RadonDB

新一代分布式关系型数据库

张雁飞

青云QingCloud 数据库高级技术专家

12.09.2017

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

主办方 Geekbang InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

SPEAKER

- TokuDB内核维护者、XeLabs核心成员
- 淘宝核心系统/阿里云数据库内核组/青云数据库团队
- 目前在青云从事新一代数据库产品设计与研发工作

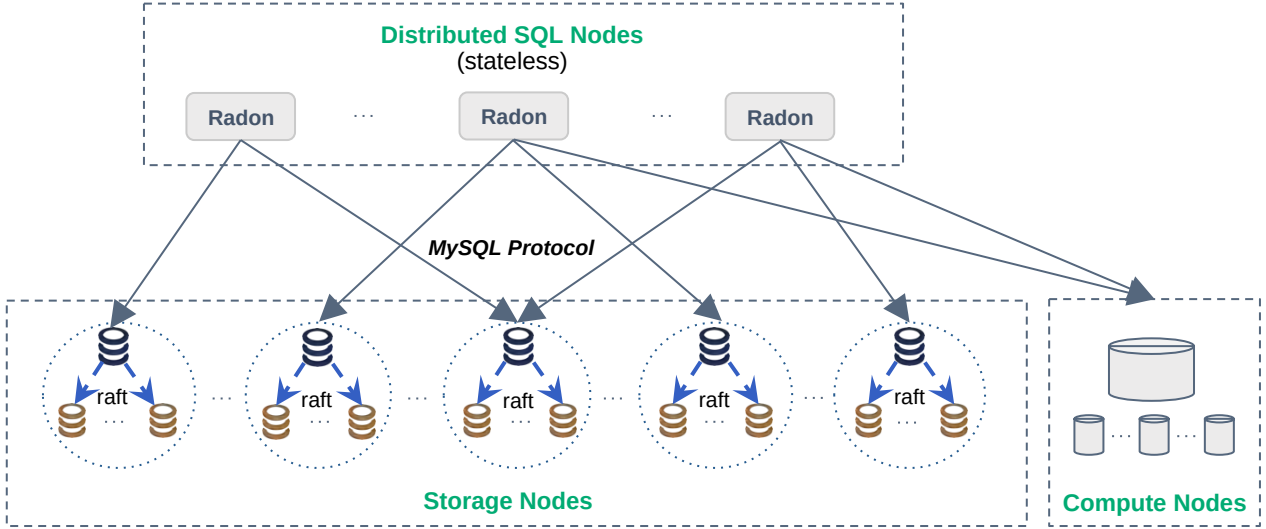
@BohuTANG

QINGCLOUD青云

RadonDB

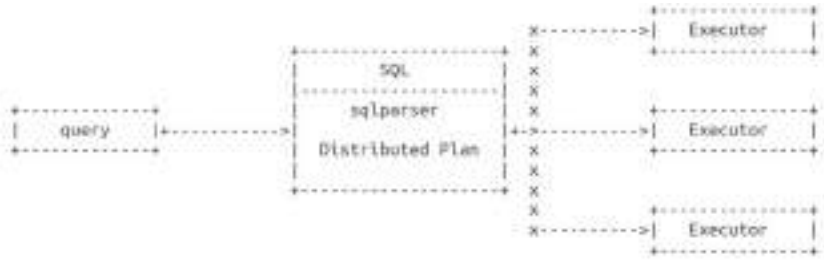
- 可扩展
- 高可用
- 强一致
- 易部署
- MyNewSQL

Architecture



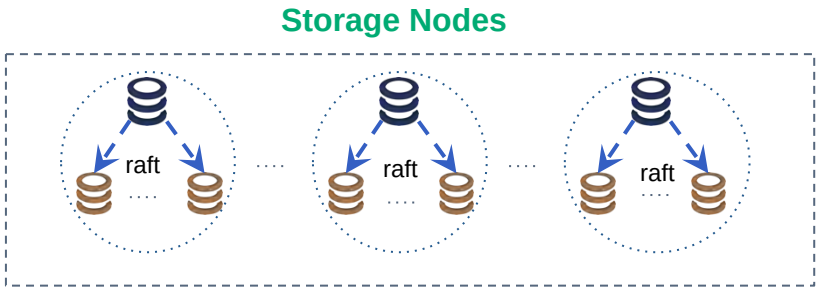
Distributed SQL

- 生成分布式执行计划和执行器
- 执行器并行执行
- 分布式事务协调器
- orderby/limit/groupby/aggregation/join …
- 无中心化设计

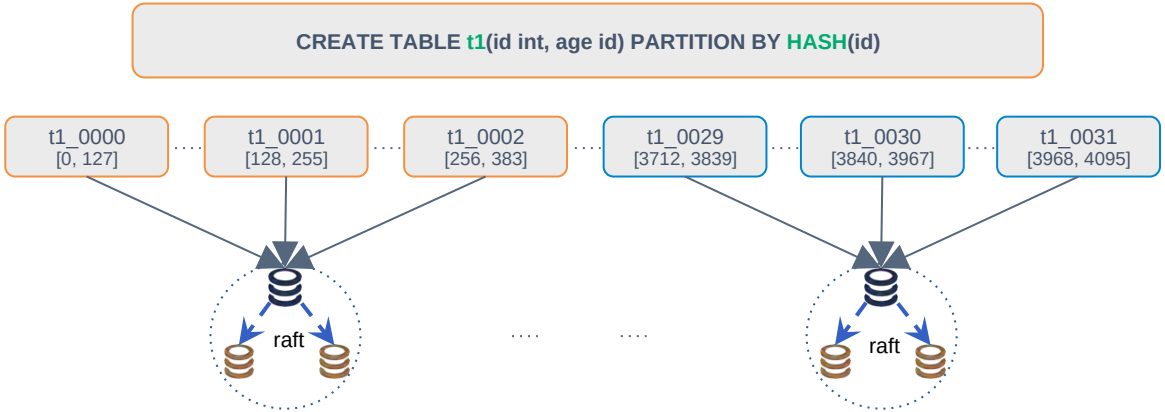


Storage Nodes

- 存储层由多个 node 组成
- 每个 node 负责部分数据存储
- node 由多副本组成
- 每个副本为一个 *MySQL*
- 不仅存储还有计算能力



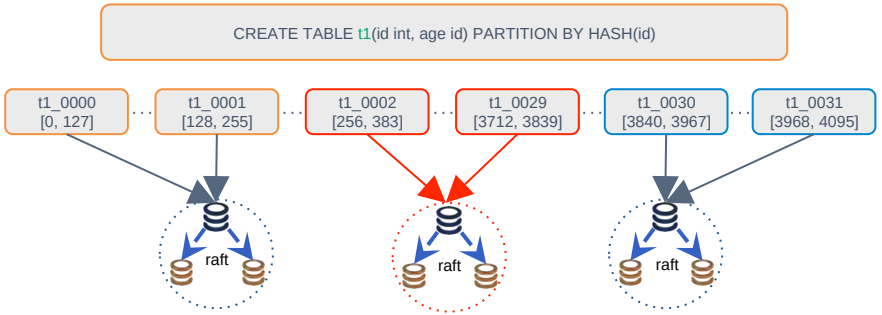
数据分布



- 整张表共 4096 slots
- 每个小表 128 slots
- 小表均匀分散在 node节点

扩容

- 小表可动态漂移
- 先全量后增量
- 较大/热度高者优先
- 资源分配最优化



副本

- 为什么不是KV? MySQL!
- 稳定可靠、多索引写原子保证
- 计算下推，数据就近计算原则
- SQL 与 Storage 数据传输最小化
- MySQL 8.0更加强大...


VOLTDB

FoundationDB's Lesson: A Fast Key-Value Store is Not Enough


Spanner: Becoming a SQL System

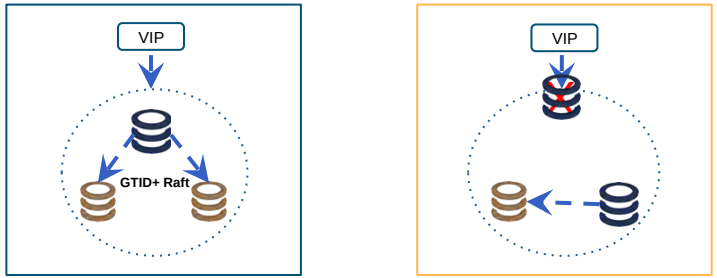
David F. Bacon, Nathan Bales, Nico Bruno, Brian F. Cooper, Adam Dickinson, Andrew Fikes, Campbell Fraser, Andrey Gubarev, Milind Joshi, Eugene Kogan, Alexander Lloyd, Sergey Melnik, Rajesh Rao, David Shue, Christopher Taylor, Marcel van der Holst, Dale Woodford

Google, Inc.

ABSTRACT

副本高可用

- GTID 作为 Raft Log Index
- Raft 协议选主、Log 并行复制
- 主副本故障秒级切换即可服务
- 强 Semi-Sync 确保事务不丢失
- 副本故障可快速流式重建
- 无中心化，可跨机房部署



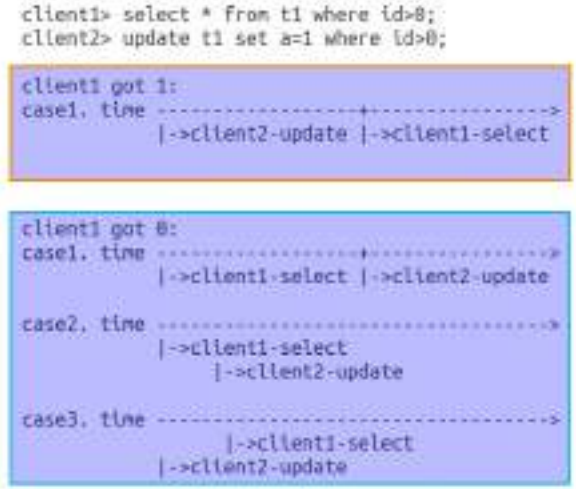
Raft+MySQL = Raft 选主+GTID 并行复制+强 Semi-Sync 数据强一致、切换零丢失

分布式事务

- 事务管理
- 事务可靠性
- Snapshot Isolation 隔离级别

SI隔离级别

- 未提交不可见
- 部分提交不可见



SI检测

- XeLabs/go-jepsen
- 1个更新线程，16个扫表线程
- 100多亿次操作和检测
- 随机 kill 存储节点主副本

```
Thread1:
update jepsen_si set score=0;
```

```
ThreadN:
select score from jepsen_si;
for cur := row.next() {
   if pre != cur {
      errors++
   }
}
```

Radon - Binlog

- Statement + GTID格式
- 可被订阅用于数据同步(计算节点)

OLTP + OLAP

- 独立计算节点(Compute Node)
- 数据通过 Radon Binlog 进行快速同步
- SQL 层自动路由复杂查询到计算节点
- 优点: 高并发事务与复杂查询资源隔离
- 缺点: 存储 2 份，目前使用压缩解决

Backup & restore

- XeLabs/go-mydumper
- 批量并行流式导出，snapshot备份
- 批量并行导入

性能

sysbench: 16表，512线程，随机写，5000万条数据

	Transaction Per Second(TPS)	Response Time(avg)	规格
RadonDB (1SQL节点，4 存储节点)	**26,589**	**20ms**	4 存储节点(16C64G超高性能主机) sync_binlog=1 innodb_flush_log_at_trx_commit=1
单机 MySQL (QingCloud RDB)	9,346	73ms	RDB(16C64G超高性能主机) sync_binlog=1 innodb_flush_log_at_trx_commit=1

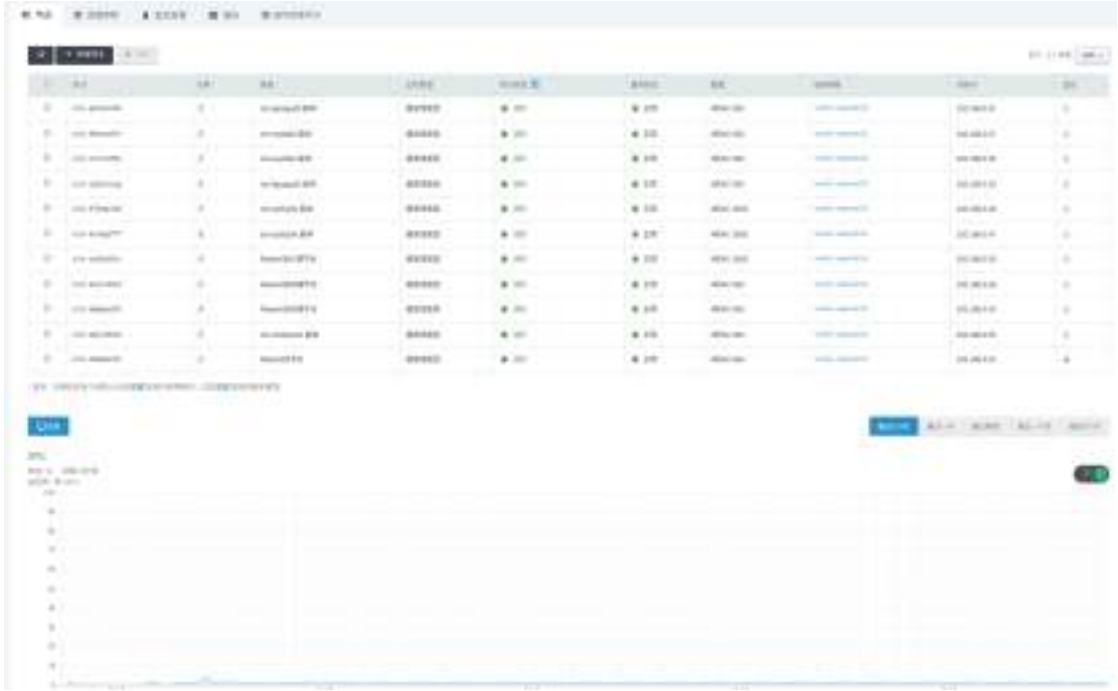
部署

- 云化部署
- 简单快捷

QINGCLOUD青云

资源监控

- CPU
- 内存
- 硬盘IOPS/使用率...



监控

- 全链路监控
- mysql> show processlist;
- mysql> show txnz;
- mysql> show queryz;

展望

- MyNewSQL刚刚开始
- Hybrid Row/Column Data Storage
- 分布式事务改进(Linearizability)
- RadonDB将全部开源(2018)

Thank you.

Array@yunify.com