2019
05
08-10
北京新云南皇冠假日酒店

数据风云十年变迁

DTCC 第十届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2019

IT168 ChinaUnix ITPUB

CockroachDB 2.x特性详解

Baidu
严龙

目录

CockroachDB 简介
CockroachDB 2.x关键特性
CockroachDB 2.x百度实践
Roadmap

PART 01

CockroachDB 介绍

Release Timeline

14年Github开源，17 年Release 1.0，目前版本2.1.6，即将release 19.1

1.0 — 2017.5.10

- DistSQL
- Online DDL
- Multi-Active
- Cloud Native

…

1.1 — 2017.10.12

- 支持滚动升级
- 支持import/dump
- Show/Cancel Query

…

2.0 — 2018.4.4

- GEO-Partition
- 支持Backup/Restore
- RBAC
- 支持JSON类型
- 支持SQL审计

…

2.1 — 2018.10.30

- CBO
- Read From Follower
- CDC
- 任意时间点备份恢复
- SQL可视化诊断
- 透明加密

…

PART 02

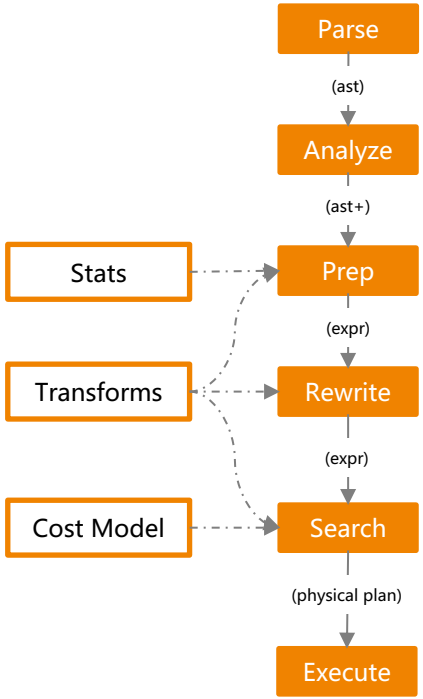
CockroachDB 2.x关键特性

- ✓ Cost Base Optimizer
- ✓ Read From Follower
- ✓ CDC

Cost Base Optimizer

- RBO :
  - ✓ Rule-based Optimizer
  - ✓ 根据通用的确定的优化规则变换重写执行计划
- CBO :
  - ✓ Cost-based Optimizer
  - ✓ 枚举可能的执行计划，计算每个执行计划的COST，最终选择COST最低的执行计划
- Why choose CBO?
  - ✓ RBO只根据静态规则优化执行计划，而CBO可以从时间和资源的角度评估，为每个SQL选择最优的执行计划
  - ✓ 随着优化规则越来越多，RBO添加和修改规则会越来越复杂

CBO——优化过程



- SQL解析
- 语法检查
- …

- 常量展开、类型检查、resolve names
- 语义检查
- …

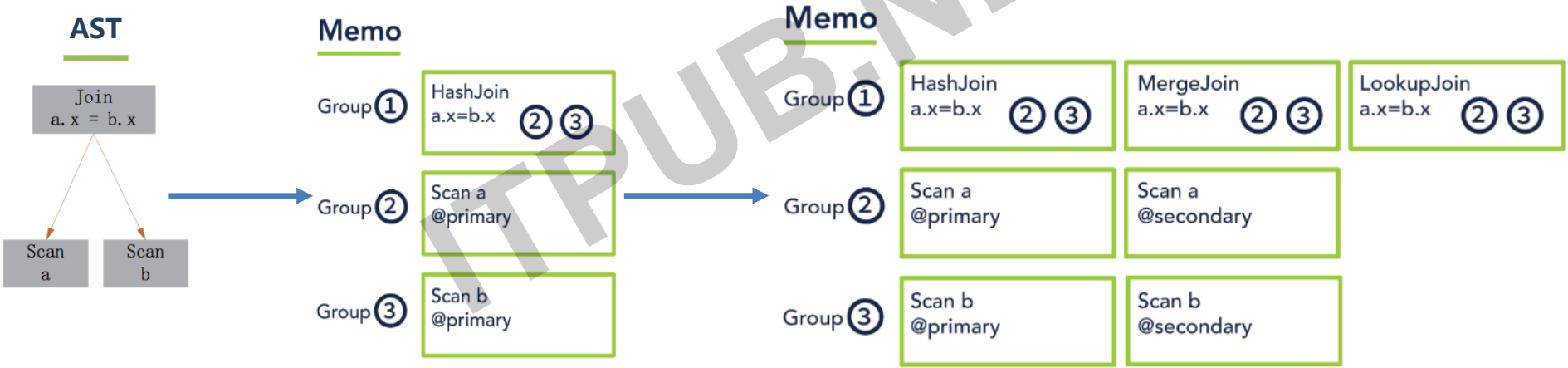
- 把AST变换成Memo expressions tree
- 填充统计信息
- …

- placeholder展开
- 根据规则重写执行计划
- …

- 根据规则产生等价表达式
- 生成执行代价，并查找最小代价的执行路径
  - ✓ Cost of a Expr Tree = Cost of Child Expr + Cost of Parent Expr （自底向上）
- 生成分布式执行计划

CBO——等价表达式管理

【Memo】- 管理查询所有等价计划树
- 跟踪每个Memo group的最低代价
└── Memo Group：一组 **logically equivalent的表达式**
└── Memo Expr：每个Memo表达式拥有一组Memo子表达式



- There are 12 possible plans (3 * 2 * 2) encoded in this memo.

CBO——统计信息生成

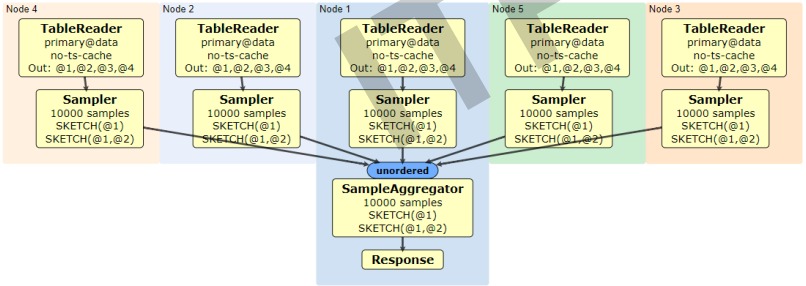
- 目标：供coster使用，计算每个Expr的计算成本（代价）
- 分类
  - Table statistics
    - 从metadata获取
    - 持久化在系统表system.table_statistics
  - Relational expression statistics
    - 从table statistics延伸
    - 仅对某一特定memo group有效
- 自动收集的统计信息

number of rows	number of columns
number of null value	Histogram (data value distribution)

Next Release:

Sketch
Range-level statistic

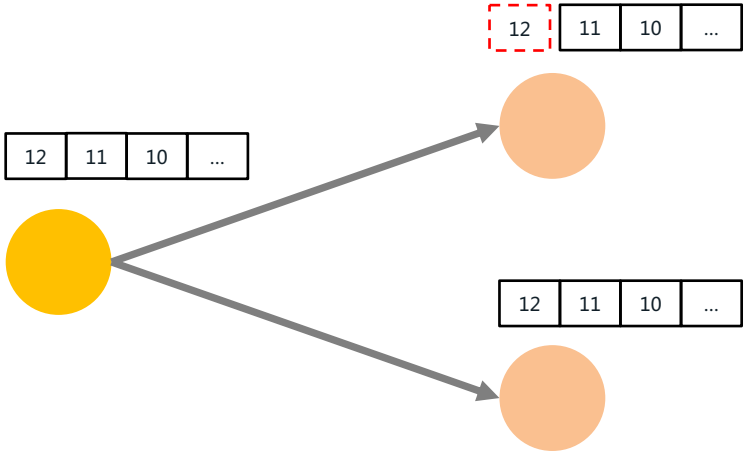
- 如何收集统计信息？



为什么要Read From Follower

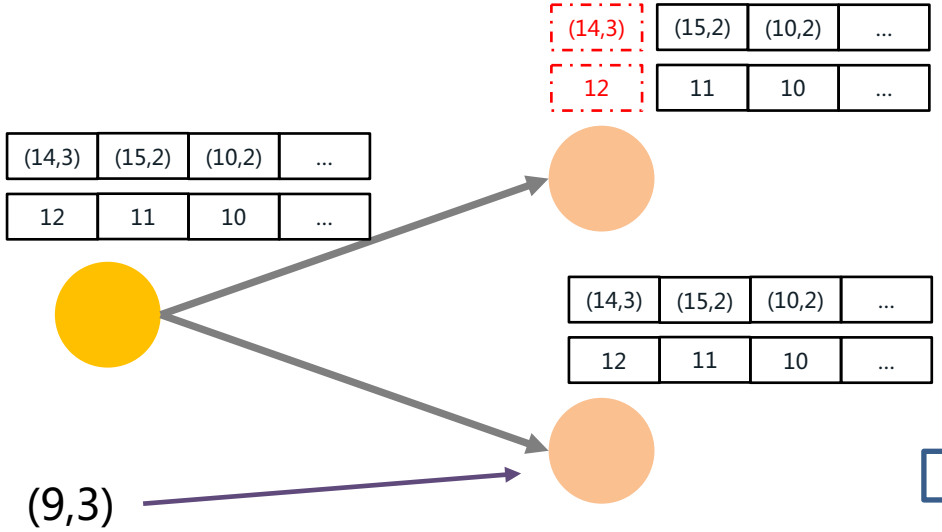
- Read from Leader/Lease holder：
  - ✓ Leader/Lease Holder 能提供一致性读
  - ✓ Follower的delay无法预期

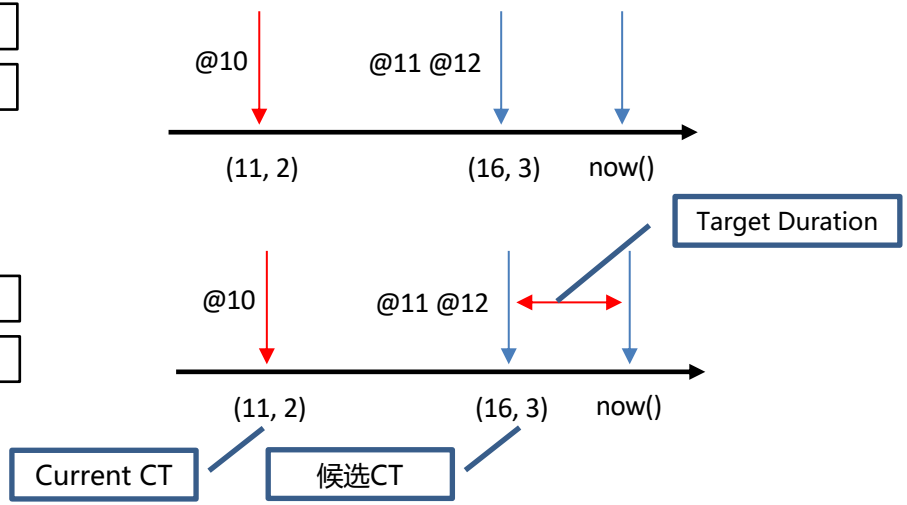
- Read from Follower：
  - ✓ Long Run类型查询
  - ✓ Historical reads
  - ✓ CDC
  - ✓ Immutable tables



一方有难，八方围观！

RFF——在Follower上一致性读

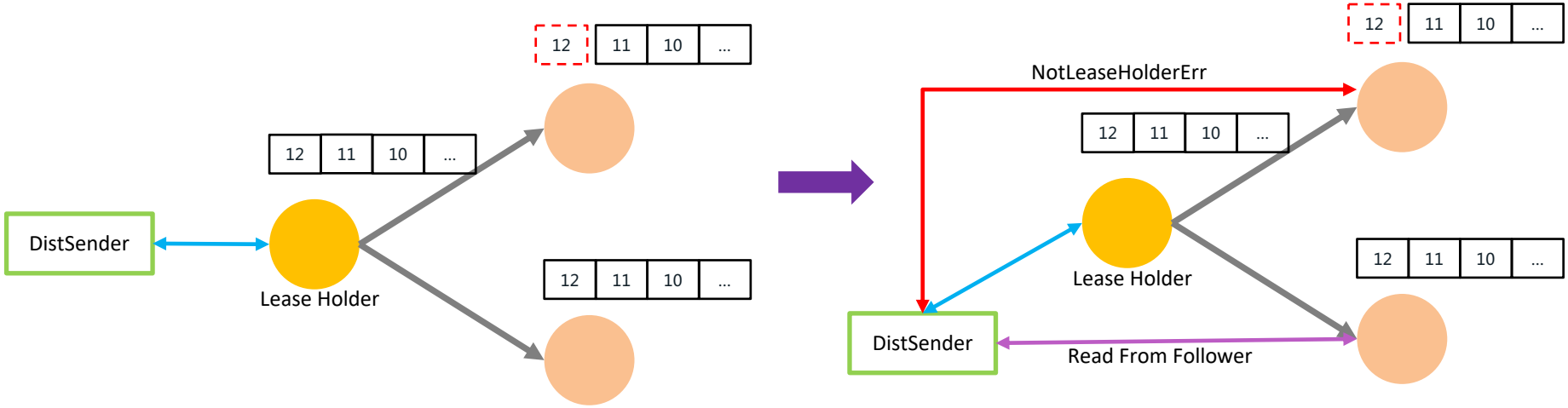




Follower一致性读：

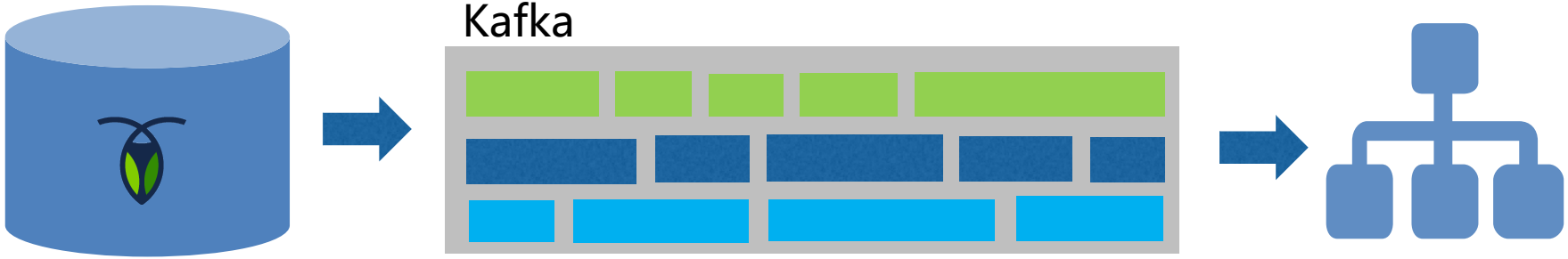
1. Log 的Applied Index 为 X，如果T满足： >= 提交Log X 的事务时间戳，且后续Applied Index > I 的Log的事务时间戳 > T
2. 时间戳<=T的事务可Read From Follower
3. 时间T称为 *Closed Timestamp*

RFF——数据路由



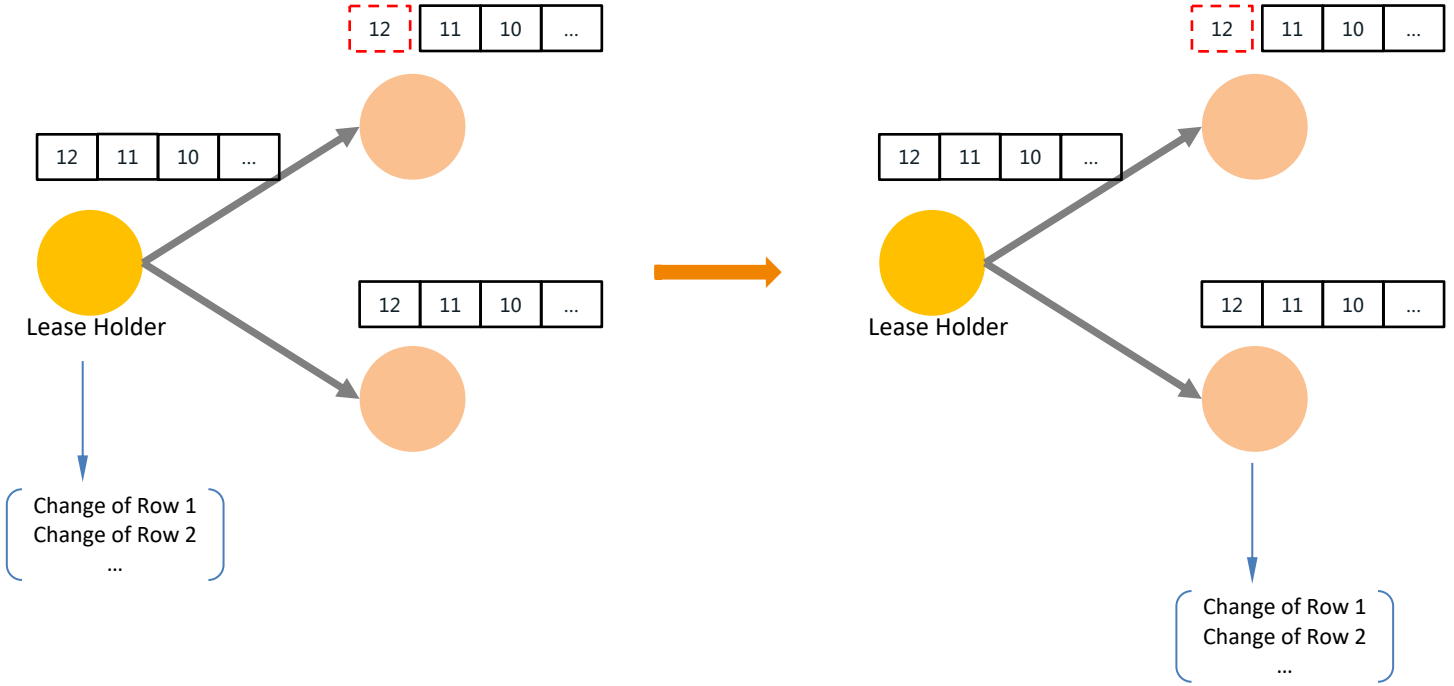
- 优先选择满足条件的最近的Follower
  - ✓ health
  - ✓ latency
  - ✓ locality
- Follower无法满足RFF，则返回NotLeaseHolderErr，请求重新路由到Lease Holder

CDC



- 集群间数据同步，可用于灾备，模拟沙盒环境等。
- 与其他数据生态互通，可用于数据分析，数据挖掘等。

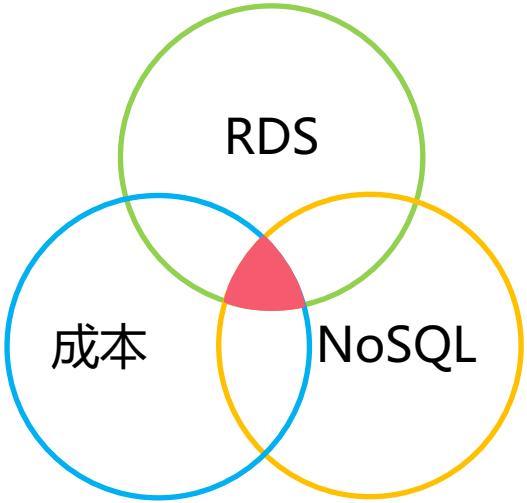
CDC

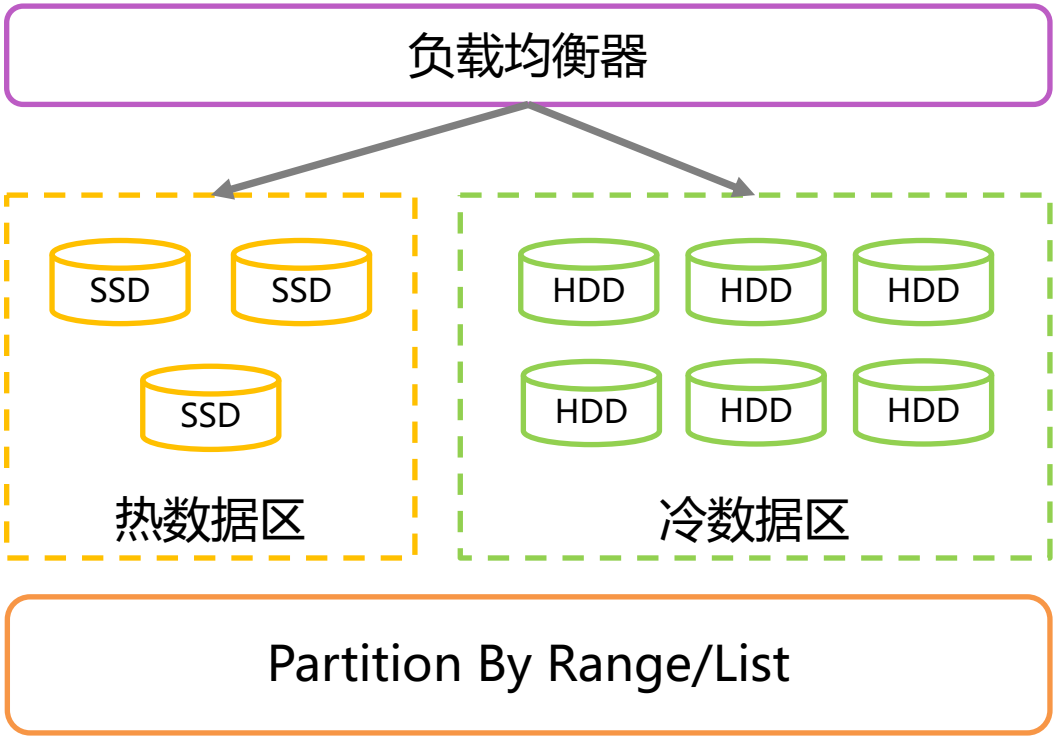

12
11
10
...
12
11
10
...
12
11
10
...
Lease Holder
Change of Row 1
Change of Row 2
...
12
11
10
...
12
11
10
...
12
11
10
...
Lease Holder
Change of Row 1
Change of Row 2
...

PART 03

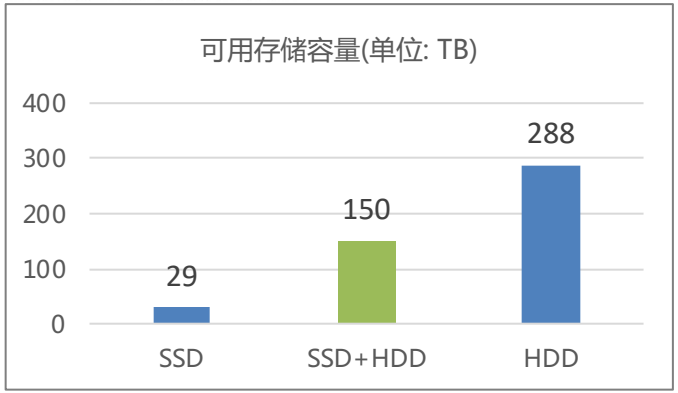
CockroachDB 2.x百度实践

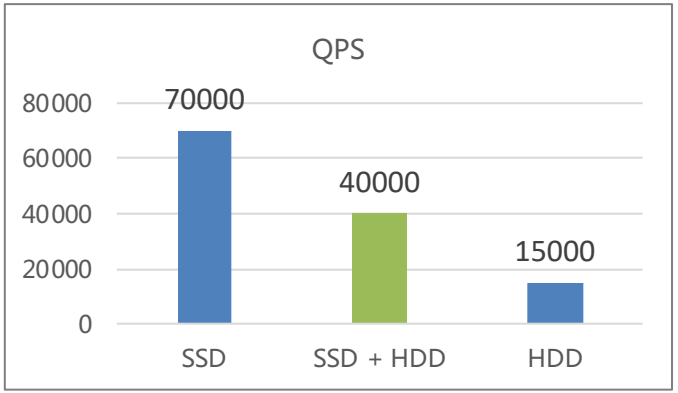
CockroachDB 实践——冷热分离

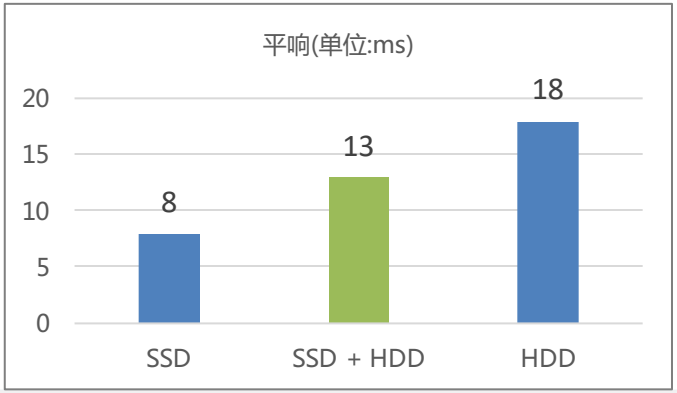




冷热分离


可用存储容量(单位: TB)
400
300
200
100
0
29
150
288
SSD
SSD+HDD
HDD

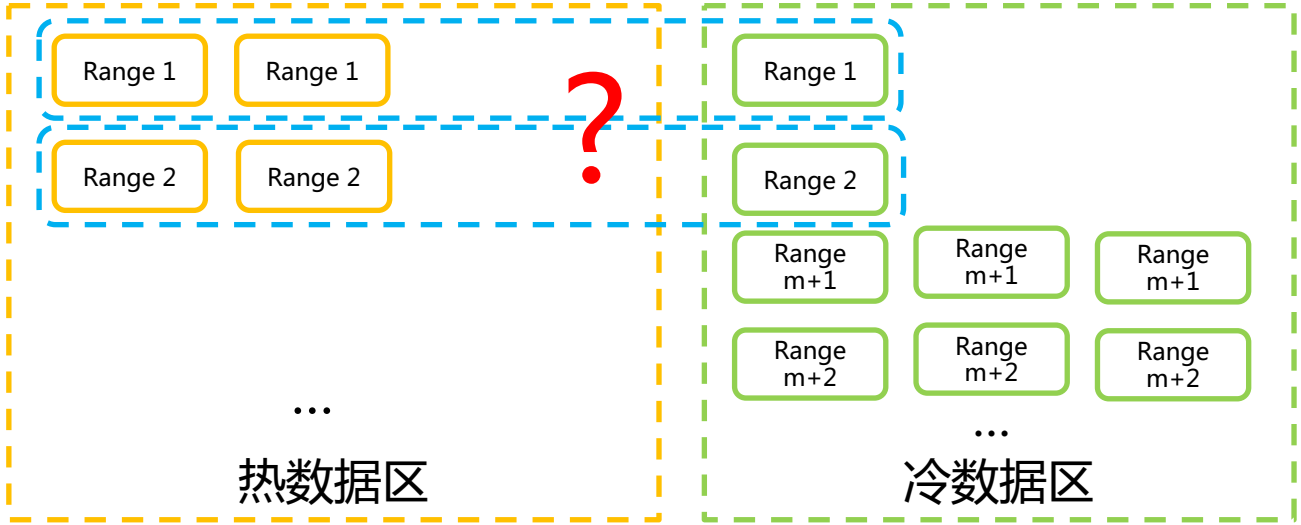

QPS
80000
60000
40000
20000
0
70000
40000
15000
SSD
SSD + HDD
HDD


平响(单位:ms)
20
15
10
5
0
8
13
18
SSD
SSD + HDD
HDD

下一步思考

如何进一步降低热数据长查询、复杂查询对写入的影响

- ✓ Read From Follower 分摊热数据读请求
- ✓ Read From Learner 分摊热数据读请求



PART 04

Roadmap

Roadmap

Schema Change

- Bulk处理，提升性能
- 可视化 – Admin UI “JOBS”
- 可在事务中使用

CBO更多优化

- 有效处理关联子查询
- 引入Plan Cache，加速CBO处理效率
- 优化决策考虑locality就近原则
- Query Optimizer Hints

Core

- Parallel Range Scans
- Load-based splitting
- changefeed推送支持PostgreSQL协议

Security

- 支持Kerberos集成
- Encryption

THANKS

中国社区网站：www.cockroachchina.cn