PingCAP

划时代的 NewSQL 分布式关系型数据库

技术团队

刘奇 | CEO Co-Founder

前京东 / 豌豆荚资深架构师，知名开源分布式缓存项目 Codis 作者，国内 Go 语言社区知名技术领袖之一

崔秋 | Co-Founder

前搜狗 / 豌豆荚资深后端工程师

黄东旭 | CTO Co-Founder

前微软亚洲研究院 / 网易有道 / 豌豆荚全栈工程师，架构师，Codis 共同作者，Open Source Hacker，业界知名程序员

王鉴 (John Wang)

前 LinkedIn 搜索架构师 / Twitter 早期员工，Tech Lead，分布式搜索数据库 SenseiDB 创始人.

CEO/CTO 是分布式计算领域的旗帜性人物和开源社区领袖，同时是国际知名开源项目 Codis 作者。Codis 是基于 Redis 的分布式集群解决方案，为用户提供在线弹性伸缩和高性能的缓存服务，该项目目前已广泛被各大互联网公司（百度 / 小米 / 滴滴 / 猎豹 等）采用。

核心技术团队均为一线互联网公司基础架构部资深架构师，有技术改变世界的强烈抱负。

数据库技术发展演进

PingCAP

2008年以前

单机关系型(SQL)

- 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

- 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

2008年至2013年

分布式非关系型(NoSQL)

- 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

- 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

2013年以后

分布式关系型(NewSQL)

- 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

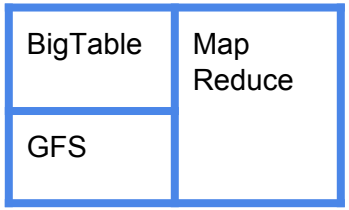
- 挑战:基于 Google Spanner/F1 论文，基础软件最前沿的领域之一，技术门槛最高

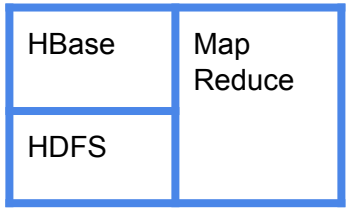
资料来源:公开资料整理

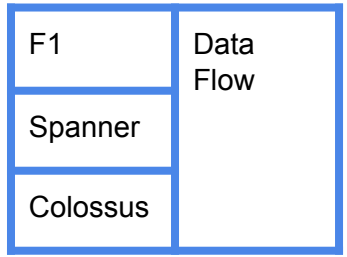
Google - 大规模分布式计算领域的领跑者

PingCAP

2003 GFS
2004 MapReduce
2006 BigTable
2012 Spanner
2013 F1







Google

hadoop

Google

NewSQL | TiDB

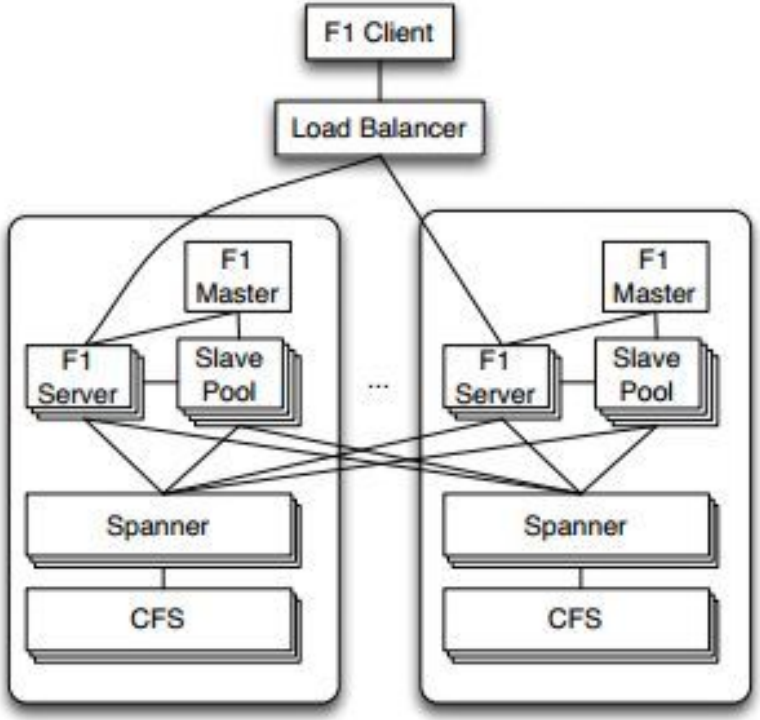
Google 十年前基于内部分布式 处理框架发表的三篇论文奠定了大数据分析 处理基石；开源社区以此为基础打造了Hadoop

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 - The First NewSQL

- 全球级别分布式 / 跨数据中心复制
  - Paxos
- ACID 事务支持
  - 两阶段提交
- 无锁快照读 / 无锁只读事务
  - MVCC
- External Consistency
  - TrueTime API
- SQL 支持
- the Next BigTable Powered By Jeff Dean



我们在做什么

PingCAP

新一代NewSQL分布式关系型数据库 Ti Project（TiDB + TiKV）

- 基于 2013 年 Google Spanner / F1 论文
- 基于 2014 年 Stanford 工业级分布式一致性协议实现 Raft 论文

概括：

无限水平线性扩展、高并发高吞吐、跨数据中心多活、MySQL 兼容的真正意义上的分布式数据库

- 我们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）
- **完全从头打造，并非基于 MySQL 或数据库中间件进行改造、封装**
- 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
- 开源模式保证技术创新、高效和领先性，天然的国际化基因

我们的数据库能解决什么问题 - 1

PingCAP

- **无限线性水平扩展(Scale Out)**

  无论多大的数据量，都可以轻松通过增加节点来解决，写入和读取时延固定(毫秒级别)，**无需分库分表或者搭建复杂的 Hadoop 集群**，完整的 MySQL 兼容接口轻松处理**高并发实时写入、实时查询**和分析，极大的简化程序设计、应用维护，轻松应对大数据存储问题。

- **高并发、高吞吐、完整的跨行事务支持、强一致性**

  通过简单的增加节点，提供无上限的、线性扩展的的高并发、高吞吐的处理能力，卓越的集群处理能力。同时提供跨行事务处理能力。

我们的数据库能解决什么问题 - 2

PingCAP

- **高可用、跨数据中心多活**

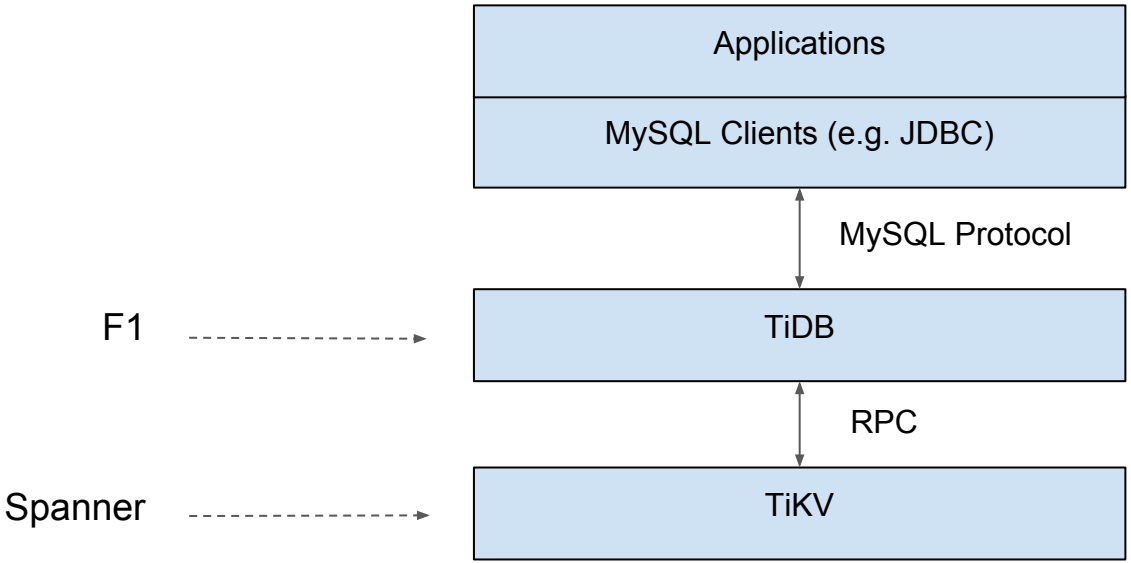
  分布式算法 Raft 自动完成多副本写入、数据分片（region）的拆分、聚合、重分布，从而保证数据高可用，天然支持跨数据中心多活且可配置。

- **底层数据打通，集中管控**

  通过增加节点即可获得无限数据存储能力，从架构层面轻松支持多个业务系统底层数据打通，便于集中管控，避免信息孤岛，提升数据价值。

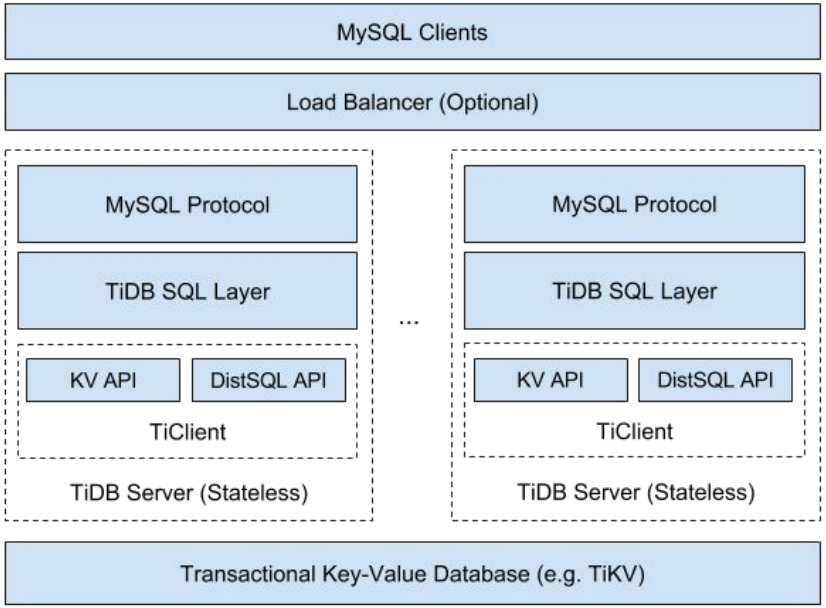
PingCAP

Architecture overview (Software)



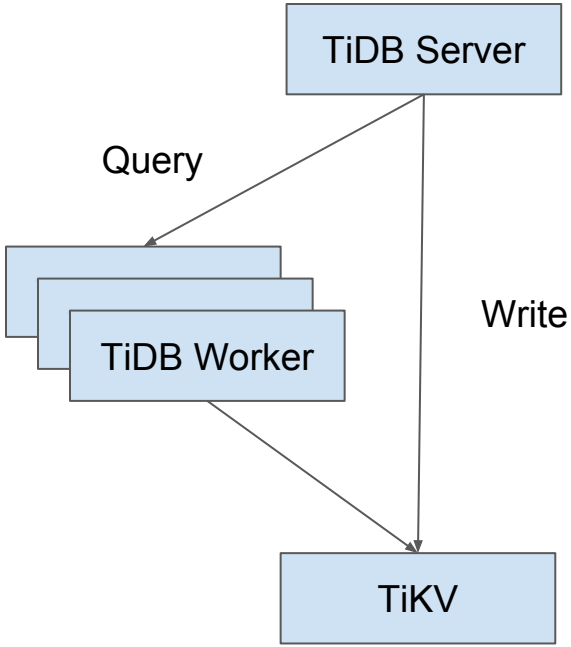
TiDB

- 开源 F1 实现
- 无状态的分布式 SQL 层
- MySQL 协议兼容
- 针对分布式场景的执行计划，优化器
  - Push-down / MPP
- Online DDL
  - 业务不中断，进行表结构变更



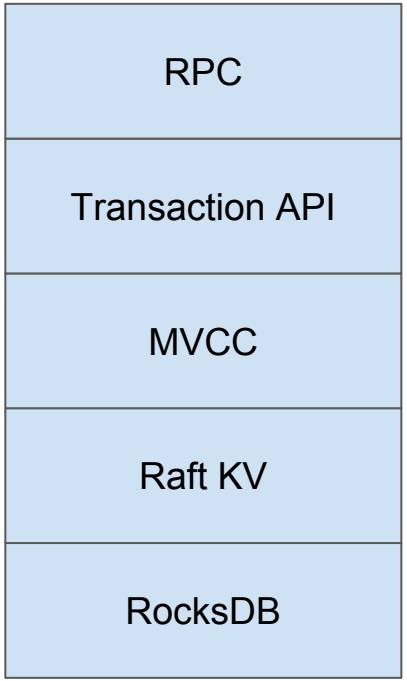
TiDB 逻辑架构

- 支持关系型的表结构
- 全局一致索引
- 通过 Map-Reduce 或 SQL 并发无锁读

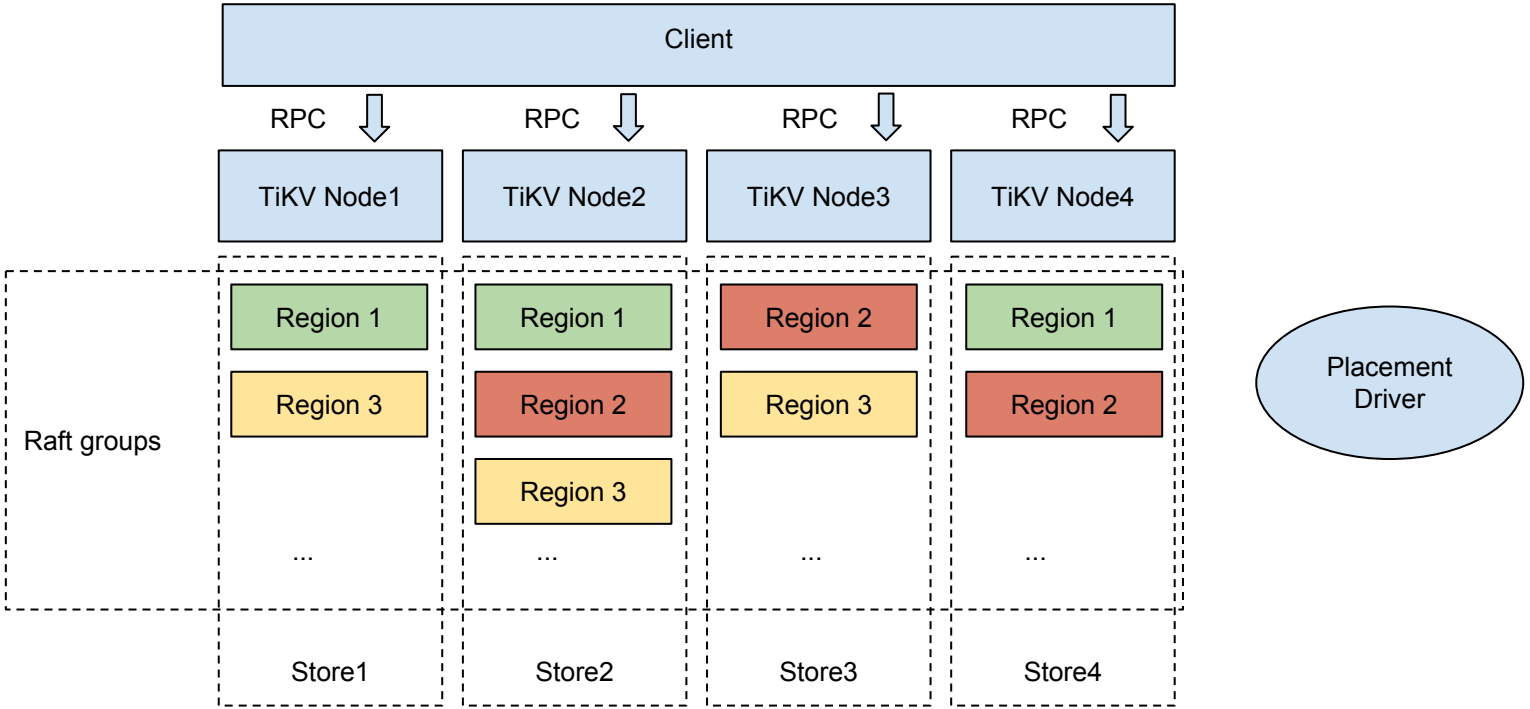


TiKV

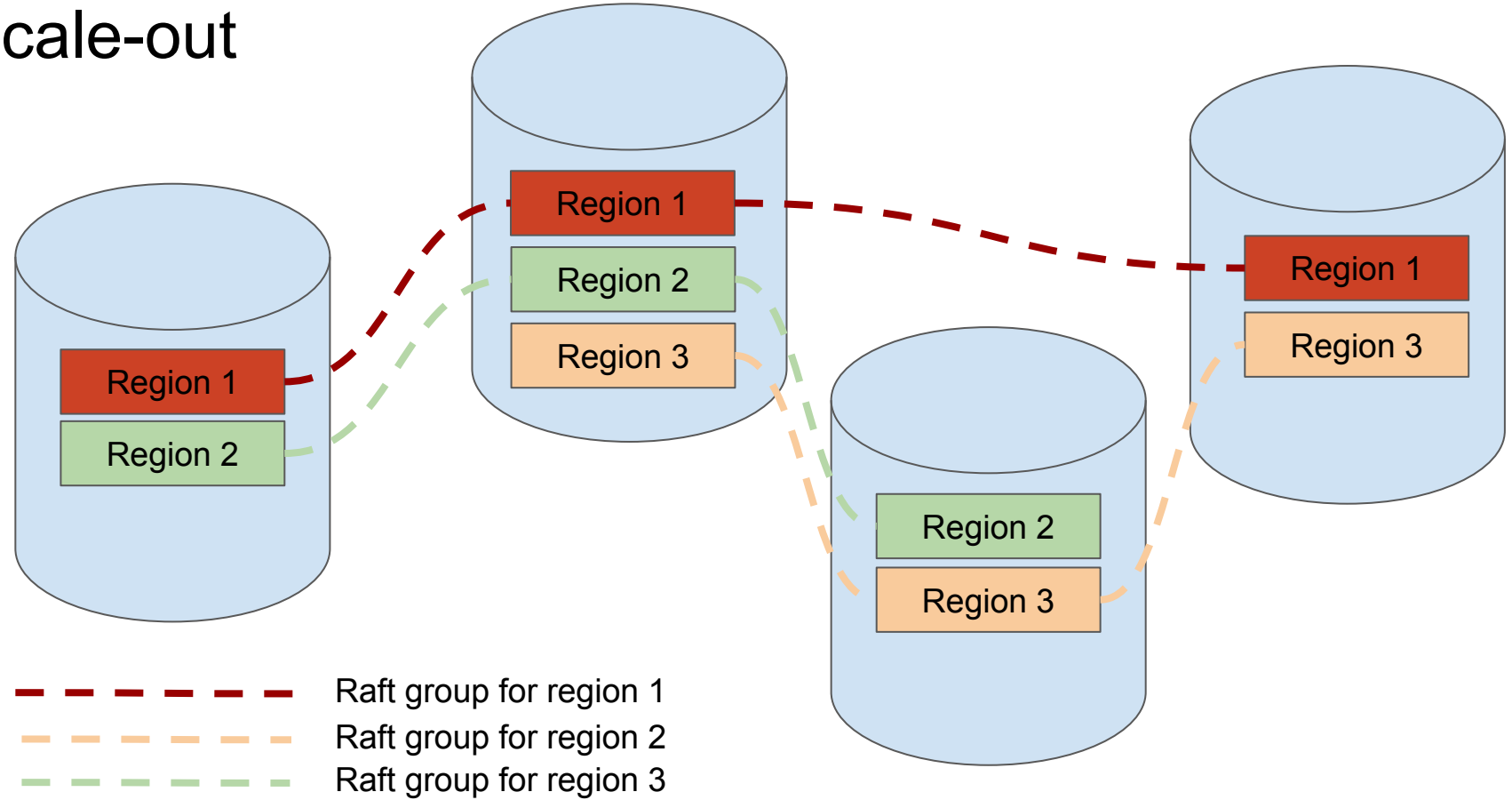
- 开源的 Spanner 实现
- 基于 Raft 多副本一致性算法
  - 使用Multi Raft支持Dynamic Scale
- 支持类 Percolator 分布式事务
- 提供 MVCC 支持
  - RocksDB



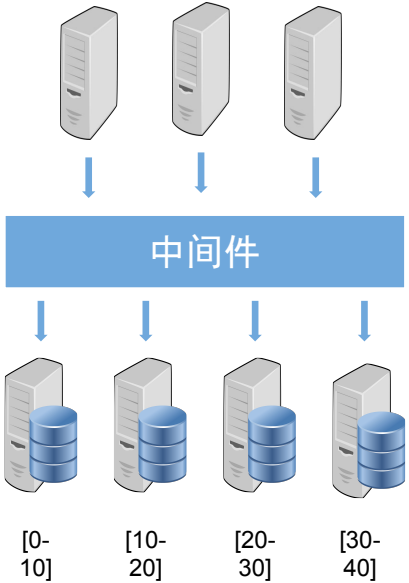
TiKV Overview



Scale-out

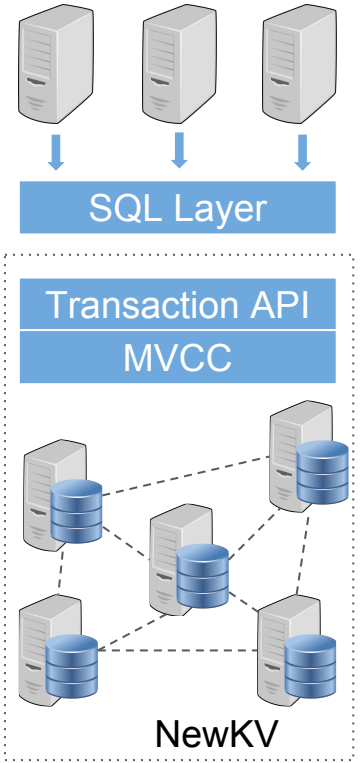

Region 1
Region 2
Region 1
Region 2
Region 3
Region 2
Region 3
Region 1
Region 3
Raft group for region 1
Raft group for region 2
Raft group for region 3

NewSQL - 数据库无限水平扩展的完美解决方案



DB Sharding

大数据时代，当单机数据库容量及处理能力达到瓶颈时，由于没有完美的分布式解决方案，业界普遍采用妥协的数据库分库分表（Sharding）方案



NewSQL | Ti Project

DB Sharding vs NewSQL

PingCAP

	DB Sharding		NewSQL | TiDB	
	工作内容	工作量	工作内容	工作量
项目设计阶段	分库分表设计 数据库中间件路由配置 数据库主从备份设计	月级别	类似单机数据库，不需要做任何设计、修改	0
项目开发阶段	事务处理机制 - 由应用保证 跨库、跨表查询 - 由应用保证	月级别	类似单机数据库，事务由底层数据库提供，支持透明分布式事务	0
扩容、缩容、维护	制作分库数据镜像 暂停业务 检验所有数据和原有数据的一致性 切换路由配置 上线后测试验证 开始业务 （中间有问题都需要进行相应的重试或者回滚方案）	周级别	人工方式：简单命令行，按需增加服务器、删除服务器 自动方式：通过API接口，自动增加或删减服务器	人工方式 1分钟 自动方式 0

NewSQL - 数据中心容灾、多活的完美解决方案

PingCAP

方案	方案特点
传统硬件数据复制方案	1、造价昂贵（存储、带宽），两中心硬件配置需完全一致（存储、主机） 2、冷备，资源闲置，主生产中心故障时，需手动切换业务 3、主生产中心出现写错误，可能导致数据同步出错，备中心无法启动
Oracle Active Data Guard	1、额外购买 ADG License 2、备机只能读，不能写入，不是真正意义上的双活 3、主生产中心故障时，会有秒级的数据无法同步，数据一致性无法 100% 保障（RPO > 0） 4、主生产中心故障时，需手动切换业务
MySQL	没有原生的安全的同步方案
NewSQL | TiDB	简单配置即可保障所有的数据中心 100% 同步，真正意义上的多活 完全实现 国标（信息安全技术-信息系统灾难恢复规范GBT 20988-2007 ）最高等级第六级的要求（RTO = 0，RPO = 0）

NewSQL - 大数据量下高压力实时处理的完美解决方案

PingCAP

随着业务增长，需要处理的数据量不断增加，系统架构师面临多种挑战：

- 传统的关系型数据库无法满足高并发写入？
- 传统的关系型数据库查询缓慢？
- 想引入 NoSQL 获取水平扩展能力，但是又不愿牺牲 SQL 和事务处理？
- 分库分表又进入另外一个坑

NewSQL | TiDB 以便捷的接口（MySQL 兼容），让你像使用单机数据库一样轻松获得大数据处理能力，满足大数据量下的高压力的实时处理的各种需求。

商业模式：开源

PingCAP

- 开源的模式在硅谷已经验证成功，作为**基础软件领域正确的商业模式**
  - Cloudera / Hortonworks / OpenStack / Docker / RedHat ...

代表	Datastax	CoreOS	Docker	Mesosphere	MongoDB	Cloudera
估值($)	10亿	10亿	20亿	10亿	20亿	70亿

- 开源正在蚕食数据库市场
  - MongoDB / Cassandra / MySQL / PostgreSQL / Spark …
- 开源 != 免费
  - 社区版
  - 企业版（监控管理插件、优化插件、数据安全插件、企 业服务及培训）

TiDB 短时间内已获巨头高度认可

PingCAP

在市场费用为0的情况下，TiDB 已经成为明星项目，获得广泛的认可和关注

合作	Github	3200+ Commits / 4000+ stars / 44 contributors / 2 Partners
	华为	全职工程师（6人团队）深度参与，将会在手机云平台（类似 Apple iCloud 业务）和华为商城（VMall）等大型项目落地
	京东	全职工程师（10人团队）深度参与，将会在京东云平台落地，替换内部 Oracle，继阿里去IOE之后又一重大标志性事件
	百度	已经开始整合 TiDB 的技术调研
	滴滴、360、乐视、去哪儿	均有浓厚兴趣，后续合作正在洽谈
渠道	多个Pass服务商	主动希望为其客户提供分布式数据库解决方案（云服务）
客户	多个客户	有明确和迫切的需求痛点，主动提供场景进行 PoC

PingCAP

附：

SDCC 对我司 CEO 刘奇的采访
http://dwz.cn/3qW56Y

DTCC（中国数据库技术大会）对我司 CTO 黄东旭的采访
http://dwz.cn/3gcZLC

项目地址
TiDB: http://github.com/pingcap/tidb
TiKV: https://github.com/pingcap/tikv