HBase简介与实践分享

毕玄

技术嘉年华 2011 Taobao Open Developer Conference

About me

- 毕玄（bluedavy）
- 目前就职于淘宝网
- 07—10年负责淘宝服务框架
- 今年开始负责HBase
- 关键词
  - JVM、SOA、网络通信、高并发、高可用、海量数据存储与分析

Simple Introduce

- Yet Another NoSQL
- Bigtable implementation
- Primary Contributors: Yahoo!,Facebook,Cloudera

Basic Concepts

- Table in HBase
  - Schema: TableName & Column Family Name ;
  - value is stored in column with version as byte[];

Column Family		Column Family		
Column Label	Column Qualifier	CL	CL	CL

- Example

// Schema	name			contact	
// CL	firstname	lastname	nickname	email	phone
bixuan	hao	lin	bluedavy	**@gmail	186******
bluedavy	hao	lin			158******

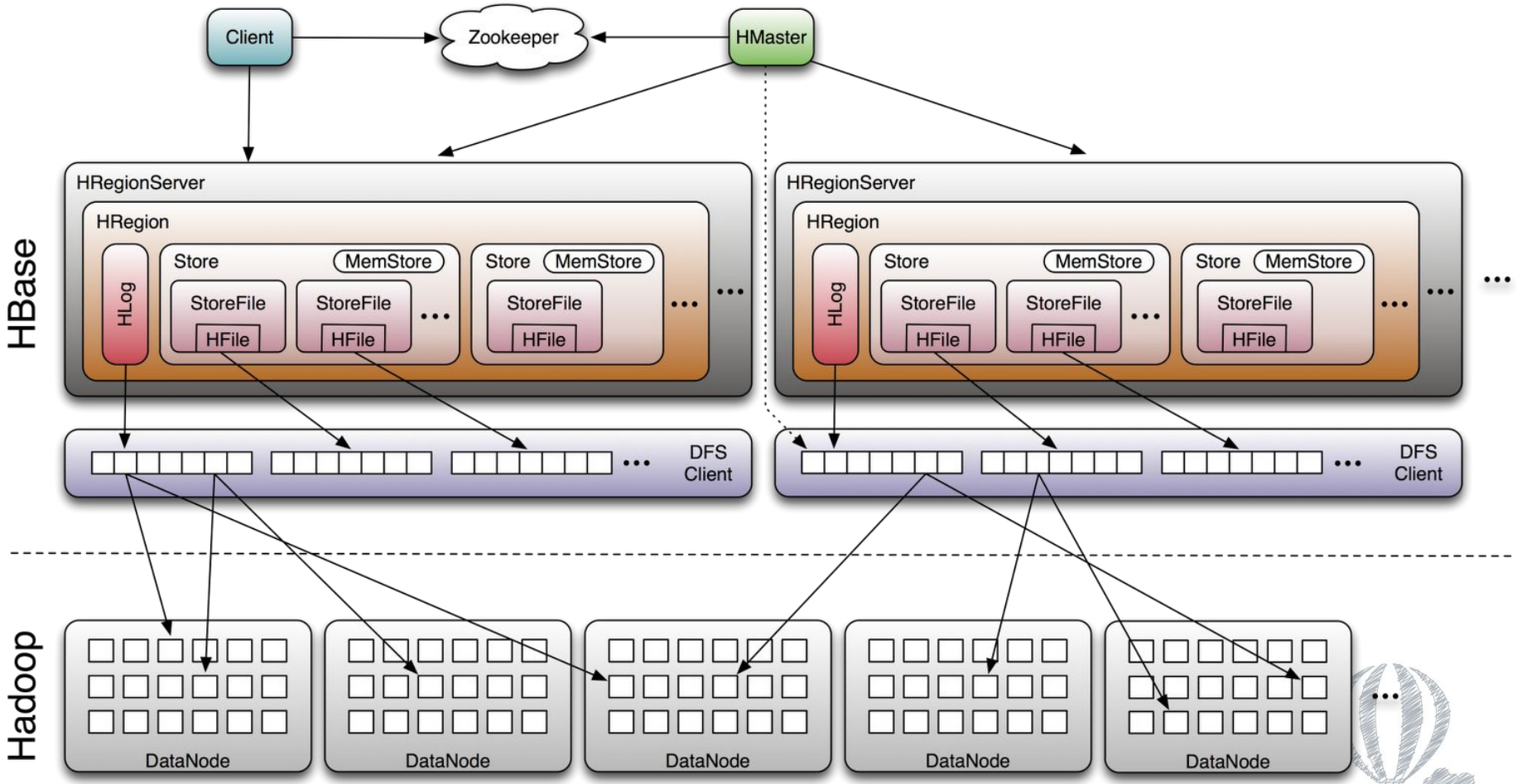
Basic Concepts

- Table in HBase
  - 以Region为单位管理region(startKey,endKey)；
  - 每个Column Family单独存储： storeFile；
  - 当某个Column Family累积的大小 > 某阈值时，自动分裂成两个Region；
    - 如何找到某行属于哪个region呢？
      - -ROOT- & .META.

Basic Concepts

- RegionServer
  - Region读写操作的场所；
- Master
  - 管理Region的分配；
  - 基于zookeeper来保证HA；

HBase Architecture



图来源于：http://goo.gl/d1T3Q

HBase Features

- 强一致性
  - 同一行数据的读写只在同一台regionserver上进行；
- 水平伸缩
  - region的自动分裂以及master的balance；
  - 只用增加datanode机器即可增加容量；
  - 只用增加regionserver机器即可增加读写吞吐量；

HBase Features(Cont.)

- 行事务
  - 同一行的列的写入是原子的；
- Column Oriented + 三维有序
  - SortedMap(RowKey,
    List(SortedMap(Column,
    List(Value,Timestamp))
)
)
  - rowKey (ASC) + columnLabel(ASC) + Version (DESC) --> value

HBase Features(Cont.)

- 支持范围查询
  - `Scan scan=new Scan(Bytes.toBytes("0"), Bytes.toBytes("20"));`
- 高性能随机写
  - WAL (Write Ahead Log)

技术嘉年华 2011 Taobao Open Developer Conference

HBase Features(Cont.)

- 和Hadoop无缝集成
  - Hadoop分析后的结果可直接写入HBase；
  - 存放在HBase的数据可直接通过Hadoop来进行分析。

Why not? just see users and performance

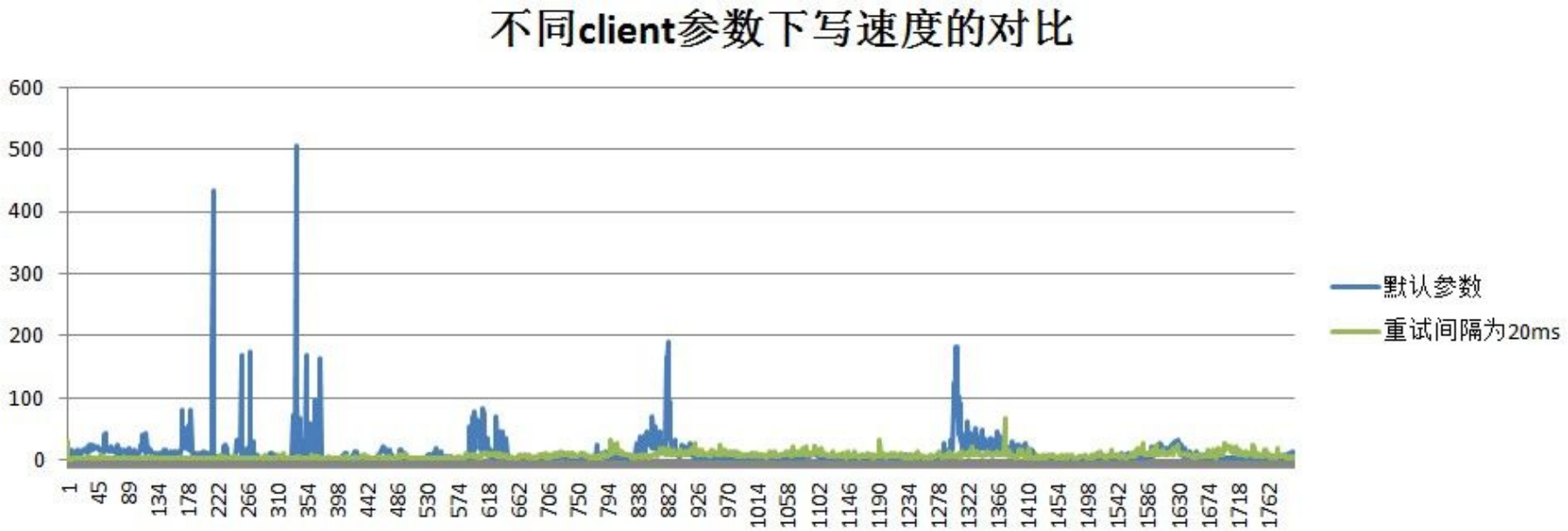
HBase能用于Online场景吗？

HBase Users

- Facebook
  - Messages
  - Realtime analytics for Big Data
- Trend Micro
- Adobe
- Twitter
- Yahoo!

HBase性能

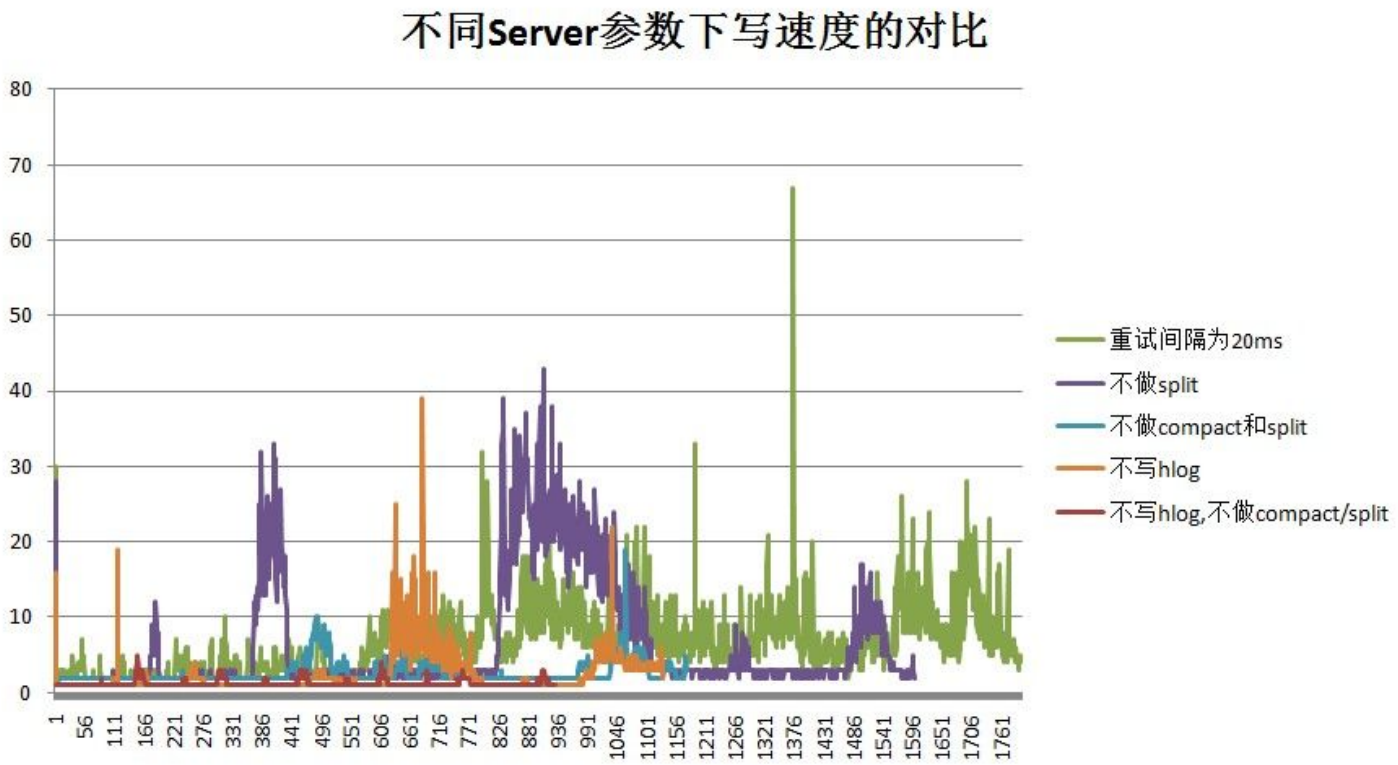
- 随机写（K: 200 byte V: 1024 byte）



region server机器：12块1T的SATA盘7200rpm，No Raid，24G，Xeon5620 2.4GHz，千兆网卡； Java Heap: 12G

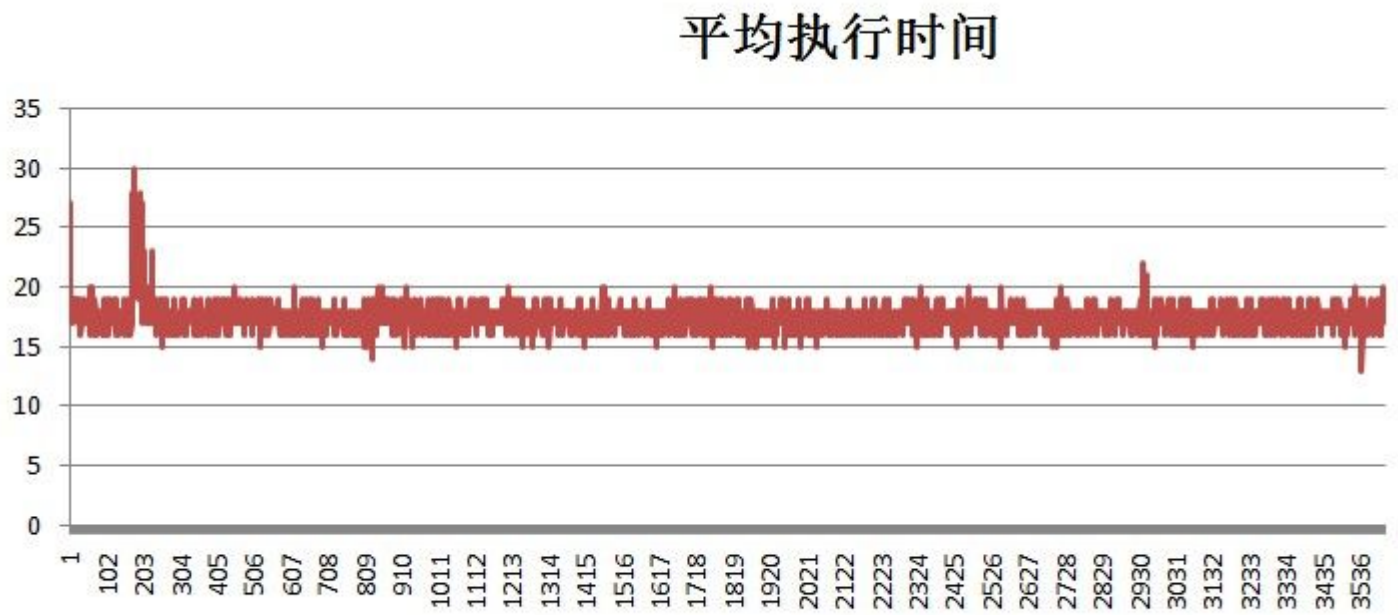
HBase性能(Cont.)

- 随机写（K: 200 byte V: 1024 byte）
  - avg: 3ms，吞吐量：1w tps/rs



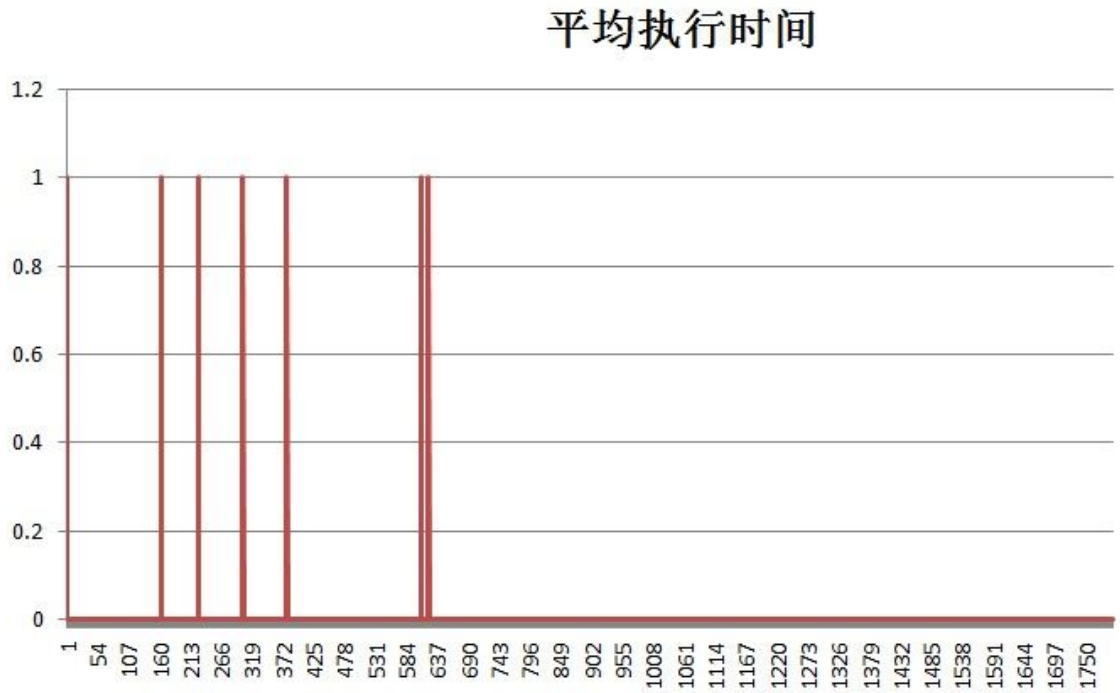
HBase性能(Cont.)

- 10亿key（1T数据量）随机读，12台rs，cache: 4g/台；



HBase性能(Cont.)

- 之前的场景，cache全命中的状况：
  - avg: <1ms



HBase不支持的

- 二级索引；
- sql/join/跨行跨表等RDBMS特性；

技术嘉年华 2011 Taobao Open Developer Conference

HBase适用场景

- storing large amounts of data (100s of TBs)
- need high write throughput
- need efficient random access (key lookups) within large data sets
- need to scale gracefully with data
- for structured and semi-structured data
- don't need full RDMS capabilities (cross row/cross table transactions, joins, etc.)

技术嘉年华 2011 Taobao Open Developer Conference

HBase @ Taobao

- 已上线的online项目
  - 4个，数据量为70T；
- 即将上线的online项目
  - 6个，到时数据量将上涨到200T+；

- More Details pls attend
  - Java版的存储和搜索介绍 [15:50 – 17:30]

我们的HBase

- 自动的测试体系
  - more powerful then YCSB!

```
* Usage:
*   BenchmarkTest
*       -tc  [表以及列的相关信息，格式为
*                例如hbasetest(120){cac
*       -tn [并发数]
*       -r  [读比率]
*       -w  [写比率]
*       -rn [数据行数]
*       -rt [测试运行时间(s)]
*       -kn [option: 读时可选择的key的数
```

1.添加测试作业

2.查看作业信息

3.配置系统数据

1.查看已经执行完毕的job情况

JobID

1	执行完成	详细情况	删除作业	bashtest3 0.25billion write random
2	执行完成	详细情况	删除作业	bashtest3 write 20 threads
3	执行完成	详细情况	删除作业	bashtest3 500w write 20 threads
5	执行完成	详细情况	删除作业	bashtest3 write 5000w 5clients 100threads
6	执行完成	详细情况	删除作业	bashtest3 write 5000w 5clients 150threads
7	执行完成	详细情况	删除作业	bashtest3 write 5000w 10clients 50threads
8	执行完成	详细情况	删除作业	bashtest3 write 5000w 10clients 100threads
9	执行完成	详细情况	删除作业	bashtest3 write 5000w 10clients 150threads
10	执行完成	详细情况	删除作业	bashtest3 write 5000w 5clients 20threads
11	执行完成	详细情况	删除作业	bashtest3 write 5000w 5clients 50threads
12	执行完成	详细情况	删除作业	bashtest3 write 5000w 5clients 20threads
13	执行完成	详细情况	删除作业	bashtest3 write 5000w 15clients 100threads
14	执行完成	详细情况	删除作业	bashtest3 write 5000w 15clients 50threads
15	执行完成	详细情况	删除作业	bashtest3 write 5000w 15clients 50threads
16	执行完成	详细情况	删除作业	bashtest3 read 5clients 20threads 10min
17	执行完成	详细情况	删除作业	bashtest3 read 5clients 20threads 10min
18	执行完成	详细情况	删除作业	bashtest3 write 5000w 15clients 50threads

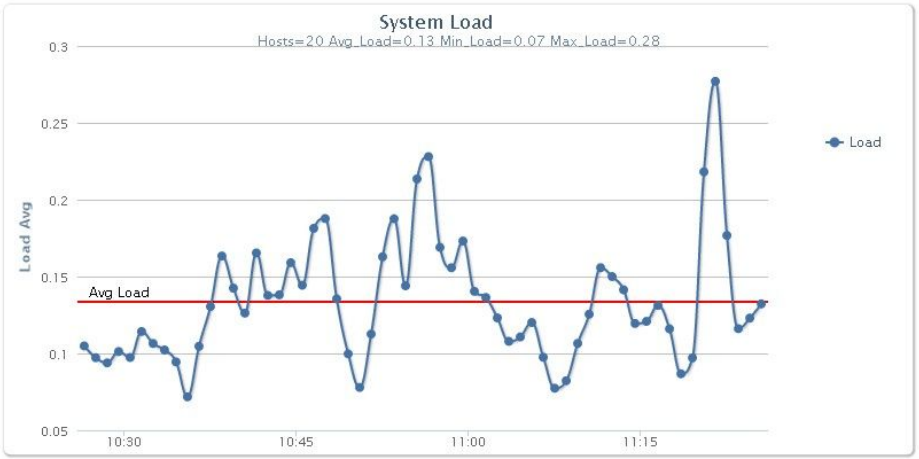
2.查看当前队列中job情况

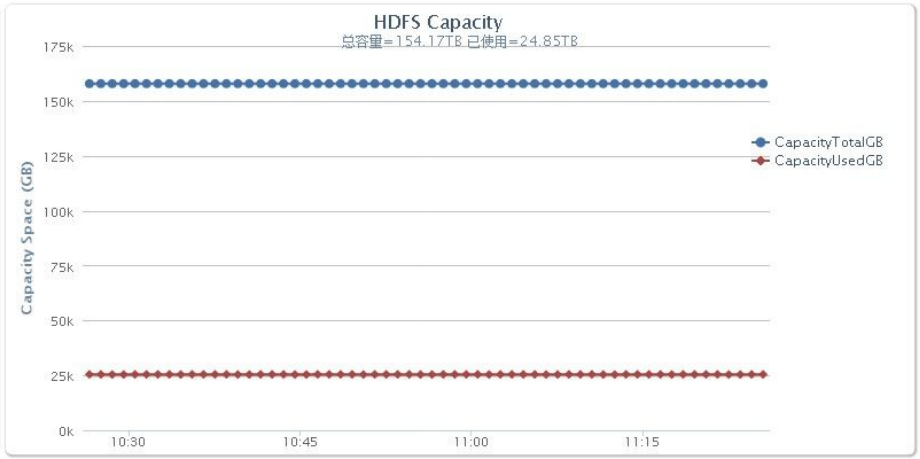
我们的HBase(Cont.)

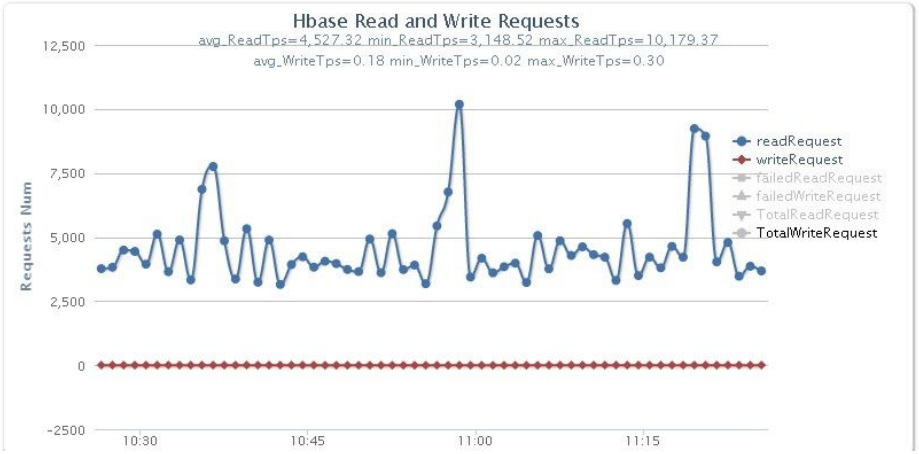
当前数据显示：1 小时前 至 0 小时前 刷新

时间选择： 天 小时前 至 天 小时前 查看

跳到详情页面

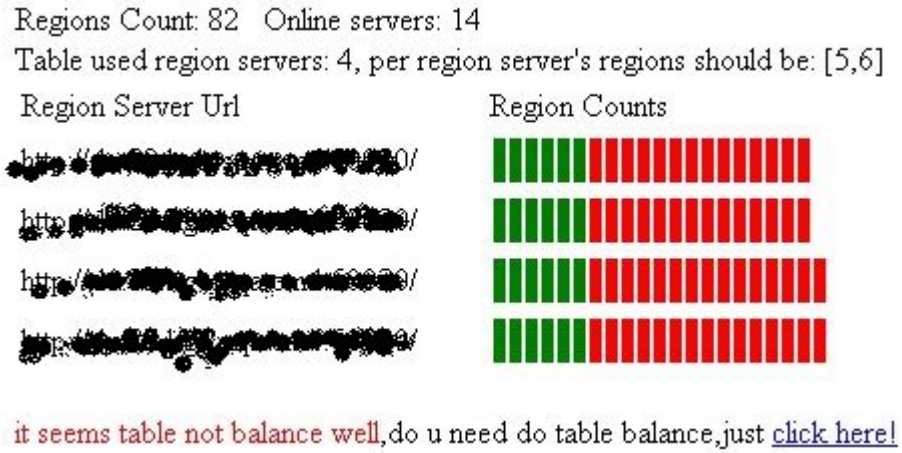






我们的HBase(Cont.)

- 改进HBase
  - Table Balance
  - Client put bug
  - Compact优化



  - Master恢复时间过长

技术嘉年华 2011 Taobao Open Developer Conference

我们的HBase(Cont.)

- 扩展HBase
  - 支持Server端计算
    - group by,avg,sum等
  - 按Group资源隔离和优先级控制

List Group

Group name	Special Configuration	Server num	Region num	Action	Change Group Configuration Or Restart Group
group: 2	true	(5)	(4618)	Delete	Change
group: 1	true	(4)	(3775)	Delete	Change
group: 0	false	(5)	(10931)	default group	
Add Group:	(must be int)	Special ?		AddNew	

RegionServer Group

Region Server name	Request num	Region num	Original Group	Update to New Group
[redacted]	0	923	2	
[redacted]	0	923	2	
[redacted]	0	924	2	
[redacted]	0	925	2	
[redacted]	0	923	2	
[redacted]	0	1188	1	
[redacted]	0	1188	1	
[redacted]	1	1181	1	
[redacted]	0	218	1	
[redacted]	0	2181	0	
[redacted]	0	2183	0	
[redacted]	0	2183	0	
[redacted]	0	2198	0	
[redacted]	10483	2186	0	
				UpdateRegionserverGroup

技术嘉年华 2011 Taobao Open Developer Conference

HBase实践经验

- 合理设计rowKey & Pre-Sharding
  - 避免仅操作集群中的少数几台机器；
  - 根据数据量、region server数合理pre-sharding。

技术嘉年华 2011 Taobao Open Developer Conference

HBase实践经验(Cont.)

- 容量影响因素
  - 开启压缩
    - lzo
    - create table ‘t1’,{NAME => ‘cf1’, COMPRESSION => ‘lzo’}

技术嘉年华 2011 Taobao Open Developer Conference

HBase实践经验(Cont.)

- 写速度关键因素
  - Table region分布均衡；
  - 单台region server的region数；
  - hbase.regionserver.handler.count
  - hbase.regionserver.global.memstore.upperLimit
  - hbase.hregion.memstore.block.multiplier
  - hbase.hstore.blockingStoreFiles
  - hbase.hregion.max.filesize

技术嘉年华 2011 Taobao Open Developer Conference

HBase实践经验(Cont.)

- 读速度关键因素
  - 单台Region Server上的Region数；
  - StoreFile数；
  - bloomfilter；
  - in-memory flag;
  - blockcache设置；
  - hfile.block.cache.size；

技术嘉年华 2011 Taobao Open Developer Conference

HBase实践经验(Cont.)

- 二级索引
  - 合理使用三维有序
    - More details pls attend 技术沙龙
  - 冗余
  - 离线


Any
Questions


The End