hbase在淘宝的应用与优化

邓明鉴(花名：竹庄)

2012.3

DTCC2012

OVERVIEW

- 使用hbase的动机
- 版本选择
- 应用状况
- 监控、运维及测试
- 发展

动机

- 数据量增加，需要TB/PB级别的在线服务
- 数据量增长速度快，对水平扩展能力有需求
- 简单kv读写，响应延迟低
- 要求强一致性
- 写入很频繁，吞吐量大
- 批量读取数据需求
- schema多变
- 良好java接口

hbase满足

- 水平线性扩展
- 强一致读写
- 原子和可配置地切分表
- 自动容错
- 可满足实时读写响应，吞吐量大
- 容易使用的java api
- 方便地运行mapreduce
- 可扩展的thrift/rest/avro接口

hbase不满足

- 暂时只有主索引
- 服务单点(区别于数据单点)
- namenode单点
- 难以实现真正的事务

版本选择

- 2007.4 first version → by Mike Cafarella
- 2008.3 0.1.0 → initial release,moving from hadoop-contrib
- 2009.9 0.20.0 → based on hadoop-0.20, add zookeeper, target for online
- 2010.10 0.89fb → production version, add master
- 2011.4 0.90.2 → taobao-hbase-0.90 based on
- 2012.1 0.92.0 → add coprocessor,hfileV2,security, taobao-hbase-0.92 based on 0.92.1
- 2012.3 0.94.0 → performance release
- ??? 0.96.0 → add PB

应用状况

- 2011.3开始研究hbase的在线服务
- 2011.5上线第一个应用
- 目前已上线20+在线应用，集群约15个，服务器 > 300
- tps高峰期约120k+/s(不算notify应用)，已使用容量超过300TB
- 还在持续增加，预计2012年会翻倍
- 离线应用另有约300台服务器
- 在线应用稳定性有所保障，达到生产标准，最近8个月无数据丢失

应用状况

- cube
  - 导入数据？ → 离线导入，在线读取
  - 需要有sum/group等接口？ → coprocessor
  - 非java开发人员？ → rest
  - 无二级索引，性能跟不上？ → 索引用冗余解决
  - 数据完整性？ → 数据对比，性能对比

应用状况

- timetunel
  - 在线读写
  - 读刚写入的数据
  - 数据量大，每天20TB? → seta盘机器，hdfs层够大
  - split很频繁，且导致数据丢失? → fix掉很多split的bug,0.92以上版本基本杜绝了该类情况
  - 负载不均衡？ → rowkey设计以及pre-sharding很重要
  - 业务以天为单位？ → 每天建一张表
  - 应用端不能长时间挂起？→ 合理的rpc超时时间以及重试次数
  - region数接近10万(不合理) → 增大单个region的大小

应用状况

- 实时计算与实时推荐
  - hbase做存储层
  - 上层可以是storm/akka等
  - 有数据热点，读性能好
  - 一次读取很多数据？→ rowkey前缀相同，批量读取数据
  - 负载不均衡？ → 表级别balance
  - 减少应用中断时间？ → graceful_stop

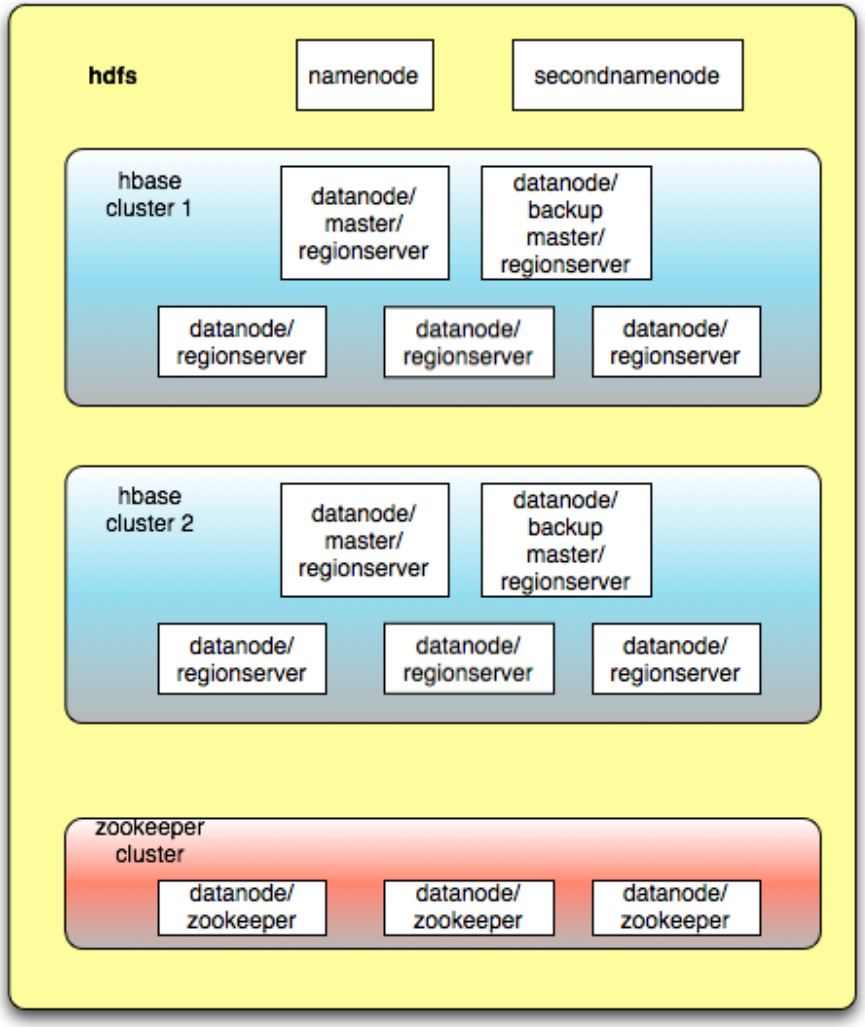
应用状况

- 历史交易订单
  - 接近核心应用，运维要求级别高
  - 晚上导入数据，白天关闭split及compact来提高读性能
  - 分页查询
  - 批量读取
  - >2千亿的索引记录
  - 0故障率

应用状况

- notify日志
  - 尝试性项目，对稳定性要求不太高，对性能要求极高
  - hbase 0.94+cdh3u3+zookeeper 3.4.3
  - 每天250亿带hlog写入(每秒超过50w)
  - 前缀压缩(HBASE-4218, 适用于k>>v的场景)

监控、运维及测试


hdfs
namenode
secondnamenode
hbase cluster 1
datanode/ master/ regionserver
datanode/ backup master/ regionserver
datanode/ regionserver
datanode/ regionserver
datanode/ regionserver
hbase cluster 2
datanode/ master/ regionserver
datanode/ backup master/ regionserver
datanode/ regionserver
datanode/ regionserver
datanode/ regionserver
zookeeper cluster
datanode/ zookeeper
datanode/ zookeeper
datanode/ zookeeper

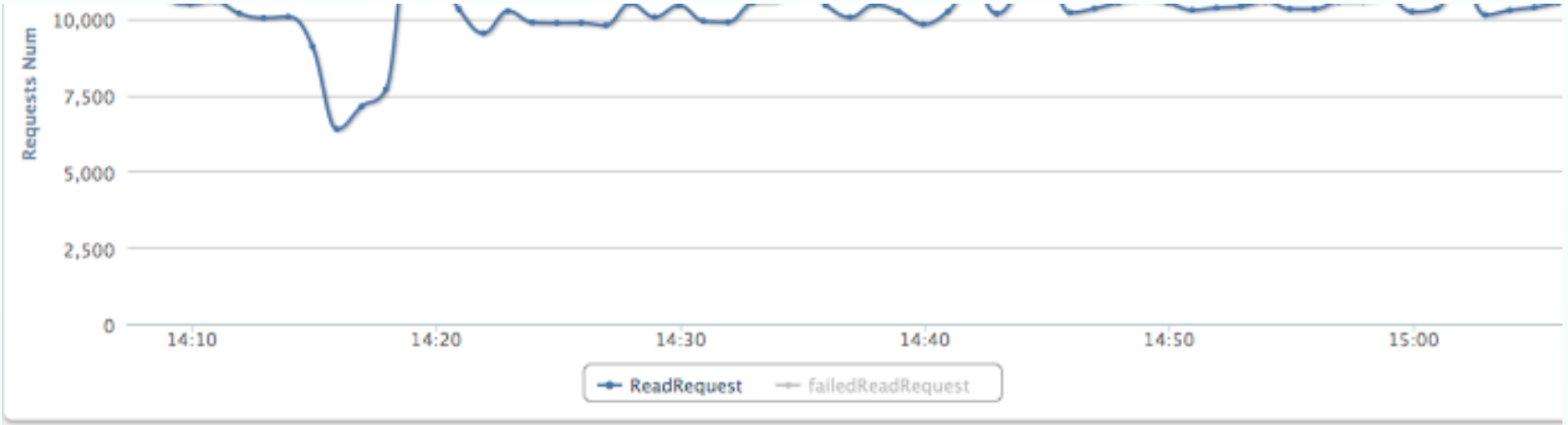
监控、运维及测试

- 原因：
  - zookeeper共享是为了运维方便
  - hbase集群尽量独立是为了应用不相互干扰，另一个原因是可以最大程度发挥伸缩性
  - hdfs合并是为了尽量减小compact影响，另外可以平摊存储成本
  - hmaster有backupmaster
  - namenode目前为单点，但元数据会写两份，另外在secondary namenode上有一份冷数据

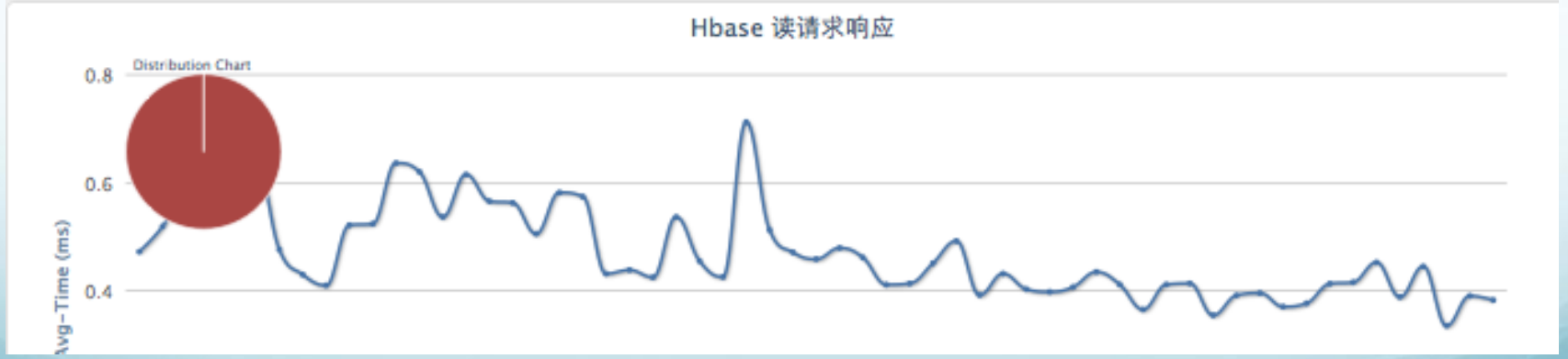
监控、运维及测试

- 监控：
  - hackcode，发送自定义的metrics
  - 通过ganglia收集到rrd中，编写监控程序读取rrd数据并持久化，展现相应的数据
  - 读写请求数、平均响应时间、最大响应时间、各个接口调用时间、写log时间、compact/splie/balance情况、hlog队列等
  - 监控粒度到达region级别
  - 考虑监控数据写入hbase

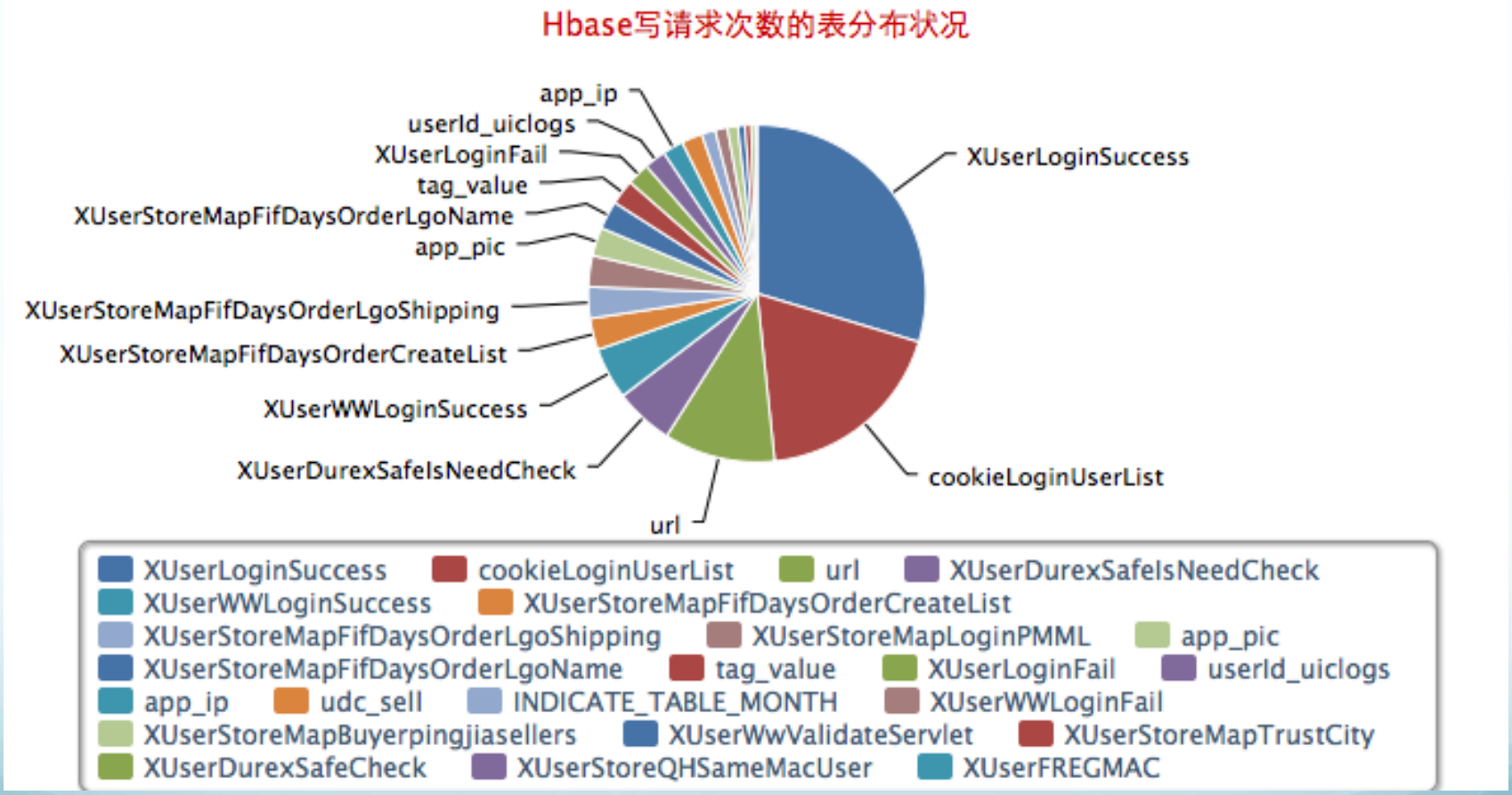
监控、运维及测试



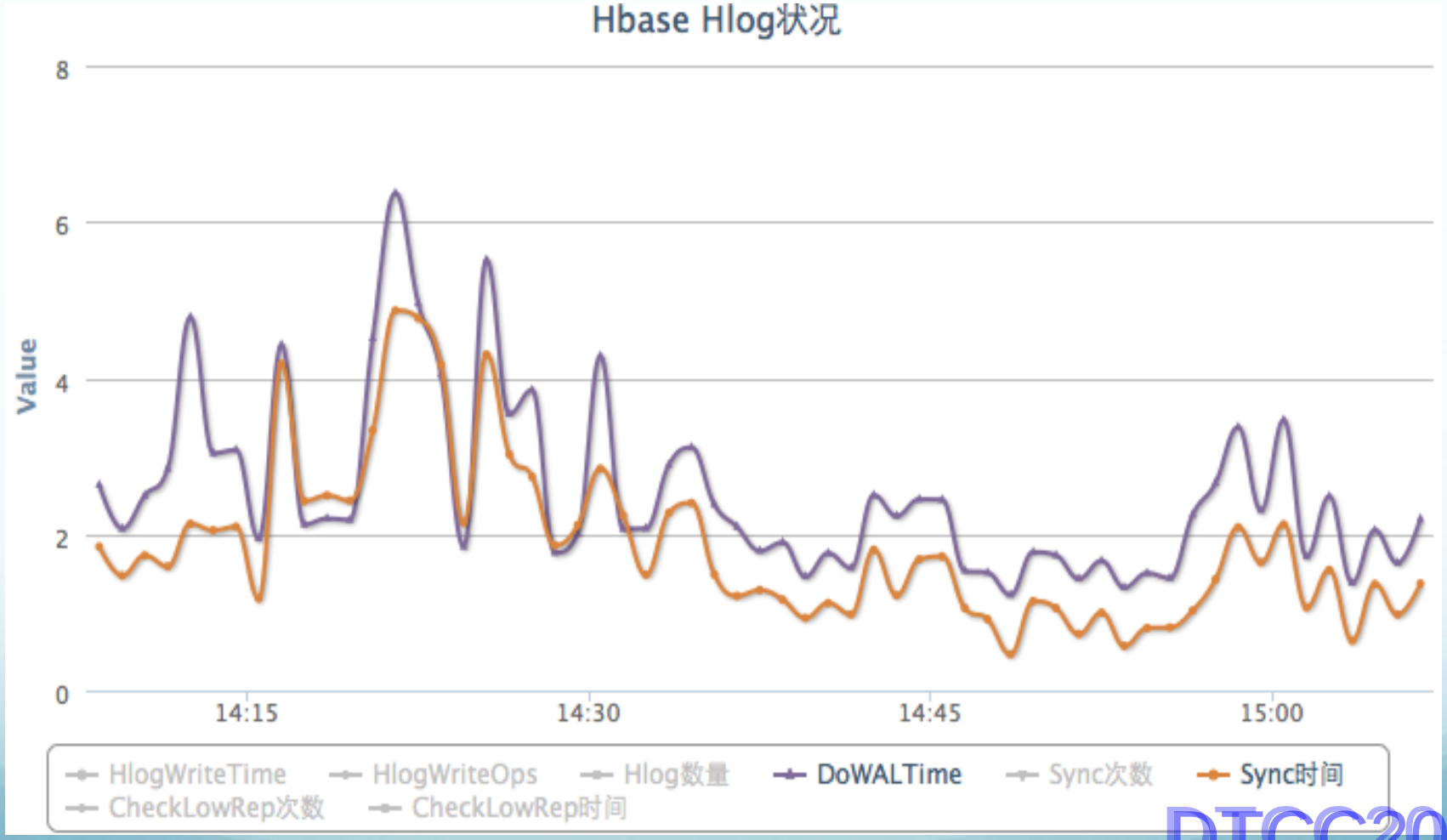
Server端读请求响应时间 平均： 0.47ms 最大： 0.71ms 最小： 0.33ms 查看详情



监控、运维及测试


Hbase写请求次数的表分布状况
app_ip
userId_uiclogs
XUserLoginFail
tag_value
XUserStoreMapFifDaysOrderLgoName
app_pic
XUserStoreMapFifDaysOrderLgoShipping
XUserStoreMapFifDaysOrderCreateList
XUserWWLoginSuccess
XUserDurexSafeIsNeedCheck
url
XUserLoginSuccess
cookieLoginUserList
XUserLoginSuccess
cookieLoginUserList
url
XUserDurexSafeIsNeedCheck
XUserWWLoginSuccess
XUserStoreMapFifDaysOrderCreateList
XUserStoreMapFifDaysOrderLgoShipping
XUserStoreMapLoginPMML
app_pic
XUserStoreMapFifDaysOrderLgoName
tag_value
XUserLoginFail
userId_uiclogs
app_ip
udc_sell
INDICATE_TABLE_MONTH
XUserWWLoginFail
XUserStoreMapBuyerpingjiasellers
XUserWwValidateServlet
XUserStoreMapTrustCity
XUserDurexSafeCheck
XUserStoreQHSameMacUser
XUserFREGMAC

监控、运维及测试


Hbase Hlog状况
8
6
4
2
0
Value
14:15
14:30
14:45
15:00
HlogWriteTime
HlogWriteOps
Hlog数量
DoWALTime
Sync次数
Sync时间
CheckLowRep次数
CheckLowRep时间

监控、运维及测试

- 关于测试
  - 基准性能测试
  - 分布式client
  - 单元测试/功能测试
  - 异常测试
  - 每日回归测试
- 关于发布
  - 每2-3周release一个版本
  - 每两个月线上更新一个版本
  - 分阶段发布

监控、运维及测试

- 关于运维
  - 定期hbck
  - graceful_stop
  - 定期fsck
  - 日志实时跟踪
  - 错误隔离(连接级别、rpc级别)
  - 注意full gc/cms gc
  - 重要指标实时报警/基线报警

发展

- coprocessor
- 权限
- replication
- online ddl
- 更好支持hive
- HFileV2(V3?)
- 单机事务
- 二级索引
- 向下兼容
- bug fix

红色字体代表还在做

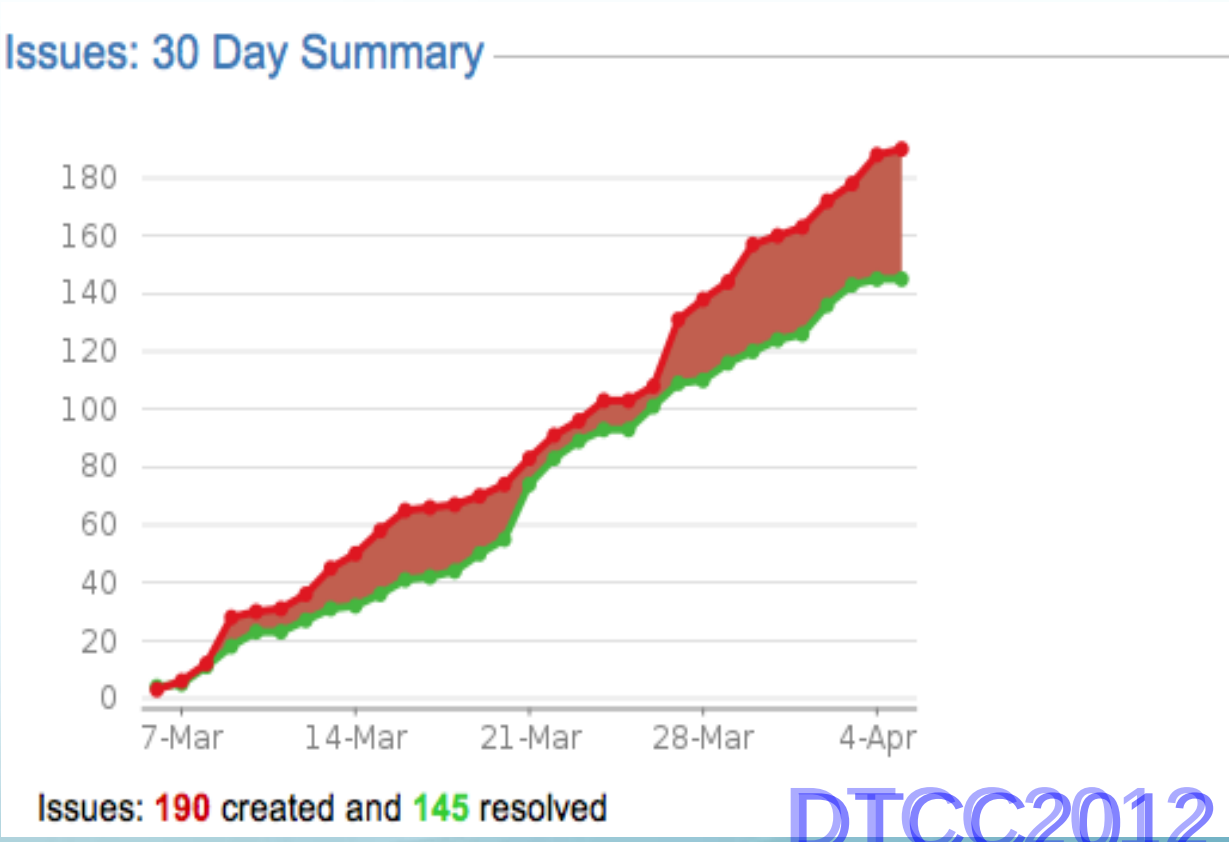
发展

- 本地化读
- 通讯协议(protobuf)
- client(nio/异步)
- 恢复速度
- 前缀压缩/hlog压缩
- gc问题
- blockcache独立

红色字体代表还在做

发展

- hdfs层面的优化
- omit(by yahoo)
- more...



Q & A

谢谢！

koven2049@gmail.com
taobao.com