滴滴高性能KV存储系统实践

Rockstable简介

盛克华
2016-10-1

滴滴

Agenda

简介
目标问题
使用情况
实现概述

Rockstable简介

分布式列式KV存储系统

NoSQL、Schemaless

Rocksdb
Table Rocks!

参考:
bigtable、hbase、cassandra、voldemort
mola、tera、ckv、tair、bada

目标问题

滴滴后台各种实时特征存取需求
- 多更新源
- 多使用方

目标问题

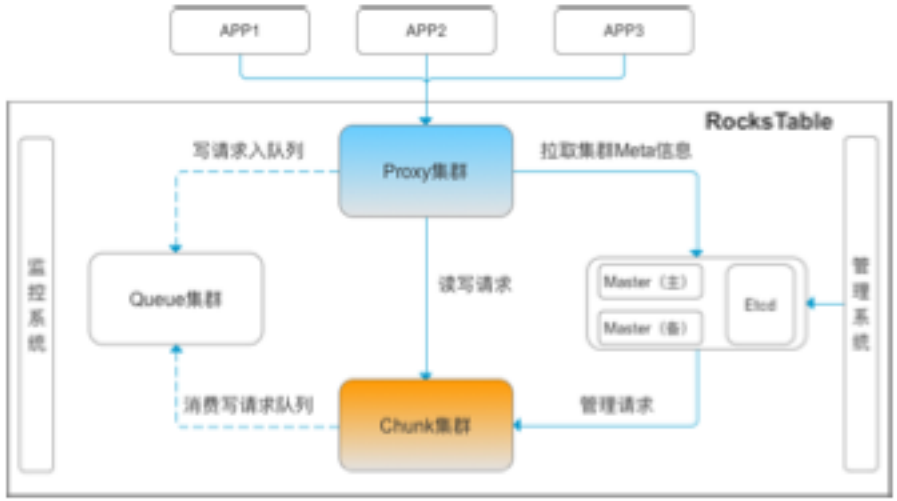
滴滴后台各种实时特征存取需求

- 多更新源
- 多使用方
- 高度稀疏
- 批量读写
- 实时&批量
- 高并发、低延迟、高可用

使用情况

多套集群，总计接近千台机器
高峰期2kw key/s读请求，50w key/s写请求
列数在百级别
三副本，TB级数据量
存储主要使用内存与SSD

实现概述-总体结构



实现概述-主要特性

最终一致性（N副本）
支持表
支持列（Schemaless）
支持批量读、批量写、列式Append
支持TTL（Table、Key、Column）
支持平滑扩容、自动修复

最终一致性（N副本）
支持表
支持列（Schemaless）
支持批量读、批量写、列式Append
支持TTL（Table、Key、Column）
支持平滑扩容、自动修复
支持Quata、白名单
支持Mc、Thrift两种接口
支持批量数据扫描、数据修复
支持数据持久化
支持集群间数据同步

实现概述-集群管理

命令行的管理入口

- 创建、删除、查看表
- 添加、摘除、查看chunk
- 表平衡、表修复

```
Support Cmd:
./MasterCmd [-h host_ip] [-p port] [-t time_out] add_chunk_server
./MasterCmd [-h host_ip] [-p port] [-t time_out] add_trp_meta
./MasterCmd [-h host_ip] [-p port] [-t time_out] create_table
./MasterCmd [-h host_ip] [-p port] [-t time_out] del_all_copy_tasks
./MasterCmd [-h host_ip] [-p port] [-t time_out] del_chunk_server
./MasterCmd [-h host_ip] [-p port] [-t time_out] del_table
./MasterCmd [-h host_ip] [-p port] [-t time_out] del_trp_meta
./MasterCmd [-h host_ip] [-p port] [-t time_out] force_balance
./MasterCmd [-h host_ip] [-p port] [-t time_out] force_table_balance
./MasterCmd [-h host_ip] [-p port] [-t time_out] get_chunk_trps
./MasterCmd [-h host_ip] [-p port] [-t time_out] host_master
./MasterCmd [-h host_ip] [-p port] [-t time_out] repair_all_table
./MasterCmd [-h host_ip] [-p port] [-t time_out] repair_table
./MasterCmd [-h host_ip] [-p port] [-t time_out] restart_trp
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_chunks
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_copy_tasks
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_metas
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_sick_trps
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_tables
./MasterCmd [-h host_ip] [-p port] [-t time_out] stop_copying_tasks
./MasterCmd [-h host_ip] [-p port] [-t time_out] update_chunk
./MasterCmd [-h host_ip] [-p port] [-t time_out] update_table_bypass
```

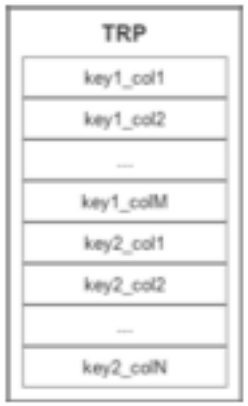

实现概述-数据模型

```
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_sick_trps
./MasterCmd [-h host_ip] [-p port] [-t time_out] show_tables
./MasterCmd [-h host_ip] [-p port] [-t time_out] stop_copying_tasks
./MasterCmd [-h host_ip] [-p port] [-t time_out] update_chunk
./MasterCmd [-h host_ip] [-p port] [-t time_out] update_table_bypass
```


实现概述-数据模型

每个表根据ID进行均匀分片，每个分片存储三个副本
TRP：Tablename_ReplicaID_PartitionID，代表一个表的一个副本的一个分片

Key组成：colfamily#key#colname



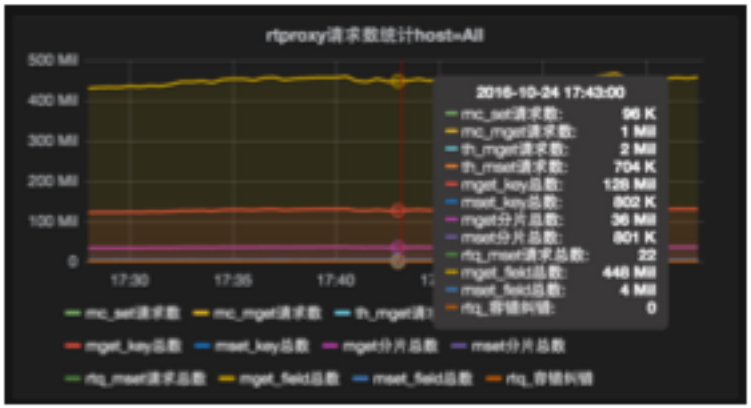
实现概述-Balance

1、停止一个TRP
2、Proxy开始写队列
3、拷贝TRP
4、启动新TRP，消费队列，追齐数据
5、Proxy访问新TRP

3、拷贝TRP
4、启动新TRP，消费队列，追齐数据
5、Proxy访问新TRP

实现概述-监控

200+根监控曲线，包括Qps统计、错误统计、耗时统计、实例数统计等等

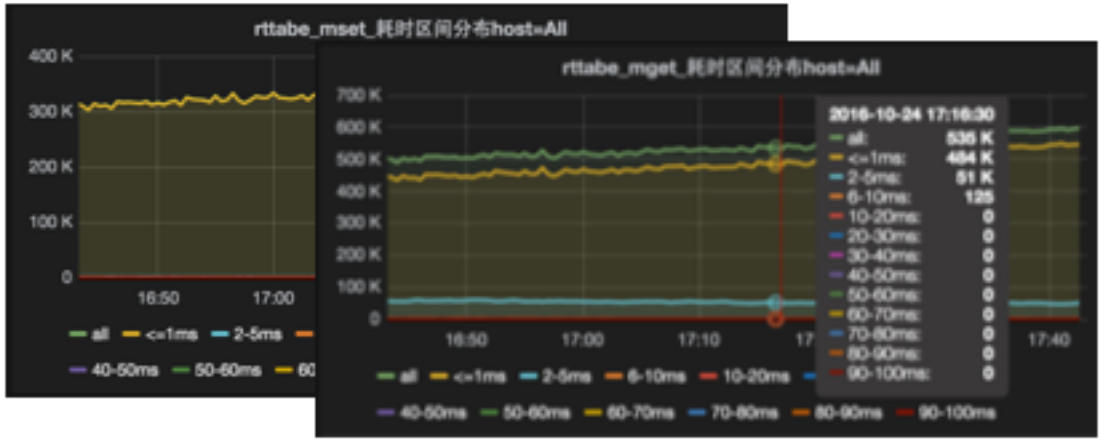


实现概述-性能

线上真实响应时长如截图，读写耗时99.9%在5ms以内

实现概述-性能

线上真实响应时长如截图，读写耗时99.9%在5ms以内



实现概述-NodeMgr

下游节点动态管理器，支持自动摘除与恢复，实现了golang/c++/php版本

接口

- init/uninit
- getNode
- vote

THANK YOU

与卓越的人一起共事，成为卓越的人
欢迎自荐: shengkehua@didichuxing.com

滴滴

北京嘀嘀无限科技发展有限公司
北京市海淀区东北旺路9号软件园 数字山谷B1号楼