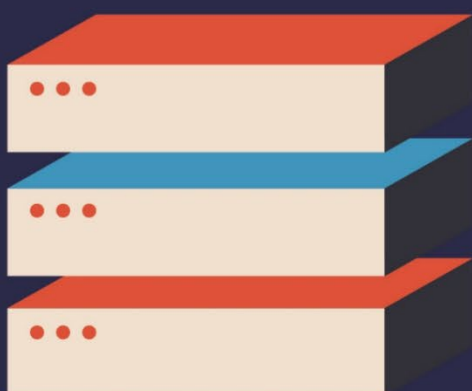




TiDB 技术白皮书（企业版）

产品版本： v6.5 | 文档版本： 20221229



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。

目 录

TiDB 简介	7
技术优势	8
产品架构	9
分布式数据库计算层	9
分布式数据库存储层	11
分布式数据库调度层	12
TiDB HTAP 架构	13
TiDB 生态工具	14
数据迁移	14
数据迁入 - TiDB Data Migration	14
增量数据迁出 - TiCDC	15
数据导入 - TiDB Lightning	16
数据导出 - Dumpling	16
备份和恢复 - Backup & Restore	16
安装部署	17
TiUP	17
TiDB Operator	18
运维和可视化管理 - TiDB Dashboard	18
企业级高级功能	20
弹性扩缩容	20
多副本、数据强一致	22
故障自动转移与恢复	23
备份与恢复	24
同城和异地容灾	25
同城三中心容灾	25
两地三中心容灾	26

TiDB 简介

平凯星辰（北京）科技有限公司（PingCAP）是一家企业级开源软件服务供应商，提供包括开源分布式数据库产品、解决方案及相关的咨询、支持与培训服务。致力于为用户打造稳定高效、安全可靠、开放兼容的新型数据基础设施，加速企业数字化转型升级。

PingCAP 自主研发的开源分布式关系型数据库 TiDB，为企业关键业务打造，具备「分布式强一致性事务、在线弹性水平扩展、故障自恢复的高可用、跨数据中心多活」等核心特性，助力企业最大化发挥数据价值，释放企业增长空间。

目前，PingCAP 已经向包括中国、美国、欧洲、日本、东南亚等国家和地区，超过 1500 家企业提供服务，涉及金融、运营商、制造、零售、互联网、政府等多个行业。

技术优势

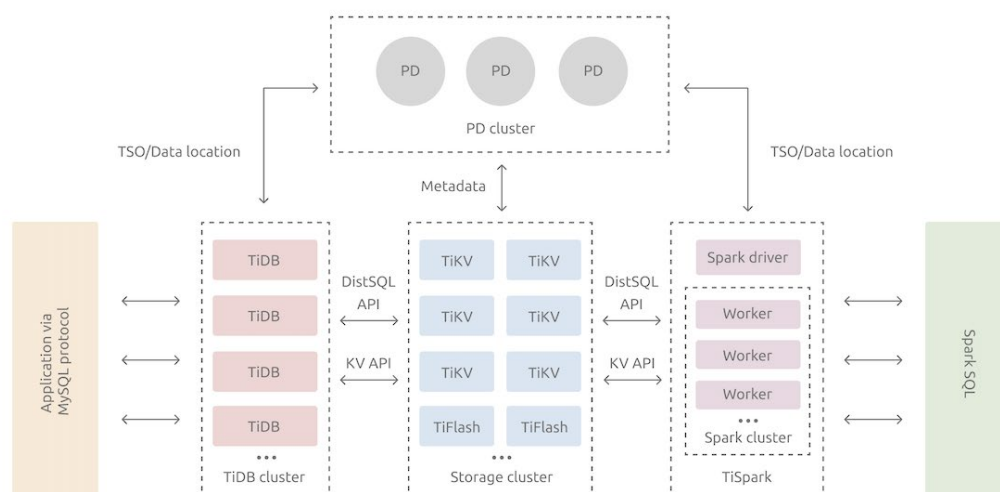
与传统的单机数据库相比，TiDB 具有以下优势：

- 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
- 支持实时 在线事务与在线分析处理 (Hybrid Transactional and Analytical Processing, HTAP) 功能

产品架构

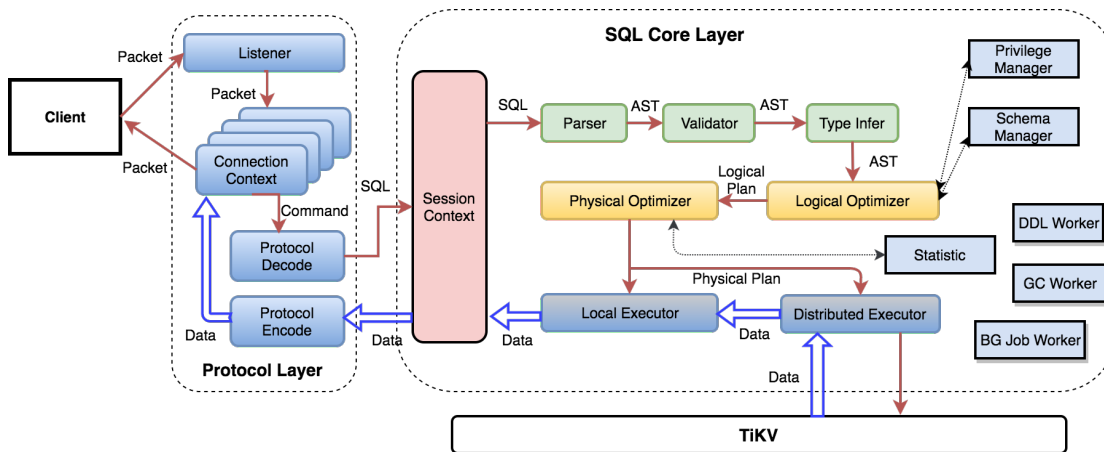
近年来，随着移动互联网、云计算、大数据和人工智能等技术的飞速发展，给各行业带来了深刻的影响和变革，使得企业的数据量越来越庞大，应用的规模也越来越复杂。在这个背景之下，传统的单机数据库已经在很多场景下表现的力不从心，为了解决海量数据平台的扩展性的问题，TiDB 分布式数据库应运而生。

TiDB 分布式数据库最初的设计受到 Google 内部开发的知名分布式数据库 Spanner 和 F1 的启发，在内核设计上将整体的架构拆分成多个大的模块，大的模块之间互相通信，组成完整的 TiDB 系统，TiDB 采用 Shared-Nothing、计算存储分离的分布式集群架构，主要包括TiDB Server、TiKV Server、TiFlash Server、PD Server、TiSpark、TiCDC 等多个核心组件。

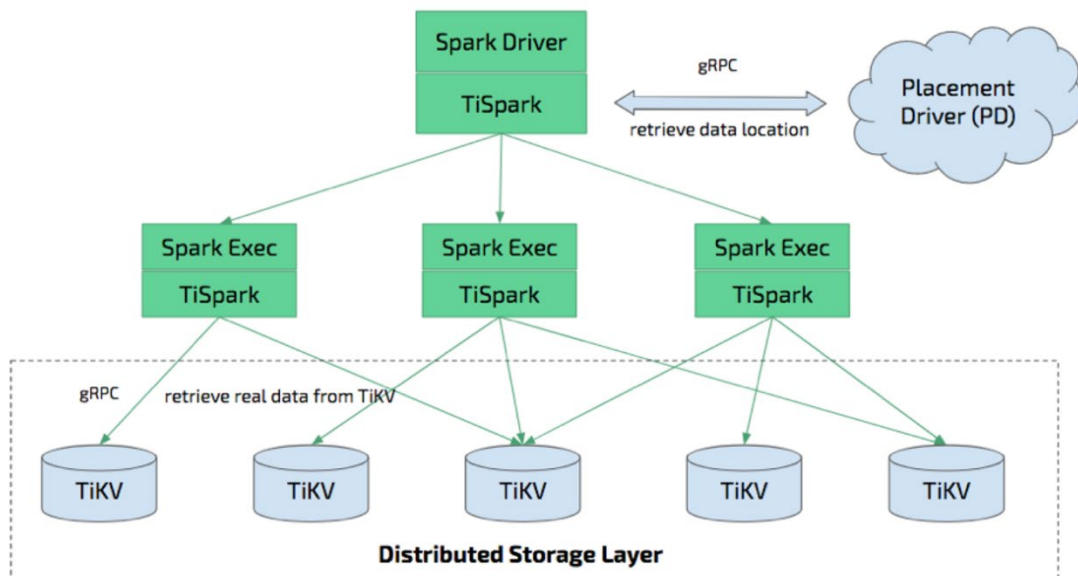


分布式数据库计算层

计算层由 TiDB Server 和 TiSpark 组成，如下：



TiDB Server 是 TiDB 的 SQL 层，由多个 TiDB Server 组件组成，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 或者 TiFlash。

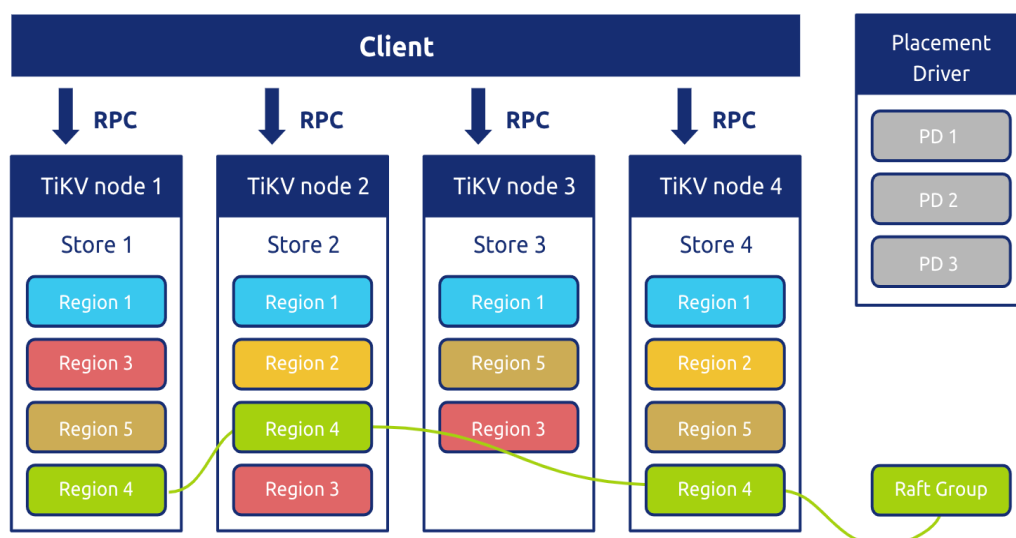


TiSpark 是解决用户复杂 OLAP 需求而推出的产品。它借助 Spark 平台，同时融合 TiKV 分布式集群的优势，和 TiDB 一起为用户一站式解决 HTAP (Hybrid Transactional/Analytical Processing) 的需求。将 Spark SQL 直接运行在分布式存储引擎 TiKV 上的 OLAP 解决方案。深度整合了 Spark Catalyst 引擎, 可以对计算提供精确的控制, 使 Spark 能够高效的读取 TiKV 中的数据, 提供索引支持以实现高速的点查。通过多种计算下推

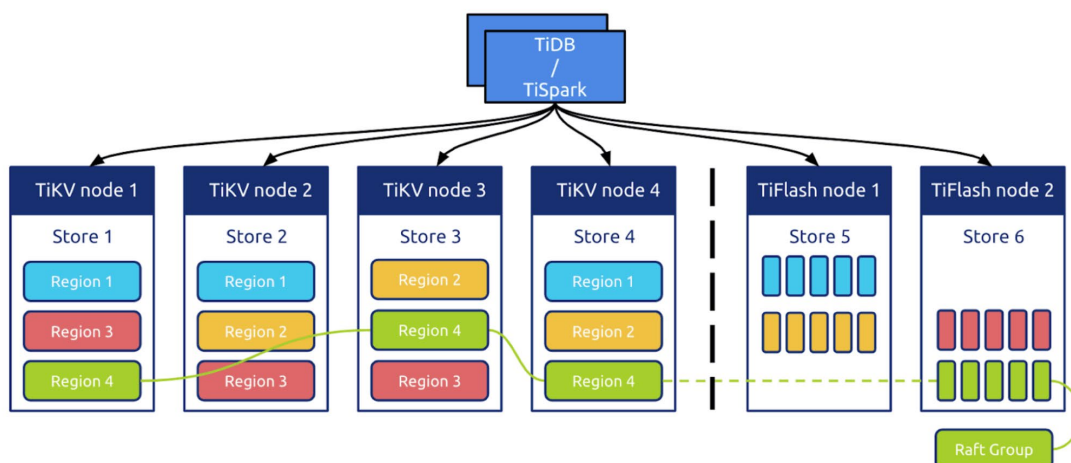
减少 Spark SQL 需要处理的数据大小，以加速查询；利用 TiDB 的内建的统计信息选择更优的查询计划。从数据集群的角度看，TiSpark + TiDB 可以让用户无需进行脆弱和难以维护的 ETL。

分布式数据库存储层

存储层由行存储和列存储服务组成，如下：

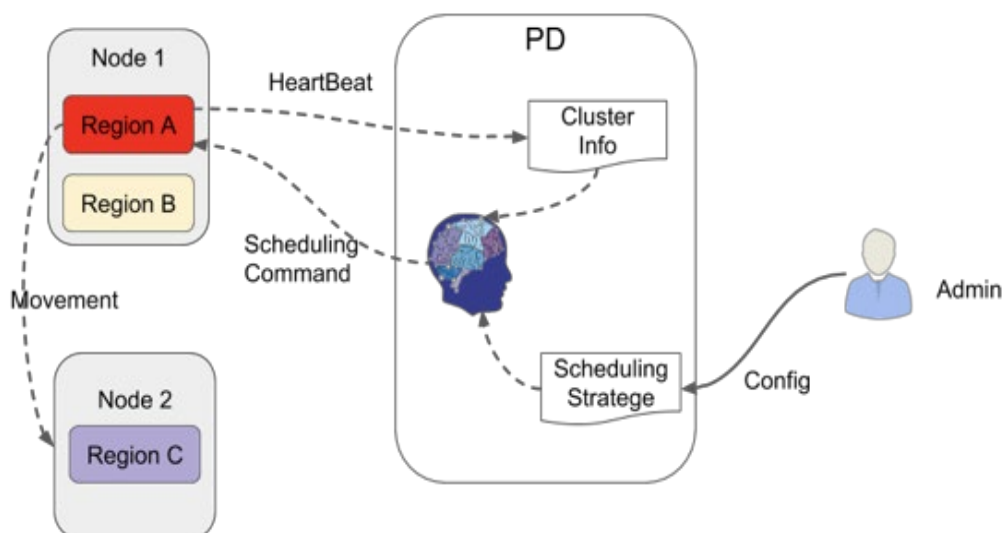


TiKV 是行存储服务，由多个 tikv-server 组成，从外部看 TiKV 是一个分布式的提供事务的 Key-Value 存储引擎。存储数据的基本单位是 Region，每个 Region 负责存储一个 Key Range（从 StartKey 到 EndKey 的左闭右开区间）的数据，每个 TiKV 节点会负责多个 Region。TiKV 的 API 在 KV 键值对层面提供对分布式事务的原生支持，默认提供了 SI (Snapshot Isolation) 的隔离级别，这也是 TiDB 在 SQL 层面支持分布式事务的核心。TiDB 的 SQL 层做完 SQL 解析后，会将 SQL 的执行计划转换为对 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 是列存储服务，由多个 tiflash-server 组件组成，TiFlash 与普通的 TiKV 基本是一样的，不同之处在于，TiFlash 提供列式存储，借助 ClickHouse 实现高效的协处理器层、MPP 架构使分析的性能能得到很好的加速。TiFlash 以 Raft Learner 角色加入 TiKV 并进行异步的数据复制，当 TiFlash 节点宕机或者网络高延迟等状况发生时，TiKV 的业务仍然能确保正常进行。TiFlash 支持快照隔离，保证读取到最新的数据。读取数据时 TiDB 根据统计信息并计算出代价，自动选择从 TiFlash 或者 TiKV 读取数据，用户对此完全无感知。

分布式数据库调度层



调度层是整个 TiDB 集群的元信息管理模块，由 pd-server 组件组成，负责存储每个 TiKV 或者 TiFlash 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

或者 TiFlash 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 或者 TiFlash 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。

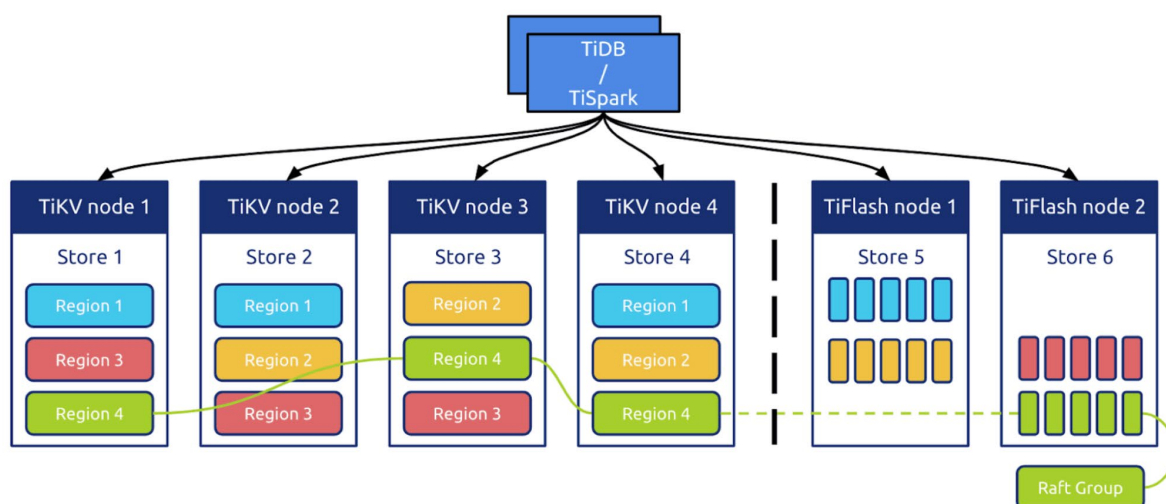
TiDB HTAP 架构

TiDB HTAP 可以满足企业海量数据的增产需求、降低运维的风险成本、与现有的大数据栈无缝缝合，从而实现数据资产价值的实时变现。

以下是三种 HTAP 典型适用场景：

- 混合负载场景
当将 TiDB 应用于在线实时分析处理的混合负载场景时，开发人员只需要提供一个入口，TiDB 将自动根据业务类型选择不同的处理引擎。
- 实时流处理场景
当将 TiDB 应用于实时流处理场景时，TiDB 能保证源源不断流入系统的数据实时可查，同时可兼顾高并发数据服务与 BI 查询。
- 数据中枢场景
当将 TiDB 应用于数据中枢场景时，TiDB 作为数据中枢可以无缝连接数据业务层和数据仓库层，满足不同业务的需求。

下图为 TiDB HTAP 形态架构，其中包含 TiFlash 节点。



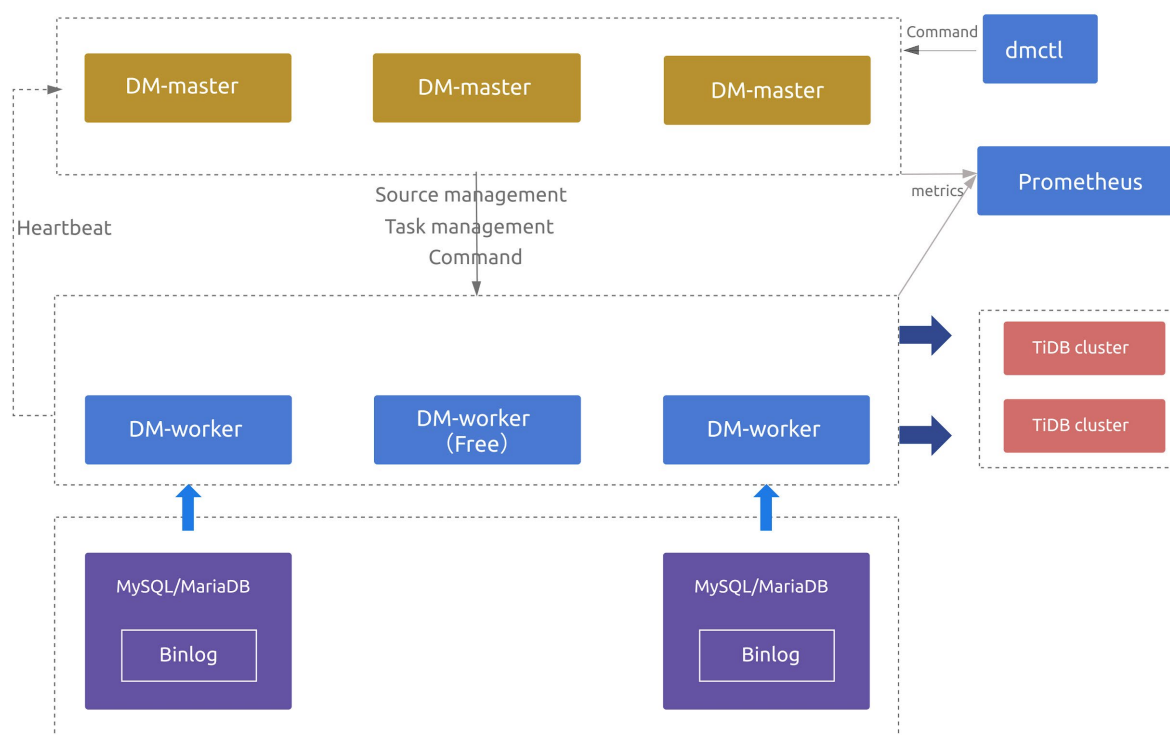
TiDB 生态工具

数据迁移

数据迁入 - TiDB Data Migration

TiDB Data Migration (DM) 是一体化的数据迁移任务管理平台，支持从与 MySQL 协议兼容的数据库（MySQL、MariaDB、Aurora MySQL）到 TiDB 的数据迁移。使用 DM 工具有利于简化错误处理流程，降低运维成本，主要优势如下：

- 数据迁移任务的高可用
- 先进的分库分表合并技术
 - 分表表数据同步合并
 - 独有的分库分表 DDL 合并技术
- 易用的错误处理机制
- 支持各组件间数据传输加密

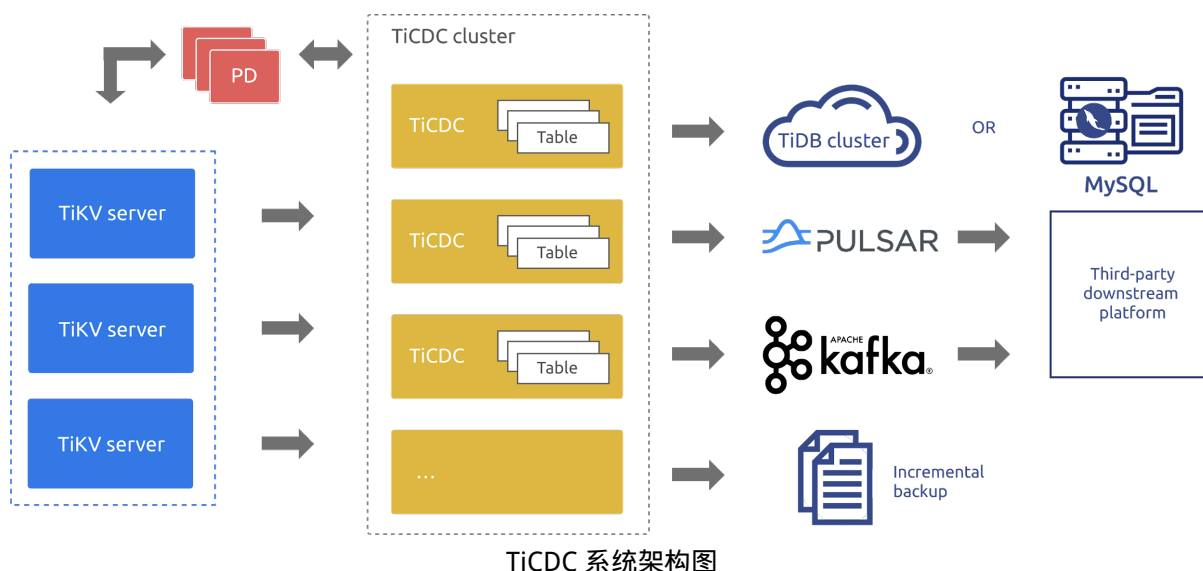


DM 架构图

增量数据迁出 - TiCDC

TiCDC (CDC, Change Data Capture) 是一款通过拉取 TiKV 变更日志实现的 TiDB 增量数据同步工具，提供多种数据协议 (TiCDC Open Protocol / Apache Avro / Canal)，支持其他系统订阅数据变更，并已获得 Confluent Standard 认证。

TiCDC 运行时通过 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。



通过 TiCDC 工具，用户可以非常方便及安全可靠地实现：

- TiDB 分布式数据库的数据实时同步
- 匹配任何规模 TiDB 集群的水平扩展能力
- 数据复制的高可用
- 对接丰富的第三方生态 (Canal / Kafka Connect)

数据导入 - TiDB Lightning

TiDB Lightning 是一个用于将全量数据高速导入到 TiDB 集群的工具，速度可达到 SQL 导入方式的至少 10 倍。基于 TPCC 数据的导入测试显示，导入速度可达到每小时 400 ~ 500 GB，适合上线前迁移现有的大型数据库到全新的 TiDB 集群。

- 支持导入 MySQL dump files / CSV / Aurora snapshot / Parquet 格式的数据
- 支持直接从本地存储及与 Amazon S3 协议兼容的存储导入数据

数据导出 - Dumpling

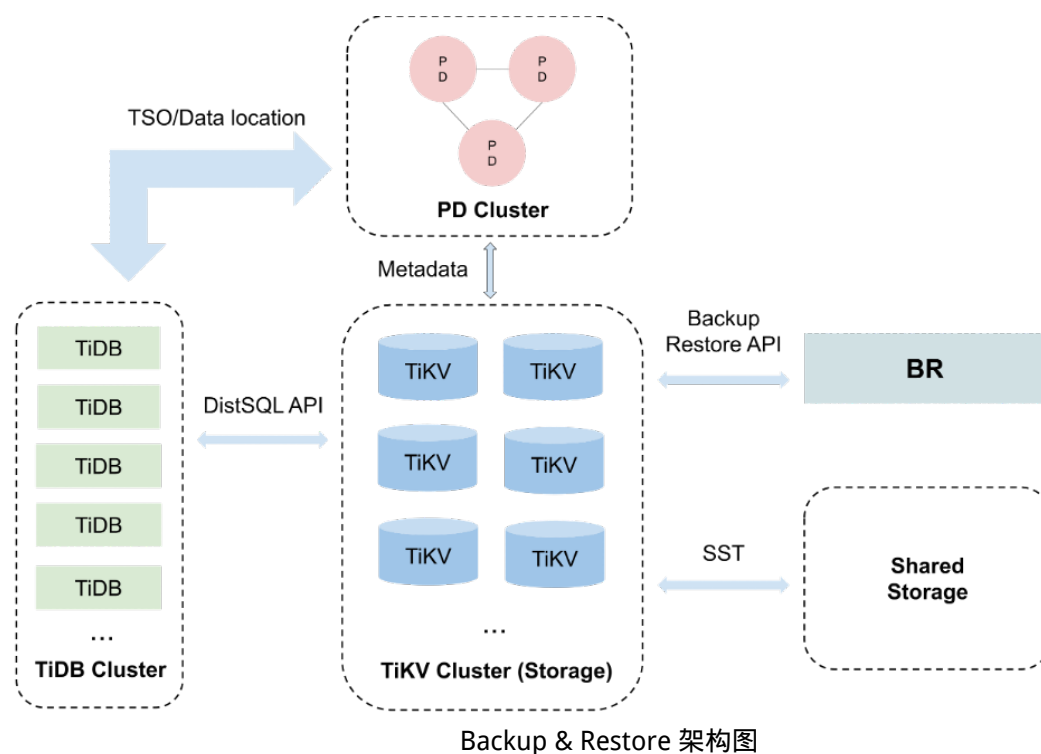
Dumpling 是由 Go 语言编写的用于对数据库进行数据导出的工具，目前支持 MySQL 协议的数据库，并且针对 TiDB 的特性进行了优化，Dumpling 的主要特点包括：

- 适配 Mydumper，轻松上手
- 支持自定义导出过滤条件
- 支持自动设置 TiDB GC 时间
- 支持自动调节 TiDB 导出参数以防止 OOM
- 多种导出格式，目前支持 SQL、CSV 格式的导出

备份和恢复 - Backup & Restore

Backup & Restore (简称 BR) 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。功能和特点如下

- 性能可水平扩展：
 - 备份速度：~150MB/s * TiKV 节点数 (基于 Intel Xeon E5 + NVMe SSD + 3 副本)
 - 恢复速度：~50MB/s * TiKV 节点数 (配置同上)
- 支持增量备份
- 支持备份到 Amazon S3、Google Cloud Storage 以及持久卷



安装部署

TiUP

TiUP 是 TiDB 的组件管理平台，用于扩展 TiDB 生态边界，集成并管理 TiDB 生态的所有组件，方便用户进行多组件多版本管理，目前已经广泛用于客户生产环境。主要特点：

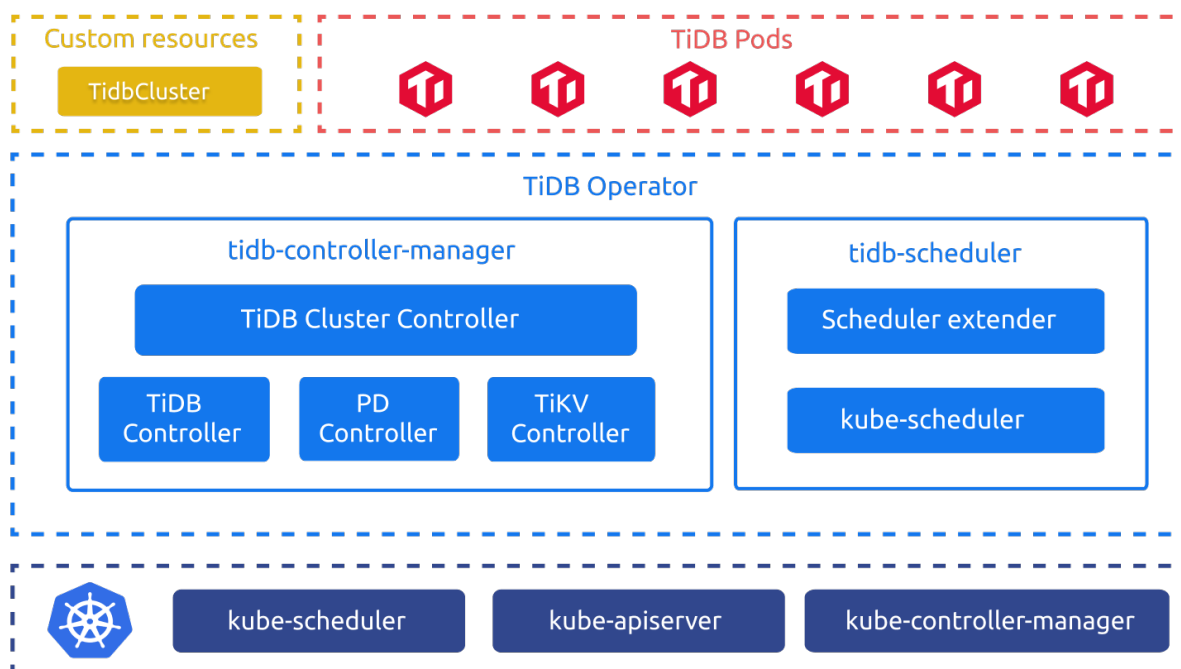
- 通过 `tiup playground` 一键运行一个本地 TiDB 集群，可以让用户快速上手 TiDB 尝鲜，同时也为开发者提供了极大的便利运行测试集群。
- 通过 `tiup cluster` 组件快速部署、升级、扩缩容、配置变更 On-Premises 部署的 TiDB 生产集群。TiDB v5.0 版本中，`tiup cluster` 组件的集群部署操作逻辑对用户更友好，升级集群操作更加稳定，升级时长更短，同时支持断点功能，并增强了集群运维功能。
- 可以通过 `tiup dm` 组件快速部署、升级、扩缩容、配置变更 On-Premises 部署的 DM 数据迁移集群。
- 完备的离线镜像功能，为企业级客户的内网环境提供一致的用户体验。

- 支持 x86/arm 单架构部署、多架构异构部署。

TiDB Operator

TiDB Operator 是 Kubernetes 上的 TiDB 集群自动运维系统，为云原生大趋势提供了标准的数据库上云解决方案，主要特点：

- 借助 TiDB Operator，TiDB 可以无缝运行在公有云（支持 AWS EKS、GCP GKE、Azure AKS、阿里云 ACK）或私有部署的 Kubernetes 集群上。
- 部署、升级、扩缩容、备份恢复、配置变更的 TiDB 全生命周期管理的全自动化运维。
- 弹性扩缩容，根据实际负载自动扩容、缩容计算节点和存储节点。



TiDB Operator 架构图

运维和可视化管理 - TiDB Dashboard

TiDB Dashboard 是从 TiDB v4.0 版本起内置的图形化界面，可以在线监控、诊断和管理 TiDB 集群。

- TiDB 原生集成无需独立部署、可详细分析各类 SQL 语句的性能指标和执行情况、**并提供**独特的热点可视化诊断工具 KeyViz。
- 与 Navicat 商业工具类似，支持图形化查看和管理数据库用户、创建或删除数据库和数据表、编辑数据表的列、索引和分区结构信息、查询表中数据、添加或删除数据行、在线导入导出等。
- 支持全流程追踪 SQL 执行技术，用于性能和问题诊断。
- 支持更细粒度的 TiDB 集群状态可视化展示。

企业级高级功能

高可用特性是企业级系统关键特性之一，是确保企业的业务连续性的关键。

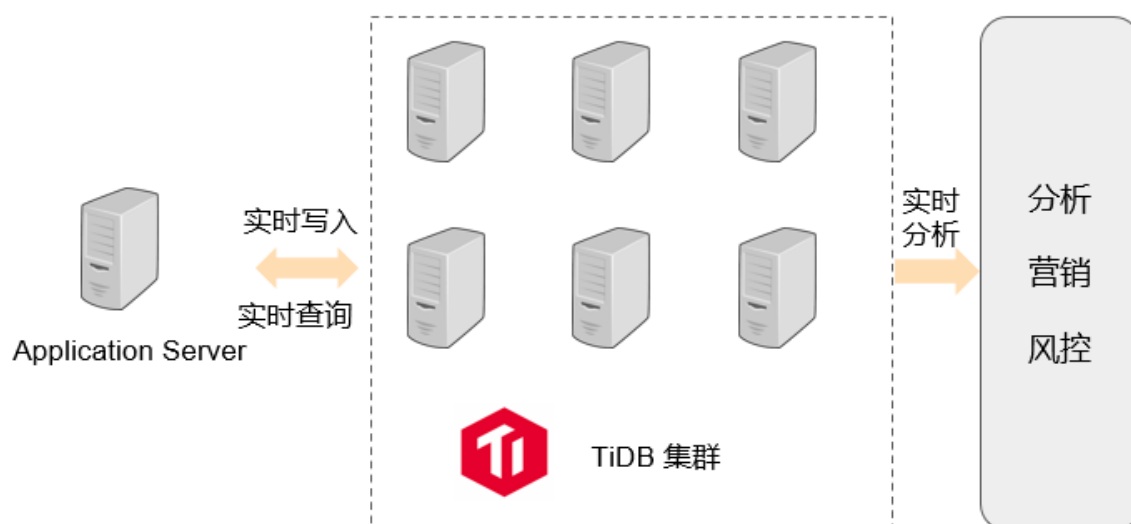
TiDB 是一款分布式数据库，分布式系统在设计时出于对成本的考量，会将服务部署在比较廉价硬件上，其可靠性相对小型机、专有硬件有很大的不足，通过数据分片、副本等技术将超大数据集分散到众多的廉价机器上，而众多廉价会部在不同的机架、机房、数据中心，如此众多的机器的操作系统、软件、存储介质、网络等都可能会出现不可用的状态。

TiDB 在设计之初就考虑到此类的问题，主要采用存储计算分离的分布式架构、确保不会因任意硬件、软件、网络故障，导致系统的元数据、数据、计算服务、存储服务不可用，其主要的技术包括：多副本、数据强一致、故障自动转移和恢复、备份与恢复、同城和异地容灾等。

弹性扩缩容

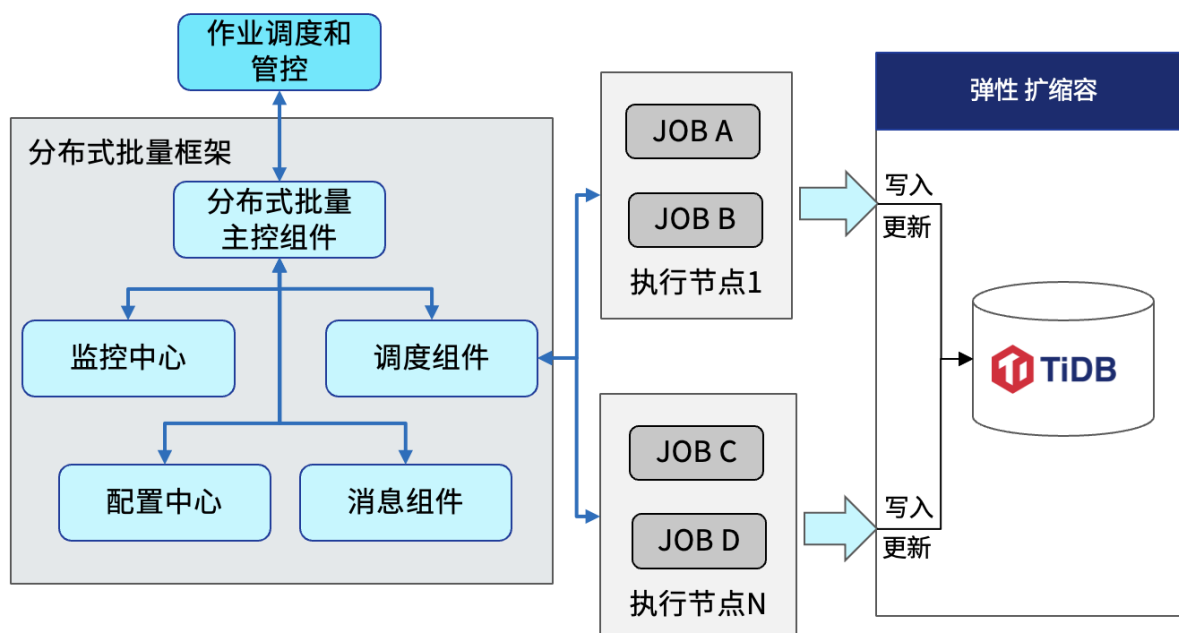
TiDB 支持海量数据的高性能写入与实时分析：

- TiDB 提供可扩展、无上限的实时写入和实时查询能力
- TiDB 具备强大的分布式查询引擎，可用于各种实时分析场景



TiDB 分布式架构提供整体的弹性扩缩容能力

- TiDB 水平弹性扩展、分布式数据库强一致性等特性适合批量计算场景
- 解决传统数据库并发支持低、分析能力差、无法线性扩展等痛点
- 结合应用系统的场景能力，打造整体分布式弹性扩缩容的业务架构

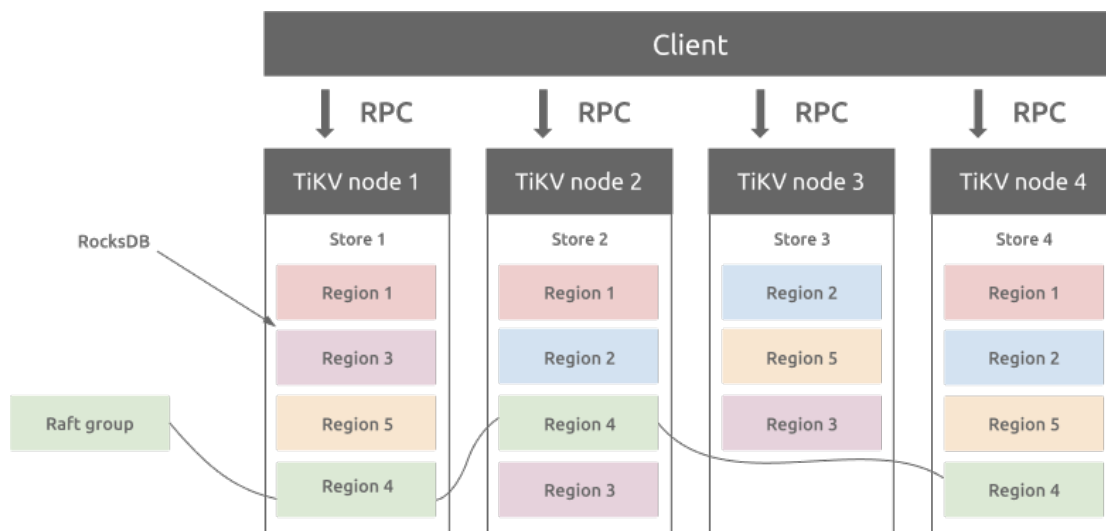


具备以下优势：

- 高并发、可扩展、无上限的实时写入和查询能力；
- 采用存储计算分离的架构设计，可按需对计算、存储分别进行在线扩容或者缩容，且过程对应用运维人员透明；

- 数据均匀分散到所有存储节点，对应用完全透明。

多副本、数据强一致



TiKV 参考 Spanner 设计了 multi-raft-group 的副本机制。将数据按照 key 的范围划分成大致相等的切片（下文统称为 Region），每一个切片会有多个副本（通常是 3 个），其中一个副本是 Leader，提供读写服务。TiKV 通过 PD 对这些 Region 以及副本进行调度，以保证数据和读写负载都均匀地分散在各个 TiKV 上，这样的设计保证了整个集群资源的充分利用并且可以随着机器数量的增加水平扩展。

虽然 TiKV 将数据按照范围切割成了多个 Region，但是同一个节点的所有 Region 数据仍然是不加区分地存储于同一个 RocksDB 实例上，而用于 Raft 协议复制所需要的日志则存储于另一个 RocksDB 实例。这样设计的原因是因为随机 I/O 的性能远低于顺序 I/O，所以 TiKV 使用同一个 RocksDB 实例来存储这些数据，以便不同 Region 的写入可以合并在一次 I/O 中。

Region 与副本之间通过 Raft 协议来维持数据一致性，任何写请求都只能在 Leader 上

写入，并且需要写入多数副本后（默认配置为 3 副本，即所有请求必须至少写入两个副本成功）才会返回客户端写入成功。

当某个 Region 的大小超过一定限制（默认是 144MB）后，TiKV 会将它分裂为两个或者更多个 Region，以保证各个 Region 的大小是大致接近的，这样更有利于 PD 进行调度决策。同样，当某个 Region 因为大量的删除请求导致 Region 的大小变得更小时，TiKV 会将比较小的两个相邻 Region 合并为一个。

当 PD 需要把某个 Region 的一个副本从一个 TiKV 节点调度到另一个上面时，PD 会先为这个 Raft Group 在目标节点上增加一个 Learner 副本（虽然会复制 Leader 的数据，但是不会计入写请求的多数副本中）。当这个 Learner 副本的进度大致追上 Leader 副本时，Leader 会将它变更为 Follower，之后再移除操作节点的 Follower 副本，这样就完成了 Region 副本的一次调度。

Leader 副本的调度原理也类似，不过需要在目标节点的 Learner 副本变为 Follower 副本后，再执行一次 Leader Transfer，让该 Follower 主动发起一次选举成为新 Leader，之后新 Leader 负责删除旧 Leader 这个副本。

故障自动转移与恢复

元数据管理组件 PD 通常是由 3 或者 5 个节点组成的集群，每一个节点都存储了全量的元数据信息，任意节点的故障都不会影响服务的正常运转。

数据存储层由多个 TiKV 存储节点组成，通常 TiKV 的节点数大于等于 3 个。默认情况下每一份数据都有 3 个副本，分别存储在不同的 3 个 TiKV 存储节点上，当某一个 TiKV 存储节点出现故障的时候，其他两个 TiKV 存储节点可以保证对这份数据继续提供服务。当某个 TiKV 存储节点长时间（超过 30 分钟）处于故障的时候，原信息管理组件

PD 会自动发起补充数据副本的动作，让这部分数据从 2 副本状态恢复到 3 副本的健康状态。

计算层（SQL 层）由多个无状态的节点 TiDB 组成，通常会在应用和 TiDB 组件之间使用负载均衡中间件，如 HAProxy，当某个 TiDB 计算节点出现故障时，HAProxy 会自动将请求路由到其他存活的 TiDB 节点上，保证服务的高可用性。

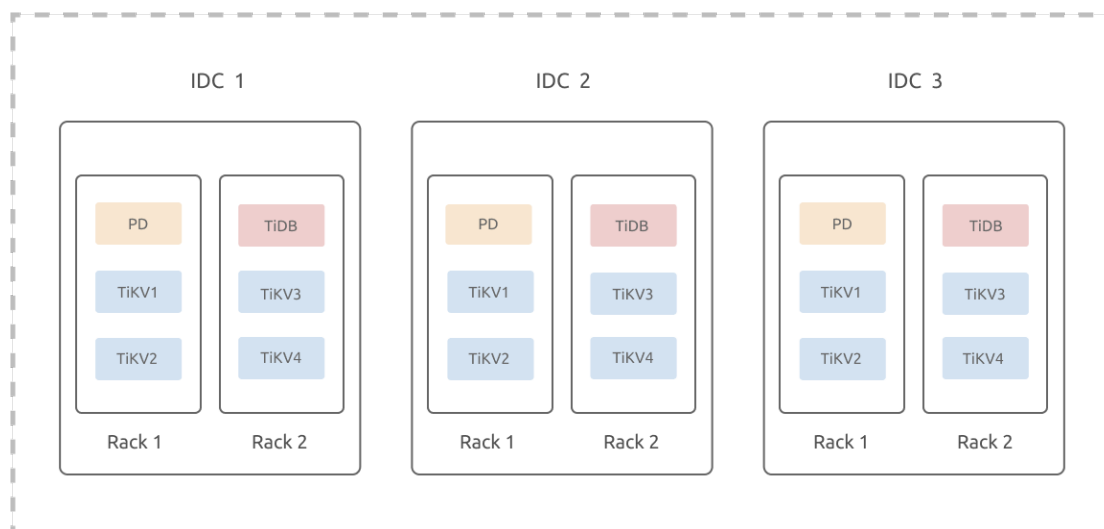
备份与恢复

TiDB 分布式数据库软件提供在线的，一致性的全量备份、增量备份，全量恢复、增量恢复，同时可以通过 TiDB 提供自身的 Binlog 输出，实时生成数据增量备份，也可以实时同步数据到 MySQL 或者其他数据存储便捷的实现 TiDB 到其它数据和处理系统的数据交互：

- 具备在线全量备份、在线增量备份和在线增量备份的功能，增量备份不依赖全量备份。
- 支持到任意时间点的备份还原操作。全量备份和增量备份数据强一致、时间点强一致、不阻塞业务服务。
- 可将任意库表数据恢复到任意时间点，而不影响其他库表数据。

同城和异地容灾

同城三中心容灾



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。

优点，如下：

- RPO == 0，任何一个数据中心失效后，不会产生任何数据丢失。
- RTO <= 30s，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 30s 合理长的时间内自动恢复服务。
- 副本或者节点宕机，自己补齐副本，确保有三个副本。

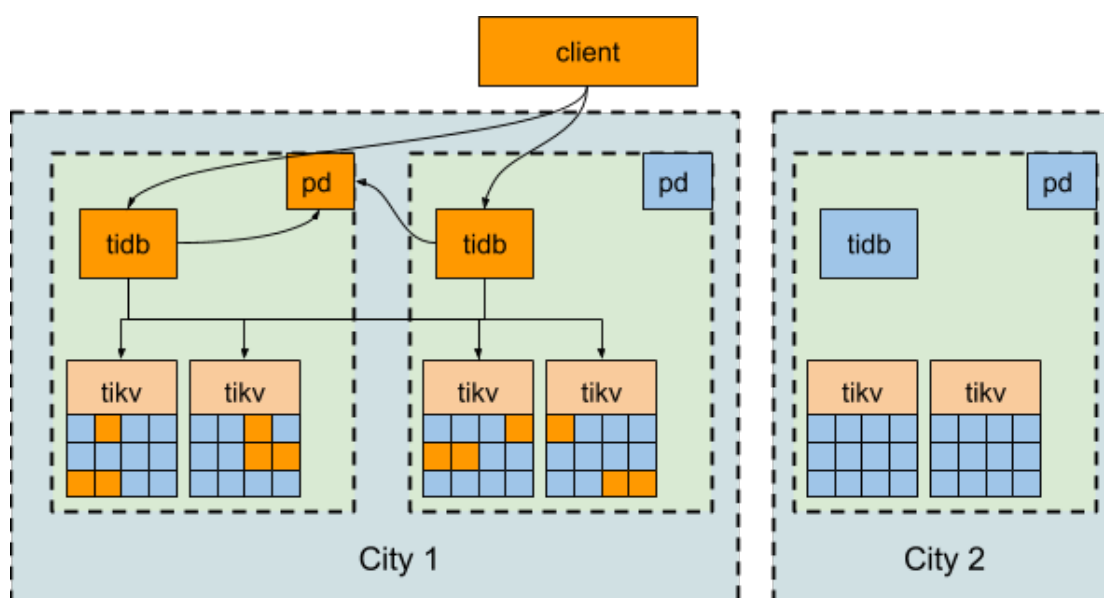
缺点，如下：

- 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。

两地三中心容灾



两地三中心容灾主案，即同城主集群两机房 5 副本，异地备集群 3 副本，主备集群之间使用 Binlog 同步，主 TiDB 数据库集群持续对外提供服务，同时通过 TiDB Binlog 将数据库的日志流式发送给灾备数据中心的 TiDB 从数据库集群进行数据同步。主站点发生不可抗力的故障导致整体不可用后，TiDB 从数据库集群拥有全量备份数据，可以立刻开始接受灾备切换后的业务访问。

优点，如下：

- 同城 2 副本单机房 或者 异地机房宕机不影响业务的连续性
- RPO < 30s、RTO <= 15min