

2018PostgreSQL中国技术大会



# 开源分布式NewSQL数据库 CockroachDB架构及最佳实践

赖宝华

laibaohua@baidu.com

百度云 & CockroachDB中国社区



Cockroach DB



Baidu 百度



百度云

# CockroachDB

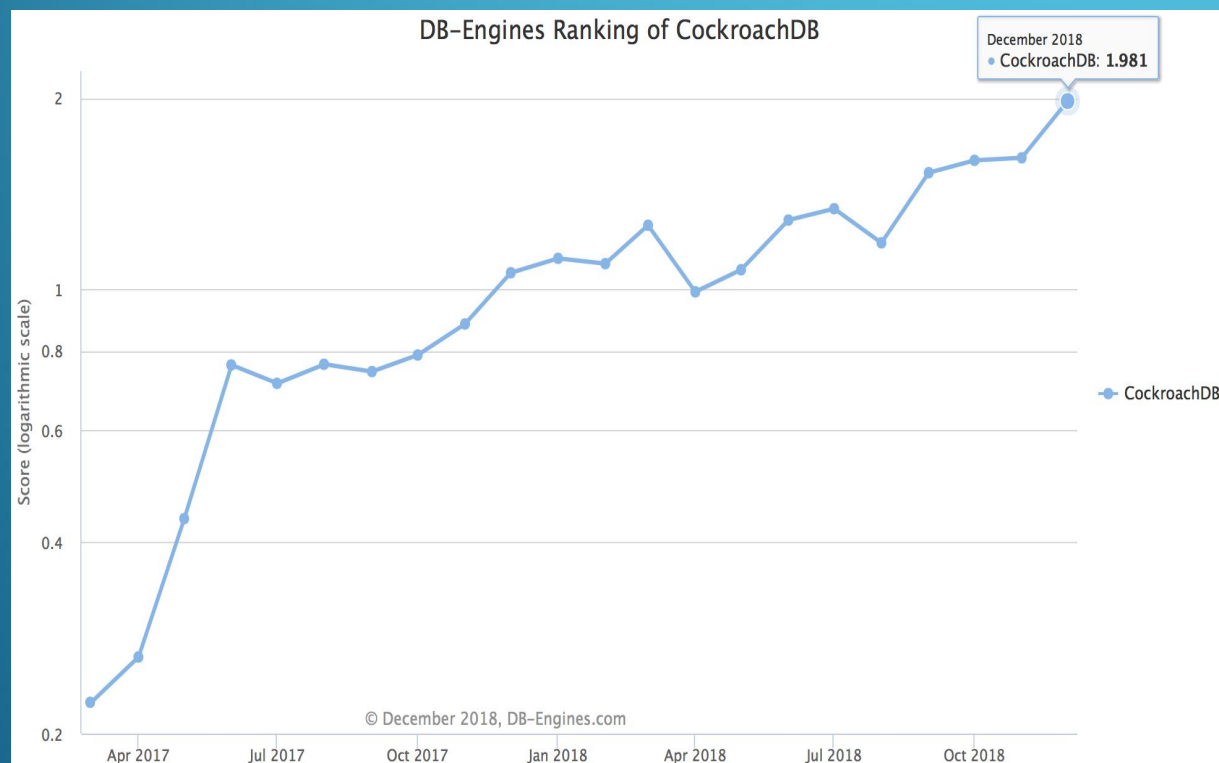
# CockroachDB



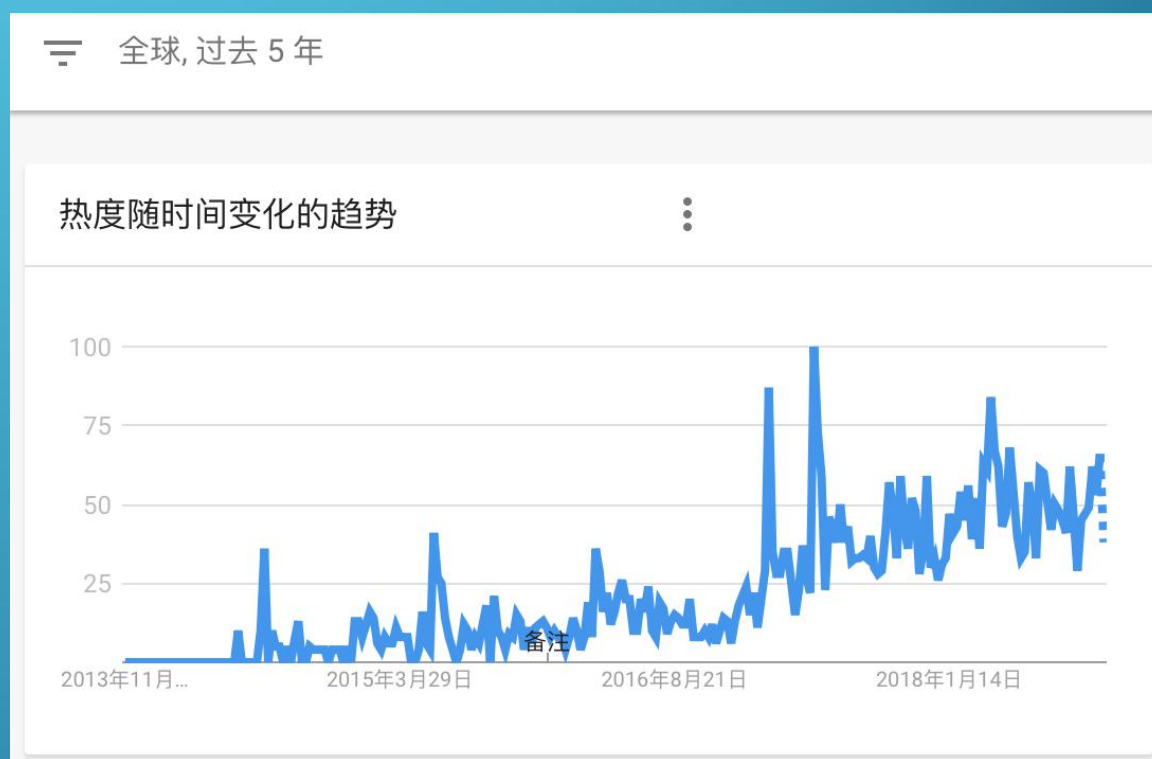
- ✓ Github开源项目，遵循Apache License，基于Golang实现：
  - 2014年开源，2017年5月发布1.0，目前最新2.1版本；
  - Spanner开源实现，目前GitHub Star数量14700+，Contributor数量220+；
  - 采用PostgreSQL协议，InforWorld 2018最佳开源数据平台；
- ✓ 母公司Cockroach Labs：
  - 三位创始人全部来自Google，有 BigTable, GFS, Colossus, Gmail项目背景；
  - 已获得来自Benchmark, Google Venture 等共计近\$5325万的融资；
  - Cockroach Labs Base在纽约，有50+研发人员；
- ✓ 行业应用：
  - Comcast、Bose、Express、ExonMobil、Metro、Vistaprint、Mesosphere、MINDBODY；

# CockroachDB

## DB-Engines



## Google Trends



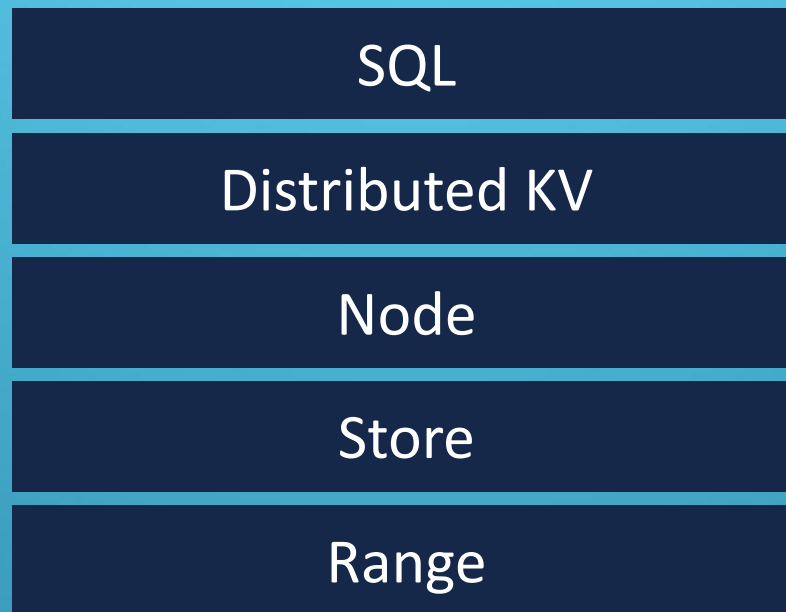
# NewSQL起源

- ✓ 分析师Matthew Aslett在2011年发表的报告中首次提出了NewSQL概念;
- ✓ 维基百科对NewSQL的定义: 具有NoSQL对海量数据的存储管理能力, 还保持了传统数据库支持ACID和SQL等特性;



参考: <https://cs.brown.edu/courses/cs227/archives/2012/papers/newsq/aslett-newsq.pdf>  
<https://en.wikipedia.org/wiki/NewSQL>

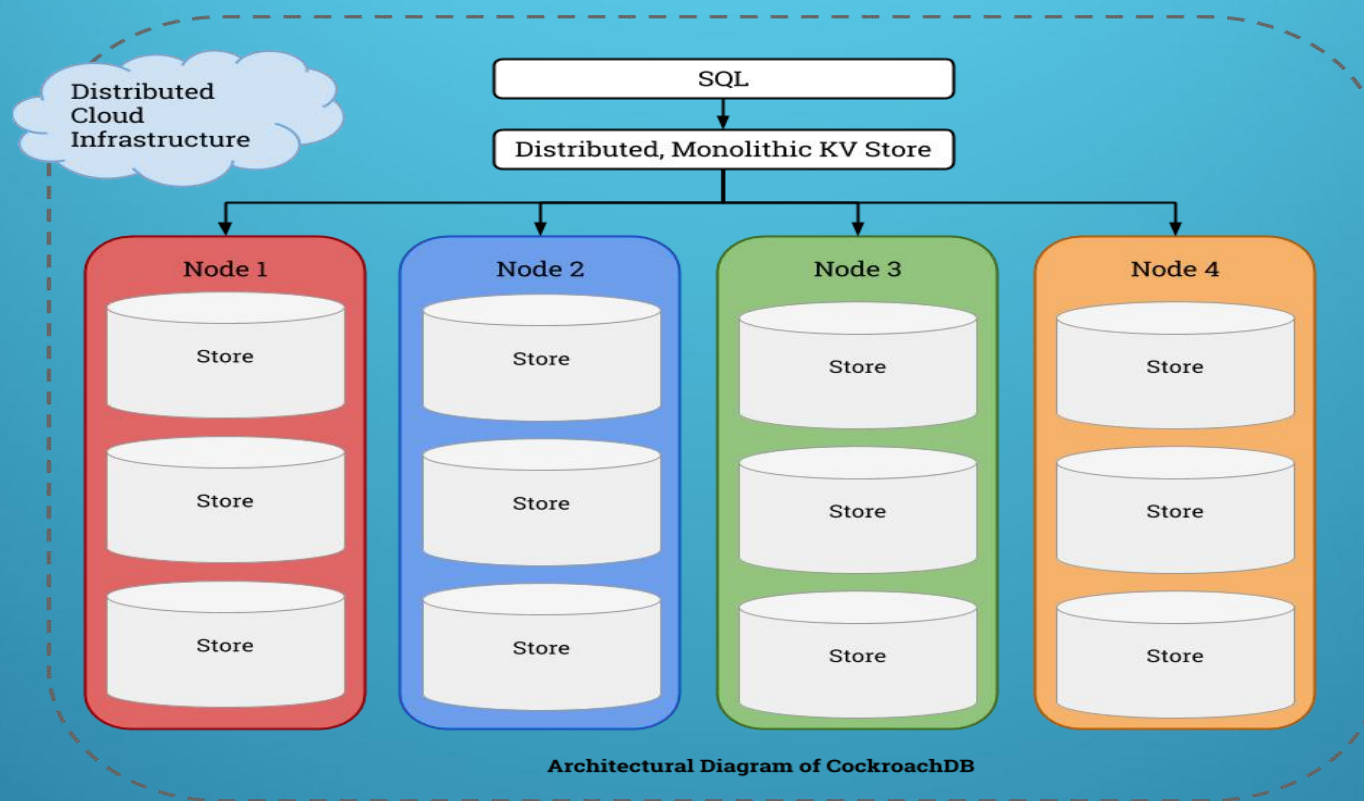
# CockroachDB架构



- ✓ SQL on Distributed KV, 是NewSQL数据库的典型架构
- ✓ F1/Spanner、FoundationDB都采用类似架构

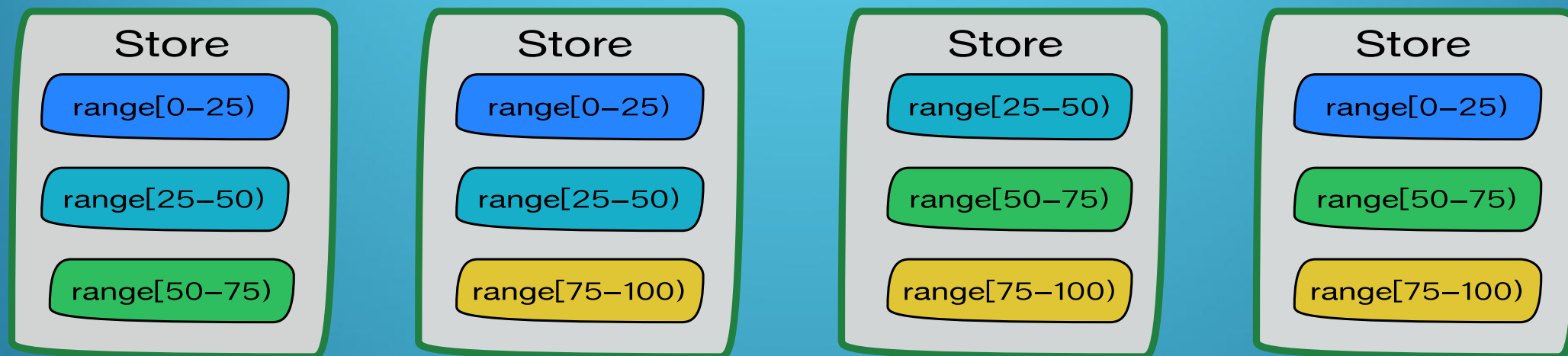


# CockroachDB架构（二） - Node & Store



- ✓ Node是CockroachDB的进程实例，一台物理服务器启动一个Node即可；
- ✓ 一个物理存储介质（例如一块硬盘）一般配置一个Store，一个Node中有多个Store；

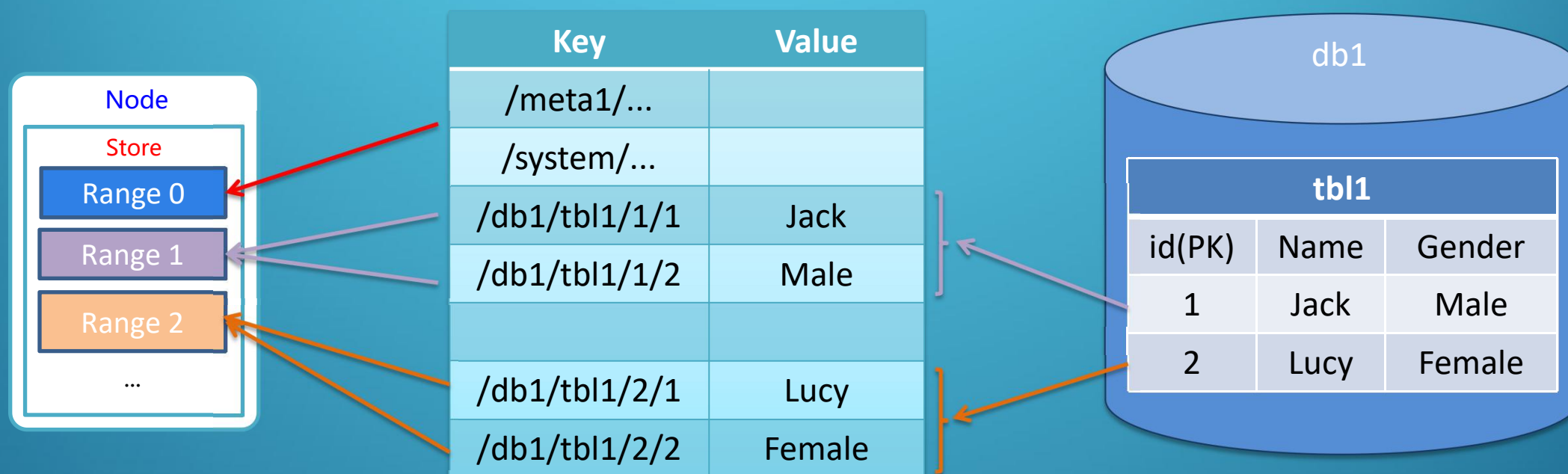
# CockroachDB架构（三） - Range



- ✓ 一个Range是一段键值区间  $[K1, K2)$  的数据分片，是CockroachDB存储管理的最小单位；
- ✓ 一个Store中有多个Range；
- ✓ 每个Range分片默认为64MB，默认存在3个副本，分布在不同的Node上；



# CockroachDB架构（四） - 物理逻辑映射



有序KV Map，支持Column Family

# CockroachDB内核实现

SQL/KV模型映射  
分布式事务

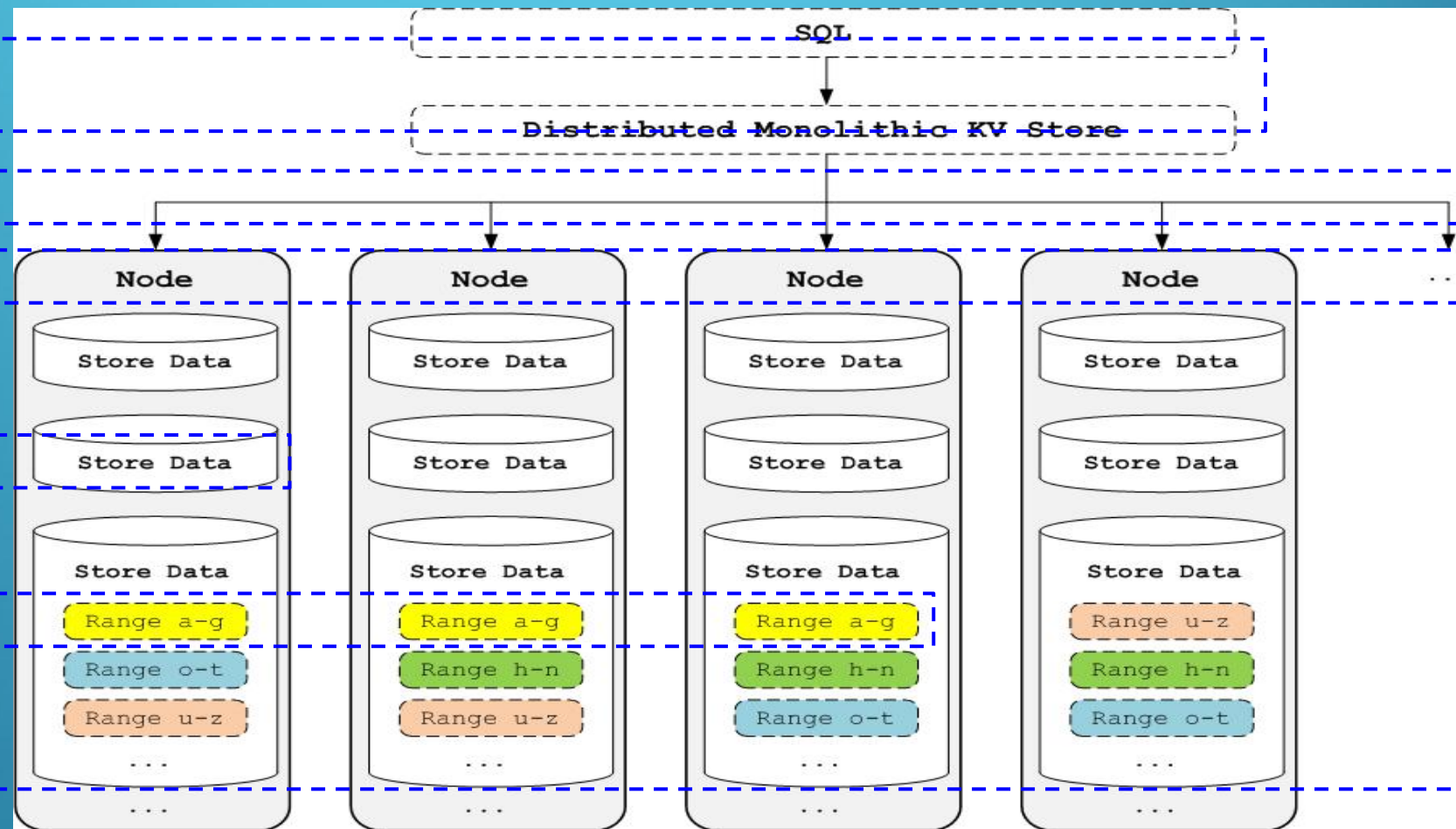
分片数据(Range)路由

集群节点状态管理Gossip

单节点存储引擎RocksDB

多副本数据一致性Raft

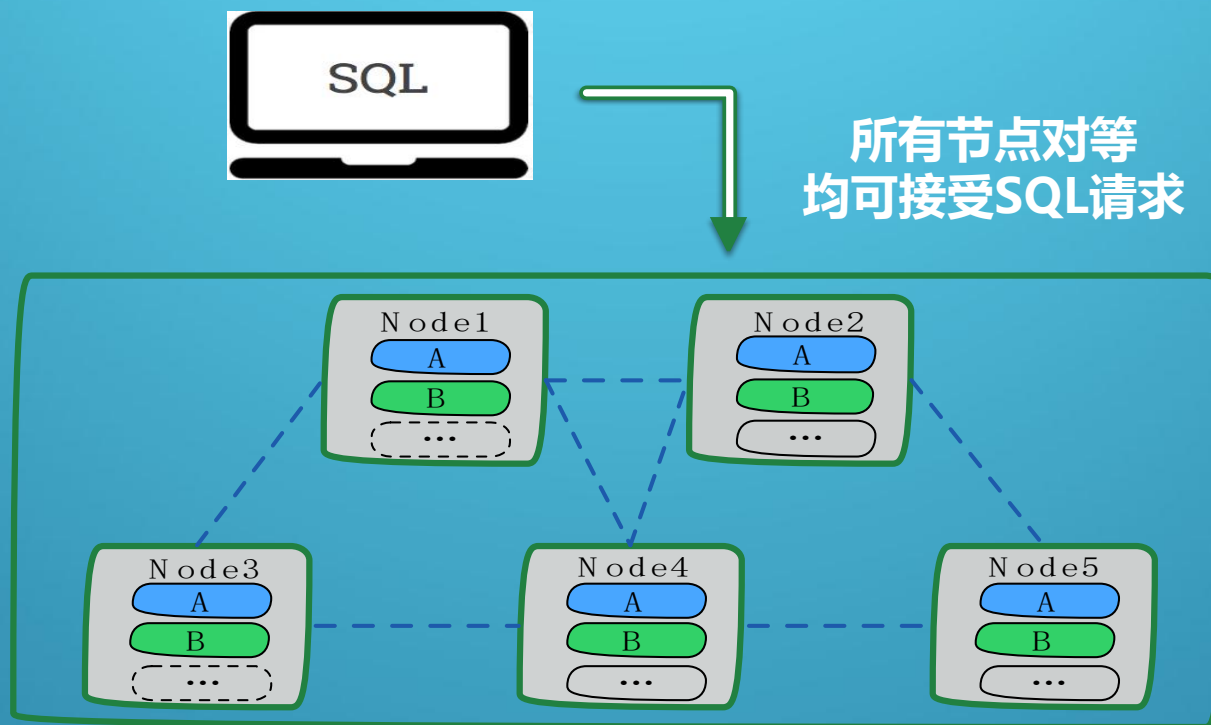
HLC时钟算法



# CockroachDB特性

- 标准SQL接口
- 扩展能力强、高并发
- 弹性扩容
- 多副本强一致
- 服务高可用
- 分布式事务
- 多区域部署

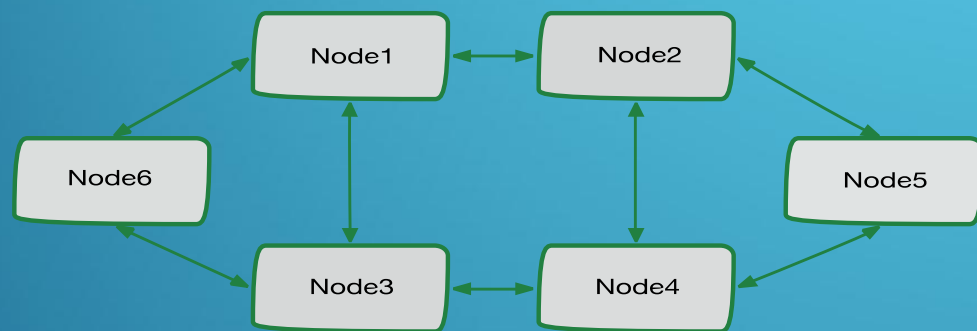
# 标准SQL接口



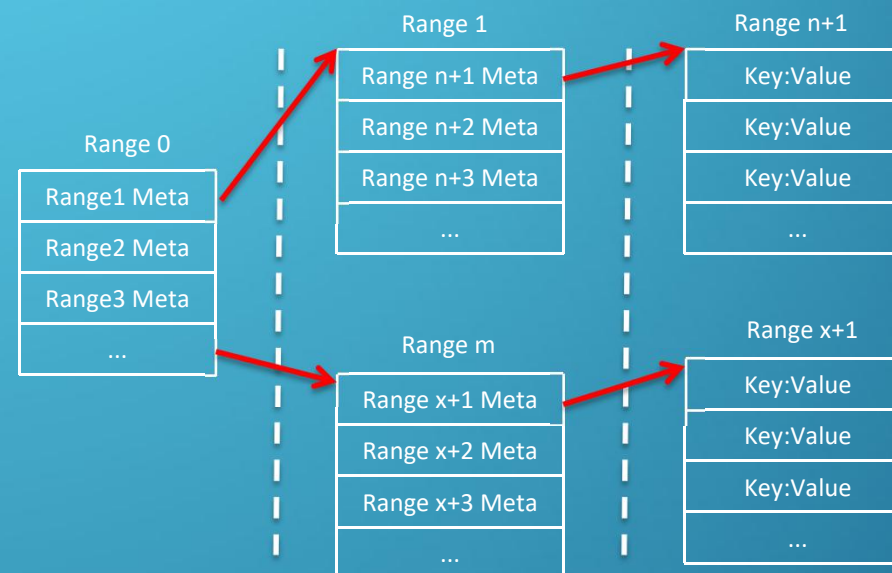
- ✓ 使用PostgreSQL协议，支持 标准SQL接口，兼容关系型数据库SQL生态；
- ✓ 支持 事务、二级索引、Join 等NoSQL欠缺的特性；
- ✓ 支持 类MPP并行查询框架；

# 扩展能力强，高并发

## 去中心化架构



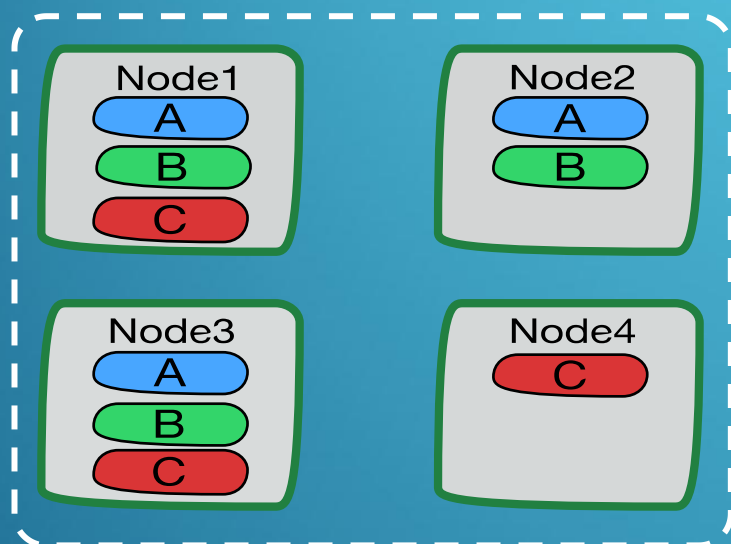
## 两级路由元数据



- ✓ Gossip协议实现节点状态管理，理论上 单集群支持10K节点规模；
- ✓ 两级路由元数据，单集群支撑最大 4EB用户数据存储；
- ✓ 架构中子模块都采用分布式设计，无单点瓶颈，支持多节点并发写入；

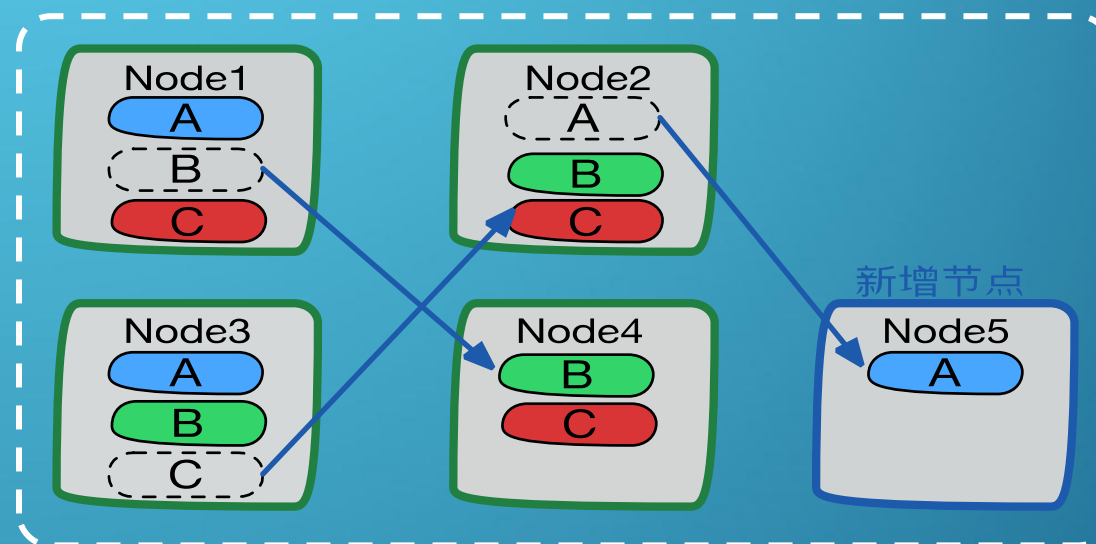
# 弹性扩容

扩容前



加入节点

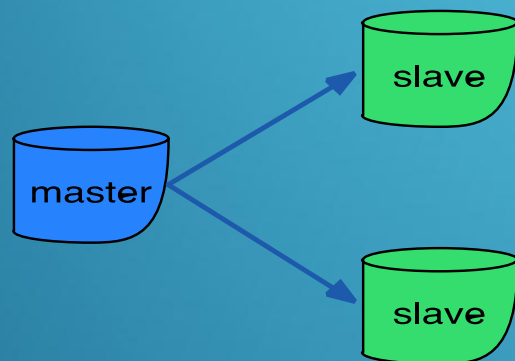
扩容后



- ✓ 支持按需扩容：数据管理最小分片粒度为64MB，可按需线性扩展；
- ✓ 支持在线扩容：新节点加入 自动负载均衡（支持存储和状态均衡模式）；

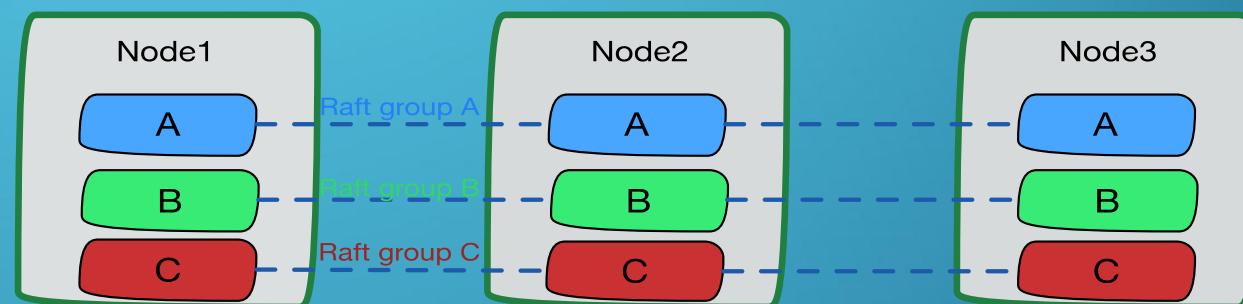
# 多副本强一致

## MySQL数据同步



VS

## CRDB数据同步

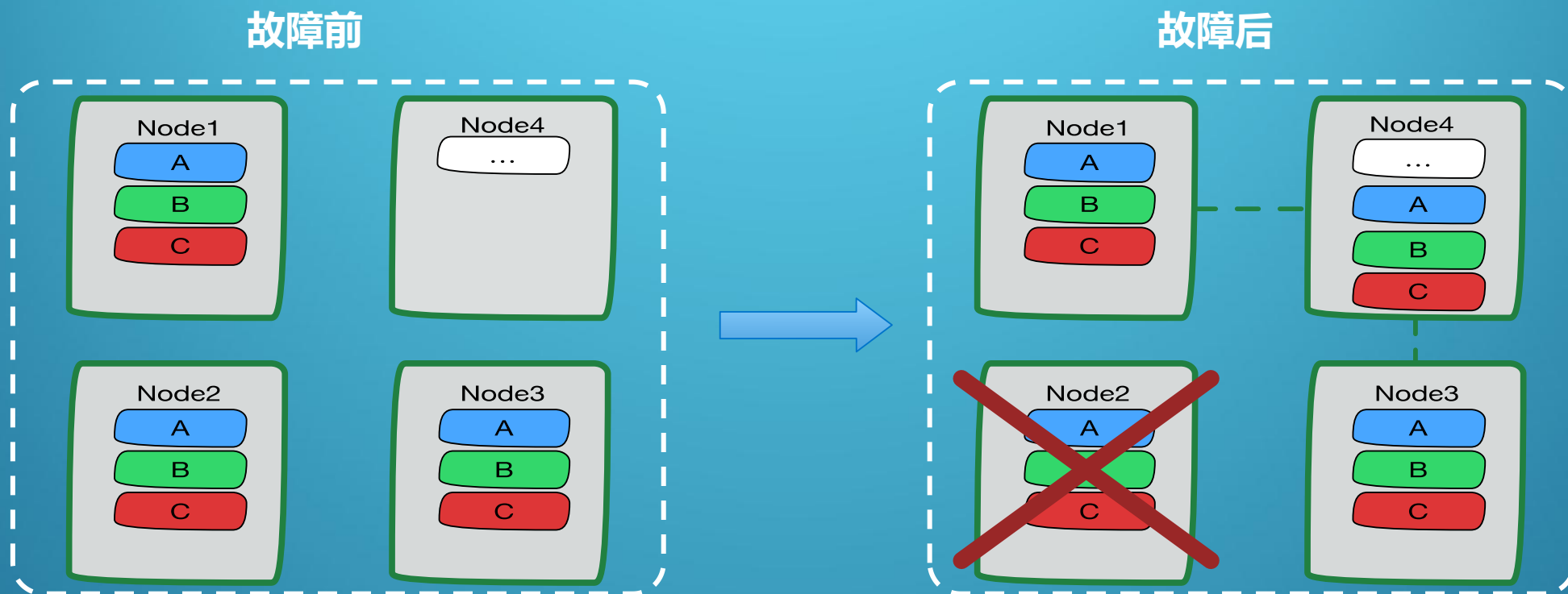


主从之间异步复制，弱一致性

副本数据同步基于Raft协议，强一致性，  
节点故障无丢失数据风险

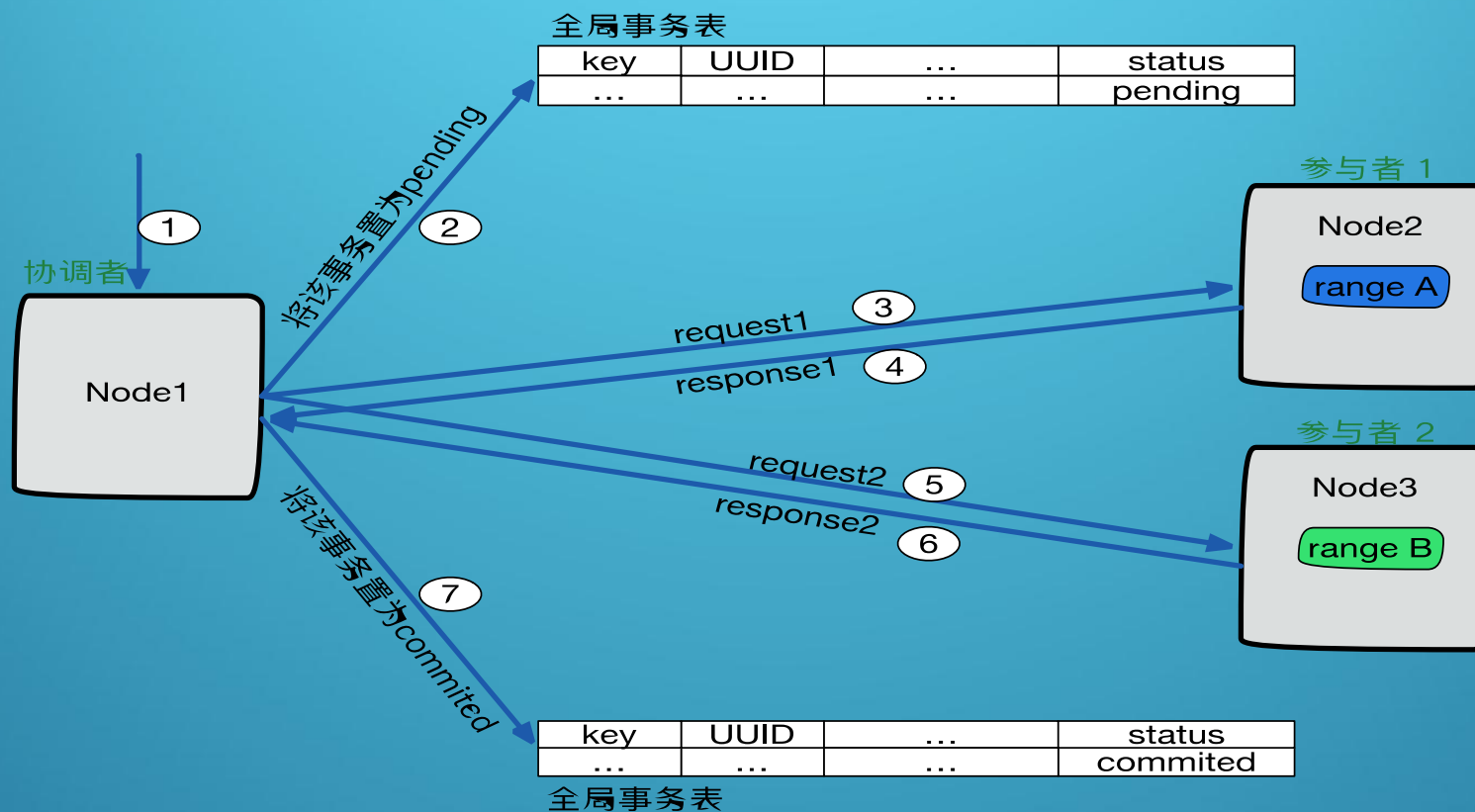


# 高可用



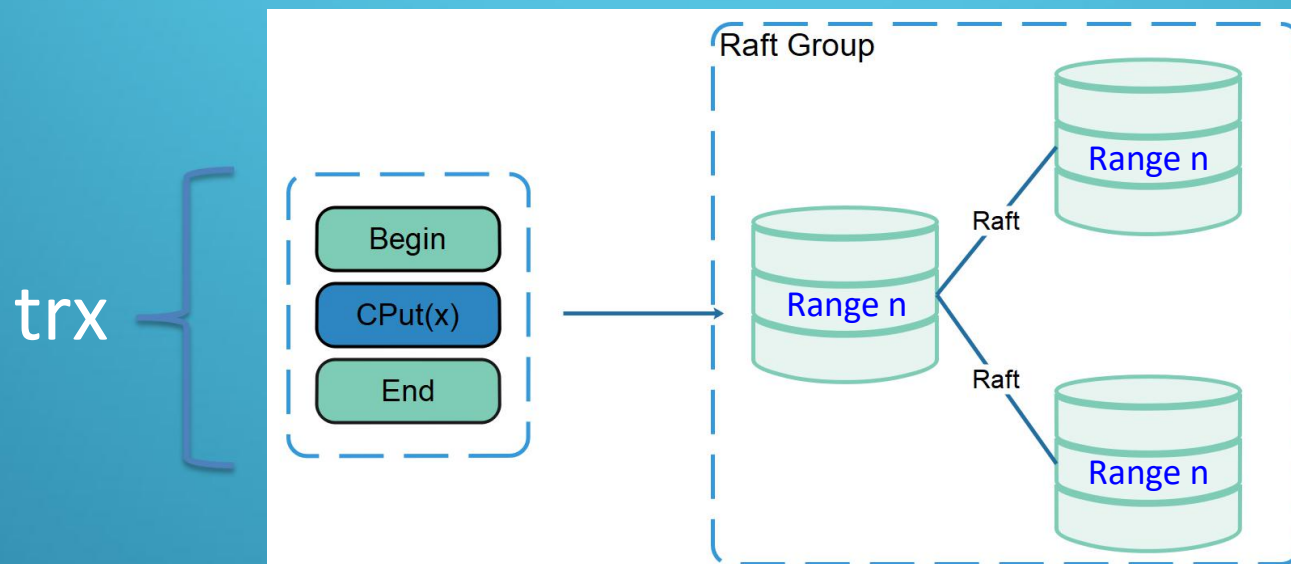
- ✓ **架构上去中心化，无SPOF**：架构不存在类似HBaseHMaster和Percolator oracle等集中式模块，单节点故障不影响集群整体的可用性；
- ✓ **故障自愈**：基于Raft协议，只要半数以上副本存活，则服务可用；当节点异常，数据副本数量少于指定阈值时，自动补齐副本；

# 分布式事务



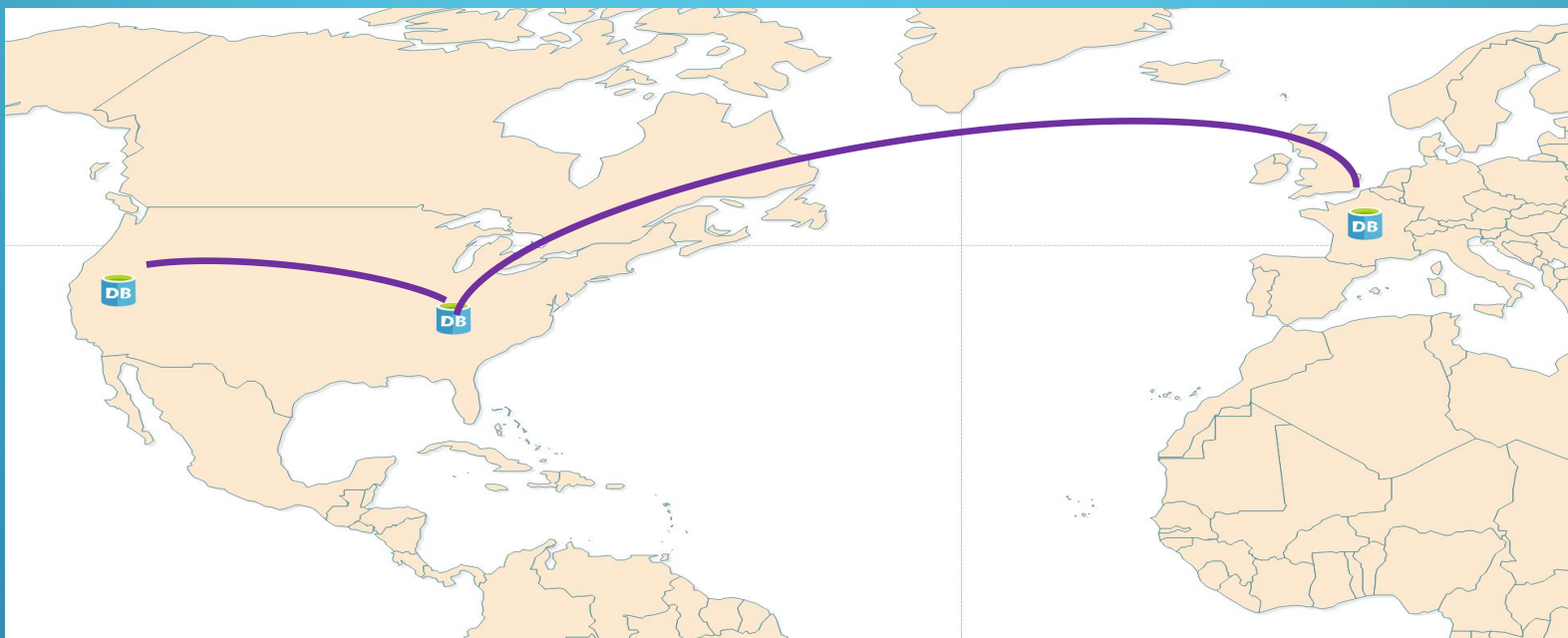
- ✓ 支持完整的分布式事务和ACID语义，对外呈现统一完整的的数据库视图；
- ✓ 基于MVCC实现事务控制，支持SI和SSI两种隔离级别；
- ✓ 基于HLC分布式时钟同步算法，任意节点皆可充当事务管理节点，无中心事务管理器；

# 分布式事务—1PC



- ✓ 优化非跨Range写事务性能
- ✓ 利用Raft保证原子性，一次完成数据写入
- ✓ 减少RPC通信

# 多区域部署



- ✓ 支持按业务需求对用户数据水平分区
- ✓ 各个分区可按地理位置设置分布策略
- ✓ 数据就近访问
- ✓ 冷热数据分离

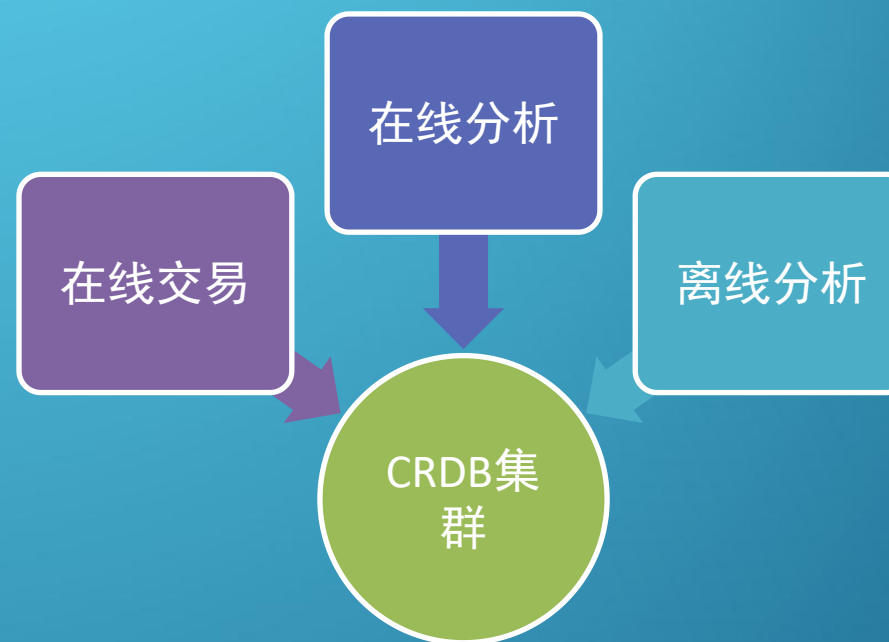
# CockroachDB适用场景

## ✓ HTAP场景

- 高并发读写，支持多点写入，自动负载均衡
- 大数据量存储
- 随时按需扩展、在线扩容
- 跨数据中心容灾，多副本数据强一致
- SQL 接口，事务能力

## ✓ NoSQL场景：

- 大部分NoSQL类型业务
- 日志、账单类数据（特征：海量数据、持续增量、高速入库，多维低时延查询）



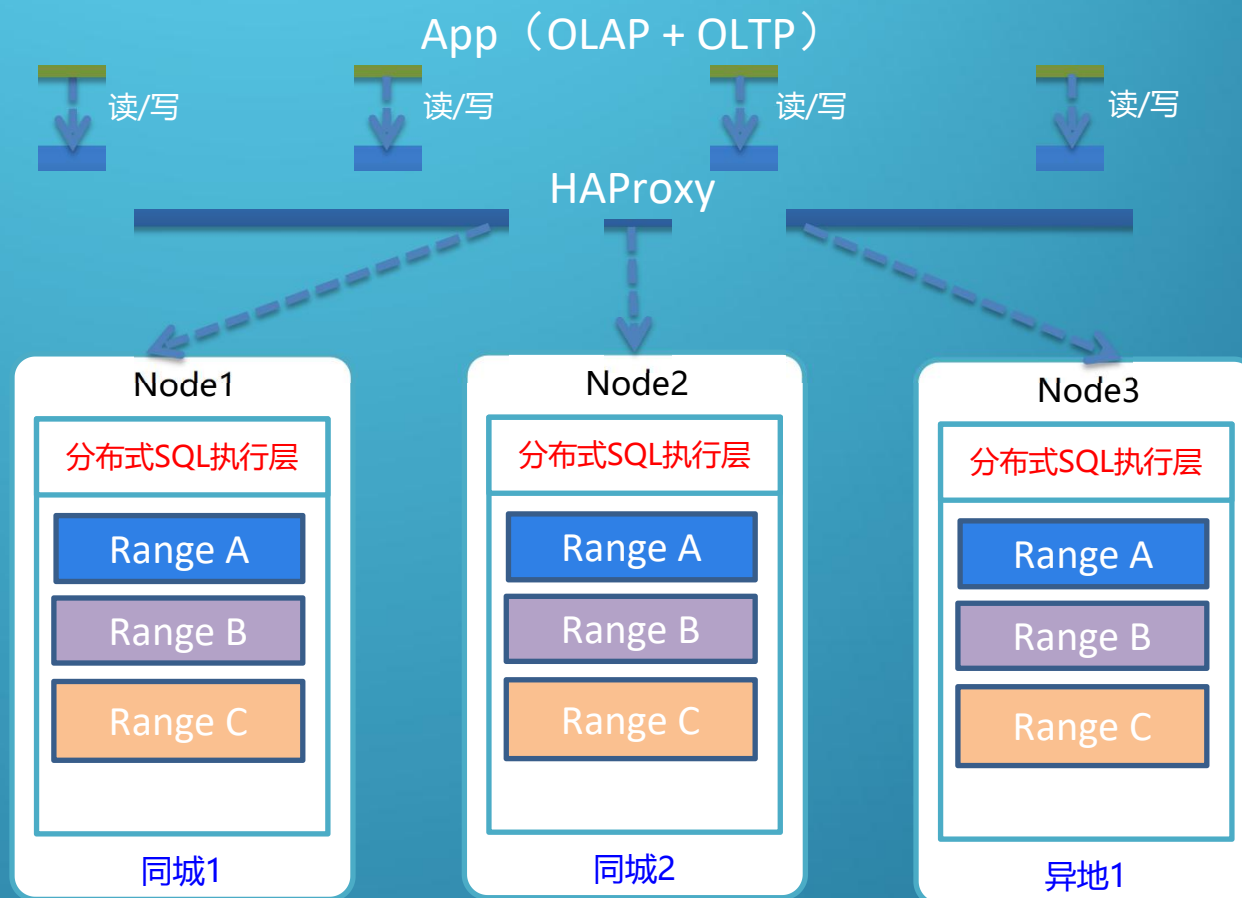
# CockroachDB最佳实践

## ✓ 硬件

- SSD / NVME
- 万兆网卡 / RDMA
- NTP / GPS

## ✓ 部署:

- 两地三中心三副本
- 两地三中心五副本
- 按地域分区 (GDPR)





# CockroachDB && Baidu



# CockroachDB & Baidu

- ✓ 百度智能云DBA团队在CockroachDB Github及社区贡献
  - 2015年初开始, 目前 2 PMC Member, 1 Contributor;
  - 30+个Commit, 15000+代码;
  - 受邀组建CockroachDB中国社区、权威发布中文社区网站;
- ✓ 外部技术分享与交流
  - 与Cockroach Labs保持良好互动, 多次互访交流, 双方达成战略合作;
  - 受邀在DTCC2018、平安SMART大会、Oracle嘉年华等大会分享;
- ✓ 业务实践
  - 业务覆盖内部多个产品线, 规模业界领先;



# 百度NewSQL解决方案

高可扩展  
EB级数据存储能力

高可用  
5个9, 秒级恢复

高性能  
百万TPS

高可靠  
多副本强一致

去中心化

分布式事务

MySQL & PG兼容

全局二级索引

滚动升级

Online Schema Change

Row-Level  
Partition

Distributed Query

Change data  
Capture

异地多活

多副本强一致

Column Family

计算虚拟化

存储虚拟化

网络虚拟化

CPU

NVMe

GPS原子钟

RDMA

FPGA

监控管理

上线管理

灾备管理

容量管理

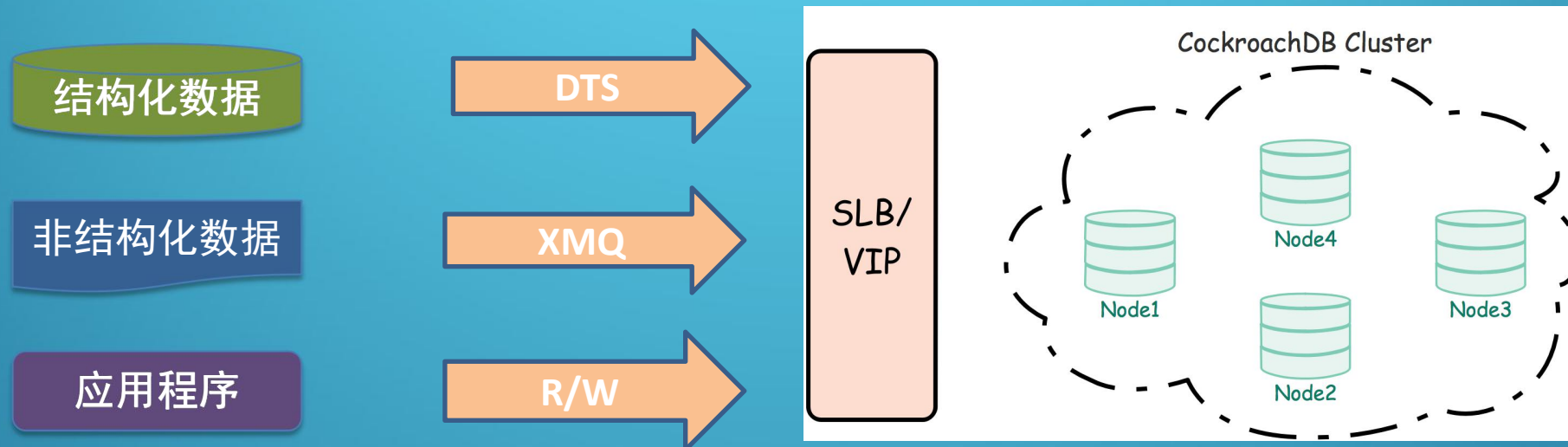
权限管理

安全管理

数据迁移

1. 兼容MySQL和PostgreSQL协议语法
2. 基于CockroachDB的HTAP方案
3. 软硬件联合优化
4. 完善的运维管理平台

# 百度NewSQL应用案例



- ✓ 对业务提供统一的数据库视图，SQL接口
- ✓ 使用更简单，平滑扩缩容，高并发
- ✓ 更易于运维，故障自动恢复
- ✓ HTAP融合

快速入库 海量数据

核心  
诉求

扩缩容 SQL接口

# CockroachDB 中国社区



# CockroachDB中国社区发展

**CockroachDB中国社区成立**

2017.11 @ 深圳 Meetup

**CockroachDB中文官网上线**

2018.4

[www.cockroachchina.cn](http://www.cockroachchina.cn)

**CockroachDB 2018 CAB**

2018.9 @ NewYork

2017.08

2017.1  
1

2017.12

2018.03

2018.07

2018.09

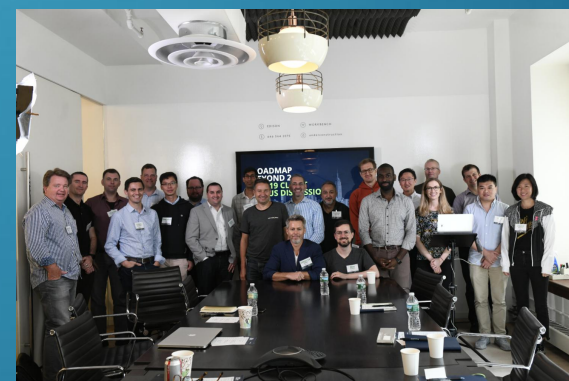
2019.1

**CockroachDB公众号上线**  
2017.8

**CockroachLabs再访中国**  
2017.12 @ 北京, 深圳

**CockroachDB 2.0 Meetup**  
2018.7 @ 深圳南山区软件产业基地

**CockroachDB 社区大会**  
2019.1 @ 北京



# CockroachDB中国社区发展

## ■ 官方社区

- 2015年初开始，目前 2 PMC Member, 1 Contributor;
- 提交30+个Commit, 15000+代码;
- 社区战略合作伙伴，受邀组建中国社区、权威发布中文社区网站

## ■ 中文社区官网

- 发布技术文章: 30+篇;
- 翻译文章: 200+篇;
- 日访问量: 5000+
- 总访问量: 500,000+;

## ■ 中文社区组织

- 社区核心成员: 10+, 来自业界TOP;
- 社区志愿者: 30+
- 专业合作媒体

## ■ 微信公众号

- 关注人数: 3000+
- 每周发布技术文章
- 访问总计: 500, 000+

## ■ 社区用户群

- 群成员: 210+
- 成员分布: 来自业界TOP及各行业
- 日活跃度: 50+ MSG

 **CockroachDB**  
著名的开源NewSQL数据库CockroachDB—中国社区

### CockroachDB中国社区

CockroachDB是一款开源的分布式数据库，具有NoSQL对海量数据的存储管理能力，又保持了传统数据库支持的ACID和SQL等，还支持跨地域、去中心、高并发、多副本强一致和高可用等特性。支持OLTP场景，同时支持轻量级OLAP场景。

2017年11月4日，受Cockroach Labs邀请，百度DBA团队在深圳百度国际大厦举办了中国区首届CockroachDB中国社区大会，标志着CockroachDB中国社区正式成立。大会邀请了众多数据界专家分享CockroachDB及数据库相关技术，同时吸引了来自各界的数据库开发爱好者及科技媒体参加。大会详细信息请点击参考：[CockroachDB中国社区成立大会](#)。

大会宣布了中国社区荣誉会长及组委会成员名单：[CockroachDB中国社区组委会名单](#)。

中国社区后续将负责在中国地区推广和应用CockroachDB以及组织CockroachDB在中国区的相关活动。

### 最近文章

#### CockroachDB设计文档 (下)

📅 2018年7月25日 📁 设计文档

 **CockroachDB**  
CockroachDB动态

发消息

#### CockroachDB存储引擎介绍（一）

2018年10月24日



#### 国内首发CockroachDB百度云数据库CRDB现已开放邀测

2018年10月10日



#### CockroachDB 2018 CAB Meetup @NewYork

2018年10月9日



#### CockroachDB 用户手册中文版（上）

2018年9月21日



#### CockroachDB Key-Value编码解析

2018年9月5日



# 社区RoadMap

- CBO优化
- Plan Cache
- 关联子查询
- Read From Follower
- Security

2.1

- 任意时间点恢复
- 导入导出
- SQL诊断
- CBO
- 透明加密
- ...

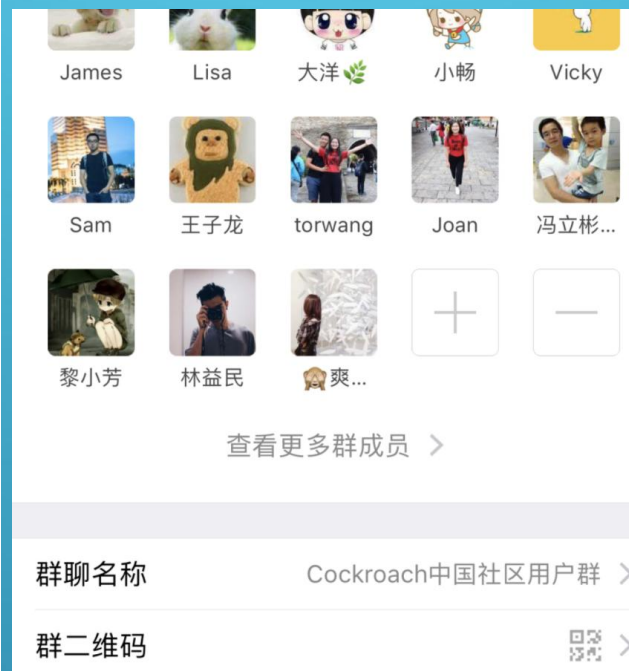
2.2

Future

- Load-Based Split
- Parallel Commit
- 存储过程
- 存储引擎改造
- ...



# 欢迎加入CockroachDB中国区



CockroachDB

<http://www.cockroachchina.cn>

<https://cloud.baidu.com/product/crdb.html>

[cockroach-china@baidu.com](mailto:cockroach-china@baidu.com)

# Thanks



Cockroach DB



百度云