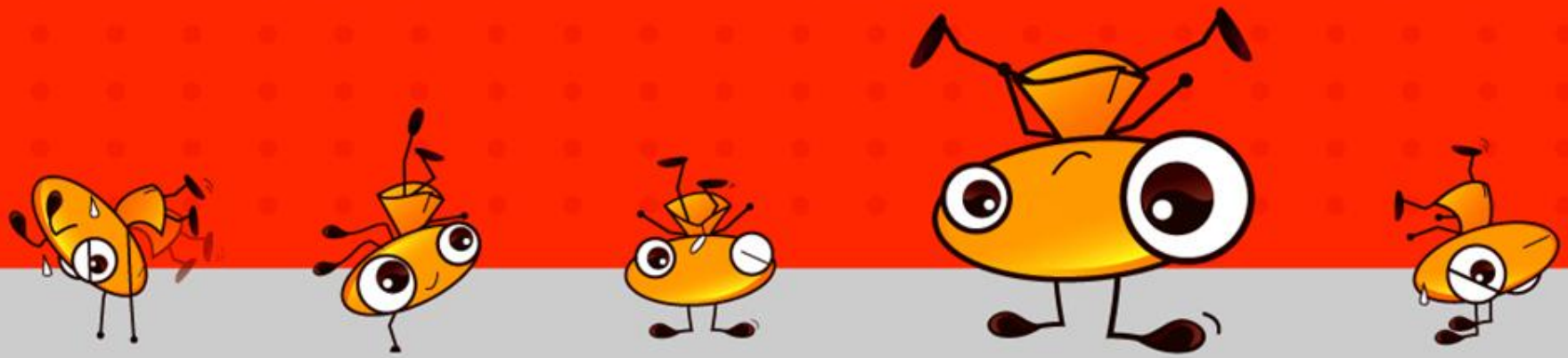
# 淘宝Hbase应用和改善

核心系统部 张毅
wuting@taobao.com



淘宝网
Taobao.com

# 目录

- Hbase 在淘宝
- Hbase 应用场景举例
- 问题和措施
- 今后的工作

淘宝网
Taobao.com

# 什么是Hbase

- 分布式NOSQL存储系统
- 底层数据存储基于HDFS
- 高可用，高性能
- 列存储，多版本
- 百亿行×百万列×上万个版本

淘宝网
Taobao.com

# 淘宝选择HBASE的原因

- 和Hadoop一样的海量数据处理的能力
- 易于横向扩展
- 随机读写的高性能
- 高可靠性和稳定性
- 在互联网公司有较多的适用场景
- 单行写入的强一致性
- 开源，社区活跃并有大公司支持

# Hbase 在淘宝的规模

- 于2011年上半年开始使用
- 版本基于HBase 0.90.3+Patch
- 10个集群，300台Region Server
  - 16 core,24G/48G, SATA 1T * 12/ SAS 300G * 12
- 200k ops/sec, 70% write, 30% read
- 应用于多个Online和Offline系统
- 百亿行Table规模

# Region Server读写性能

- 读性能
  - 表现稳定，hdfs为主要瓶颈
  - 响应时间6.63ms(TPS 7546)
- 写性能
  - 纯读过程中表现稳定，
  - 命中率非常重要（TPS: 2K-40K）

淘宝网 Taobao.com

# 读写影响因素

- Region Server规模
- 缓存的大小和命中率
- HDFS规模和读写效率
- 安全级别的配置
- Region Split 和Compact设置
- Row Key的设置

淘宝网 Taobao.com

# 应用举例

- 淘宝实时传输平台
  - 数据每天TB级的数据写入应用
  - 旧的存储模型（内存+硬盘 ）
  - 发布和订阅的使用场景
- 淘宝指数
  - 倒排索引的属性查询（Redis ->Hbase）
  - 实时/性能
  - 客户端Join
  - 高冗余，每行百兆级的数据应用

- 交易历史记录查询系统
  - 百亿行数据表，千亿级二级索引表
  - 每天千万行更新
  - 查询场景简单，检索条件较少
  - 关系型数据库所带来的问题
  - 基于userId + time + id rowkey设计
  - 成本考虑

# 数据量增大遇到的问题

- 随着region数目的增多写性能下降
  - 7W region , 20 台服务器，写TPS降至1500以下
  - Region Server 遍历onlineregions所致 HBASE-3694
- Region Server OOM
  - 行的版本过多
  - RowKey 设计错误
  - HBASE-3290
- Master OOM
  - N个region，M台server，则Hmaster中AssignmentManager最多会包含N*N/M个regionLoad对象 HBASE-3906

淘宝网 Taobao.com

# 数据丢失及读写异常

- Splits 可能会造成.META.表存在临时holes
  - HBASE-4335
- Master Split Hlog失败导致数据丢失
  - Master处理Hlog和写入恢复日志的动作是并行所致
- 频繁重启集群会导致root和meta被不同服务器占有
  - HBASE-3914
- Region 关闭忽视了IOE导致Memstore数据丢失
  - HBASE-4270
- Split失败导致数据丢失
  - split在offlineParentInMeta失败 ,但Meta所在的regionserver成功
  - HBASE-4562和HBASE-4563

淘宝网 Taobao.com

# 性能

- Split Region带来的影响
- HDFS 客户端的改进
  - 短连接的影响
  - 本地化读Block
- 压缩算法的比较
- 大量的读取导致写性能下降
- 批量写入
- RowKey的合理设计
  - 分散写，连续读

淘宝网 Taobao.com

# 功能

- 目前版本不支持二级索引
- 不支持多行的事务
- 用户权限
- 备份
- HDFS NameNode 持久化存储
- Hbase 集群管理

淘宝网 Taobao.com

# 监控和运维

- 使用Ganglia 对集群状况监控
- 增加对Hbase内部状态的监控
  - Region 详细状况
  - 读写状态和时延
  - Cache状态
  - 数据文件监控
  - Split 和Compact状态
  - JVM
- 网页管理工具针对HBASE的操作

淘宝网 Taobao.com

# 总结

- 瞬间写入量很大，数据库不好支撑或需要很高成本支撑的场景
- 数据需要长久保存，且量会持续增长到比较大的场景
- Hbase 不适用于有Join，多级索引，表关系复杂的数据模型
- 合理设计 RowKey，非常重要
- 数据最好是可恢复的
- 生产环境关闭split，region数不要太多

淘宝网
Taobao.com

# 后期的工作

- 数据的稳定性,数据零丢失
- 性能提升
- 备份/回复
- 多数据中心

淘宝网
Taobao.com

# THANK YOU