360 Cassandra 实践分享

唐会军

2012年9月14日

SACC2012

提纲

- 现状
- 使用&改进
- 总结

现状

- Cassandra在360
  - 总服务器规模 超过1500台
  - 最大单个集群150台
  - 版本：基于0.7.3改进

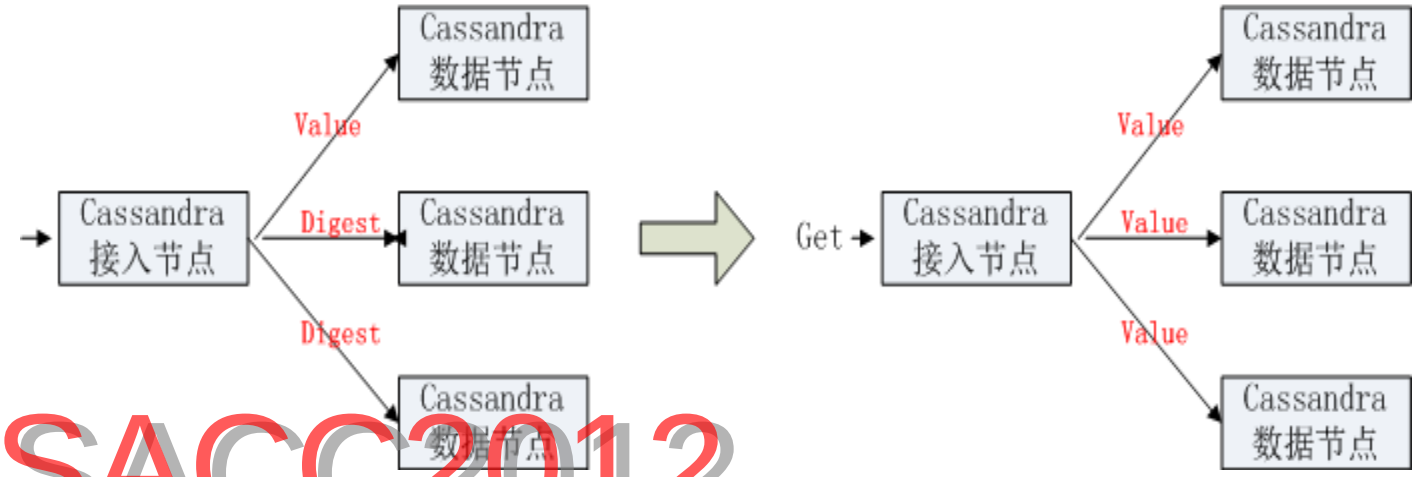
SACC2012

现状

- Why Cassandra?
  - 团队人员少，需求紧， 选择开源项目
  - 无单点，无中心，适合在线业务
  - 代码易懂，团队成员有代码基础
  - 社区比较活跃

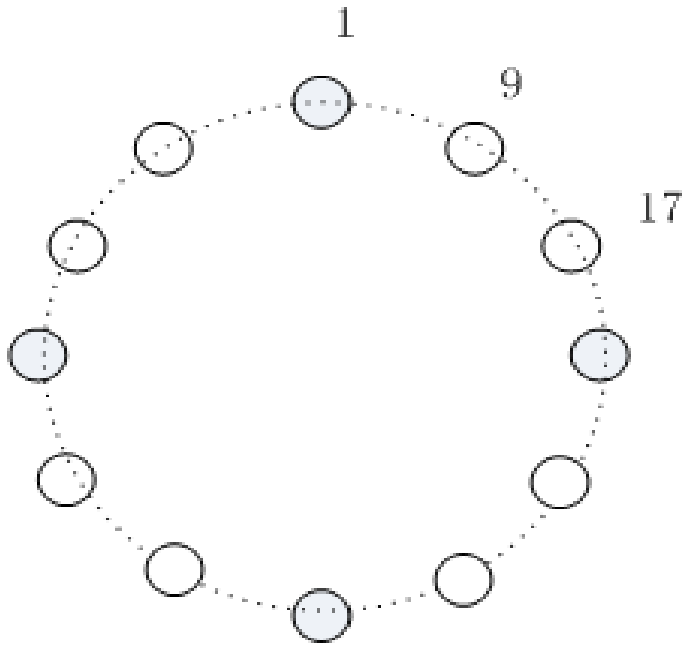
使用&改进

- 读&写
  - 数据 3 份拷贝，一致性级别为 quorum
  - 支持读操作发3个value请求
  - 增加一系列 digest接口（get_digest……）
  - 100% 开启读修复



使用&改进

- Compaction
  - 关闭自动触发，脚本定期触发
  - 数据量大的keyspace每天一次
  - 数据量小且重复key较多每两小时一次
  - 增加指定参与Compaction的SStable大小
  - Skip异常record

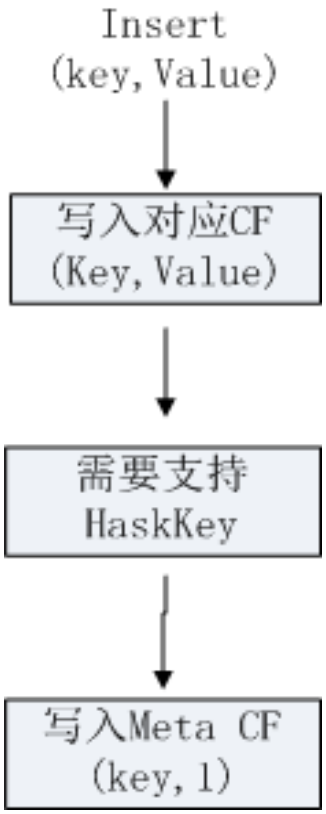


使用&改进

- 扩容(Bootstrap)
  - 增加限速功能（参考高版本）
  - 修改选取算法，保证从三个节点拖数据
  - 拖取数据时跳过有问题无法读取的Sstable
  - 重建Index时跳过异常record
- 顶替节点(Replace token)
  - Replace过程中正常接收写请求

使用&改进

- Haskey
  - 高效判断某个key是否存在



使用&改进

- 数据备份检查&修复
  - 接入节点记录写成功不足3的KEY，后台线程修复
  - 增加Hinted off保存key的时间
  - 后台线程定期全局扫描检查并修复

使用&改进

- Commitlog
  - 降低写操作较少的Keyspace刷memtable时间
  - 节点stop前手动flush memtable

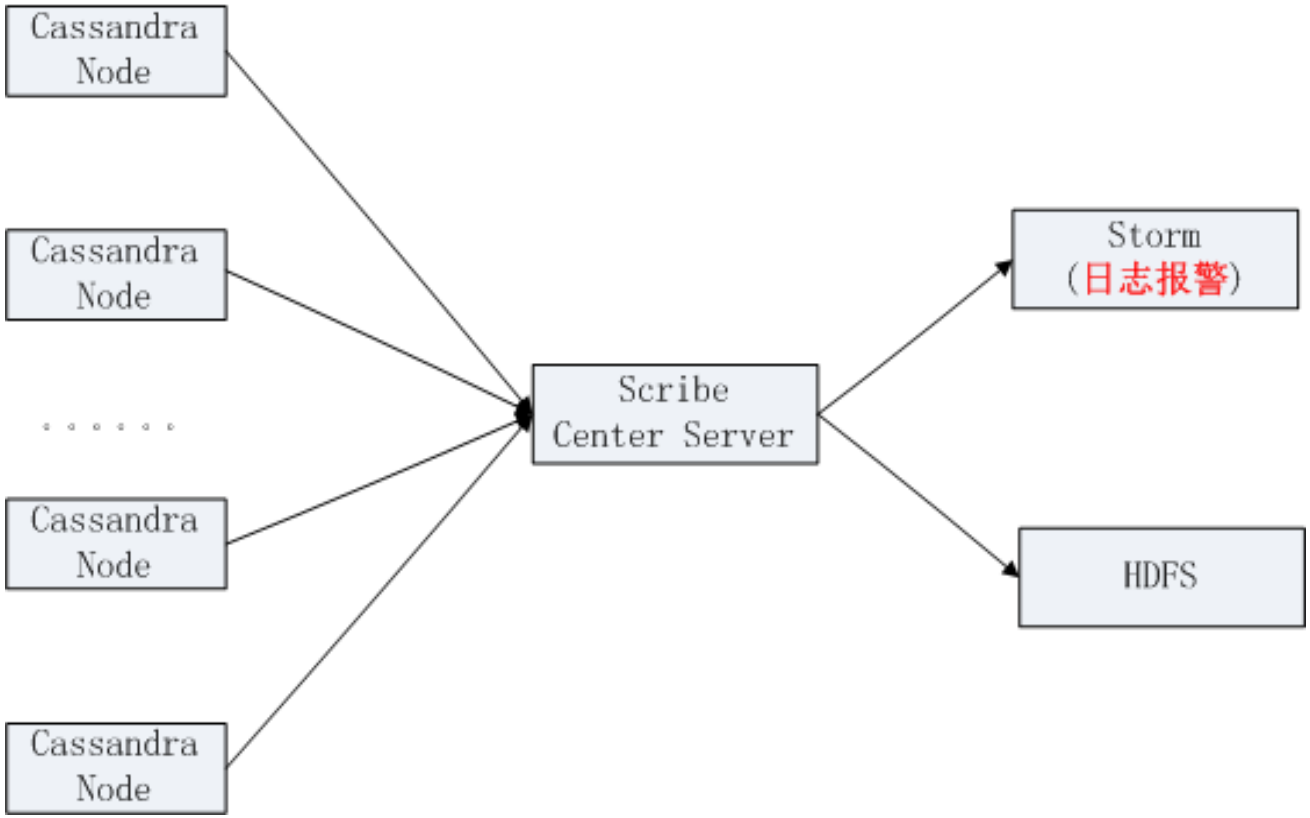
使用&改进

- Java GC
  - Full GC 使用CMS
  - 增加MaxDirectIO内存限制

SACC2012

使用&改进

- 日志级监控


Cassandra Node
Cassandra Node
Cassandra Node
Cassandra Node
Scribe Center Server
Storm (日志报警)
HDFS

总结

- Cassandra缺点
  - 不支持多版本
  - 扩容麻烦
  - 不清楚数据的备份情况

SACC2012

总结

- Cassandra VS Hbase
  - Cassandra无中心，服务可靠性强，但扩展和管理相对复杂，目前在对可靠性要求高的在线存储业务上使用
  - HBase有中心，服务无冗余，但扩展和管理容易，目前在离线存储业务上使用

总结

- 心得
  - 重视运维
    - （监控，流程，规范，预案）
  - 熟悉原理和代码
    - 规模大后，产生各个各样的问题，定位和解决需要了解大量的原理和代码
  - 重视问题
    - 不要轻易放弃任何问题，刨根问底

SACC2012

Thanks!