平凯数据库产品简介

平凯星辰

20230713

目录

1 平凯数据库企业版产品简介

1.1 法律声明

平凯星辰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. 您应当通过平凯星辰网站或本公司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。本文档的内容视为平凯星辰的保密信息，您应当严格遵守保密义务；未经平凯星辰事先书面同意，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. 未经平凯星辰事先书面许可，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。平凯星辰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，并在平凯星辰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平凯星辰网站或平凯星辰授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用平凯星辰产品及服务的参考性指引，平凯星辰以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。平凯星辰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，但平凯星辰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，平凯星辰不承担任何法律责任。在任何情况下，平凯星辰均不对任何间接性、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（即使平凯星辰已被告知该等损失的可能性）。
5. 本文档中及平凯星辰网站上所有内容，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，均由平凯星辰和/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。非经平凯星辰和/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、发行或公开发表平凯星辰网站、产品程序或包括本文档在内的内容。此外，未经平凯星辰事先书面同意，任何人不得以任何目的使用平凯星辰商标（包括但不限于单独为或以组合形式包含“平凯星辰”等平凯星辰和/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平凯星辰和/或其关联公司）。
6. 平凯星辰可能拥有涵盖本文档中描述的主题的专利或者专利申请，未经平凯星辰事先书面许可，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. 本文档中如有关于平凯星辰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，请与平凯星辰取得直接联系。平凯星辰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。

1.2 平凯数据库简介

平凯星辰（北京）科技有限公司是一家企业级开源软件服务供应商，提供包括分布式数据库产品、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。

平凯星辰自主研发的关系型分布式数据库企业级产品 – 平凯数据库，作为一款面向海量数据在线链接交易、业务数据实时分析、数据强一致、跨数据中心高可用的金融级分布式数据库软件，为企业关键业务打造，助力企业最大化发挥数据价值，释放企业增长空间。

目前，平凯数据库自主开源的知名开源项目 TiDB Open Core，有超过 1400 多位全球范围的活跃贡献者，同时超过 1500 个企业的真实生产场景，不断地使分布式数据库产品“快速迭代，持续创新”，在经历了社区的全球技术智慧的贡献和整合，以及众多行业用户实际应用反馈与持续优化的过程，从而在国内市场上造就了高稳定与高可靠的产品质量部署认可度，能够在信息创新发展趋势下持续打破技术的壁垒，真实有效的打消企业用户对于开源数据库安全可信使用的顾虑，行业应用涉及金融、运营商、制造、零售、互联网、政府等多个行业。

1.3 产品优势

1.3.1 一键水平扩容或缩容

得益于平凯数据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，扩容或者缩容过程中对应用运维人员透明。

1.3.2 金融级高可用

采用计算与存储分离的多副本存储，保证系统持续高可用。

- 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，保证系统持续高可用。

1.3.3 实时 HTAP

提供行存储引擎 TiKV 和列存储引擎 TiFlash，TiFlash 通过 Multi-Raft Learner 协议实时从 TiKV 复制数据，保证二者之间数据强一致。TiKV 和 TiFlash 可按需部署在不同的机器，解决 HTAP 资源隔离的问题。

1.3.4 高度兼容 MySQL 及 MySQL 生态

高度兼容 MySQL 及 MySQL 生态, 应用无需或者修改少量代码即可从 MySQL 迁移到平凯数据库。

- 兼容 MySQL 协议
- 提供 MySQL 常用功能
- 兼容 MySQL 生态
- 提供丰富的数据迁移工具帮助应用便捷完成从 MySQL 无缝迁移到平凯数据库

1.3.5 云原生的分布式数据库服务

为客户提供云环境下的分布式数据库服务，支持与亚马逊、国内主流云厂商的部署与集成，充分利用和考虑云平台的特性，实现按需使用、规模化扩展以及运维简化。

- 在云环境简单点击即可部署和管理平凯数据库集群，大大提升初创公司的业务效率。
- 计算能力和存储容量可以分别独立扩展，适应更多样的业务需求。
- 完整支持平凯数据库的内核功能特性，为生产级别的 OLTP 和 OLAP 工作负载提供完整的 HTAP 支持。
- 也可通过专用的云主机、云上网络，以及企业级加密保障安全；通过跨可用区的部署和备份策略保障高可用和高弹性。

1.4 产品概述

1.4.1 内核模块

在内核设计上，平凯数据库将系统拆分成了多个模块，各模块之间互相通信，组成完整的分布式数据库系统。对应的内核功能模块如下图所示。

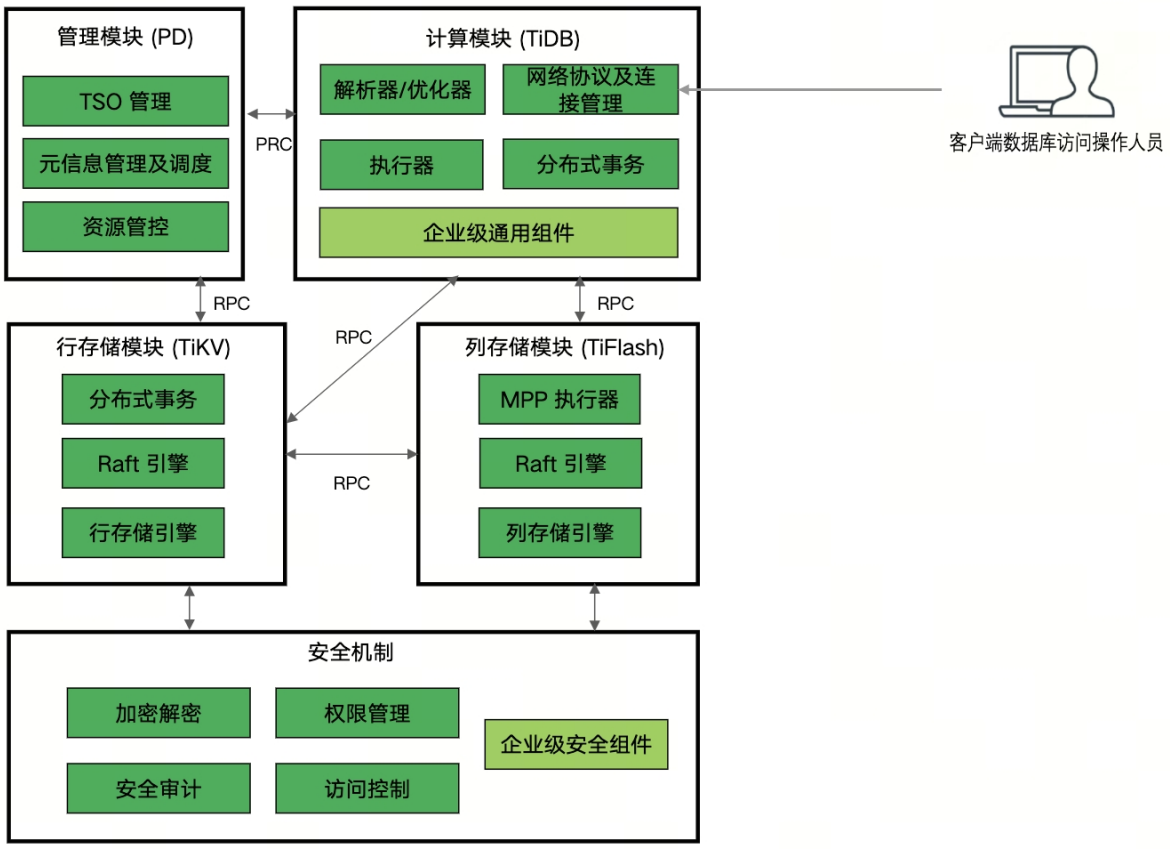

图 1: TiDB OpenCore Architecture

1.4.1.1 计算模块 TiDB

计算模块的 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。计算层本身是无状态的，实践中可以启动多个计算实例，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计算实例上以达到负载均衡的效果。计算模块本身并不存储数据，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。

1.4.1.2 管理模块 PD (Placement Driver) Server

整个平凯数据库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 Dashboard 仪表盘界面，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。

1.4.1.3 存储模块 TiKV 和 TiFlash

提供行存储引擎 TiKV 和列存储引擎 TiFlash。

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，这也是平凯数据库在 SQL 层面支持分布式事务的核心。平凯数据库的 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。

1.4.2 企业版模块

在企业级交付能力设计上，平凯数据库提供了企业级图形化组件、企业级安全组件、企业级通用组件，以及周边工具等，组成完整的数据库系统。对应的功能模块如下图所示。

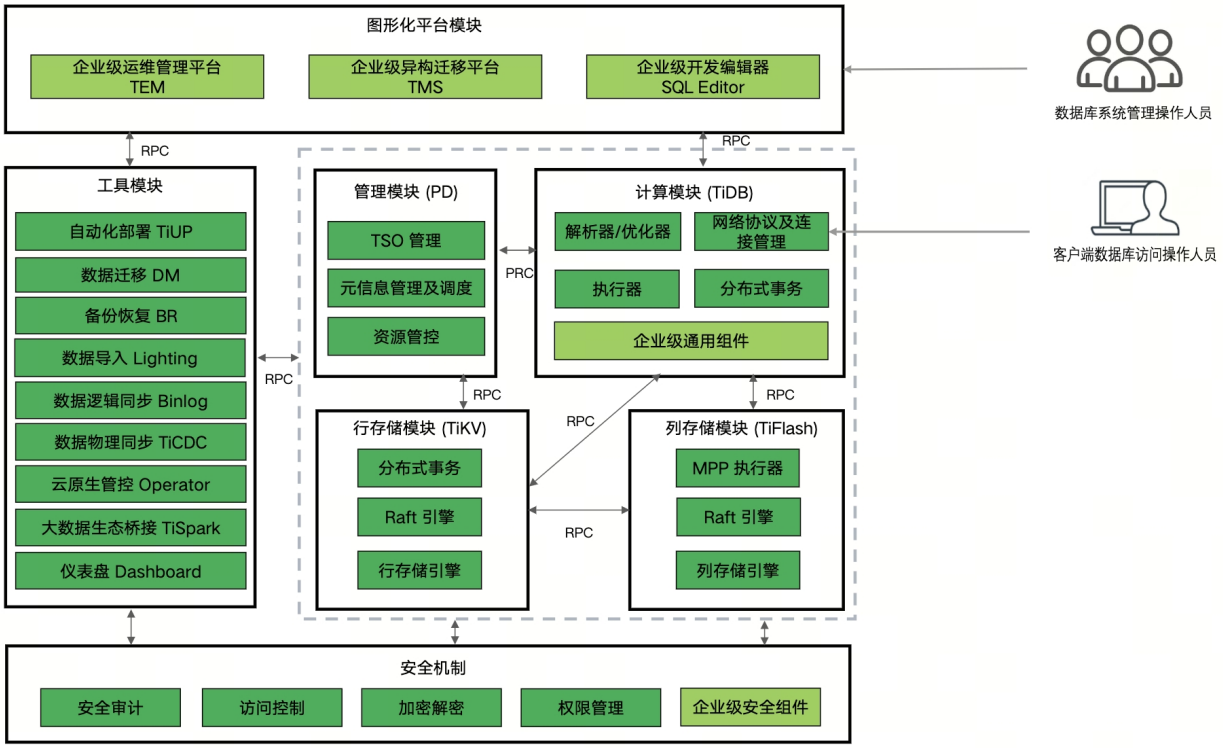

图 2: 平凯数据库企业级组件架构

1.4.2.1 企业级图形化组件

平凯数据库集群为企业级用户提供了图形化操控组件，包含：企业级运维管理平台（TEM）、企业级异构数据迁移平台（TMS）、企业级开发编辑器（SQL Editor），从运维管理，迁移，开发多个方面降低用户的使用难度和技术学习成本，可通过一个平台窗口，对多个平凯数据库集群进行集中的管理、差异化策略的数据全链路迁移、高效的 SQL 开发编辑，满足大规模、复杂和安全性要求较高的企业环境中的数据管理需求。

1.4.2.2 企业级安全组件

平凯数据库集群为企业级用户提供了增强安全组件，包含：完整的审计安全、增强的访问安全管理、数据的加解密、更细粒度的权限控制。这些组件涵盖了从身份验证和授权到审计和加密等各个方面，确保平凯数据库在客户的是生产系统使用过程中，数据库系统免受潜在威胁和攻击，数据的保密性、完整性和可用性，提供全面的安全性。

1.4.2.3 企业级通用组件

平凯数据库集群为企业级用户提供了商业化需求对技术内核基础能力之上的企业级通用组件，包含：兼容 MySQL 的存储过程、中文字符集 GB18030-2022、国密算法、存储级数据备份、内核性能提升与优化等方面，在符合产业技术标准对数据库的企业级能力的要求下，面向严肃的企业关键业务生产场景，提供易用、可靠、安全、高效的数据库系统服务能力，有助于支撑客户的战略目标、业务需求和合规性要求。

1.4.2.4 周边工具

平凯数据库集群内置了多种数据库操作的便捷工具，包含：自动化部署、逻辑数据同步、物理数据同步、备份恢复、数据导入导出、云原生管控、大数据库生态桥接、仪表盘等数据库全生命周期链路上的实操工具，极大的提升了平凯数据库的使用效率，降低了操作使用的错误率，让用户可以轻松快速的上手使用平凯数据库产品。

1.5 使用场景

- 一键水平扩容或者缩容

  得益于平凯数据库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

  为云设计的分布式数据库，通过 Operator 可在公有云、私有云、混合云中实现部署工具化、自动化，依托公有云提供开箱即用的云原生数据库服务。

- 兼容 MySQL 协议和 MySQL 生态

  兼容 MySQL 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 迁移到平凯数据库，提供丰富的数据迁移工具帮助应用便捷完成数据迁移。

1.6 基本概念

1.6.1 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。平凯数据库通过 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。平凯数据库在写入数据之前，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。平凯数据库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。在平凯数据库中，事务一旦提交成功，数据全部持久化存储到存储节点，此时即使平凯数据库服务器宕机也不会出现数据丢失。

1.6.2 GC

平凯数据库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。

1.6.3 MVCC

Multiversion Concurrency Control，简称 MVCC。MVCC 意图解决读写锁造成的多个、长时间的读操作饿死写操作问题。每个事务读到的数据项都是一个历史快照，并依赖于实现的隔离级别。写操作不覆盖已有数据项，而是创建一个新的版本，直至所在操作提交时才变为可见。快照隔离使得事务看到它启动时的数据状态。

1.6.4 PD (Placement Driver)

管理节点：整个平凯数据库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，提供仪表盘（Dashboard）操控界面，并为分布式事务分配事务 ID。

1.6.5 TiDB

计算节点：在 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。

1.6.6 TiFlash

列存储节点：平凯数据库 HTAP 形态的关键组件，它是 TiKV 的列存扩展，在提供了良好的隔离性的同时，也兼顾了强一致性。

1.6.7 TiKV

行存储节点：一个分布式事务型的键值数据库，提供了满足 ACID 约束的分布式事务接口，并且通过 Raft 协议保证了多副本数据一致性以及高可用。TiKV 作为平凯数据库的存储层，为用户写入平凯数据库的数据提供了持久化以及读写服务，同时还存储了平凯数据库的统计信息数据。

1.6.8 Region/Peer/Raft Group

行存储节点 TiKV Server 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。

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。

1.6.9 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。

1.6.10 TiUP

平凯数据库的包管理器，管理着平凯数据库众多的组件，如计算节点、管理节点、存储节点等。你想要运行平凯数据库中任何组件时，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。

1.6.11 Data Migration (DM)

一个一体化的数据迁移任务管理工具，支持从与 MySQL 协议兼容的数据库（MySQL、MariaDB、Aurora MySQL）到平凯数据库的数据迁移，可以降低数据迁移的运维成本。

1.6.12 Backup Restore (BR)

一个物理备份恢复工具，可以高速的将平开数据库的海量数据进行全量和增量的文件级备份，效率远远快于逻辑备份的执行效果，是平凯数据库在大数据量环境下的备份恢复的首选工具。

1.6.13 Lightning

Lightning 是用于从静态文件导入 TB 级数据到平凯数据库集群的工具，常用于平凯数据库集群的初始化数据导入。

1.6.14 TiCDC

TiCDC 是一款平凯数据库增量数据同步工具，通过拉取上游 TiKV 的数据变更日志，可以将数据解析为有序的行级变更数据输出到下游。提供多个平凯数据库集群，跨区域数据高可用和容灾方案，保证在灾难发生时保证主备集群数据的最终一致性。同步实时变更数据到异构系统的服务，为监控、缓存、全文索引、数据分析、异构数据库使用等场景提供数据源。

1.6.15 Binlog

Binlog 是一个用于收集平凯数据库集群的 binlog，并提供准实时备份和同步功能的商业工具。可以同步平凯数据库集群数据到其他数据库，以及备份平凯数据库集群数据，同时可以用于平凯数据库集群故障时恢复。

1.6.16 Operator

一个 Kubernetes 上的平凯数据库集群自动运维系统，提供包括部署、升级、扩缩容、备份恢复、配置变更的平凯数据库全生命周期管理。借助平凯数据库 Operator，平凯数据库可以无缝运行在公有云或私有部署的 Kubernetes 集群上。

1.6.17 TiSpark

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。
