":{\"t\":15,\"v\":\"YWE=\"}}}"]
6. [partition=1] [key="{\"ts\":415508878783938562,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":2},\"val\":{\"t\":15,\"v\":\"YmI=\"}}}"]
7. [partition=0] [key="{\"ts\":415508878783938562,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":3},\"val\":{\"t\":15,\"v\":\"Y2M=\"}}}"]
8. [partition=0] [key="{\"ts\":415508878783938562,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":3},\"val\":{\"t\":15,\"v\":\"Y2M=\"}}}"]
```

在上游执行以下 SQL 语句：

```
BEGIN;
DELETE FROM test.t1 WHERE id = 1;
UPDATE test.t1 SET val = 'dd' WHERE id = 3;
UPDATE test.t1 SET id = 4, val = 'ee' WHERE id = 2;
COMMIT;
```

- Log 9 是 Delete 类型的 Row Changed Event，这种类型的 Event 只包含主键列或唯一索引列。
- Log 13 和 Log 14 是 Resolved Event。Resolved Event 意味着在这个 Partition 中，任意小于 Resolved TS 的 Event（包括 Row Changed Event 和 DDL Event）已经发送完毕。

```
9. [partition=0] [key="{\"ts\":415508881418485761,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"d\":{\"id\":{\"t\":3,\"h\":true,\"v\":1}}}"]
10. [partition=1] [key="{\"ts\":415508881418485761,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"d\":{\"id\":{\"t\":3,\"h\":true,\"v\":2}}}"]
11. [partition=0] [key="{\"ts\":415508881418485761,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":3},\"val\":{\"t\":15,\"v\":\"ZGQ=\"}}}"]
12. [partition=0] [key="{\"ts\":415508881418485761,\"scm\":\"test\",\"tbl\":\"t1\",\"t\":1}"] [
    ↪ value="{\"u\":{\"id\":{\"t\":3,\"h\":true,\"v\":4},\"val\":{\"t\":15,\"v\":\"ZWU=\"}}}"]
13. [partition=0] [key="{\"ts\":415508881038376963,\"t\":3}"] [value=]
14. [partition=1] [key="{\"ts\":415508881038376963,\"t\":3}"] [value=]
```

12.11.7.6 消费端协议解析

目前 TiCDC 没有提供 Open Protocol 协议解析的标准实现，但是提供了 Golang 版本和 Java 版本的解析例子。你可以参考本文档提供的数据格式和以下例子实现消费端协议解析。

- Golang 例子
- Java 例子

12.11.7.7 Column 的类型码

Column 的类型码用于标识 Row Changed Event 中列的数据类型。

类型	Code	输出示例	说明
TINYINT/BOOL	1	{“t”:1,“v”:1}	
SMALLINT	2	{“t”:2,“v”:1}	
INT	3	{“t”:3,“v”:123}	
FLOAT	4	{“t”:4,“v”:153.123}	
DOUBLE	5	{“t”:5,“v”:153.123}	
NULL	6	{“t”:6,“v”:null}	
TIMESTAMP	7	{“t”:7,“v”:“1973-12-30 15:30:00”}	
BIGINT	8	{“t”:8,“v”:123}	
MEDIUMINT	9	{“t”:9,“v”:123}	
DATE	10/14	{“t”:10,“v”:“2000-01-01”}	
TIME	11	{“t”:11,“v”:“23:59:59”}	
DATETIME	12	{“t”:12,“v”:“2015-12-20 23:58:58”}	
YEAR	13	{“t”:13,“v”:1970}	

类型	Code	输出示例	说明
VARCHAR/VARBINARY	15/253	{ “t” :15, “v” : “测试” } / { “t” :15, “v” : “\\x89PNG\\r\\n\\x1a\\n” }	编码为 UTF-8；当上游类型为 VARBINARY 时，将对不可见的字符转义
BIT	16	{ “t” :16, “v” :81}	
JSON	245	{ “t” :245, “v” : “{\” key1\ “: \” value1\ “}” }	
DECIMAL	246	{ “t” :246, “v” : “129012.1230000” }	
ENUM	247	{ “t” :247, “v” :1}	
SET	248	{ “t” :248, “v” :3}	
TINYTEXT/TINYBLOB	249	{ “t” :249, “v” : “5rWL6K+VdGV4dA==” }	编码为 Base64
MEDIUMTEXT/MEDIUMBLOB	250	{ “t” :250, “v” : “5rWL6K+VdGV4dA==” }	编码为 Base64

类型	Code	输出示例	说明
LONGTEXT/LONGBLOB	251	{ “t” :251, “v”:”5rWL6K+VdGV4dA==” }	编码为 Base64
TEXT/BLOB	252	{ “t” :252, “v”:”5rWL6K+VdGV4dA==” }	编码为 Base64
CHAR/BINARY	254	{ “t” :254, “v”:”测试” } / { “t” :254, “v”:”\\x89PNG\\r\\n\\x1a\\n” }	编码为 UTF-8；当上游类型为 BINARY 时，将对不可见的字符转义
GEOMETRY	255		尚不支持

12.11.7.8 DDL 的类型码

DDL 的类型码用于标识 DDL Event 中的 DDL 语句的类型。

类型	Code
Create Schema	1
Drop Schema	2
Create Table	3
Drop Table	4
Add Column	5
Drop Column	6
Add Index	7
Drop Index	8
Add Foreign Key	9
Drop Foreign Key	10
Truncate Table	11
Modify Column	12
Rebase Auto ID	13
Rename Table	14
Set Default Value	15
Shard RowID	16
Modify Table Comment	17
Rename Index	18
Add Table Partition	19
Drop Table Partition	20
Create View	21
Modify Table Charset And Collate	22
Truncate Table Partition	23
Drop View	24
Recover Table	25
Modify Schema Charset And Collate	26
Lock Table	27
Unlock Table	28
Repair Table	29
Set TiFlash Replica	30
Update TiFlash Replica Status	31
Add Primary Key	32
Drop Primary Key	33
Create Sequence	34
Alter Sequence	35
Drop Sequence	36

12.11.7.9 列标志位

列标志位以 Bit flags 形式标记列的相关属性。

位移	值	名称	说明
1	0x01	BinaryFlag	该列是否为二进制编码列
2	0x02	HandleKeyFlag	该列是否为 Handle 列
3	0x04	GeneratedColumnFlag	该列是否为生成列
4	0x08	PrimaryKeyFlag	该列是否为主键列
5	0x10	UniqueKeyFlag	该列是否为唯一索引列
6	0x20	MultipleKeyFlag	该列是否为组合索引列
7	0x40	NullableFlag	该列是否为可空列
8	0x80	UnsignedFlag	该列是否为无符号列

示例：

若某列 Flag 值为 85，则代表这一列为可空列、唯一索引列、生成列、二进制编码列。

```
85 == 0b_101_0101
   == NullableFlag | UniqueKeyFlag | GeneratedColumnFlag | BinaryFlag
```

若某列 Flag 值为 46，则代表这一列为组合索引列、主键列、生成列、Handle 列。

```
46 == 0b_010_1110
   == MultipleKeyFlag | PrimaryKeyFlag | GeneratedColumnFlag | HandleKeyFlag
```

> 注意：
>
> - BinaryFlag 仅在列为 BLOB/TEXT（包括 TINYBLOB/TINYTEXT、BINARY/CHAR 等）类型时才有意义。当上游列为 BLOB 类型时，BinaryFlag 置 1；当上游列为 TEXT 类型时，BinaryFlag 置 0。
> - 若要同步上游的一张表，TiCDC 会选择一个有效索引作为 Handle Index。Handle Index 包含的列的 HandleKeyFlag 置 1。

12.11.8 TiCDC Canal-JSON Protocol

Canal-JSON 是由 Alibaba Canal 定义的一种数据交换格式协议。通过本文，你可以了解 TiCDC 对 Canal-JSON 数据格式的实现，包括 TiDB 扩展字段、Canal-JSON 数据格式定义，以及和官方实现进行对比等相关内容。

12.11.8.1 使用 Canal-JSON

当使用 MQ (Message Queue) 作为下游 Sink 时，你可以在 `sink-uri` 中指定使用 Canal-JSON，TiCDC 将以 Event 为基本单位封装构造 Canal-JSON Message，向下游发送 TiDB 的数据变更事件。

Event 分为三类：

- DDL Event：代表 DDL 变更记录，在上游成功执行 DDL 语句后发出，DDL Event 会被发送到索引为 0 的 MQ Partition。

- DML Event：代表一行数据变更记录，在行变更发生时该类 Event 被发出，包含变更后该行的相关信息。
- WATERMARK Event：代表一个特殊的时间点，表示在这个时间点前收到的 Event 是完整的。仅适用于 TiDB 扩展字段，当你在 `sink-uri` 中设置 `enable-tidb-extension=true` 时生效。

使用 Canal-JSON 时的配置样例如下所示：

```
cdc cli changefeed create --pd=http://127.0.0.1:2379 --changefeed-id="kafka-canal-json" --sink-
    ↪ uri="kafka://127.0.0.1:9092/topic-name?kafka-version=2.6.0&protocol=canal-json"
```

12.11.8.2 TiDB 扩展字段

Canal-JSON 协议本是为 MySQL 设计的，其中并不包含 TiDB 专有的 CommitTS 事务唯一标识等重要字段。为了解决这个问题，TiCDC 在 Canal-JSON 协议格式中附加了 TiDB 扩展字段。在 `sink-uri` 中设置 `enable-tidb-extension` 为 `true` 后，TiCDC 生成 Canal-JSON 消息时的行为如下：

- TiCDC 发送的 DML Event 和 DDL Event 类型消息中，将会含有一个名为 `_tidb` 的字段。
- TiCDC 将会发送 WATERMARK Event 消息。·

配置样例如下所示：

```
cdc cli changefeed create --pd=http://127.0.0.1:2379 --changefeed-id="kafka-canal-json-enable-
    ↪ tidb-extension" --sink-uri="kafka://127.0.0.1:9092/topic-name?kafka-version=2.6.0&
    ↪ protocol=canal-json&enable-tidb-extension=true"
```

`enable-tidb-extension` 默认为 `false`，仅当使用 Canal-JSON 时生效。

12.11.8.3 Message 格式定义

下面介绍 DDL Event、 DML Event 和 WATERMARK Event 的格式定义，以及消费端的数据解析。

12.11.8.3.1 DDL Event

TiCDC 会把一个 DDL Event 编码成如下 Canal-JSON 格式：

```
{
    "id": 0,
    "database": "test",
    "table": "",
    "pkNames": null,
    "isDdl": true,
    "type": "QUERY",
    "es": 1639633094670,
    "ts": 1639633095489,
    "sql": "drop database if exists test",
    "sqlType": null,
    "mysqlType": null,
    "data": null,
```

```
    "old": null,
    "_tidb": {     // TiDB 的扩展字段
        "commitTs": 163963309467037594
    }
}
```

以上 JSON 数据的字段解释如下：

字段	类型	说明
id	Number	TiCDC 默认值为 0
database	String	Row 所在的 Database 的名字
table	String	Row 所在的 Table 的名字
pkNames	Array	组成 primary key 的所有列的名字
isDdl	Bool	该条消息是否为 DDL 事件
type	String	Canal-JSON 定义的事件类型
es	Number	产生该条消息的事件发生时的 13 位（毫秒级）时间戳
ts	Number	TiCDC 生成该条消息时的 13 位（毫秒级）时间戳
sql	String	当 isDdl 为 true 时，记录对应的 DDL 语句
sqlType	Object	当 isDdl 为 false 时，记录每一列数据类型在 Java 中的类型表示
mysqlType	object	当 isDdl 为 false 时，记录每一列数据类型在 MySQL 中的类型表示
data	Object	当 isDdl 为 false 时，记录每一列的名字及其数据值
old	Object	仅当该条消息由 Update 类型事件产生时，记录每一列的名字，和 Update 之前的数据值
_tidb	Object	TiDB 扩展字段，仅当 `enable-tidb-extension` 为 true 时才会存在。其中的 `commitTs` 值为造成 Row 变更的事务的 TSO

12.11.8.3.2 DML Event

对于一行 DML 数据变更事件，TiCDC 会将其编码成如下形式:

```
{
    "id": 0,
    "database": "test",
    "table": "tp_int",
    "pkNames": [
        "id"
    ],
    "isDdl": false,
    "type": "INSERT",
    "es": 1639633141221,
    "ts": 1639633142960,
    "sql": "",
    "sqlType": {
        "c_bigint": -5,
        "c_int": 4,
```

```
        "c_mediumint": 4,
        "c_smallint": 5,
        "c_tinyint": -6,
        "id": 4
    },
    "mysqlType": {
        "c_bigint": "bigint",
        "c_int": "int",
        "c_mediumint": "mediumint",
        "c_smallint": "smallint",
        "c_tinyint": "tinyint",
        "id": "int"
    },
    "data": [
        {
            "c_bigint": "9223372036854775807",
            "c_int": "2147483647",
            "c_mediumint": "8388607",
            "c_smallint": "32767",
            "c_tinyint": "127",
            "id": "2"
        }
    ],
    "old": null,
    "_tidb": {     // TiDB 的扩展字段
        "commitTs": 163963314122145239
    }
}
```

12.11.8.3.3 WATERMARK Event

仅当 `enable-tidb-extension` 为 `true` 时，TiCDC 才会发送 WATERMARK Event，其 `type` 字段值为 `TIDB_WATERMARK`。该类型事件具有 `_tidb` 字段，当前只含有 `watermarkTs`，其值为该 Event 发送时的 TSO。

当你收到一个该类型的事件，所有 `commitTs` 小于 `watermarkTs` 的事件均已发送完毕。因为 TiCDC 提供 At Least Once 语义，可能出现重复发送数据的情况。如果后续收到有 `commitTs` 小于 `watermarkTs` 的事件，可以忽略。

WATERMARK Event 的示例如下：

```
{
    "id": 0,
    "database": "",
    "table": "",
    "pkNames": null,
    "isDdl": false,
    "type": "TIDB_WATERMARK",
    "es": 1640007049196,
```

```
    "ts": 1640007050284,
    "sql": "",
    "sqlType": null,
    "mysqlType": null,
    "data": null,
    "old": null,
    "_tidb": {     // TiDB 的扩展字段
        "watermarkTs": 429918007904436226
    }
}
```

12.11.8.3.4 消费端数据解析

从上面的示例中可知，Canal-JSON 具有统一的数据格式，针对不同的事件类型，有不同的字段填充规则。消费者可以使用统一的方法对该 JSON 格式的数据进行解析，然后通过判断字段值的方式，来确定具体事件类型：

- 当 `isDdl` 为 true 时，该消息含有一条 DDL Event。
- 当 `isDdl` 为 false 时，需要对 `type` 字段加以判断。如果 `type` 为 `TIDB_WATERMARK`，可得知其为 WATERMARK Event，否则就是 DML Event。

12.11.8.4 字段说明

Canal-JSON 格式会在 `mysqlType` 字段和 `sqlType` 字段中记录对应的数据类型。

12.11.8.4.1 MySQL Type 字段

Canal-JSON 格式会在 `mysqlType` 字段中记录每一列的 MySQL Type 的字符串表示。相关详情可以参考 TiDB Data Types。

12.11.8.4.2 SQL Type 字段

Canal-JSON 格式会在 `sqlType` 字段中记录每一列的 Java SQL Type，即每条数据在 JDBC 中对应的数据类型，其值可以通过 MySQL Type 和具体数据值计算得到。具体对应关系如下:

MySQL Type	Java SQL Type Code
Boolean	-6
Float	7
Double	8
Decimal	3
Char	1
Varchar	12
Binary	2004
Varbinary	2004
Tinytext	2005
Text	2005

MySQL Type	Java SQL Type Code
Mediumtext	2005
Longtext	2005
Tinyblob	2004
Blob	2004
Mediumblob	2004
Longblob	2004
Date	91
Datetime	93
Timestamp	93
Time	92
Year	12
Enum	4
Set	-7
Bit	-7
JSON	12

12.11.8.5 整数类型

你需要考虑整数类型是否有 `Unsigned` 约束，以及当前取值大小，分别对应不同的 Java SQL Type Code。如下表所示。

MySQL Type String	Value Range	Java SQL Type Code
tinyint	[-128, 127]	-6
tinyint unsigned	[0, 127]	-6
tinyint unsigned	[128, 255]	5
smallint	[-32768, 32767]	5
smallint unsigned	[0, 32767]	5
smallint unsigned	[32768, 65535]	4
mediumint	[-8388608, 8388607]	4
mediumint unsigned	[0, 8388607]	4
mediumint unsigned	[8388608, 16777215]	4
int	[-2147483648, 2147483647]	4
int unsigned	[0, 2147483647]	4
int unsigned	[2147483648, 4294967295]	-5
bigint	[-9223372036854775808, 9223372036854775807]	-5
bigint unsigned	[0, 9223372036854775807]	-5
bigint unsigned	[9223372036854775808, 18446744073709551615]	3

TiCDC 涉及的 Java SQL Type 及其 Code 映射关系如下表所示。

Java SQL Type	Java SQL Type Code
CHAR	1

Java SQL Type	Java SQL Type Code
DECIMAL	3
INTEGER	4
SMALLINT	5
REAL	7
DOUBLE	8
VARCHAR	12
DATE	91
TIME	92
TIMESTAMP	93
BLOB	2004
CLOB	2005
BIGINT	-5
TINYINT	-6
Bit	-7

想要了解 Java SQL Type 的更多信息，请参考 Java SQL Class Types。

12.11.8.6 TiCDC Canal-JSON 和 Canal 官方实现对比

TiCDC 对 Canal-JSON 数据格式的实现，包括 `Update` 类型事件和 `mysqlType` 字段，和官方有些许不同。主要差异见下表。

差异点	TiCDC	Canal
`Update` 类型事件	`old` 字段包含所有列数据	`old` 字段仅包含被修改的列数据
`mysqlType` 字段	对于含有参数的类型，没有类型参数信息	对于含有参数的类型，会包含完整的参数信息

12.11.8.6.1 `Update` 类型事件

对于 `Update` 类型事件，Canal 官方实现中，`old` 字段仅包含被修改的列数据，而 TiCDC 的实现则包含所有列数据。

假设在上游 TiDB 按顺序执行如下 SQL 语句:

```
create table tp_int
(
    id          int auto_increment,
    c_tinyint   tinyint   null,
    c_smallint  smallint  null,
    c_mediumint mediumint null,
    c_int       int       null,
    c_bigint    bigint    null,
    constraint pk
        primary key (id)
);
```

```
insert into tp_int(c_tinyint, c_smallint, c_mediumint, c_int, c_bigint)
values (127, 32767, 8388607, 2147483647, 9223372036854775807);

update tp_int set c_int = 0, c_tinyint = 0 where c_smallint = 32767;
```

对于 update 语句，TiCDC 将会输出一条 type 为 UPDATE 的事件消息，如下所示。该 update 语句仅对 c_int 和 c_tinyint 两列进行了修改。输出事件消息的 old 字段，则包含所有列数据。

```
{
    "id": 0,
    ...
    "type": "UPDATE",
    ...
    "sqlType": {
        ...
    },
    "mysqlType": {
        ...
    },
    "data": [
        {
            "c_bigint": "9223372036854775807",
            "c_int": "0",
            "c_mediumint": "8388607",
            "c_smallint": "32767",
            "c_tinyint": "0",
            "id": "2"
        }
    ],
    "old": [                                // TiCDC 输出事件消息的 `old` 字段，则包含所有列数据
        ↪ 。
        {
            "c_bigint": "9223372036854775807",
            "c_int": "2147483647",          // 修改的列
            "c_mediumint": "8388607",
            "c_smallint": "32767",
            "c_tinyint": "127",             // 修改的列
            "id": "2"
        }
    ]
}
```

官方 Canal 输出事件消息的 old 字段仅包含被修改的列数据。示例如下。

```
{
```

```
    "id": 0,
    ...
    "type": "UPDATE",
    ...
    "sqlType": {
        ...
    },
    "mysqlType": {
        ...
    },
    "data": [
        {
            "c_bigint": "9223372036854775807",
            "c_int": "0",
            "c_mediumint": "8388607",
            "c_smallint": "32767",
            "c_tinyint": "0",
            "id": "2"
        }
    ],
    "old": [                                    // Canal 输出事件消息的 `old` 字段，
        ↪ 仅包含被修改的列的数据。
        {
            "c_int": "2147483647",              // 修改的列
            "c_tinyint": "127",                 // 修改的列
        }
    ]
}
```

12.11.8.6.2 mysqlType 字段

对于 mysqlType 字段，Canal 官方实现中，对于含有参数的类型，会包含完整的参数信息，TiCDC 实现则没有类型参数信息。

在下面示例的表定义 SQL 语句中，如 decimal / char / varchar / enum 等类型，都含有参数。对比 TiCDC 和 Canal 官方实现分别生成的 Canal-JSON 格式数据可知，在 mysqlType 字段中的数据，TiCDC 实现只包含基本 MySQL Type。如果业务需要类型参数信息，需要你自行通过其他方式实现。

假设在上游数据库按顺序执行如下 SQL 语句:

```
create table t (
    id     int auto_increment,
    c_decimal    decimal(10, 4) null,
    c_char       char(16)       null,
    c_varchar    varchar(16)    null,
    c_binary     binary(16)     null,
    c_varbinary  varbinary(16) null,
```

```
    c_enum enum('a','b','c') null,
    c_set  set('a','b','c')  null,
    c_bit  bit(64)            null,
    constraint pk
        primary key (id)
);

insert into t (c_decimal, c_char, c_varchar, c_binary, c_varbinary, c_enum, c_set, c_bit)
values (123.456, "abc", "abc", "abc", "abc", 'a', 'a,b', b'1000001');
```

TiCDC 输出内容如下：

```
{
    "id": 0,
    ...
    "isDdl": false,
    "sqlType": {
        ...
    },
    "mysqlType": {
        "c_binary": "binary",
        "c_bit": "bit",
        "c_char": "char",
        "c_decimal": "decimal",
        "c_enum": "enum",
        "c_set": "set",
        "c_varbinary": "varbinary",
        "c_varchar": "varchar",
        "id": "int"
    },
    "data": [
        {
            ...
        }
    ],
    "old": null,
}
```

Canal 官方实现输出内容如下：

```
{
    "id": 0,
    ...
    "isDdl": false,
    "sqlType": {
```

```
        ...
    },
    "mysqlType": {
        "c_binary": "binary(16)",
        "c_bit": "bit(64)",
        "c_char": "char(16)",
        "c_decimal": "decimal(10, 4)",
        "c_enum": "enum('a','b','c')",
        "c_set": "set('a','b','c')",
        "c_varbinary": "varbinary(16)",
        "c_varchar": "varchar(16)",
        "id": "int"
    },
    "data": [
        {
            ...
        }
    ],
    "old": null,
}
```

12.11.9 TiDB 集成 Confluent Platform 快速上手指南

本文档介绍如何使用TiCDC 将 TiDB 集成到 Confluent Platform。

> **警告：**
>
> 当前该功能为实验特性，请勿在生产环境中使用。

Confluent Platform 是一个以 Apache Kafka 为核心的流数据处理平台，可以借助官方或第三方的 sink connector 将数据源连接到关系型或非关系型数据库。

你可以使用 TiCDC 组件和 Avro 协议来集成 TiDB 和 Confluent Platform。TiCDC 能将数据更改以 Confluent Platform 能识别的格式流式传输到 Kafka。下文详细介绍了集成的操作步骤。

12.11.9.1 环境准备

> **注意：**
>
> 本教程使用 JDBC sink connector 将 TiDB 的数据同步到下游的关系型数据库中。为了简化操作，此处以 SQLite 为例。

- 确保 Zookeeper、Kafka 和 Schema Registry 已正确安装。推荐参照 Confluent Platform 快速入门指南部署本地测试环境。
- 通过以下命令确保 JDBC sink connector 已安装。返回结果中预期包含 `jdbc-sink`。

```
confluent local services connect connector list
```

12.11.9.2 集成步骤

1. 将下方的配置样例保存为 `jdbc-sink-connector.json` 文件：

```
{
  "name": "jdbc-sink-connector",
  "config": {
    "connector.class": "io.confluent.connect.jdbc.JdbcSinkConnector",
    "tasks.max": "1",
    "topics": "testdb_test",
    "connection.url": "sqlite:test.db",
    "connection.ds.pool.size": 5,
    "table.name.format": "test",
    "auto.create": true,
    "auto.evolve": true
  }
}
```

2. 运行下方命令新建一个 JDBC sink connector 实例（假设 Kafka 监听的 IP 地址与端口是 `127.0.0.1:8083`）：

```
curl -X POST -H "Content-Type: application/json" -d @jdbc-sink-connector.json http
    ↪ ://127.0.0.1:8083/connectors
```

3. 通过以下任一方式部署 TiCDC。如果已经部署了 TiCDC，可以跳过这一步。

    - 使用 TiUP 部署包含 TiCDC 组件的全新 TiDB 集群
    - 使用 TiUP 在原有 TiDB 集群上新增 TiCDC 组件
    - 使用 binary 在原有 TiDB 集群上新增 TiCDC 组件（不推荐）

    在继续接下来的操作之前，请先确保 TiDB 和 TiCDC 集群处于健康状态。

4. 运行下面的 `cdc cli` 命令，新建一个同步任务 `changefeed`：

```
./cdc cli changefeed create --pd="http://127.0.0.1:2379" --sink-uri="kafka://127.0.0.1:9092/
    ↪ testdb_test?protocol=avro" --opts "registry=http://127.0.0.1:8081"
```

> **注意：**
>
> 请确保 PD、Kafka 和 Schema Registry 在各自的默认端口上运行。

12.11.9.3 数据同步测试

TiDB 与 Confluent Platform 成功集成后，你可以按照以下步骤来测试数据同步功能。

1. 在 TiDB 集群中新建 `testdb` 数据库：

```
CREATE DATABASE IF NOT EXISTS testdb;
```

在 `testdb` 数据库中创建 `test` 数据表：

```
USE testdb;
CREATE TABLE test (
    id INT PRIMARY KEY,
    v TEXT
);
```

> 注意：
> 如果需要修改数据库名或数据表名，要相应地修改 `jdbc-sink-connector.json` 文件中 `topics` 的参数值。

2. 向 `test` 数据表中插入数据：

```
INSERT INTO test (id, v) values (1, 'a');
INSERT INTO test (id, v) values (2, 'b');
INSERT INTO test (id, v) values (3, 'c');
INSERT INTO test (id, v) values (4, 'd');
```

3. 等待数据被同步到下游数据库中，并检查下游的数据：

```
sqlite3 test.db
sqlite> SELECT * from test;
```

12.11.10 TiCDC 术语表

本术语表提供 TiCDC 相关的术语和定义，这些术语会出现在 TiCDC 的日志、监控指标、配置和文档中。

12.11.10.1 B

12.11.10.1.1 变更数据

从上游 TiDB 集群写入 TiCDC 的数据，包括 DML 操作引发的数据变更和 DDL 操作引发的表结构变更。

12.11.10.2 C

12.11.10.2.1 Capture

单个 TiCDC 实例。多个 Capture 组成一个 TiCDC 集群，Capture 上运行集群中的同步任务。

12.11.10.2.2 Changefeed

TiCDC 中的单个同步任务。同步任务将一个 TiDB 集群中数张表的变更数据输出到一个指定的下游中。

12.11.10.3 O

12.11.10.3.1 Owner

一个特殊角色的Capture，负责管理 TiCDC 集群和调度 TiCDC 集群中的同步任务。该角色由 Capture 选举产生，在任意时刻最多只存在一个。

12.11.10.4 P

12.11.10.4.1 Processor

TiCDC 同步任务会在 TiCDC 实例上分配数据表，Processor 指这些数据表的同步处理单元。处理任务包括变更数据的拉取、排序、还原和分发。

12.12 sync-diff-inspector

12.12.1 sync-diff-inspector 用户文档

sync-diff-inspector 是一个用于校验 MySQL / TiDB 中两份数据是否一致的工具。该工具提供了修复数据的功能 (适用于修复少量不一致的数据)。

主要功能：

- 对比表结构和数据
- 如果数据不一致，则生成用于修复数据的 SQL 语句
- 支持不同库名或表名的数据校验
- 支持分库分表场景下的数据校验
- 支持TiDB 主从集群的数据校验
- 支持从 TiDB DM 拉取配置的数据校验

你可通过以下方式下载 sync-diff-inspector：

- Binary 包。点击 tidb-enterprise-tools-nightly-linux-amd64 进行下载。
- Docker 镜像。执行以下命令进行下载：

```
docker pull pingcap/tidb-enterprise-tools:nightly
```

12.12.1.1 sync-diff-inspector 的使用限制

- 对于 MySQL 和 TiDB 之间的数据同步不支持在线校验，需要保证上下游校验的表中没有数据写入，或者保证某个范围内的数据不再变更，通过配置 `range` 来校验这个范围内的数据。
- 不支持 JSON、 BIT、 BINARY、 BLOB 等类型的数据，在校验时需要设置 `ignore-columns` 忽略检查这些类型的数据。
- FLOAT、 DOUBLE 等浮点数类型在 TiDB 和 MySQL 中的实现方式不同，在计算 checksum 时会分别取 6 位和 15 位有效数字。如果不使用该特性，需要设置 `ignore-columns` 忽略这些列的检查。
- 支持对不包含主键或者唯一索引的表进行校验，但是如果数据不一致，生成的用于修复的 SQL 可能无法正确修复数据。

12.12.1.2 sync-diff-inspector 所需的数据库权限

sync-diff-inspector 需要获取表结构信息、查询数据，需要的数据库权限如下：

- 上游数据库
  - SELECT（查数据进行对比）
  - SHOW_DATABASES（查看库名）
  - RELOAD（查看表结构）
- 下游数据库
  - SELECT（查数据进行对比）
  - SHOW_DATABASES（查看库名）
  - RELOAD（查看表结构）

12.12.1.3 配置文件说明

sync-diff-inspector 的配置总共分为五个部分：

- Global config: 通用配置，包括校验的线程数量、是否输出修复 SQL、是否比对数据等。
- Datasource config: 配置上下游数据库实例。
- Routes: 上游多表名通过正则匹配下游单表名的规则。（可选）
- Task config: 配置校验哪些表，如果有的表在上下游有一定的映射关系或者有一些特殊要求，则需要对指定的表进行配置。
- Table config: 对具体表的特殊配置，例如指定范围、忽略的列等等。（可选）

下面是一个完整配置文件的说明：

- 提示：配置名后带 s 的配置项允许拥有多个配置值，因此需要使用方括号 `[]` 来包含配置值。

```
### Diff Configuration.

############################ Global config ############################

### 检查数据的线程数量，上下游数据库的连接数会略大于该值
check-thread-count = 4

### 如果开启，若表存在不一致，则输出用于修复的 SQL 语句。
export-fix-sql = true

### 只对比表结构而不对比数据
check-struct-only = false


############################ Datasource config ############################
[data-sources]
[data-sources.mysql1] # mysql1 是该数据库实例唯一标识的自定义 id，用于下面 task.source-instances/
    ↪ task.target-instance 中
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    # ( 可选 ) 使用映射规则来匹配上游多个分表，其中 rule1 和 rule2 在下面 Routes 配置栏中定义
    route-rules = ["rule1", "rule2"]

[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
    password = ""
    # ( 可选 ) 使用 TiDB 的 snapshot 功能，如果开启的话会使用历史数据进行对比
    # snapshot = "386902609362944000"

############################ Routes ############################
### 如果需要对比大量的不同库名或者表名的表的数据，或者用于校验上游多个分表与下游总表的数据，
    ↪ 可以通过 table-rule 来设置映射关系
### 可以只配置 schema 或者 table 的映射关系，也可以都配置
[routes]
[routes.rule1] # rule1 是该配置的唯一标识的自定义 id，用于上面 data-sources.route-rules 中
schema-pattern = "test_*"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "t_*"          # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test"         # 目标库名
target-table = "t" # 目标表名
```

```
[routes.rule2]
schema-pattern = "test2_*"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "t2_*"          # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test2"         # 目标库名
target-table = "t2" # 目标表名

########################### Task config ###########################
### 配置需要对比的*目标数据库*中的表
[task]
    # output-dir 会保存如下信息
    # 1 sql: 检查出错误后生成的修复 SQL 文件，并且一个 chunk 对应一个文件
    # 2 log: sync-diff.log 保存日志信息
    # 3 summary: summary.txt 保存总结
    # 4 checkpoint: a dir 保存断点续传信息
    output-dir = "./output"

    # 上游数据库，内容是 data-sources 声明的唯一标识 id
    source-instances = ["mysql1"]

    # 下游数据库，内容是 data-sources 声明的唯一标识 id
    target-instance = "tidb0"

    # 需要比对的下游数据库的表，每个表需要包含数据库名和表名，两者由 `.` 隔开
    # 使用 ? 来匹配任意一个字符；使用 * 来匹配任意；详细匹配规则参考 golang regexp pkg: https://
        ↪ github.com/google/re2/wiki/Syntax
    target-check-tables = ["schema*.table*", "!c.*", "test2.t2"]

    # ( 可选 ) 对部分表的额外配置，其中 config1 在下面 Table config 配置栏中定义
    target-configs = ["config1"]

########################### Table config ###########################
### 对部分表进行特殊的配置，配置的表必须包含在 task.target-check-tables 中
[table-configs.config1] # config1 是该配置的唯一标识自定义 id，用于上面 task.target-configs 中
### 目标表名称，可以使用正则来匹配多个表，但不允许存在一个表同时被多个特殊配置匹配。
target-tables = ["schema*.test*", "test2.t2"]
# ( 可选 ) 指定检查的数据的范围，需要符合 sql 中 where 条件的语法
range = "age > 10 AND age < 20"
# ( 可选 ) 指定用于划分 chunk 的列，如果不配置该项，sync-diff-inspector 会选取一些合适的列 (
    ↪ 主键 / 唯一键 / 索引 )
index-fields = ["col1","col2"]
# ( 可选 ) 忽略某些列的检查，例如 sync-diff-inspector 目前还不支持的一些类型 ( json，bit，blob 等
    ↪ ) ,
### 或者是浮点类型数据在 TiDB 和 MySQL 中的表现可能存在差异，可以使用 ignore-columns
    ↪ 忽略检查这些列
```

```
ignore-columns = ["",""]
# ( 可选 ) 指定划分该表的 chunk 的大小，若不指定可以删去或者将其配置为 0。
chunk-size = 0
# ( 可选 ) 指定该表的 collation，若不指定可以删去或者将其配置为空字符串。
collation = ""
```

12.12.1.4 运行 sync-diff-inspector

执行如下命令：

```
./bin/sync_diff_inspector --config=./config.toml
```

该命令最终会在 config.toml 中的 output-dir 输出目录输出本次比对的检查报告 summary.txt 和日志 sync_diff.log。在输出目录下还会生成由 config.toml 文件内容哈希值命名的文件夹，该文件夹下包括断点续传 checkpoint 结点信息以及数据存在不一致时生成的 SQL 修复数据。

12.12.1.4.1 前台输出

sync-diff-inspector 在执行过程中会往 stdout 发送进度信息。进度信息包括表的结构比较结果、表的数据比较结果以及进度条。

> 注意：
>
> 为了达成显示效果，请保持显示窗口宽度在 80 字符以上。

```
A total of 2 tables need to be compared

Comparing the table structure of ``sbtest`.`sbtest96`` ... equivalent
Comparing the table structure of ``sbtest`.`sbtest99`` ... equivalent
Comparing the table data of ``sbtest`.`sbtest96`` ... failure
Comparing the table data of ``sbtest`.`sbtest99`` ...
_____________________________________________________________________________
Progress [============================================================>--] 98% 193/200
```

```
A total of 2 tables need to be compared

Comparing the table structure of ``sbtest`.`sbtest96`` ... equivalent
Comparing the table structure of ``sbtest`.`sbtest99`` ... equivalent
Comparing the table data of ``sbtest`.`sbtest96`` ... failure
Comparing the table data of ``sbtest`.`sbtest99`` ... failure
_____________________________________________________________________________
Progress [============================================================>] 100% 0/0
The data of `sbtest`.`sbtest99` is not equal
```

```
The data of `sbtest`.`sbtest96` is not equal

The rest of tables are all equal.
The patch file has been generated in
        'output/fix-on-tidb2/'
You can view the comparision details through 'output/sync_diff.log'
```

12.12.1.4.2 输出文件

输出文件目录结构如下：

```
output/
|-- checkpoint # 保存断点续传信息
| |-- bbfec8cc8d1f58a5800e63aa73e5 # config hash 占位文件，标识该输出目录（output/）
    ↪ 对应的配置文件
| |-- DO_NOT_EDIT_THIS_DIR
| └-- sync_diff_checkpoints.pb # 断点续传信息
|
|-- fix-on-target # 保存用于修复不一致的 SQL 文件
| |-- xxx.sql
| |-- xxx.sql
| └-- xxx.sql
|
|-- summary.txt # 保存校验结果的总结
└-- sync_diff.log # 保存 sync-diff-inspector 执行过程中输出的日志信息
```

日志

sync-diff-inspector 的日志存放在 `${output}/sync_diff.log` 中，其中 `${output}` 是 `config.toml` 文件中 `output-dir` 的值。

校验进度

sync-diff-inspector 会在运行时定期（间隔 10s）输出校验进度到 checkpoint 中 (位于 `${output}/checkpoint/
↪ sync_diff_checkpoints.pb` 中，其中 `${output}` 是 `config.toml` 文件中 `output-dir` 的值。

校验结果

当校验结束时，sync-diff-inspector 会输出一份校验报告，位于 `${output}/summary.txt` 中，其中 `${output}` 是 `config.toml` 文件中 `output-dir` 的值。

```
+---------------------+--------------------+----------------+
|        TABLE        | STRUCTURE EQUALITY | DATA DIFF ROWS |
+---------------------+--------------------+----------------+
| `sbtest`.`sbtest99` | true               | +97/-97        |
| `sbtest`.`sbtest96` | true               | +0/-101        |
+---------------------+--------------------+----------------+
Time Cost: 16.75370462s
Average Speed: 113.277149MB/s
```

- TABLE: 该列表示对应的数据库及表名
- STRUCTURE EQUALITY: 表结构是否相同
- DATA DIFF ROWS: 即 rowAdd / rowDelete ，表示该表修复需要增加/删除的行数

SQL 修复

校验过程中遇到不同的行，会生成修复数据的 SQL 语句。一个 chunk 如果出现数据不一致，就会生成一个以 chunk.Index 命名的 SQL 文件。文件位于 ${output}/fix-on-${instance} 文件夹下。其中 ${instance} 为 config.toml 中 task.target-instance 的值。

一个 SQL 文件会包含该 chunk 的所属表以及表示的范围信息。对每个修复 SQL 语句，有三种情况：

- 下游数据库缺失行，则是 REPLACE 语句
- 下游数据库冗余行，则是 DELETE 语句
- 下游数据库行部分数据不一致，则是 REPLACE 语句，但会在 SQL 文件中通过注释的方法标明不同的列

```
-- table: sbtest.sbtest99
-- range in sequence: (3690708) < (id) <= (3720581)
/*
  DIFF COLUMNS │   `K`   │                 `C`                 │               `PAD`
───────────────┼─────────┼─────────────────────────────────────┼─────────────────────────────────────
    ↪
  source data  │ 2501808 │ 'hello'                             │ 'world'
───────────────┼─────────┼─────────────────────────────────────┼─────────────────────────────────────
    ↪
  target data  │ 5003616 │ '0709824117-9809973320-4456050422' │
      ↪ '1714066100-7057807621-1425865505'
───────────────┼─────────┼─────────────────────────────────────┼─────────────────────────────────────
    ↪
*/
REPLACE INTO `sbtest`.`sbtest99`(`id`,`k`,`c`,`pad`) VALUES (3700000,2501808,'hello','world');
```

12.12.1.5 注意事项

- sync-diff-inspector 在校验数据时会消耗一定的服务器资源，需要避免在业务高峰期间校验。
- 在数据对比前，需要注意表中的 collation 设置。如果表的主键或唯一键为 varchar 类型，且上下游数据库中 collation 设置不同，可能会因为排序问题导致最终校验结果不正确，需要在 sync-diff-inspector 的配置文件中增加 collation 设置。
- sync-diff-inspector 会优先使用 TiDB 的统计信息来划分 chunk，需要尽量保证统计信息精确，可以在业务空闲期手动执行 analyze table {table_name}。
- table-rule 的规则需要特殊注意，例如设置了 schema-pattern="test1"，table-pattern = "t_1"，target ↪ -schema="test2"，target-table = "t_2"，会对比 source 中的表 test1.t_1 和 target 中的表 test2.t_2。sync-diff-inspector 默认开启 sharding，如果 source 中还有表 test2.t_2，则会把 source 端的表 test1.t_1 和表 test2.t_2 作为 sharding 与 target 中的表 test2.t_2 进行一致性校验。
- 生成的 SQL 文件仅作为修复数据的参考，需要确认后再执行这些 SQL 修复数据。

12.12.2 不同库名或表名的数据校验

当你在使用TiDB DM 等同步工具时，可以设置 route-rules 将数据同步到下游指定表中。sync-diff-inspector 通过设置 rules 提供了校验不同库名、表名的表的功能。

下面是一个简单的配置文件说明，要了解完整配置，请参考sync-diff-inspector 用户文档。

```
########################### Datasource config ###########################
[data-sources.mysql1]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule1"]

[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
    password = ""

############################## Routes ############################
[routes.rule1]
schema-pattern = "test_1"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "t_1"          # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test_2"         # 目标库名
target-table = "t_2" # 目标表名
```

使用该配置会对下游的 test_2.t_2 与实例 mysql1 中的 test_1.t_1 进行校验。

如果需要校验大量的不同库名或者表名的表，也可以通过 rules 设置映射关系来简化配置。可以只配置 schema 或者 table 的映射关系，也可以都配置。例如上游库 test_1 中的所有表都同步到了下游的 test_2 库中，可以使用如下配置进行校验：

```
########################### Datasource config ###########################
[data-sources.mysql1]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule1"]

[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
```

```
    password = ""

############################ Routes ##########################
[routes.rule1]
schema-pattern = "test_1"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "*"            # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test_2"       # 目标库名
target-table = "t_2"           # 目标表名
```

12.12.2.1 注意事项

如果上游数据库有 `test_2.t_2` 也会被下游数据库匹配到。

12.12.3 分库分表场景下的数据校验

sync-diff-inspector 支持对分库分表场景进行数据校验。例如有多个 MySQL 实例，当你使用同步工具TiDB DM 同步到一个 TiDB 时，可以使用 sync-diff-inspector 对上下游数据进行校验。

12.12.3.1 使用 datasource config 进行配置

使用 `Datasource config` 对 `table-0` 进行特殊配置，设置对应 `rules`，配置上游表与下游表的映射关系。这种配置方式需要对所有分表进行设置，适合上游分表数量较少，且分表的命名规则没有规律的场景。场景如图所示：

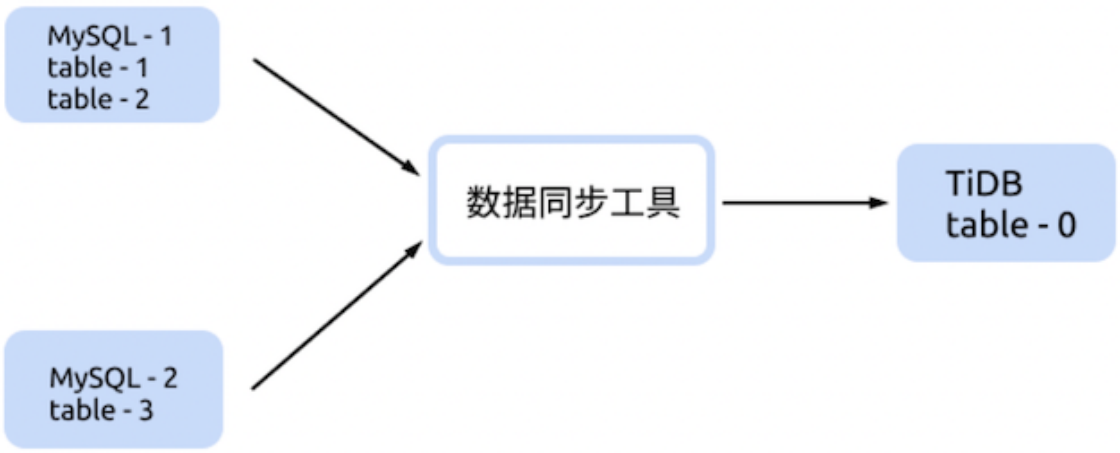


图 325: shard-table-sync-1

sync-diff-inspector 完整的示例配置如下：

```
### Diff Configuration.

############################ Global config ############################

### 检查数据的线程数量，上下游数据库的连接数会略大于该值
check-thread-count = 4

### 如果开启，若表存在不一致，则输出用于修复的 SQL 语句
export-fix-sql = true

### 只对比表结构而不对比数据
check-struct-only = false


############################ Datasource config ############################
[data-sources.mysql1]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule1"]

[data-sources.mysql2]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule2"]

[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
    password = ""

############################## Routes ############################
[routes.rule1]
schema-pattern = "test"        # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "table-[1-2]"  # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test"         # 目标库名
target-table = "table-0"       # 目标表名
```

```
[routes.rule2]
schema-pattern = "test"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "table-3"    # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test"       # 目标库名
target-table = "table-0"     # 目标表名

########################### Task config ###########################
[task]
    output-dir = "./output"

    source-instances = ["mysql1", "mysql2"]

    target-instance = "tidb0"

    # 需要比对的下游数据库的表，每个表需要包含数据库名和表名，两者由 `.` 隔开
    target-check-tables = ["test.table-0"]
```

当上游分表较多，且所有分表的命名都符合一定的规则时，则可以使用 `table-rules` 进行配置。场景如图所示：

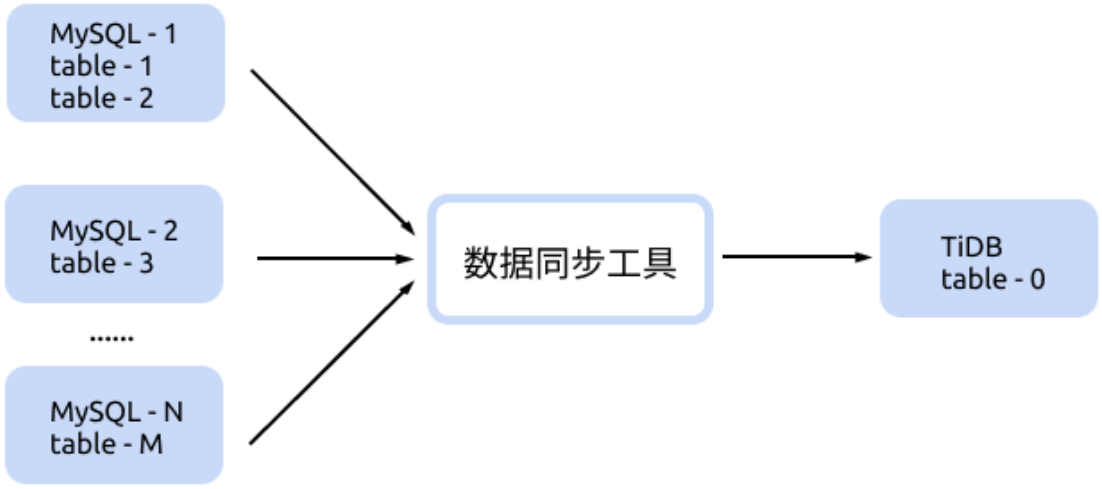


图 326: shard-table-sync-2

sync-diff-inspector 完整的示例配置如下：

```
### Diff Configuration.

########################### Global config ###########################

### 检查数据的线程数量，上下游数据库的连接数会略大于该值
check-thread-count = 4
```

```
### 如果开启，若表存在不一致，则输出用于修复的 SQL 语句
export-fix-sql = true

### 只对比表结构而不对比数据
check-struct-only = false


########################### Datasource config ########################
[data-sources.mysql1]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule1"]

[data-sources.mysql2]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""

    route-rules = ["rule1"]

[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
    password = ""

############################## Routes ############################
[routes.rule1]
schema-pattern = "test"      # 匹配数据源的库名，支持通配符 "*" 和 "?"
table-pattern = "table-*"    # 匹配数据源的表名，支持通配符 "*" 和 "?"
target-schema = "test"       # 目标库名
target-table = "table-0"     # 目标表名

########################### Task config ##########################
[task]
    output-dir = "./output"

    source-instances = ["mysql1", "mysql2"]

    target-instance = "tidb0"
```

```
    # 需要比对的下游数据库的表，每个表需要包含数据库名和表名，两者由 `.` 隔开
    target-check-tables = ["test.table-0"]
```

12.12.3.2 注意事项

如果上游数据库有 `test.table-0` 也会被下游数据库匹配到。

12.12.4 TiDB 主从集群的数据校验

在你使用 TiDB Binlog 搭建 TiDB 的主从集群，Drainer 在把数据同步到 TiDB 时，保存 checkpoint 的同时也会将上下游的 TSO 对应关系保存为 `ts-map`。在 sync-diff-inspector 中配置 `snapshot` 即可对 TiDB 主从集群的数据进行校验。

12.12.4.1 获取 ts-map

在下游 TiDB 中执行以下 SQL 语句：

```
select * from tidb_binlog.checkpoint;
```

```
+--------------------+-----------------------------------------------------------------------------------------------
    ↪
| clusterID          | checkPoint
    ↪
    ↪ |
+--------------------+-----------------------------------------------------------------------------------------------
    ↪
| 6711243465327639221 | {"commitTS":409622383615541249,"ts-map":{"primary-ts
    ↪ ":409621863377928194,"secondary-ts":409621863377928345}} |
+--------------------+-----------------------------------------------------------------------------------------------
    ↪
```

从结果中可以获取 ts-map 信息。

12.12.4.2 配置 snapshot

使用上一步骤获取的 ts-map 信息来配置上下游数据库的 snapshot 信息。其中的 `Datasource config` 部分示例配置如下：

```
########################### Datasource config ########################
[data-sources.mysql1]
    host = "127.0.0.1"
    port = 3306
    user = "root"
    password = ""
    snapshot = "409621863377928345"
```

```
[data-sources.tidb0]
    host = "127.0.0.1"
    port = 4000
    user = "root"
    password = ""
    snapshot = "409621863377928345"
```

12.12.4.3 注意事项

- Drainer 的 db-type 需要设置为 tidb，这样才会在 checkpoint 中保存 ts-map。
- 需要调整 TiKV 的 GC 时间，保证在校验时 snapshot 对应的历史数据不会被执行 GC。建议调整为 1 个小时，在校验后再还原 GC 设置。
- 以上配置只展示 Datasource config 部分，并不完全。完整配置请参考sync-diff-inspector 用户文档。

12.12.5 基于 DM 同步场景下的数据校验

当你在使用TiDB DM 等同步工具时，需要校验 DM 同步后数据的一致性。你可以从 DM-master 拉取指定 task-name 的配置，进行数据校验。

下面是一个简单的配置文件说明，要了解完整配置，请参考sync-diff-inspector 用户文档。

```
### Diff Configuration.

########################### Global config ###########################

### 检查数据的线程数量，上下游数据库的连接数会略大于该值
check-thread-count = 4

### 如果开启，若表存在不一致，则输出用于修复的 SQL 语句
export-fix-sql = true

### 只对比表结构而不对比数据
check-struct-only = false

### dm-master 的地址，格式为 "http://127.0.0.1:8261"
dm-addr = "http://127.0.0.1:8261"

### 指定 DM 的 `task-name`
dm-task = "test"

########################### Task config ###########################
[task]
    output-dir = "./output"

    # 需要比对的下游数据库的表，每个表需要包含数据库名和表名，两者由 `.` 隔开
```

```
    target-check-tables = ["hb_test.*"]
```

该配置在 dm-task = “test” 中，会对该任务下 hb_test 库的所有表进行检验，自动从 DM 配置中获取上游对下游库名的正则匹配，以校验 DM 同步后数据的一致性。

12.13 TiSpark

12.13.1 TiSpark 快速上手

为了让大家快速体验TiSpark，通过 TiUP 安装的 TiDB 集群中默认已集成 Spark 和 TiSpark jar 包。

12.13.1.1 部署信息

- Spark 默认部署在 TiDB 实例部署目录下 spark 目录中
- TiSpark jar 包默认部署在 Spark 部署目录 jars 文件夹下：`spark/jars/tispark-${name_with_version}.jar`
- TiSpark 示例数据和导入脚本可点击 TiSpark 示例数据下载。

```
tispark-sample-data/
```

12.13.1.2 环境准备

12.13.1.2.1 在 TiDB 实例上安装 JDK

在 Oracle JDK 官方下载页面下载 JDK 1.8 当前最新版，本示例中下载的版本为 `jdk-8u141-linux-x64.tar.gz`。

解压并根据您的 JDK 部署目录设置环境变量，编辑 `~/.bashrc` 文件，比如：

```
export JAVA_HOME=/home/pingcap/jdk1.8.0_144 &&
export PATH=$JAVA_HOME/bin:$PATH
```

验证 JDK 有效性：

```
java -version
```

```
java version "1.8.0_144"
Java(TM) SE Runtime Environment (build 1.8.0_144-b01)
Java HotSpot(TM) 64-Bit Server VM (build 25.144-b01, mixed mode)
```

12.13.1.2.2 导入样例数据

假设 TiDB 集群已启动，其中一台 TiDB 实例服务 IP 为 192.168.0.2，端口为 4000，用户名为 root, 密码为空。

```
wget http://download.pingcap.org/tispark-sample-data.tar.gz && \
tar -zxvf tispark-sample-data.tar.gz && \
cd tispark-sample-data
```

修改 sample_data.sh 中 TiDB 登录信息，比如：

```
mysql --local-infile=1 -h 192.168.0.2 -P 4000 -u root < dss.ddl
```

执行脚本

```
./sample_data.sh
```

> **注意：**
>
> 执行脚本的机器上需要安装 MySQL client，CentOS 用户可通过 `yum -y install mysql`来安装。

登录 TiDB 并验证数据包含 TPCH_001 库及以下表：

```
mysql -uroot -P4000 -h192.168.0.2
```

```
show databases;
```

```
+--------------------+
| Database           |
+--------------------+
| INFORMATION_SCHEMA |
| PERFORMANCE_SCHEMA |
| TPCH_001           |
| mysql              |
| test               |
+--------------------+
5 rows in set (0.00 sec)
```

```
use TPCH_001;
```

```
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
```

```
show tables;
```

```
+--------------------+
| Tables_in_TPCH_001 |
+--------------------+
| CUSTOMER           |
| LINEITEM           |
```

```
| NATION             |
| ORDERS             |
| PART               |
| PARTSUPP           |
| REGION             |
| SUPPLIER           |
+--------------------+
8 rows in set (0.00 sec)
```

12.13.1.3 使用范例

进入 spark 部署目录启动 spark-shell:

```
cd spark &&
bin/spark-shell
```

然后像使用原生 Spark 一样查询 TiDB 表：

```
scala> spark.sql("select count(*) from lineitem").show
```

结果为

```
+--------+
|count(1)|
+--------+
|   60175|
+--------+
```

下面执行另一个复杂一点的 Spark SQL:

```
scala> spark.sql(
      """select
        |   l_returnflag,
        |   l_linestatus,
        |   sum(l_quantity) as sum_qty,
        |   sum(l_extendedprice) as sum_base_price,
        |   sum(l_extendedprice * (1 - l_discount)) as sum_disc_price,
        |   sum(l_extendedprice * (1 - l_discount) * (1 + l_tax)) as sum_charge,
        |   avg(l_quantity) as avg_qty,
        |   avg(l_extendedprice) as avg_price,
        |   avg(l_discount) as avg_disc,
        |   count(*) as count_order
        |from
        |   lineitem
        |where
        |   l_shipdate <= date '1998-12-01' - interval '90' day
        |group by
```

```
        |   l_returnflag,
        |   l_linestatus
        |order by
        |   l_returnflag,
        |   l_linestatus
      """.stripMargin).show
```

结果为：

```
+------------+------------+---------+--------------+--------------+
|l_returnflag|l_linestatus|  sum_qty|sum_base_price|sum_disc_price|
+------------+------------+---------+--------------+--------------+
|           A|           F|380456.00|  532348211.65|505822441.4861|
|           N|           F|  8971.00|   12384801.37| 11798257.2080|
|           N|           O|742802.00| 1041502841.45|989737518.6346|
|           R|           F|381449.00|  534594445.35|507996454.4067|
+------------+------------+---------+--------------+--------------+
(续)
-----------------+---------+------------+--------+-----------+
       sum_charge|  avg_qty|   avg_price|avg_disc|count_order|
-----------------+---------+------------+--------+-----------+
 526165934.000839|25.575155|35785.709307|0.050081|      14876|
  12282485.056933|25.778736|35588.509684|0.047759|        348|
1029418531.523350|25.454988|35691.129209|0.049931|      29181|
 528524219.358903|25.597168|35874.006533|0.049828|      14902|
-----------------+---------+------------+--------+-----------+
```

更多样例请参考 pingcap/tispark-test

12.13.2 TiSpark 用户指南

TiSpark 是 PingCAP 为解决用户复杂 OLAP 需求而推出的产品。它借助 Spark 平台，同时融合 TiKV 分布式集群的优势，和 TiDB 一起为用户一站式解决 HTAP (Hybrid Transactional/Analytical Processing) 的需求。TiSpark 依赖于 TiKV 集群和 Placement Driver (PD)，也需要你搭建一个 Spark 集群。

本文简单介绍如何部署和使用 TiSpark。本文假设你对 Spark 有基本认知。你可以参阅 Apache Spark 官网了解 Spark 的相关信息。

12.13.2.1 概述

TiSpark 是将 Spark SQL 直接运行在分布式存储引擎 TiKV 上的 OLAP 解决方案。其架构图如下：

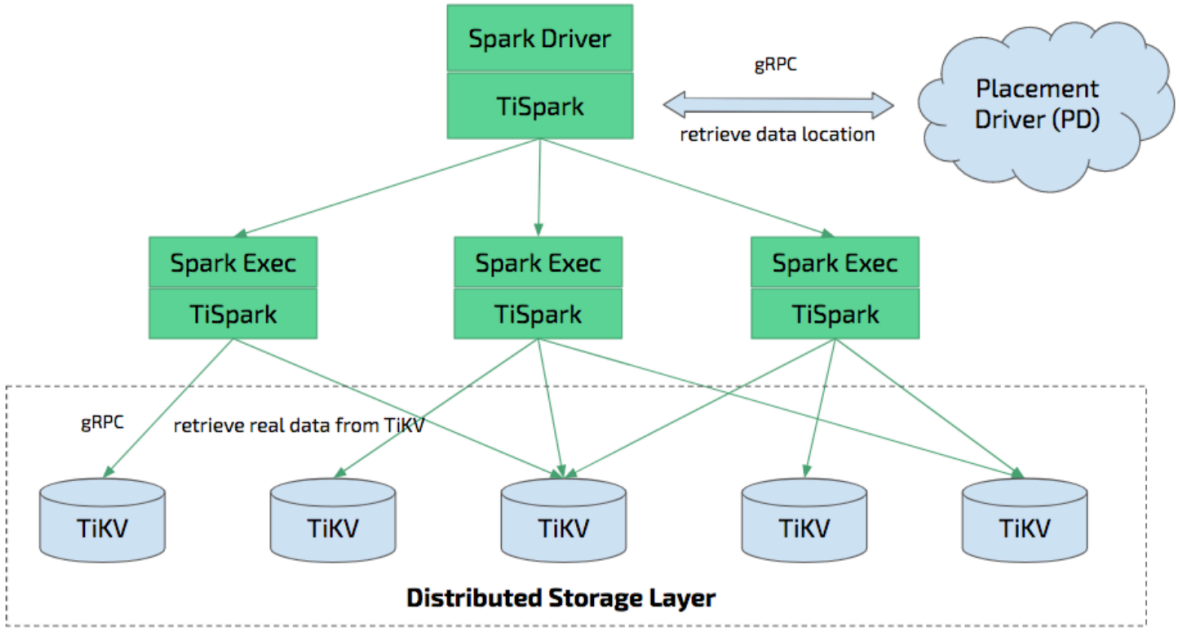


图 327: TiSpark Architecture

- TiSpark 深度整合了 Spark Catalyst 引擎，可以对计算提供精确的控制，使 Spark 能够高效的读取 TiKV 中的数据，提供索引支持以实现高速的点查。
- 通过多种计算下推减少 Spark SQL 需要处理的数据大小，以加速查询；利用 TiDB 的内建的统计信息选择更优的查询计划。
- 从数据集群的角度看，TiSpark + TiDB 可以让用户无需进行脆弱和难以维护的 ETL，直接在同一个平台进行事务和分析两种工作，简化了系统架构和运维。
- 用户借助 TiSpark 项目可以在 TiDB 上使用 Spark 生态圈提供的多种工具进行数据处理。例如，使用 TiSpark 进行数据分析和 ETL；使用 TiKV 作为机器学习的数据源；借助调度系统产生定时报表等等。
- 除此之外，TiSpark 还提供了分布式写入 TiKV 的功能。相比使用 Spark 结合 JDBC 的方式写入 TiDB，分布式写入 TiKV 可以实现事务（要么全部数据写入成功，要么全部都写入失败），并且写入速度会更快。

12.13.2.2 环境准备

目前 TiSpark 支持版本的兼容情况如下所示，你可以根据需要选择相应的 TiSpark 版本。

TiSpark 版本	TiDB、TiKV、PD 版本	Spark 版本	Scala 版本
2.4.x-scala_2.11	5.x，4.x	2.3.x，2.4.x	2.11
2.4.x-scala_2.12	5.x，4.x	2.4.x	2.12
2.5.x	5.x，4.x	3.0.x，3.1.x	2.12

TiSpark 可以在 YARN，Mesos，Standalone 等任意 Spark 模式下运行。

12.13.2.3 推荐配置

本部分描述了 TiKV 与 TiSpark 集群分开部署、Spark 与 TiSpark 集群独立部署，以及 TiSpark 与 TiKV 集群混合部署的建议配置。

12.13.2.3.1 TiKV 与 TiSpark 集群分开部署的配置

对于 TiKV 与 TiSpark 分开部署的场景，可以参考如下建议配置：

- 硬件配置建议

  普通场景可以参考TiDB 和 TiKV 硬件配置建议，但是如果是偏重分析的场景，可以将 TiKV 节点增加到至少 64G 内存。

12.13.2.3.2 Spark 与 TiSpark 集群独立部署的配置

关于 Spark 的详细硬件推荐配置请参考官网，如下是 TiSpark 所需环境的简单描述：

Spark 推荐 32G 内存以上的配额。请在配置中预留 25% 的内存给操作系统。

Spark 推荐每台计算节点配备 CPU 累计 8 到 16 核以上。你可以初始设定分配所有 CPU 核给 Spark。

Spark 的具体配置方式也请参考官方说明。以下为根据 `spark-env.sh` 配置的范例：

```
SPARK_EXECUTOR_CORES: 5
SPARK_EXECUTOR_MEMORY: 10g
SPARK_WORKER_CORES: 5
SPARK_WORKER_MEMORY: 10g
```

在 `spark-defaults.conf` 中，增加如下配置：

```
spark.tispark.pd.addresses $your_pd_servers
spark.sql.extensions org.apache.spark.sql.TiExtensions
```

在 CDH spark 版本中添加如下配置：

```
spark.tispark.pd.addresses=$your_pd_servers
spark.sql.extensions=org.apache.spark.sql.TiExtensions
```

`your_pd_servers` 是用逗号分隔的 PD 地址，每个地址使用 地址:端口 的格式。

例如你有一组 PD 在`10.16.20.1,10.16.20.2,10.16.20.3`,那么 PD 配置格式是`10.16.20.1:2379,10.16.20.2:2379,10.16.20.3:23`
↪ 。

12.13.2.3.3 TiSpark 与 TiKV 集群混合部署的配置

对于 TiKV 与 TiSpark 混合部署的场景，需在原有 TiKV 预留资源之外累加 Spark 所需部分，并分配 25% 的内存作为系统本身占用。

12.13.2.4 部署 TiSpark

TiSpark 的 jar 包可以在 TiSpark Releases 页面下载对应版本的 jar 包并拷贝到合适的目录。

12.13.2.4.1 已有 Spark 集群的部署方式

如果在已有 Spark 集群上运行 TiSpark，无需重启集群。可以使用 Spark 的 `--jars` 参数将 TiSpark 作为依赖引入：

```
spark-shell --jars $TISPARK_FOLDER/tispark-${name_with_version}.jar
```

12.13.2.4.2 没有 Spark 集群的部署方式

如果没有使用中的 Spark 集群，推荐使用 Spark Standalone 方式部署。这里简单介绍下 Standalone 部署方式。如果遇到问题，可以去官网寻求帮助；也欢迎在 GitHub 上提 issue。

下载安装包并安装

你可以在 Download Apache Spark™ 页面下载 Apache Spark。

对于 Standalone 模式且无需 Hadoop 支持，则选择 Spark 2.3.x 或者 Spark 2.4.x 且带有 Hadoop 依赖的 Pre-build with Apache Hadoop 2.x 任意版本。如有需要配合使用的 Hadoop 集群，则选择对应的 Hadoop 版本号。你也可以选择从源代码自行构建以配合官方 Hadoop 2.x 之前的版本。

如果你已经有了 Spark 二进制文件，并且当前 PATH 为 SPARKPATH，需将 TiSpark jar 包拷贝到 `${SPARKPATH}/jars` 目录下。

启动 Master

在选中的 Spark Master 节点执行如下命令：

```
cd $SPARKPATH
```

```
./sbin/start-master.sh
```

在这步完成以后，屏幕上会打印出一个 log 文件。检查 log 文件确认 Spark-Master 是否启动成功。你可以打开 http://spark-master-hostname:8080 查看集群信息（如果你没有改动 Spark-Master 默认 Port Numebr）。在启动 Spark-Worker 的时候，也可以通过这个面板来确认 Worker 是否已经加入集群。

启动 Worker

类似地，可以用如下命令启动 Spark-Worker 节点：

```
./sbin/start-slave.sh spark://spark-master-hostname:7077
```

命令返回以后，即可通过刚才的面板查看这个 Worker 是否已经正确地加入了 Spark 集群。在所有 Worker 节点重复刚才的命令。确认所有的 Worker 都可以正确连接 Master，这样你就拥有了一个 Standalone 模式的 Spark 集群。

Spark SQL shell 和 JDBC 服务器

当前版本的 TiSpark 可以直接使用 `spark-sql` 和 Spark 的 ThriftServer JDBC 服务器。

12.13.2.5 一个使用范例

假设你已经按照上述步骤成功启动了 TiSpark 集群，下面简单介绍如何使用 Spark SQL 来做 OLAP 分析。这里我们用名为 tpch 数据库中的 lineitem 表作为范例。

假设你的 PD 节点位于 192.168.1.100，端口为 2379，在 `$SPARK_HOME/conf/spark-defaults.conf` 加入：

```
spark.tispark.pd.addresses 192.168.1.100:2379
```

```
spark.sql.extensions org.apache.spark.sql.TiExtensions
```

然后在 Spark-Shell 里像原生 Spark 一样输入下面的命令：

```
spark.sql("use tpch")
```

```
spark.sql("select count(*) from lineitem").show
```

结果为：

```
+-------------+
| Count (1)   |
+-------------+
| 600000000   |
+-------------+
```

Spark SQL 交互 Shell 和原生 Spark 一致：

```
spark-sql> use tpch;
```

```
Time taken: 0.015 seconds
```

```
spark-sql> select count(*) from lineitem;
```

```
2000
Time taken: 0.673 seconds, Fetched 1 row(s)
```

SQuirreLSQL 和 hive-beeline 可以使用 JDBC 连接 Thrift 服务器。例如，使用 beeline 连接：

```
./beeline
```

```
Beeline version 1.2.2 by Apache Hive
```

```
beeline> !connect jdbc:hive2://localhost:10000
```

```
1: jdbc:hive2://localhost:10000> use testdb;
```

```
+---------+--+
| Result  |
+---------+--+
+---------+--+
No rows selected (0.013 seconds)
```

```
select count(*) from account;
```

```
+-----------+--+
| count(1)  |
+-----------+--+
| 1000000   |
+-----------+--+
1 row selected (1.97 seconds)
```

12.13.2.6 和 Hive 一起使用 TiSpark

TiSpark 可以和 Hive 混合使用。在启动 Spark 之前，需要添加 HADOOP_CONF_DIR 环境变量指向 Hadoop 配置目录并且将 `hive-site.xml` 拷贝到 `$SPARK_HOME/conf` 目录下。

```
val tisparkDF = spark.sql("select * from tispark_table").toDF
tisparkDF.write.saveAsTable("hive_table") // save table to hive
spark.sql("select * from hive_table a, tispark_table b where a.col1 = b.col1").show // join table
    ↪  across Hive and Tispark
```

12.13.2.7 通过 TiSpark 将 DataFrame 批量写入 TiDB

TiSpark 从 v2.3 版本开始原生支持将 DataFrame 批量写入 TiDB 集群，该写入模式通过 TiKV 的两阶段提交协议实现。

TiSpark 批量写入相比 Spark + JDBC 写入，有以下特点：

比较的方面	TiSpark 批量写入	Spark + JDBC 写入
原子性	DataFrame 的数据要么全部写入成功，要么全部写入失败	如果在写入过程中 spark 任务失败退出，会出现部分数据写入成功的情况
隔离性	写入过程中其他事务对正在写入的数据不可见	写入过程中其他事务能看到部分写入成功的数据
错误恢复	失败后只需要重新运行 Spark 程序	需要业务来实现幂等，例如失败后需要先清理部分写入成功的数据，再重新运行 Spark 程序，并且需要设置 `spark.task.maxFailures=1`，防止 task 内重试导致数据重复

比较的方面	TiSpark 批量写入	Spark + JDBC 写入
速度	直接写入 TiKV，速度更快	通过 TiDB 再写入 TiKV，对速度会有影响

以下通过 scala API 演示如何使用 TiSpark 批量写入：

```
// select data to write
val df = spark.sql("select * from tpch.ORDERS")

// write data to tidb
df.write.
  format("tidb").
  option("tidb.addr", "127.0.0.1").
  option("tidb.port", "4000").
  option("tidb.user", "root").
  option("tidb.password", "").
  option("database", "tpch").
  option("table", "target_orders").
  mode("append").
  save()
```

如果写入的数据量比较大，且写入时间超过 10 分钟，则需要保证 GC 时间大于写入时间。

```
update mysql.tidb set VARIABLE_VALUE="6h" where VARIABLE_NAME="tikv_gc_life_time";
```

详细使用手册请参考该文档。

12.13.2.8 通过 JDBC 将 DataFrame 写入 TiDB

除了使用 TiSpark 将 DataFrame 批量写入 TiDB 集群以外，也可以使用 Spark 原生的 JDBC 支持进行写入：

```
import org.apache.spark.sql.execution.datasources.jdbc.JDBCOptions

val customer = spark.sql("select * from customer limit 100000")
// you might repartition source to make it balance across nodes
// and increase concurrency
val df = customer.repartition(32)
df.write
.mode(saveMode = "append")
.format("jdbc")
.option("driver", "com.mysql.jdbc.Driver")
 // replace host and port as your and be sure to use rewrite batch
.option("url", "jdbc:mysql://127.0.0.1:4000/test?rewriteBatchedStatements=true")
.option("useSSL", "false")
```

```
// As tested, 150 is good practice
.option(JDBCOptions.JDBC_BATCH_INSERT_SIZE, 150)
.option("dbtable", s"cust_test_select") // database name and table name here
.option("isolationLevel", "NONE") // recommended to set isolationLevel to NONE if you have a
    ↪ large DF to load.
.option("user", "root") // TiDB user here
.save()
```

推荐将 `isolationLevel` 设置为 `NONE`，否则单一大事务有可能造成 TiDB 服务器内存溢出。

> **注意：**
>
> TiSpark 使用 JDBC 时默认 `isolationLevel` 为 `READ_UNCOMMITTED`，会造成事务隔离级别不支持的错误。推荐将 `isolationLevel` 设置为 `NONE`。

12.13.2.9 统计信息

TiSpark 可以使用 TiDB 的统计信息：

1. 选择代价最低的索引或扫表访问
2. 估算数据大小以决定是否进行广播优化

如果希望使用统计信息支持，需要确保所涉及的表已经被分析。请阅读这份文档了解如何进行表分析。

从 TiSpark 2.0 开始，统计信息将会默认被读取。

统计信息将在 Spark Driver 进行缓存，请确定 Driver 内存足够缓存统计信息。可以在`spark-defaults.conf`中开启或关闭统计信息读取：

Property Name	Default	Description
spark.tispark.statistics.auto_load	true	是否默认进行统计信息读取

12.13.2.10 安全

从 TiSpark v2.5.0 起，你可以通过 TiDB 对 TiSpark 进行鉴权与授权。

该功能默认关闭。要开启该功能，请在 Spark 配置文件 `spark-defaults.conf` 中添加以下配置项：

```
// 开启鉴权与授权功能
spark.sql.auth.enable   true

// 配置 TiDB 信息
spark.sql.tidb.addr     $your_tidb_server_address
spark.sql.tidb.port     $your_tidb_server_port
spark.sql.tidb.user     $your_tidb_server_user
```

```
spark.sql.tidb.password $your_tidb_server_password
```

更多详细信息，请参考 TiSpark 鉴权与授权文档。

> **注意：**
>
> 开启鉴权功能后，TiSpark Spark SQL 只能使用 TiDB 作为数据源，切换到其他数据源（例如 Hive）会导致数据表不可见。

12.13.2.11 TiSpark FAQ

- Q. 是独立部署还是和现有 Spark / Hadoop 集群共用资源？

  A. 可以利用现有 Spark 集群无需单独部署，但是如果现有集群繁忙，TiSpark 将无法达到理想速度。

- Q. 是否可以和 TiKV 混合部署？

  A. 如果 TiDB 以及 TiKV 负载较高且运行关键的线上任务，请考虑单独部署 TiSpark；并且考虑使用不同的网卡保证 OLTP 的网络资源不被侵占而影响线上业务。如果线上业务要求不高或者机器负载不大，可以考虑与 TiKV 混合部署。

- Q. Spark 执行中报 warning：WARN ObjectStore:568 - Failed to get database

  A. Warning 忽略即可，原因是 Spark 找不到对应的 hive 库，因为这个库是在 TIKV 中，而不是在 hive 中。可以考虑调整 log4j 日志，将该参数添加到 spark 下 conf 里 log4j 文件（如果后缀是 template 那先 mv 成后缀 properties）。

- Q. Spark 执行中报 java.sql.BatchUpdateException: Data Truncated

  A. 写入的数据长度超过了数据库定义的数据类型的长度，可以确认 target table 的字段长度，进行调整。

- Q. TiSpark 任务是否默认读取 Hive 的元数据？

  A. TiSpark 通过读取 hive-site 里的 meta 来搜寻 hive 的库。如果搜寻不到，就通过读取 tidb meta 搜寻 tidb 库。如果不需要该行为，可不在 hive site 中配置 hive 的 meta。

- Q. TiSpark 执行 Spark 任务时报：“Error：java.io.InvalidClassException: com.pingcap.tikv.region.TiRegion; local class incompatible: stream classdesc serialVersionUID …”

  A. 该报错日志中显示 serialVersionUID 冲突，说明存在不同版本的 class 和 TiRegion。因为 TiRegion 是 TiSpark 独有的，所以可能存在多个版本的 TiSpark 包。要解决该报错，请确保集群中各节点的 TiSpark 依赖包版本一致。

13 故障诊断

13.1 集群性能问题

13.1.1 使用 PLAN REPLAYER 保存和恢复集群现场信息

用户在定位排查 TiDB 集群问题时，经常需要提供系统和查询计划相关的信息。为了帮助用户更方便地获取相关信息，更高效地排查集群问题，TiDB 在 v5.3.0 中引入了 PLAN REPLAYER 命令，用于“一键”保存和恢复现场问题的相关信息，提升查询计划问题诊断的效率，同时方便将问题归档管理。

PLAN REPLAYER 主要功能如下：

- 导出排查现场 TiDB 集群的相关信息，导出为 ZIP 格式的文件用于保存。
- 在任意 TiDB 集群上导入另一 TiDB 集群现场信息的 ZIP 文件。

13.1.1.1 使用 PLAN REPLAYER 导出集群信息

你可以使用 PLAN REPLAYER 来保存 TiDB 集群的现场信息。导出接口如下：

```
PLAN REPLAYER DUMP EXPLAIN [ANALYZE] sql-statement;
```

TiDB 根据 sql-statement 整理出以下集群现场信息：

- TiDB 版本信息
- TiDB 配置信息
- TiDB Session 系统变量
- TiDB 执行计划绑定信息（SQL Binding）
- sql-statement 中所包含的表结构
- sql-statement 中所包含表的统计信息
- EXPLAIN [ANALYZE] sql-statement 的结果

> 注意：
>
> PLAN REPLAYER 不会导出表中数据

13.1.1.1.1 PLAN REPLAYER 导出示例

```
use test;
create table t(a int, b int);
insert into t values(1,1), (2, 2), (3, 3);
analyze table t;

plan replayer dump explain select * from t;
```

PLAN REPLAYER DUMP 会将以上信息打包整理成 ZIP 文件，并返回文件标识作为执行结果。该文件为一次性文件，被下载后 TiDB 会将其删除。

> **注意：**
>
> ZIP 文件最多会在 TiDB 集群中保存一个小时，超时后 TiDB 会将其删除。

```
MySQL [test]> plan replayer dump explain select * from t;
+------------------------------------------------------------------------+
| Dump_link                                                              |
+------------------------------------------------------------------------+
| replayer_single_JOGvpu4t7dssySqJfTtS4A==_1635750890568691080.zip |
+------------------------------------------------------------------------+
1 row in set (0.015 sec)
```

因为 MySQL Client 无法下载文件，所以需要通过 TiDB HTTP 接口和文件标识下载文件：

```
http://${tidb-server-ip}:${tidb-server-status-port}/plan_replayer/dump/${file_token}
```

其中，${tidb-server-ip}:${tidb-server-status-port} 是集群中任意 TiDB server 的地址。示例如下：

```
curl http://127.0.0.1:10080/plan_replayer/dump/replayer_single_JOGvpu4t7dssySqJfTtS4A==
    ↪ _1635750890568691080.zip > plan_replayer.zip
```

13.1.1.2 使用 PLAN REPLAYER 导入集群信息

> **警告：**
>
> PLAN REPLAYER 在一个 TiDB 集群上导入另一集群的现场信息，会修改导入集群的 TiDB Session 系统变量、执行计划绑定信息、表结构和统计信息。

有 PLAN REPLAYER 导出的 ZIP 文件后，用户便可以通过 PLAN REPLAYER 导入接口在任意 TiDB 集群上恢复另一集群地现场信息。语法如下：

```
PLAN REPLAYER LOAD 'file_name';
```

以上语句中，file_name 为要导入的 ZIP 文件名。

示例如下：

```
PLAN REPLAYER LOAD 'plan_replayer.zip';
```

13.1.2 TiDB 集群故障诊断

当试用 TiDB 遇到问题时，请先参考本篇文档。如果问题未解决，请按文档要求收集必要的信息通过 Github 提供给 TiDB 开发者。

13.1.2.1 如何给 TiDB 开发者报告错误

当使用 TiDB 遇到问题并且通过后面所列信息无法解决时，请收集以下信息并创建新 Issue:

- 具体的出错信息以及正在执行的操作
- 当前所有组件的状态
- 出问题组件 log 中的 error/fatal/panic 信息
- 机器配置以及部署拓扑
- dmesg 中 TiDB 组件相关的问题

13.1.2.2 数据库连接不上

首先请确认集群的各项服务是否已经启动，包括 tidb-server、 pd-server、 tikv-server。请用 ps 命令查看所有进程是否在。如果某个组件的进程已经不在了，请参考对应的章节排查错误。

如果所有的进程都在，请查看 tidb-server 的日志，看是否有报错？常见的错误包括：

- InformationSchema is out of date

  无法连接 tikv-server，请检查 pd-server 以及 tikv-server 的状态和日志。

- panic

  程序有错误，请将具体的 panic log 提供给 TiDB 开发者。

  如果是清空数据并重新部署服务，请确认以下信息：

- pd-server、 tikv-server 数据都已清空

  tikv-server 存储具体的数据，pd-server 存储 tikv-server 中数据的元信息。如果只清空 pd-server 或只清空 tikv-server 的数据，会导致两边数据不匹配。

- 清空 pd-server 和 tikv-server 的数据并重启后，也需要重启 tidb-server

  集群 ID 是由 pd-server 在集群初始化时随机分配，所以重新部署集群后，集群 ID 会发生变化。tidb-server 业务需要重启以获取新的集群 ID。

13.1.2.3 tidb-server 启动报错

tidb-server 无法启动的常见情况包括：

- 启动参数错误

  请参考 TiDB 命令行参数。

- 端口被占用：`lsof -i:port`

  请确保 tidb-server 启动所需要的端口未被占用。

- 无法连接 pd-server

  首先检查 pd-server 的进程状态和日志，确保 pd-server 成功启动，对应端口已打开：`lsof -i:port`。

  若 pd-server 正常，则需要检查 tidb-server 机器和 pd-server 对应端口之间的连通性，确保网段连通且对应服务端口已添加到防火墙白名单中，可通过 nc 或 curl 工具检查。

  例如，假设 tidb 服务位于 `192.168.1.100`，无法连接的 pd 位于 `192.168.1.101`，且 2379 为其 client port，则可以在 tidb 机器上执行 `nc -v -z 192.168.1.101 2379`，测试是否可以访问端口。或使用 `curl -v 192.168.1.101:2379/pd/api/v1/leader` 直接检查 pd 是否正常服务。

13.1.2.4 tikv-server 启动报错

- 启动参数错误

  请参考 TiKV 启动参数文档。

- 端口被占用：`lsof -i:port`

  请确保 tikv-server 启动所需要的端口未被占用：`lsof -i:port`。

- 无法连接 pd-server

  首先检查 pd-server 的进程状态和日志。确保 pd-server 成功启动，对应端口已打开：`lsof -i:port`。

  若 pd-server 正常，则需要检查 tikv-server 机器和 pd-server 对应端口之间的连通性，确保网段连通且对应服务端口已添加到防火墙白名单中，可通过 nc 或 curl 工具检查。具体命令参考上一节。

- 文件被占用

  不要在一个数据库文件目录上打开两个 tikv。

13.1.2.5 pd-server 启动报错

- 启动参数错误

  请参考 PD 命令行参数文档。

- 端口被占用：`lsof -i:port`

  请确保 pd-server 启动所需要的端口未被占用：`lsof -i:port`。

13.1.2.6 TiDB/TiKV/PD 进程异常退出

- 进程是否是启动在前台

  当前终端退出给其所有子进程发送 HUP 信号，从而导致进程退出。

- 是否是在命令行用过 nohup+& 方式直接运行

  这样依然可能导致进程因终端连接突然中断，作为终端 SHELL 的子进程被杀掉。

  推荐将启动命令写在脚本中，通过脚本运行（相当于二次 fork 启动）。

13.1.2.7 TiKV 进程异常重启

- 检查 dmesg 或者 syslog 里面是否有 OOM 信息

  如果有 OOM 信息并且杀掉的进程为 TiKV，请减少 TiKV 的 RocksDB 的各个 CF 的 `block-cache-size` 值。

- 检查 TiKV 日志是否有 panic 的 log

  提交 Issue 并附上 panic 的 log。

13.1.2.8 TiDB panic

请提供 panic 的 log。

13.1.2.9 连接被拒绝

- 请确保操作系统的网络参数正确，包括但不限于
  - 连接字符串中的端口和 tidb-server 启动的端口需要一致
  - 请保证防火墙的配置正确

13.1.2.10 Too many open files

在启动进程之前，请确保 `ulimit -n` 的结果足够大，推荐设为 unlimited 或者是大于 1000000。

13.1.2.11 数据库访问超时，系统负载高

首先检查SLOW-QUERY 日志，判断是否是因为某条 SQL 语句导致。如果未能解决，请提供如下信息：

- 部署的拓扑结构
  - tidb-server/pd-server/tikv-server 部署了几个实例
  - 这些实例在机器上是如何分布的
- 机器的硬件配置
  - CPU 核数
  - 内存大小
  - 硬盘类型（SSD 还是机械硬盘）
  - 是实体机还是虚拟机
- 机器上除了 TiDB 集群之外是否还有其他服务
- pd-server 和 tikv-server 是否分开部署
- 目前正在进行什么操作
- 用 `top -H` 命令查看当前占用 CPU 的线程名
- 最近一段时间的网络/IO 监控数据是否有异常

13.1.3 TiDB 集群问题导图

本篇文档总结了使用 TiDB 及其组件时的常见错误。遇到相关错误时，可以通过本文档的问题导图来排查错误原因并进行处理。

13.1.3.1 1. 服务不可用

13.1.3.1.1 1.1 客户端报 Region is Unavailable 错误

- 1.1.1 Region is Unavailable 一般是由于 Region 在一段时间不可用（可能会遇到 TiKV server is busy；或者发送给 TiKV 的请求由于 not leader 或者 epoch not match 等原因被打回；又或者请求 TiKV 超时等），TiDB 内部会进行 backoff 重试。backoff 的时间超过一定阈值（默认 20s）后就会报错给客户端。如果 backoff 在阈值内，客户端对该错误无感知。
- 1.1.2 多台 TiKV 同时内存不足 (OOM)，导致 Region 在一定时期内没有 Leader，见案例 case-991。
- 1.1.3 TiKV 报 TiKV server is busy 错误，超过 backoff 时间，参考4.3 客户端报 server is busy 错误。TiKV server is busy 属于内部流控机制，后续可能不计入 backoff 时间。
- 1.1.4 多台 TiKV 启动不了，导致 Region 没有 Leader。单台物理主机部署多个 TiKV 实例，一个物理机挂掉，由于 label 配置错误导致 Region 没有 Leader，见案例 case-228。
- 1.1.5 follower apply 落后，成为 Leader 之后把收到的请求以 epoch not match 理由打回，见案例 case-958（TiKV 内部需要优化该机制）。

13.1.3.1.2 1.2 PD 异常导致服务不可用

查看本文档5. PD 问题。

13.1.3.2 2. 延迟明显升高

13.1.3.2.1 2.1 延迟短暂升高

- 2.1.1 TiDB 执行计划不对导致延迟升高，请参考3.3 执行计划不对。
- 2.1.2 PD 出现选举问题或者 OOM 问题，请参考5.2 PD 选举问题和5.3 PD OOM 问题。
- 2.1.3 某些 TiKV 大量掉 Leader，请参考4.4 某些 TiKV 大量掉 Leader。

13.1.3.2.2 2.2 Latency 持续升高

- 2.2.1 TiKV 单线程瓶颈
  - 单个 TiKV Region 过多，导致单个 gRPC 线程成为瓶颈（查看监控：Grafana -> TiKV-details -> Thread CPU/gRPC CPU Per Thread），v3.x 以上版本可以开启 Hibernate Region 特性解决该问题，见案例 case-612。
  - v3.0 之前版本 Raftstore 单线程或者 apply 单线程到达瓶颈（查看监控：Grafana -> TiKV-details -> Thread CPU/raft store CPU 和 Async apply CPU 超过 80%）。可以选择扩容 TiKV（v2.x 版本）实例，或者升级到多线程模型的 v3.x 版本。
- 2.2.2 CPU load 升高。
- 2.2.3 TiKV 写入慢，请参考4.5 TiKV 写入慢。
- 2.2.4 TiDB 执行计划不对，请参考3.3 执行计划不对。

13.1.3.3 3. TiDB 问题

13.1.3.3.1 3.1 DDL

- 3.1.1 修改 `decimal` 字段长度时报错 `"ERROR 1105 (HY000): unsupported modify decimal column precision"`。TiDB 暂时不支持修改 `decimal` 字段长度。
- 3.1.2 TiDB DDL job 卡住不动/执行很慢（通过 `admin show ddl jobs` 可以查看 DDL 进度）：
  - 原因 1：与外部组件 (PD/TiKV) 的网络问题。
  - 原因 2：早期版本（v3.0.8 之前）TiDB 内部自身负载很重（高并发下可能产生了很多协程）。
  - 原因 3：早期版本（v2.1.15 & v3.0.0-rc1 之前）PD 实例删除 TiDB key 无效的问题，会导致每次 DDL 变更都需要等 2 个 lease（很慢）。
  - 其他未知原因，请上报 bug。
  - 解决方法：原因 1 需要检查与外部组件的网络问题；原因 2 和 3 已经修复，需要升级到高版本；其他原因，可选择以下兜底方案进行 DDL owner 迁移。
  - DDL owner 迁移方案：
    * 如果与该 TiDB 集群可以网络互通，执行重新进行 owner 选举命令：`curl -X POST http://{TiDBIP}:10080/ddl/owner/resign`
    * 如果与该 TiDB 集群不可以网络互通，需旁路下线，通过 `tidb-ctl` 工具，从 PD 集群的 etcd 中直接删除 DDL owner，之后也会重新选举：`tidb-ctl etcd delowner [LeaseID] [flags] + ownerKey`
- 3.1.3 TiDB 日志中报 `information schema is changed` 的错误：
  - 原因 1：正在执行的 DML 所涉及的表和正在执行 DDL 的表相同，可以通过命令 `admin show ddl job` 查看正在运行的 DDL 操作。
  - 原因 2：当前执行的 DML 时间太久，且这段时间内执行了很多 DDL（新版本 `lock table` 也会导致 schema 版本变化），导致中间 `schema version` 变更超过 1024 个版本数。
  - 原因 3：当前执行 DML 请求的 TiDB 实例长时间不能加载到新的 `schema information`（与 PD 或者 TiKV 网络问题等都会导致此问题），而这段时间内执行了很多 DDL 语句（也包括 `lock table` 语句），导致中间 `schema version` 变更超过 1024 个版本数。
  - 解决方法：前 2 种原因都不会导致业务问题，相应的 DML 会在失败后重试；第 3 种原因需要检查 TiDB 实例和 PD 及 TiKV 的网络情况。
  - 背景知识：`schema version` 的增长数量与每个 DDL 变更操作的 `schema state` 个数一致，例如 `create table` 操作会有 1 个版本变更，`add column` 操作会有 4 个版本变更（详情可以参考 online schema change），所以太多的 column 变更操作会导致 `schema version` 增长得很快。
- 3.1.4 TiDB 日志中报 `information schema is out of date` 的错误：
  - 原因 1：执行 DML 的 TiDB 被 `graceful kill` 后准备退出，且此 DML 对应的事务执行时间超过一个 DDL lease，在事务提交的时候会报此错。
  - 原因 2：TiDB 在执行 DML 时，有一段时间连不上 PD 和 TiKV，导致 TiDB 在一个 DDL Lease（默认 45s）内没有加载新的 schema，或者 TiDB 断开与 PD 之间带 `keep alive` 设置的连接。

- 原因 3：TiKV 压力大或网络超时，通过监控 Grafana -> TiDB 和 TiKV 节点的负载情况来确认是否是该原因。
- 解决方法：第 1 种原因，在 TiDB 启动时手动重试该 DML 即可；第 2 种原因，需要检查 TiDB 实例和 PD 及 TiKV 的网络波动情况；第 3 种原因，需要检查 TiKV 为什么繁忙，参考4. TiKV 问题。

13.1.3.3.2 3.2 OOM 问题

- 3.2.1 现象
  - 客户端：客户端收到 TiDB server 报错 `ERROR 2013 (HY000): Lost connection to MySQL server during query`
  - 日志：
    - `dmesg -T | grep tidb-server` 结果中有事故发生附近时间点的 OOM-killer 的日志。
    - tidb.log 中可以 grep 到事故发生后附近时间的 `"Welcome to TiDB"` 的日志（即 TiDB server 发生重启）。
    - tidb_stderr.log 中能 grep 到 `fatal error: "runtime: out of memory"` 或 `"cannot allocate memory"`。
    - v2.1.8 及其之前的版本，tidb_stderr.log 中能 grep 到 `fatal error: stack overflow`。
  - 监控：TiDB server 实例所在机器可用内存迅速回升
- 3.2.2 定位造成 OOM 的 SQL（目前所有版本都无法完成精准定位，需要在发现 SQL 后再做进一步分析，确认 OOM 是否的确由该 SQL 造成）：
  - `>= v3.0.0` 的版本，可以在 tidb.log 中 `grep "expensive_query"`，该 log 会记录运行超时、或使用内存超过阈值的 SQL。
  - `< v3.0.0` 的版本，通过 `grep "memory exceeds quota"` 定位运行时内存超限的 SQL。

> 注意：
> 单条 SQL 内存阈值的默认值为 1GB，可通过 `tidb_mem_quota_query` 系统变量进行设置，作用域为 SESSION，单位为 Byte。也可以通过配置项热加载的方式，对配置文件中的 `mem-quota-query` 项进行修改，单位为 Byte。

- 3.2.3 缓解 OOM 问题
  - 通过开启 SWAP 的方式，可以缓解由于大查询使用内存过多而造成的 OOM 问题。但该方法由于存在 I/O 开销，会在内存空间不足时对大查询性能造成一定影响。性能回退程度受剩余内存量、读写盘速度影响。
- 3.2.4 OOM 常见原因
  - SQL 中包含 join，通过 `explain` 查看发现该 join 选用 HashJoin 算法且 `inner` 端的表很大。
  - 单条 UPDATE/DELETE 涉及的查询数据量太大，见案例 case-882。
  - SQL 中包含 Union 连接的多条子查询，见案例 case-1828。

13.1.3.3.3 3.3 执行计划不对

- 3.3.1 现象
  - SQL 相比于之前的执行时间有较大程度变慢，执行计划突然发生改变。如果慢日志中输出了执行计划，可以直接对比执行计划。
  - SQL 执行时间相比于其他数据库（例如 MySQL）有较大差距。可以对比其他数据库执行计划，例如 `Join Order` 是否不同。
  - 慢日志中 SQL 执行时间 `Scan Keys` 数目较大。
- 3.3.2 排查执行计划问题
  - `explain analyze {SQL}` 在执行时间可以接受的情况下，对比 `explain analyze` 结果中 `count` 和 execution info 中 `rows` 的数目差距。如果在 `TableScan/IndexScan` 行上发现比较大的差距，很大可能是统计信息出问题；如果在其他行上发现较大差距，则也有可能是非统计信息问题。
  - `select count(*)` 在执行计划中包含 `join` 等情况下，explain analyze 可能耗时过长；此时可以通过对 `TableScan/IndexScan` 上的条件进行 `select count(*)`，并对比 `explain` 结果中的 `row count` 信息，确定是不是统计信息的问题。
- 3.3.3 缓解问题
  - v3.0 及以上版本可以使用 `SQL Bind` 功能固定执行计划。
  - 更新统计信息。在大致确定问题是由统计信息导致的情况下，先dump 统计信息保留现场。如果是由于统计信息过期导致，例如 `show stats_meta` 中 modify count/row count 大于某个值（例如 0.3）或者表中存在时间列的索引情况下，可以先尝试 analyze table 恢复；如果配置了 auto analyze，可以查看系统变量 `tidb_auto_analyze_ratio` 是否过大（例如大于 0.3），以及当前时间是否在 `tidb_auto_analyze_start_time` 和 `tidb_auto_analyze_end_time` 范围内。
  - 其他情况，请上报 bug。

13.1.3.3.4 3.4 SQL 执行报错

- 3.4.1 客户端报 `ERROR 1265(01000)Data Truncated` 错误。原因是 TiDB 内在计算 `Decimal` 类型处理精度的时候，和 MySQL 不兼容。该错误已于 v3.0.10 中修复 (#14438)，具体原因如下：

  在 MySQL 内，如果两个大精度 `Decimal` 做除法运算，超出最大小数精度时 (30)，会只保留 30 位且不报错。TiDB 在计算结果上，也是这样实现的，但是在内部表示 `Decimal` 的结构体内，有一个表示小数精度的字段，还是保留的真实精度。

  比如 `(0.1^30)/ 10`，TiDB 和 MySQL 的结果都为 0，是正确的，因为精度最多 30；但是 TiDB 内表示精度的那个字段，还是 31；

  多次 `Decimal` 除法计算后，虽然结果正确，但是这个精度可能越来越大，最终超过 TiDB 内的另一个阈值 72，此时就会报 `Data Truncated` 的错误；`Decimal` 的乘法计算就不会有这个问题，因为绕过越界，会直接把精度设置为最大精度限制。

  解决方法：可以通过手动加 `Cast(xx as decimal(a, b))` 来绕过这个问题，a 和 b 就是目标的精度。

13.1.3.4 4. TiKV 问题

13.1.3.4.1 4.1 TiKV panic 启动不了

- 4.1.1 `sync-log = false`,机器断电之后出现 `unexpected raft log index: last_index X < applied_index Y` 错误。符合预期，需通过 `tikv-ctl` 工具恢复 Region。
- 4.1.2 虚拟机部署 TiKV，`kill` 虚拟机或物理机断电，出现 `entries[X, Y] is unavailable from storage` 错误。符合预期，虚拟机的 fsync 不可靠，需通过 `tikv-ctl` 工具恢复 Region。
- 4.1.3 其他原因（非预期，需报 bug）。

13.1.3.4.2 4.2 TiKV OOM

- 4.2.1 `block-cache` 配置太大导致 OOM：
  - 在监控 Grafana -> TiKV-details 选中对应的 instance 后，查看 RocksDB 的 `block cache size` 监控来确认是否是该问题。
  - 同时，请检查 `[storage.block-cache] capacity = # "1GB"` 参数是否设置合理，默认情况下 TiKV 的 `block-cache` 设置为机器总内存的 45%；在 container 部署时，需要显式指定该参数，因为 TiKV 获取的是物理机的内存，可能会超出 container 的内存限制。
- 4.2.2 Coprocessor 收到大量大查询，返回的数据量太大，gRPC 的发送速度跟不上 Coprocessor 往外输出数据的速度，导致 OOM：
  - 可以通过检查监控：Grafana -> TiKV-details -> coprocessor overview 的 `response size` 是否超过 `network outbound` 流量来确认是否属于这种情况。
- 4.2.3 其他部分占用太多内存（非预期，需报 bug）。

13.1.3.4.3 4.3 客户端报 `server is busy` 错误

通过查看监控：Grafana -> TiKV -> errors 确认具体 busy 原因。`server is busy` 是 TiKV 自身的流控机制，TiKV 通过这种方式告知 `tidb/ti-client` 当前 TiKV 的压力过大，稍后再尝试。

- 4.3.1 TiKV RocksDB 出现 `write stall`。一个 TiKV 包含两个 RocksDB 实例，一个用于存储 Raft 日志，位于 `data/raft`。另一个用于存储真正的数据，位于 `data/db`。通过 `grep "Stalling" RocksDB` 日志查看 stall 的具体原因，RocksDB 日志是 LOG 开头的文件，LOG 为当前日志。
  - `level0 sst` 太多导致 stall，可以添加参数 `[rocksdb] max-sub-compactions = 2`（或者 3），加快 level0 sst 往下 compact 的速度。该参数的意思是将从 level0 到 level1 的 compaction 任务最多切成 `max-sub-compactions` 个子任务交给多线程并发执行，见案例 case-815。
  - `pending compaction bytes` 太多导致 stall，磁盘 I/O 能力在业务高峰跟不上写入，可以通过调大对应 Column Family (CF) 的 `soft-pending-compaction-bytes-limit` 和 `hard-pending-compaction-bytes-limit` 参数来缓解：
    * 如果 `pending compaction bytes` 达到该阈值，RocksDB 会放慢写入速度。默认值 64GB，`[rocksdb.defaultcf] soft-pending-compaction-bytes-limit = "128GB"`。

    * 如果 pending compaction bytes 达到该阈值，RocksDB 会 stop 写入，通常不太可能触发该情况，因为在达到 soft-pending-compaction-bytes-limit 的阈值之后会放慢写入速度。默认值 256GB，hard-pending-compaction-bytes-limit = "512GB"。
    * 如果磁盘 IO 能力持续跟不上写入，建议扩容。如果磁盘的吞吐达到了上限（例如 SATA SSD 的吞吐相对 NVME SSD 会低很多）导致 write stall，但是 CPU 资源又比较充足，可以尝试采用压缩率更高的压缩算法来缓解磁盘的压力，用 CPU 资源换磁盘资源。
    * 比如 default cf compaction 压力比较大，调整参数 [rocksdb.defaultcf] compression-per-level = ["no", "no", "lz4", "lz4", "lz4", "zstd", "zstd"] 改成 compression-per-level = ["no", "no", "zstd", "zstd", "zstd", "zstd", "zstd"]。
  - memtable 太多导致 stall。该问题一般发生在瞬间写入量比较大，并且 memtable flush 到磁盘的速度比较慢的情况下。如果磁盘写入速度不能改善，并且只有业务峰值会出现这种情况，可以通过调大对应 CF 的 max-write-buffer-number 来缓解：
    * 例如 [rocksdb.defaultcf] max-write-buffer-number = 8（默认值 5），同时请求注意在高峰期可能会占用更多的内存，因为可能存在于内存中的 memtable 会更多。

- 4.3.2 scheduler too busy
  - 写入冲突严重，latch wait duration 比较高，查看监控：Grafana -> TiKV-details -> scheduler prewrite 或者 scheduler commit 的 latch wait duration。scheduler 写入任务堆积，导致超过了 [storage] scheduler -pending-write-threshold = "100MB" 设置的阈值。可通过查看 MVCC_CONFLICT_COUNTER 对应的 metric 来确认是否属于该情况。
  - 写入慢导致写入堆积，该 TiKV 正在写入的数据超过了 [storage] scheduler-pending-write-threshold = "100MB" 设置的阈值。请参考4.5 TiKV 写入慢。

- 4.3.3 raftstore is busy，主要是消息的处理速度没有跟上接收消息的速度。短时间的 channel full 不会影响服务，长时间持续出现该错误可能会导致 Leader 切换走。
  - append log 遇到了 stall，参考4.3.1 客户端报 server is busy 错误。
  - append log duration 比较高，导致处理消息不及时，可以参考4.5 TiKV 写入慢分析为什么 append log duration 比较高。
  - 瞬间收到大量消息（查看 TiKV Raft messages 面板），Raftstore 没处理过来，通常情况下短时间的 channel full 不会影响服务。

- 4.3.4 TiKV Coprocessor 排队，任务堆积超过了 Coprocessor 线程数 * readpool.coprocessor.max-tasks-per-worker-[normal|low|high]。大量大查询导致 Coprocessor 出现了堆积情况，需要确认是否由于执行计划变化而导致了大量扫表操作，请参考3.3 执行计划不对。

13.1.3.4.4 4.4 某些 TiKV 大量掉 Leader

- 4.4.1 TiKV 重启，导致重新选举。
  - TiKV panic 之后又被 systemd 重新拉起正常运行，可以通过查看 TiKV 的日志来确认是否有 panic，这种情况属于非预期，需报 bug。
  - 被第三者 stop/kill，被 systemd 重新拉起。查看 dmesg 和 TiKV log 确认原因。
  - TiKV 发生 OOM 导致重启了，参考4.2 TiKV OOM 问题。
  - 动态调整 THP 导致 hung 住，见案例 case-500。

- 4.4.2 查看监控：Grafana -> TiKV-details -> errors 面板 `server is busy`，看到 TiKV RocksDB 出现 write stall 导致发生重新选举，请参考4.3.1。
- 4.4.3 网络隔离导致重新选举。

13.1.3.4.5 4.5 TiKV 写入慢

- 4.5.1 通过查看 TiKV gRPC 的 `prewrite/commit/raw-put`（仅限 raw kv 集群）duration 确认确实是 TiKV 写入慢了。通常情况下可以按照 performance-map 来定位到底哪个阶段慢了，下面列出几种常见的情况。
- 4.5.2 scheduler CPU 繁忙（仅限 transaction kv）。prewrite/commit 的 `scheduler command duration` 比 `scheduler latch wait duration` + `storage async write duration` 更长，并且 scheduler worker CPU 比较高，例如超过 `scheduler-worker-pool-size` * 100% 的 80%，并且或者整个机器的 CPU 资源比较紧张。如果写入量很大，确认下是否 `[storage] scheduler-worker-pool-size` 配置得太小。其他情况，需报 bug。
- 4.5.3 Append log 慢。TiKV Grafana 的 Raft IO/`append log duration` 比较高，通常情况下是由于写盘慢了，可以检查 RocksDB - Raft 的 `WAL Sync Duration max` 值来确认，否则可能需要报 bug。
- 4.5.4 Raftstore 线程繁忙。TiKV Grafana 的 Raft Propose/`propose wait duration` 明显高于 `append log duration`。请查看以下情况：
  - `[raftstore] store-pool-size` 配置是否过小（该值建议在 [1,5] 之间，不建议太大）。
  - 机器的 CPU 是不是不够。
- 4.5.5 apply 慢了。TiKV Grafana 的 Raft IO/`apply log duration` 比较高，通常会伴随着 Raft Propose/`apply wait duration` 比较高。可能是以下原因引起的：
  - `[raftstore] apply-pool-size` 配置过小（建议在 [1, 5] 之间，不建议太大），Thread CPU/apply cpu 比较高；
  - 机器的 CPU 资源不够了；
  - Region 写入热点问题，单个 apply 线程 CPU 使用率比较高（通过修改 Grafana 表达式，加上 by (instance, name) 来看各个线程的 CPU 使用情况），暂时对于单个 Region 的热点写入没有很好的方式，最近在优化该场景；
  - 写 RocksDB 比较慢，RocksDB kv/`max write duration` 比较高（单个 Raft log 可能包含很多个 kv，写 RocksDB 的时候会把 128 个 kv 放在一个 write batch 写入到 RocksDB，所以一次 apply log 可能涉及到多次 RocksDB 的 write）；
  - 其他情况，需报 bug。
- 4.5.6 Raft commit log 慢了。
  - TiKV Grafana 的 Raft IO/`commit log duration` 比较高（4.x 版本的 Grafana 才有该 metric）。每个 Region 对应一个独立的 Raft group，Raft 本身是有流控机制的，类似 TCP 的滑动窗口机制，通过参数 `[raftstore] raft-max-inflight-msgs = 256` 来控制滑动窗口的大小，如果有热点写入并且 `commit log duration` 比较高可以适度调大该参数，比如 1024。
- 4.5.7 其他情况，请参考 Performance Map 上的写入路径来分析。

13.1.3.5 5. PD 问题

13.1.3.5.1 5.1 PD 调度问题

- 5.1.1 merge 问题：
  - 跨表空 Region 无法 merge，需要修改 TiKV 的 `[coprocessor] split-region-on-table = false` 参数来解决，4.x 版本该参数默认为 false，见案例 case-896。
  - Region merge 慢，可检查监控 Grafana -> PD -> operator 面板是否有 merge 的 operator 产生，可以适当调大 `merge-schedule-limit` 参数来加速 merge。
- 5.1.2 补副本/上下线问题：
  - TiKV 磁盘使用 80% 容量，PD 不会进行补副本操作，miss peer 数量上升，需要扩容 TiKV，见案例 case-801。
  - 下线 TiKV，有 Region 长时间迁移不走。v3.0.4 版本已经修复改问题，见 #5526 和案例 case-870。
- 5.1.3 Balance 问题：
  - Leader/Region count 分布不均，见案例 case-394, case-759。主要原因是 balance 是依赖 Region/leader 的 size 去调度的，所以可能会造成 count 数量的不均衡，v4.0 新增了一个参数 `[leader-schedule-policy]`，可以调整 Leader 的调度策略，根据 “count” 或者是 “size” 进行调度。

13.1.3.5.2 5.2 PD 选举问题

- 5.2.1 PD 发生 Leader 切换：
  - 磁盘问题，PD 所在的节点 I/O 被打满，排查是否有其他 I/O 高的组件与 PD 混部以及盘的健康情况，可通过监控 Grafana -> disk performance -> latency 和 load 等指标进行验证，必要时可以使用 fio 工具对盘进行检测，见案例 case-292。
  - 网络问题，PD 日志中有 `lost the TCP streaming connection`，排查 PD 之间网络是否有问题，可通过监控 Grafana -> PD -> etcd 的 `round trip` 来验证，见案例 case-177。
  - 系统 load 高，日志中能看到 `server is likely overloaded`，见案例 case-214。
- 5.2.2 PD 选不出 Leader 或者选举慢：
  - 选不出 Leader，PD 日志中有 `lease is not expired`，见 #10355。v3.0.x 版本和 v2.1.19 版本已修复该问题，见案例 case-875。
  - 选举慢，Region 加载时间长。从 PD 日志中 `grep "regions cost"`（例如日志中可能是 “load 460927 regions cost 11.77099s”），如果出现秒级，则说明较慢，v3.0 版本可开启 Region storage（设置 `use-region-storage` 为 `true`），该特性能极大缩短加载 Region 的时间，见案例 case-429。
- 5.2.3 TiDB 执行 SQL 时报 PD timeout：
  - PD 没 Leader 或者有切换，参考5.2.1 PD 选举问题和5.2.2 PD 选举问题。
  - 网络问题，排查网络相关情况。通过监控 Grafana -> blackbox_exporter -> ping latency 确定 TiDB 到 PD Leader 的网络是否正常。
  - PD panic，需报 bug。
  - PD OOM，参考5.3 PD OOM 问题。

- 其他原因，通过 `curl http://127.0.0.1:2379/debug/pprof/goroutine?debug=2` 抓 goroutine，报 bug。

- 5.2.4 其他问题
  - PD 报 FATAL 错误，日志中有 `range failed to find revision pair`，v3.0.8 已经修复该问题，见 #2040。详情参考案例 case-947。
  - 其他原因，需报 bug。

13.1.3.5.3 5.3 PD OOM

- 5.3.1 使用 `/api/v1/regions` 接口时 Region 数量过多，可能会导致 PD OOM，在 v3.0.8 版本中修复，见 #1986。
- 5.3.2 滚动升级的时候 PD OOM，gRPC 消息大小没限制，监控可看到 TCP InSegs 较大，在 v3.0.6 版本中修复，见 #1952。

13.1.3.5.4 5.4 Grafana 显示问题

- 5.4.1 监控 Grafana -> PD -> cluster -> role 显示 follower，Grafana 表达式问题，在 v3.0.8 版本修复，见 #1065。详情请参考案例 case-1022。

13.1.3.6 6. 生态 Tools 问题

13.1.3.6.1 6.1 TiDB Binlog 问题

- 6.1.1 TiDB Binlog 是将 TiDB 的修改同步给下游 TiDB 或者 MySQL 的工具，见 TiDB Binlog on GitHub。
- 6.1.2 Pump/Drainer Status 中 Update Time 正常更新，日志中也没有异常，但下游没有数据写入。
  - TiDB 配置中没有开启 binlog，需要修改 TiDB 配置 `[binlog]`。
- 6.1.3 Drainer 中的 sarama 报 EOF 错误。
  - Drainer 使用的 Kafka 客户端版本和 Kafka 版本不匹配，需要修改配置 `[syncer.to] kafka-version` 来解决。
- 6.1.4 Drainer 写 Kafka 失败然后 panic，Kafka 报 `Message was too large` 错误。
  - binlog 数据太大，造成写 Kafka 的单条消息太大，需要修改 Kafka 的下列配置来解决：

    ```
    message.max.bytes=1073741824
    replica.fetch.max.bytes=1073741824
    fetch.message.max.bytes=1073741824
    ```

    见案例 case-789。
- 6.1.5 上下游数据不一致

  - 部分 TiDB 节点没有开启 binlog。v3.0.6 及之后的版本可以通过访问 http://127.0.0.1:10080/info/all 接口可以检查所有节点的 binlog 状态。之前的版本可以通过查看配置文件来确认是否开启了 binlog。
  - 部分 TiDB 节点进入 ignore binlog 状态。v3.0.6 及之后的版本可以通过访问 http://127.0.0.1:10080/info/all 接口可以检查所有节点的 binlog 状态。之前的版本需要看 TiDB 的日志中是否有 ignore binlog 关键字来确认是该问题。
  - 上下游 Timestamp 列的值不一致：
    * 时区问题，需要确保 Drainer 和上下游数据库时区一致，Drainer 通过 `/etc/localtime` 获取时区，不支持 TZ 环境变量，见案例 case-826。
    * TiDB 中 Timestamp 的默认值为 `null`，MySQL 5.7 中 Timestamp 默认值为当前时间（MySQL 8 无此问题），因此当上游写入 `null` 的 Timestamp 且下游是 MySQL 5.7 时，Timestamp 列数据不一致。在开启 binlog 前，在上游执行 `set @@global.explicit_defaults_for_timestamp=on;` 可解决此问题。
  - 其他情况需报 bug。

- 6.1.6 同步慢
  - 下游是 TiDB/MySQL，上游频繁进行 DDL 操作，见案例 case-1023。
  - 下游是 TiDB/MySQL，需要同步的表中存在没有主键且没有唯一索引的表，这种情况会导致 binlog 性能下降，建议加主键或唯一索引。
  - 下游输出到文件，检查目标磁盘/网络盘是否慢。
  - 其他情况需报 bug。

- 6.1.7 Pump 无法写 binlog，报 `no space left on device` 错误。
  - 本地磁盘空间不足，Pump 无法正常写 binlog 数据。需要清理磁盘空间，然后重启 Pump。

- 6.1.8 Pump 启动时报错 `fail to notify all living drainer`。
  - Pump 启动时需要通知所有 Online 状态的 Drainer，如果通知失败则会打印该错误日志。
  - 可以使用 binlogctl 工具查看所有 Drainer 的状态是否有异常，保证 Online 状态的 Drainer 都在正常工作。如果某个 Drainer 的状态和实际运行情况不一致，则使用 binlogctl 修改状态，然后再重启 Pump。见案例fail-to-notify-all-living-drainer。

- 6.1.9 Drainer 报错 `gen update sqls failed: table xxx: row data is corruption []`。
  - 触发条件：上游做 Drop Column DDL 的时候同时在做这张表的 DML。已经在 v3.0.6 修复，见 case-820。

- 6.1.10 Drainer 同步卡住，进程活跃但 checkpoint 不更新。
  - 已知 bug 在 v3.0.4 修复，见案例 case-741。

- 6.1.11 任何组件 `panic`。
  - 需报 bug。

13.1.3.6.2 6.2 DM 问题

- 6.2.1 TiDB Data Migration (DM) 是能将 MySQL/MariaDB 的数据迁移到 TiDB 的迁移工具，详情见 DM on GitHub。
- 6.2.2 执行 `query-status` 或查看日志时出现 `Access denied for user 'root'@'172.31.43.27' (using password: YES)`。
  - 在所有 DM 配置文件中，数据库相关的密码都必须使用经 dmctl 加密后的密文（若数据库密码为空，则无需加密）。在 v1.0.6 及以后的版本可使用明文密码。
  - 在 DM 运行过程中，上下游数据库的用户必须具备相应的读写权限。在启动同步任务过程中，DM 会自动进行相应权限的检查。
  - 同一套 DM 集群，混合部署不同版本的 DM-worker/DM-master/dmctl，见案例 AskTUG-1049。
- 6.2.3 DM 同步任务中断并包含 `driver: bad connection` 错误。
  - 发生 `driver: bad connection` 错误时，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启等）且当前请求的数据暂时未能发送到 TiDB。
    * 1.0.0 GA 之前的版本，DM 发生该类型错误时，需要先使用 `stop-task` 命令停止任务后再使用 `start-task` 命令重启任务。
    * 1.0.0 GA 版本，增加对此类错误的自动重试机制，见 #265。
- 6.2.4 同步任务中断并包含 `invalid connection` 错误。
  - 发生 `invalid connection` 错误时，通常表示 DM 到下游 TiDB 的数据库连接出现了异常（如网络故障、TiDB 重启、TiKV busy 等）且当前请求已有部分数据发送到了 TiDB。由于 DM 中存在同步任务并发向下游复制数据的特性，因此在任务中断时可能同时包含多个错误（可通过 `query-status` 或 `query-error` 查询当前错误）：
    * 如果错误中仅包含 `invalid connection` 类型的错误，且当前处于增量复制阶段，则 DM 会自动进行重试。
    * 如果 DM 由于版本问题（v1.0.0-rc.1 后引入自动重试）等未自动进行重试或自动重试未能成功，则可尝试先使用 `stop-task` 停止任务，然后再使用 `start-task` 重启任务。
- 6.2.5 Relay 处理单元报错 `event from * in * diff from passed-in event *` 或同步任务中断并包含 `get binlog error ERROR 1236 (HY000)`、`binlog checksum mismatch, data may be corrupted` 等 binlog 获取或解析失败错误。
  - 在 DM 进行 relay log 拉取与增量同步过程中，如果遇到了上游超过 4 GB 的 binlog 文件，就可能出现这两个错误。原因是 DM 在写 relay log 时需要依据 binlog position 及文件大小对 event 进行验证，且需要保存同步的 binlog position 信息作为 checkpoint。但是 MySQL binlog position 官方定义使用 uint32 存储，所以超过 4 GB 部分的 binlog position 的 offset 值会溢出，进而出现上面的错误。
    * 对于 relay 处理单元，可通过官网步骤进行手动处理。
    * 对于 binlog replication 处理单元，可通过官网步骤进行手动处理。
- 6.2.6 DM 同步中断，日志报错 `ERROR 1236 (HY000)The slave is connecting using CHANGE MASTER TO MASTER_AUTO_POSITION = 1, but the master has purged binary logs containing GTIDs that the slave requires.`。

  - 检查 master 的 binlog 是否被 purge。
  - 检查 relay.meta 中记录的位点信息。
    * relay.meta 中记录空的 GTID 信息，DM-worker 进程在退出时、以及定时 (30s) 会把内存中的 GTID 信息保存到 relay.meta 中，在没有获取到上游 GTID 信息的情况下，把空的 GTID 信息保存到了 relay.meta 中。见案例 case-772。
    * relay.meta 中记录的 binlog event 不完整触发 recover 流程后记录错误的 GTID 信息，该问题可能会在 1.0.2 之前的版本遇到，已在 1.0.2 版本修复。

- 6.2.7 DM 同步报错 `Error 1366: incorrect utf8 value eda0bdedb29d(\ufffd\ufffd\ufffd\ufffd\ufffd\ufffd)`。
  - 该值 MySQL 8.0 和 TiDB 都不能写入成功，但是 MySQL 5.7 可以写入成功。可以开启 TiDB 动态参数 `tidb_skip_utf8_check` 参数，跳过数据格式检查。

13.1.3.6.3 6.3 TiDB Lightning 问题

- 6.3.1 TiDB Lightning 是快速的全量数据导入工具，见 TiDB Lightning on GitHub。
- 6.3.2 导入速度太慢。
  - `region-concurrency` 设定太高，线程间争用资源反而减低了效率。排查方法如下：
    * 从日志的开头搜寻 `region-concurrency` 能知道 TiDB Lightning 读到的参数是多少；
    * 如果 TiDB Lightning 与其他服务（如 TiKV Importer）共用一台服务器，必需手动将 `region-concurrency` 设为该服务器 CPU 数量的 75%；
    * 如果 CPU 设有限额（例如从 Kubernetes 指定的上限），TiDB Lightning 可能无法自动判断出来，此时亦需要手动调整 `region-concurrency`。
  - 表结构太复杂。每条索引都会额外增加 KV 对，如果有 N 条索引，实际导入的大小就差不多是 Mydumper 文件的 N+1 倍。如果索引不太重要，可以考虑先从 schema 去掉，待导入完成后再使用 `CREATE INDEX` 加回去。
  - TiDB Lightning 版本太旧。尝试使用最新的版本，可能会有改善。
- 6.3.3 `checksum failed: checksum mismatched remote vs local`
  - 原因 1：这张表可能本身已有数据，影响最终结果。
  - 原因 2：如果目标数据库的校验和全是 0，表示没有发生任何导入，有可能是集群太忙无法接收任何数据。
  - 原因 3：如果数据源是由机器生成而不是从 Mydumper 备份的，需确保数据符合表的限制。例如：
    * 自增 (AUTO_INCREMENT) 的列需要为正数，不能为 0。
    * 单一键和主键 (UNIQUE and PRIMARY KEYs) 不能有重复的值。
  - 解决办法：参考官网步骤处理。
- 6.3.4 `Checkpoint for … has invalid status:(错误码)`
  - 原因：断点续传已启用。TiDB Lightning 或 TiKV Importer 之前发生了异常退出。为了防止数据意外损坏，TiDB Lightning 在错误解决以前不会启动。错误码是小于 25 的整数，可能的取值是 0、3、6、9、12、14、15、17、18、20、21。整数越大，表示异常退出所发生的步骤在导入流程中越晚。

  - 解决办法：参考官网步骤处理。

- 6.3.5 `ResourceTemporarilyUnavailable("Too many open engines …: 8")`

  - 原因：并行打开的引擎文件 (engine files) 超出 tikv-importer 里的限制。这可能由配置错误引起。即使配置没问题，如果 tidb-lightning 曾经异常退出，也有可能令引擎文件残留在打开的状态，占据可用的数量。
  - 解决办法：参考官网步骤处理。

- 6.3.6 `cannot guess encoding for input file, please convert to UTF-8 manually`

  - 原因：TiDB Lightning 只支持 UTF-8 和 GB-18030 编码的表架构。此错误代表数据源不是这里任一个编码。也有可能是文件中混合了不同的编码，例如在不同的环境运行过 `ALTER TABLE`，使表架构同时出现 UTF-8 和 GB-18030 的字符。
  - 解决办法：参考官网步骤处理。

- 6.3.7 `[sql2kv] sql encode error = [types:1292]invalid time format: '{1970 1 1 0 45 0 0}'`

  - 原因：一个 timestamp 类型的时间戳记录了不存在的时间值。时间值不存在是由于夏令时切换或超出支持的范围（1970 年 1 月 1 日至 2038 年 1 月 19 日）。
  - 解决办法：参考官网步骤处理。

13.1.3.7 7. 常见日志分析

13.1.3.7.1 7.1 TiDB

- 7.1.1 `GC life time is shorter than transaction duration`。事务执行时间太长，超过了 GC lifetime（默认为 10 分钟），可以通过修改系统变量`tidb_gc_life_time` 来延长 life time，通常情况下不建议修改，因为延长时限可能导致大量老版本数据的堆积（如果有大量 `UPDATE` 和 `DELETE` 语句）。

- 7.1.2 `txn takes too much time`。事务太长时间（超过 590s）没有提交，准备提交的时候报该错误。可以通过调大 `[tikv-client] max-txn-time-use = 590` 参数，以及调大 `GC life time` 来绕过该问题（如果确实有这个需求）。通常情况下，建议看看业务是否真的需要执行这么长时间的事务。

- 7.1.3 coprocessor.go 报 `request outdated`。发往 TiKV 的 Coprocessor 请求在 TiKV 端排队时间超过了 60s，直接返回该错误。需要排查 TiKV Coprocessor 为什么排队这么严重。

- 7.1.4 region_cache.go 大量报 `switch region peer to next due to send request fail` 且 error 信息是 `context deadline exceeded`。请求 TiKV 超时触发 region cache 切换请求到其他节点，可以对日志中的 addr 字段继续 `grep "<addr> cancelled"`，根据 grep 结果：

  - `send request is cancelled`。请求发送阶段超时，可以排查监控 Grafana -> TiDB -> Batch Client/Pending Request Count by TiKV 是否大于 128，确定是否因发送远超 KV 处理能力导致发送堆积。如果 Pending Request 不多，需要排查日志确认是否因为对应 KV 有运维变更，导致短暂报出；否则非预期，需报 bug。
  - `wait response is cancelled`。请求发送到 TiKV 后超时未收到 TiKV 响应。需要排查对应地址 TiKV 的响应时间和对应 Region 在当时的 PD 和 KV 日志，确定为什么 KV 未及时响应。

- 7.1.5 distsql.go 报 `inconsistent index`。数据索引疑似发生不一致，首先对报错的信息中 index 所在表执行 `admin check table <TableName>` 命令，如果检查失败，则先通过以下命令禁用 GC，然后报 bug。

```
SET GLOBAL tidb_gc_enable = 0;
```

13.1.3.7.2 7.2 TiKV

- 7.2.1 `key is locked` 读写冲突，读请求碰到还未提交的数据，需要等待其提交之后才能读。少量这个错误对业务无影响，大量出现这个错误说明业务读写冲突比较严重。
- 7.2.2 `write conflict` 乐观事务中的写写冲突，同时多个事务对相同的 key 进行修改，只有一个事务会成功，其他事务会自动重取 timestamp 然后进行重试，不影响业务。如果业务冲突很严重可能会导致重试多次之后事务失败，这种情况下建议使用悲观锁。
- 7.2.3 TxnLockNotFound 事务提交太慢，过了 TTL（小事务默认 3s）时间之后被其他事务回滚了，该事务会自动重试，通常情况下对业务无感知。
- 7.2.4 PessimisticLockNotFound 类似 TxnLockNotFound，悲观事务提交太慢被其他事务回滚了。
- 7.2.5 `stale_epoch` 请求的 epoch 太旧了，TiDB 会更新路由之后再重新发送请求，业务无感知。epoch 在 Region 发生 split/merge 以及迁移副本的时候会变化。
- 7.2.6 `peer is not leader` 请求发到了非 Leader 的副本上，TiDB 会根据该错误更新本地路由（如果错误 response 里携带了最新 Leader 是哪个副本这一信息），并且重新发送请求到最新 Leader，一般情况下业务无感知。在 v3.0 后 TiDB 在原 Leader 请求失败时会尝试其他 peer，也会导致 TiKV 频繁出现 `not leader` 日志，可以通过查看 TiDB 对应 Region 的 `switch region peer to next due to send request fail` 日志，排查发送失败根本原因，参考7.1.4 TiDB。另外也可能是由于其他原因导致一些 Region 一直没有 Leader，请参考4.4 某些 TiKV 大量掉 Leader。

13.1.4 TiDB 热点问题处理

本文介绍如何定位和解决读写热点问题。

TiDB 作为分布式数据库，内建负载均衡机制，尽可能将业务负载均匀地分布到不同计算或存储节点上，更好地利用上整体系统资源。然而，机制不是万能的，在一些场景下仍会有部分业务负载不能被很好地分散，影响性能，形成单点的过高负载，也称为热点。

TiDB 提供了完整的方案用于排查、解决或规避这类热点。通过均衡负载热点，可以提升整体性能，包括提高 QPS 和降低延迟等。

13.1.4.1 常见热点场景

13.1.4.1.1 TiDB 编码规则回顾

TiDB 对每个表分配一个 TableID，每一个索引都会分配一个 IndexID，每一行分配一个 RowID（默认情况下，如果表使用整数型的 Primary Key，那么会用 Primary Key 的值当做 RowID）。其中 TableID 在整个集群内唯一，IndexID/RowID 在表内唯一，这些 ID 都是 int64 类型。

每行数据按照如下规则进行编码成 Key-Value pair：

```
Key: tablePrefix{tableID}_recordPrefixSep{rowID}
Value: [col1, col2, col3, col4]
```

其中 Key 的 tablePrefix 和 recordPrefixSep 都是特定的字符串常量，用于在 KV 空间内区分其他数据。

对于 Index 数据，会按照如下规则编码成 Key-Value pair：

```
Key: tablePrefix{tableID}_indexPrefixSep{indexID}_indexedColumnsValue
Value: rowID
```

Index 数据还需要考虑 Unique Index 和非 Unique Index 两种情况，对于 Unique Index，可以按照上述编码规则。但是对于非 Unique Index，通过这种编码并不能构造出唯一的 Key，因为同一个 Index 的 tablePrefix{tableID} ↪ _indexPrefixSep{indexID} 都一样，可能有多行数据的 ColumnsValue 是一样的，所以对于非 Unique Index 的编码做了一点调整：

```
Key: tablePrefix{tableID}_indexPrefixSep{indexID}_indexedColumnsValue_rowID
Value: null
```

13.1.4.1.2 表热点

从 TiDB 编码规则可知，同一个表的数据会在以表 ID 开头为前缀的一个 range 中，数据的顺序按照 RowID 的值顺序排列。在表 insert 的过程中如果 RowID 的值是递增的，则插入的行只能在末端追加。当 Region 达到一定的大小之后会进行分裂，分裂之后还是只能在 range 范围的末端追加，永远只能在一个 Region 上进行 insert 操作，形成热点。

常见的 increment 类型自增主键就是顺序递增的，默认情况下，在主键为整数型时，会用主键值当做 RowID，此时 RowID 为顺序递增，在大量 insert 时形成表的写入热点。

同时，TiDB 中 RowID 默认也按照自增的方式顺序递增，主键不为整数类型时，同样会遇到写入热点的问题。

13.1.4.1.3 索引热点

索引热点与表热点类似，常见的热点场景出现在时间顺序单调递增的字段，或者插入大量重复值的场景。

13.1.4.2 确定存在热点问题

性能问题不一定是热点造成的，也可能存在多个因素共同影响，在排查前需要先确认是否与热点相关。

- 判断写热点依据：打开监控面板 TiKV-Trouble-Shooting 中 Hot Write 面板，观察 Raftstore CPU 监控是否存在个别 TiKV 节点的指标明显高于其他节点的现象。
- 判断读热点依据：打开监控面板 TIKV-Details 中 Thread_CPU，查看 coprocessor cpu 有没有明显的某个 TiKV 特别高。

13.1.4.3 使用 TiDB Dashboard 定位热点表

TiDB Dashboard 中的流量可视化功能可帮助用户缩小热点排查范围到表级别。以下是流量可视化功能展示的一个热力图样例，该图横坐标是时间，纵坐标是各个表和索引，颜色越亮代表其流量越大。可在工具栏中切换显示读或写流量。

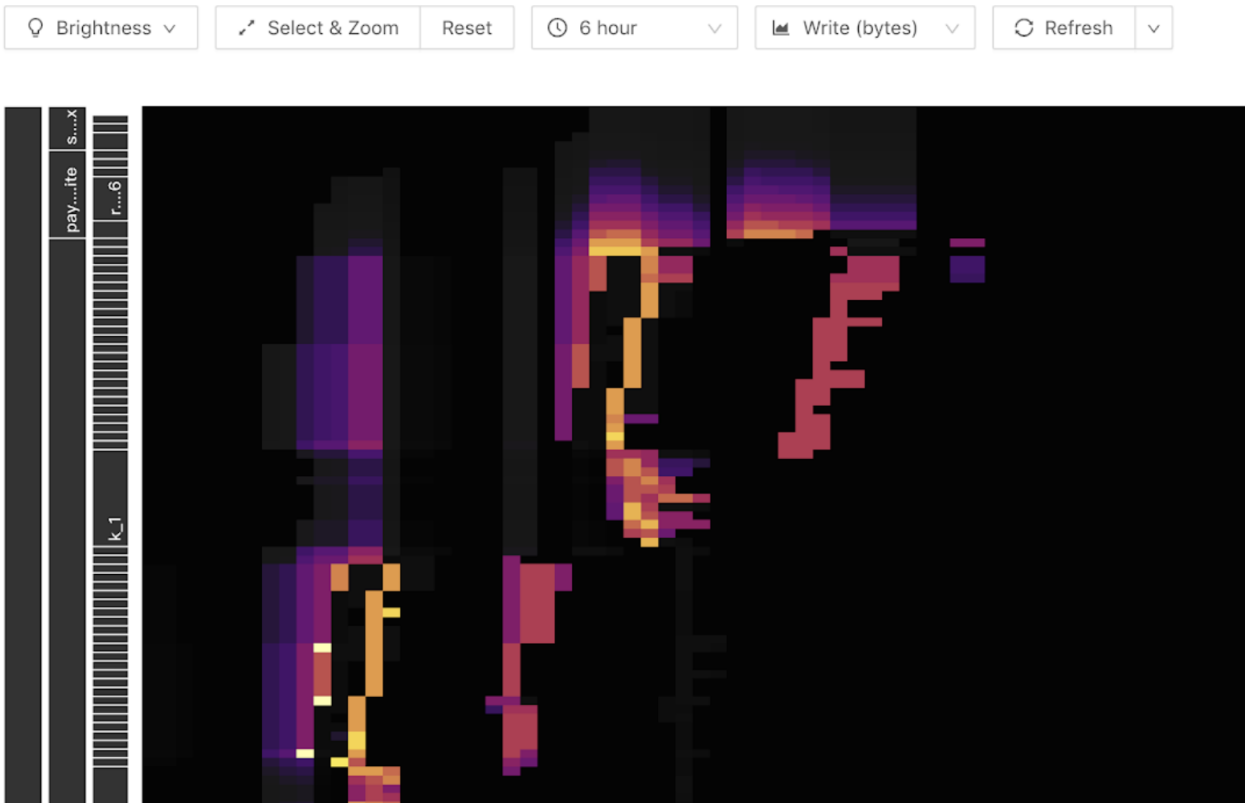


图 328: Dashboard 示例 1

当图中写入流量图出现以下明亮斜线（斜向上或斜向下）时，由于写入只出现在末端，随着表 Region 数量变多，呈现出阶梯状。此时说明该表构成了写入热点：

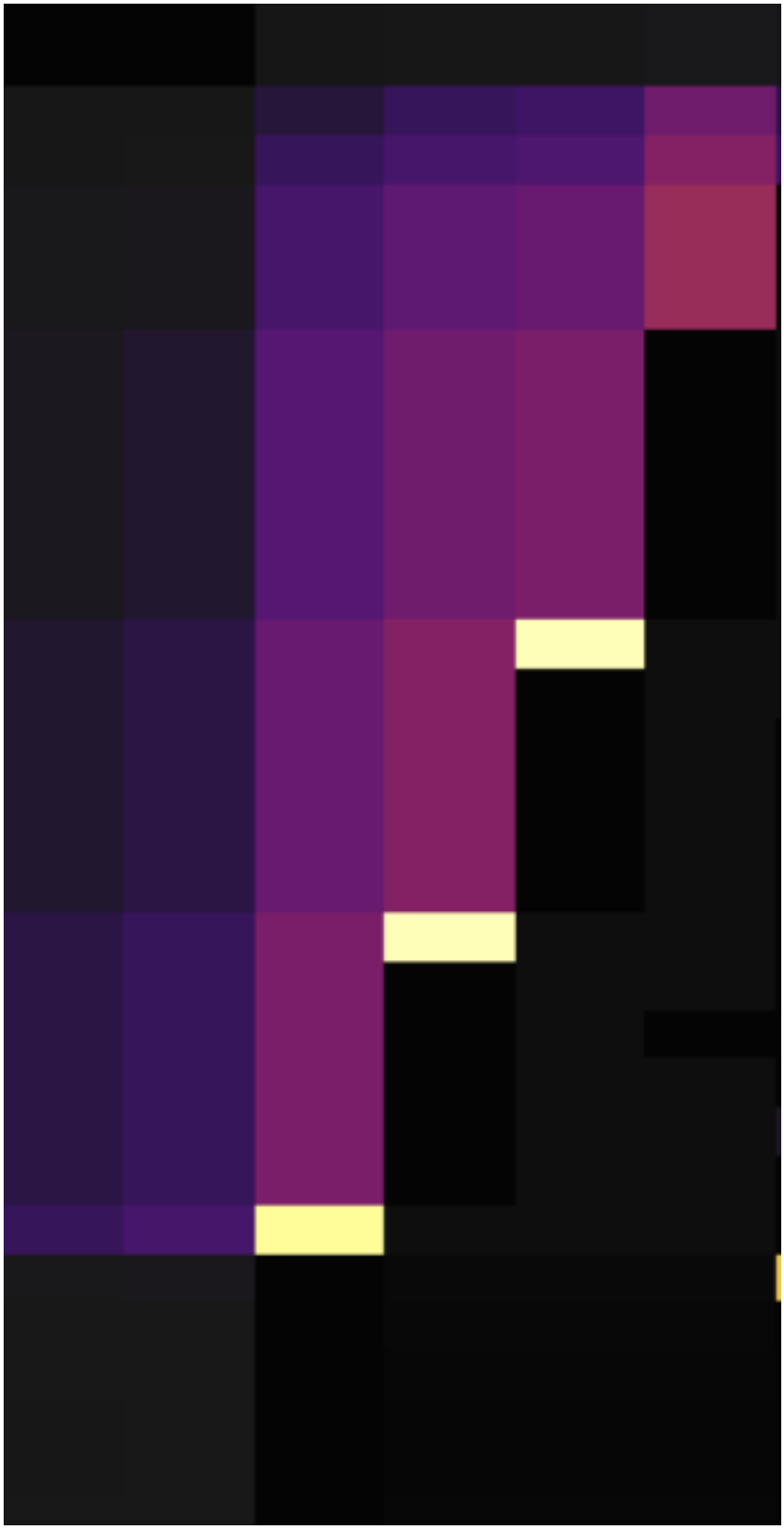


图 329: Dashboard 示例 2

对于读热点，在热力图中一般表现为一条明亮的横线，通常是有大量访问的小表，如下图所示：

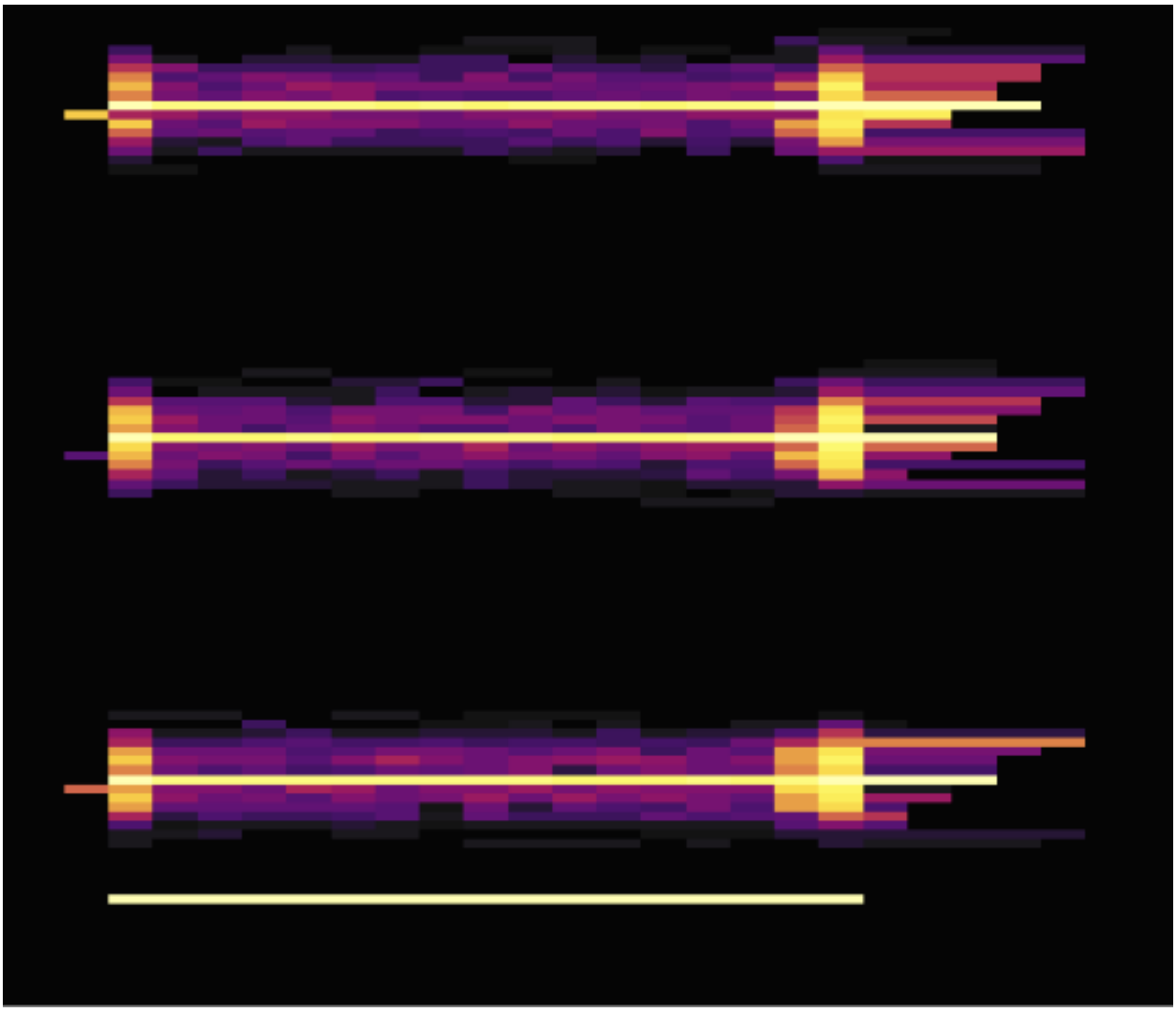


图 330: Dashboard 示例 3

将鼠标移到亮色块上，即可看到是什么表或索引具有大流量，如下图所示：

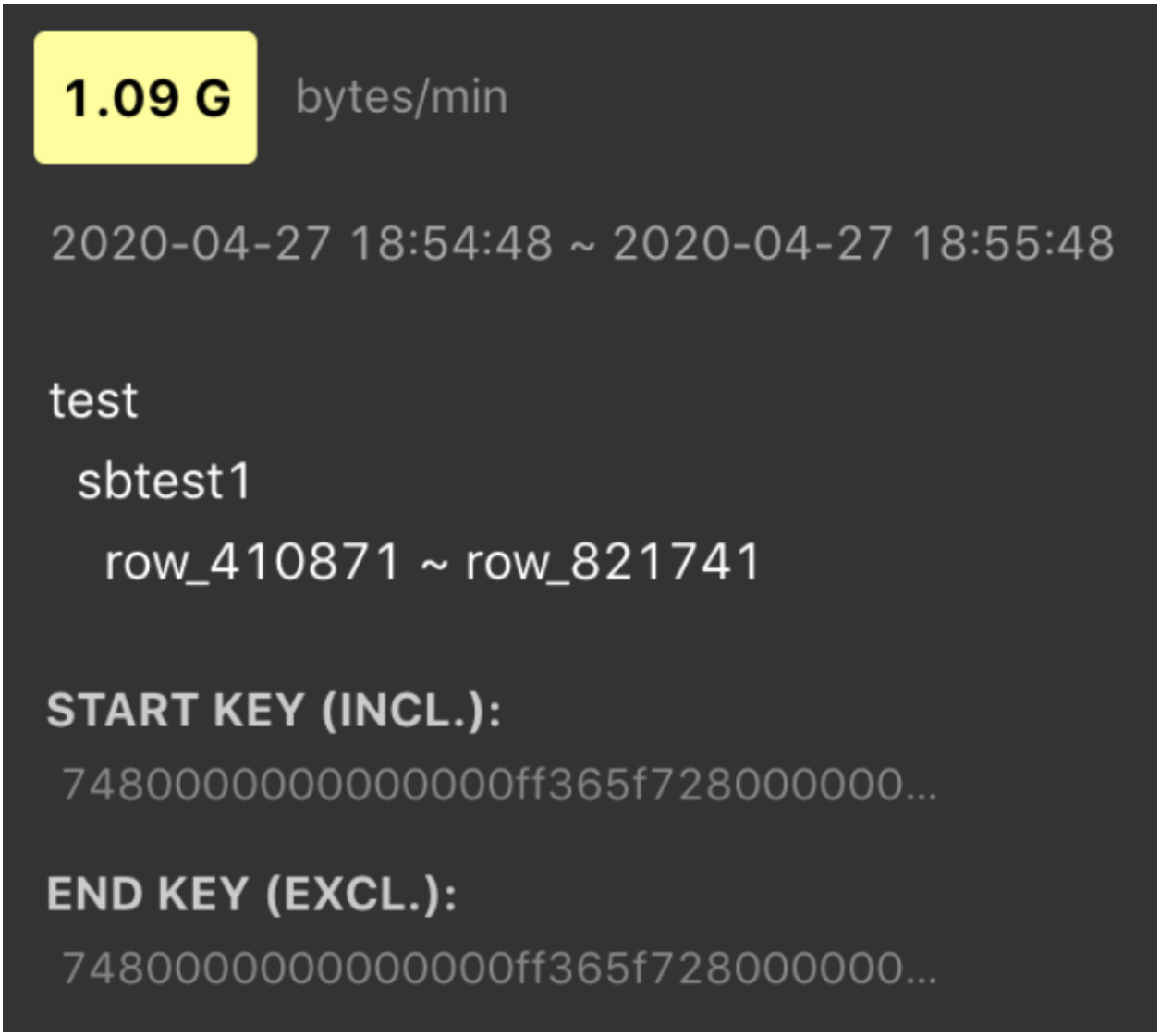


图 331: Dashboard 示例 4

13.1.4.4 使用 SHARD_ROW_ID_BITS 处理热点表

对于主键非整数或没有主键的表或者是联合主键，TiDB 会使用一个隐式的自增 RowID，大量 INSERT 时会把数据集中写入单个 Region，造成写入热点。

通过设置 SHARD_ROW_ID_BITS，可以把 RowID 打散写入多个不同的 Region，缓解写入热点问题。但是设置的过大会造成 RPC 请求数放大，增加 CPU 和网络开销。

```
SHARD_ROW_ID_BITS = 4 表示 16 个分片\
SHARD_ROW_ID_BITS = 6 表示 64 个分片\
SHARD_ROW_ID_BITS = 0 表示默认值 1 个分片
```

语句示例：

```
CREATE TABLE: CREATE TABLE t (c int) SHARD_ROW_ID_BITS = 4;
ALTER TABLE: ALTER TABLE t SHARD_ROW_ID_BITS = 4;
```

SHARD_ROW_ID_BITS 的值可以动态修改，每次修改之后，只对新写入的数据生效。

对于含有 CLUSTERED 主键的表，TiDB 会使用表的主键作为 RowID，因为 SHARD_ROW_ID_BITS 会改变 RowID 生成规则，所以此时无法使用 SHARD_ROW_ID_BITS 选项。而对于使用 NONCLUSTERED 主键的表，TiDB 会使用自动分配的 64 位整数作为 RowID，此时也可以使用 SHARD_ROW_ID_BITS 特性。要了解关于 CLUSTERED 主键的详细信息，请参考聚簇索引。

以下是两张无主键情况下使用 SHARD_ROW_ID_BITS 打散热点后的流量图，第一张展示了打散前的情况，第二张展示了打散后的情况。

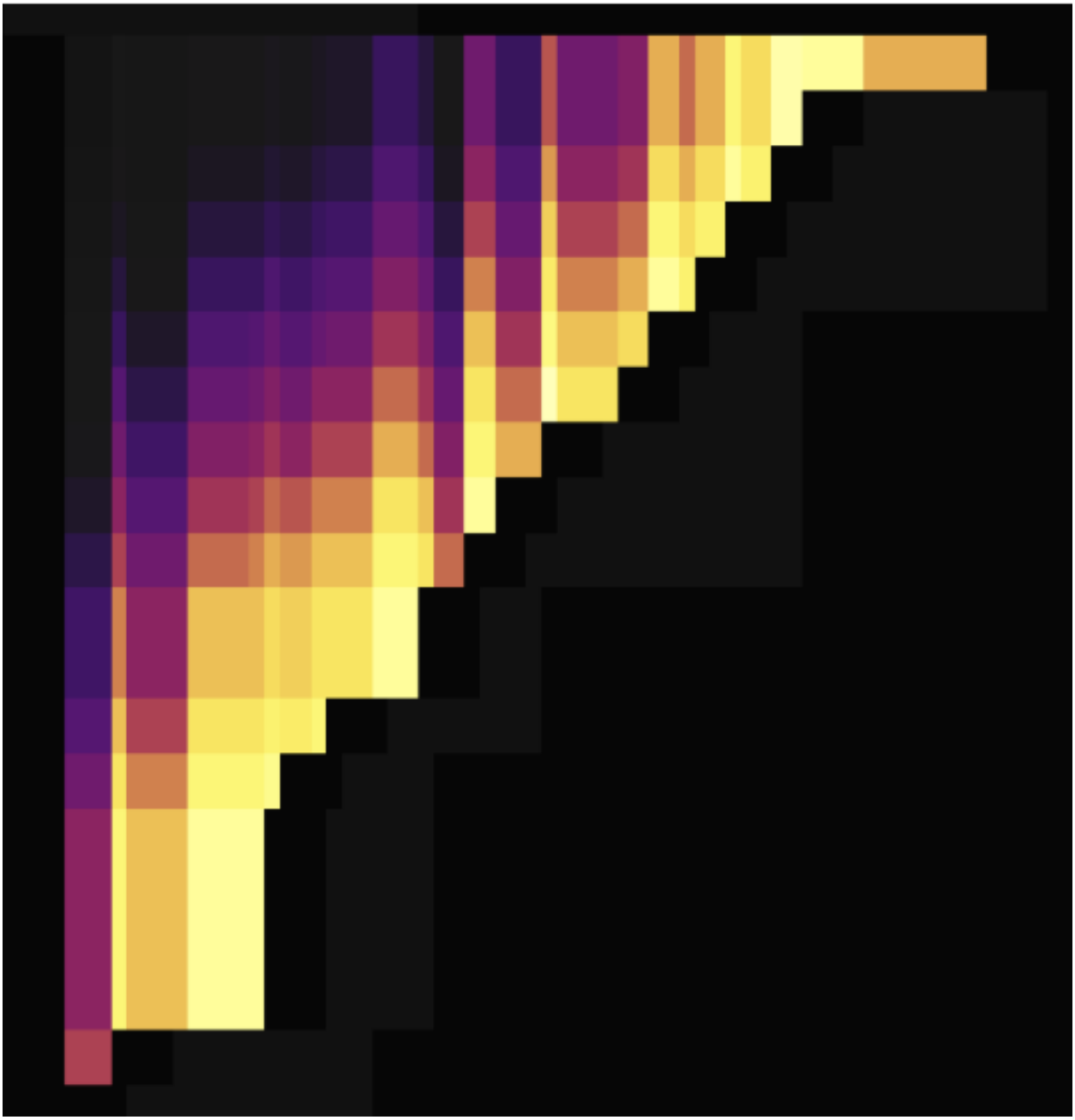


图 332: Dashboard 示例 5

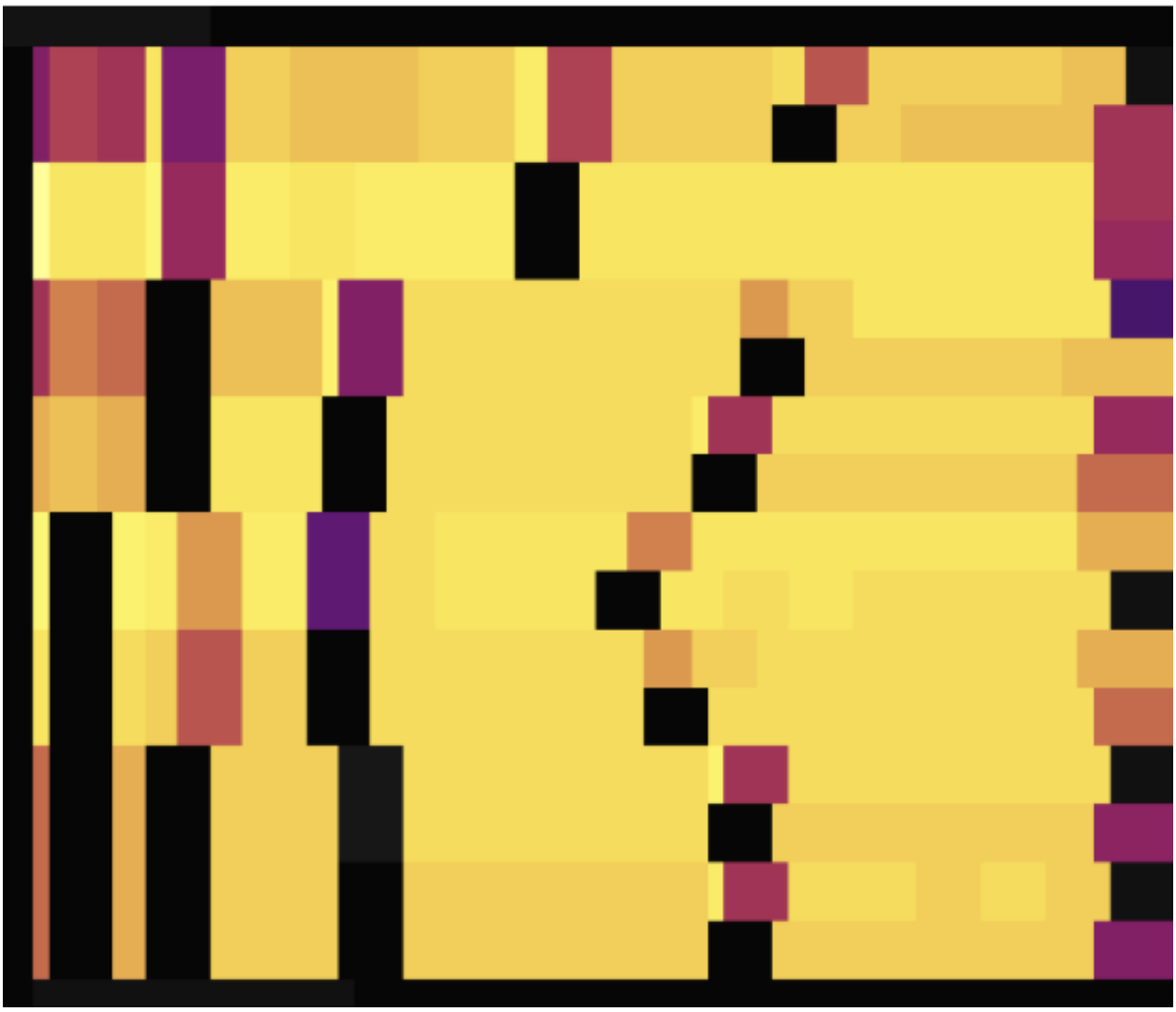


图 333: Dashboard 示例 6

从流量图可见，设置 SHARD_ROW_ID_BITS 后，流量热点由之前的只在一个 Region 上变得很分散。

13.1.4.5 使用 AUTO_RANDOM 处理自增主键热点表

使用 AUTO_RANDOM 处理自增主键热点表，适用于代替自增主键，解决自增主键带来的写入热点。

使用该功能后，将由 TiDB 生成随机分布且空间耗尽前不重复的主键，达到离散写入、打散写入热点的目的。

注意 TiDB 生成的主键不再是自增的主键，可使用 LAST_INSERT_ID() 获取上次分配的主键值。

将建表语句中的 AUTO_INCREMENT 改为 AUTO_RANDOM 即可使用该功能，适用于主键只需要保证唯一，不包含业务意义的场景。示例如下：

```
CREATE TABLE t (a BIGINT PRIMARY KEY AUTO_RANDOM, b varchar(255));
INSERT INTO t (b) VALUES ("foo");
SELECT * FROM t;
```

```
+------------+---+
| a          | b |
+------------+---+
| 1073741825 | b |
+------------+---+
```

```
SELECT LAST_INSERT_ID();
```

```
+------------------+
| LAST_INSERT_ID() |
+------------------+
| 1073741825       |
+------------------+
```

以下是将 AUTO_INCREMENT 表改为 AUTO_RANDOM 打散热点后的流量图，第一张是 AUTO_INCREMENT，第二张是 AUTO_RANDOM。

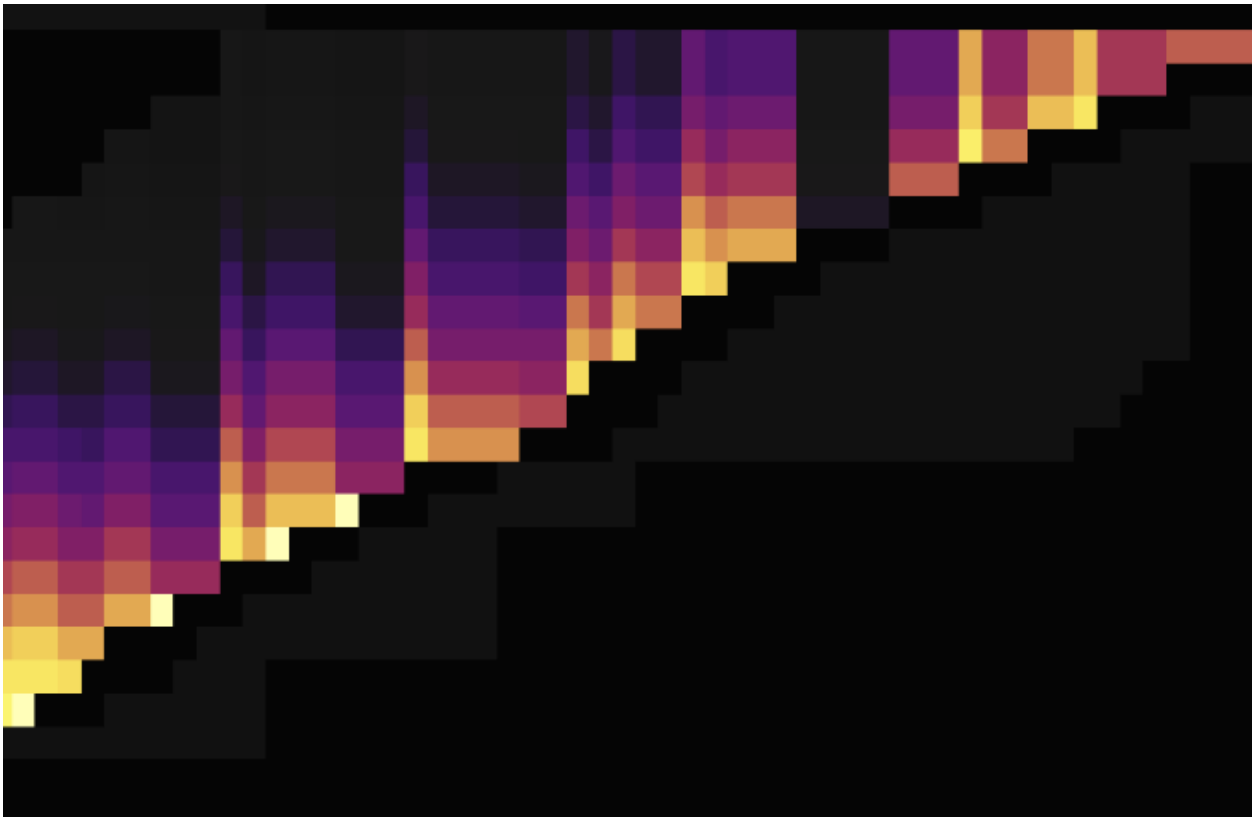


图 334: Dashboard 示例 7

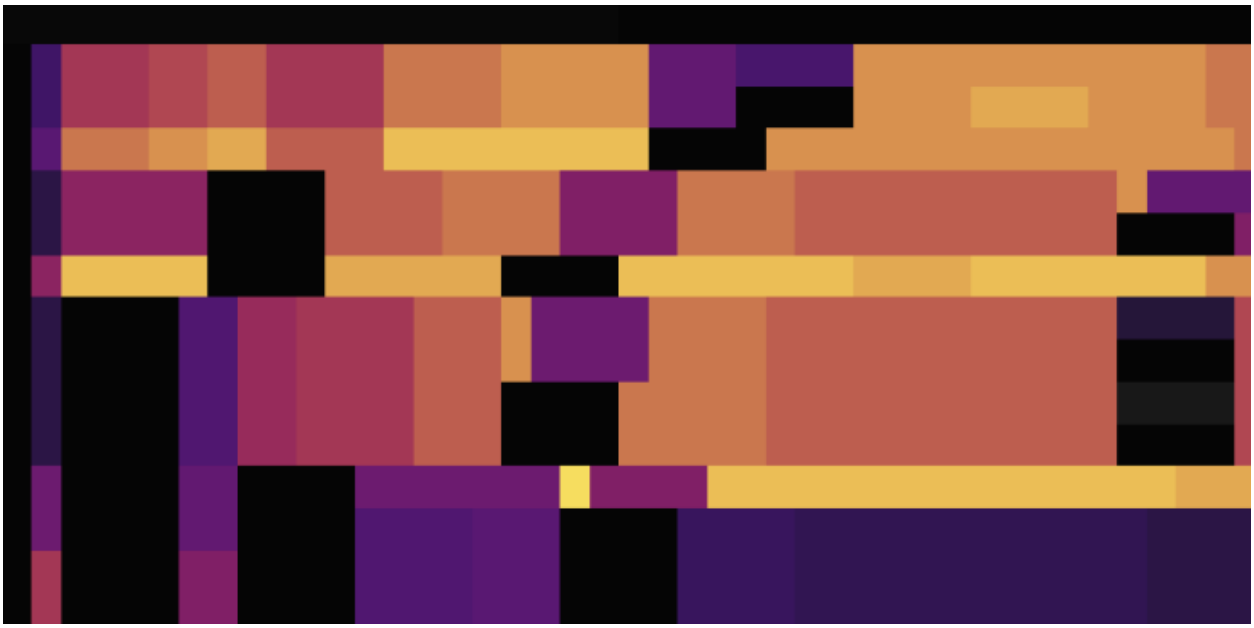


图 335: Dashboard 示例 8

由流量图可见，使用 AUTO_RANDOM 代替 AUTO_INCREMENT 能很好地打散热点。

更详细的说明参见 AUTO_RANDOM 文档。

13.1.4.6 小表热点的优化

TiDB 的 Coprocessor Cache 功能支持下推计算结果缓存。开启该功能后，将在 TiDB 实例侧缓存下推给 TiKV 计算的结果，对于小表读热点能起到比较好的效果。

更详细的说明参见下推计算结果缓存文档。

其他相关资料：

- TiDB 高并发写入场景最佳实践
- Split Region 使用文档

13.1.5 读写延迟增加

本文档介绍读写延迟增加、抖动时的排查思路，可能的原因和解决方法。

13.1.5.1 常见原因

13.1.5.1.1 TiDB 执行计划不对导致延迟增高

查询语句的执行计划不稳定，偶尔执行计划选择错误的索引，导致查询延迟增加。

现象：

- 如果慢日志中输出了执行计划，可以直接查看执行计划。用 `select tidb_decode_plan('xxx...')` 语句可以解析出具体的执行计划。
- 监控中的 key 扫描异常升高；慢日志中 SQL 执行时间 `Scan Keys` 数目较大。
- SQL 执行时间相比于其他数据库（例如 MySQL）有较大差距。可以对比其他数据库执行计划，例如 `Join Order` 是否不同。

可能的原因：

- 统计信息不准确

解决方案：

- 更新统计信息
  - 手动 `analyze table`，配合 crontab 定期 `analyze`，维持统计信息准确度。
  - 自动 `auto analyze`，调低 `analyze ratio` 阈值，提高收集频次，并设置运行时间窗口。示例如下：
    * `set global tidb_auto_analyze_ratio=0.2;`
    * `set global tidb_auto_analyze_start_time='00:00 +0800';`
    * `set global tidb_auto_analyze_end_time='06:00 +0800';`
- 绑定执行计划
  - 修改业务 SQL，使用 `use index` 固定使用列上的索引。
  - 3.0 版本下，业务可以不用修改 SQL，使用 `create global binding` 创建 `force index` 的绑定 SQL。
  - 4.0 版本支持 SQL Plan Management，可以避免因执行计划不稳定导致的性能下降。

13.1.5.1.2 PD 异常

现象：

监控中 PD TSO 的 wait duration 异常升高。wait duration 代表从开始等待 PD 返回，到等待结束的时间。

可能的原因：

- 磁盘问题。PD 所在的节点 I/O 被占满，排查是否有其他 I/O 高的组件与 PD 混合部署以及磁盘的健康情况，可通过监控 Grafana -> disk performance -> latency 和 load 等指标进行验证，必要时可以使用 fio 工具对盘进行检测，见案例 case-292。
- PD 之间的网络问题。PD 日志中有 `"lost the TCP streaming connection"`，排查 PD 之间网络是否有问题，可通过监控 Grafana -> PD -> etcd 的 round trip 来验证，见案例 case-177。
- 系统负载高，日志中能看到 `"server is likely overloaded"`，见案例 case-214。
- 选举不出 leader。PD 日志中有 `"lease is not expired"`，见 issues https://github.com/etcd-io/etcd/issues/10355。v3.0.x 版本和 v2.1.19 版本已解决该问题，见案例 case-875。
- 选举慢。Region 加载时间长，从 PD 日志中 `grep "regions cost"`（例如日志中可能是 `load 460927 regions cost 11.77099s`），如果出现秒级，则说明较慢，v3.0 版本可开启 Region Storage（设置 `use-region-storage` 为 `true`），该特性能极大缩短加载 Region 的时间，见案例 case-429。

- TiDB 与 PD 之间的网络问题，应排查网络相关情况。通过监控 Grafana -> blackbox_exporter -> ping latency 确定 TiDB 到 PD leader 的网络是否正常。
- PD 报 FATAL 错误，日志中有 "range failed to find revision pair"。v3.0.8 中已经解决问题，见 PR https://github.com/pingcap/pd/pull/2040。详情请参考案例 case-947。
- 使用 /api/v1/regions 接口时 Region 数量过多可能会导致 PD OOM。已于 v3.0.8 版本修复，见 https://github.com/pingcap/pd/pull/1986。
- 滚动升级的时候 PD OOM，gRPC 消息大小没限制，监控可看到 TCP InSegs 较大，已于 v3.0.6 版本修复，见 https://github.com/pingcap/pd/pull/1952。
- PD panic，请提交 bug。
- 其他原因，通过 curl http://127.0.0.1:2379/debug/pprof/goroutine?debug=2 抓取 goroutine，并提交 bug。

13.1.5.1.3 TiKV 异常

现象：

监控中 KV Cmd Duration 异常升高。KV Cmd Duration 是 TiDB 发送请求给 TiKV 到收到回复的延迟。

可能的原因：

- 查看 gRPC duration。gRPC duration 是请求在 TiKV 端的总耗时。通过对比 TiKV 的 gRPC duration 以及 TiDB 中的 KV duration 可以发现潜在的网络问题。比如 gRPC duration 很短但是 TiDB 的 KV duration 显示很长，说明 TiDB 和 TiKV 之间网络延迟可能很高，或者 TiDB 和 TiKV 之间的网卡带宽被占满。
- TiKV 重启了导致重新选举
  - TiKV panic 之后又被 systemd 重新拉起正常运行，可以通过查看 TiKV 的日志来确认是否有 panic，这种情况属于非预期，需要报 bug。
  - 被第三者 stop/kill，被 systemd 重新拉起。查看 dmesg 和 TiKV log 确认原因。
  - TiKV 发生 OOM 导致重启了。
  - 动态调整 THP 导致 hung 住，见案例 case-500。
- 查看监控：Grafana -> TiKV-details -> errors 面板 server is busy 看到 TiKV RocksDB 出现 write stall 导致发生重新选举。
- TiKV 发生网络隔离导致重新选举。
- block-cache 配置太大导致 OOM，在监控 Grafana -> TiKV-details 选中对应的 instance 之后查看 RocksDB 的 block cache size 监控来确认是否是该问题。同时请检查 [storage.block-cache] capacity = # "1GB" 参数是否设置合理，默认情况下 TiKV 的 block-cache 设置为机器总内存的 45%。在容器化部署时需要显式指定该参数，因为 TiKV 获取的是物理机的内存，可能会超出单个 container 的内存限制。
- Coprocessor 收到大量大查询，返回的数据量太大，gRPC 发送速度跟不上 Coprocessor 向客户端输出数据的速度导致 OOM。可以通过检查监控：Grafana -> TiKV-details -> coprocessor overview 的 response size 是否超过 network outbound 流量来确认是否属于这种情况。

13.1.5.1.4 TiKV 单线程瓶颈

TiKV 中存在一些单线程线程，可能会成为瓶颈。

- 单个 TiKV Region 过多导致单个 gRPC 线程成为瓶颈（查看 Grafana -> TiKV-details -> Thread CPU/gRPC CPU Per Thread 监控），v3.x 以上版本可以开启 Hibernate Region 特性来解决，见案例 case-612。
- v3.0 之前版本 Raftstore 单线程或者 Apply 单线程到达瓶颈（Grafana -> TiKV-details -> Thread CPU/raft store CPU 和 Async apply CPU 超过 80%），可以选择扩容 TiKV（v2.x 版本）实例或者升级到多线程模型的 v3.x 版本。

13.1.5.1.5 CPU Load 升高

现象：

CPU 资源使用到达瓶颈

可能的原因：

- 热点问题。
- 整体负载高，排查 TiDB 的 slow query 和 expensive query。对运行的 query 进行优化，如果缺索引就加索引，如果可以批量执行就批量执行。另一个方案是对集群进行扩容。

13.1.5.2 其它原因

13.1.5.2.1 集群维护

通常大多数的线上集群有 3 或 5 个 PD 节点，如果维护的主机上有 PD 组件，需要具体考虑节点是 leader 还是 follower，关闭 follower 对集群运行没有任何影响，关闭 leader 需要先切换，并在切换时有 3 秒左右的性能抖动。

13.1.5.2.2 少数派副本离线

TiDB 集群默认配置为 3 副本，每一个 Region 都会在集群中保存 3 份，它们之间通过 Raft 协议来选举 Leader 并同步数据。Raft 协议可以保证在数量小于副本数（注意：不是节点数）一半的节点挂掉或者隔离的情况下，仍然能够提供服务，并且不丢失任何数据。对于 3 副本集群，挂掉一个节点可能会导致性能抖动，可用性和正确性理论上不会受影响。

13.1.5.2.3 新增索引

由于创建索引在扫表回填索引的时候会消耗大量资源，甚至与一些频繁更新的字段会发生冲突导致正常业务受到影响。大表创建索引的过程往往会持续很长时间，所以要尽可能地平衡执行时间和集群性能之间的关系，比如选择非高频更新时间段。

参数调整：

目前主要使用 `tidb_ddl_reorg_worker_cnt` 和 `tidb_ddl_reorg_batch_size` 这两个参数来动态调整索引创建速度，通常来说它们的值越小对系统影响越小，但是执行时间越长。

一般情况下，先将值保持为默认的 4 和 256，观察集群资源使用情况和响应速度，再逐渐调大 `tidb_ddl_reorg_worker_cnt` 参数来增加并发，观察监控如果系统没有发生明显的抖动，再逐渐调大 `tidb_ddl_reorg_batch_size` 参数，但如果索引涉及的列更新很频繁的话就会造成大量冲突造成失败重试。

另外还可以通过调整参数 tidb_ddl_reorg_priority 为 PRIORITY_HIGH 来让创建索引的任务保持高优先级来提升速度，但在通用 OLTP 系统上，一般建议保持默认。

13.1.5.2.4 GC 压力大

TiDB 的事务的实现采用了 MVCC（多版本并发控制）机制，当新写入的数据覆盖旧的数据时，旧的数据不会被替换掉，而是与新写入的数据同时保留，并以时间戳来区分版本。GC 的任务便是清理不再需要的旧数据。

- Resolve Locks 阶段在 TiKV 一侧会产生大量的 scan_lock 请求，可以在 gRPC 相关的 metrics 中观察到。scan_lock 请求会对全部的 Region 调用。
- Delete Ranges 阶段会往 TiKV 发送少量的 unsafe_destroy_range 请求，也可能没有。可以在 gRPC 相关的 metrics 中和 GC 分类下的 GC tasks 中观察到。
- Do GC 阶段，默认每台 TiKV 会自动扫描本机上的 leader Region 并对每一个 leader 进行 GC，这一活动可以在 GC 分类下的 GC tasks 中观察到。

13.1.6 乐观事务模型下写写冲突问题排查

本文介绍 TiDB 中乐观锁下写写冲突出现的原因以及解决方案。

在 v3.0.8 版本之前，TiDB 默认采用乐观事务模型，在事务执行过程中并不会做冲突检测，而是在事务最终 COMMIT 提交时触发两阶段提交，并检测是否存在写写冲突。当出现写写冲突，并且开启了事务重试机制，则 TiDB 会在限定次数内进行重试，最终重试成功或者达到重试次数上限后，会给客户端返回结果。因此，如果 TiDB 集群中存在大量的写写冲突情况，容易导致集群的 Duration 比较高。

13.1.6.1 出现写写冲突的原因

TiDB 中使用 Percolator 事务模型来实现 TiDB 中的事务。Percolator 总体上就是一个二阶段提交的实现。具体的二阶段提交过程可参考乐观事务文档。

当客户端发起 COMMIT 请求的时候，TiDB 开始两阶段提交：

1. TiDB 从所有要写入的 Key 中选择一个作为当前事务的 Primary Key
2. TiDB 向所有的本次提交涉及到的 TiKV 发起 prewrite 请求，TiKV 判断是否所有 Key 都可以 prewrite 成功
3. TiDB 收到所有 Key 都 prewrite 成功的消息
4. TiDB 向 PD 请求 commit_ts
5. TiDB 向 Primary Key 发起第二阶段提交。Primary Key 所在的 TiKV 收到 commit 操作后，检查数据合法性，清理 prewrite 阶段留下的锁
6. TiDB 收到两阶段提交成功的信息

写写冲突发生在 prewrite 阶段，当发现有其他的事务在写当前 Key (data.commit_ts > txn.start_ts)，则会发生写写冲突。

TiDB 会根据 tidb_disable_txn_auto_retry 和 tidb_retry_limit 参数设置的情况决定是否进行重试，如果设置了不重试，或者重试次数达到上限后还是没有 prewrite 成功，则向 TiDB 返回 Write Conflict 错误。

13.1.6.2 如何判断当前集群存在写写冲突

可以通过 Grafana 监控查看集群写写冲突的情况：

- 通过 TiDB 监控面板中 KV Errors 监控栏中 KV Backoff OPS 监控指标项，查看 TiKV 中返回错误信息的数量

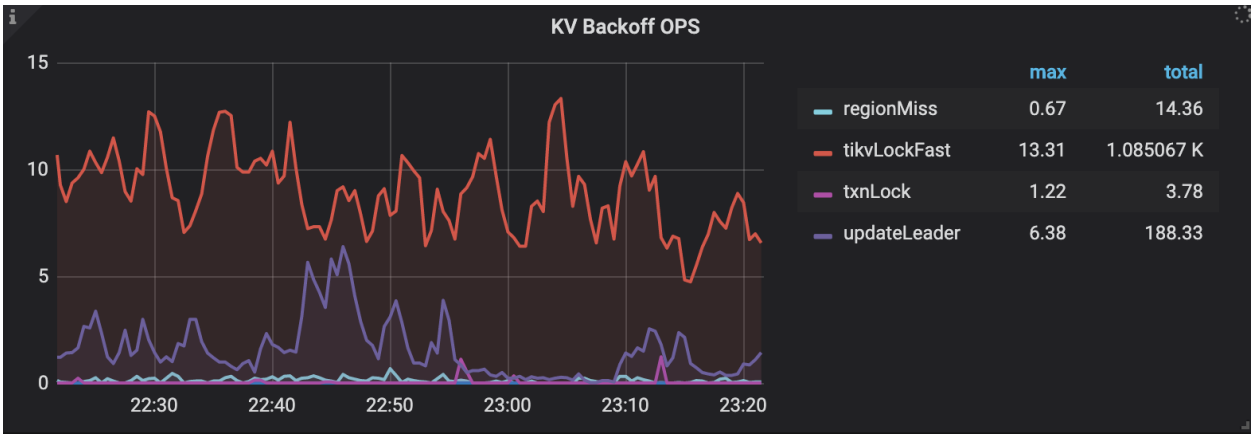


图 336: kv-backoff-ops

txnlock 表示集群中存在写写冲突，txnLockFast 表示集群中存在读写冲突。

- 通过 TiDB 监控面板中 KV Errors 监控栏中 Lock Resolve OPS 监控指标项，查看事务冲突相关的数量

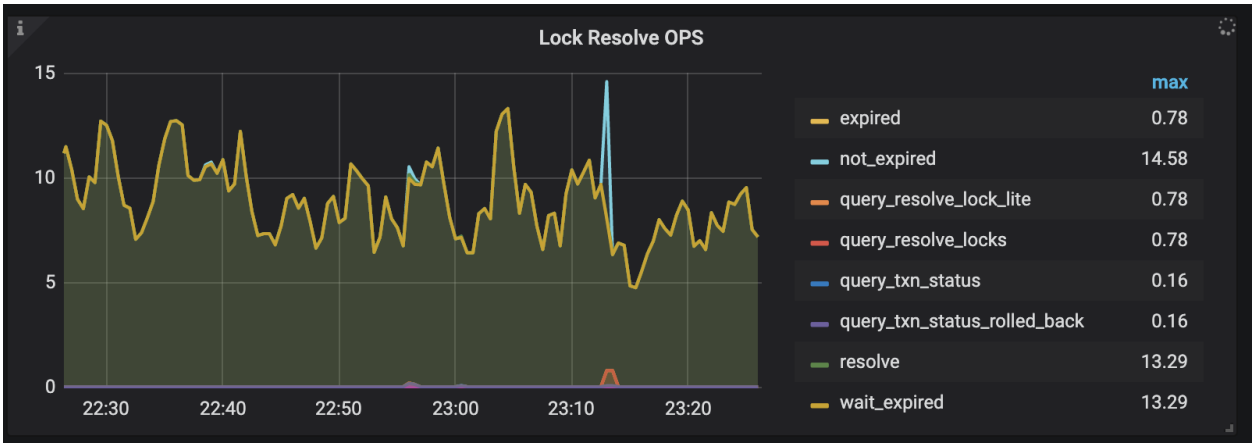


图 337: lock-resolve-ops

expired、 not_expired、 wait_expired 表示对应的 lock 状态

- 通过 TiDB 监控面板中 KV Errors 监控栏中 KV Retry Duration 监控指标项，查看 KV 重试请求的时间

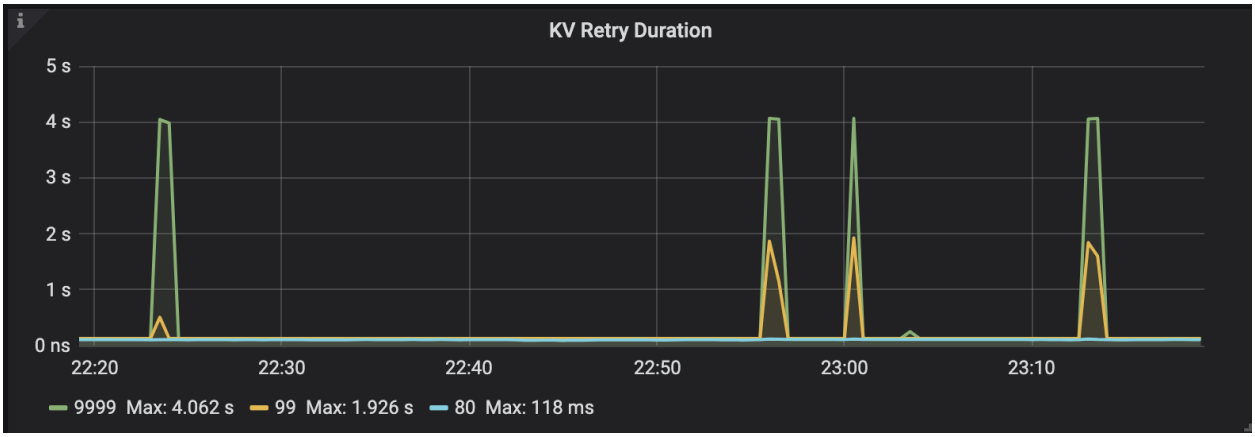


图 338: kv-retry-duration

也可以通过 TiDB 日志查看是否有 `[kv:9007]Write conflict` 关键字，如果搜索到对应关键字，则可以表明集群中存在写写冲突。

13.1.6.3 如何解决写写冲突问题

如果通过以上方式判断出集群中存在大量的写写冲突，建议找到冲突的数据，以及写写冲突的原因，看是否能从应用程序修改逻辑，加上重试的逻辑。当出现写写冲突的时候，可以在 TiDB 日志中看到类似的日志：

```
[2020/05/12 15:17:01.568 +08:00] [WARN] [session.go:446] ["commit failed"] [conn=3] ["finished
    ↪ txn"="Txn{state=invalid}"] [error="[kv:9007]Write conflict, txnStartTS
    ↪ =416617006551793665, conflictStartTS=416617018650001409, conflictCommitTS
    ↪ =416617023093080065, key={tableID=47, indexID=1, indexValues={string, }} primary={tableID
    ↪ =47, indexID=1, indexValues={string, }} [try again later]"]
```

关于日志的解释如下：

- `[kv:9007]Write conflict`：表示出现了写写冲突
- `txnStartTS=416617006551793665`：表示当前事务的 start_ts 时间戳，可以通过 pd-ctl 工具将时间戳转换为具体时间
- `conflictStartTS=416617018650001409`：表示冲突事务的 start_ts 时间戳，可以通过 pd-ctl 工具将时间戳转换为具体时间
- `conflictCommitTS=416617023093080065`：表示冲突事务的 commit_ts 时间戳，可以通过 pd-ctl 工具将时间戳转换为具体时间
- `key={tableID=47, indexID=1, indexValues={string, }}`：表示当前事务中冲突的数据，tableID 表示发生冲突的表的 ID，indexID 表示是索引数据发生了冲突。如果是数据发生了冲突，会打印 `handle=x` 表示对应哪行数据发生了冲突，indexValues 表示发生冲突的索引数据
- `primary={tableID=47, indexID=1, indexValues={string, }}`：表示当前事务中的 Primary Key 信息

通过 pd-ctl 将时间戳转换为可读时间：

```
./pd-ctl -u https://127.0.0.1:2379 tso {TIMESTAMP}
```

通过 tableID 查找具体的表名：

```
curl http://{TiDBIP}:10080/db-table/{tableID}
```

通过 indexID 查找具体的索引名：

```
SELECT * FROM INFORMATION_SCHEMA.TIDB_INDEXES WHERE TABLE_SCHEMA='{db_name}' AND TABLE_NAME='{
    ↪ table_name}' AND INDEX_ID={indexID};
```

另外在 v3.0.8 及之后版本默认使用悲观事务模式，从而避免在事务提交的时候因为冲突而导致失败，无需修改应用程序。悲观事务模式下会在每个 DML 语句执行的时候，加上悲观锁，用于防止其他事务修改相同 Key，从而保证在最后提交的 prewrite 阶段不会出现写写冲突的情况。

13.1.7 TiDB 磁盘 I/O 过高的处理办法

本文主要介绍如何定位和处理 TiDB 存储 I/O 过高的问题。

13.1.7.1 确认当前 I/O 指标

当出现系统响应变慢的时候，如果已经排查了 CPU 的瓶颈、数据事务冲突的瓶颈后，就需要从 I/O 来入手来辅助判断目前的系统瓶颈点。

13.1.7.1.1 从监控定位 I/O 问题

最快速的定位手段是从监控来查看整体的 I/O 情况，可以从集群部署工具 (TiUP) 默认会部署的监控组件 Grafana 来查看对应的 I/O 监控，跟 I/O 相关的 Dashboard 有 `Overview`，`Node_exporter`，`Disk-Performance`。

第一类面板

在 `Overview` > `System Info` > `IO Util` 中，可以看到集群中每个机器的 I/O 情况，该指标和 Linux iostat 监控中的 util 类似，百分比越高代表磁盘 I/O 占用越高：

- 如果监控中只有一台机器的 I/O 高，那么可以辅助判断当前有读写热点。
- 如果监控中大部分机器的 I/O 都高，那么集群现在有高 I/O 负载存在。

如果发现某台机器的 I/O 比较高，可以从监控 `Disk-Performance Dashboard` 进一步观察 I/O 的使用情况，结合 `Disk Latency`，`Disk Load` 等 metric 判断是否存在异常，必要时可以使用 fio 工具对磁盘进行检测。

第二类面板

TiDB 集群主要的持久化组件是 TiKV 集群，一个 TiKV 包含两个 RocksDB 实例：一个用于存储 Raft 日志，位于 data/raft，一个用于存储真正的数据，位于 data/db。

在 `TiKV-Details` > `Raft IO` 中，可以看到这两个实例磁盘写入的相关 metric：

- `Append log duration`：该监控表明了存储 Raft 日志的 RocksDB 写入的响应时间，.99 响应应该在 50ms 以内。
- `Apply log duration`：该监控表明了存储真正数据的 RocksDB 写入的响应时间，.99 响应应该在 100ms 以内。

这两个监控还有 `.. per server` 的监控面板来提供辅助查看热点写入的情况。

第三类面板

在 `TiKV-Details > Storage` 中，有关于 storage 相关情况的监控：

- `Storage command total`：收到的不同命令的个数。
- `Storage async write duration`：包括了磁盘 sync duration 等监控项，可能和 Raft IO 有关。如遇到异常情况，需要通过 log 来检查相关组件的工作状态是否正常。

其他面板

此外，可能还需要一些其它内容来辅助确认瓶颈是否为 I/O，并可以尝试调整一些参数。通过查看 TiKV gRPC 的 prewrite/commit/raw-put（仅限 raw kv 集群）duration，确认确实是 TiKV 写入慢了。常见的几种情况如下：

- append log 慢。TiKV Grafana 的 Raft I/O 和 append log duration 比较高，通常情况下是由于写盘慢了，可以检查 RocksDB - raft 的 WAL Sync Duration max 值来确认，否则可能需要报 bug。
- raftstore 线程繁忙。TiKV grafana 的 Raft Propose/propose wait duration 明显高于 append log duration。请查看以下两点：
  - `[raftstore]` 的 `store-pool-size` 配置是否过小（该值建议在 [1,5] 之间，不建议太大）。
  - 机器的 CPU 是不是不够了。
- apply log 慢。TiKV Grafana 的 Raft I/O 和 apply log duration 比较高，通常会伴随着 Raft Propose/apply wait duration 比较高。可能的情况如下：
  - `[raftstore]` 的 `apply-pool-size` 配置过小（建议在 [1, 5] 之间，不建议太大），Thread CPU/apply cpu 比较高；
  - 机器的 CPU 资源不够了。
  - Region 写入热点问题，单个 apply 线程 CPU 使用率比较高（通过修改 Grafana 表达式，加上 by (instance, name) 来看各个线程的 cpu 使用情况），暂时对于单个 Region 的热点写入没有很好的方式，最近在优化该场景。
  - 写 RocksDB 比较慢，RocksDB kv/max write duration 比较高（单个 Raft log 可能包含很多个 kv，写 RocksDB 的时候会把 128 个 kv 放在一个 write batch 写入到 RocksDB，所以一次 apply log 可能涉及到多次 RocksDB 的 write）。
  - 其他情况，需要报 bug。
- raft commit log 慢。TiKV Grafana 的 Raft I/O 和 commit log duration 比较高（4.x 版本的 Grafana 才有该 metric）。每个 Region 对应一个独立的 Raft group，Raft 本身是有流控机制的，类似 TCP 的滑动窗口机制，通过参数 [raftstore] raft-max-inflight-msgs = 256 来控制滑动窗口的大小，如果有热点写入并且 commit log duration 比较高可以适度调大改参数，比如 1024。

13.1.7.1.2 从 log 定位 I/O 问题

- 如果客户端报 `server is busy` 错误，特别是 `raftstore is busy` 的错误信息，会和 I/O 有相关性。

  可以通过查看监控：grafana -> TiKV -> errors 监控确认具体 busy 原因。其中，`server is busy` 是 TiKV 自身的流控机制，TiKV 通过这种方式告知 `tidb/ti-client` 当前 TiKV 的压力过大，过一会儿再尝试。

- TiKV RocksDB 日志出现 `write stall`。

  可能是 `level0 sst` 太多导致 stall。可以添加参数 `[rocksdb] max-sub-compactions = 2`（或者 `3`）加快 level0 sst 往下 compact 的速度，该参数的意思是将从 level0 到 level1 的 compaction 任务最多切成 `max-sub-compactions` 个子任务交给多线程并发执行。

  如果磁盘 I/O 能力持续跟不上写入，建议扩容。如果磁盘的吞吐达到了上限（例如 SATA SSD 的吞吐相对 NVME SSD 会低很多）导致 write stall，但是 CPU 资源又比较充足，可以尝试采用压缩率更高的压缩算法来缓解磁盘的压力，用 CPU 资源换磁盘资源。

  比如 `default cf compaction` 压力比较大时，可以调整参数 `[rocksdb.defaultcf] compression-per-level = ["no", "no", "lz4", "lz4", "lz4", "zstd", "zstd"]` 改成 `compression-per-level = ["no", "no", "zstd", "zstd", "zstd", "zstd", "zstd"]`。

13.1.7.1.3 从告警发现 I/O 问题

集群部署工具 (TiUP) 默认部署的告警组件，官方已经预置了相关的告警项目和阈值，I/O 相关项包括：

- TiKV_write_stall
- TiKV_raft_log_lag
- TiKV_async_request_snapshot_duration_seconds
- TiKV_async_request_write_duration_seconds
- TiKV_raft_append_log_duration_secs
- TiKV_raft_apply_log_duration_secs

13.1.7.2 I/O 问题处理方案

1. 当确认为热点 I/O 问题的时候，需要参考TiDB 热点问题处理来消除相关的热点 I/O 情况。
2. 当确认整体 I/O 已经到达瓶颈的时候，且从业务侧能够判断 I/O 的能力会持续的跟不上，那么就可以利用分布式数据库的 scale 的能力，采用扩容 TiKV 节点数量的方案来获取更大的整体 I/O 吞吐量。
3. 调整上述说明中的一些参数，使用计算/内存资源来换取磁盘的存储资源。

13.1.8 TiDB 锁冲突问题处理

TiDB 支持完整的分布式事务，自 v3.0 版本起，提供乐观事务与悲观事务两种事务模式。本文介绍如何使用 Lock View 排查锁相关的问题，以及如何处理使用乐观事务或者悲观事务的过程中常见的锁冲突问题。

13.1.8.1 使用 Lock View 排查锁相关的问题

自 v5.1 版本起，TiDB 支持 Lock View 功能。该功能在 `information_schema` 中内置了若干系统表，用于提供更多关于锁冲突和锁等待的信息。

> **注意：**
>
> Lock View 功能目前仅提供悲观锁的冲突和等待信息。

关于这些系统表的详细说明，请参考以下文档：

- TIDB_TRX 与 CLUSTER_TIDB_TRX：提供当前 TiDB 节点上或整个集群上所有运行中的事务的信息，包括事务是否处于等锁状态、等锁时间和事务曾经执行过的语句的 Digest 等信息。
- DATA_LOCK_WAITS：提供关于 TiKV 内的悲观锁等锁信息，包括阻塞和被阻塞的事务的 start_ts、被阻塞的 SQL 语句的 Digest 和发生等待的 key。
- DEADLOCKS 与 CLUSTER_DEADLOCKS：提供当前 TiDB 节点上或整个集群上最近发生过的若干次死锁的相关信息，包括死锁环中事务之间的等待关系、事务当前正在执行的语句的 Digest 和发生等待的 key。

> 注意：
>
> Lock View 所属的系统表中展示的 SQL 语句为归一化的 SQL 语句（即去除了格式和参数的 SQL 语句），通过内部查询从 SQL Digest 获得，因而无法获取包括格式和参数在内的完整语句。有关 SQL Digest 和归一化 SQL 语句的详细介绍，请参阅Statement Summary Tables。

以下为排查部分问题的示例。

13.1.8.1.1 死锁错误

要获取最近发生的死锁错误的信息，可查询 DEADLOCKS 或 CLUSTER_DEADLOCKS 表。

以查询 DEADLOCKS 表为例，请执行以下 SQL 语句：

```
select * from information_schema.deadlocks;
```

示例输出：

```
+--
    ↪ -----------+----------------------------+-----------+--------------------+-----------------------------
    ↪
| DEADLOCK_ID | OCCUR_TIME                 | RETRYABLE | TRY_LOCK_TRX_ID    | CURRENT_SQL_DIGEST
    ↪                                                 | CURRENT_SQL_DIGEST_TEXT                 |
    ↪  KEY                                    | KEY_INFO
    ↪
    ↪ | TRX_HOLDING_LOCK   |
+--
    ↪ -----------+----------------------------+-----------+--------------------+-----------------------------
    ↪
|           1 | 2021-08-05 11:09:03.230341 |         0 | 426812829645406216 | 22230766411
    ↪ edb40f27a68dadefc63c6c6970d5827f1e5e22fc97be2c4d8350d | update `t` set `v` = ? where `id`
    ↪  = ? ; | 748000000000000355F728000000000000002 | {"db_id":1,"db_name":"test","table_id"
    ↪ :53,"table_name":"t","handle_type":"int","handle_value":"2"} | 426812829645406217 |
|           1 | 2021-08-05 11:09:03.230341 |         0 | 426812829645406217 | 22230766411
    ↪ edb40f27a68dadefc63c6c6970d5827f1e5e22fc97be2c4d8350d | update `t` set `v` = ? where `id`
    ↪  = ? ; | 748000000000000355F728000000000000001 | {"db_id":1,"db_name":"test","table_id"
    ↪ :53,"table_name":"t","handle_type":"int","handle_value":"1"} | 426812829645406216 |
```

```
+--
    ↪ -----------+----------------------------+-----------+--------------------+-------------------------------
    ↪
```

查询结果会显示死锁错误中多个事务之间的等待关系和各个事务当前正在执行的 SQL 语句的归一化形式（即去掉参数和格式的形式），以及发生冲突的 key 及其从 key 中解读出的一些信息。

例如在上述例子中，第一行意味着 ID 为 426812829645406216 的事务当前正在执行形如 `update `t` set `v` = ? where `id` = ?` 的语句，被另一个 ID 为 426812829645406217 的事务阻塞；而 426812829645406217 同样也在执行一条形如 `update `t` set `v` = ? where `id` = ?` 的语句，并被 ID 为 426812829645406216 的事务阻塞，两个事务因而构成死锁。

13.1.8.1.2 少数热点 key 造成锁排队

DATA_LOCK_WAITS 系统表提供 TiKV 节点上的等锁情况。查询该表时，TiDB 将自动从所有 TiKV 节点上获取当前时刻的等锁信息。当少数热点 key 频繁被上锁并阻塞较多事务时，你可以查询 DATA_LOCK_WAITS 表并按 key 对结果进行聚合，以尝试找出经常发生问题的 key：

```
select `key`, count(*) as `count` from information_schema.data_lock_waits group by `key` order by
    ↪  `count` desc;
```

示例输出：

```
+----------------------------------------+-------+
| key                                    | count |
+----------------------------------------+-------+
| 748000000000000415F728000000000000001 |     2 |
| 748000000000000415F728000000000000002 |     1 |
+----------------------------------------+-------+
```

为避免偶然性，你可考虑进行多次查询。

如果已知频繁出问题的 key，可尝试从 TIDB_TRX 或 CLUSTER_TIDB_TRX 表中获取试图上锁该 key 的事务的信息。

需要注意 TIDB_TRX 和 CLUSTER_TIDB_TRX 表所展示的信息也是对其进行查询的时刻正在运行的事务的信息，并不展示已经结束的事务。如果并发的事务数量很大，该查询的结果集也可能很大，可以考虑添加 limit 子句，或用 where 子句筛选出等锁时间较长的事务。需要注意，对 Lock View 中的多张表进行 join 时，不同表之间的数据并不保证在同一时刻获取，因而不同表中的信息可能并不同步。

以用 where 子句筛选出等锁时间较长的事务为例，请执行以下 SQL 语句：

```
select trx.* from information_schema.data_lock_waits as l left join information_schema.tidb_trx
    ↪ as trx on l.trx_id = trx.id where l.key = "748000000000000415F728000000000000001"\G
```

示例输出：

```
*************************** 1. row ***************************
                    ID: 426831815660273668
            START_TIME: 2021-08-06 07:16:00.081000
    CURRENT_SQL_DIGEST: 06da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7
```

```
CURRENT_SQL_DIGEST_TEXT: update `t` set `v` = `v` + ? where `id` = ? ;
                  STATE: LockWaiting
     WAITING_START_TIME: 2021-08-06 07:16:00.087720
        MEM_BUFFER_KEYS: 0
       MEM_BUFFER_BYTES: 0
             SESSION_ID: 77
                   USER: root
                     DB: test
        ALL_SQL_DIGESTS: ["0fdc781f19da1c6078c9de7eadef8a307889c001e05f107847bee4cfc8f3cdf3","06
            ↪ da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7"]
*************************** 2. row ***************************
                   ID: 426831818019569665
           START_TIME: 2021-08-06 07:16:09.081000
     CURRENT_SQL_DIGEST: 06da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7
CURRENT_SQL_DIGEST_TEXT: update `t` set `v` = `v` + ? where `id` = ? ;
                  STATE: LockWaiting
     WAITING_START_TIME: 2021-08-06 07:16:09.290271
        MEM_BUFFER_KEYS: 0
       MEM_BUFFER_BYTES: 0
             SESSION_ID: 75
                   USER: root
                     DB: test
        ALL_SQL_DIGESTS: ["0fdc781f19da1c6078c9de7eadef8a307889c001e05f107847bee4cfc8f3cdf3","06
            ↪ da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7"]
2 rows in set (0.00 sec)
```

13.1.8.1.3 事务被长时间阻塞

如果已知一个事务被另一事务（或多个事务）阻塞，且已知当前事务的 start_ts（即事务 ID），则可使用如下方式获取导致该事务阻塞的事务的信息。注意对 Lock View 中的多张表进行 join 时，不同表之间的数据并不保证在同一时刻获取，因而可能不同表中的信息可能并不同步。

```
select l.key, trx.*, tidb_decode_sql_digests(trx.all_sql_digests) as sqls from information_schema
    ↪ .data_lock_waits as l join information_schema.cluster_tidb_trx as trx on l.
    ↪ current_holding_trx_id = trx.id where l.trx_id = 426831965449355272\G
```

示例输出：

```
*************************** 1. row ***************************
                   key: 7480000000000004D5F728000000000000001
              INSTANCE: 127.0.0.1:10080
                   ID: 426832040186609668
           START_TIME: 2021-08-06 07:30:16.581000
     CURRENT_SQL_DIGEST: 06da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7
CURRENT_SQL_DIGEST_TEXT: update `t` set `v` = `v` + ? where `id` = ? ;
                  STATE: LockWaiting
```

```
    WAITING_START_TIME: 2021-08-06 07:30:16.592763
       MEM_BUFFER_KEYS: 1
      MEM_BUFFER_BYTES: 19
            SESSION_ID: 113
                  USER: root
                    DB: test
       ALL_SQL_DIGESTS: ["0fdc781f19da1c6078c9de7eadef8a307889c001e05f107847bee4cfc8f3cdf3","
           ↪ a4e28cc182bdd18288e2a34180499b9404cd0ba07e3cc34b6b3be7b7c2de7fe9","06
           ↪ da614b93e62713bd282d4685fc5b88d688337f36e88fe55871726ce0eb80d7"]
                  sqls: ["begin ;","select * from `t` where `id` = ? for update ;","update `t`
                      ↪ set `v` = `v` + ? where `id` = ? ;"]
1 row in set (0.01 sec)
```

上述查询中，对 CLUSTER_TIDB_TRX 表的 ALL_SQL_DIGESTS 列使用了TIDB_DECODE_SQL_DIGESTS 函数，目的是将该列（内容为一组 SQL Digest）转换为其对应的归一化 SQL 语句，便于阅读。

如果当前事务的 start_ts 未知，可以尝试从 TIDB_TRX / CLUSTER_TIDB_TRX 表或者 PROCESSLIST / CLUSTER_PROCESSLIST 表中的信息进行判断。

13.1.8.2 处理乐观锁冲突问题

以下介绍乐观事务模式下常见的锁冲突问题的处理方式。

13.1.8.2.1 读写冲突

在 TiDB 中，读取数据时，会获取一个包含当前物理时间且全局唯一递增的时间戳作为当前事务的 start_ts。事务在读取时，需要读到目标 key 的 commit_ts 小于这个事务的 start_ts 的最新的数据版本。当读取时发现目标 key 上存在 lock 时，因为无法知道上锁的那个事务是在 Commit 阶段还是 Prewrite 阶段，所以就会出现读写冲突的情况，如下图：

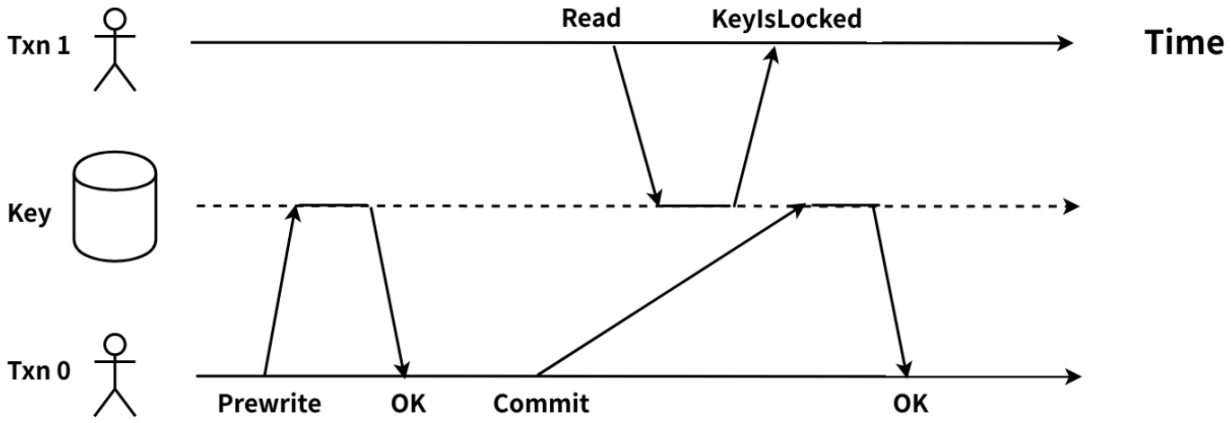


图 339: 读写冲突

分析：

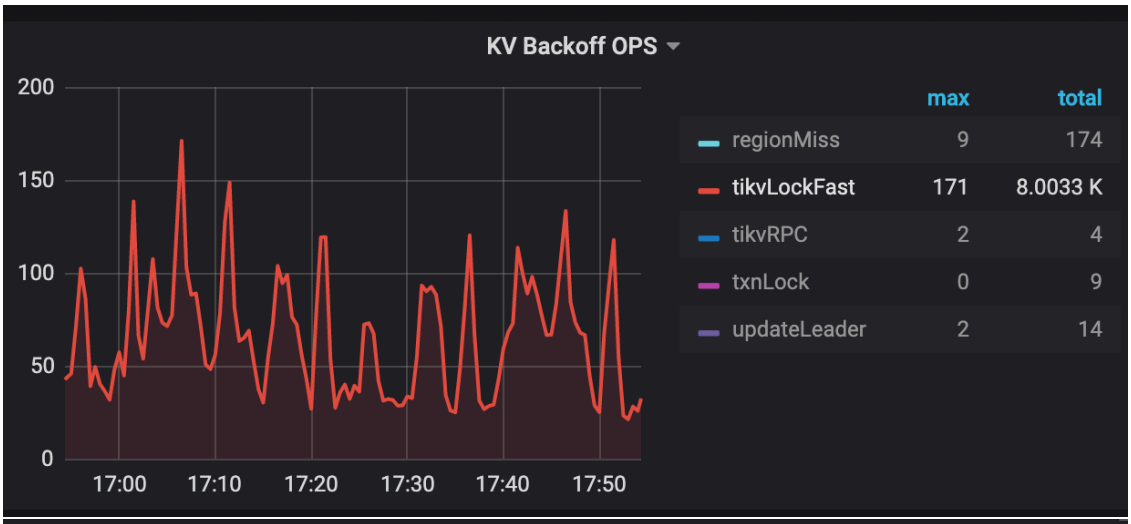
Txn0 完成了 Prewrite，在 Commit 的过程中 Txn1 对该 key 发起了读请求，Txn1 需要读取 start_ts > commit_ts 最近的 key 的版本。此时，Txn1 的 `start_ts` > Txn0 的 lock_ts，需要读取的 key 上的锁信息仍未清理，故无法判断 Txn0 是否提交成功，因此 Txn1 与 Txn0 出现读写冲突。

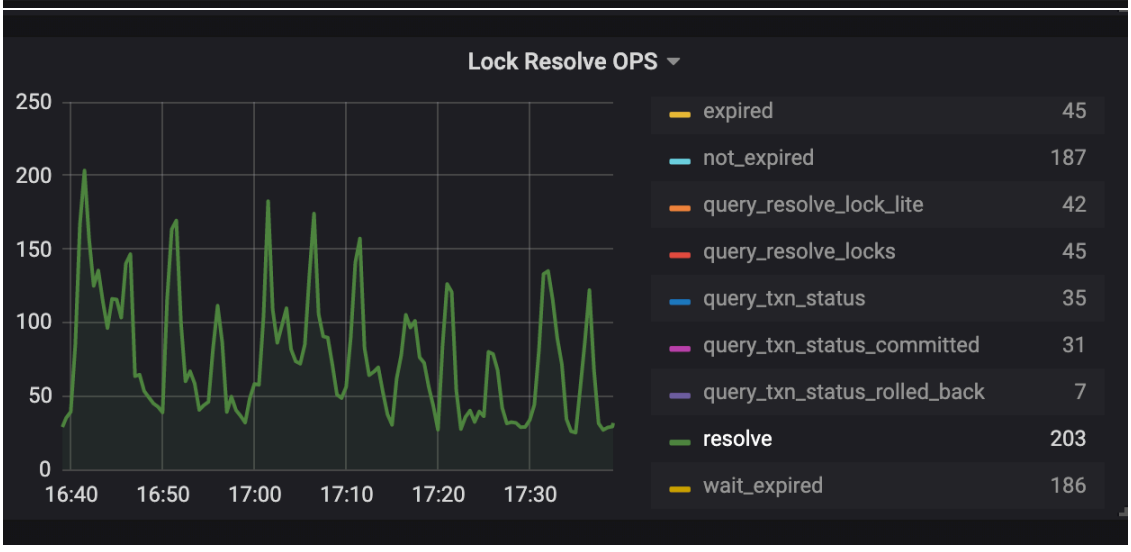
你可以通过如下两种途径来检测当前环境中是否存在读写冲突：

1. TiDB 监控及日志

    - 通过 TiDB Grafana 监控分析：

      观察 KV Errors 下 Lock Resolve OPS 面板中的 not_expired/resolve 监控项以及 KV Backoff OPS 面板中的 txnLockFast 监控项，如果有较为明显的上升趋势，那么可能是当前的环境中出现了大量的读写冲突。其中，not_expired 是指对应的锁还没有超时，resolve 是指尝试清锁的操作，txnLockFast 代表出现了读写冲突。





    - 通过 TiDB 日志分析：

      在 TiDB 的日志中可以看到下列信息：

```
[INFO] [coprocessor.go:743] ["[TIME_COP_PROCESS] resp_time:406.038899ms txnStartTS
    ↪ :416643508703592451 region_id:8297 store_addr:10.8.1.208:20160 backoff_ms:255
    ↪ backoff_types:[txnLockFast,txnLockFast] kv_process_ms:333 scan_total_write:0
```

```
    ↪ scan_processed_write:0 scan_total_data:0 scan_processed_data:0 scan_total_lock:0
    ↪  scan_processed_lock:0"]
```

- txnStartTS：发起读请求的事务的 start_ts，如上面示例中的 416643508703592451
- backoff_types：读写发生了冲突，并且读请求进行了 backoff 重试，重试的类型为 txnLockFast
- backoff_ms：读请求 backoff 重试的耗时，单位为 ms ，如上面示例中的 255
- region_id：读请求访问的目标 region 的 id

2. 通过 TiKV 日志分析：

   在 TiKV 的日志可以看到下列信息：

```
[ERROR] [endpoint.rs:454] [error-response] [err=""locked primary_lock:7480000000000004
    ↪ D35F6980000000000000010380000000004C788E0380000000004C0748 lock_version:
    ↪ 411402933858205712 key: 7480000000000004D35F7280000000004C0748 lock_ttl: 3008
    ↪ txn_size: 1""]
```

   这段报错信息表示出现了读写冲突，当读数据时发现 key 有锁阻碍读，这个锁包括未提交的乐观锁和未提交的 prewrite 后的悲观锁。

   - primary_lock：锁对应事务的 primary lock。
   - lock_version：锁对应事务的 start_ts。
   - key：表示被锁的 key。
   - lock_ttl: 锁的 TTL。
   - txn_size：锁所在事务在其 Region 的 key 数量，指导清锁方式。

处理建议：

- 在遇到读写冲突时会有 backoff 自动重试机制，如上述示例中 Txn1 会进行 backoff 重试，单次初始 100 ms，单次最大 3000 ms，总共最大 20000 ms
- 可以使用 TiDB Control 的子命令 decoder 来查看指定 key 对应的行的 table id 以及 rowid：

```
./tidb-ctl decoder -f table_row -k "t\x00\x00\x00\x00\x00\x00\x00\x1c_r\x00\x00\x00\x00\x00\
    ↪ x00\x00\xfa"

table_id: -9223372036854775780
row_id: -9223372036854775558
```

13.1.8.2.2 KeyIsLocked 错误

事务在 Prewrite 阶段的第一步就会检查是否有写写冲突，第二步会检查目标 key 是否已经被另一个事务上锁。当检测到该 key 被 lock 后，会在 TiKV 端报出 KeyIsLocked。目前该报错信息没有打印到 TiDB 以及 TiKV 的日志中。与读写冲突一样，在出现 KeyIsLocked 时，后台会自动进行 backoff 重试。

你可以通过 TiDB Grafana 监控检测 KeyIsLocked 错误：

观察 KV Errors 下 Lock Resolve OPS 面板中的 resolve 监控项以及 KV Backoff OPS 面板中的 txnLock 监控项，会有比较明显的上升趋势，其中 resolve 是指尝试清锁的操作，txnLock 代表出现了写冲突。



处理建议：

- 监控中出现少量 txnLock，无需过多关注。后台会自动进行 backoff 重试，单次初始 200 ms，单次最大 3000 ms。
- 如果出现大量的 txnLock，需要从业务的角度评估下冲突的原因。
- 使用悲观锁模式。

13.1.8.2.3 锁被清除 (LockNotFound) 错误

TxnLockNotFound 错误是由于事务提交的慢了，超过了 TTL 的时间。当要提交时，发现被其他事务给 Rollback 掉了。在开启 TiDB 自动重试事务的情况下，会自动在后台进行事务重试（注意显示和隐式事务的差别）。

你可以通过如下两种途径来查看 LockNotFound 报错信息：

1. 查看 TiDB 日志

如果出现了 TxnLockNotFound 的报错，会在 TiDB 的日志中看到下面的信息：

```
[WARN] [session.go:446] ["commit failed"] [conn=149370] ["finished txn"="Txn{state=invalid
    ↪ }"] [error="[kv:6]Error: KV error safe to retry tikv restarts txn: Txn(Mvcc(
    ↪ TxnLockNotFound{ start_ts: 412720515987275779, commit_ts: 412720519984971777, key:
    ↪ [116, 128, 0, 0, 0, 0, 1, 111, 16, 95, 114, 128, 0, 0, 0, 0, 0, 2] })) [try again
    ↪  later]"]
```

- start_ts：出现 TxnLockNotFound 报错的事务的 start_ts，如上例中的 412720515987275779
- commit_ts：出现 TxnLockNotFound 报错的事务的 commit_ts，如上例中的 412720519984971777

2. 查看 TiKV 日志

   如果出现了 TxnLockNotFound 的报错，在 TiKV 的日志中同样可以看到相应的报错信息：

```
Error: KV error safe to retry restarts txn: Txn(Mvcc(TxnLockNotFound)) [ERROR [Kv.rs:708] ["
    ↪ KvService::batch_raft send response fail"] [err=RemoteStoped]
```

处理建议：

- 通过检查 start_ts 和 commit_ts 之间的提交间隔，可以确认是否超过了默认的 TTL 的时间。

  查看提交间隔：

```
./pd-ctl tso [start_ts]
./pd-ctl tso [commit_ts]
```

- 建议检查下是否是因为写入性能的缓慢导致事务提交的效率差，进而出现了锁被清除的情况。
- 在关闭 TiDB 事务重试的情况下，需要在应用端捕获异常，并进行重试。

13.1.8.3 处理悲观锁冲突问题

以下介绍悲观事务模式下常见的锁冲突问题的处理方式。

> 注意：
>
> 即使设置了悲观事务模式，autocommit 事务仍然会优先尝试使用乐观事务模式进行提交，并在发生冲突后、自动重试时切换为悲观事务模式。

13.1.8.3.1 读写冲突

报错信息以及处理建议同乐观锁模式。

13.1.8.3.2 pessimistic lock retry limit reached

在冲突非常严重的场景下，或者当发生 write conflict 时，乐观事务会直接终止，而悲观事务会尝试用最新数据重试该语句直到没有 write conflict。因为 TiDB 的加锁操作是一个写入操作，且操作过程是先读后写，需要 2 次 RPC。如果在这中间发生了 write conflict，那么会重试。每次重试都会打印日志，不用特别关注。重试次数由pessimistic-txn.max-retry-count 定义。

可通过查看 TiDB 日志查看报错信息：

悲观事务模式下，如果发生 write conflict，并且重试的次数达到了上限，那么在 TiDB 的日志中会出现含有下述关键字的报错信息。如下：

```
err="pessimistic lock retry limit reached"
```

处理建议：

- 如果上述报错出现的比较频繁，建议从业务的角度进行调整。

13.1.8.3.3 Lock wait timeout exceeded

在悲观锁模式下，事务之间出现会等锁的情况。等锁的超时时间由 TiDB 的 innodb_lock_wait_timeout 参数来定义，这个是 SQL 语句层面的最大允许等锁时间，即一个 SQL 语句期望加锁，但锁一直获取不到，超过这个时间，TiDB 不会再尝试加锁，会向客户端返回相应的报错信息。

可通过查看 TiDB 日志查看报错信息：

当出现等锁超时的情况时，会向客户端返回下述报错信息：

```
ERROR 1205 (HY000): Lock wait timeout exceeded; try restarting transaction
```

处理建议：

- 如果出现的次数非常频繁，建议从业务逻辑的角度来进行调整。

13.1.8.3.4 TTL manager has timed out

除了有不能超出 GC 时间的限制外，悲观锁的 TTL 有上限，默认为 1 小时，所以执行时间超过 1 小时的悲观事务有可能提交失败。这个超时时间由 TiDB 参数 performance.max-txn-ttl 指定。

可通过查看 TiDB 日志查看报错信息：

当悲观锁的事务执行时间超过 TTL 时，会出现下述报错：

```
TTL manager has timed out, pessimistic locks may expire, please commit or rollback this
    ↪ transaction
```

处理建议：

- 当遇到该报错时，建议确认下业务逻辑是否可以进行优化，如将大事务拆分为小事务。在未使用大事务的前提下，大事务可能会触发 TiDB 的事务限制。
- 可适当调整相关参数，使其符合事务要求。

13.1.8.3.5 Deadlock found when trying to get lock

死锁是指两个或两个以上的事务在执行过程中，由于竞争资源而造成的一种阻塞的现象，若无外力作用，它们都将无法推进下去，将永远在互相等待。此时，需要终止其中一个事务使其能够继续推进下去。

TiDB 在使用悲观锁的情况下，多个事务之间出现了死锁，必定有一个事务 abort 来解开死锁。在客户端层面行为和 MySQL 一致，在客户端返回表示死锁的 Error 1213。如下：

```
[err="[executor:1213]Deadlock found when trying to get lock; try restarting transaction"]
```

处理建议：

- 如果难以确认产生死锁的原因，对于 v5.1 及以后的版本，建议尝试查询 `INFORMATION_SCHEMA.DEADLOCKS` 或 `INFORMATION_SCHEMA.CLUSTER_DEADLOCKS` 系统表来获取死锁的等待链信息。详情请参考死锁错误小节和 `DEADLOCKS` 表文档。
- 如果死锁出现非常频繁，需要调整业务代码来降低发生概率。

13.2 SQL 性能问题

13.2.1 慢查询日志

TiDB 会将执行时间超过slow-threshold（默认值为 300 毫秒）的语句输出到slow-query-file（默认值：“tidb-slow.log”）日志文件中，用于帮助用户定位慢查询语句，分析和解决 SQL 执行的性能问题。

TiDB 默认启用慢查询日志，可以修改配置enable-slow-log 来启用或禁用它。

13.2.1.1 日志示例

```
### Time: 2019-08-14T09:26:59.487776265+08:00
### Txn_start_ts: 410450924122144769
### User@Host: root[root] @ localhost [127.0.0.1]
### Conn_ID: 3086
### Exec_retry_time: 5.1 Exec_retry_count: 3
### Query_time: 1.527627037
### Parse_time: 0.000054933
### Compile_time: 0.000129729
### Rewrite_time: 0.000000003 Preproc_subqueries: 2 Preproc_subqueries_time: 0.000000002
### Process_time: 0.07 Request_count: 1 Total_keys: 131073 Process_keys: 131072 Prewrite_time:
    ↪ 0.335415029 Commit_time: 0.032175429 Get_commit_ts_time: 0.000177098
    ↪ Local_latch_wait_time: 0.106869448 Write_keys: 131072 Write_size: 3538944 Prewrite_region
    ↪ : 1
### DB: test
### Is_internal: false
### Digest: 50a2e32d2abbd6c1764b1b7f2058d428ef2712b029282b776beb9506a365c0f1
### Stats: t:pseudo
### Num_cop_tasks: 1
### Cop_proc_avg: 0.07 Cop_proc_p90: 0.07 Cop_proc_max: 0.07 Cop_proc_addr: 172.16.5.87:20171
```

```
### Cop_wait_avg: 0 Cop_wait_p90: 0 Cop_wait_max: 0 Cop_wait_addr: 172.16.5.87:20171
### Cop_backoff_regionMiss_total_times: 200 Cop_backoff_regionMiss_total_time: 0.2
    ↪ Cop_backoff_regionMiss_max_time: 0.2 Cop_backoff_regionMiss_max_addr: 127.0.0.1
    ↪ Cop_backoff_regionMiss_avg_time: 0.2 Cop_backoff_regionMiss_p90_time: 0.2
### Cop_backoff_rpcPD_total_times: 200 Cop_backoff_rpcPD_total_time: 0.2
    ↪ Cop_backoff_rpcPD_max_time: 0.2 Cop_backoff_rpcPD_max_addr: 127.0.0.1
    ↪ Cop_backoff_rpcPD_avg_time: 0.2 Cop_backoff_rpcPD_p90_time: 0.2
### Cop_backoff_rpcTiKV_total_times: 200 Cop_backoff_rpcTiKV_total_time: 0.2
    ↪ Cop_backoff_rpcTiKV_max_time: 0.2 Cop_backoff_rpcTiKV_max_addr: 127.0.0.1
    ↪ Cop_backoff_rpcTiKV_avg_time: 0.2 Cop_backoff_rpcTiKV_p90_time: 0.2
### Mem_max: 525211
### Disk_max: 65536
### Prepared: false
### Plan_from_cache: false
### Succ: true
### Plan: tidb_decode_plan('
    ↪ ZJAwCTMyXzcJMAkyMAlkYXRhOlRhYmxlU2Nhbl82CjEJMTBfNgkxAR0AdAEY1Dp0LCByYW5nZTpbLWluZiwraW5mXSwga2VlcCBvcmRlc
    ↪ ==')
use test;
insert into t select * from t;
```

13.2.1.2 字段含义说明

> **注意：**
>
> 慢查询日志中所有时间相关字段的单位都是“秒”

Slow Query 基础信息：

- `Time`：表示日志打印时间。
- `Query_time`：表示执行这个语句花费的时间。
- `Parse_time`：表示这个语句在语法解析阶段花费的时间。
- `Compile_time`：表示这个语句在查询优化阶段花费的时间。
- `Query`：表示 SQL 语句。慢日志里面不会打印 `Query`，但映射到内存表后，对应的字段叫 `Query`。
- `Digest`：表示 SQL 语句的指纹。
- `Txn_start_ts`：表示事务的开始时间戳，也是事务的唯一 ID，可以用这个值在 TiDB 日志中查找事务相关的其他日志。
- `Is_internal`：表示是否为 TiDB 内部的 SQL 语句。`true` 表示 TiDB 系统内部执行的 SQL 语句，`false` 表示用户执行的 SQL 语句。
- `Index_ids`：表示语句涉及到的索引的 ID。
- `Succ`：表示语句是否执行成功。
- `Backoff_time`：表示语句遇到需要重试的错误时在重试前等待的时间。常见的需要重试的错误有以下几种：遇到了 lock、Region 分裂、`tikv server is busy`。

- `Plan`：表示语句的执行计划，用 `select tidb_decode_plan('xxx...')` SQL 语句可以解析出具体的执行计划。
- `Prepared`：表示这个语句是否是 `Prepare` 或 `Execute` 的请求。
- `Plan_from_cache`：表示这个语句是否命中了执行计划缓存。
- `Rewrite_time`：表示这个语句在查询改写阶段花费的时间。
- `Preproc_subqueries`：表示这个语句中被提前执行的子查询个数，如 `where id in (select if from t)` 这个子查询就可能被提前执行。
- `Preproc_subqueries_time`：表示这个语句中被提前执行的子查询耗时。
- `Exec_retry_count`：表示这个语句执行的重试次数。一般出现在悲观事务中，上锁失败时重试执行该语句。
- `Exec_retry_time`：表示这个语句的重试执行时间。例如某个查询一共执行了三次（前两次失败），则 `Exec_retry_time` 表示前两次的执行时间之和，`Query_time` 减去 `Exec_retry_time` 则为最后一次执行时间。

和事务执行相关的字段：

- `Prewrite_time`：表示事务两阶段提交中第一阶段（prewrite 阶段）的耗时。
- `Commit_time`：表示事务两阶段提交中第二阶段（commit 阶段）的耗时。
- `Get_commit_ts_time`：表示事务两阶段提交中第二阶段（commit 阶段）获取 commit 时间戳的耗时。
- `Local_latch_wait_time`：表示事务两阶段提交中第二阶段（commit 阶段）发起前在 TiDB 侧等锁的耗时。
- `Write_keys`：表示该事务向 TiKV 的 Write CF 写入 Key 的数量。
- `Write_size`：表示事务提交时写 key 或 value 的总大小。
- `Prewrite_region`：表示事务两阶段提交中第一阶段（prewrite 阶段）涉及的 TiKV Region 数量。每个 Region 会触发一次远程过程调用。

和内存使用相关的字段：

- `Mem_max`：表示执行期间 TiDB 使用的最大内存空间，单位为 byte。

和硬盘使用相关的字段：

- `Disk_max`: 表示执行期间 TiDB 使用的最大硬盘空间，单位为 byte。

和 SQL 执行的用户相关的字段：

- `User`：表示执行语句的用户名。
- `Conn_ID`：表示用户的链接 ID，可以用类似 `con:3` 的关键字在 TiDB 日志中查找该链接相关的其他日志。
- `DB`：表示执行语句时使用的 database。

和 TiKV Coprocessor Task 相关的字段：

- `Request_count`：表示这个语句发送的 Coprocessor 请求的数量。
- `Total_keys`：表示 Coprocessor 扫过的 key 的数量。
- `Process_time`：执行 SQL 在 TiKV 的处理时间之和，因为数据会并行的发到 TiKV 执行，这个值可能会超过 `Query_time`。

- Wait_time：表示这个语句在 TiKV 的等待时间之和，因为 TiKV 的 Coprocessor 线程数是有限的，当所有的 Coprocessor 线程都在工作的时候，请求会排队；当队列中有某些请求耗时很长的时候，后面的请求的等待时间都会增加。
- Process_keys：表示 Coprocessor 处理的 key 的数量。相比 total_keys，processed_keys 不包含 MVCC 的旧版本。如果 processed_keys 和 total_keys 相差很大，说明旧版本比较多。
- Cop_proc_avg：cop-task 的平均执行时间。
- Cop_proc_p90：cop-task 的 P90 分位执行时间。
- Cop_proc_max：cop-task 的最大执行时间。
- Cop_proc_addr：执行时间最长的 cop-task 所在地址。
- Cop_wait_avg：cop-task 的平均等待时间。
- Cop_wait_p90：cop-task 的 P90 分位等待时间。
- Cop_wait_max：cop-task 的最大等待时间。
- Cop_wait_addr：等待时间最长的 cop-task 所在地址。
- Cop_backoff_{backoff-type}_total_times：因某种错误造成的 backoff 总次数。
- Cop_backoff_{backoff-type}_total_time：因某种错误造成的 backoff 总时间。
- Cop_backoff_{backoff-type}_max_time：因某种错误造成的最大 backoff 时间。
- Cop_backoff_{backoff-type}_max_addr：因某种错误造成的最大 backoff 时间的 cop-task 地址。
- Cop_backoff_{backoff-type}_avg_time：因某种错误造成的平均 backoff 时间。
- Cop_backoff_{backoff-type}_p90_time：因某种错误造成的 P90 分位 backoff 时间。

13.2.1.3 相关系统变量

- tidb_slow_log_threshold：设置慢日志的阈值，执行时间超过阈值的 SQL 语句将被记录到慢日志中。默认值是 300 ms。
- tidb_query_log_max_len：设置慢日志记录 SQL 语句的最大长度。默认值是 4096 byte。
- tidb_redact_log：设置慢日志记录 SQL 时是否将用户数据脱敏用 ? 代替。默认值是 0，即关闭该功能。
- tidb_enable_collect_execution_info：设置是否记录执行计划中各个算子的物理执行信息，默认值是 1。该功能对性能的影响约为 3%。开启该项后查看 Plan 的示例如下：

```
> select tidb_decode_plan('jAOIMAk1XzE3CTAJMQlmdW5jczpjb3VudChDb2x1bW4jNyktPkMJC/
    ↪ BMNQkxCXRpbWU6MTAuOTMxNTA1bXMsIGxvb3BzOjIJMzcyIEJ5dGVzCU4vQQoxCTMyXzE4CTAJMQlpbmRleDpTdHJlYW1BZ2dfOQkxCXQ
    ↪ ');
+--
    ↪ ----------------------------------------------------------------------------------------------------
    ↪
| tidb_decode_plan('jAOIMAk1XzE3CTAJMQlmdW5jczpjb3VudChDb2x1bW4jNyktPkMJC/
    ↪ BMNQkxCXRpbWU6MTAuOTMxNTA1bXMsIGxvb3BzOjIJMzcyIEJ5dGVzCU4vQQoxCTMyXzE4CTAJMQlpbmRleDpTdHJlYW1BZ2dfOQkxCXQ
    ↪  |
+----------------------------------------------------------------------------------------------------
    ↪
|     id                    task    estRows               operator info
    ↪                                                   actRows    execution info
    ↪                                                                memory       disk
    ↪                                |
```

```
|     StreamAgg_17          root     1                    funcs:count(Column#7)->Column#5
    ↪                                  1          time:10.931505ms, loops:2
    ↪                                                     372 Bytes     N/A
    ↪                                  |
|     └─IndexReader_18      root     1                    index:StreamAgg_9
    ↪                                                  1          time:10.927685ms, loops:2, rpc
    ↪ num: 1, rpc time:10.884355ms, proc keys:25007    206 Bytes     N/A
    ↪                                  |
|       └─StreamAgg_9       cop      1                    funcs:count(1)->Column#7
    ↪                                           1          time:11ms, loops:25
    ↪                                                               N/A           N/A
    ↪                                  |
|         └─IndexScan_16    cop      31281.857819905217   table:t, index:idx(a), range:[-inf
    ↪ ,50000), keep order:false    25007      time:11ms, loops:25
    ↪                                                               N/A           N/A
    ↪                                  |
+-------------------------------------------------------------------------------------------------------------------------------------------
    ↪
```

在性能测试中可以关闭自动收集算子的执行信息：

```
set @@tidb_enable_collect_execution_info=0;
```

Plan 字段显示的格式和 EXPLAIN 或者 EXPLAIN ANALYZE 大致一致。可以查看 EXPLAIN 或者 EXPLAIN ANALYZE 文档了解更多关于执行计划的信息。

更多详细信息，可以参见 TiDB 专用系统变量和语法。

13.2.1.4 慢日志内存映射表

用户可通过查询 INFORMATION_SCHEMA.SLOW_QUERY 表来查询慢查询日志中的内容，表中列名和慢日志中字段名一一对应，表结构可查看 SLOW_QUERY 表中的介绍。

> 注意：
>
> 每次查询 SLOW_QUERY 表时，TiDB 都会去读取和解析一次当前的慢查询日志。

TiDB 4.0 中，SLOW_QUERY 已经支持查询任意时间段的慢日志，即支持查询已经被 rotate 的慢日志文件的数据。用户查询时只需要指定 TIME 时间范围即可定位需要解析的慢日志文件。如果查询不指定时间范围，则仍然只解析当前的慢日志文件，示例如下：

不指定时间范围时，只会解析当前 TiDB 正在写入的慢日志文件的慢查询数据：

```
select count(*),
       min(time),
       max(time)
```

```
from slow_query;
```

```
+----------+----------------------------+----------------------------+
| count(*) | min(time)                  | max(time)                  |
+----------+----------------------------+----------------------------+
| 122492   | 2020-03-11 23:35:20.908574 | 2020-03-25 19:16:38.229035 |
+----------+----------------------------+----------------------------+
```

指定查询 2020-03-10 00:00:00 到 2020-03-11 00:00:00 时间范围后，会定位指定时间范围内的慢日志文件后解析慢查询数据：

```
select count(*),
       min(time),
       max(time)
from slow_query
where time > '2020-03-10 00:00:00'
  and time < '2020-03-11 00:00:00';
```

```
+----------+----------------------------+----------------------------+
| count(*) | min(time)                  | max(time)                  |
+----------+----------------------------+----------------------------+
| 2618049  | 2020-03-10 00:00:00.427138 | 2020-03-10 23:00:22.716728 |
+----------+----------------------------+----------------------------+
```

> 注意：
>
> 如果指定时间范围内的慢日志文件被删除，或者并没有慢查询，则查询结果会返回空。

TiDB 4.0 中新增了CLUSTER_SLOW_QUERY 系统表，用来查询所有 TiDB 节点的慢查询信息，表结构在 SLOW_QUERY 的基础上多增加了 INSTANCE 列，表示该行慢查询信息来自的 TiDB 节点地址。使用方式和 SLOW_QUERY 系统表一样。

关于查询 CLUSTER_SLOW_QUERY 表，TiDB 会把相关的计算和判断下推到其他节点执行，而不是把其他节点的慢查询数据都取回来在一台 TiDB 上执行。

13.2.1.5 查询 SLOW_QUERY/CLUSTER_SLOW_QUERY 示例

13.2.1.5.1 搜索 Top N 的慢查询

查询 Top 2 的用户慢查询。is_internal=false 表示排除 TiDB 内部的慢查询，只看用户的慢查询：

```
select query_time, query
from information_schema.slow_query
where is_internal = false  -- 排除 TiDB 内部的慢查询 SQL
```

```
order by query_time desc
limit 2;
```

输出样例：

```
+--------------+-----------------------------------------------------------------+
| query_time   | query                                                           |
+--------------+-----------------------------------------------------------------+
| 12.77583857  | select * from t_slim, t_wide where t_slim.c0=t_wide.c0;         |
|  0.734982725 | select t0.c0, t1.c1 from t_slim t0, t_wide t1 where t0.c0=t1.c0; |
+--------------+-----------------------------------------------------------------+
```

13.2.1.5.2 搜索某个用户的 Top N 慢查询

下面例子中搜索 test 用户执行的慢查询 SQL，且按执行消耗时间逆序排序显式前 2 条：

```
select query_time, query, user
from information_schema.slow_query
where is_internal = false  -- 排除 TiDB 内部的慢查询 SQL
  and user = "test"        -- 查找的用户名
order by query_time desc
limit 2;
```

输出样例：

```
+-------------+-----------------------------------------------------------------+----------------+
    ↪
| Query_time  | query                                                           | user
    ↪           |
+-------------+-----------------------------------------------------------------+----------------+
    ↪
| 0.676408014 | select t0.c0, t1.c1 from t_slim t0, t_wide t1 where t0.c0=t1.c1; | test
    ↪           |
+-------------+-----------------------------------------------------------------+----------------+
    ↪
```

13.2.1.5.3 根据 SQL 指纹搜索同类慢查询

在得到 Top N 的慢查询 SQL 后，可通过 SQL 指纹继续搜索同类慢查询 SQL。

先获取 Top N 的慢查询和对应的 SQL 指纹：

```
select query_time, query, digest
from information_schema.slow_query
where is_internal = false
order by query_time desc
limit 1;
```

输出样例：

```
+-------------+----------------------------+-------------------------------------------------------------------
    ↪
| query_time  | query                      | digest
    ↪                                                                 |
+-------------+----------------------------+-------------------------------------------------------------------
    ↪
| 0.302558006 | select * from t1 where a=1; | 4751
    ↪ cb6008fda383e22dacb601fde85425dc8f8cf669338d55d944bafb46a6fa |
+-------------+----------------------------+-------------------------------------------------------------------
    ↪
```

再根据 SQL 指纹搜索同类慢查询：

```
select query, query_time
from information_schema.slow_query
where digest = "4751cb6008fda383e22dacb601fde85425dc8f8cf669338d55d944bafb46a6fa";
```

输出样例：

```
+----------------------------+-------------+
| query                      | query_time  |
+----------------------------+-------------+
| select * from t1 where a=1; | 0.302558006 |
| select * from t1 where a=2; | 0.401313532 |
+----------------------------+-------------+
```

13.2.1.5.4 搜索统计信息为 pseudo 的慢查询 SQL 语句

```
select query, query_time, stats
from information_schema.slow_query
where is_internal = false
  and stats like '%pseudo%';
```

输出样例：

```
+----------------------------+-------------+---------------------------------+
| query                      | query_time  | stats                           |
+----------------------------+-------------+---------------------------------+
| select * from t1 where a=1; | 0.302558006 | t1:pseudo                       |
| select * from t1 where a=2; | 0.401313532 | t1:pseudo                       |
| select * from t1 where a>2; | 0.602011247 | t1:pseudo                       |
| select * from t1 where a>3; | 0.50077719  | t1:pseudo                       |
| select * from t1 join t2;   | 0.931260518 | t1:407872303825682445,t2:pseudo |
+----------------------------+-------------+---------------------------------+
```

13.2.1.5.5 查询执行计划发生变化的慢查询

由于统计信息过时，或者统计信息因为误差无法精确反映数据的真实分布情况时，可能导致同类型 SQL 的执行计划发生改变导致执行变慢，可以用以下 SQL 查询哪些 SQL 具有不同的执行计划：

```
select count(distinct plan_digest) as count,
       digest,
       min(query)
from cluster_slow_query
group by digest
having count > 1
limit 3\G
```

输出样例：

```
***************************[ 1. row ]***************************
count      | 2
digest     | 17b4518fde82e32021877878bec2bb309619d384fca944106fcaf9c93b536e94
min(query) | SELECT DISTINCT c FROM sbtest25 WHERE id BETWEEN ? AND ? ORDER BY c [arguments:
    ↪ (291638, 291737)];
***************************[ 2. row ]***************************
count      | 2
digest     | 9337865f3e2ee71c1c2e740e773b6dd85f23ad00f8fa1f11a795e62e15fc9b23
min(query) | SELECT DISTINCT c FROM sbtest22 WHERE id BETWEEN ? AND ? ORDER BY c [arguments:
    ↪ (215420, 215519)];
***************************[ 3. row ]***************************
count      | 2
digest     | db705c89ca2dfc1d39d10e0f30f285cbbadec7e24da4f15af461b148d8ffb020
min(query) | SELECT DISTINCT c FROM sbtest11 WHERE id BETWEEN ? AND ? ORDER BY c [arguments:
    ↪ (303359, 303458)];
```

然后可以用查询结果中的 SQL 指纹进一步查询不同的 plan

```
select min(plan),
       plan_digest
from cluster_slow_query
where digest='17b4518fde82e32021877878bec2bb309619d384fca944106fcaf9c93b536e94'
group by plan_digest\G
```

输出样例：

```
*************************** 1. row ***************************
  min(plan):    Sort_6                  root    100.00131380758702      sbtest.sbtest25.c:asc
        └─HashAgg_10            root    100.00131380758702      group by:sbtest.sbtest25.c,
            ↪ funcs:firstrow(sbtest.sbtest25.c)->sbtest.sbtest25.c
          └─TableReader_15      root    100.00131380758702      data:TableRangeScan_14
            └─TableScan_14      cop     100.00131380758702      table:sbtest25, range
                ↪ :[502791,502890], keep order:false
```

```
plan_digest: 6afbbd21f60ca6c6fdf3d3cd94f7c7a49dd93c00fcf8774646da492e50e204ee
*************************** 2. row ***************************
  min(plan):    Sort_6                 root    1                       sbtest.sbtest25.c:asc
        └─HashAgg_12            root    1                       group by:sbtest.sbtest25.c,
          ↪ funcs:firstrow(sbtest.sbtest25.c)->sbtest.sbtest25.c
          └─TableReader_13      root    1                       data:HashAgg_8
            └─HashAgg_8         cop     1                       group by:sbtest.sbtest25.c,
              └─TableScan_11    cop     1.2440069558121831      table:sbtest25, range
                ↪ :[472745,472844], keep order:false
```

13.2.1.5.6 查询集群各个 TIDB 节点的慢查询数量

```
select instance, count(*) from information_schema.cluster_slow_query where time >= "2020-03-06
    ↪ 00:00:00" and time < now() group by instance;
```

输出样例：

```
+---------------+----------+
| instance      | count(*) |
+---------------+----------+
| 0.0.0.0:10081 | 124      |
| 0.0.0.0:10080 | 119771   |
+---------------+----------+
```

13.2.1.5.7 查询仅出现在异常时间段的慢日志

假如发现 2020-03-10 13:24:00 ~ 2020-03-10 13:27:00 的 QPS 降低或者延迟上升等问题，可能是由于突然出现大查询导致的，可以用下面 SQL 查询仅出现在异常时间段的慢日志，其中 2020-03-10 13:20:00 ~ 2020-03-10 13:23:00 为正常时间段。

```
SELECT * FROM
    (SELECT /*+ AGG_TO_COP(), HASH_AGG() */ count(*),
         min(time),
         sum(query_time) AS sum_query_time,
         sum(Process_time) AS sum_process_time,
         sum(Wait_time) AS sum_wait_time,
         sum(Commit_time),
         sum(Request_count),
         sum(process_keys),
         sum(Write_keys),
         max(Cop_proc_max),
         min(query),min(prev_stmt),
         digest
    FROM information_schema.CLUSTER_SLOW_QUERY
    WHERE time >= '2020-03-10 13:24:00'
```

```
            AND time < '2020-03-10 13:27:00'
            AND Is_internal = false
    GROUP BY  digest) AS t1
WHERE t1.digest NOT IN
    (SELECT /*+ AGG_TO_COP(), HASH_AGG() */ digest
    FROM information_schema.CLUSTER_SLOW_QUERY
    WHERE time >= '2020-03-10 13:20:00'
            AND time < '2020-03-10 13:23:00'
    GROUP BY  digest)
ORDER BY  t1.sum_query_time DESC limit 10\G
```

输出样例：

```
***************************[ 1. row ]***************************
count(*)            | 200
min(time)           | 2020-03-10 13:24:27.216186
sum_query_time      | 50.114126194
sum_process_time    | 268.351
sum_wait_time       | 8.476
sum(Commit_time)    | 1.044304306
sum(Request_count) | 6077
sum(process_keys)   | 202871950
sum(Write_keys)     | 319500
max(Cop_proc_max)   | 0.263
min(query)          | delete from test.tcs2 limit 5000;
min(prev_stmt)      |
digest              | 24bd6d8a9b238086c9b8c3d240ad4ef32f79ce94cf5a468c0b8fe1eb5f8d03df
```

13.2.1.6 解析其他的 TiDB 慢日志文件

TiDB 通过 session 变量 `tidb_slow_query_file` 控制查询 `INFORMATION_SCHEMA.SLOW_QUERY` 时要读取和解析的文件，可通过修改改 session 变量的值来查询其他慢查询日志文件的内容：

```
set tidb_slow_query_file = "/path-to-log/tidb-slow.log"
```

13.2.1.7 用 pt-query-digest 工具分析 TiDB 慢日志

可以用 `pt-query-digest` 工具分析 TiDB 慢日志。

> **注意：**
>
> 建议使用 pt-query-digest 3.0.13 及以上版本。

示例如下：

```
pt-query-digest --report tidb-slow.log
```

输出样例：

```
### 320ms user time, 20ms system time, 27.00M rss, 221.32M vsz
### Current date: Mon Mar 18 13:18:51 2019
### Hostname: localhost.localdomain
### Files: tidb-slow.log
### Overall: 1.02k total, 21 unique, 0 QPS, 0x concurrency _____________________
### Time range: 2019-03-18-12:22:16 to 2019-03-18-13:08:52
### Attribute          total     min     max     avg     95%  stddev  median
### ============     ======= ======= ======= ======= ======= ======= =======
### Exec time            218s    10ms     13s   213ms    30ms       1s    19ms
### Query size       175.37k       9   2.01k  175.89  158.58  122.36  158.58
### Commit time          46ms     2ms     7ms     3ms     7ms     1ms     3ms
### Conn ID                71       1      16    8.88   15.25    4.06    9.83
### Process keys     581.87k       2 103.15k  596.43  400.73   3.91k  400.73
### Process time          31s     1ms     10s    32ms    19ms   334ms    16ms
### Request coun        1.97k       1      10    2.02    1.96    0.33    1.96
### Total keys       636.43k       2 103.16k  652.35  793.42   3.97k  400.73
### Txn start ts     374.38E       0  16.00E 375.48P   1.25P  89.05T   1.25P
### Wait time            943ms     1ms    19ms     1ms     2ms     1ms   972us
.
.
.
```

13.2.1.7.1 定位问题语句的方法

并不是所有 SLOW_QUERY 的语句都是有问题的。会造成集群整体压力增大的，是那些 process_time 很大的语句。wait_time 很大，但 process_time 很小的语句通常不是问题语句，是因为被问题语句阻塞，在执行队列等待造成的响应时间过长。

13.2.1.8 admin show slow 命令

除了获取 TiDB 日志，还有一种定位慢查询的方式是通过 admin show slow SQL 命令：

```
admin show slow recent N;
```

```
admin show slow top [internal | all] N;
```

recent N 会显示最近的 N 条慢查询记录，例如：

```
admin show slow recent 10;
```

top N 则显示最近一段时间（大约几天）内，最慢的查询记录。如果指定 internal 选项，则返回查询系统内部 SQL 的慢查询记录；如果指定 all 选项，返回系统内部和用户 SQL 汇总以后的慢查询记录；默认只返回用户 SQL 中的慢查询记录。

```
admin show slow top 3;
admin show slow top internal 3;
admin show slow top all 5;
```

由于内存限制，保留的慢查询记录的条数是有限的。当命令查询的 N 大于记录条数时，返回的结果记录条数会小于 N。

输出内容详细说明，如下：

列名	描述
start	SQL 语句执行开始时间
duration	SQL 语句执行持续时间
details	执行语句的详细信息
succ	SQL 语句执行是否成功，1: 成功，0: 失败
conn_id	session 连接 ID
transcation_ts	事务提交的 commit ts
user	执行该语句的用户名
db	执行该 SQL 涉及到 database
table_ids	执行该 SQL 涉及到表的 ID
index_ids	执行该 SQL 涉及到索引 ID
internal	表示为 TiDB 内部的 SQL 语句
digest	表示 SQL 语句的指纹
sql	执行的 SQL 语句

13.2.2 分析慢查询

处理慢查询分为两步：

1. 从大量查询中定位出哪一类查询比较慢
2. 分析这类慢查询的原因

第一步可以通过慢日志、statement-summary 方便地定位，推荐直接使用TiDB Dashboard，它整合了这两个功能，且能方便直观地在浏览器中展示出来。本文聚焦第二步。

首先将慢查询归因成两大类：

- 优化器问题：如选错索引，选错 Join 类型或顺序。
- 系统性问题：将非优化器问题都归结于此类。如：某个 TiKV 实例忙导致处理请求慢，Region 信息过期导致查询变慢。

实际中，优化器问题可能造成系统性问题。例如对于某类查询，优化器应使用索引，但却使用了全表扫。这可能导致这类 SQL 消耗大量资源，造成某些 KV 实例 CPU 飚高等问题。表现上看就是一个系统性问题，但本质是优化器问题。

分析优化器问题需要有判断执行计划是否合理的能力，而系统性问题的定位相对简单，因此面对慢查询推荐的分析过程如下：

1. 定位查询瓶颈：即查询过程中耗时多的部分
2. 分析系统性问题：根据瓶颈点，结合当时的监控/日志等信息，分析可能的原因
3. 分析优化器问题：分析是否有更好的执行计划

接下来会分别介绍上面几点。

13.2.2.1 定位查询瓶颈

定位查询瓶颈需要对查询过程有一个大致理解，TiDB 处理查询过程的关键阶段都在 performance-map 图中了。

查询的耗时信息可以从下面几种方式获得：

- 慢日志（推荐直接在TiDB Dashboard 中查看）
- `explain analyze` 语句

他们的侧重点不同：

- 慢日志记录了 SQL 从解析到返回，几乎所有阶段的耗时，较为全面（在 TiDB Dashboard 中可以直观地查询和分析慢日志）；
- `explain analyze` 可以拿到 SQL 实际执行中每个执行算子的耗时，对执行耗时有更细分的统计；

总的来说，利用慢日志和 `explain analyze` 可以比较准确地定位查询的瓶颈点，帮助你判断这条 SQL 慢在哪个模块 (TiDB/TiKV)，慢在哪个阶段，下面会有一些例子。

另外在 4.0.3 之后，慢日志中的 `Plan` 字段也会包含 SQL 的执行信息，也就是 `explain analyze` 的结果，这样一来 SQL 的所有耗时信息都可以在慢日志中找到。

13.2.2.2 分析系统性问题

对于系统性问题，我们根据执行阶段，分成三个大类：

1. TiKV 处理慢：如 coprocessor 处理数据慢
2. TiDB 执行慢：主要指执行阶段，如某个 Join 算子处理数据慢
3. 其他关键阶段慢：如取时间戳慢

拿到一个慢查询，我们应该先根据已有信息判断大致是哪个大类，再具体分析。

13.2.2.2.1 TiKV 处理慢

如果是 TiKV 处理慢，可以很明显的通过 `explain analyze` 中看出来。例如下面这个例子，可以看到 `StreamAgg_8` 和 `TableFullScan_15` 这两个 `tikv-task`（在 `task` 列可以看出这两个任务类型是 `cop[tikv]`）花费了 `170ms`，而 TiDB 部分的算子耗时，减去这 `170ms` 后，耗时占比非常小，说明瓶颈在 TiKV。

```
+--
    ↪ -------------------------+---------+---------+-----------+---------------+--------------------------------
    ↪
```

```
| id                          | estRows | actRows | task      | access object | execution info
    ↪                                                                | operator info
    ↪                    | memory    | disk |
+--
    ↪ -------------------------+---------+---------+-----------+---------------+-------------------------------
    ↪
| StreamAgg_16                | 1.00    | 1       | root      |               | time:170.08572ms,
    ↪ loops:2                                                      | funcs:count(Column#5)->
    ↪ Column#3 | 372 Bytes | N/A   |
| └─TableReader_17            | 1.00    | 1       | root      |               | time:170.080369ms
    ↪ , loops:2, rpc num: 1, rpc time:17.023347ms, proc keys:28672 | data:StreamAgg_8
    ↪                 | 202 Bytes | N/A   |
|   └─StreamAgg_8             | 1.00    | 1       | cop[tikv] |               | time:170ms, loops
    ↪ :29                                                          | funcs:count(1)->Column#5
    ↪         | N/A       | N/A   |
|     └─TableFullScan_15      | 7.00    | 28672   | cop[tikv] | table:t       | time:170ms, loops
    ↪ :29                                                          | keep order:false, stats:
    ↪ pseudo  | N/A       | N/A   |
+--
    ↪ -------------------------+---------+---------+-----------+---------------+-------------------------------
    ↪
```

另外在慢日志中，Cop_process 和 Cop_wait 字段也可以帮助判断，如下面这个例子，查询整个耗时是 180.85ms 左右，而最大的那个 coptask 就消耗了 171ms，可以说明对这个查询而言，瓶颈在 TiKV 侧。

慢日志中的各个字段的说明可以参考慢查询日志中的字段含义说明

```
### Query_time: 0.18085
...
### Num_cop_tasks: 1
### Cop_process: Avg_time: 170ms P90_time: 170ms Max_time: 170ms Max_addr: 10.6.131.78
### Cop_wait: Avg_time: 1ms P90_time: 1ms Max_time: 1ms Max_Addr: 10.6.131.78
```

根据上述方式判断是 TiKV 慢后，可以依次排查 TiKV 慢的原因。

TiKV 实例忙

一条 SQL 可能会去从多个 TiKV 上拿数据，如果某个 TiKV 响应慢，可能拖慢整个 SQL 的处理速度。

慢日志中的 Cop_wait 可以帮忙判断这个问题：

```
### Cop_wait: Avg_time: 1ms P90_time: 2ms Max_time: 110ms Max_Addr: 10.6.131.78
```

如上图，发给 10.6.131.78 的一个 cop-task 等待了 110ms 才被执行，可以判断是当时该实例忙，此时可以打开当时的 CPU 监控辅助判断。

过期 key 多

如果 TiKV 上过期的数据比较多，在扫描的时候则需要处理这些不必要的数据，影响处理速度。

这可以通过 Total_keys 和 Processed_keys 判断，如果两者相差较大，则说明旧版本的 key 太多：

```
...
### Total_keys: 2215187529 Processed_keys: 1108056368
...
```

13.2.2.2.2 其他关键阶段慢

取 TS 慢

可以对比慢日志中的 Wait_TS 和 Query_time，因为 TS 有预取操作，通常来说 Wait_TS 应该很低。

```
### Query_time: 0.0300000
...
### Wait_TS: 0.02500000
```

Region 信息过期

TiDB 侧 Region 信息可能过期，此时 TiKV 可能返回 regionMiss 的错误，然后 TiDB 会从 PD 去重新获取 Region 信息，这些信息会被反应在 Cop_backoff 信息内，失败的次数和总耗时都会被记录下来。

```
### Cop_backoff_regionMiss_total_times: 200 Cop_backoff_regionMiss_total_time: 0.2
    ↪ Cop_backoff_regionMiss_max_time: 0.2 Cop_backoff_regionMiss_max_addr: 127.0.0.1
    ↪ Cop_backoff_regionMiss_avg_time: 0.2 Cop_backoff_regionMiss_p90_time: 0.2
### Cop_backoff_rpcPD_total_times: 200 Cop_backoff_rpcPD_total_time: 0.2
    ↪ Cop_backoff_rpcPD_max_time: 0.2 Cop_backoff_rpcPD_max_addr: 127.0.0.1
    ↪ Cop_backoff_rpcPD_avg_time: 0.2 Cop_backoff_rpcPD_p90_time: 0.2
```

子查询被提前执行

对于带有非关联子查询的语句，子查询部分可能被提前执行，如：select * from t1 where a = (select max(a)from t2)，select max(a)from t2 部分可能在优化阶段被提前执行。这种查询用 explain analyze 看不到对应的耗时，如下：

```
mysql> explain analyze select count(*) from t where a=(select max(t1.a) from t t1, t t2 where t1.
    ↪ a=t2.a);
+--
    ↪ ----------------------------+----------+---------+-----------+---------------+-------------------------------+
    ↪
| id                            | estRows  | actRows | task      | access object | execution info
    ↪           | operator info                     | memory    | disk |
+--
    ↪ ----------------------------+----------+---------+-----------+---------------+-------------------------------+
    ↪
| StreamAgg_59                  | 1.00     | 1       | root      |               | time:4.69267ms,
    ↪  loops:2  | funcs:count(Column#10)->Column#8 | 372 Bytes | N/A  |
| └─TableReader_60              | 1.00     | 1       | root      |               | time:4.690428
    ↪ ms, loops:2 | data:StreamAgg_48                | 141 Bytes | N/A  |
|   └─StreamAgg_48              | 1.00     |         | cop[tikv] |               | time:0ns,
    ↪ loops:0        | funcs:count(1)->Column#10        | N/A       | N/A  |
```

```
|     └─Selection_58           | 16384.00 |         | cop[tikv] |               | time:0ns,
    ↪ loops:0        | eq(test.t.a, 1)                 | N/A       | N/A  |
|       └─TableFullScan_57     | 16384.00 | -1      | cop[tikv] | table:t       | time:0s, loops
    ↪ :0         | keep order:false                | N/A       | N/A  |
+--
    ↪ ---------------------------+----------+---------+-----------+---------------+-------------------------+
    ↪
5 rows in set (7.77 sec)
```

不过可以从慢日志中排查这种情况：

```
### Query_time: 7.770634843
...
### Rewrite_time: 7.765673663 Preproc_subqueries: 1 Preproc_subqueries_time: 7.765231874
```

可以看到有 1 个子查询被提前执行，花费了 7.76s。

13.2.2.2.3 TiDB 执行慢

这里我们假设 TiDB 的执行计划正确（不正确的情况在分析优化器问题这一节中说明），但是执行上很慢；

解决这类问题主要靠调整参数或利用 hint，并结合 explain analyze 对 SQL 进行调整。

并发太低

如果发现瓶颈在有并发的算子上，可以通过调整并发度来尝试提速，如下面的执行计划中：

```
mysql> explain analyze select sum(t1.a) from t t1, t t2 where t1.a=t2.a;
+--
    ↪ -------------------------------+--------------+-----------+-----------+---------------+-----------------
    ↪
| id                              | estRows      | actRows   | task      | access object |
    ↪ execution info                                                                      |
    ↪ operator info                                 | memory           | disk    |
+--
    ↪ -------------------------------+--------------+-----------+-----------+---------------+-----------------
    ↪
| HashAgg_11                      | 1.00         | 1         | root      |               | time
    ↪ :9.666832189s, loops:2, PartialConcurrency:4, FinalConcurrency:4               | funcs:
    ↪ sum(Column#6)->Column#5                  | 322.125 KB       | N/A     |
| └─Projection_24                 | 268435456.00 | 268435456 | root      |               | time
    ↪ :9.098644711s, loops:262145, Concurrency:4                                     | cast(
    ↪ test.t.a, decimal(65,0) BINARY)->Column#6 | 199 KB           | N/A     |
|   └─HashJoin_14                 | 268435456.00 | 268435456 | root      |               | time
    ↪ :6.616773501s, loops:262145, Concurrency:5, probe collision:0, build:881.404µs  | inner
    ↪ join, equal:[eq(test.t.a, test.t.a)]     | 131.75 KB        | 0 Bytes |
|     ├─TableReader_21(Build)     | 16384.00     | 16384     | root      |               | time
    ↪ :6.553717ms, loops:17                                                          | data:
    ↪ Selection_20                             | 33.6318359375 KB | N/A     |
```

```
|     | └─Selection_20             | 16384.00     |           | cop[tikv] |               | time
    ↪ :0ns, loops:0                                                                  | not(
    ↪ isnull(test.t.a))                          | N/A              | N/A      |
|     |   └─TableFullScan_19       | 16384.00     | -1        | cop[tikv] | table:t2      | time
    ↪ :0s, loops:0                                                                   | keep
    ↪ order:false                                | N/A              | N/A      |
|     └─TableReader_18(Probe)      | 16384.00     | 16384     | root      |               | time
    ↪ :6.880923ms, loops:17                                                          | data:
    ↪ Selection_17                               | 33.6318359375 KB | N/A      |
|       └─Selection_17             | 16384.00     |           | cop[tikv] |               | time
    ↪ :0ns, loops:0                                                                  | not(
    ↪ isnull(test.t.a))                          | N/A              | N/A      |
|         └─TableFullScan_16       | 16384.00     | -1        | cop[tikv] | table:t1      | time
    ↪ :0s, loops:0                                                                   | keep
    ↪ order:false                                | N/A              | N/A      |
+--
    ↪ ---------------------------------+--------------+-----------+-----------+---------------+----------------
    ↪
9 rows in set (9.67 sec)
```

发现耗时主要在 HashJoin_14 和 Projection_24，可以酌情通过 SQL 变量来提高他们的并发度进行提速。

system-variables 中有所有的系统变量，如想提高 HashJoin_14 的并发度，则可以修改变量 tidb_hash_join_concurrency
↪ 。

产生了落盘

执行慢的另一个原因是执行过程中，因为到达内存限制，产生了落盘，这点在执行计划和慢日志中都能看到：

```
+--
    ↪ ----------------------+----------+---------+----------+--------------+------------------------------+
    ↪
| id                      | estRows    | actRows | task      | access object | execution info
    ↪               | operator info        | memory                | disk           |
+--
    ↪ ----------------------+----------+---------+----------+--------------+------------------------------+
    ↪
| Sort_4                  | 462144.00 | 462144  | root      |               | time:2.02848898s,
    ↪ loops:453  | test.t.a             | 149.68795776367188 MB | 219.3203125 MB |
| └─TableReader_8         | 462144.00 | 462144  | root      |               | time:616.211272ms,
    ↪  loops:453 | data:TableFullScan_7 | 197.49601364135742 MB | N/A            |
|   └─TableFullScan_7     | 462144.00 | -1      | cop[tikv] | table:t       | time:0s, loops:0
    ↪              | keep order:false     | N/A                   | N/A            |
+--
    ↪ ----------------------+----------+---------+----------+--------------+------------------------------+
    ↪
```

```
...
### Disk_max: 229974016
...
```

做了笛卡尔积的 Join

做笛卡尔积的 Join 会产生 左边孩子行数 * 右边孩子行数 这么多数据，效率较低，应该尽量避免；

目前对于产生笛卡尔积的 Join 会在执行计划中显示的标明 CARTESIAN，如下：

```
mysql> explain select * from t t1, t t2 where t1.a>t2.a;
+--
    ↪ ---------------------------+-------------+-----------+---------------+--------------------------------
    ↪
| id                           | estRows     | task      | access object | operator info
    ↪                                        |
+--
    ↪ ---------------------------+-------------+-----------+---------------+--------------------------------
    ↪
| HashJoin_8                   | 99800100.00 | root      |               | CARTESIAN inner join,
    ↪ other cond:gt(test.t.a, test.t.a) |
| ├─TableReader_15(Build)      | 9990.00     | root      |               | data:Selection_14
    ↪                                       |
| │ └─Selection_14             | 9990.00     | cop[tikv] |               | not(isnull(test.t.a))
    ↪                                    |
| │   └─TableFullScan_13       | 10000.00    | cop[tikv] | table:t2      | keep order:false,
    ↪ stats:pseudo                         |
| └─TableReader_12(Probe)      | 9990.00     | root      |               | data:Selection_11
    ↪                                       |
|   └─Selection_11             | 9990.00     | cop[tikv] |               | not(isnull(test.t.a))
    ↪                                    |
|     └─TableFullScan_10       | 10000.00    | cop[tikv] | table:t1      | keep order:false,
    ↪ stats:pseudo                         |
+--
    ↪ ---------------------------+-------------+-----------+---------------+--------------------------------
    ↪
```

13.2.2.3 分析优化器问题

分析优化问题需要有判断执行计划是否合理的能力，这需要对优化过程和各算子有一定了解。

下面是一组例子，假设表结构为 `create table t (id int, a int, b int, c int, primary key(id), key(a)
↪ , key(b, c))`：

1. `select * from t`: 没有过滤条件，会扫全表，所以会用 TableFullScan 算子读取数据；
2. `select a from t where a=2`：有过滤条件且只读索引列，所以会用 IndexReader 算子读取数据；

3. `select * from t where a=2`：在 a 有过滤条件，但索引 a 不能完全覆盖需要读取的内容，因此会采用 IndexLookup；
4. `select b from t where c=3`：多列索引没有前缀条件就用不上，所以会用 IndexFullScan；
5. …

上面举例了数据读入相关的算子，在理解 TiDB 执行计划中描述了更多算子的情况。

另外阅读SQL 性能调优整个小节能增加你对 TiDB 优化器的了解，帮助判断执行计划是否合理。

由于大多数优化器问题在SQL 性能调优已经有解释，这里就直接列举出来跳转过去：

1. 索引选择错误
2. Join 顺序错误
3. 表达式未下推

13.2.3 SQL 诊断

> 警告：
>
> 该功能目前为实验特性，不建议在生产环境中使用。

SQL 诊断功能是在 TiDB 4.0 版本中引入的特性，用于提升 TiDB 问题定位的效率。TiDB 4.0 版本以前，用户需要使用不同的工具以异构的方式获取不同信息。新的 SQL 诊断系统对这些离散的信息进行了整体设计，它整合系统各个维度的信息，通过系统表的方式向上层提供一致的接口，提供监控汇总与自动诊断，方便用户查询集群信息。

SQL 诊断共分三大块：

- 集群信息表：TiDB 4.0 诊断系统添加了集群信息表，为原先离散的各实例信息提供了统一的获取方式。它将整个集群的集群拓扑、硬件信息、软件信息、内核参数、监控、系统信息、慢查询、语句、日志完全整合在表中，让用户能够统一使用 SQL 进行查询。
- 集群监控表：TiDB 4.0 诊断系统添加了集群监控系统表，所有表都在 `metrics_schema` 中，可以通过 SQL 语句来查询监控信息。比起原先的可视化监控，SQL 查询监控允许用户对整个集群的所有监控进行关联查询，并对比不同时间段的结果，迅速找出性能瓶颈。由于 TiDB 集群的监控指标数量较大，SQL 诊断还提供了监控汇总表，让用户能够更便捷地从众多监控中找出异常的监控项。
- 自动诊断：尽管用户可以手动执行 SQL 来查询集群信息表、集群监控表与汇总表来定位问题，但自动诊断可以快速对常见异常进行定位。SQL 诊断基于已有的集群信息表和监控表，提供了与之相关的诊断结果表与诊断汇总表来执行自动诊断。

13.2.3.1 集群信息表

集群信息表将一个集群中的所有实例的信息都汇聚在一起，让用户仅通过一条 SQL 就能查询整个集群相关信息。集群信息表列表如下：

- 集群拓扑表 information_schema.cluster_info 用于获取集群当前的拓扑信息，以及各个实例的版本、版本对应的 Git Hash、各实例的启动时间、各实例的运行时间。
- 集群配置表 information_schema.cluster_config 用于获取集群当前所有实例的配置。对于 TiDB 4.0 之前的版本，用户必须逐个访问各个实例的 HTTP API 才能获取这些配置信息。
- 集群硬件表 information_schema.cluster_hardware 用于快速查询集群硬件信息。
- 集群负载表 information_schema.cluster_load 用于查询集群不同实例以及不同硬件类型的负载信息。
- 内核参数表 information_schema.cluster_systeminfo 用于查询集群不同实例的内核配置信息。目前支持查询 sysctl 的信息。
- 集群日志表 information_schema.cluster_log 用于集群日志查询，通过将查询条件下推到各个实例，降低日志查询对集群的影响，性能影响小于等 grep 命令。

TiDB 4.0 之前的系统表，只能查看当前实例信息，TiDB 4.0 实现了对应的集群表，可以在单个 TiDB 实例上拥有整个集群的全局视图。这些表目前都位于 information_schema 中，查询方式与其他 information_schema 系统表一致。

13.2.3.2 集群监控表

为了能够动态地观察并对比不同时间段的集群情况，TiDB 4.0 诊断系统添加了集群监控系统表。所有监控表都在 metrics_schema 中，可以通过 SQL 的方式查询监控信息。SQL 查询监控允许用户对整个集群的所有监控进行关联查询，并对比不同时间段的结果，迅速找出性能瓶颈。

- information_schema.metrics_tables：由于目前添加的系统表数量较多，因此用户可以通过该表查询这些监控表的相关元信息。

由于 TiDB 集群的监控指标数量较大，因此 TiDB 4.0 提供以下监控汇总表：

- 监控汇总表 information_schema.metrics_summary 用于汇总所有监控数据，以提升用户排查各监控指标的效率。
- 监控汇总并按 label 聚合表 information_schema.metrics_summary_by_label 同样用于汇总所有监控数据，但该表会对各项监控的不同的 label 进行聚合统计。

13.2.3.3 自动诊断

以上集群信息表和集群监控表均需要用户手动执行 SQL 语句来排查集群问题。TiDB 4.0 中添加了自动诊断功能，根据已有的基础信息表，提供诊断相关的系统表，使诊断自动执行。自动诊断相关的系统表如下：

- 诊断结果表 information_schema.inspection_result 用于展示对系统的诊断结果。诊断是惰性触发，使用 select * from inspection_result 会触发所有诊断规则对系统进行诊断，并在结果中展示系统中的故障或风险。
- 诊断汇总表 information_schema.inspection_summary 用于对特定链路或模块的监控进行汇总，用户可以根据整个模块或链路的上下文来排查定位问题。

13.2.4 定位消耗系统资源多的查询

TiDB 会将执行时间超过tidb_expensive_query_time_threshold 限制（默认值为 60s），或使用内存超过mem-
↪ quota-query 限制（默认值为 1 GB）的语句输出到tidb-server 日志文件（默认文件为 “tidb.log”）中，用于

在语句执行结束前定位消耗系统资源多的查询语句（以下简称为 expensive query），帮助用户分析和解决语句执行的性能问题。

注意，expensive query 日志和慢查询日志的区别是，慢查询日志是在语句执行完后才打印，expensive query 日志可以将正在执行的语句的相关信息打印出来。当一条语句在执行过程中达到资源使用阈值时（执行时间/使用内存量），TiDB 会即时将这条语句的相关信息写入日志。

13.2.4.1 Expensive query 日志示例

```
[2020/02/05 15:32:25.096 +08:00] [WARN] [expensivequery.go:167] [expensive_query] [cost_time
    ↪ =60.008338935s] [wait_time=0s] [request_count=1] [total_keys=70] [process_keys=65] [
    ↪ num_cop_tasks=1] [process_avg_time=0s] [process_p90_time=0s] [process_max_time=0s] [
    ↪ process_max_addr=10.0.1.9:20160] [wait_avg_time=0.002s] [wait_p90_time=0.002s] [
    ↪ wait_max_time=0.002s] [wait_max_addr=10.0.1.9:20160] [stats=t:pseudo] [conn_id=60026] [
    ↪ user=root] [database=test] [table_ids="[122]"] [txn_start_ts=414420273735139329] [mem_max
    ↪ ="1035 Bytes (1.0107421875 KB)"] [sql="insert into t select sleep(1) from t"]
```

13.2.4.2 字段含义说明

基本字段：

- `cost_time`：日志打印时语句已经花费的执行时间。
- `stats`：语句涉及到的表或索引使用的统计信息版本。值为 pesudo 时表示无可用统计信息，需要对表或索引进行 analyze。
- `table_ids`：语句涉及到的表的 ID。
- `txn_start_ts`：事务的开始时间戳，也是事务的唯一 ID，可以用这个值在 TiDB 日志中查找事务相关的其他日志。
- `sql`：SQL 语句。

和内存使用相关的字段：

- `mem_max`：日志打印时语句已经使用的内存空间。该项使用两种单位标识内存使用量，分别为 Bytes 以及易于阅读的自适应单位（比如 MB、GB 等）。

和 SQL 执行的用户相关的字段：

- `user`：执行语句的用户名。
- `conn_id`：用户的连接 ID，可以用类似 `con:60026` 的关键字在 TiDB 日志中查找该连接相关的其他日志。
- `database`：执行语句时使用的 database。

和 TiKV Coprocessor Task 相关的字段：

- `wait_time`：该语句在 TiKV 的等待时间之和，因为 TiKV 的 Coprocessor 线程数是有限的，当所有的 Coprocessor 线程都在工作的时候，请求会排队；当队列中有某些请求耗时很长的时候，后面的请求的等待时间都会增加。

- request_count：该语句发送的 Coprocessor 请求的数量。
- total_keys：Coprocessor 扫过的 key 的数量。
- processed_keys：Coprocessor 处理的 key 的数量。与 total_keys 相比，processed_keys 不包含 MVCC 的旧版本。如果 processed_keys 和 total_keys 相差很大，说明旧版本比较多。
- num_cop_tasks：该语句发送的 Coprocessor 请求的数量。
- process_avg_time：Coprocessor 执行 task 的平均执行时间。
- process_p90_time：Coprocessor 执行 task 的 P90 分位执行时间。
- process_max_time：Coprocessor 执行 task 的最长执行时间。
- process_max_addr：task 执行时间最长的 Coprocessor 所在地址。
- wait_avg_time：Coprocessor 上 task 的等待时间。
- wait_p90_time：Coprocessor 上 task 的 P90 分位等待时间。
- wait_max_time：Coprocessor 上 task 的最长等待时间。
- wait_max_addr：task 等待时间最长的 Coprocessor 所在地址。

13.2.5 Statement Summary Tables

针对 SQL 性能相关的问题，MySQL 在 performance_schema 提供了 statement summary tables，用来监控和统计 SQL。例如其中的一张表 events_statements_summary_by_digest，提供了丰富的字段，包括延迟、执行次数、扫描行数、全表扫描次数等，有助于用户定位 SQL 问题。

为此，从 4.0.0-rc.1 版本开始，TiDB 在 information_schema（而不是 performance_schema）中提供与 events_statements_summary_by_digest 功能相似的系统表：

- statements_summary
- statements_summary_history
- cluster_statements_summary
- cluster_statements_summary_history

本文将详细介绍这些表，以及如何利用它们来排查 SQL 性能问题。

13.2.5.1 statements_summary

statements_summary 是 information_schema 里的一张系统表，它把 SQL 按 SQL digest 和 plan digest 分组，统计每一组的 SQL 信息。

此处的 SQL digest 与 slow log 里的 SQL digest 一样，是把 SQL 规一化后算出的唯一标识符。SQL 的规一化会忽略常量、空白符、大小写的差别。即语法一致的 SQL 语句，其 digest 也相同。

例如：

```
SELECT * FROM employee WHERE id IN (1, 2, 3) AND salary BETWEEN 1000 AND 2000;
select * from EMPLOYEE where ID in (4, 5) and SALARY between 3000 and 4000;
```

归一化后都是：

```
select * from employee where id in (...) and salary between ? and ?;
```

此处的 plan digest 是把执行计划规一化后算出的唯一标识符。执行计划的规一化会忽略常量的差别。由于相同的 SQL 可能产生不同的执行计划，所以可能分到多个组，同一个组内的执行计划是相同的。

statements_summary 用于保存 SQL 监控指标聚合后的结果。一般来说，每一项监控指标都包含平均值和最大值。例如执行延时对应 AVG_LATENCY 和 MAX_LATENCY 两个字段，分别是平均延时和最大延时。

为了监控指标的即时性，statements_summary 里的数据定期被清空，只展现最近一段时间内的聚合结果。清空周期由系统变量 tidb_stmt_summary_refresh_interval 设置。如果刚好在清空之后进行查询，显示的数据可能很少。

以下为查询 statements_summary 的部分结果：

```
  SUMMARY_BEGIN_TIME: 2020-01-02 11:00:00
    SUMMARY_END_TIME: 2020-01-02 11:30:00
           STMT_TYPE: Select
         SCHEMA_NAME: test
              DIGEST: 0611cc2fe792f8c146cc97d39b31d9562014cf15f8d41f23a4938ca341f54182
         DIGEST_TEXT: select * from employee where id = ?
         TABLE_NAMES: test.employee
         INDEX_NAMES: NULL
         SAMPLE_USER: root
          EXEC_COUNT: 3
         SUM_LATENCY: 1035161
         MAX_LATENCY: 399594
         MIN_LATENCY: 301353
         AVG_LATENCY: 345053
   AVG_PARSE_LATENCY: 57000
   MAX_PARSE_LATENCY: 57000
 AVG_COMPILE_LATENCY: 175458
 MAX_COMPILE_LATENCY: 175458
 ...........
             AVG_MEM: 103
             MAX_MEM: 103
            AVG_DISK: 65535
            MAX_DISK: 65535
   AVG_AFFECTED_ROWS: 0
          FIRST_SEEN: 2020-01-02 11:12:54
           LAST_SEEN: 2020-01-02 11:25:24
   QUERY_SAMPLE_TEXT: select * from employee where id=3100
    PREV_SAMPLE_TEXT:
         PLAN_DIGEST: f415b8d52640b535b9b12a9c148a8630d2c6d59e419aad29397842e32e8e5de3
                PLAN:  Point_Get_1     root    1       table:employee, handle:3100
```

> **注意：**

在 TiDB 中，statement summary tables 中字段的时间单位是纳秒 (ns)，而 MySQL 中的时间单位是皮秒 (ps)。

13.2.5.2 statements_summary_history

statements_summary_history 的表结构与 statements_summary 完全相同，用于保存历史时间段的数据。通过历史数据，可以排查过去出现的异常，也可以对比不同时间的监控指标。

字段 SUMMARY_BEGIN_TIME 和 SUMMARY_END_TIME 代表历史时间段的开始时间和结束时间。

13.2.5.3 statements_summary_evicted

statements_summary 表的容量受 tidb_stmt_summary_max_stmt_count 配置控制，内部使用 LRU 算法，一旦接收到的 SQL 种类超过了 tidb_stmt_summary_max_stmt_count，表中最久未被命中的记录就会被驱逐出表。TiDB 引入了 statements_summary_evicted 表，该表记录了各个时段被驱逐 SQL 语句的具体数量。

只有当 SQL 语句被 statement summary 表驱逐的时候，statements_summary_evicted 表的内容才会更新。statements_summary_evicted 表记录发生驱逐的时间段和被驱逐 SQL 的数量。

13.2.5.4 statement summary 的 cluster 表

statements_summary、statements_summary_history 和 statements_summary_evicted 仅显示单台 TiDB server 的 statement summary 数据。若要查询整个集群的数据，需要查询 cluster_statements_summary、cluster_statements_summary_history 或 cluster_statements_summary_evicted 表。

cluster_statements_summary 显示各台 TiDB server 的 statements_summary 数据，cluster_statements_summary_history 显示各台 TiDB server 的 statements_summary_history 数据，而 cluster_statements_summary_evicted 则显示各台 TiDB server 的 statements_summary_evicted 数据。这三张表用字段 INSTANCE 表示 TiDB server 的地址，其他字段与 statements_summary 相同。

13.2.5.5 参数配置

以下系统变量用于控制 statement summary：

- tidb_enable_stmt_summary：是否打开 statement summary 功能。1 代表打开，0 代表关闭，默认打开。statement summary 关闭后，系统表里的数据会被清空，下次打开后重新统计。经测试，打开后对性能几乎没有影响。
- tidb_stmt_summary_refresh_interval：statements_summary 的清空周期，单位是秒 (s)，默认值是 1800。
- tidb_stmt_summary_history_size：statements_summary_history 保存每种 SQL 的历史的数量，也是 statements_summary_evicted 的表容量，默认值是 24。
- tidb_stmt_summary_max_stmt_count：statement summary tables 保存的 SQL 种类数量，默认 3000 条。当 SQL 种类超过该值时，会移除最近没有使用的 SQL。这些 SQL 将会被 statements_summary_evicted 统计记录。
- tidb_stmt_summary_max_sql_length：字段 DIGEST_TEXT 和 QUERY_SAMPLE_TEXT 的最大显示长度，默认值是 4096。
- tidb_stmt_summary_internal_query：是否统计 TiDB 的内部 SQL。1 代表统计，0 代表不统计，默认不统计。

> **注意：**
>
> 当一种 SQL 因为达到 `tidb_stmt_summary_max_stmt_count` 限制要被移除时，TiDB 会移除该 SQL 语句种类在所有时间段的数据。因此，即使一个时间段内的 SQL 种类数量没有达到上限，显示的 SQL 语句数量也会比实际的少。如遇到该情况，对性能也有一些影响，建议调大 `tidb_stmt_summary_max_stmt_count` 的值。

statement summary 配置示例如下：

```
set global tidb_enable_stmt_summary = true;
set global tidb_stmt_summary_refresh_interval = 1800;
set global tidb_stmt_summary_history_size = 24;
```

以上配置生效后，`statements_summary` 每 30 分钟清空一次，所以 `statements_summary_history` 保存最近 12 小时的历史数。`statements_summary_evicted` 保存最近 24 个发生了 evict 的时间段记录；`statements_summary_evicted` 则以 30 分钟为一个记录周期，表容量为 24 个时间段。

> **注意：**
>
> `tidb_stmt_summary_history_size`、`tidb_stmt_summary_max_stmt_count`、`tidb_stmt_summary_max_sql_length` 这些配置都影响内存占用，建议根据实际情况调整，不宜设置得过大。

13.2.5.5.1 为 statement summary 设定合适的大小

在系统运行一段时间后，可以查看 `statements_summary` 表检查是否发生了 evict，例如：

```
select @@global.tidb_stmt_summary_max_stmt_count;
select count(*) from information_schema.statements_summary;
```

```
+-----------------------------------------+
| @@global.tidb_stmt_summary_max_stmt_count |
+-----------------------------------------+
| 3000                                    |
+-----------------------------------------+
1 row in set (0.001 sec)

+----------+
| count(*) |
+----------+
|     3001 |
+----------+
```

```
1 row in set (0.001 sec)
```

可以发现 statements_summary 表已经满了。再查看 statements_summary_evicted 表检查 evict 的数据。

```
select * from information_schema.statements_summary_evicted;
```

```
+---------------------+---------------------+---------------+
| BEGIN_TIME          | END_TIME            | EVICTED_COUNT |
+---------------------+---------------------+---------------+
| 2020-01-02 16:30:00 | 2020-01-02 17:00:00 |            59 |
+---------------------+---------------------+---------------+
| 2020-01-02 16:00:00 | 2020-01-02 16:30:00 |            45 |
+---------------------+---------------------+---------------+
2 row in set (0.001 sec)
```

由上可知，对最多 59 种 SQL 发生了 evict，也就是说最少应将 statement summary 的容量增大至 59 条记录。

13.2.5.6 目前的限制

Statement summary tables 现在还存在以下限制：

- TiDB server 重启后以上 4 张表的 statement summary 会全部丢失。因为 statement summary tables 全部都是内存表，不会持久化数据，所以一旦 server 被重启，statement summary 随之丢失。

13.2.5.7 排查示例

下面用两个示例问题演示如何利用 statement summary 来排查。

13.2.5.7.1 SQL 延迟比较大，是不是服务端的问题？

例如客户端显示 employee 表的点查比较慢，那么可以按 SQL 文本来模糊查询：

```
SELECT avg_latency, exec_count, query_sample_text
    FROM information_schema.statements_summary
    WHERE digest_text LIKE 'select * from employee%';
```

结果如下，avg_latency 是 1 ms 和 0.3 ms，在正常范围，所以可以判定不是服务端的问题，继而排查客户端或网络问题。

```
+-------------+-----------+-----------------------------------------+
| avg_latency | exec_count | query_sample_text                      |
+-------------+-----------+-----------------------------------------+
|     1042040 |          2 | select * from employee where name='eric' |
|      345053 |          3 | select * from employee where id=3100     |
+-------------+-----------+-----------------------------------------+
2 rows in set (0.00 sec)
```

13.2.5.7.2 哪类 SQL 的总耗时最高？

假如上午 10:00 到 10:30 的 QPS 明显下降，可以从历史表中找出当时耗时最高的三类 SQL：

```
SELECT sum_latency, avg_latency, exec_count, query_sample_text
    FROM information_schema.statements_summary_history
    WHERE summary_begin_time='2020-01-02 10:00:00'
    ORDER BY sum_latency DESC LIMIT 3;
```

结果显示以下三类 SQL 的总延迟最高，所以这些 SQL 需要重点优化。

```
+-------------+-------------+-----------+-----------------------------------------------------------------------------
    ↪
| sum_latency | avg_latency | exec_count | query_sample_text
    ↪                                                          |
+-------------+-------------+-----------+-----------------------------------------------------------------------------
    ↪
|     7855660 |     1122237 |          7 | select avg(salary) from employee where company_id=2013
    ↪                  |
|     7241960 |     1448392 |          5 | select * from employee join company on employee.
    ↪ company_id=company.id |
|     2084081 |     1042040 |          2 | select * from employee where name='eric'
    ↪                                |
+-------------+-------------+-----------+-----------------------------------------------------------------------------
    ↪
3 rows in set (0.00 sec)
```

13.2.5.8 表的字段介绍

13.2.5.8.1 `statements_summary` 字段介绍

下面介绍 `statements_summary` 表中各个字段的含义。

SQL 的基础信息：

- `STMT_TYPE`：SQL 语句的类型
- `SCHEMA_NAME`：执行这类 SQL 的当前 schema
- `DIGEST`：这类 SQL 的 digest
- `DIGEST_TEXT`：规一化后的 SQL
- `QUERY_SAMPLE_TEXT`：这类 SQL 的原 SQL 语句，多条语句只取其中一条
- `TABLE_NAMES`：SQL 中涉及的所有表，多张表用，分隔
- `INDEX_NAMES`：SQL 中使用的索引名，多个索引用，分隔
- `SAMPLE_USER`：执行这类 SQL 的用户名，多个用户名只取其中一个
- `PLAN_DIGEST`：执行计划的 digest
- `PLAN`：原执行计划，多条语句只取其中一条的执行计划
- `PLAN_CACHE_HITS`：这类 SQL 语句命中 plan cache 的总次数
- `PLAN_IN_CACHE`：这类 SQL 语句的上次执行是否命中了 plan cache

执行时间相关的信息：

- SUMMARY_BEGIN_TIME：当前统计的时间段的开始时间
- SUMMARY_END_TIME：当前统计的时间段的结束时间
- FIRST_SEEN：这类 SQL 的首次出现时间
- LAST_SEEN：这类 SQL 的最后一次出现时间

在 TiDB server 上的执行数据：

- EXEC_COUNT：这类 SQL 的总执行次数
- SUM_ERRORS：执行过程中遇到的 error 的总数
- SUM_WARNINGS：执行过程中遇到的 warning 的总数
- SUM_LATENCY：这类 SQL 的总延时
- MAX_LATENCY：这类 SQL 的最大延时
- MIN_LATENCY：这类 SQL 的最小延时
- AVG_LATENCY：这类 SQL 的平均延时
- AVG_PARSE_LATENCY：解析器的平均延时
- MAX_PARSE_LATENCY：解析器的最大延时
- AVG_COMPILE_LATENCY：优化器的平均延时
- MAX_COMPILE_LATENCY：优化器的最大延时
- AVG_MEM：使用的平均内存，单位 byte
- MAX_MEM：使用的最大内存，单位 byte
- AVG_DISK：使用的平均硬盘空间，单位 byte
- MAX_DISK：使用的最大硬盘空间，单位 byte

和 TiKV Coprocessor Task 相关的字段：

- SUM_COP_TASK_NUM：发送 Coprocessor 请求的总数
- MAX_COP_PROCESS_TIME：cop-task 的最大处理时间
- MAX_COP_PROCESS_ADDRESS：执行时间最长的 cop-task 所在地址
- MAX_COP_WAIT_TIME：cop-task 的最大等待时间
- MAX_COP_WAIT_ADDRESS：等待时间最长的 cop-task 所在地址
- AVG_PROCESS_TIME：SQL 在 TiKV 的平均处理时间
- MAX_PROCESS_TIME：SQL 在 TiKV 的最大处理时间
- AVG_WAIT_TIME：SQL 在 TiKV 的平均等待时间
- MAX_WAIT_TIME：SQL 在 TiKV 的最大等待时间
- AVG_BACKOFF_TIME：SQL 遇到需要重试的错误时在重试前的平均等待时间
- MAX_BACKOFF_TIME：SQL 遇到需要重试的错误时在重试前的最大等待时间
- AVG_TOTAL_KEYS：Coprocessor 扫过的 key 的平均数量
- MAX_TOTAL_KEYS：Coprocessor 扫过的 key 的最大数量
- AVG_PROCESSED_KEYS：Coprocessor 处理的 key 的平均数量。相比 avg_total_keys，avg_processed_keys 不包含 MVCC 的旧版本。如果 avg_total_keys 和 avg_processed_keys 相差很大，说明旧版本比较多
- MAX_PROCESSED_KEYS：Coprocessor 处理的 key 的最大数量

和事务相关的字段：

- AVG_PREWRITE_TIME：prewrite 阶段消耗的平均时间
- MAX_PREWRITE_TIME prewrite 阶段消耗的最大时间
- AVG_COMMIT_TIME：commit 阶段消耗的平均时间
- MAX_COMMIT_TIME：commit 阶段消耗的最大时间
- AVG_GET_COMMIT_TS_TIME：获取 commit_ts 的平均时间
- MAX_GET_COMMIT_TS_TIME：获取 commit_ts 的最大时间
- AVG_COMMIT_BACKOFF_TIME：commit 时遇到需要重试的错误时在重试前的平均等待时间
- MAX_COMMIT_BACKOFF_TIME：commit 时遇到需要重试的错误时在重试前的最大等待时间
- AVG_RESOLVE_LOCK_TIME：解决事务的锁冲突的平均时间
- MAX_RESOLVE_LOCK_TIME：解决事务的锁冲突的最大时间
- AVG_LOCAL_LATCH_WAIT_TIME：本地事务等待的平均时间
- MAX_LOCAL_LATCH_WAIT_TIME：本地事务等待的最大时间
- AVG_WRITE_KEYS：写入 key 的平均数量
- MAX_WRITE_KEYS：写入 key 的最大数量
- AVG_WRITE_SIZE：写入的平均数据量，单位 byte
- MAX_WRITE_SIZE：写入的最大数据量，单位 byte
- AVG_PREWRITE_REGIONS：prewrite 涉及的平均 Region 数量
- MAX_PREWRITE_REGIONS：prewrite 涉及的最大 Region 数量
- AVG_TXN_RETRY：事务平均重试次数
- MAX_TXN_RETRY：事务最大重试次数
- SUM_BACKOFF_TIMES：这类 SQL 遇到需要重试的错误后的总重试次数
- BACKOFF_TYPES：遇到需要重试的错误时的所有错误类型及每种类型重试的次数，格式为 类型:次数。如有多种错误则用 , 分隔，例如 txnLock:2,pdRPC:1
- AVG_AFFECTED_ROWS：平均影响行数
- PREV_SAMPLE_TEXT：当 SQL 是 COMMIT 时，该字段为 COMMIT 的前一条语句；否则该字段为空字符串。当 SQL 是 COMMIT 时，按 digest 和 prev_sample_text 一起分组，即不同 prev_sample_text 的 COMMIT 也会分到不同的行

13.2.5.8.2 statements_summary_evicted 字段介绍

- BEGIN_TIME: 记录的开始时间；
- END_TIME: 记录的结束时间；
- EVICTED_COUNT：在记录的时间段内 evict 了多少种 SQL。

13.3 TiFlash 常见问题

本文介绍了一些 TiFlash 常见问题、原因及解决办法。

13.3.1 TiFlash 未能正常启动

该问题可能由多个因素构成，可以通过以下步骤依次排查：

1. 检查系统环境是否是 CentOS8。

   CentOS8 中缺少 libnsl.so 系统库，可以通过手动安装的方式解决：

```
dnf install libnsl
```

2. 检查系统的 ulimit 参数设置。

```
ulimit -n 1000000
```

3. 使用 PD Control 工具检查在该节点（相同 IP 和 Port）是否有之前未成功下线的 TiFlash 实例，并将它们强制下线。（下线步骤参考手动缩容 TiFlash 节点）

如果遇到上述方法无法解决的问题，可以打包 TiFlash 的 log 文件夹，并在 AskTUG 社区中提问。

13.3.2 TiFlash 副本始终处于不可用状态

该问题一般由于配置错误或者环境问题导致 TiFlash 处于异常状态，可以先通过以下步骤定位问题组件：

1. 使用 pd-ctl 检查 PD 的Placement Rules 功能是否开启：

```
echo 'config show replication' | /path/to/pd-ctl -u http://${pd-ip}:${pd-port}
```

   - 如果返回 true，进入下一步。
   - 如果返回 false，你需要先开启 Placement Rules 特性 后再进入下一步。

2. 通过 TiFlash-Summary 监控面板下的 UpTime 检查操作系统中 TiFlash 进程是否正常。

3. 通过 pd-ctl 查看 TiFlash proxy 状态是否正常：

```
echo "store" | /path/to/pd-ctl -u http://${pd-ip}:${pd-port}
```

   store.labels 中含有 {"key": "engine", "value": "tiflash"} 信息的为 TiFlash proxy。

4. 查看 pd buddy 是否正常打印日志（日志路径的对应配置项 [flash.flash_cluster] log 设置的值，默认为 TiFlash 配置文件配置的 tmp 目录下）。

5. 检查配置的副本数是否小于等于集群 TiKV 节点数。若配置的副本数超过 TiKV 节点数，则 PD 不会向 TiFlash 同步数据；

```
echo 'config placement-rules show' | /path/to/pd-ctl -u http://${pd-ip}:${pd-port}
```

   再确认 “default: count” 参数值。

> **注意：**
>
> 开启 Placement Rules 后，原先的 max-replicas 及 location-labels 配置项将不再生效。如果需要调整副本策略，应当使用 Placement Rules 相关接口。

6. 检查 TiFlash 节点对应 store 所在机器剩余的磁盘空间是否充足。默认情况下当磁盘剩余空间小于该 store 的 capacity 的 20%（通过 low-space-ratio 参数控制）时，PD 不会向 TiFlash 调度数据。

13.3.3 TiFlash 查询时间不稳定，同时错误日志中打印出大量的 Lock Exception

该问题是由于集群中存在大量写入，导致 TiFlash 查询时遇到锁并发生查询重试。

可以在 TiDB 中将查询时间戳设置为 1 秒前（例如：假设当前时间为 '2020-04-08 20:15:01'，可以在执行 query 前执行 `set @@tidb_snapshot='2020-04-08 20:15:00';`），来减小 TiFlash 查询碰到锁的可能性，从而减轻查询时间不稳定的程度。

13.3.4 部分查询返回 Region Unavailable 的错误

如果在 TiFlash 上的负载压力过大，会导致 TiFlash 数据同步落后，部分查询可能会返回 `Region Unavailable` 的错误。

在这种情况下，可以增加 TiFlash 节点分担负载压力。

13.3.5 数据文件损坏

可依照如下步骤进行处理：

1. 参照下线 TiFlash 节点下线对应的 TiFlash 节点。
2. 清除该 TiFlash 节点的相关数据。
3. 重新在集群中部署 TiFlash 节点。

13.3.6 TiFlash 分析慢

如果语句中含有 MPP 模式不支持的算子或函数等，TiDB 不会选择 MPP 模式，可能导致分析慢。此时，可以执行 EXPLAIN 语句检查 SQL 中是否含有 TiFlash 不支持的函数或算子。

```
create table t(a datetime);
alter table t set tiflash replica 1;
insert into t values('2022-01-13');
set @@session.tidb_enforce_mpp=1;
explain select count(*) from t where subtime(a, '12:00:00') > '2022-01-01' group by a;
show warnings;
```

示例中，warning 消息显示，因为 TiDB 5.4 及更早的版本尚不支持 `subtime` 函数的下推，因此 TiDB 没有选择 MPP 模式。

```
+---------+------+-----------------------------------------------------------------------------------+
> | Level   | Code | Message
    ↪ |
+---------+------+-----------------------------------------------------------------------------------+
| Warning | 1105 | Scalar function 'subtime'(signature: SubDatetimeAndString, return type:
    ↪ datetime) is not supported to push down to tiflash now.       |
+---------+------+-----------------------------------------------------------------------------------+
```

13.3.7 TiFlash 数据不同步

在部署完 TiFlash 节点且进行了数据的同步操作（ALTER 操作）之后，如果实际没有数据同步到 TiFlash 节点，你可以通过以下步骤确认或解决问题：

1. 检查同步操作是否执行。

   执行 `ALTER table <tbl_name> set tiflash replica <num>` 操作，查看是否有正常返回:

   - 如果有正常返回，进入下一步。
   - 如果无正常返回，请执行 `SELECT * FROM information_schema.tiflash_replica` 检查是否已经创建 TiFlash replica。如果没有，请重新执行 `ALTER table ${tbl_name} set tiflash replica ${num}`，查看是否有其他执行语句（如 `add index`），或者检查 DDL 操作是否正常。

2. 检查 TiFlash 进程是否正常。

   查看 `progress` 和 `tiflash_cluster_manager.log` 日志中的 `flash_region_count` 参数以及 TiFlash Uptime 是否有变化:

   - 如果有变化，说明 TiFlash 进程正常，进入下一步。
   - 如果没有变化，说明 TiFlash 进程异常，请通过 `tiflash` 日志进一步排查。

3. 使用 pd-ctl 检查 PD 的Placement Rules 功能是否开启：

   ```
   echo 'config show replication' | /path/to/pd-ctl -u http://<pd-ip>:<pd-port>
   ```

   - 如果返回 `true`，进入下一步。
   - 如果返回 `false`，你需要先开启 Placement Rules 特性，然后进入下一步。

4. 检查集群副本数 `max-replicas` 配置是否合理。

   - 如果 `max-replicas` 取值未超过 TiKV 节点数，进入下一步。
   - 如果 `max-replicas` 超过 TiKV 节点数，PD 不会向 TiFlash 同步数据。此时，请将 `max-replicas` 修改为小于等于 TiKV 节点数的整数。

> **注意：**
>
> `max-replicas` 的默认值是 3。在生产环境中，TiKV 节点数一般大于该值；在测试环境中，可以修改为 1。

```
curl -X POST -d '{
    "group_id": "pd",
    "id": "default",
    "start_key": "",
    "end_key": "",
    "role": "voter",
    "count": 3,
    "location_labels": [
```

```
    "host"

    ]
}' <http://172.16.x.xxx:2379/pd/api/v1/config/rule>
```

5. 检查 TiFlash 与 TiDB 或 PD 的连接是否正常。

   检查 `flash_cluster_manager.log` 日志，查找关键字 `ERROR`。

   - 如果没有关键字 `ERROR`，说明连接正常，进入下一步。
   - 如果有关键字 `ERROR`，说明连接异常，请进行以下检查。
   - 检查日志文件中是否出现了 PD 相关关键字。

     如果出现，请检查 TiFlash 配置文件中 `raft.pd_addr` 对应 PD 地址是否有效。检查方法：执行 `curl '{pd-addr}/pd/api/v1/config/rules'`，检查 5 秒内是否有正常返回。
   - 检查日志文件中是否出现了 TiDB 相关关键字。

     如果出现，请检查 TiFlash 配置文件中 `flash.tidb_status_addr` 对应 TiDB 的 status 服务地址是否有效。检查方法：执行 `curl '{tidb-status-addr}/tiflash/replica'`, 检查 5 秒内是否有正常返回。
   - 检查节点间能否相互连通。

> 注意：
> 如问题依然无法解决，收集相关组件的日志进行排查。

6. 检查表是否创建 `placement-rule`。

   检查 `flash_cluster_manager.log` 日志，查找是否存在关键字 `Set placement rule ⋯ table-<table_id >-r`。

   - 有关键字，进入下一步。
   - 没有关键字，收集相关组件的日志进行排查。

7. 检查 PD 是否正常发起调度。

   查看 `pd.log` 日志是否出现 `table-<table_id>-r` 关键字，且之后是否出现 `add operator` 之类的调度行为。

   - 是，PD 调度正常。
   - 否，PD 调度异常，请联系 PingCAP 技术支持人员协助排查。

> 注意：
> 当需要同步的表有很多小 Region 且 Region merge 参数已开启或取值较大时，可能会出现同步进度一段时间不变化或者变小的现象。

13.3.8 TiFlash 数据同步卡住

如果 TiFlash 数据一开始可以正常同步，过一段时间后全部或者部分数据无法继续同步，你可以通过以下步骤确认或解决问题：

1. 检查磁盘空间。

   检查磁盘使用空间比例是否高于 `low-space-ratio` 的值（默认值 0.8，即当节点的空间占用比例超过 80% 时，为避免磁盘空间被耗尽，PD 会尽可能避免往该节点迁移数据）。

   - 如果磁盘使用率大于等于 `low-space-ratio`，说明磁盘空间不足。此时，请删除不必要的文件，如 `${data}/flash/` 目录下的 `space_placeholder_file` 文件（必要时可在删除文件后将 `reserve-space` 设置为 0MB）。
   - 如果磁盘使用率小于 `low-space-ratio`，说明磁盘空间正常，进入下一步。

2. 检查 TiKV、PD、TiFlash 之间的网络连接情况。

   在 `flash_cluster_manager.log` 中，检查同步卡住的表对应的 `flash_region_count` 是否有更新。

   - 没有更新，检查下一步。
   - 有更新，检查是否有 down peer（down peer 没有清理干净可能会导致同步卡住）。
   - 执行 `pd-ctl region check-down-peer` 命令检查是否有 down peer。
   - 如果存在 down peer，执行 `pd-ctl operator add remove-peer\<region-id> \<tiflash-store-id>` 命令将其清除。

3. 查看 CPU 使用率。

   在 Grafana 界面，选择 TiFlash-Proxy-Details > Thread CPU > Region task worker pre-handle/generate snapshot CPU，查看监控中 `<instance-ip>:<instance-port>-region-worker` 对应线程的 CPU 使用率。

   若曲线为一条直线，表示 TiFlash 发生卡死，可强制杀进程重启或联系 PingCAP 技术支持人员。

13.3.9 数据同步慢

同步慢可能由多种原因引起，你可以按以下步骤进行排查。

1. 调整调度参数取值。

   - 调大`store limit`，加快同步速度。
   - 调小`config set patrol-region-interval 10ms`，增加 TiKV 侧 checker 扫描 Region 的频率。
   - 调大`region merge` 参数，减少 Region 数量，从而减少扫描数量，提高检查频次。

2. 调整 TiFlash 侧负载。

   TiFlash 负载过大会引起同步慢，可通过 Grafana 中的 TiFlash-Summary 面板查看各个指标的负载情况：

   - `Applying snapshots Count`: `TiFlash-summary` > `raft` > `Applying snapshots Count`
   - `Snapshot Predecode Duration`: `TiFlash-summary` > `raft` > `Snapshot Predecode Duration`
   - `Snapshot Flush Duration`: `TiFlash-summary` > `raft` > `Snapshot Flush Duration`
   - `Write Stall Duration`: `TiFlash-summary` > `Storage Write Stall` > `Write Stall Duration`

- generate snapshot CPU: TiFlash-Proxy-Details > Thread CPU > Region task worker pre-handle/ generate snapshot CPU

根据业务优先级，调整负载情况。

13.4 错误码与故障诊断

本篇文档描述在使用 TiDB 过程中会遇到的问题以及解决方法。

13.4.1 错误码

TiDB 兼容 MySQL 的错误码，在大多数情况下，返回和 MySQL 一样的错误码。关于 MySQL 的错误码列表，详见 [MySQL 5.7 Error Message Reference。另外还有一些 TiDB 特有的错误码：

> 注意：
>
> 有一部分错误码属于内部错误，正常情况下 TiDB 会自行处理不会直接返回给用户，故没有在此列出。
>
> 如果您遇到了这里没有列出的错误码，请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8001

  请求使用的内存超过 TiDB 内存使用的阈值限制。出现这种错误，可以通过调整 `mem-quota-query` 来增大单个 SQL 使用的内存上限。

- Error Number: 8002

  带有 `SELECT FOR UPDATE` 语句的事务，在遇到写入冲突时，为保证一致性无法进行重试，事务将进行回滚并返回该错误。出现这种错误，应用程序可以安全地重新执行整个事务。

- Error Number: 8003

  `ADMIN CHECK TABLE` 命令在遇到行数据跟索引不一致的时候返回该错误，在检查表中数据是否有损坏时常出现。出现该错误时，请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8004

  单个事务过大，原因及解决方法请参考这里

- Error Number: 8005

  事务在 TiDB 中遇到了写入冲突，原因及解决方法请参考这里

- Error Number: 8018

  当执行重新载入插件时，如果之前插件没有载入过，则会出现该错误。出现该错误，进行插件首次载入即可。

- Error Number: 8019

  重新载入的插件版本与之前的插件版本不一致，无法重新载入插件并报告该错误。可重新载入插件，确保插件的版本与之前载入的插件版本一致。

- Error Number: 8020

  当表被加锁时，如果对该表执行写入操作，将出现该错误。请将表解锁后，再进行尝试写入。

- Error Number: 8021

  当向 TiKV 读取的 key 不存在时将出现该错误，该错误用于内部使用，对外表现为读到的结果为空。

- Error Number: 8022

  事务提交失败，已经回滚，应用程序可以安全的重新执行整个事务。

- Error Number: 8023

  在事务内，写入事务缓存时，设置了空值，将返回该错误。这是一个内部使用的错误，将由内部进行处理，不会返回给应用程序。

- Error Number: 8024

  非法的事务。当事务执行时，发现没有获取事务的 ID (Start Timestamp)，代表正在执行的事务是一个非法的事务，将返回该错误。通常情况下不会出现该问题，当发生时，请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8025

  写入的单条键值对过大。TiDB 默认支持最大 6MB 的单个键值对，超过该限制可适当调整`txn-entry-size-limit`配置项以放宽限制。

- Error Number: 8026

  使用了没有实现的接口函数。该错误仅用于数据库内部，应用程序不会收到这个错误。

- Error Number: 8027

  表结构版本过期。TiDB 采用在线变更表结构的方法。当 TiDB server 表结构版本落后于整个系统的时，执行 SQL 将遇到该错误。遇到该错误，请检查该 TiDB server 与 PD leader 之间的网络。

- Error Number: 8028

  TiDB 没有表锁（在 MySQL 中称为元数据锁，在其他数据库中可能称为意向锁）。当事务执行时，TiDB 表结构发生了变化是无法被事务感知到的。因此，TiDB 在事务提交时，会对事务涉及表的结构进行检查。如果事务执行中表结构发生了变化，则事务将提交失败，并返回该错误。遇到该错误，应用程序可以安全地重新执行整个事务。

- Error Number: 8029

  当数据库内部进行数值转换发生错误时，将会出现该错误，该错误仅在内部使用，对外部应用将转换为具体类型的错误。

- Error Number: 8030

  将值转变为带符号正整数时发生了越界，将结果显示为负数。多在告警信息里出现。

- Error Number: 8031

  将负数转变为无符号数时，将负数转变为了正数。多在告警信息里出现。

- Error Number: 8032

  使用了非法的 year 格式。year 只允许 1 位、2 位和 4 位数。

- Error Number: 8033

  使用了非法的 year 值。year 的合法范围是 (1901, 2155)。

- Error Number: 8037

  week 函数中使用了非法的 mode 格式。mode 必须是一位数字，范围 [0, 7]。

- Error Number: 8038

  字段无法获取到默认值。一般作为内部错误使用，转换成其他具体错误类型后，返回给应用程序。

- Error Number: 8040

  尝试进行不支持的操作，比如在 View 和 Sequence 上进行 lock table。

- Error Number: 8047

  设置了不支持的系统变量值，通常在用户设置了数据库不支持的变量值后的告警信息里出现。

- Error Number: 8048

  设置了不支持的隔离级别，如果是使用第三方工具或框架等无法修改代码进行适配的情况，可以考虑通过`tidb_skip_isolation_level_check` 来绕过这一检查。

```
set @@tidb_skip_isolation_level_check = 1;
```

- Error Number: 8050

  设置了不支持的权限类型，遇到该错误请参考 TiDB 权限说明进行调整。

- Error Number: 8051

  TiDB 在解析客户端发送的 Exec 参数列表时遇到了未知的数据类型。如果遇到这个错误，请检查客户端是否正常，如果客户端正常请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8052

  来自客户端的数据包的序列号错误。如果遇到这个错误，请检查客户端是否正常，如果客户端正常请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8055

  当前快照过旧，数据可能已经被 GC。可以调大`tidb_gc_life_time` 的值来避免该问题。从 TiDB v4.0.8 版本起，TiDB 会自动为长时间运行的事务保留数据，一般不会遇到该错误。

  有关 GC 的介绍和配置可以参考 GC 机制简介和 GC 配置文档。

- Error Number: 8059

  自动随机量可用次数用尽无法进行分配。当前没有恢复这类错误的方法。建议在使用 auto random 功能时使用 bigint 以获取最大的可分配次数，并尽量避免手动给 auto random 列赋值。相关的介绍和使用建议可以参考 auto random 功能文档。

- Error Number: 8060

  非法的自增列偏移量。请检查 `auto_increment_increment` 和 `auto_increment_offset` 的取值是否符合要求。

- Error Number: 8061

  不支持的 SQL Hint。 请参考Optimizer Hints 检查和修正 SQL Hint。

- Error Number: 8062

  SQL Hint 中使用了非法的 token，与 Hint 的保留字冲突。 请参考Optimizer Hints 检查和修正 SQL Hint。

- Error Number: 8063

  SQL Hint 中限制内存使用量超过系统设置的上限，设置被忽略。请参考Optimizer Hints 检查和修正 SQL Hint。

- Error Number: 8064

  解析 SQL Hint 失败。 请参考Optimizer Hints 检查和修正 SQL Hint。

- Error Number: 8065

  SQL Hint 中使用了非法的整数。 请参考Optimizer Hints 检查和修正 SQL Hint。

- Error Number: 8066

  JSON_OBJECTAGG 函数的第二个参数是非法参数。

- Error Number: 8101

  插件 ID 格式错误，正确的格式是 `[name]-[version]` 并且 name 和 version 中不能带有 ‘-’ 。

- Error Number: 8102

  无法读取插件定义信息。 请检查插件相关的配置。

- Error Number: 8103

  插件名称错误，请检查插件的配置。

- Error Number: 8104

  插件版本不匹配，请检查插件的配置。

- Error Number: 8105

  插件被重复载入。

- Error Number: 8106

  插件定义的系统变量名称没有以插件名作为开头，请联系插件的开发者进行修复。

- Error Number: 8107

  载入的插件未指定版本或指定的版本过低，请检查插件的配置。

- Error Number: 8108

  不支持的执行计划类型。 该错误为内部处理的错误，如果遇到该报错请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8109

  analyze 索引时找不到指定的索引。

- Error Number: 8110

  不能进行笛卡尔积运算，需要将配置文件里的 `cross-join` 设置为 `true`。

- Error Number: 8111

  execute 语句执行时找不到对应的 prepare 语句。

- Error Number: 8112

  execute 语句的参数个数与 prepare 语句不符合。

- Error Number: 8113

  execute 语句涉及的表结构在 prepare 语句执行后发生了变化。

- Error Number: 8115

  不支持 prepare 多行语句。

- Error Number: 8116

  不支持 prepare DDL 语句。

- Error Number: 8120

  获取不到事务的 start tso，请检查 PD Server 状态/监控/日志以及 TiDB Server 与 PD Server 之间的网络。

- Error Number: 8121

  权限检查失败，请检查数据库的权限配置。

- Error Number: 8122

  指定了通配符，但是找不到对应的表名。

- Error Number: 8123

  带聚合函数的 SQL 中返回非聚合的列，违反了 `only_full_group_by` 模式。请修改 SQL 或者考虑关闭 `only_full_group_by` 模式。

- Error Number: 8129

  TiDB 尚不支持键长度 >= 65536 的 JSON 对象。

- Error Number: 8200

  尚不支持的 DDL 语法。请参考与 MySQL DDL 的兼容性。

- Error Number: 8214

  DDL 操作被 admin cancel 操作终止。

- Error Number: 8215

  Admin Repair 表失败，如果遇到该报错请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8216

  自动随机列使用的方法不正确，请参考 auto random 功能文档进行修改。

- Error Number: 8223

  检测出数据与索引不一致的错误，如果遇到该报错请向 PingCAP 工程师或通过官方论坛寻求帮助。

- Error Number: 8224

  找不到 DDL job，请检查 restore 操作指定的 job id 是否存在。

- Error Number: 8225

  DDL 已经完成，无法被取消。

- Error Number: 8226

  DDL 几乎要完成了，无法被取消。

- Error Number: 8227

  创建 Sequence 时使用了不支持的选项，支持的选项的列表可以参考 Sequence 使用文档。

- Error Number: 8228

  在 Sequence 上使用 `setval` 时指定了不支持的类型，该函数的示例可以在 Sequence 使用文档中找到。

- Error Number: 8229

  事务超过存活时间，遇到该问题可以提交或者回滚当前事务，开启一个新事务。

- Error Number: 8230

  TiDB 目前不支持在新添加的列上使用 Sequence 作为默认值，如果尝试进行这类操作会返回该错误。

- Error Number: 9001

  请求 PD 超时，请检查 PD Server 状态/监控/日志以及 TiDB Server 与 PD Server 之间的网络。

- Error Number: 9002

  请求 TiKV 超时，请检查 TiKV Server 状态/监控/日志以及 TiDB Server 与 TiKV Server 之间的网络。

- Error Number: 9003

  TiKV 操作繁忙，一般出现在数据库负载比较高时，请检查 TiKV Server 状态/监控/日志。

- Error Number: 9004

  当数据库上承载的业务存在大量的事务冲突时，会遇到这种错误，请检查业务代码。

- Error Number: 9005

  某个 Raft Group 不可用，如副本数目不足，出现在 TiKV 比较繁忙或者是 TiKV 节点停机的时候，请检查 TiKV Server 状态/监控/日志。

- Error Number: 9006

  GC Life Time 间隔时间过短，长事务本应读到的数据可能被清理了，应增加 GC Life Time。

- Error Number: 9007

  事务在 TiKV 中遇到了写入冲突，原因及解决方法请参考这里。

- Error Number: 9008

  同时向 TiKV 发送的请求过多，超过了限制。请调大 `tidb_store_limit` 或将其设置为 0 来取消对请求流量的限制。

- Error Number: 9010

  TiKV 无法处理这条 raft log，请检查 TiKV Server 状态/监控/日志。

- Error Number: 9012

  请求 TiFlash 超时。请检查 TiFlash Server 状态/监控/日志以及 TiDB Server 与 TiFlash Server 之间的网络。

- Error Number: 9013

  TiFlash 操作繁忙。该错误一般出现在数据库负载比较高时。请检查 TiFlash Server 的状态/监控/日志。

13.4.2 故障诊断

参见故障诊断文档以及 FAQ。

14 常见问题解答 (FAQ)

14.1 安装部署 FAQ

本文介绍 TiDB 集群安装部署的常见问题、原因及解决方法。

14.1.1 环境准备 FAQ

操作系统版本要求如下表：

Linux 操作系统平台	版本
Red Hat Enterprise Linux	7.3 及以上
CentOS	7.3 及以上
Oracle Enterprise Linux	7.3 及以上

14.1.1.1 为什么要在 CentOS 7 上部署 TiDB 集群？

TiDB 作为一款开源分布式 NewSQL 数据库，可以很好的部署和运行在 Intel 架构服务器环境及主流虚拟化环境，并支持绝大多数的主流硬件网络，作为一款高性能数据库系统，TiDB 支持主流的 Linux 操作系统环境，具体可以参考 TiDB 的官方部署要求。

其中 TiDB 在 CentOS 7.3 的环境下进行大量的测试，同时也有很多这个操作系统的部署最佳实践，因此，我们推荐客户在部署 TiDB 的时候使用 CentOS 7.3+ 以上的 Linux 操作系统。

14.1.2 硬件要求 FAQ

TiDB 支持部署和运行在 Intel x86-64 架构的 64 位通用硬件服务器平台。对于开发、测试、及生产环境的服务器硬件配置有以下要求和建议：

14.1.2.1 开发及测试环境

组件	CPU	内存	本地存储	网络	实例数量 (最低要求)
TiDB	8 核 +	16 GB+	SAS, 200 GB+	千兆网卡	1（可与 PD 同机器）

组件	CPU	内存	本地存储	网络	实例数量 (最低要求)
PD	8 核 +	16 GB+	SAS, 200 GB+	千兆网卡	1（可与 TiDB 同机器）
TiKV	8 核 +	32 GB+	SSD, 200 GB+	千兆网卡	3
				服务器总计	4

14.1.2.2 生产环境

组件	CPU	内存	硬盘类型	网络	实例数量 (最低要求)
TiDB	16 核 +	48 GB+	SAS	万兆网卡（2 块最佳）	2
PD	8 核 +	16 GB+	SSD	万兆网卡（2 块最佳）	3
TiKV	16 核 +	48 GB+	SSD	万兆网卡（2 块最佳）	3
监控	8 核 +	16 GB+	SAS	千兆网卡	1
				服务器总计	9

14.1.2.3 两块网卡的目的是？万兆的目的是？

作为一个分布式集群，TiDB 对时间的要求还是比较高的，尤其是 PD 需要分发唯一的时间戳，如果 PD 时间不统一，如果有 PD 切换，将会等待更长的时间。两块网卡可以做 bond，保证数据传输的稳定，万兆可以保证数据传输的速度，千兆网卡容易出现瓶颈，我们强烈建议使用万兆网卡。

14.1.2.4 SSD 不做 RAID 是否可行？

资源可接受的话，我们建议做 RAID 10，如果资源有限，也可以不做 RAID。

14.1.2.5 TiDB 集群各个组件的配置推荐？

- TiDB 需要 CPU 和内存比较好的机器，参考官网配置要求，如果后期需要开启 TiDB Binlog，根据业务量的评估和 GC 时间的要求，也需要本地磁盘大一点，不要求 SSD 磁盘；
- PD 里面存了集群元信息，会有频繁的读写请求，对磁盘 I/O 要求相对比较高，磁盘太差会影响整个集群性能，推荐 SSD 磁盘，空间不用太大。另外集群 Region 数量越多对 CPU、内存的要求越高；
- TiKV 对 CPU、内存、磁盘要求都比较高，一定要用 SSD 磁盘。

详情可参考TiDB 软硬件环境需求。

14.1.3 安装部署 FAQ

如果用于生产环境，推荐使用 TiUP 使用 TiUP 部署 TiDB 集群。

14.1.3.1 为什么修改了 TiKV/PD 的 toml 配置文件，却没有生效？

这种情况一般是因为没有使用 `--config` 参数来指定配置文件（目前只会出现在 binary 部署的场景），TiKV/PD 会按默认值来设置。如果要使用配置文件，请设置 TiKV/PD 的 `--config` 参数。对于 TiKV 组件，修改配置后重启服务即可；对于 PD 组件，只会在第一次启动时读取配置文件，之后可以使用 pd-ctl 的方式来修改配置，详情可参考 PD 配置参数。

14.1.3.2 TiDB 监控框架 Prometheus + Grafana 监控机器建议单独还是多台部署？

监控机建议单独部署。建议 CPU 8 core，内存 16 GB 以上，硬盘 500 GB 以上。

14.1.3.3 有一部分监控信息显示不出来？

查看访问监控的机器时间跟集群内机器的时间差，如果比较大，更正时间后即可显示正常。

14.1.3.4 supervise/svc/svstat 服务具体起什么作用？

- supervise 守护进程
- svc 启停服务
- svstat 查看进程状态

14.1.3.5 inventory.ini 变量参数解读

变量	含义
cluster_name	集群名称，可调整
tidb_version	TiDB 版本
deployment_method	部署方式，默认为 binary，可选 docker
process_supervision	进程监管方式，默认为 systemd，可选 supervise

变量	含义
timezone	修改部署目标机器时区，默认为 Asia/Shanghai, 可调整，与 set_timezone 变量结合使用
set_timezone	默认为 True，即修改部署目标机器时区，关闭可修改为 False
enable_elk	目前不支持，请忽略
enable_firewalld	开启防火墙，默认不开启

变量	含义
enable_ntpd	检测部署目标机器 NTP 服务，默认为 True，请勿关闭
machine_benchmark	检测部署目标机器磁盘 IOPS，默认为 True，请勿关闭
set_hostname	根据 IP 修改部署目标机器主机名，默认为 False

变量	含义
enable_binlog	是否部署 pump 并开启 bin-log，默认为 False，依赖 Kafka 集群，参见 zookeeper_addrs 变量
zookeeper_addrs	Kafka 集群的 zookeeper 地址
enable_slow_query_log	TiDB 慢查询日志记录到单独文件 ({{ deploy_dir }}/log/tidb_slow_query.log)，默认为 False，记录到 tidb 日志

变量	含义
deploy_without_tidb	KV 模式，不部署 TiDB 服务，仅部署 PD、TiKV 及监控服务，请将 inventory.ini 文件中 tidb_servers 主机组 IP 设置为空。

14.1.3.6 如何单独记录 TiDB 中的慢查询日志，如何定位慢查询 SQL？

1. TiDB 中，对慢查询的定义在 TiDB 的配置文件中。`slow-threshold: 300`，这个参数是配置慢查询记录阈值的，单位是 ms。
2. 如果出现了慢查询，可以从 Grafana 监控定位到出现慢查询的 tidb-server 以及时间点，然后在对应节点查找日志中记录的 SQL 信息。
3. 除了日志，还可以通过 `admin show slow` 命令查看，详情可参考admin show slow 命令。

14.1.3.7 首次部署 TiDB 集群时，没有配置 tikv 的 Label 信息，在后续如何添加配置 Label？

TiDB 的 Label 设置是与集群的部署架构相关的，是集群部署中的重要内容，是 PD 进行全局管理和调度的依据。如果集群在初期部署过程中没有设置 Label，需要在后期对部署结构进行调整，就需要手动通过 PD 的管理工具 pd-ctl 来添加 location-labels 信息，例如：`config set location-labels "zone,rack,host"`（根据实际的 label 层级名字配置）。

pd-ctl 的使用参考 PD Control 使用说明。

14.1.3.8 为什么测试磁盘的 dd 命令用 oflag=direct 这个选项？

Direct 模式就是把写入请求直接封装成 I/O 指令发到磁盘，这样是为了绕开文件系统的缓存，可以直接测试磁盘的真实的 I/O 读写能力。

14.1.3.9 如何用 fio 命令测试 TiKV 实例的磁盘性能？

- 随机读测试：

```
./fio -ioengine=psync -bs=32k -fdatasync=1 -thread -rw=randread -size=10G -filename=
    ↪ fio_randread_test.txt -name='fio randread test' -iodepth=4 -runtime=60 -numjobs=4 -
    ↪ group_reporting --output-format=json --output=fio_randread_result.json
```

- 顺序写和随机读混合测试：

```
./fio -ioengine=psync -bs=32k -fdatasync=1 -thread -rw=randrw -percentage_random=100,0 -size
    ↪ =10G -filename=fio_randread_write_test.txt -name='fio mixed randread and sequential
    ↪ write test' -iodepth=4 -runtime=60 -numjobs=4 -group_reporting --output-format=json
    ↪ --output=fio_randread_write_test.json
```

14.2 迁移常见问题

本文介绍 TiDB 数据迁移中的常见问题。

如果要查看迁移相关工具的常见问题，请参考以下链接：

- Backup & Restore 常见问题
- TiDB Binlog 常见问题
- TiDB Lightning 常见问题
- Data Migration 常见问题
- TiCDC 常见问题和故障处理

14.2.1 全量数据导出导入

14.2.1.1 如何将一个运行在 MySQL 上的应用迁移到 TiDB 上？

TiDB 支持绝大多数 MySQL 语法，一般不需要修改代码。

14.2.1.2 导入导出速度慢，各组件日志中出现大量重试、EOF 错误并且没有其他错误

在没有其他逻辑出错的情况下，重试、EOF 可能是由网络问题引起的，建议首先使用相关工具排查网络连通状况。以下示例使用 iperf 进行排查：

- 在出现重试、EOF 错误的服务器端节点执行以下命令：

```
iperf3 -s
```

- 在出现重试、EOF 错误的客户端节点执行以下命令：

```
iperf3 -c <server-IP>
```

下面是一个网络连接良好的客户端节点的输出：

```
$ iperf3 -c 192.168.196.58
Connecting to host 192.168.196.58, port 5201
[  5] local 192.168.196.150 port 55397 connected to 192.168.196.58 port 5201
[ ID] Interval           Transfer     Bitrate
[  5]   0.00-1.00   sec  18.0 MBytes   150 Mbits/sec
[  5]   1.00-2.00   sec  20.8 MBytes   175 Mbits/sec
[  5]   2.00-3.00   sec  18.2 MBytes   153 Mbits/sec
[  5]   3.00-4.00   sec  22.5 MBytes   188 Mbits/sec
[  5]   4.00-5.00   sec  22.4 MBytes   188 Mbits/sec
[  5]   5.00-6.00   sec  22.8 MBytes   191 Mbits/sec
[  5]   6.00-7.00   sec  20.8 MBytes   174 Mbits/sec
[  5]   7.00-8.00   sec  20.1 MBytes   168 Mbits/sec
[  5]   8.00-9.00   sec  20.8 MBytes   175 Mbits/sec
[  5]   9.00-10.00  sec  21.8 MBytes   183 Mbits/sec
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bitrate
[  5]   0.00-10.00  sec   208 MBytes   175 Mbits/sec                  sender
[  5]   0.00-10.00  sec   208 MBytes   174 Mbits/sec                  receiver

iperf Done.
```

如果输出显示网络带宽较低、带宽波动大，各组件日志中就可能出现大量重试、EOF 错误。此时你需要咨询网络服务供应商以提升网络质量。

如果输出的各指标良好，请尝试更新各组件版本。如果更新后仍无法解决问题，请移步 AskTUG 论坛寻求帮助。

14.2.1.3 不小心把 MySQL 的 user 表导入到 TiDB 了，或者忘记密码，无法登录，如何处理？

重启 TiDB 服务，配置文件中增加 `-skip-grant-table=true` 参数，无密码登录集群后，可以根据情况重建用户，或者重建 mysql.user 表，具体表结构搜索官网。

14.2.1.4 如何导出 TiDB 数据？

你可以通过以下方式导出 TiDB 数据：

- 参考 MySQL 使用 mysqldump 导出某个表的部分数据，使用 mysqldump 加 where 条件导出。
- 使用 MySQL client 将 select 的结果输出到一个文件。

14.2.1.5 如何从 DB2、 Oracle 数据库迁移到 TiDB？

DB2、 Oracle 到 TiDB 数据迁移（增量 + 全量），通常做法有：

- 使用 Oracle 官方迁移工具，如 OGG、 Gateway（透明网关）、 CDC (Change Data Capture)。
- 自研数据导出导入程序实现。
- 导出 (Spool) 成文本文件，然后通过 Load infile 进行导入。
- 使用第三方数据迁移工具。

目前看来 OGG 最为合适。

14.2.1.6 用 Sqoop 批量写入 TiDB 数据，虽然配置了 `--batch` 选项，但还是会遇到 `java.sql.BatchUpdateExecption:statement count 5001 exceeds the transaction limitation` 的错误，该如何解决？

- 在 Sqoop 中，`--batch` 是指每个批次提交 100 条 statement，但是默认每个 statement 包含 100 条 SQL 语句，所以此时 100 * 100 = 10000 条 SQL 语句，超出了 TiDB 的事务限制 5000 条，可以增加选项 `-Dsqoop.export.records.per.statement=10` 来解决这个问题，完整的用法如下：

```
sqoop export \
    -Dsqoop.export.records.per.statement=10 \
    --connect jdbc:mysql://mysql.example.com/sqoop \
    --username sqoop ${user} \
    --password ${passwd} \
    --table ${tab_name} \
    --export-dir ${dir} \
    --batch
```

- 也可以选择增大 tidb 的单个事物语句数量限制，不过这个会导致内存上涨。

14.2.1.7 Dumpling 导出时引发上游数据库 OOM 或报错 “磁盘空间不足”

该问题可能有如下原因：

- 数据库主键分布不均匀，例如启用了 SHARD_ROW_ID_BITS
- 上游数据库为 TiDB，导出表是分区表

在上述情况下，Dumpling 划分导出子范围时，会划分出过大的子范围，从而向上游发送结果过大的查询。请联系 AskTUG 社区专家获取实验版本的 Dumpling。

14.2.1.8 TiDB 有像 Oracle 那样的 Flashback Query 功能么，DDL 支持么？

有，也支持 DDL。详细参考 TiDB 历史数据回溯。

14.2.2 在线数据同步

14.2.2.1 有没有现成的同步方案，可以将数据同步到 Hbase、 Elasticsearh 等其他存储？

没有，目前依赖程序自行实现。

14.2.3 业务流量迁入

14.2.3.1 如何快速迁移业务流量？

我们建议通过TiDB Data Migration 进行 MySQL -> TiDB 的业务数据的迁移；业务读写可以按照需求分阶段通过修改网络配置进行流量迁移，建议 DB 上层部署一个稳定的网络 LB（HAproxy、LVS、F5、DNS 等），这样直接修改网络配置就能实现无缝流量迁移。

14.2.3.2 TiDB 总读写流量有限制吗？

TiDB 读流量可以通过增加 TiDB server 进行扩展，总读容量无限制，写流量可以通过增加 TiKV 节点进行扩容，基本上写容量也没有限制。

14.2.3.3 Transaction too large 是什么原因，怎么解决？

TiDB 限制了单条 KV entry 不超过 6MB，可以修改配置文件中的`txn-entry-size-limit` 配置项进行调整，最大可以修改到 120MB。

分布式事务要做两阶段提交，而且底层还需要做 Raft 复制。如果一个事务非常大，提交过程会非常慢，事务写冲突概率会增加，而且事务失败后回滚会导致不必要的性能开销。所以我们设置了 key-value entry 的总大小默认不超过 100MB。如果业务需要使用大事务，可以修改配置文件中的 `txn-total-size-limit` 配置项进行调整，最大可以修改到 10G。实际的大小限制还受机器的物理内存影响。

在 Google 的 Cloud Spanner 上面，也有类似的限制。

14.2.3.4 如何批量导入？

导入数据的时候，可以分批插入，每批最好不要超过 1w 行。

14.2.3.5 TiDB 中删除数据后会立即释放空间吗？

DELETE，TRUNCATE 和 DROP 都不会立即释放空间。对于 TRUNCATE 和 DROP 操作，在达到 TiDB 的 GC (garbage collection) 时间后（默认 10 分钟），TiDB 的 GC 机制会删除数据并释放空间。对于 DELETE 操作 TiDB 的 GC 机制会删除数据，但不会释放空间，而是当后续数据写入 RocksDB 且进行 compact 时对空间重新利用。

14.2.3.6 Load 数据时可以对目标表执行 DDL 操作吗？

不可以，加载数据期间不能对目标表执行任何 DDL 操作，这会导致数据加载失败。

14.2.3.7 TiDB 是否支持 replace into 语法？

支持，但是 load data 不支持 replace into 语法。

14.2.3.8 数据删除后查询速度为何会变慢？

大量删除数据后，会有很多无用的 key 存在，影响查询效率。目前正在开发 Region Merge 功能，完善之后可以解决这个问题，具体看参考最佳实践中的删除数据部分。

14.2.3.9 数据删除最高效最快的方式？

在删除大量数据的时候，建议使用 `Delete from t where xx limit 5000`（xx 建议在满足业务过滤逻辑下，尽量加上强过滤索引列或者直接使用主键选定范围，如 `id >= 5000*n+m and id <= 5000*(n+1)+m` 这样的方案，通过循环来删除，用 `Affected Rows == 0` 作为循环结束条件，这样避免遇到事务大小的限制。如果一次删除的数据量非常大，这种循环的方式会越来越慢，因为每次删除都是从前向后遍历，前面的删除之后，短时间内会残留不少删除标记（后续会被 GC 掉），影响后面的 Delete 语句。如果有可能，建议把 Where 条件细化。可以参考官网最佳实践。

14.2.3.10 TiDB 如何提高数据加载速度？

主要有两个方面：

- 目前已开发分布式导入工具TiDB Lightning，需要注意的是数据导入过程中为了性能考虑，不会执行完整的事务流程，所以没办法保证导入过程中正在导入的数据的 ACID 约束，只能保证整个导入过程结束以后导入数据的 ACID 约束。因此适用场景主要为新数据的导入（比如新的表或者新的索引），或者是全量的备份恢复（先 Truncate 原表再导入）。
- TiDB 的数据加载与磁盘以及整体集群状态相关，加载数据时应关注该主机的磁盘利用率，TiClient Error/Backoff/Thread CPU 等相关 metric，可以分析相应瓶颈。

14.3 升级与升级后常见问题

本文介绍 TiDB 升级与升级后的常见问题与解决办法。

14.3.1 升级常见问题

本小节列出了 TiDB 升级相关的常见问题与解决办法。

14.3.1.1 滚动升级有那些影响？

滚动升级 TiDB 服务，滚动升级期间不影响业务运行。需要配置最小集群拓扑 (TiDB * 2、PD * 3、TiKV * 3)，如果集群环境中有 Pump 和 Drainer 服务，建议先停止 Drainer，然后滚动升级（升级 TiDB 时会升级 Pump）。

14.3.1.2 集群在执行 DDL 请求期间可以进行升级操作吗？

集群中有 DDL 语句正在被执行时（通常为 ADD INDEX 和列类型变更等耗时较久的 DDL 语句），请勿进行升级操作。在升级前，建议使用 `ADMIN SHOW DDL` 命令查看集群中是否有正在进行的 DDL Job。如需升级，请等待 DDL 执行完成或使用 `ADMIN CANCEL DDL` 命令取消该 DDL Job 后再进行升级。

另外，在升级 TiDB 集群的过程中，请勿执行 DDL 语句，否则可能会出现行为未定义的问题。

14.3.1.3 Binary 如何升级？

Binary 不是建议的安装方式，对升级支持也不友好，建议换成TiUP 部署。

14.3.2 升级后常见问题

本小节列出了一些升级后可能会遇到的问题与解决办法。

14.3.2.1 执行 DDL 操作时遇到的字符集 (charset) 问题

TiDB 在 v2.1.0 以及之前版本（包括 v2.0 所有版本）中，默认字符集是 UTF8。从 v2.1.1 开始，默认字符集变更为 UTF8MB4。如果在 v2.1.0 及之前版本中，建表时显式指定了 table 的 charset 为 UTF8，那么升级到 v2.1.1 之后，执行 DDL 操作可能会失败。

要避免该问题，需注意以下两个要点：

- 在 v2.1.3 之前，TiDB 不支持修改 column 的 charset。所以，执行 DDL 操作时，新 column 的 charset 需要和旧 column 的 charset 保持一致。
- 在 v2.1.3 之前，即使 column 的 charset 和 table 的 charset 不一样，`show create table` 也不会显示 column 的 charset，但可以通过 HTTP API 获取 table 的元信息来查看 column 的 charset，下文提供了示例。

14.3.2.1.1 `unsupported modify column charset utf8mb4 not match origin utf8`

- 升级前：v2.1.0 及之前版本

```
create table t(a varchar(10)) charset=utf8;
```

```
Query OK, 0 rows affected
Time: 0.106s
```

```
show create table t
```

```
+-------+-------------------------------------------------------+
| Table | Create Table                                          |
+-------+-------------------------------------------------------+
| t     | CREATE TABLE `t` (                                    |
|       |   `a` varchar(10) DEFAULT NULL                        |
|       | ) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_bin |
+-------+-------------------------------------------------------+
1 row in set
Time: 0.006s
```

- 升级后：v2.1.1、v2.1.2 会出现下面的问题，v2.1.3 以及之后版本不会出现下面的问题。

```
alter table t change column a a varchar(20);
```

```
ERROR 1105 (HY000): unsupported modify column charset utf8mb4 not match origin utf8
```

解决方案：显式指定 column charset，保持和原来的 charset 一致即可。

```
alter table t change column a a varchar(22) character set utf8;
```

- 根据要点 1，此处如果不指定 column 的 charset，会用默认的 UTF8MB4，所以需要指定 column charset 保持和原来一致。
- 根据要点 2，用 HTTP API 获取 table 元信息，然后根据 column 名字和 Charset 关键字搜索即可找到 column 的 charset。

```
curl "http://$IP:10080/schema/test/t" | python -m json.tool
```

这里用了 python 的格式化 json 的工具，也可以不加，此处只是为了方便注释。

```
{
    "ShardRowIDBits": 0,
    "auto_inc_id": 0,
    "charset": "utf8",   # table 的 charset
    "collate": "",
    "cols": [            # 从这里开始列举 column 的相关信息
        {
            ...
            "id": 1,
            "name": {
                "L": "a",
                "O": "a"   # column 的名字
            },
            "offset": 0,
            "origin_default": null,
            "state": 5,
            "type": {
                "Charset": "utf8",   # column a 的 charset
                "Collate": "utf8_bin",
                "Decimal": 0,
                "Elems": null,
                "Flag": 0,
                "Flen": 10,
                "Tp": 15
            }
        }
    ],
    ...
}
```

14.3.2.1.2 unsupported modify charset from utf8mb4 to utf8

- 升级前：v2.1.1，v2.1.2

```
create table t(a varchar(10)) charset=utf8;
```

```
Query OK, 0 rows affected
Time: 0.109s
```

```
show create table t;
```

```
+-------+-------------------------------------------------------+
| Table | Create Table                                          |
+-------+-------------------------------------------------------+
| t     | CREATE TABLE `t` (                                    |
|       |   `a` varchar(10) DEFAULT NULL                        |
|       | ) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_bin |
+-------+-------------------------------------------------------+
```

上面 show create table 只显示出了 table 的 charset，但其实 column 的 charset 是 UTF8MB4，这可以通过 HTTP API 获取 schema 来确认。这是一个 bug，即此处建表时 column 的 charset 应该要和 table 保持一致为 UTF8，该问题在 v2.1.3 中已经修复。

- 升级后：v2.1.3 及之后版本

```
show create table t;
```

```
+-------+------------------------------------------------------------------------+
| Table | Create Table                                                           |
+-------+------------------------------------------------------------------------+
| t     | CREATE TABLE `t` (                                                     |
|       |   `a` varchar(10) CHARSET utf8mb4 COLLATE utf8mb4_bin DEFAULT NULL |
|       | ) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_bin                  |
+-------+------------------------------------------------------------------------+
1 row in set
Time: 0.007s
```

```
alter table t change column a a varchar(20);
```

```
ERROR 1105 (HY000): unsupported modify charset from utf8mb4 to utf8
```

解决方案：

- 因为在 v2.1.3 之后，TiDB 支持修改 column 和 table 的 charset，所以这里推荐修改 table 的 charset 为 UTF8MB4。

```
alter table t convert to character set utf8mb4;
```

- 也可以像问题 1 一样指定 column 的 charset，保持和 column 原来的 charset (UTF8MB4) 一致即可。

```
alter table t change column a a varchar(20) character set utf8mb4;
```

14.3.2.1.3 ERROR 1366 (HY000): incorrect utf8 value f09f8c80( ) for column a

TiDB 在 v2.1.1 及之前版本中，如果 charset 是 UTF8，没有对 4-byte 的插入数据进行 UTF8 Unicode encoding 检查。在 v2.1.2 及之后版本中，添加了该检查。

- 升级前：v2.1.1 及之前版本

```
create table t(a varchar(100) charset utf8);
```

```
Query OK, 0 rows affected
```

```
insert t values (unhex('f09f8c80'));
```

```
Query OK, 1 row affected
```

- 升级后：v2.1.2 及之后版本

```
insert t values (unhex('f09f8c80'));
```

```
ERROR 1366 (HY000): incorrect utf8 value f09f8c80( ) for column a
```

解决方案：

- v2.1.2 版本：该版本不支持修改 column charset，所以只能跳过 UTF8 的检查。

```
set @@session.tidb_skip_utf8_check=1;
```

```
Query OK, 0 rows affected
```

```
insert t values (unhex('f09f8c80'));
```

```
Query OK, 1 row affected
```

- v2.1.3 及之后版本：建议修改 column 的 charset 为 UTF8MB4。或者也可以设置 `tidb_skip_utf8_check` 变量跳过 UTF8 的检查。如果跳过 UTF8 的检查，在需要将数据从 TiDB 同步回 MySQL 的时候，可能会失败，因为 MySQL 会执行该检查。

```
alter table t change column a a varchar(100) character set utf8mb4;
```

```
Query OK, 0 rows affected
```

```
insert t values (unhex('f09f8c80'));
```

```
Query OK, 1 row affected
```

关于 tidb_skip_utf8_check 变量，具体来说是指跳过 UTF8 和 UTF8MB4 类型对数据的合法性检查。如果跳过这个检查，在需要将数据从 TiDB 同步回 MySQL 的时候，可能会失败，因为 MySQL 执行该检查。如果只想跳过 UTF8 类型的检查，可以设置 tidb_check_mb4_value_in_utf8 变量。

tidb_check_mb4_value_in_utf8 在 v2.1.3 版本加入 config.toml 文件，可以修改配置文件里面的 check-mb4-value-in-utf8 后重启集群生效。

tidb_check_mb4_value_in_utf8 在 v2.1.5 版本开始可以用 HTTP API 来设置，也可以用 session 变量来设置。

- HTTP API（HTTP API 只在单台服务器上生效）
  * 执行下列命令启用 HTTP API：

```
curl -X POST -d "check_mb4_value_in_utf8=1" http://{TiDBIP}:10080/settings
```

  * 执行下列命令禁用 HTTP API：

```
curl -X POST -d "check_mb4_value_in_utf8=0" http://{TiDBIP}:10080/settings
```

- Session 变量
  * 执行下列命令启用 Session 变量：

```
set @@session.tidb_check_mb4_value_in_utf8 = 1;
```

  * 执行下列命令禁用 Session 变量：

```
set @@session.tidb_check_mb4_value_in_utf8 = 0;
```

- v2.1.7 及之后版本，如果对表和 column 的字符集没有严格要求为 UTF8，也不想修改客户端代码去跳过 UTF8 检查或者手动修改 column 的 charset，可以在配置文件中把 treat-old-version-utf8-as-utf8mb4 打开。该配置的作用是自动把 v2.1.7 版本之前创建的旧版本的表和 column 的 UTF8 字符集转成 UTF8MB4。这个转换是在 TiDB load schema 时在内存中将 UTF8 转成 UTF8MB4，不会对实际存储的数据做任何修改。在配置文件中关闭 treat-old-version-utf8-as-utf8mb4 并重启 TiDB 后，以前字符集为 UTF8 的表和 column 的字符集仍然还是 UTF8。

> **注意：**
>
> treat-old-version-utf8-as-utf8mb4 参数默认打开，如果客户端强制需要用 UTF8 而不用 UTF8MB4，需要在配置文件中关闭。

14.4 监控 FAQ

本文介绍在监控 TiDB 集群时的常见问题、原因及解决方法。

- Prometheus 监控框架详情可见TiDB 监控框架概述。
- 监控指标解读详细参考重要监控指标详解。

14.4.1 目前的监控使用方式及主要监控指标，有没有更好看的监控？

TiDB 使用 Prometheus + Grafana 组成 TiDB 数据库系统的监控系统。用户在 Grafana 上通过 dashboard 可以监控到 TiDB 的各类运行指标，包括

- 系统资源的监控指标
- 客户端连接与 SQL 运行的指标
- 内部通信和 Region 调度的指标

通过这些指标，可以让数据库管理员更好的了解到系统的运行状态，运行瓶颈等内容。在监控指标的过程中，我们按照 TiDB 不同的模块，分别列出了各个模块重要的指标项，一般用户只需要关注这些常见的指标项。具体指标请参见官方文档。

14.4.2 Prometheus 监控数据默认 15 天自动清除一次，可以自己设定成 2 个月或者手动删除吗？

可以的，在 Prometheus 启动的机器上，找到启动脚本，然后修改启动参数，然后重启 Prometheus 生效。

```
--storage.tsdb.retention="60d"
```

14.4.3 Region Health 监控项

TiDB-2.0 版本中，PD metric 监控页面中，对 Region 健康度进行了监控，其中 Region Health 监控项是对所有 Region 副本状况的一些统计。其中 miss 是缺副本，extra 是多副本。同时也增加了按 Label 统计的隔离级别，level-1 表示这些 Region 的副本在第一级 Label 下是物理隔离的，没有配置 location label 时所有 Region 都在 level-0。

14.4.4 Statement Count 监控项中的 selectsimplefull 是什么意思？

代表全表扫，但是可能是很小的系统表。

14.4.5 监控上的 QPS 和 Statement OPS 有什么区别？

QPS 会统计执行的所有 SQL 命令，包括 use database、 load data、 begin、 commit、 set、 show、 insert、 select 等。

Statement OPS 只统计 select、 update、 insert 等业务相关的，所以 Statement OPS 的统计和业务比较相符。

14.5 集群管理 FAQ

本文介绍管理 TiDB 集群时的常见问题、原因及解决方法。

14.5.1 集群日常管理

14.5.1.1 TiDB 如何登录？

和 MySQL 登录方式一样，可以按照下面例子进行登录。

```
mysql -h 127.0.0.1 -uroot -P4000
```

14.5.1.2 TiDB 如何修改数据库系统变量？

和 MySQL 一样，TiDB 也分为静态参数和固态参数，静态参数可以直接通过`set global xxx = n`的方式进行修改，不过新参数值只限于该实例生命周期有效。

14.5.1.3 TiDB (TiKV) 有哪些数据目录？

默认在`--data-dir` 目录下，其中包括 backup、db、raft、snap 四个目录，分别存储备份、数据、raft 数据及镜像数据。

14.5.1.4 TiDB 有哪些系统表？

和 MySQL 类似，TiDB 中也有系统表，用于存放数据库运行时所需信息，具体信息参考 TiDB 系统表文档。

14.5.1.5 TiDB 各节点服务器下是否有日志文件，如何管理？

默认情况下各节点服务器会在日志中输出标准错误，如果启动的时候通过 `--log-file` 参数指定了日志文件，那么日志会输出到指定的文件中，并且按天做 rotation。

14.5.1.6 TiDB、TiKV、PD 节点的各类文件存放在哪里？

如需快速了解 TiDB 节点、TiKV 节点、PD 节点的配置文件、数据文件及日志文件的相关介绍与其存放位置，建议观看下面的培训视频（时长 9 分钟）。

14.5.1.7 如何规范停止 TiDB？

可以直接 kill 掉所有服务。如果使用 kill 命令，TiDB 的组件会做 graceful 的 shutdown。

14.5.1.8 TiDB 里面可以执行 kill 命令吗？

- 可以 kill DML 语句，首先使用 `show processlist`，找到对应 session 的 id，然后执行 `kill tidb [session id]`。
- 可以 kill DDL 语句，首先使用 `admin show ddl jobs`，查找需要 kill 的 DDL job ID，然后执行 `admin cancel ddl jobs 'job_id' [, 'job_id'] ...`。具体可以参考 admin 操作。

14.5.1.9 TiDB 是否支持会话超时？

TiDB 暂不支持数据库层面的会话超时，目前想要实现超时，在没 LB (Load Balancing) 的时候，需要应用侧记录发起的 Session 的 ID，通过应用自定义超时，超时以后需要到发起 Query 的节点上用 `kill tidb [session id]` 来杀掉 SQL。目前建议使用应用程序来实现会话超时，当达到超时时间，应用层就会抛出异常继续执行后续的程序段。

14.5.1.10 TiDB 生产环境的版本管理策略是怎么样的？如何尽可能避免频繁升级？

TiDB 版本目前逐步标准化，每次 Release 都包含详细的 Change log，版本功能变化详情，生产环境是否有必要升级取决于业务系统，建议升级之前详细了解前后版本的功能差异。

版本号说明参考：Release Version: `v1.0.3-1-ga80e796`

- v1.0.3 表示 GA 标准版
- 1 表示该版本 commit 1 次
- ga80e796 代表版本的 git-hash

14.5.1.11 分不清 TiDB master 版本之间的区别，应该怎么办？

TiDB 目前社区非常活跃，在 1.0 GA 版本发布后，还在不断的优化和修改 BUG，因此 TiDB 的版本更新周期比较快，会不定期有新版本发布，请关注我们的新版本发布官方网站。此外 TiDB 安装推荐使用 TiUP 进行安装。此外，在 TiDB 1.0 GA 版本后，对 TiDB 的版本号进行了统一管理，TiDB 的版本可以通过以下两种方式进行查看：

- 通过 `select tidb_version()` 进行查看
- 通过执行 `tidb-server -V` 进行查看

14.5.1.12 有没有图形化部署 TiDB 的工具？

暂时没有。

14.5.1.13 TiDB 如何进行水平扩展？

当您的业务不断增长时，数据库可能会面临三方面瓶颈，第一是存储空间，第二是计算资源，第三是读写容量，这时可以对 TiDB 集群做水平扩展。

- 如果是存储资源不够，可以通过添加 TiKV Server 节点来解决，新节点启动后，PD 会自动将其他节点的部分数据迁移过去，无需人工介入。
- 如果是计算资源不够，可以查看 TiDB Server 和 TiKV Server 节点的 CPU 消耗情况，再考虑添加 TiDB Server 节点或者是 TiKV Server 节点来解决，如添加 TiDB Server 节点，将其添加到前端 Load Balancer 配置之中即可。
- 如果是容量跟不上，一般可以考虑同时增加 TiDB Server 和 TiKV Server 节点。

14.5.1.14 Percolator 用了分布式锁，crash 的客户端会保持锁，会造成锁没有 release？

详细可参考 Percolator 和 TiDB 事务算法。

14.5.1.15 TiDB 为什么选用 gRPC 而不选用 Thrift，是因为 Google 在用吗？

不只是因为 Google 在用，有一些比较好的特性我们需要，比如流控、加密还有 Streaming。

14.5.1.16 like(bindo.customers.name, jason%, 92) 这个 92 代表什么？

那个是转义字符，默认是 (ASCII 92)。

14.5.1.17 为什么 `information_schema.tables.data_length` 记录的大小和 TiKV 监控面板上的 store size 不一样？

这是因为两者计算的角度不一样。`information_schema.tables.data_length` 是通过统计信息（平均每行的大小）得到的估算值。TiKV 监控面板上的 store size 是单个 TiKV 实例的数据文件（RocksDB 的 SST 文件）的大小总和。由于多版本和 TiKV 会压缩数据，所以两者显示的大小不一样。

14.5.1.18 为什么事务没有使用异步提交或一阶段提交？

在以下情况中，即使通过系统变量开启了异步提交和一阶段提交，TiDB 也不会使用这些特性：

- 如果开启了 TiDB Binlog，受 TiDB Binlog 的实现原理限制，TiDB 不会使用异步提交或一阶段提交特性。
- TiDB 只在事务写入不超过 256 个键值对，以及所有键值对里键的总大小不超过 4 KB 时，才会使用异步提交或一阶段提交特性。这是因为对于写入量大的事务，异步提交不能明显提升执行性能。

14.5.2 PD 管理

14.5.2.1 访问 PD 报错：TiKV cluster is not bootstrapped

PD 的大部分 API 需要在初始化 TiKV 集群以后才能使用，如果在部署新集群的时候只启动了 PD，还没有启动 TiKV，这时候访问 PD 就会报这个错误。遇到这个错误应该先把要部署的 TiKV 启动起来，TiKV 会自动完成初始化工作，然后就可以正常访问 PD。

14.5.2.2 PD 启动报错：etcd cluster ID mismatch

PD 启动参数中的 `--initial-cluster` 包含了某个不属于该集群的成员。遇到这个错误时请检查各个成员的所属集群，剔除错误的成员后即可正常启动。

14.5.2.3 PD 能容忍的时间同步误差是多少？

理论上，时间同步误差越小越好。PD 可容忍任意时长的误差，但是，时间同步误差越大意味着 PD 分配的时间戳与真实的物理时间相差越大，这个差距会影响读历史版本等功能。

14.5.2.4 Client 连接是如何寻找 PD 的？

Client 连接只能通过 TiDB 访问集群，TiDB 负责连接 PD 与 TiKV，PD 与 TiKV 对 Client 透明。当 TiDB 连接任意一台 PD 的时候，PD 会告知 TiDB 当前的 leader 是谁，如果此台 PD 不是 leader，TiDB 将会重新连接至 leader PD。

14.5.2.5 TiKV 节点 (Store) 各状态 (Up, Disconnect, Offline, Down, Tombstone) 之间的关系是什么？

使用 `pd-ctl` 可以查看 TiKV 节点的状态信息。如需查看各个状态之间的关系，请参考 TiKV Store 状态之间的关系。

14.5.2.6 PD 参数中 leader-schedule-limit 和 region-schedule-limit 调度有什么区别？

- leader-schedule-limit 调度是用来均衡不同 TiKV 的 leader 数，影响处理查询的负载。
- region-schedule-limit 调度是均衡不同 TiKV 的副本数，影响不同节点的数据量。

14.5.2.7 每个 region 的 replica 数量可配置吗？调整的方法是？

可以，目前只能调整全局的 replica 数量。首次启动时 PD 会读配置文件 (conf/pd.yml)，使用其中的 max-replicas 配置，之后修改需要使用 pd-ctl 配置命令 `config set max-replicas $num`，配置后可通过 `config show all` 来查看已生效的配置。调整的时候，不会影响业务，会在后台添加，注意总 TiKV 实例数总是要大于等于设置的副本数，例如 3 副本需要至少 3 个 TiKV。增加副本数量之前需要预估额外的存储需求。pd-ctl 的详细用法可参考 PD Control 使用说明。

14.5.2.8 缺少命令行集群管理工具，整个集群的健康度当前是否正常，不好确认？

可以通过 pd-ctl 等工具来判断集群大概的状态，详细的集群状态还是需要通过监控来确认。

14.5.2.9 集群下线节点后，怎么删除老集群节点监控信息？

下线节点一般指 TiKV 节点通过 pd-ctl 或者监控判断节点是否下线完成。节点下线完成后，手动停止下线节点上相关的服务。从 Prometheus 配置文件中删除对应节点的 node_exporter 信息。

14.5.3 TiDB server 管理

14.5.3.1 TiDB 的 lease 参数应该如何设置？

启动 TiDB Server 时，需要通过命令行参数设置 lease 参数 (`--lease=60`)，其值会影响 DDL 的速度（只会影响当前执行 DDL 的 session，其他的 session 不会受影响）。在测试阶段，lease 的值可以设为 1s，加快测试进度；在生产环境下，我们推荐这个值设为分钟级（一般可以设为 60），这样可以保证 DDL 操作的安全。

14.5.3.2 DDL 在正常情况下的耗时是多少？

一般情况下处理一个 DDL 操作（之前没有其他 DDL 操作在处理）的耗时基本可以分如下为三种：

- add index 操作，且此操作对应表数据行数比较少，耗时约为 3s。
- add index 操作，且此操作对应表数据行数比较多，耗时具体由表中数据行数和当时 QPS 情况定（add index 操作优先级比一般 SQL 低）。
- 其他 DDL 操作耗时约为 1s。

此外，如果接收 DDL 请求的 TiDB 和 DDL owner 所处的 TiDB 是一台，那么上面列举的第一和第三种可能的耗时应该在几十到几百毫秒。

14.5.3.3 为什么有的时候执行 DDL 会很慢？

可能原因如下：

- 多个 DDL 语句一起执行的时候，后面的几个 DDL 语句会比较慢。原因是当前 TiDB 集群中 DDL 操作是串行执行的。
- 在正常集群启动后，第一个 DDL 操作的执行时间可能会比较久，一般在 30s 左右，这个原因是刚启动时 TiDB 在竞选处理 DDL 的 leader。
- 由于停 TiDB 时不能与 PD 正常通信（包括停电情况）或者用 `kill -9` 指令停 TiDB 导致 TiDB 没有及时从 PD 清理注册数据，那么会影响 TiDB 启动后 10min 内的 DDL 语句处理时间。这段时间内运行 DDL 语句时，每个 DDL 状态变化都需要等待 2 * lease（默认 lease = 45s）。
- 当集群中某个 TiDB 与 PD 之间发生通信问题，即 TiDB 不能从 PD 及时获取或更新版本信息，那么这时候 DDL 操作的每个状态处理需要等待 2 * lease。

14.5.3.4 TiDB 可以使用 S3 作为后端存储吗？

不可以，目前 TiDB 只支持分布式存储引擎和 GolevelDB/RocksDB/BoltDB 引擎。

14.5.3.5 Information_schema 能否支持更多真实信息？

Information_schema 库里面的表主要是为了兼容 MySQL 而存在，有些第三方软件会查询里面的信息。在目前 TiDB 的实现中，里面大部分只是一些空表。后续随着 TiDB 的升级，会提供更多的参数信息。当前 TiDB 支持的 Information_schema 请参考 TiDB 系统数据库说明文档。

14.5.3.6 TiDB Backoff type 主要原因？

TiDB-server 与 TiKV-server 随时进行通信，在进行大量数据操作过程中，会出现 `Server is busy` 或者 `backoff.maxsleep 20000ms` 的日志提示信息，这是由于 TiKV-server 在处理过程中系统比较忙而出现的提示信息，通常这时候可以通过系统资源监控到 TiKV 主机系统资源使用率比较高的情况出现。如果这种情况出现，可以根据资源使用情况进行相应的扩容操作。

14.5.3.7 TiDB TiClient type 主要原因？

TiClient Region Error 该指标描述的是在 TiDB-server 作为客户端通过 KV 接口访问 TiKV-server 进行数据操作过程中，TiDB-server 操作 TiKV-server 中的 Region 数据出现的错误类型与 metric 指标，错误类型包括 not_leader、 stale_epoch。出现这些错误的情况是当 TiDB-server 根据自己的缓存信息去操作 Region leader 数据的时候，Region leader 发生了迁移或者 TiKV 当前的 Region 信息与 TiDB 缓存的路由信息不一致而出现的错误提示。一般这种情况下，TiDB-server 都会自动重新从 PD 获取最新的路由数据，重做之前的操作。

14.5.3.8 TiDB 同时支持的最大并发连接数？

默认情况下，每个 TiDB 服务器的最大连接数没有限制。如有需要，可以在 `config.toml` 文件中设置 `max-server-connections` 来限制最大连接数。如果并发量过大导致响应时间增加，建议通过添加 TiDB 节点进行扩容。

14.5.3.9 如何查看某张表创建的时间？

information_schema 库中的 tables 表里的 create_time 即为表的真实创建时间。

14.5.3.10 TiDB 的日志中 EXPENSIVE_QUERY 是什么意思？

TiDB 在执行 SQL 时，预估出来每个 operator 处理了超过 10000 条数据就认为这条 query 是 expensive query。可以通过修改 tidb-server 配置参数来对这个门限值进行调整，调整后需要重新启动 tidb-server。

14.5.4 TiKV 管理

14.5.4.1 TiKV 集群副本建议配置数量是多少，是不是最小高可用配置（3 个）最好？

如果是测试环境 3 副本足够；在生产环境中，不可让集群副本数低于 3，需根据架构特点、业务系统及恢复能力的需求，适当增加副本数。值得注意的是，副本升高，性能会有下降，但是安全性更高。

14.5.4.2 TiKV 启动报错：cluster ID mismatch

TiKV 本地存储的 cluster ID 和指定的 PD 的 cluster ID 不一致。在部署新的 PD 集群的时候，PD 会随机生成一个 cluster ID，TiKV 第一次初始化的时候会从 PD 获取 cluster ID 存储在本地，下次启动的时候会检查本地的 cluster ID 与 PD 的 cluster ID 是否一致，如果不一致则会报错并退出。出现这个错误一个常见的原因是，用户原先部署了一个集群，后来把 PD 的数据删除了并且重新部署了新的 PD，但是 TiKV 还是使用旧的数据重启连到新的 PD 上，就会报这个错误。

14.5.4.3 TiKV 启动报错：duplicated store address

启动参数中的地址已经被其他的 TiKV 注册在 PD 集群中了。造成该错误的常见情况：TiKV `--data-dir` 指定的路径下没有数据文件夹（删除或移动后没有更新 –data-dir），用之前参数重新启动该 TiKV。请尝试用 pd-ctl 的 store delete 功能，删除之前的 store，然后重新启动 TiKV 即可。

14.5.4.4 TiKV master 和 slave 用的是一样的压缩算法，为什么效果不一样？

目前来看 master 有些文件的压缩率会高一些，这个取决于底层数据的分布和 RocksDB 的实现，数据大小偶尔有些波动是正常的，底层存储引擎会根据需要调整数据。

14.5.4.5 TiKV block cache 有哪些特性？

TiKV 使用了 RocksDB 的 Column Family (CF) 特性，KV 数据最终存储在默认 RocksDB 内部的 default、write、lock 3 个 CF 内。

- default CF 存储的是真正的数据，与其对应的参数位于 `[rocksdb.defaultcf]` 项中。
- write CF 存储的是数据的版本信息 (MVCC)、索引、小表相关的数据，相关的参数位于 `[rocksdb.writecf]` 项中。
- lock CF 存储的是锁信息，系统使用默认参数。
- Raft RocksDB 实例存储 Raft log。default CF 主要存储的是 Raft log，与其对应的参数位于 `[raftdb.defaultcf]` 项中。
- 所有 CF 共享一个 Block-cache，用于缓存数据块，加速 RocksDB 的读取速度。Block-cache 的大小通过参数 `block-cache-size` 控制，`block-cache-size` 越大，能够缓存的热点数据越多，对读取操作越有利，同时占用的系统内存也会越多。
- 每个 CF 有各自的 Write-buffer，大小通过 `write-buffer-size` 控制。

14.5.4.6 TiKV channel full 是什么原因？

- Raftstore 线程太忙，或者因 I/O 而卡住。可以看一下 Raftstore 的 CPU 使用情况。
- TiKV 过忙（CPU、磁盘 I/O 等），请求处理不过来。

14.5.4.7 TiKV 频繁切换 Region leader 是什么原因？

- 网络问题导致节点间通信卡了，查看 Report failures 监控。
- 原主 Leader 的节点卡了，导致没有及时给 Follower 发送消息。
- Raftstore 线程卡了。

14.5.4.8 如果一个节点挂了会影响服务吗？影响会持续多久？

TiDB 使用 Raft 在多个副本之间做数据同步（默认为每个 Region 3 个副本）。当一份备份出现问题时，其他的副本能保证数据的安全。根据 Raft 协议，当某个节点挂掉导致该节点里的 Leader 失效时，在最大 2 * lease time（lease time 是 10 秒）时间后，通过 Raft 协议会很快将一个另外一个节点里的 Follower 选为新的 Region Leader 来提供服务。

14.5.4.9 TiKV 在分别在哪些场景下占用大量 IO、内存、CPU（超过参数配置的多倍）?

在大量写入、读取的场景中会占用大量的磁盘 IO、内存、CPU。在执行很复杂的查询，比如会产生很大中间结果集的情况下，会消耗很多的内存和 CPU 资源。

14.5.4.10 TiKV 是否可以使用 SAS/SATA 盘或者进行 SSD/SAS 混合部署？

不可以使用。TiDB 在进行 OLTP 场景中，数据访问和操作需要高 IO 磁盘的支持。TiDB 作为强一致的分布式数据库，存在一定的写放大，如副本复制、存储底层 Compaction，因此，TiDB 部署的最佳实践中推荐用户使用 NVMe SSD 磁盘作为数据存储磁盘。另外，TiKV 与 PD 不能混合部署。

14.5.4.11 数据表 Key 的 Range 范围划分是在数据接入之前就已经划分好了吗？

不是的，这个和 MySQL 分表规则不一样，需要提前设置好，TiKV 是根据 Region 的大小动态分裂的。

14.5.4.12 Region 是如何进行分裂的？

Region 不是前期划分好的，但确实有 Region 分裂机制。当 Region 的大小超过参数 `region-max-size` 或 `region-max-keys` 的值时，就会触发分裂，分裂后的信息会汇报给 PD。

14.5.4.13 TiKV 是否有类似 MySQL 的 `innodb_flush_log_trx_commit` 参数，来保证提交数据不丢失？

是的。TiKV 单机的存储引擎目前使用两个 RocksDB 实例，其中一个存储 raft-log。TiKV 有个 sync-log 参数，在 ture 的情况下，每次提交都会强制刷盘到 raft-log，如果发生 crash 后，通过 raft-log 进行 KV 数据的恢复。

14.5.4.14 对 WAL 存储有什么推荐的硬件配置，例如 SSD，RAID 级别，RAID 卡 cache 策略，NUMA 设置，文件系统选择，操作系统的 IO 调度策略等？

WAL 属于顺序写，目前我们并没有单独对他进行配置，建议 SSD。RAID 如果允许的话，最好是 RAID 10，RAID 卡 cache、操作系统 I/O 调度目前没有针对性的最佳实践，Linux 7 以上默认配置即可。NUMA 没有特别建议，NUMA 内存分配策略可以尝试使用 `interleave = all`，文件系统建议 ext4。

14.5.4.15 在最严格的 `sync-log = true` 数据可用模式下，写入性能如何？

一般来说，开启 `sync-log` 会让性能损耗 30% 左右。关闭 `sync-log` 时的性能表现，请参见 TiDB Sysbench 性能测试报告。

14.5.4.16 是否可以利用 TiKV 的 Raft + 多副本达到完全的数据可靠，单机存储引擎是否需要最严格模式？

通过使用 Raft 一致性算法，数据在各 TiKV 节点间复制为多副本，以确保某个节点挂掉时数据的安全性。只有当数据已写入超过 50% 的副本时，应用才返回 ACK（三副本中的两副本）。但理论上两个节点也可能同时发生故障，所以除非是对性能要求高于数据安全的场景，一般都强烈推荐开启 sync-log。

另外，还有一种 sync-log 的替代方案，即在 Raft group 中用五个副本而非三个。这将允许两个副本同时发生故障，而仍然能保证数据安全性。

对于单机存储引擎也同样推荐打开 sync-log 模式。否则如果节点宕机可能会丢失最后一次写入数据。

14.5.4.17 使用 Raft 协议，数据写入会有多次网络的 roundtrip，实际写入延迟如何？

理论上，和单机数据库相比，数据写入会多四个网络延迟。

14.5.4.18 有没有类似 MySQL 的 InnoDB Memcached plugin，可以直接使用 KV 接口，可以不需要独立的 Cache？

TiKV 支持单独进行接口调用，理论上也可以起个实例做为 Cache，但 TiDB 最大的价值是分布式关系型数据库，我们原则上不对 TiKV 单独进行支持。

14.5.4.19 Coprocessor 组件的主要作用？

- 减少 TiDB 与 TiKV 之间的数据传输。
- 计算下推，充分利用 TiKV 的分布式计算资源。

14.5.4.20 IO error: No space left on device While appending to file

这是磁盘空间不足导致的，需要加节点或者扩大磁盘空间。

14.5.4.21 为什么 TiKV 容易出现 OOM？

TiKV 的内存占用主要来自于 RocksDB 的 block-cache，默认为系统总内存的 40%。当 TiKV 容易出现 OOM 时，检查 block-cache-size 配置是否过高。还需要注意，当单机部署了多个 TiKV 实例时，需要显式地配置该参数，以防止多个实例占用过多系统内存导致 OOM。

14.5.4.22 TiDB 数据和 RawKV 数据可存储于同一个 TiKV 集群里吗？

不可以。TiDB 数据（或使用其他事务 API 生成的数据）依赖于一种特殊的键值格式，和 RawKV API 数据（或其他基于 RawKV 的服务生成的数据）并不兼容。

14.5.5 TiDB 测试

14.5.5.1 TiDB Sysbench 基准测试结果如何？

很多用户在接触 TiDB 都习惯做一个基准测试或者 TiDB 与 MySQL 的对比测试，官方也做了一个类似测试。我们汇总很多测试结果后，发现虽然测试的数据有一定的偏差，但结论或者方向基本一致，由于 TiDB 与 MySQL 由于架构上的差别非常大，很多方面是很难找到一个基准点，所以官方的建议两点：

- 大家不要用过多精力纠结这类基准测试上，应该更多关注 TiDB 的场景上的区别。
- 大家可以直接参考 TiDB Sysbench 性能测试报告。

14.5.5.2 TiDB 集群容量 QPS 与节点数之间关系如何，和 MySQL 对比如何？

- 在 10 节点内，TiDB 写入能力 (Insert TPS) 和节点数量基本成 40% 线性递增，MySQL 由于是单节点写入，所以不具备写入扩展能力。
- MySQL 读扩容可以通过添加从库进行扩展，但写流量无法扩展，只能通过分库分表，而分库分表有很多问题，具体参考方案虽好，成本先行：数据库 Sharding+Proxy 实践解析。
- TiDB 不管是读流量、还是写流量都可以通过添加节点快速方便的进行扩展。

14.5.5.3 我们的 DBA 测试过 MySQL 性能，单台 TiDB 的性能没有 MySQL 性能那么好？

TiDB 设计的目标就是针对 MySQL 单台容量限制而被迫做的分库分表的场景，或者需要强一致性和完整分布式事务的场景。它的优势是通过尽量下推到存储节点进行并行计算。对于小表（比如千万级以下），不适合 TiDB，因为数据量少，Region 有限，发挥不了并行的优势。其中最极端的例子就是计数器表，几行记录高频更新，这几行在 TiDB 里，会变成存储引擎上的几个 KV，然后只落在一个 Region 里，而这个 Region 只落在一个节点上。加上后台强一致性复制的开销，TiDB 引擎到 TiKV 引擎的开销，最后表现出来的就是没有单个 MySQL 好。

14.5.6 TiDB 备份恢复

14.5.6.1 TiDB 主要备份方式？

目前，数据量大时推荐使用BR 进行备份。其他场景推荐使用Dumpling 进行备份。

尽管 TiDB 也支持使用 MySQL 官方工具 `mysqldump` 进行数据备份和恢复，但其性能低于Dumpling，并且 `mysqldump` 备份和恢复大量数据的耗费更长。

15 术语表

15.1 A

15.1.1 ACID

ACID 是指数据库管理系统在写入或更新资料的过程中，为保证事务是正确可靠的，所必须具备的四个特性：原子性 (atomicity)、一致性 (consistency)、隔离性 (isolation) 以及持久性 (durability)。

- 原子性 (atomicity) 指一个事务中的所有操作，或者全部完成，或者全部不完成，不会结束在中间某个环节。TiDB 通过 Primary Key 所在Region 的原子性来保证分布式事务的原子性。
- 一致性 (consistency) 指在事务开始之前和结束以后，数据库的完整性没有被破坏。TiDB 在写入数据之前，会校验数据的一致性，校验通过才会写入内存并返回成功。
- 隔离性 (isolation) 指数据库允许多个并发事务同时对其数据进行读写和修改的能力。隔离性可以防止多个事务并发执行时由于交叉执行而导致数据的不一致，主要用于处理并发场景。TiDB 目前只支持一种隔离级别，即可重复读。
- 持久性 (durability) 指事务处理结束后，对数据的修改就是永久的，即便系统故障也不会丢失。在 TiDB 中，事务一旦提交成功，数据全部持久化存储到 TiKV，此时即使 TiDB 服务器宕机也不会出现数据丢失。

15.2 C

15.2.1 Continuous Profiling

Continuous Profiling（持续性能分析）是从 TiDB v5.3 起引入的一种从系统调用层面解读资源开销的方法。引入该方法后，TiDB 可提供数据库源码级性能观测，通过火焰图的形式帮助研发、运维人员定位性能问题的根因。详情参见TiDB Dashboard 实例性能分析 - 持续分析页面。

15.3 I

15.3.1 Index Merge

Index Merge（索引合并）是在 TiDB v4.0 版本中作为实验特性引入的一种查询执行方式的优化，可以大幅提高查询在扫描多列数据时条件过滤的效率。自 v5.4 版本起，Index Merge 成为正式功能，详情参见用 EXPLAIN 查看索引合并的 SQL 执行计划。

15.4 L

15.4.1 Leader/Follower/Learner

它们分别对应Peer 的三种角色。其中 Leader 负责响应客户端的读写请求；Follower 被动地从 Leader 同步数据，当 Leader 失效时会进行选举产生新的 Leader；Learner 是一种特殊的角色，它只参与同步 raft log 而不参与投票，在目前的实现中只短暂存在于添加副本的中间步骤。

15.5 O

15.5.1 Old value

Old value 特指在 TiCDC 输出的增量变更日志中的“原始值”。可以通过配置来指定 TiCDC 输出的增量变更日志是否包含“原始值”。

15.5.2 Operator

Operator 是应用于一个 Region 的，服务于某个调度目的的一系列操作的集合。例如“将 Region 2 的 Leader 迁移至 Store 5”，“将 Region 2 的副本迁移到 Store 1, 4, 5”等。

Operator 可以是由 Scheduler 通过计算生成的，也可以是由外部 API 创建的。

15.5.3 Operator Step

Operator Step 是 Operator 执行过程的一个步骤，一个 Operator 常常会包含多个 Operator Step。

目前 PD 可生成的 Step 包括：

- `TransferLeader`：将 Region Leader 迁移至指定 Peer

- AddPeer：在指定 Store 添加 Follower
- RemovePeer：删除一个 Region Peer
- AddLearner：在指定 Store 添加 Region Learner
- PromoteLearner：将指定 Learner 提升为 Follower
- SplitRegion：将指定 Region 一分为二

15.6 P

15.6.1 Pending/Down

Pending 和 Down 是 Peer 可能出现的两种特殊状态。其中 Pending 表示 Follower 或 Learner 的 raft log 与 Leader 有较大差距，Pending 状态的 Follower 无法被选举成 Leader。Down 是指 Leader 长时间没有收到对应 Peer 的消息，通常意味着对应节点发生了宕机或者网络隔离。

15.6.2 Predicate columns

执行 SQL 语句时，优化器在大多数情况下只会用到部分列（例如，WHERE、JOIN、ORDER BY、GROUP BY 子句中出现的列）的统计信息，这些用到的列称为 PREDICATE COLUMNS。详情参见收集部分列的统计信息。

15.7 R

15.8 Raft Engine

一种内置的持久化存储引擎，有着日志结构的设计，为 TiKV 提供 multi-Raft 日志存储。从 v5.4 起，TiDB 支持使用 Raft Engine 作为 TiKV 的日志存储引擎（实验特性，默认关闭）。详情参见 Raft Engine。

15.8.1 Region/Peer/Raft Group

每个 Region 负责维护集群的一段连续数据（默认配置下平均约 96 MiB），每份数据会在不同的 Store 存储多个副本（默认配置是 3 副本），每个副本称为 Peer。同一个 Region 的多个 Peer 通过 raft 协议进行数据同步，所以 Peer 也用来指代 raft 实例中的成员。TiKV 使用 multi-raft 模式来管理数据，即每个 Region 都对应一个独立运行的 raft 实例，我们也把这样的一个 raft 实例叫做一个 Raft Group。

15.8.2 Region Split

TiKV 集群中的 Region 不是一开始就划分好的，而是随着数据写入逐渐分裂生成的，分裂的过程被称为 Region Split。

其机制是集群初始化时构建一个初始 Region 覆盖整个 key space，随后在运行过程中每当 Region 数据达到一定量之后就通过 Split 产生新的 Region。

15.8.3 Restore

备份操作的逆过程，即利用保存的备份数据还原出原始数据的过程。

15.9 S

15.9.1 Scheduler

Scheduler（调度器）是 PD 中生成调度的组件。PD 中每个调度器是独立运行的，分别服务于不同的调度目的。常用的调度器及其调用目标有：

- `balance-leader-scheduler`：保持不同节点的 Leader 均衡。
- `balance-region-scheduler`：保持不同节点的 Peer 均衡。
- `hot-region-scheduler`：保持不同节点的读写热点 Region 均衡。
- `evict-leader-{store-id}`：驱逐某个节点的所有 Leader。（常用于滚动升级）

15.9.2 Store

PD 中的 Store 指的是集群中的存储节点，也就是 tikv-server 实例。Store 与 TiKV 实例是严格一一对应的，即使在同一主机甚至同一块磁盘部署多个 TiKV 实例，这些实例也对会对应不同的 Store。

15.10 T

15.10.1 Top SQL

从 v5.4 起引入的实验特性，用于找到一段时间内对某个 TiDB 或 TiKV 节点消耗负载较大的 SQL 查询。详情参见Top SQL 用户文档。

15.10.2 TSO

因为 TiKV 是一个分布式的储存系统，它需要一个全球性的授时服务 TSO（Timestamp Oracle），来分配一个单调递增的时间戳。这样的功能在 TiKV 中是由 PD 提供的，在 Google 的 Spanner 中是由多个原子钟和 GPS 来提供的。

©2019 PingCAP 公司保留所有权利。除非版权法允许，否则在未得到本公司事先给出的书面许可的情况下，严禁复制、改编或翻译本文。