ArchData

技术峰会北京站

主办方：中生代技术 FRESHMAN TECHNOLOGY 快CTO 互联网创业技术服务平台

2017年9月24日北京海淀区丹棱街5号微软亚太研发中心一号楼一层 故宫会议室

PingCAP

分布式数据库在金融行业的创新实践

余军 | PingCAP

PingCAP

关于我

- 余军
- Chief Solution Architect, Financial Industry
- 20y+ Open Source background
- Focus on Distributed system and High Availability

PingCAP

数据库系统的技术演进

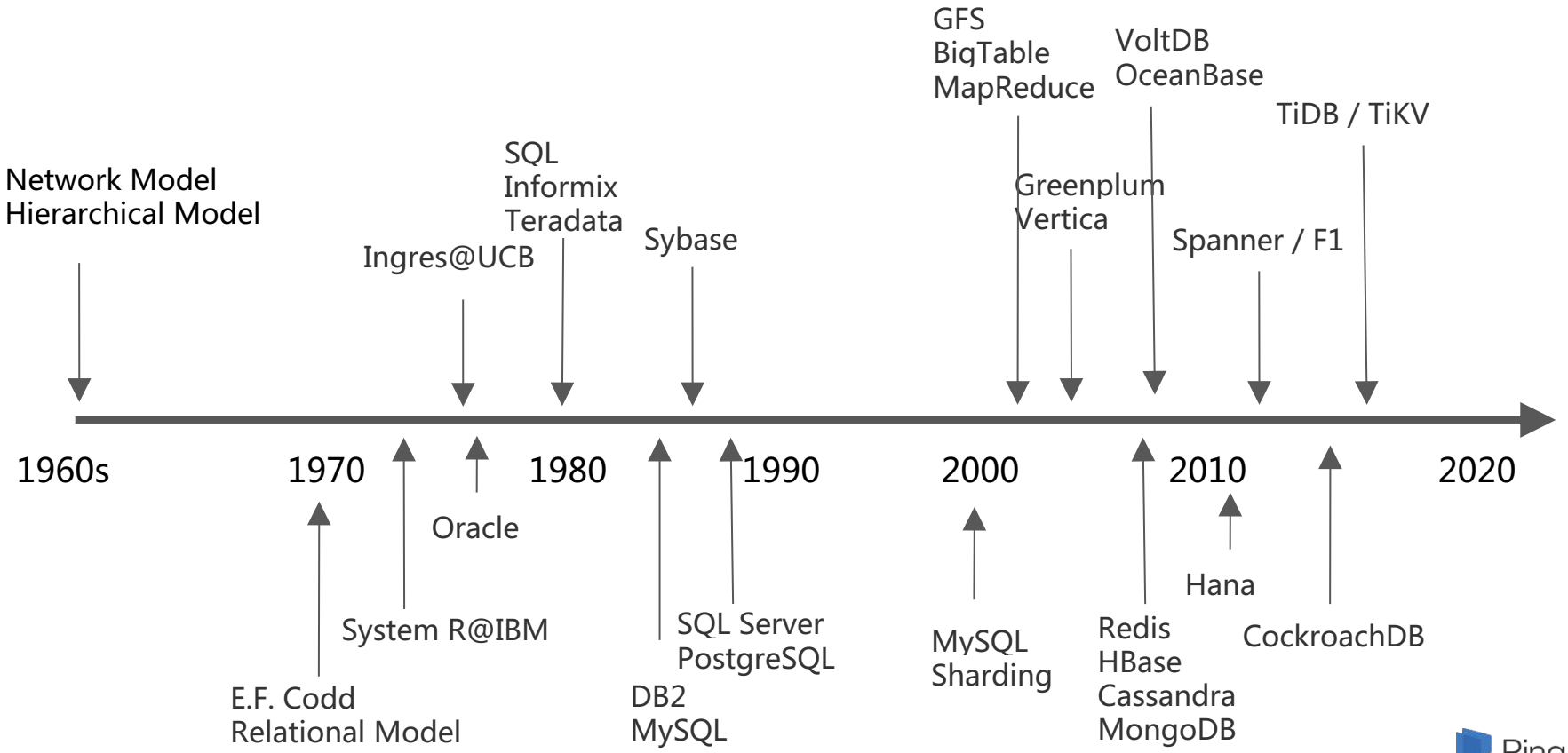
数据库技术发展演进（1/2）

2008 年以前	2008 年至 2013 年	2013 年以后
单机关系型（SQL）	**分布式非关系型（NoSQL）**	**分布式关系型（NewSQL）**
• 背景：应用最为广泛的数据库；能很好的解决复杂的数据运算及表间处理；多用于银行、电信等传统行业复杂业务逻辑场景中，以 Oracle 为代表	• 背景：随着搜索 / 社交的发展，数据量爆发增长，传统数据库高成本，无法线性扩容问题日益突显；分布式及 NoSQL 开始快速发展，如 MongoDB，HBase	• 背景：随着互联网向银行、电信、电力等方向的渗透，传统行业数据量迅速提升，需要同时满足低成本、线性扩容及能够处理交易类事务的新型数据库，大数据的存储刚需不可避免
• 挑战：**成本高**，随着数据量增加，只能通过购买更贵更好的服务器；**无法线性扩容**，海量数据下处理能力大幅下降	• 挑战：擅长简单读写，**无法处理交易类数据及复杂业务逻辑**的特性限制其在非互联网领域的发展	• 挑战：基于 Google Spanner/F1 论文，基础软件最前沿的领域之一，技术门槛最高

NewSQL: 兼具 NoSQL 扩展性又不丧失传统关系型数据库 ACID 特性的分布式数据库

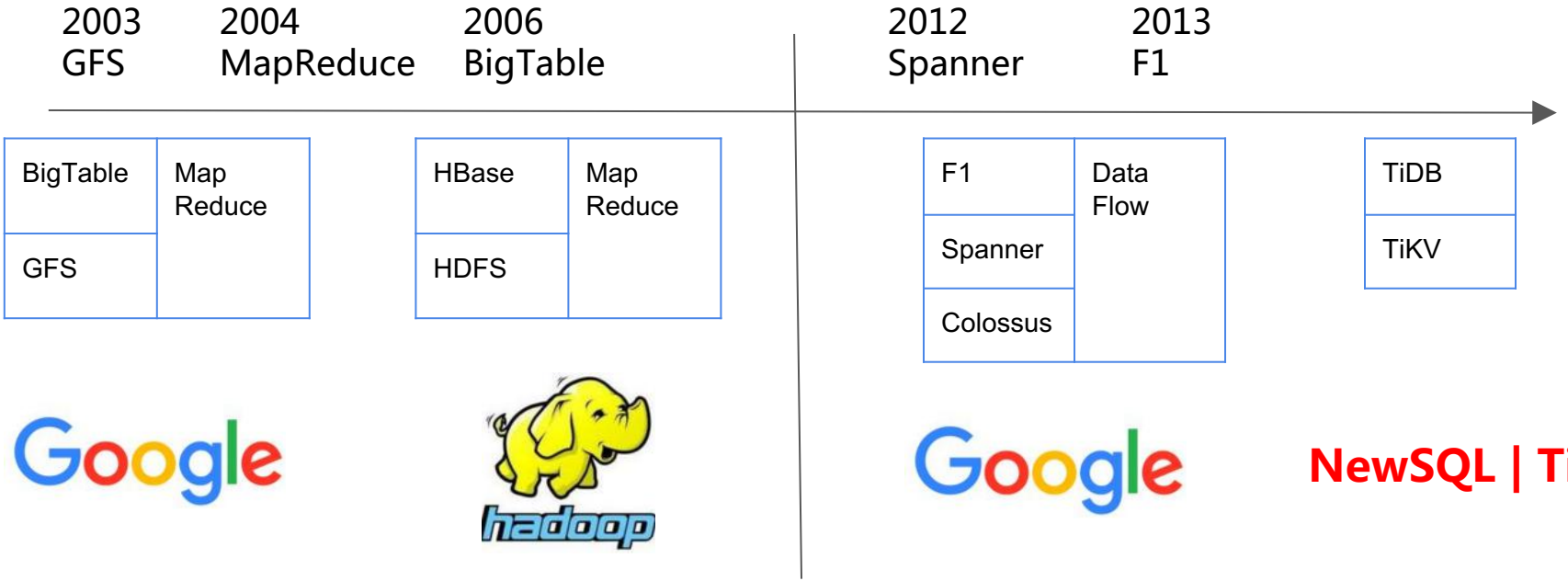
PingCAP

数据库技术发展演进（2/2）



PingCAP

Google - 大规模分布式计算领域的领跑者



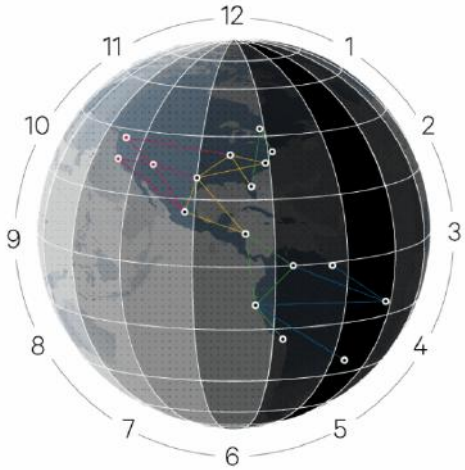
Google 十年前基于内部分布式处理框架发表的三篇论文奠定了大数据分析处理基石。
开源社区以此为基础打造了 Hadoop.

Google 内部新一代分布式处理框架，于 12/13 年发表相关论文，奠定下一代分布式 NewSQL 的理论和工程实践基石。
PingCAP 以此为基础打造了 TiDB & TiKV.

PingCAP

Google Spanner | F1 - 第一个真正意义上 NewSQL 数据库

- 全球级分布式关系型数据库，数十万机器组成一个超大的数据库集群
  - Spanner - 有状态分布式 Key-Value 数据库
  - F1 - 无状态分布式 SQL 解析器
- 支撑 Google Adwords、Wallet 等核心金融业务
- 根据业务压力，水平无限扩展或者伸缩，底层七副本，保证任意一个数据中心宕机，底层自动切换，上层业务不中断，无需人工介入
- 2017年2月，Google 在其 GCP 公有云平台正式提供 Cloud Spanner 服务，并于5月 GA。



PingCAP

PingCAP

TiDB 优势

TiDB：Google Spanner 和 F1 的开源实现

PingCAP

新一代分布式关系型 NewSQL 数据库 TiDB

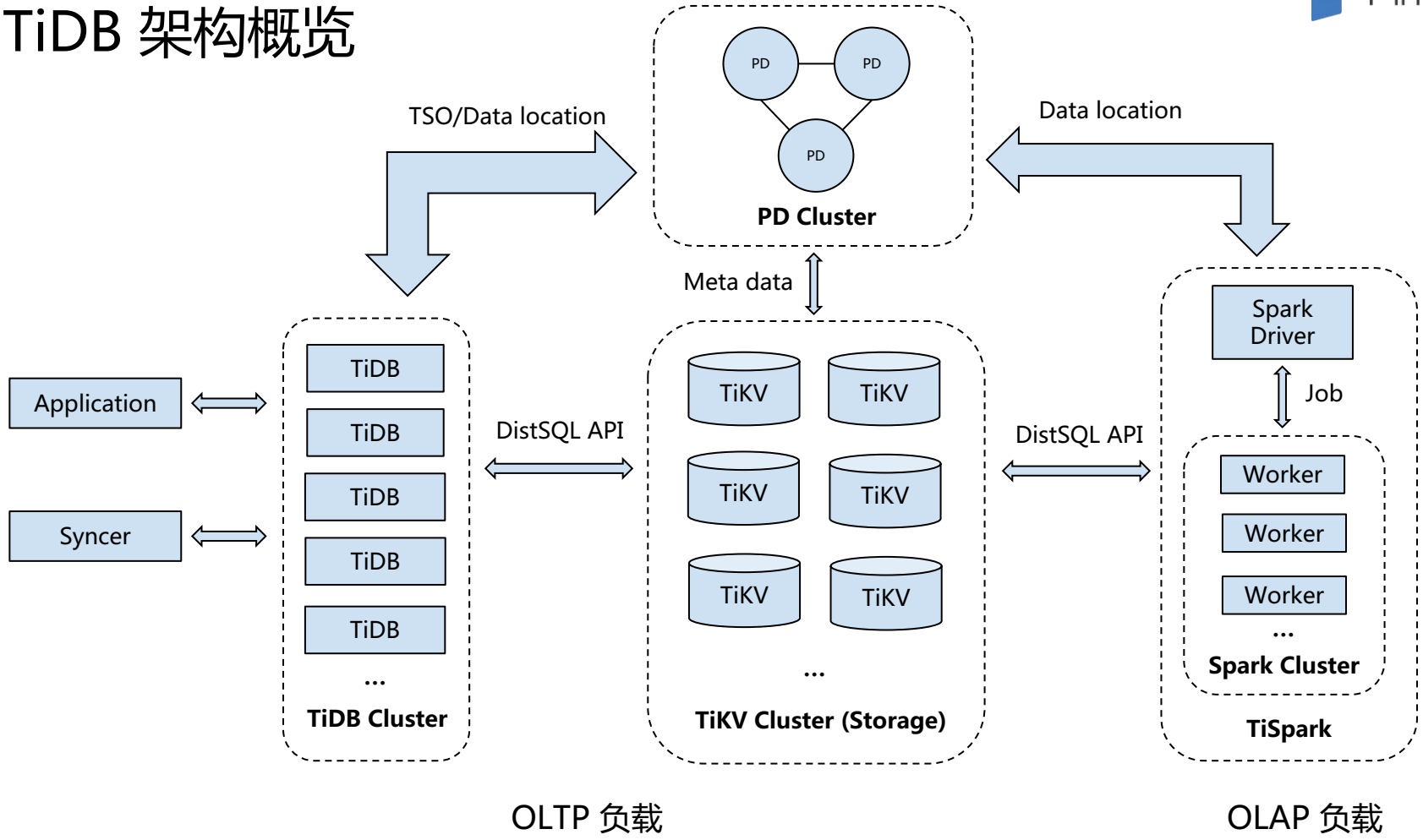
- 基于 2013 年 Google Spanner / F1 论文，在 Google 全球核心业务系统中大规模验证。
- 基于 2014 年 Stanford 工业级分布式一致性协议实现 Raft 博士论文，已成为事实工业标准。

核心 **NewSQL** 特性概括：

水平线性扩展、强一致分布式事务、故障自恢复的高可用（非主从）、真正跨数据中心多活

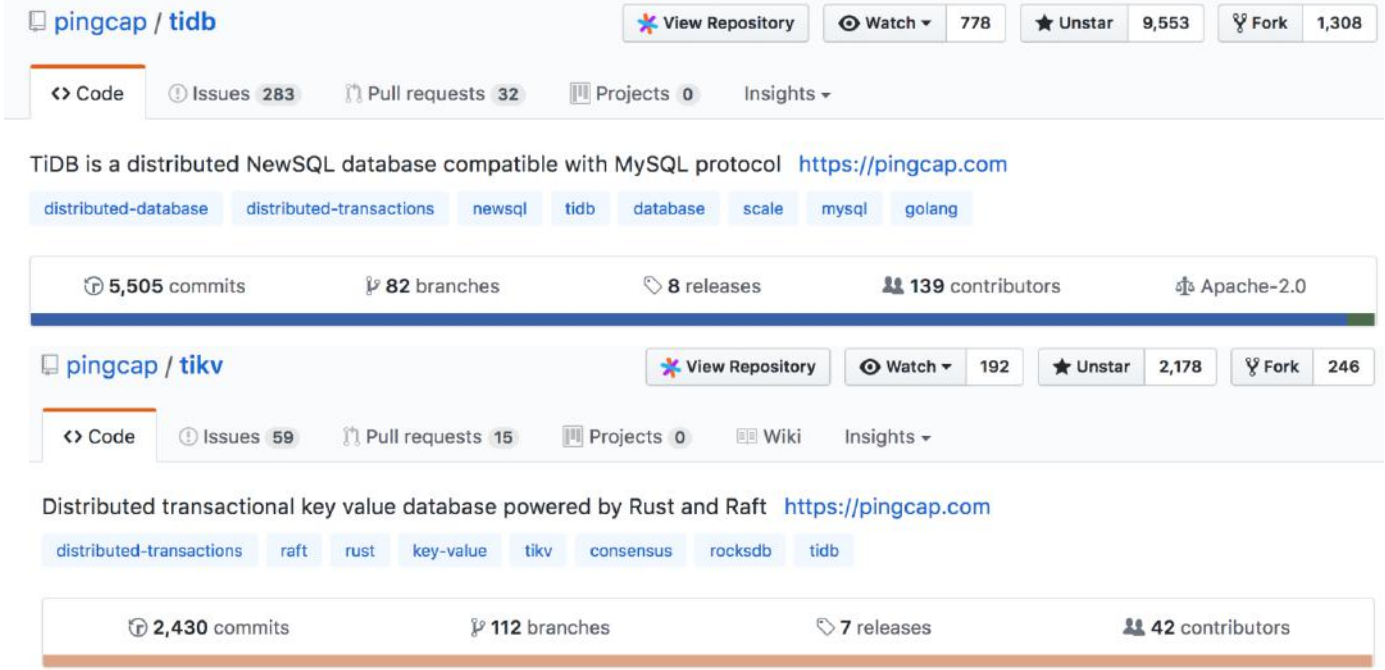
- PingCAP 是全球仅有的在该领域进行技术创新的两家公司之一（对标美国 CockroachDB ）
- 体系架构完全不同于传统的单机型数据库的理论，真正意义上的分布式架构
- **完全从头打造，并非基于 MySQL、PG 或任何数据库中间件进行改造、封装**

TiDB 架构概览


PD
PD
PD
PD Cluster
TSO/Data location
Data location
Meta data
Application
Syncer
TiDB
TiDB
TiDB
TiDB
TiDB
...
TiDB Cluster
DistSQL API
TiKV
TiKV
TiKV
TiKV
TiKV
TiKV
...
TiKV Cluster (Storage)
DistSQL API
Spark Driver
Job
Worker
Worker
Worker
...
Spark Cluster
TiSpark
OLTP 负载
OLAP 负载

TiDB 已成为数据库领域国际顶级开源项目（1/3）

PingCAP



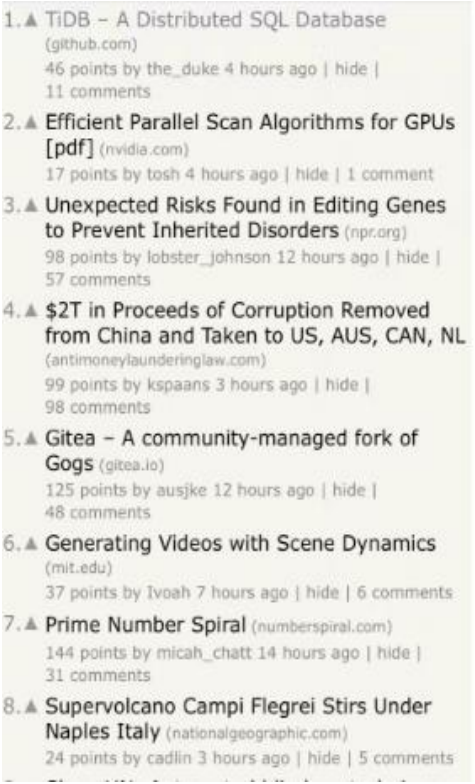
TiDB on Github（11000+ stars / 150+ contributors）

https://github.com/pingcap/tidb

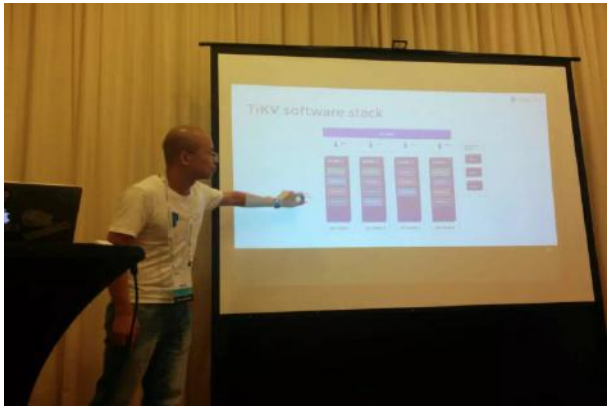
TiDB 是全球最成熟稳定的 Google Spanner / F1 的开源实现，是基础软件领域的重大创新，具有极高的工程难度，TiDB 既不是数据库中间件，也不是 SQL-On-Hadoop，是真正意义上的 NewSQL。

TiDB 已成为数据库领域国际顶级开源项目（2/3）

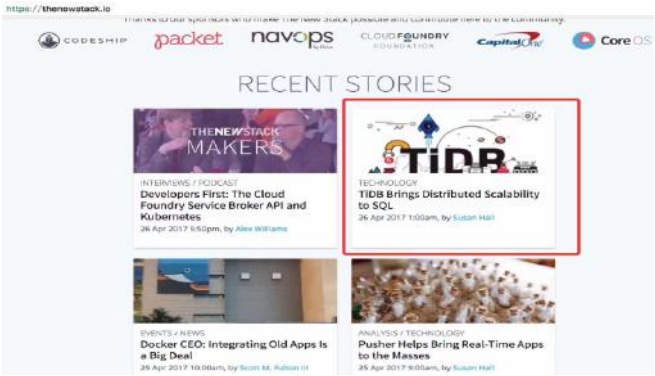
PingCAP



多次登上 Hacker News 首页



多次海外顶级会议分享

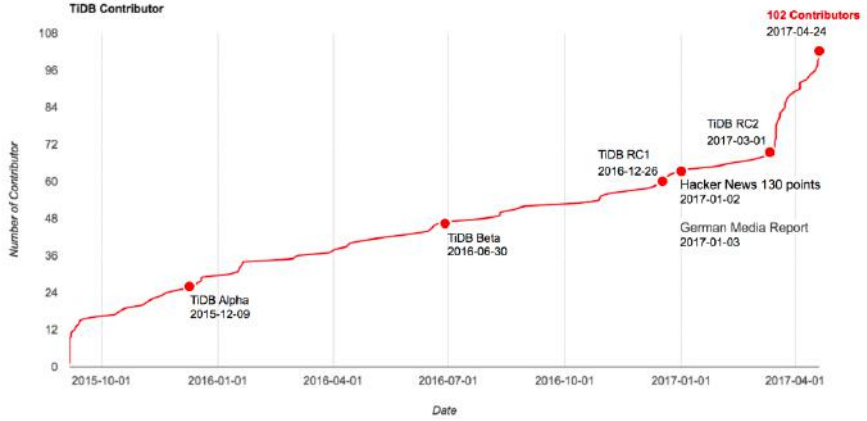


海外媒体广泛报道

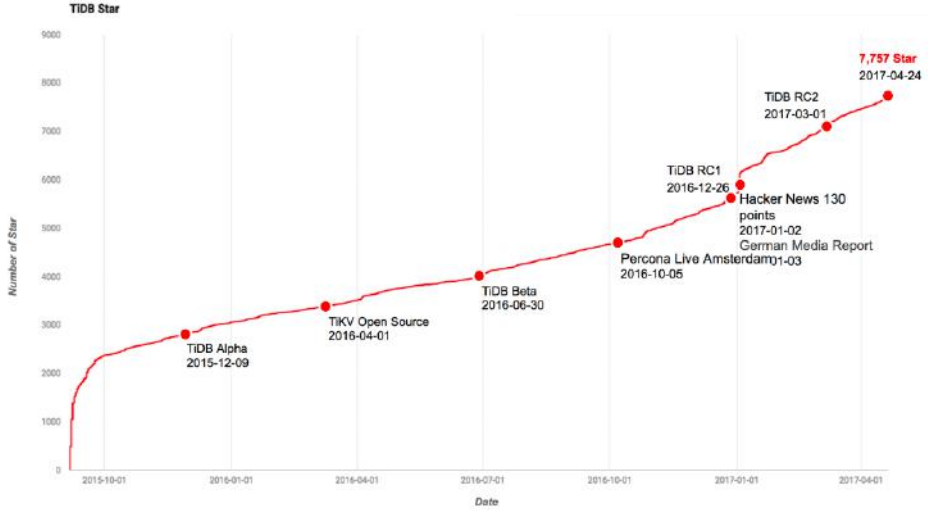
TiDB 已成为数据库领域国际顶级开源项目（3/3）

PingCAP

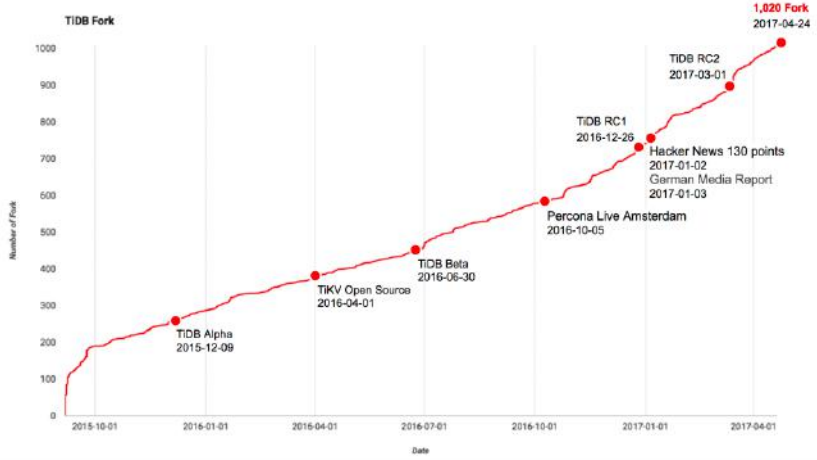
全球代码贡献者增幅



Github 软件项目 星级评定增幅



Github 项目 fork 数量增幅



TiDB 得到了国内外金融用户的高度关注

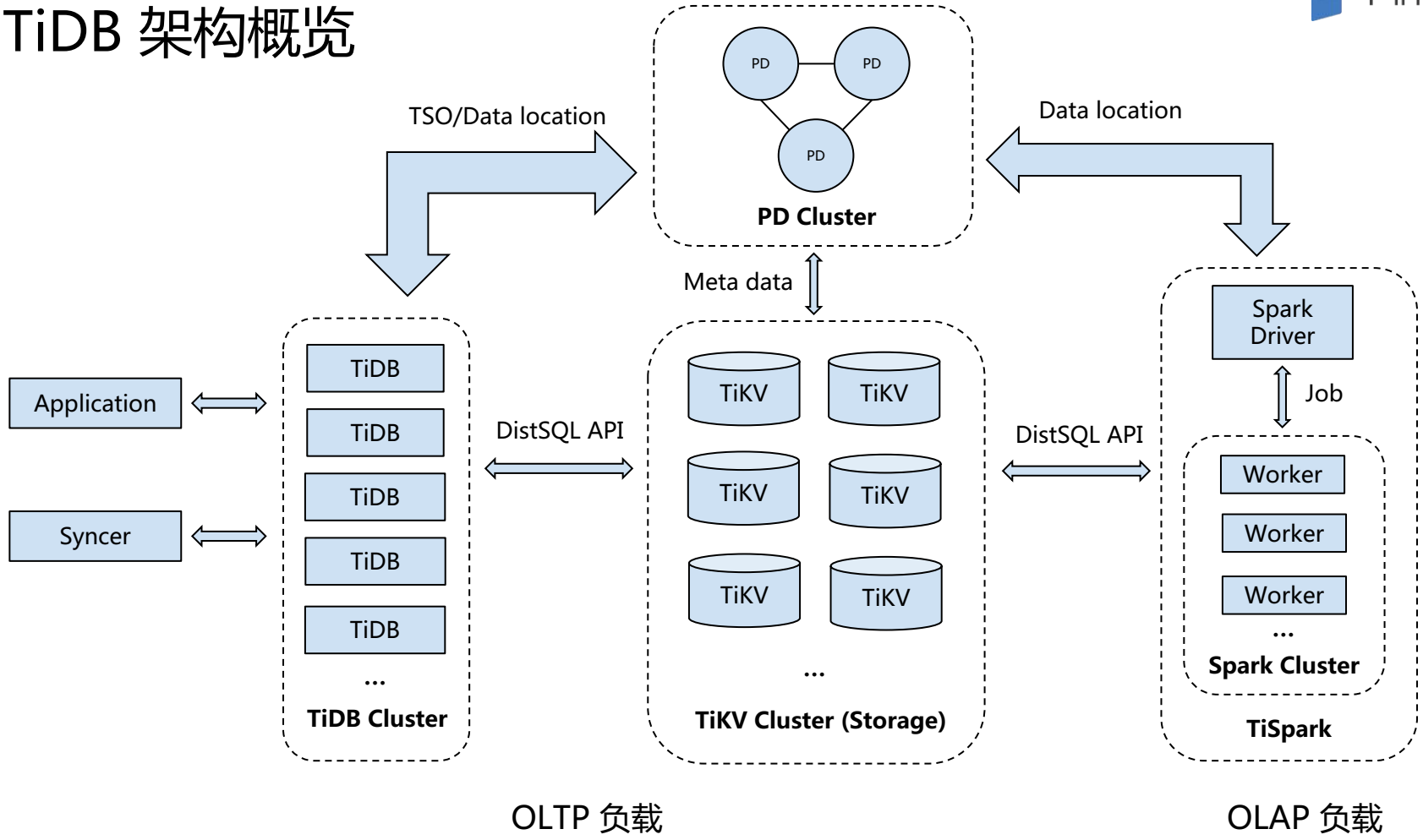
一组数据：

- 2016 年 7 月 推出 Beta2 版本
- 2017 年 9 月 推出 Pre-GA (准正式版)
- 以上 15 个月 迭代发布总计 8 个大版本，约 11114 次重要代码递交和更新。
- 来自国内外 30+ 金融客户的测试，验证和生产使用。
- 他们有大型国有商业银行，股份制商业银行，大型券商和保险，第三方电子支付企业，互联网金融企业 (个贷 / 理财 / 投顾)，科技金融(Fintech)企业以及金融征信企业等。
- 也有美国知名的移动支付集团及亚洲某大国最大的金融支付集团企业。
- 典型数据量规模在 5000万+/单表 至 50亿+/单表不等。
- 这些用户的应用 TiDB 的业务场景，主要聚焦在：在线交易，在线支付，移动支付，在线信贷，营销积分，实时风控，投资者服务，金融征信管理。

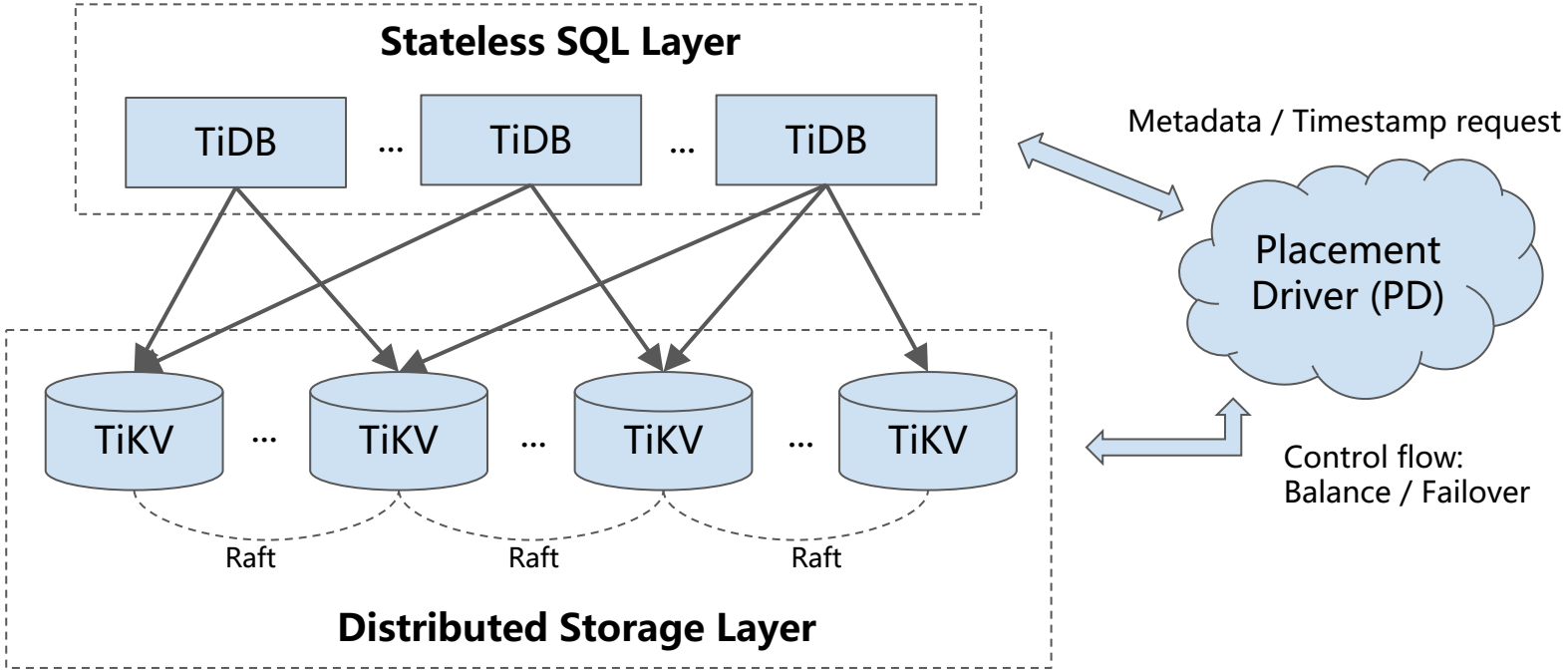
PingCAP

TiDB 技术精读

TiDB 架构概览


PD
PD
PD
PD Cluster
TSO/Data location
Data location
Meta data
Application
Syncer
TiDB
TiDB
TiDB
TiDB
TiDB
...
TiDB Cluster
DistSQL API
TiKV
TiKV
TiKV
TiKV
TiKV
TiKV
...
TiKV Cluster (Storage)
DistSQL API
Spark Driver
Job
Worker
Worker
Worker
...
Spark Cluster
TiSpark
OLTP 负载
OLAP 负载

TiDB OLTP 分布式架构



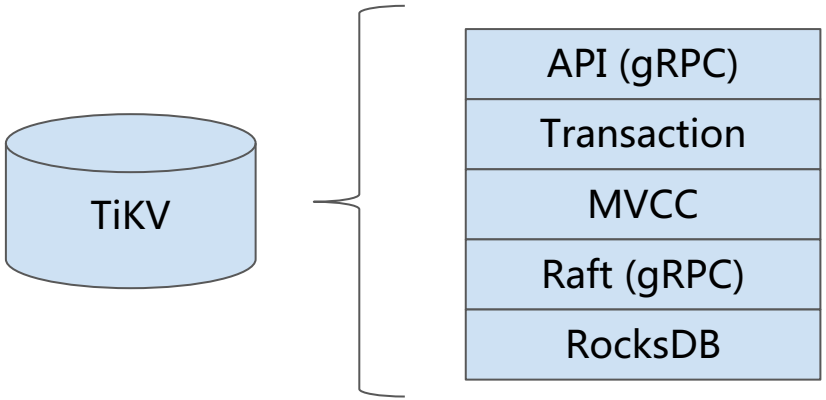
TiDB - **无状态的 SQL 层**（对标 F1）

TiKV - **分布式 KV 存储引擎**（对标 Spanner）

PD - **元信息管理、集群管理和调度**（拥有全局视角的调度模块）

TiKV - 分布式存储引擎（1/2）

- TiKV 是一个分布式且支持事务的 Key-Value 存储引擎
- 数据存储在 RocksDB 中
- 节点之间通过 Raft 协议 保持数据一致性
- 事务模型采用 Google 的 Percolator

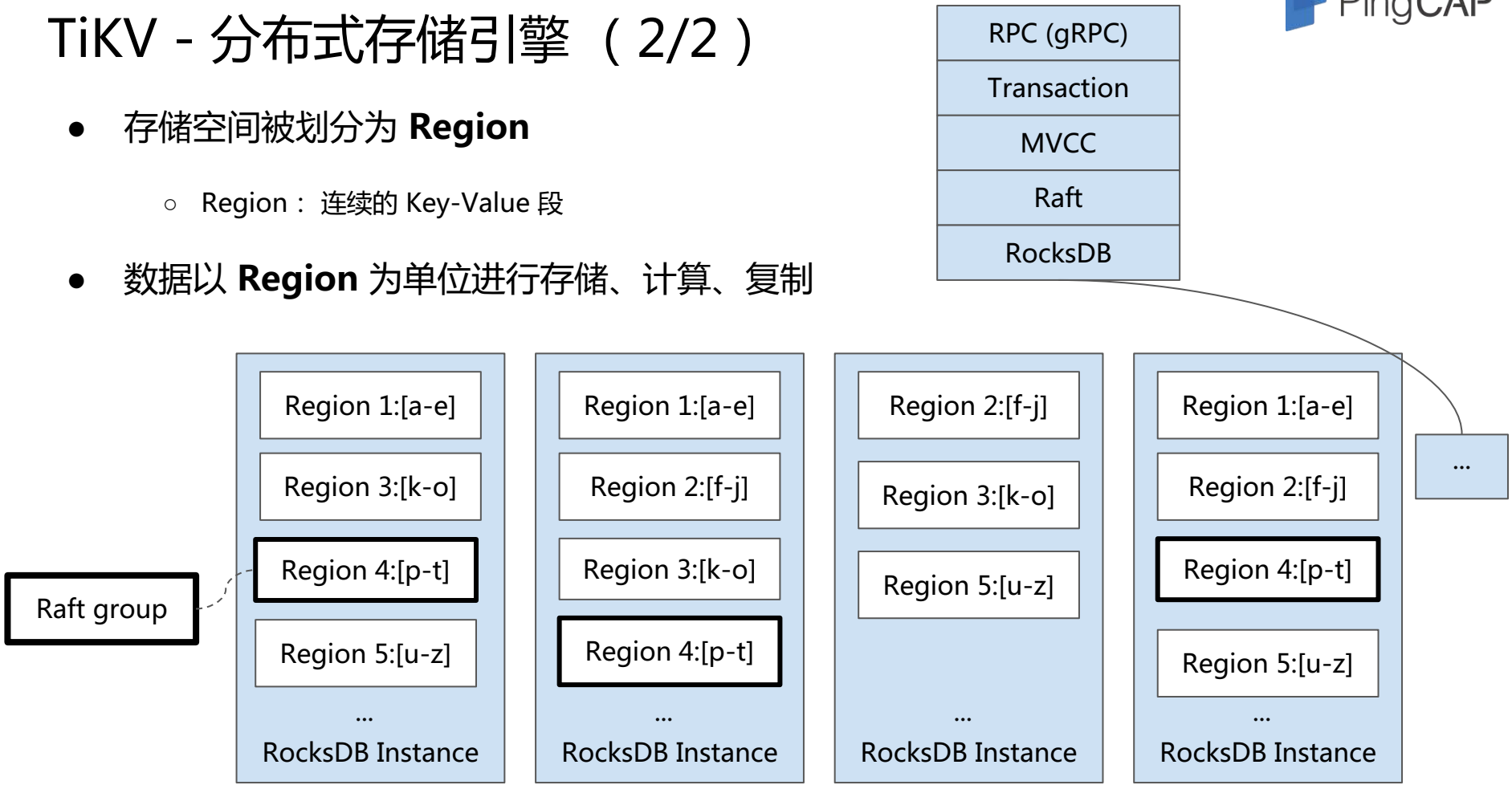


Transactional KV API
(https://github.com/pingcap/tidb/blob/master/cmd/benchkv/main.go)

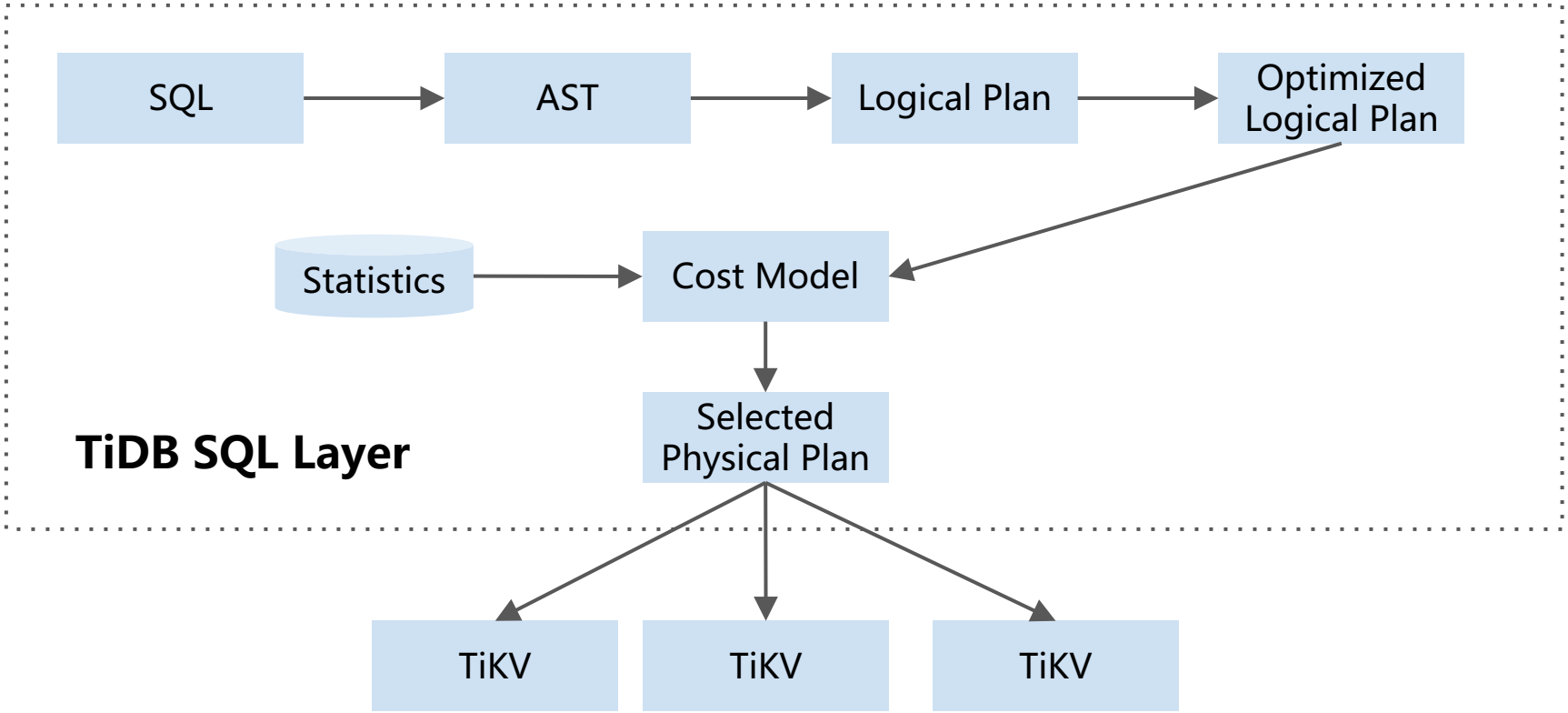
Raw KV API
(https://github.com/pingcap/tidb/blob/master/cmd/benchraw/main.go)

TiKV - 分布式存储引擎（2/2）

- 存储空间被划分为 **Region**
  - Region： 连续的 Key-Value 段
- 数据以 **Region** 为单位进行存储、计算、复制

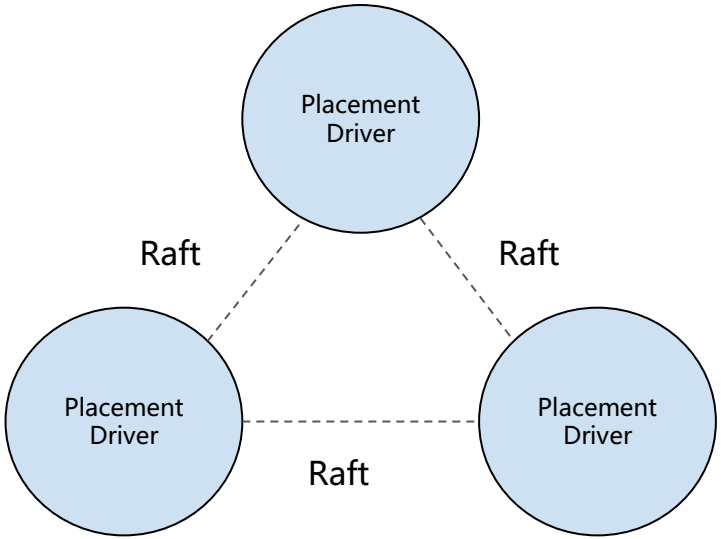


TiDB SQL - 分布式SQL引擎



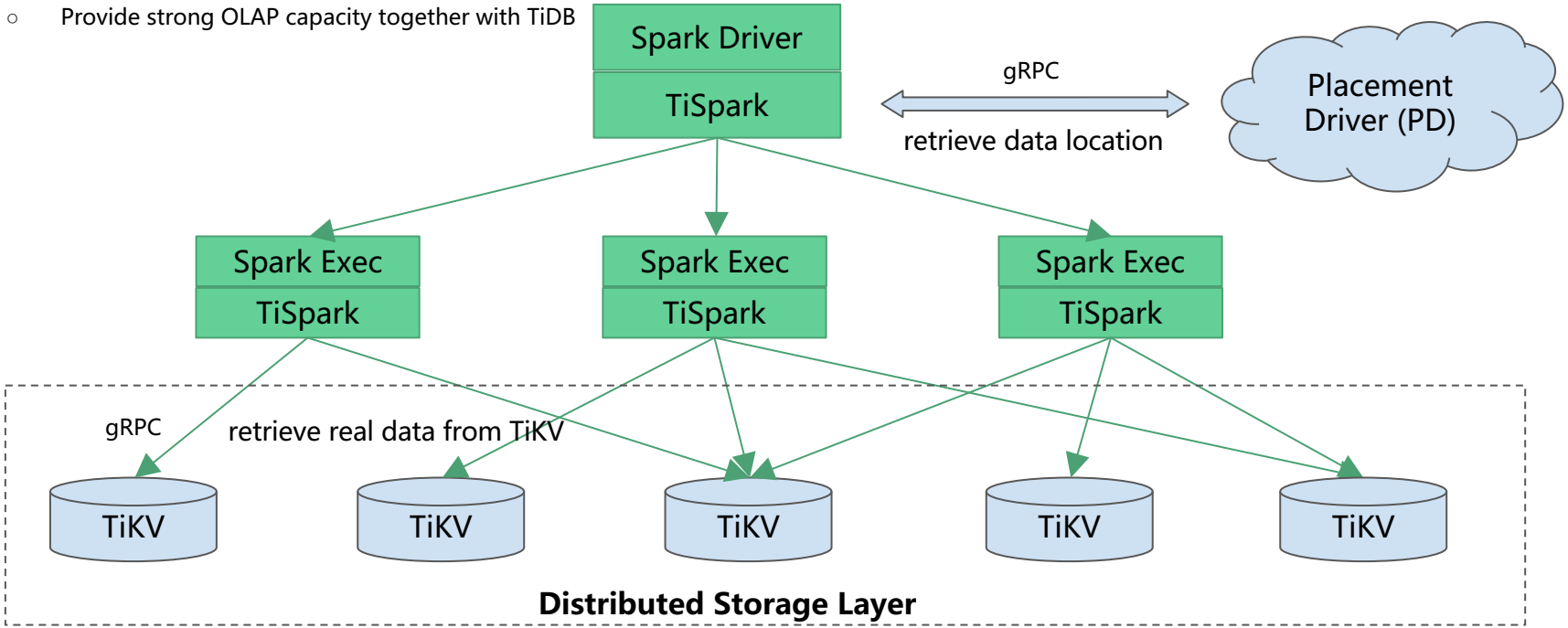
PD - 分布式集群调度和管理

- 和 Google Spanner 类似的设计
- 为整个集群的管理提供 - “上帝视角”
- 存储集群元数据 meta data
- 维护复制副本的约束
- 集群数据的迁移，自动平衡和调度
- 全局时间戳分配
- Leader region 的高性能调度
- 自身也是无单点故障的集群



TiDB OLAP 分布式架构

- TiSpark = Spark SQL on TiKV
  - Spark SQL directly on top of a distributed Database Storage
- Hybrid Transactional/Analytical Processing(HTAP) rocks
  - Provide strong OLAP capacity together with TiDB



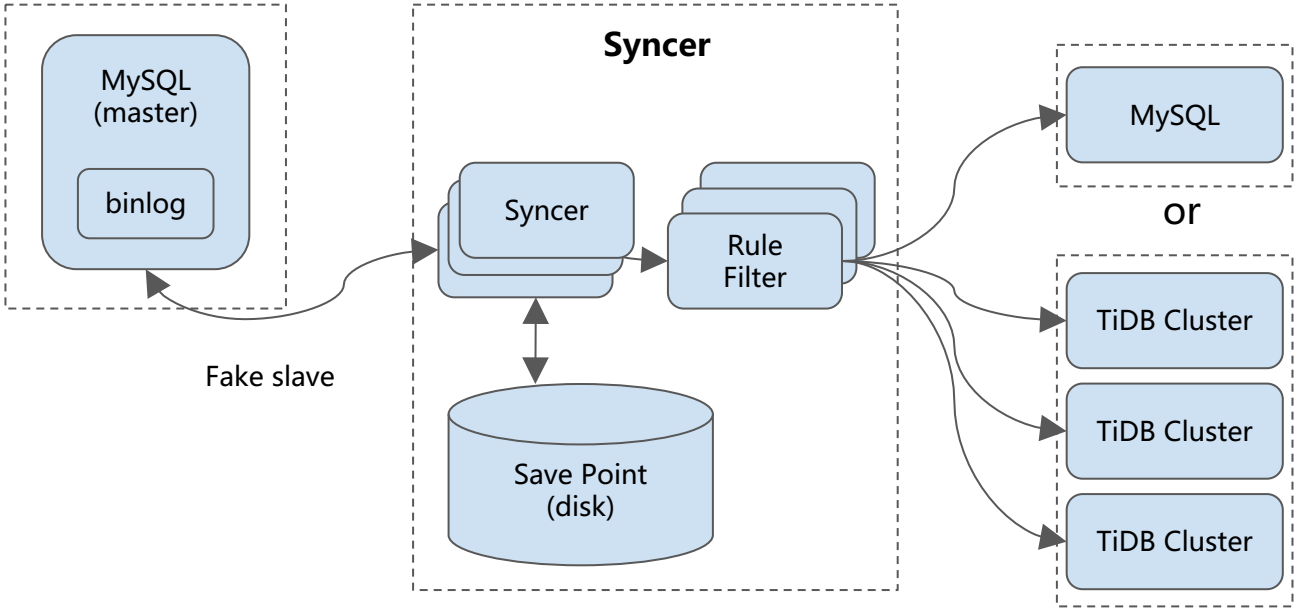
TiDB 生态工具

PingCAP

MySQL 兼容性	TiDB 是 MySQL 兼容的，MySQL 社区所有的周边工具都可以使用（驱动、管理工具、应用框架），比如 myloader / mydumper / MySQL JDBC Driver / Navicat / Workbench / WordPress / 主流 ORM 框架 等。
数据迁移 **Mydumper / Loader**	全量导出：建议使用 Mydumper 进行 MySQL 数据库的导出（多线程导出） 全量导入：建议使用 PingCAP 提供的 Loader 进行数据导入（支持断点续传、错误重试）
数据实时同步 **TiDB Syncer**	TiDB Syncer 可以实时读取 MySQL 的 binlog 作为 MySQL 的 Slave，便于用户同步对比 TiDB 的各种性能，利于用户的同步测试、验证、平滑上线。TiDB Syncer 同时支持正则表达式，可以把已经分库分表后的 MySQL 合并同步到一个 TiDB 集群。
数据实时备份 **TiDB binlog**	TiDB binlog 具备开放接口（protobuf），可以同时输出到 MySQL、TiDB 或者其他数据存储，安全备份的同时，给用户最大的自由度，避免 Vendor-Lockin
实时复杂分析和查询 **TiSpark**	**TiSpark 是 PingCAP 推出的为了解决用户复杂 OLAP 需求的产品。**借助 Spark 平台本身的优势，同时融合 TiKV 分布式集群的优势，和 TiDB 一起为用户一站解决 HTAP （Hybrid Transactional / Analytical Processing）需求。
云数据库接口 **TiDB-Operator**	TiDB-Operator 把 TiDB 和 Kubernetes 连接在一起。借助 TiDB-Operator，使得 K8S 可以调度有状态的服务，能方便快捷的同公有云 / 私有云 进行整合，提供 DBaaS 服务。

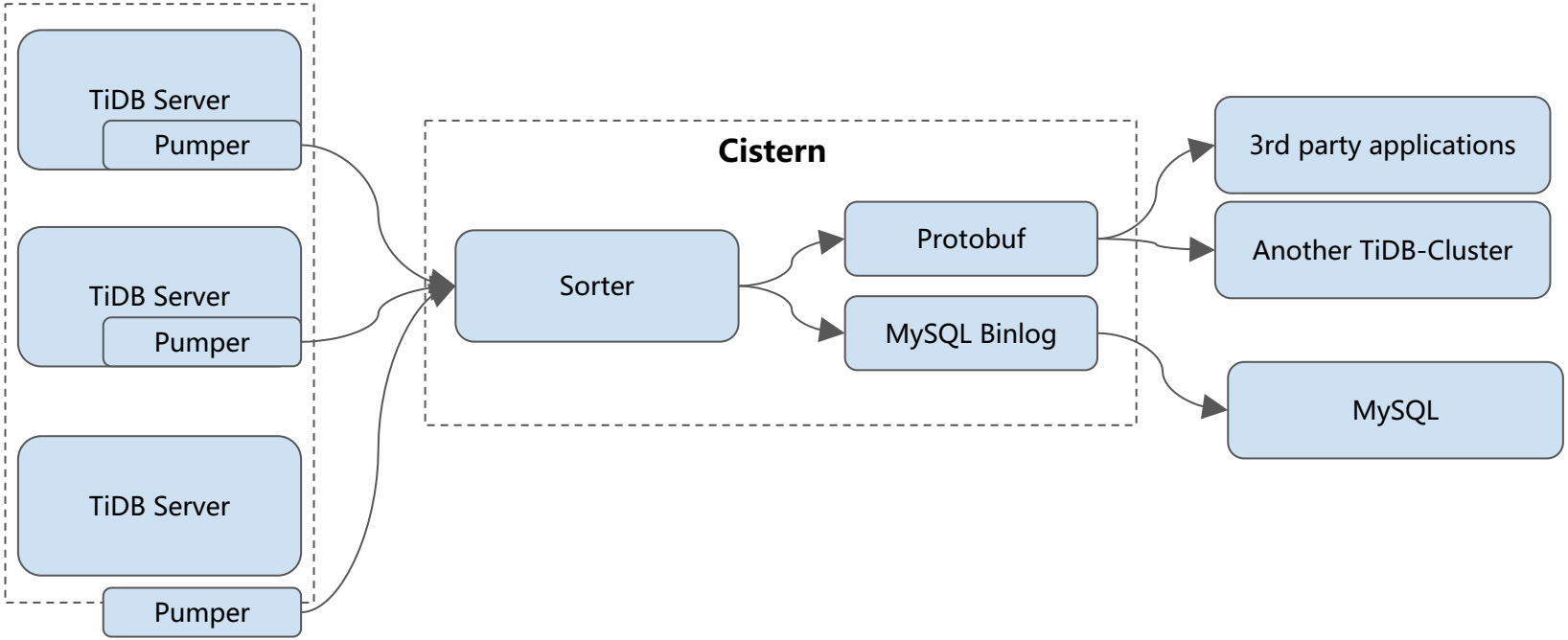
TiDB 生态工具：Syncer （数据实时同步）

- 从 MySQL 实时同步数据到 TiDB
- 使 TiDB 作为 MySQL 的一个从库
- 支持正则表达式，可以合并分库分表后 MySQL 集群



TiDB 生态工具：TiDB-Binlog（数据实时备份）

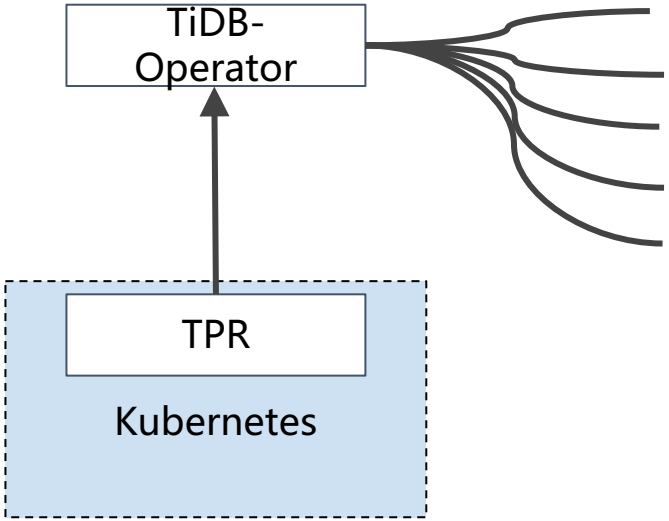
- 订阅 TiDB 的增量数据
- 输出格式可以是 protobuf 或者 是 MySQL binlog (WIP)



TiDB 生态工具：TiDB-Operator （云数据库接口）

PingCAP

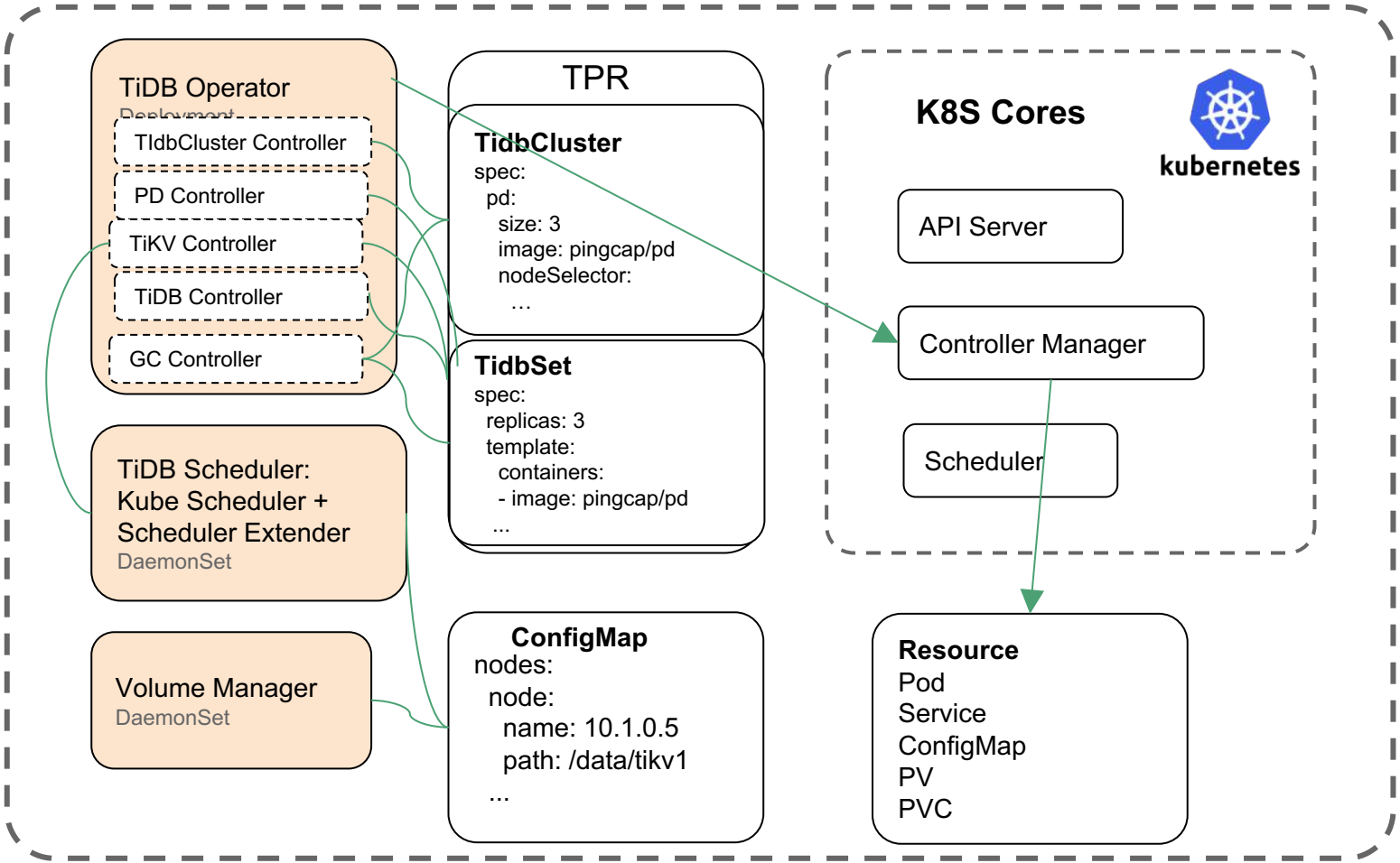
- 连接 TiDB 和 Kubernetes，将 TiDB 运维知识写入 TiDB-Operator，使得 K8S 可以更好地调度有状态的服务
- 轻松整合公有云 / 私有云，提供 DBaaS 服务



- Create
- Rolling update
- Scale in/out
- Failover
- Backup/Restore

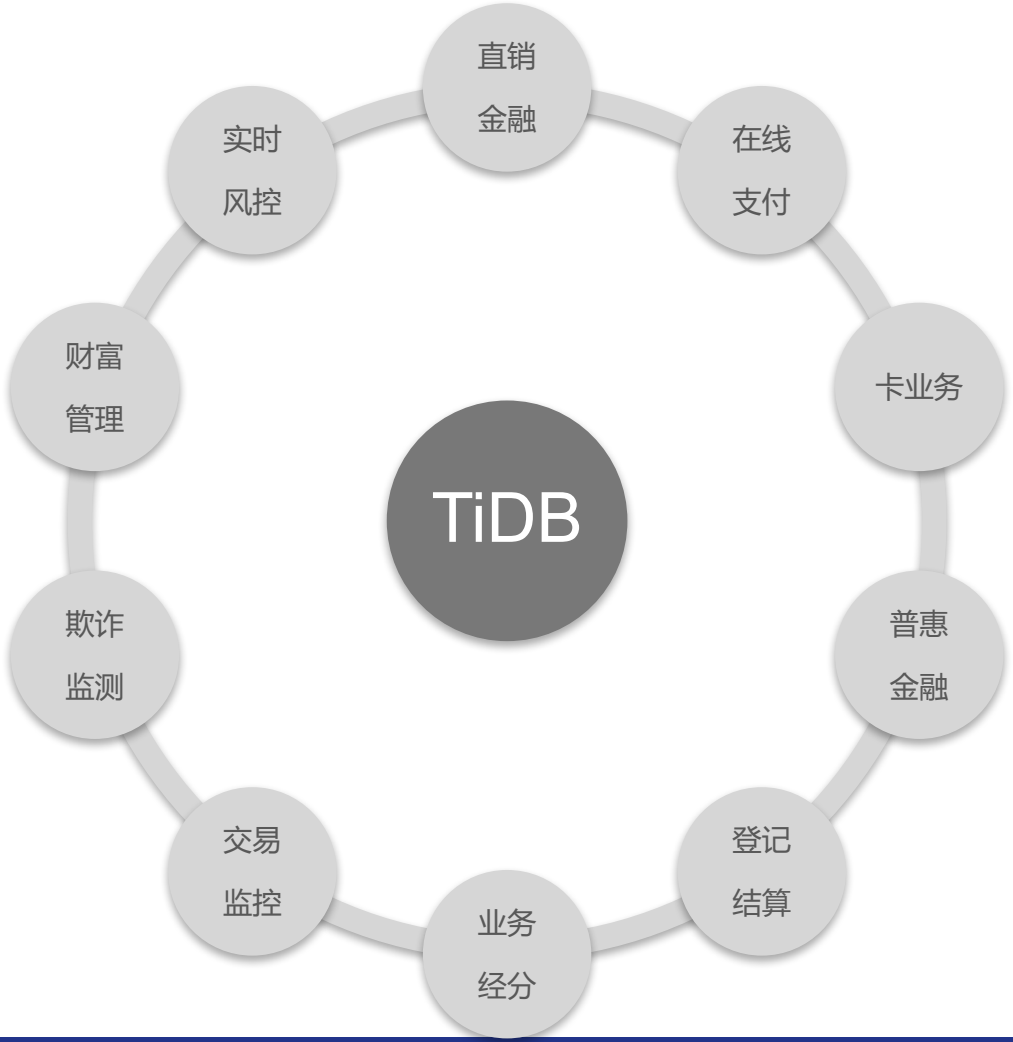
TiDB-Operator 整体架构

PingCAP



PingCAP

TiDB 金融业务场景实践



PingCAP

1. OLTP - 直销金融：营销活动平台

PingCAP

业务背景

某商业银行开展直销银行的业务，依托互联网开展各种金融产品和服务的直销业务。通过其营销平台完成广告，承揽，交易和客服闭环。营销平台的重要功能之一就是支持业务方开展各种营销推广活动。如针对客户的各种促销，商户联合的推广等。营销活动平台承担营销推广中各种在线的短时高并发交易，并结合实时数据分析技术跟踪和分析目标用户群参与活动的各项业务分析指标。

问题和痛点

该营销活动平台主要有互联网/移动端 Web 接入集群(Nginx/OpenResty)，中间件集群(Tomcat)，数据库集群(MySQL 分片集群），以及消息中间件集群 (ActiveMQ) 构成。在业务活动开展后，多次发现在短时高峰交易到来时，接入和中间件层可以通过负载均衡快速扩容，而整体的性能瓶颈频繁出现在数据库端。而目前采用的 MySQL 主从复制+ MyCAT 中间件分片的架构，无法动态扩容，且中间件本身的性能和可靠性问题在几次高峰时给业务带来了严重的问题。预估在未来几个月内，随着营销业务覆盖面增加，短时高峰活动的访问并发将从当前的8万/小时，增加到12万/小时。用户方迫切寻求更好的数据库集群方案。

TiDB 的解决之道：

- 面向业务应用侧，完全采用了兼容 MySQL 的协议，应用迁移低成本，低风险，且迁移周期极短，满足了用户的周活动间隔的变更窗口要求。
- 完全透明的分布式架构，用户不用在架设中间件，也无需考虑分库分表/数据路由问题，数据库端容量和性能完全实现线性扩展。
- 一键式的在线扩容和缩容，活动高峰前 / 中 / 后阶段，均可以灵活的高度自动化的调配资源进行适配。
- 强壮的高可用设计，没有单点故障的架构保证营销活动的顺利开展。

2. OLTP - 支付业务：在线/移动支付

PingCAP

业务背景

某大型第三方支付企业，在2014年开始，就逐渐将面向互联网在线支付和移动端支付的核心数据库平台，从小型机 + DB2 平台逐渐迁移到了 MySQL 集群上以解决迅速增长的支付交易业务，逐渐摆脱了对传统集中式架构的依赖。初期采用业务侧分库分表的方案，利用 MySQL 的主从复制建立了8-10 对数据分片组。近年开始通过与外部厂商的研发合作，架构迁移到了自研的 MySQL Proxy上。集群日均TPS/QPS 在 3万/8万 的规模。

问题和痛点

切换到 MySQL Proxy 架构上后，出现了几个比较头痛的问题，主要集中在：

1) 无法很好支持分布式事务以及跨库 JOIN，目前解决方法是在应用侧进行调整适配，业务支持非常不灵活。
2) 移动端支付交易反映在使用 Proxy 架构后，交易延迟 Latency 大幅度上升，虽经过多种优化，效果不理想。
3) Proxy 自身的性能瓶颈和安全性得不到保障，目前采用多个 Proxy互备方案，Proxy 无法多节点并行服务。

TiDB 的解决之道：

- TiDB 的核心设计就是一个完整的强一致性关系数据库，支持完整的分布式事务。事务问题在在 TiDB 架构上完全不存在问题。用户可以像面对一个单机数据库那样执行和管理自己的事务操作，解放了业务开发团队。
- 由于不需要数据库中间件层，因此没有额外的延迟指标损耗，TiDB 表现出了良好的对移动端支付交易的性能支持。
- TiDB 有一个完善的分布式 SQL 层，业务通过负载均衡器，可以并行访问多个 SQL 节点，彻底解决了接入层高可用和性能横向扩展的问题。

3. OLTP - 登记结算：高性能分布式结算处理

PingCAP

业务背景

某券商企业，在保持传统经纪业务的同时，为应对激烈的市场竞争，业务上积极开拓，现在具备了包括基金，资管，OTC，投行，融资融券，质押，跨境金融等业务群。同时充分利用互联网和移动端技术优势，发展出了完整的在线/移动端开户，中间业务，在线网点业务等多种创新业务。获客和交易能力大幅提升。

问题和痛点

券商在每日盘后要开展登记结算的批量数据处理，但随着新业务的不断上线，原有集中式数据库已经无法满足性能和灵活性上的要求，问题集中反映在：

1) 集中式数据库的处理能力有限，虽已经通过硬件升级扩容到顶配，但仍旧无法满足快速增长的结算数据规模和计算作业量。结算本身和差错处理等操作往往逼近第二天开盘前的时间窗口极限，带来开市风险。
2) 随着各种新业务上线，在结算库中需要对数据库核心库表进行在线的 DDL 调整适配且不能影响业务。目前的数据库架构实现起来比较困难，导致变更窗口延长，进一步影响到了结算时间窗口的作业处理。

TiDB 的解决之道：

- TiDB 的分布式架构设计，无论是数据存储层还是 SQL 引擎层 都可以横向扩展，实现了并行的结算处理要求。结算批处理作业程序直接和多个 SQL节点交互，通过增加 SQL 节点实现作业的完全并行化。
- TiDB 是一个完整的强一致性关系数据库，支持完整的分布式事务，完全能够满足结算过程中对一致性的强要求。结算应用无需进行妥协。
- TiDB 支持在线的 DDL 操作，新业务对结算库的核心表变更可以在线完成，极大的缩小了变更窗口，确保了结算作业在规定的时间窗口内完成。

4. OLAP - 交易监控：实时交易监察与监控平台

PingCAP

业务背景

某大型金融机构，通过电子化手段，提供现货和期货类产品的交易和撮合。监管机构为保护投资者，打击金融违法行为，要求该交易机构提升交易监察和监控能力。从原来的 T+1 方式的交易记录监察，通过架构改造，提升到 T+0 的准实时监察和监控水平。交易产品和撮合记录，流式进入监察和监控平台，通过一套成体系的监察和监控业务逻辑规则及时发现交易异常和违规行为。

问题和痛点

原有的的交易监察和监控流程是通过批处理数据加载方式，T+1 方式批量从交易系统中获得交易记录数据，录入数仓后，在数仓及相关工具的基础上，进行规则过滤和演算分析。近几年，相关金融交易品类的快速增加，交易方式日趋复杂，T+1 方式所需要处理和分析的数据量越来越大，过滤和监察监控规则也越来越复杂，而各种金融违规大大的延长的交易监察的时效性，无法满足市场和监管对于交易违规和交易异常时时捕获的要求。

TiDB 的解决之道：

- 利用消息中间件，将交易系统的交易记录和撮合日志，流式写入 TiDB，利用 TiDB 的分布式存储，高性能数据写入和弹性存储。
- TiDB 分布式 SQL 引擎，提供了高性能的即席查询计算能力。包括对多类型JOIN 的支持 (Hash/Index lookup, Sort merge)，基于成本的CBO优化器框架，过滤和谓词计算下推到分布式存储引擎等一系列特性。
- TiSpark OLAP 分布式引擎，结合了 TiKV 分布式存储引擎和 Spark 分布式计算的强大能力，对于极复杂的 SQL 查询，可以通过 Spark SQL, 直接从 监察监控库中获得数据，并利用Spark 集群进行高性能计算.

5. OLAP - 风控：实时风控

PingCAP

业务背景

某大型互联网金融企业，通过先进的技术手段和服务方式，通过互联网平台提供理财，信贷和保险相关的金融产品和服务。其中信贷作为主要的业务群，提供了各种消费金融场景下的信用借贷产品。基于大数据分析和相关技术，实现各种信贷产品在平台上的智能推介，放贷和合约跟踪等。这些信贷业务的背后，依赖与一套先进的风控管理平台。

问题和痛点

原有的风控数据由业务系统产生后，导入到 Hive/Hadoop平台，但 Hive 的实时性能力完全不能满足信贷业务风控检索和计算的要求，因此被迫执行 T+1 的方式提供前端各类风控数据消费者应用查询。T+1 模式对于信贷产品的推广和销售以及用户体验上带来了负面影响。

TiDB 的解决之道：

- 风控数据通过消息中间件双写 Hive/Hadoop (历史库/历史分析) 和 TiDB (实时库/实时查询分析)
- TiDB 的分布式存储引擎架构，非常轻松的应对海量风控数据的导入，存储和查询处理。风控库内主要的核心表数据量规模在 50亿以上。
- TiDB 的分布式 SQL 引擎层，可以高性能的为前端各类消费端提供低延迟的精准查询。
- TiDB 的完整的标准 SQL 关系模型支持，方便了风控业务开发团队建模和业务侧的应用开发。

PingCAP

金融行业选择 TiDB

- 真正的分布式 SQL 数据库
- 强一致，分布式事务
- 容量和性能横向弹性扩展
- 高可用高可靠性架构，无单点

国际化的顶级开源数据库项目

中生代技术
FRESHMAN TECHNOLOGY

ArchData技术峰会全国巡回
上海9月，北京9月，成都10月，南京10月，
长沙11月，广州11月
中生代咨询内训
技术架构，研发管理，敏捷开发，大数据
微服务，AI，机器学习
中生代人才内推
对接研发主管，内推精准人才