DTCC2013

小米hadoop&hbase微实践

谢良

个人简介

- 小米科技 软件工程师
- 目前在存储组做HBase研发工作
- 之前也负责维护过一段时间MySQL线上集群
- weibo.com/bestxieliang

Agenda

- 选型依据
- upstream重要issue
- 集群check list
- 若干案例解析
- 一些微改进点与社区回馈

Agenda

- **选型依据**
- upstream重要issue
- 集群check list
- 若干案例解析
- 一些微改进点与社区回馈

- 类似业务场景下Facebook在前面趟雷成功
- 国内阿里等的实践也增强了我们的信心

Agenda

- 选型依据
- **upstream重要issue**
- 集群check list
- 若干案例解析
- 一些微改进点与社区回馈

HDFS层面

- Support hsync　HDFS-744(2.0.2-alpha)

  类fsync语义，关系数据安全性

HDFS层面

- namenode HA

FB用的AvatarNode方案，我们没跟进，因为：

HA framework	HDFS-1623(2.0.0-alpha)
Auto failover for HA	HDFS-3042(2.0.2-alpha)
QJM	HDFS-3077(2.0.3-alpha)

HDFS层面

- Shortcut a local client reads

  HDFS-2246(0.23.1, 1.0.0, 0.22.1)

  HDFS-347(截止目前尚未merge进trunk)

- parallel write and Hflush/sync

  HDFS-895(0.20-append, 0.20.205.0, 0.22.0)

HDFS层面

- concurrent readers and writer

  HDFS-1907(0.23.0)

- better handling of volume failure in datanode storage 坏少于配置上限数量的磁盘后，不需要关闭datanode实例，这样上层hbase不会丢失本地性

  HDFS-457(0.20.203.0, 0.21.0)

HBase层面

- 同行写入的原子性 HBASE-2856 (0.94.0)
- online schema update　HBASE-1730/4213
- distributed log splitting　HBASE-1364(0.92.0)
- 业务低谷期做激进压缩　HBASE-4463(0.94.0)

HBase层面

- timerange hints HBASE-5010(0.94.0)
- lazy seek　HBASE-4465(0.94.0)
- HFile v2　HBASE-3857(0.92.0)
- data block encoding
  HBASE-4218
  HBASE-4676

Agenda

- 选型依据
- upstream重要issue
- **集群check list**
- 若干案例解析
- 一些微改进点与社区回馈

- 硬件主要关注磁盘和网卡、控制节点的RAID
- OS：
  2.6.32
  ulimit
  ext4 微调mount参数,noatime, tune2fs -m 1
  ntp服务
  THP

- JVM：

  使用较新的版本 1.6.0_37

  调整VM选项参数

- zookeeper：
  3.4.4+的版本
  注意事务日志落地盘
  autopurge.snapRetainCount
  autopurge.purgeInterval

Agenda

- 选型依据
- upstream重要issue
- 集群check list
- **若干案例解析**
- 一些微改进点与社区回馈

- 现象：测试集群节点随机OOM
- 日志：OutOfMemoryError: unable to create new native thread
- 取thread dump
- Centos6上ulimit设置和5不一样，常见坑！

- 心得：基础软件checklist要做完善！

- 现象：测试集群Region Server偶发挂掉
- 日志显示有30多秒的长暂停，大于集群配置的zk检测超时30s；通过GC日志显示app确实被stop了30多秒，但之前木有打印堆相关信息(我们配置了参数会打印)，结合PrintSafepointStatistics的输出，定位可能与VM偏特锁相关，尝试禁掉，再未发生
- 心得：对VM要有敬畏之心

- 现象：某台RS的RPC队列堵塞
- jstack确认是锁相关，很快可以对应到代码
  zkw.saslLatch.await()
  通过查社区svn提交历史，找到相应change
  backport到内部代码库，发布，解决

- 这坨代码之前就是workaround，被坑了...

- 现象：RS OOM: Direct buffer memory
- 经过社区讨论，在启用本地读特性后，如果hfile句柄多了，每个有1MB的本地读相关联的buffer pool引用，容易超限

  可以参考HBase-8143讨论

- 心得：我们之前其实就遇到了但分析的不够，功力待提升：）

Agenda

- 选型依据
- upstream重要issue
- 集群check list
- 若干案例解析
- **一些微改进点与社区回馈**

- 结合metric机制，开发自己的监控、报警和dashboard系统：汇聚、突出重点
- 线上enable Kerberos认证，为此hack了zk/hadoop/hbase若干代码，坑无数。。。
- 接入业务时 修改scribe来满足需求

- JVM："*Use CLOCK_MONOTONIC_RAW for nanoTime if available on Linux*" 该patch用于更合理的处理时间跳变问题
- HADOOP:

  libhdfs implementation of hsync API

  Set daemon for HttpServer's QueuedThreadPool

  Add ability to reset topologies on master nodes

  Slow RPC can prevent metrics collection on DNs

  ...

- HBASE:

  optimize hfile index

  parallel seek in StoreScanner

  key range hints

  tweak in-memory evict algorithm

  reverse scan

  ...

- 小米在Hadoop/HBase领域是新人，成立相关专职研发团队才近半年，希望以后在开源社区会有更多更重要的贡献，更希望社区生态圈会越来越好^-^

- 欢迎各位来小米做技术交流和指导工作！

Question?