云计算开源产业联盟

全球云计算开源峰会2017 中国·北京
聚合云计算新势力，拥抱全世界新开源
GLOBLE CLOUD COMPUTING OPEN SOURCE SUMMIT

小米结构化存储系统及融合云平台的设计与实践

林尚泉
linshangquan@xiaomi.com

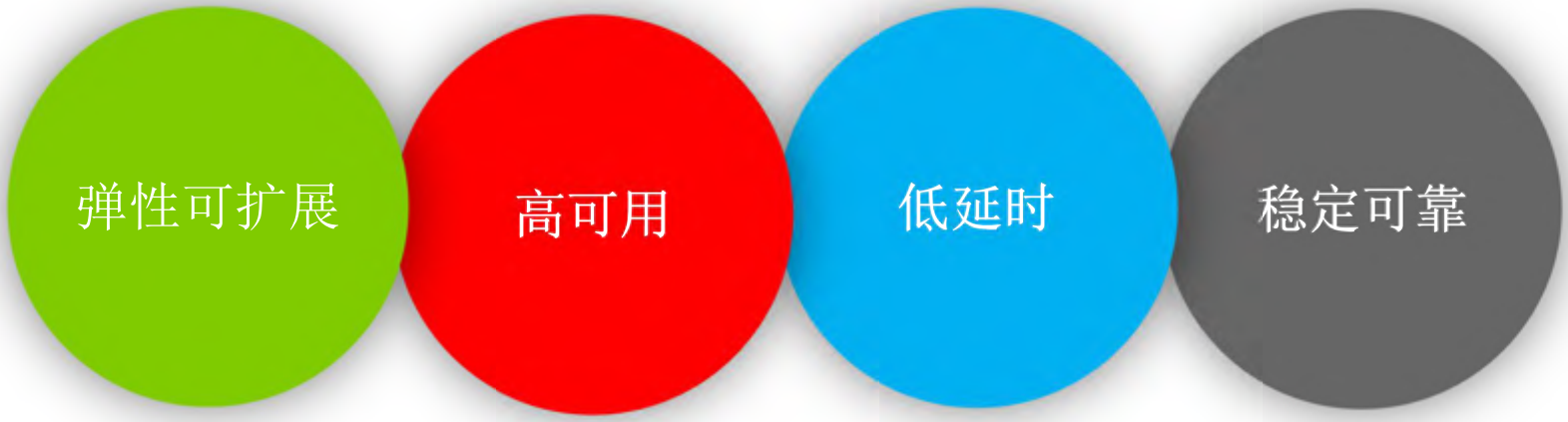
大纲

- 小米结构化数据存储
- 融合云平台

结构化存储

分布式NoSQL数据库服务

对标AWS DynamoDB



结构化存储

SDS

(Structured Data Store)

基于HBAS

无状态公网访

简化认证和配

多语言SD

多租户

功能扩展

ACL

流量控制

数据类型

二级索引

STREAM

软删除

数据冷备

结构化存储

小米内部业务

生态链业务

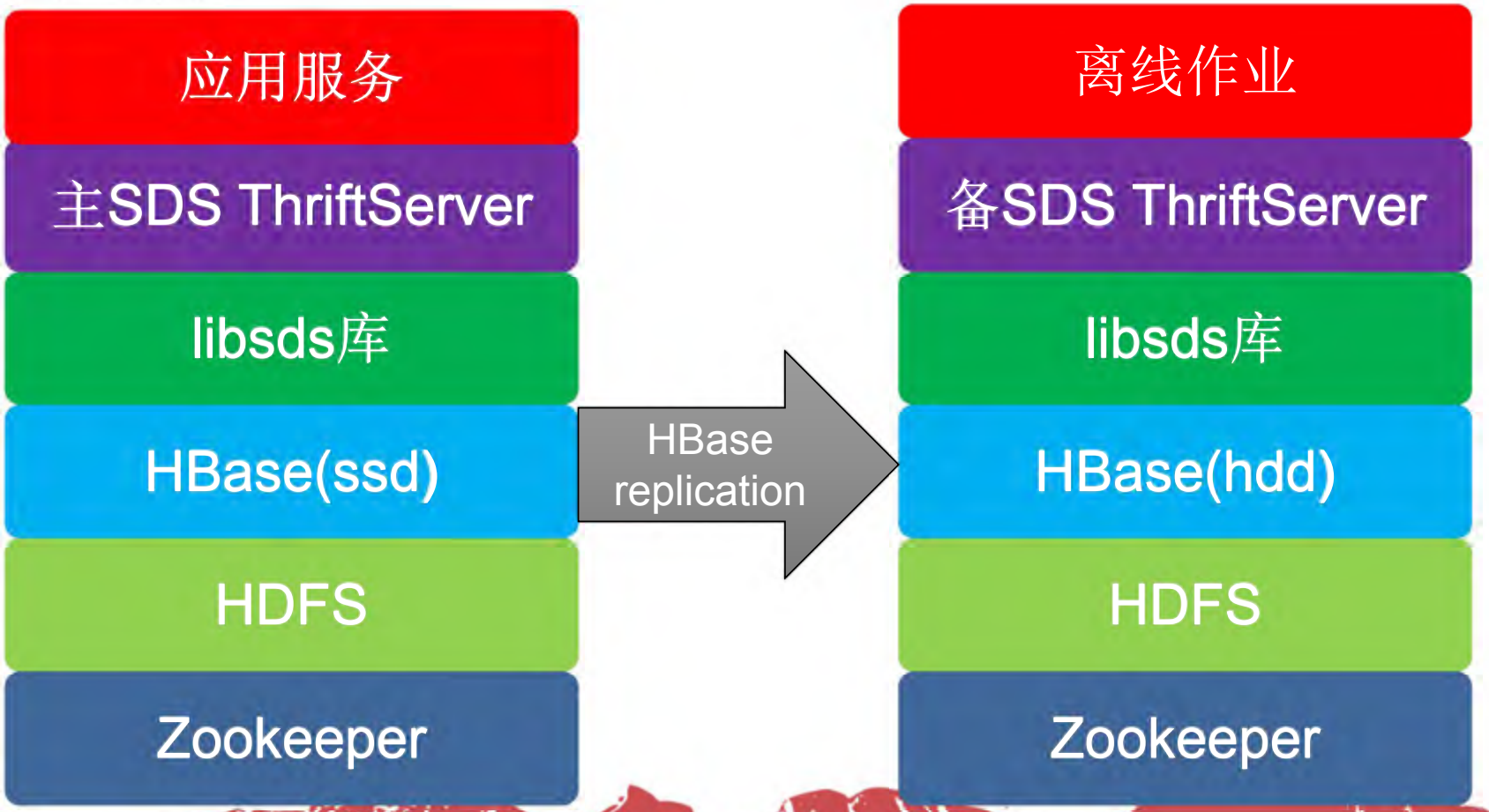
结构化存储

- 应用规模

国际化	• 北京，天津，美国，新加坡
机器规模	• 100+
业务数	• 20+
数据量	• 100+TB，数千亿行
单集群QPS	• 10万级

结构化存储

典型部署

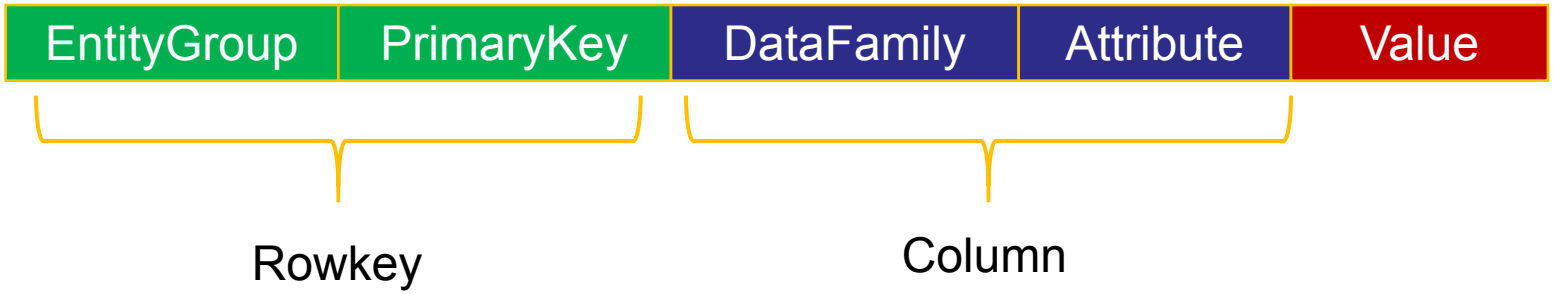

应用服务
主SDS ThriftServer
libsds库
HBase(ssd)
HDFS
Zookeeper
HBase
replication
离线作业
备SDS ThriftServer
libsds库
HBase(hdd)
HDFS
Zookeeper

结构化存储

- libsds库

结构化存储

- 规范化数据模型
  - Rowkey前缀：实体组键，支持哈希分布
  - Rowkey后缀：主键
  - Column Qualifier：数据列



结构化存储

数据类型支持

- bool/int8/int16/int32/int64/float/double/string/binary/set

HBASE-8201和SQLite编码方案[1]

- 顺序保持编码

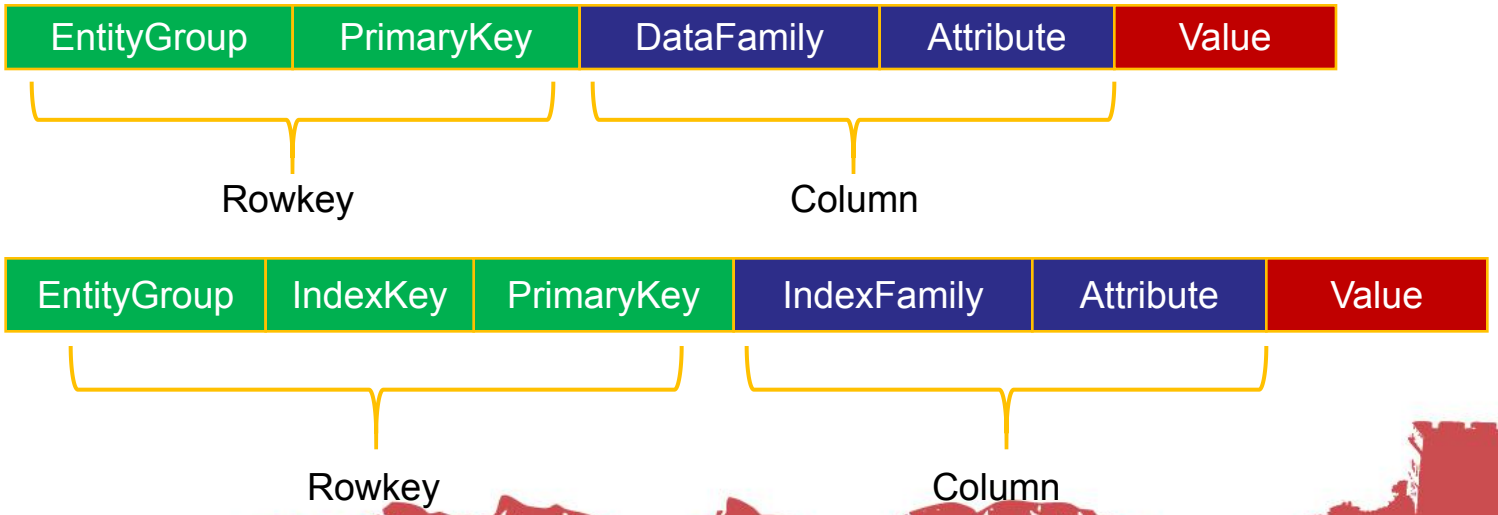
支持逆序

- 按位取反

[1]http://sqlite.org/src4/doc/trunk/www/key_encoding.wiki

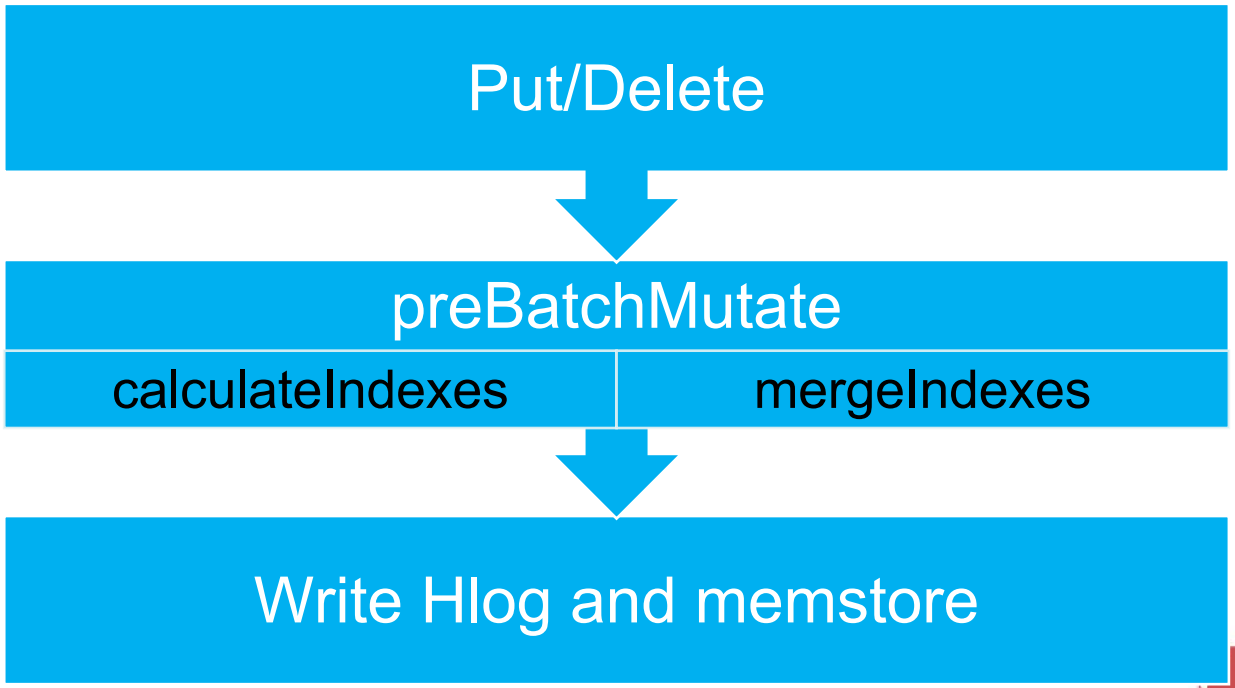
结构化存储

- 局部二级索引：实体组内部
  - 同一张表，不同Column Family
  - 基于正则表达式的前缀分割策略(RegexPrefixRegionSplitPolicy)：保证同一个实体组键数据不能split



结构化存储

- 局部二级索引：HBase Coprocessor——IndexObserver



结构化存储

- 多种类型局部二级索引

EAGER

更新删除时同时删除失效索引，适合写少读多

LAZY

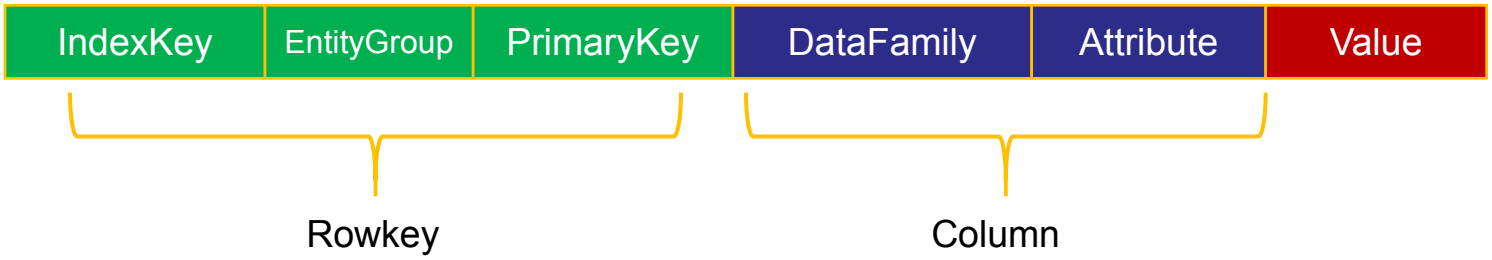
读取时判断索引有效性，更新时不做额外操作，适合写多读少

IMMUTABLE

适合只读数据一次性写入，读写均无需额外判断

结构化存储

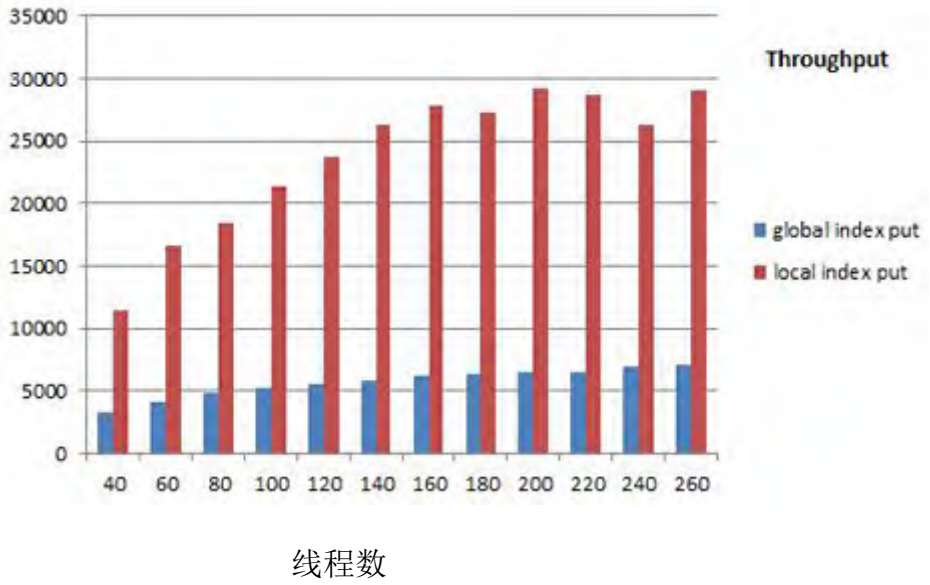
- 全局二级索引
  - 使用单独的HBase 表存储索引
  - Google percolator[2]的 HBase实现：Themis[3]，保证跨表更新的原子性
  - Chronos[4]全局单调递增时间戳



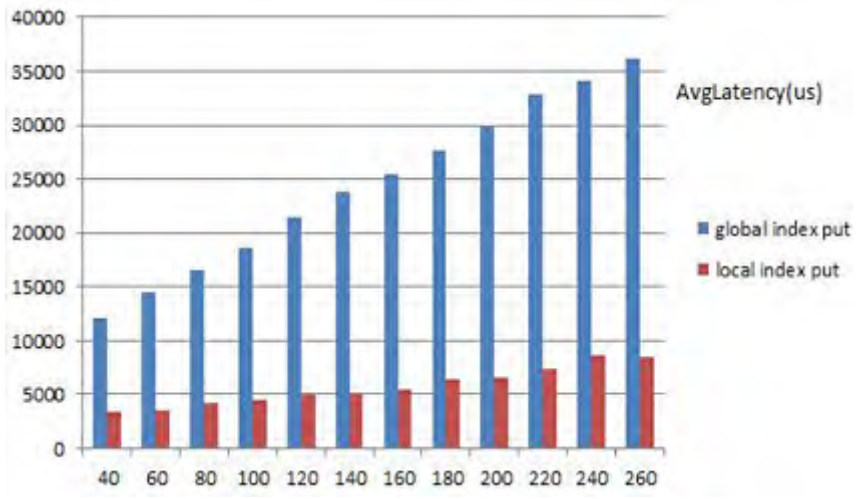
[2]Large-scale Incremental Processing Using Distributed Transactions and Notifications
[3]https://github.com/xiaomi/themis
[4]https://github.com/xiaomi/chronos

结构化存储

- 局部二级索引 VS. 全局二级索引(写)



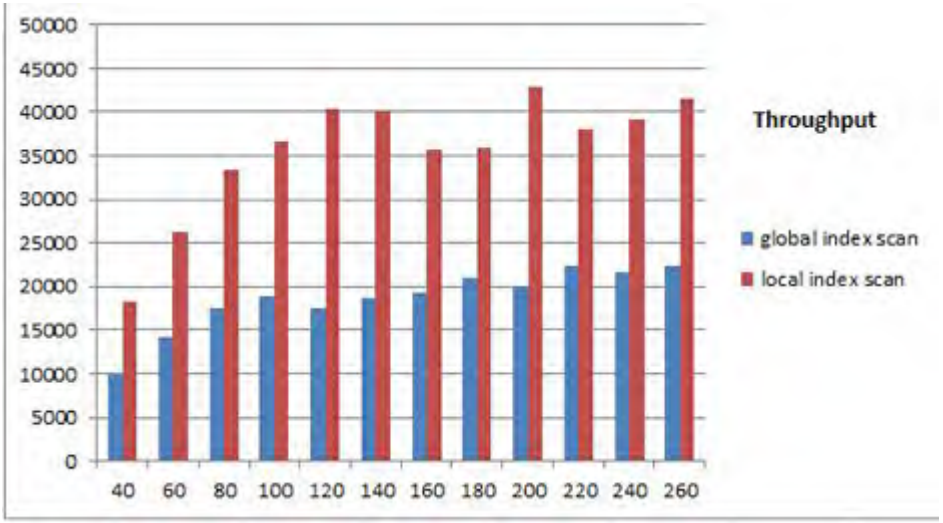
线程数



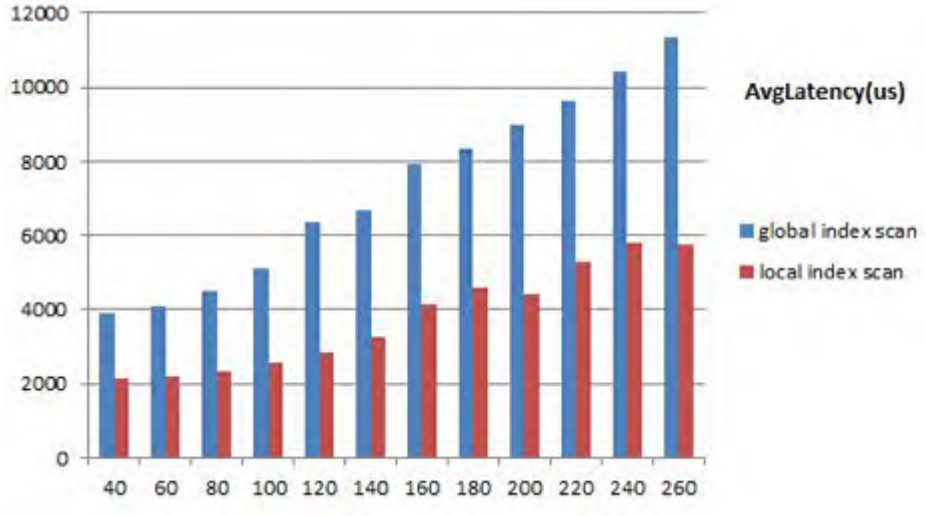
线程数

结构化存储

- 局部二级索引 VS. 全局二级索引(读)



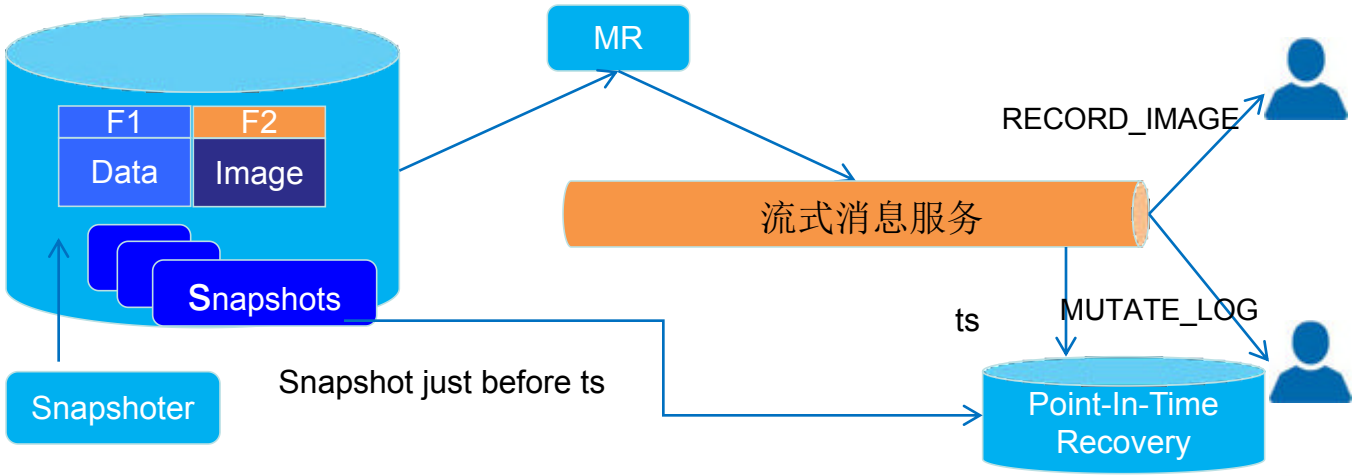
线程数



线程数

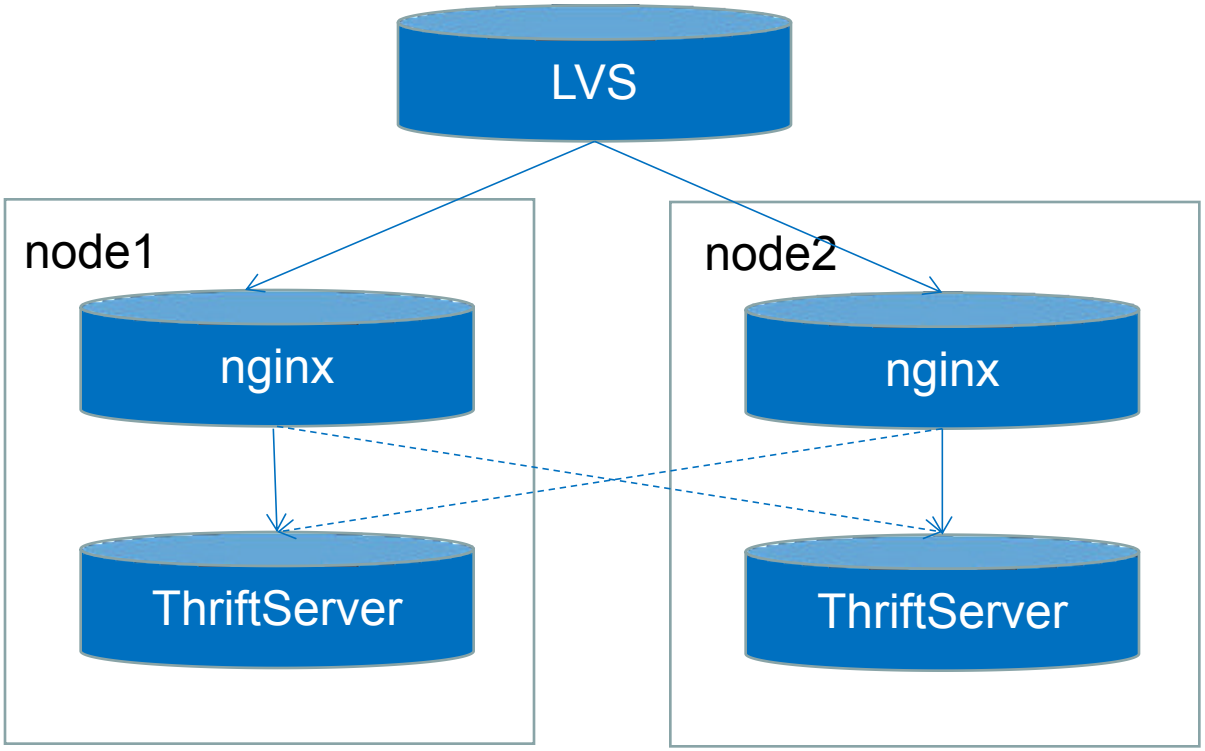
结构化存储

- stream
  - RECORD_IMAGE：增量备份
  - MUTATE_LOG：Point-In-Time Recovery



结构化存储

- SDS ThriftServer：无状态 Http Server



结构化存储

SDS ThriftServer

Http 无状态公网访问

简化认证和配置

多语言 SDK

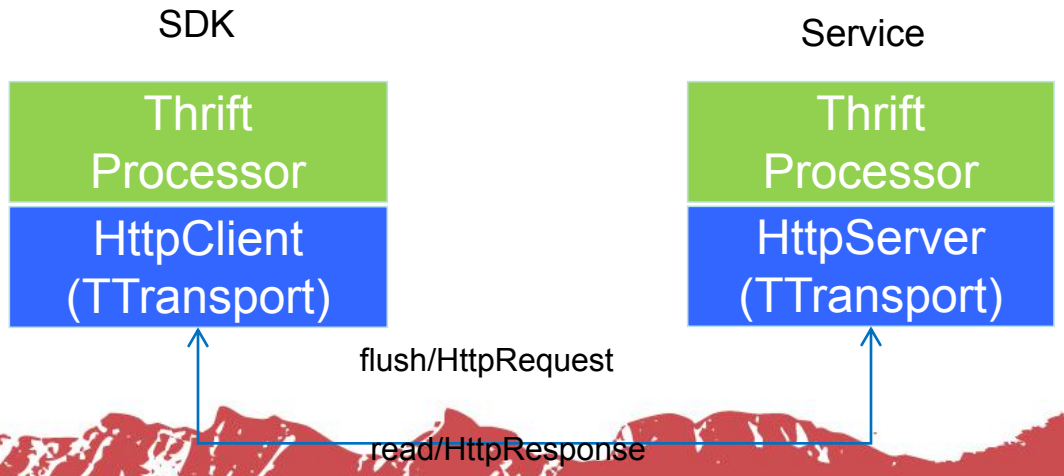
多租户

ACL

流量控制

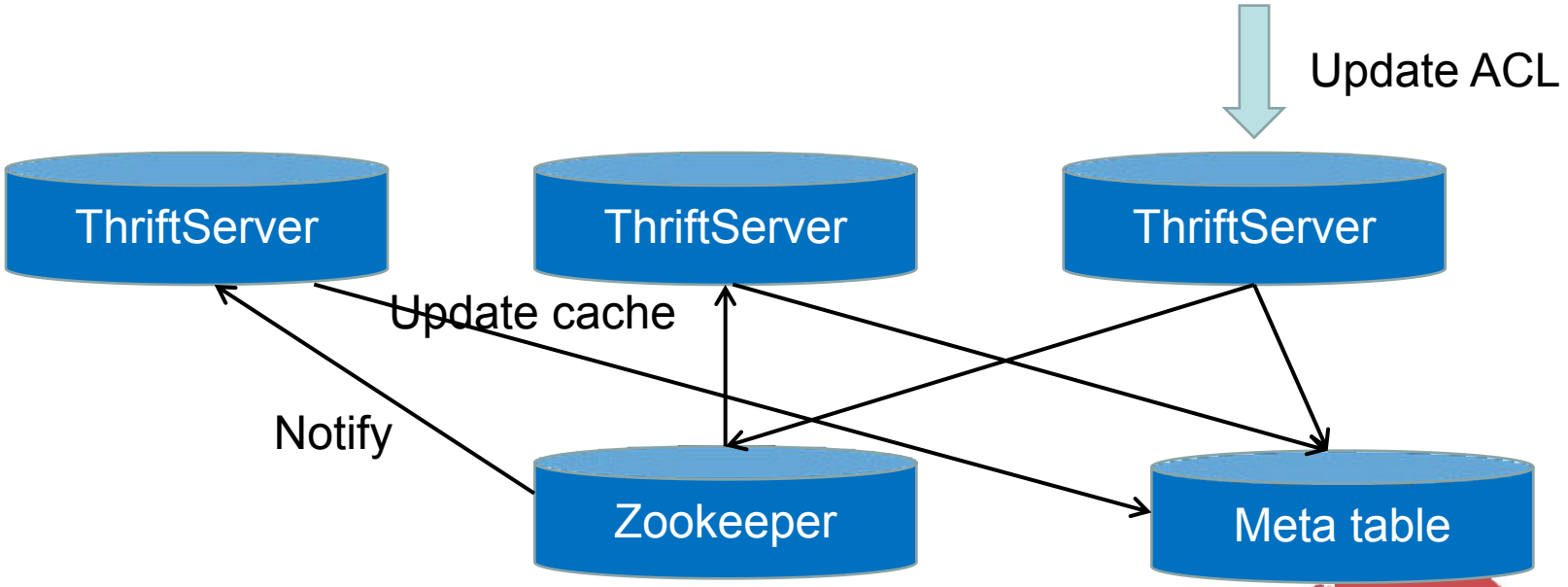
结构化存储

- SDS ThriftServer
  - Http 无状态公网访问
  - 简化认证和配置：使用小米融合云或开放平台认证，对外屏蔽了Zookeeper配置和Kerberos认证
  - 多语言SDK(java/python/php/node.js/c++/go/javascript)



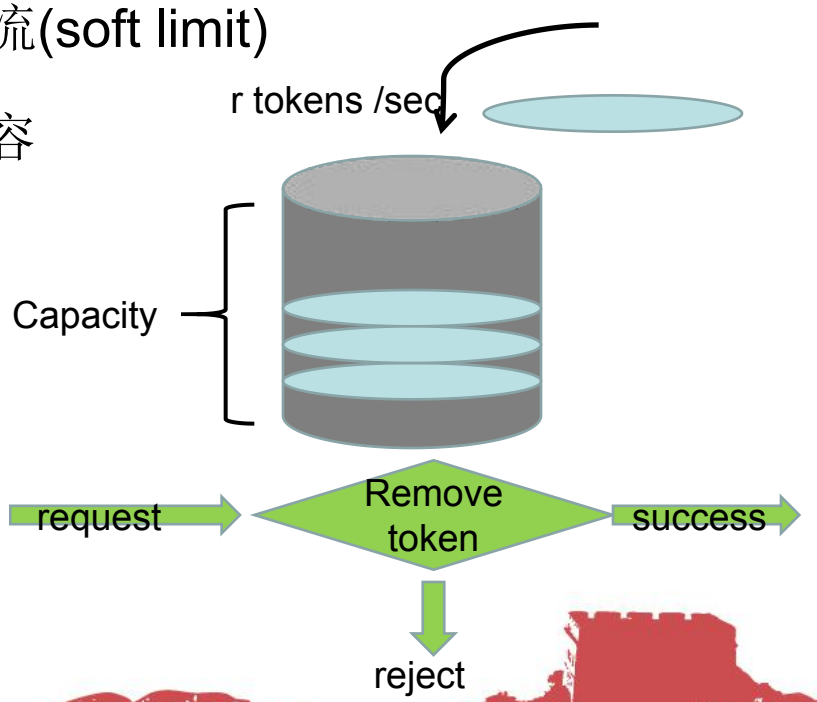
结构化存储

- SDS ThriftServer (ACL)
  - 元数据表: tableId => {ACL信息}
  - 本地cache元数据，zookeeper watcher 同步



结构化存储

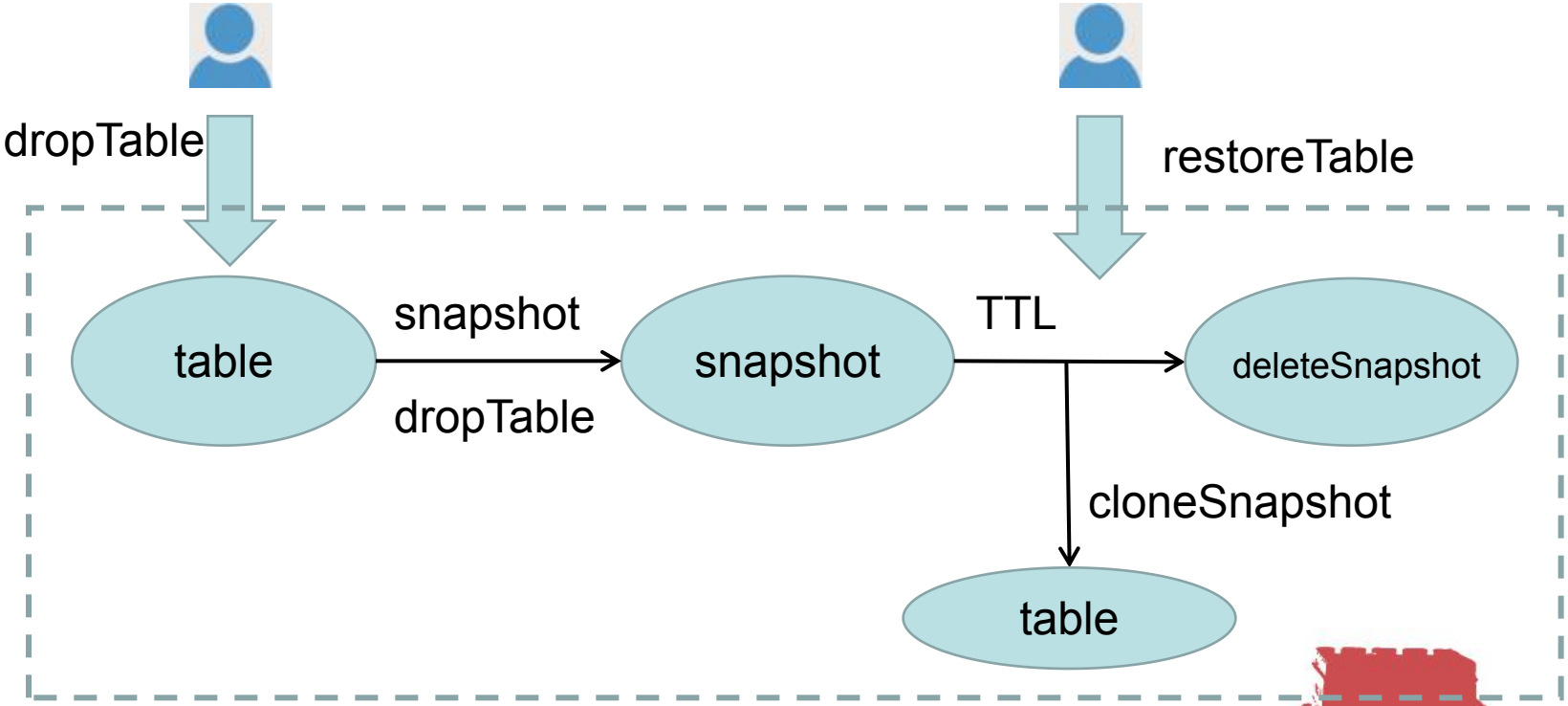
- SDS ThriftServer (流量控制)
  - 用户设置表的读写quota，设置表quota时检查集群能力
  - 基于token bucket算法[5]进行限流(soft limit)
  - 集群能力使用超80%时提醒扩容



[5]https://en.wikipedia.org/wiki/Token_bucket

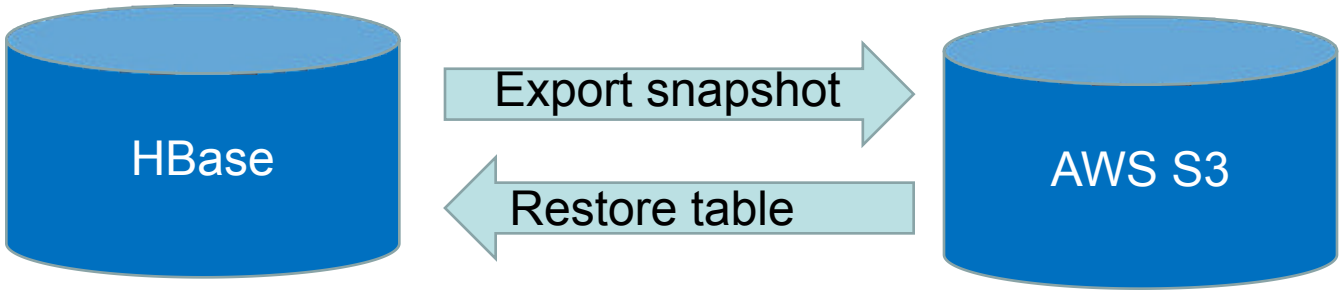
结构化存储

- 软删除


dropTable
restoreTable
table
snapshot
dropTable
snapshot
TTL
deleteSnapshot
cloneSnapshot
table

结构化存储

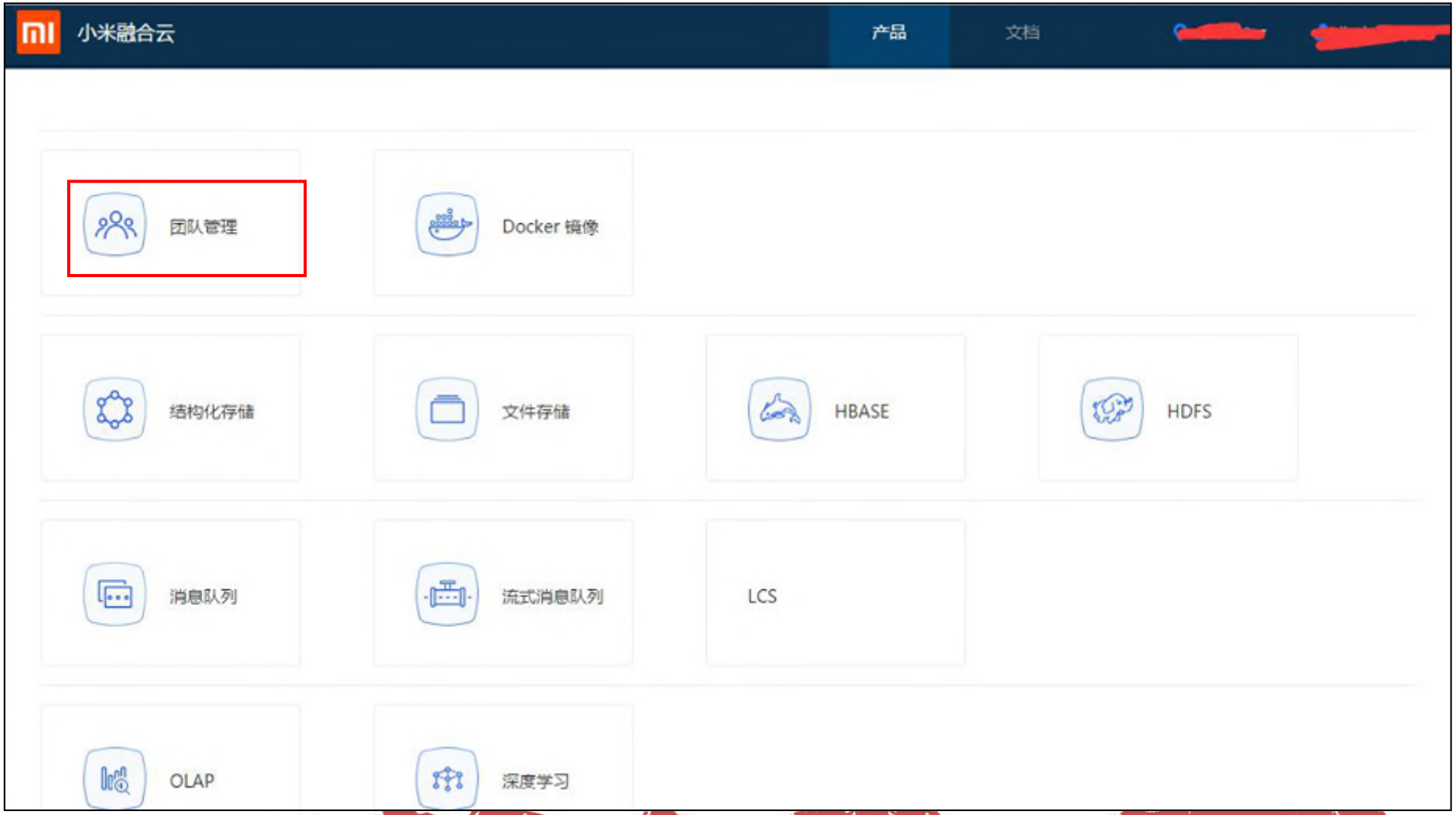
- 数据冷备



大纲

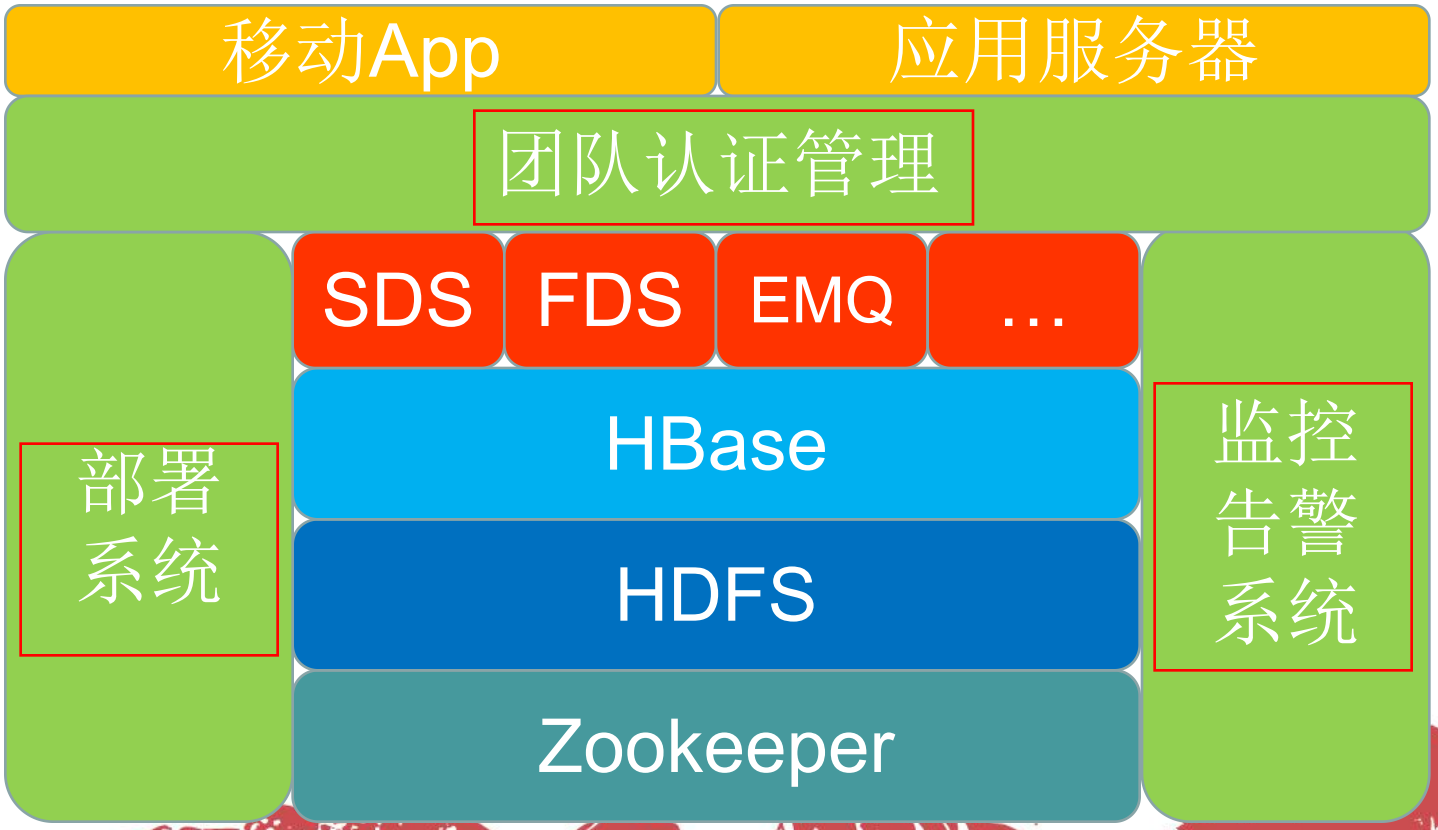
- 小米结构化数据存储
- 融合云平台

融合云


小米融合云
产品
文档
团队管理
Docker 镜像
结构化存储
文件存储
HBASE
HDFS
消息队列
流式消息队列
LCS
OLAP
深度学习

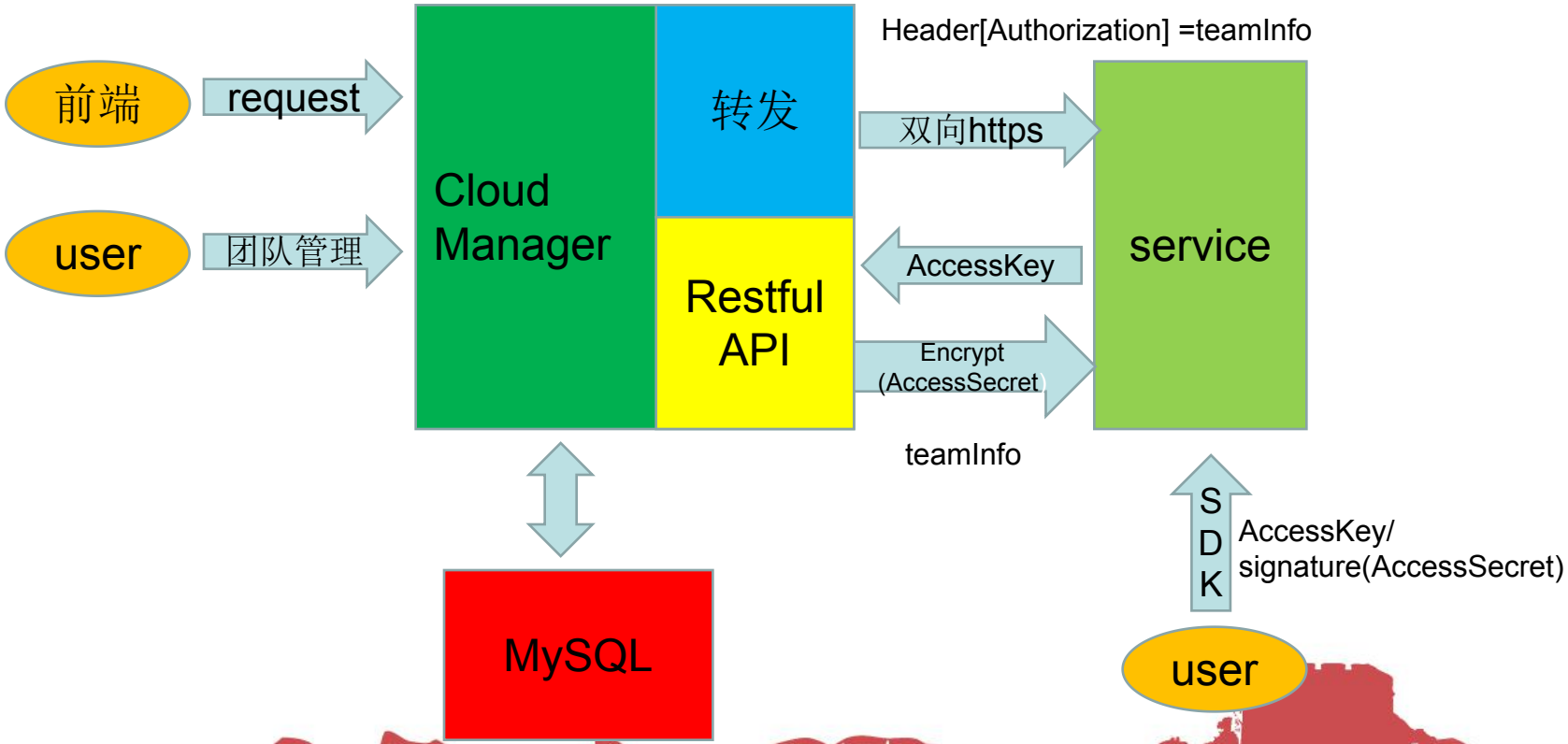
融合云

- 架构


移动App
应用服务器
团队认证管理
SDS
FDS
EMQ
…
HBase
HDFS
Zookeeper
部署系统
监控告警系统

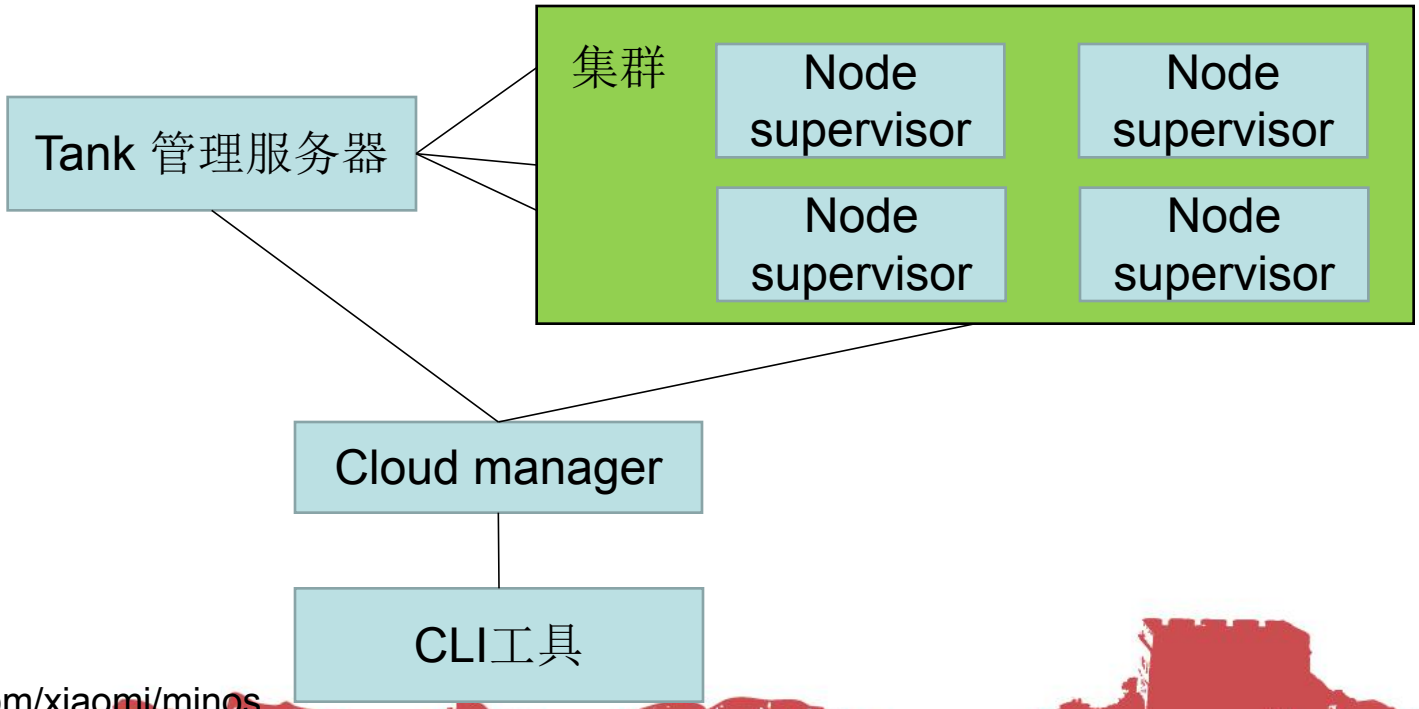
融合云

- 团队认证管理——CloudManager



融合云

- 部署系统——Minos[6] 2.0
  - 增加了认证授权模块



[6]https://github.com/xiaomi/minos

融合云

- 监控告警系统

公网访问

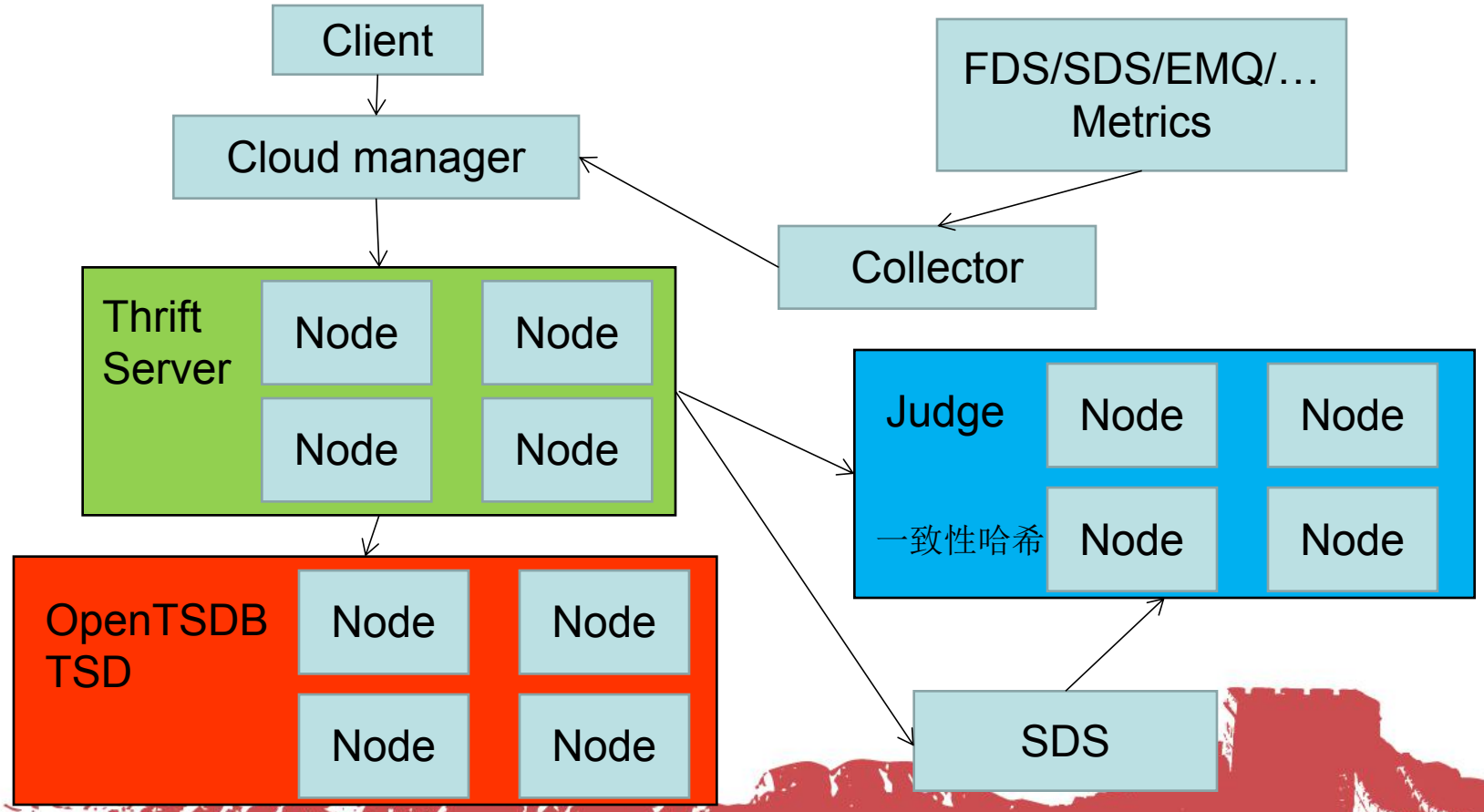
权限控制

保留原始数据，支持下采样和聚合

为用户监控其它服务的资源指标提供统一入口

融合云

- 监控告警系统——架构



THANK YOU