



# 2014 WOT全球软件技术峰会

## Software Technology Summit

2014.7.25-26 北京富力万丽酒店

## HBase 应用运维实践

许飞飞

阿里 技术保障部 数据库技术资深数据库工程师

## 大纲

- HBase简介
- HBase在阿里的应用
- HBase 运维
- HBase阿里改进

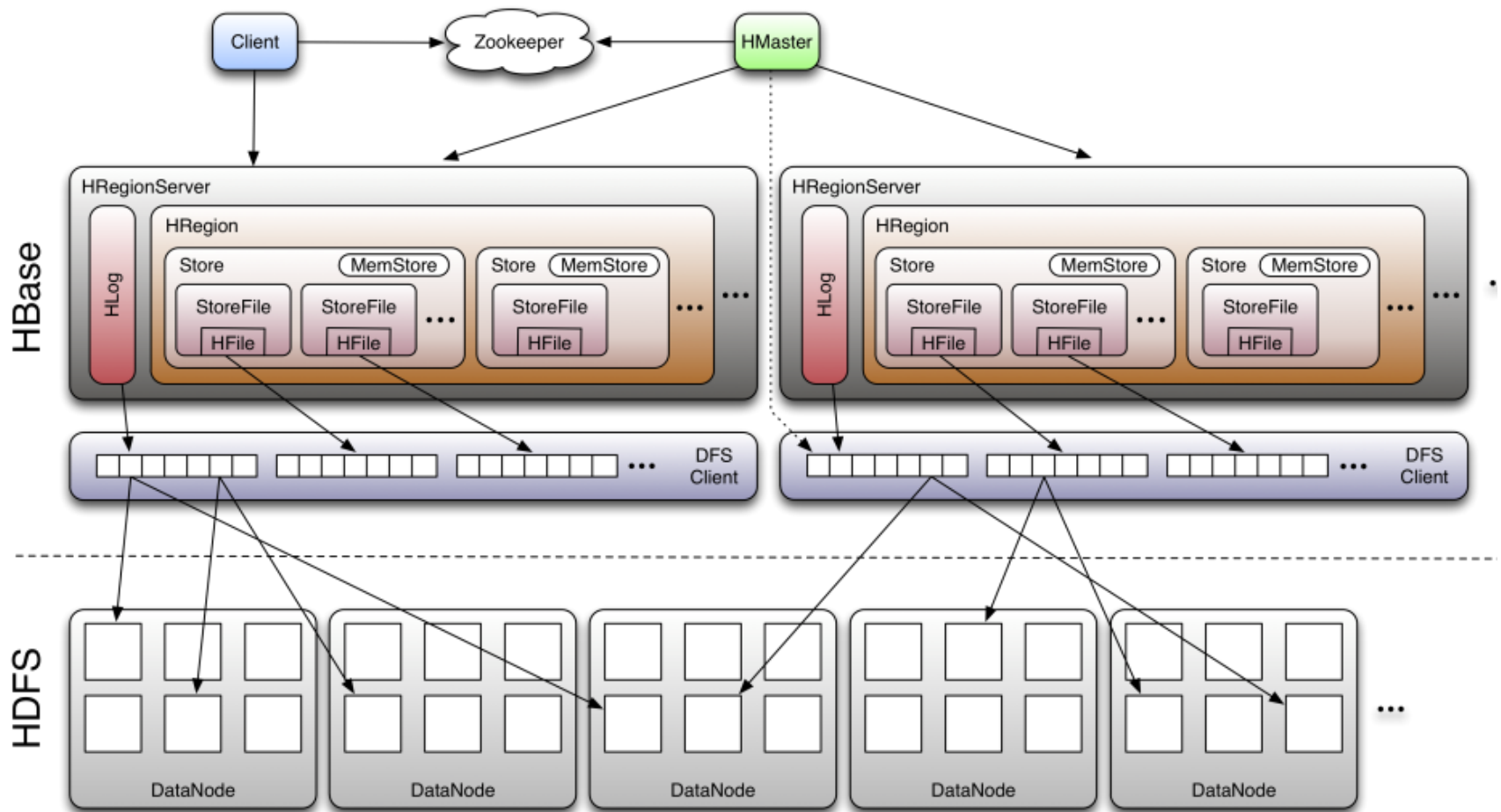
# HBase简介

- HBase 特性 架构
- HBase 数据模型和组件
- HBase 读写核心LSM Tree和HFile
- HBase的优缺点

# HBase特性

- 面向列存储的NoSQL数据库
- 水平线性扩展能力
- 强一致性 实时的读写
- 切分表的原子性配置
- 自动容错恢复(RegionServer级别)
- 方便使用Hadoop MapReduce读写
- 方便和容易使用的Java API和thrift/rest/avro接口

# HBase架构



# HBase物理结构

name	size	own
hbase	0	hbas
└─ .ROOT-	0	hbas
└─ .archive	0	hbas
└─ .corrupt	0	hbas
└─ .logs	0	hbas
└─ db135020.sqa.cm4,60020,1369742789675	0	hbas
db135020.sqa.cm4%2C60020%2C1369742789675.1369742797065	313 bytes	hbas
db135020.sqa.cm4%2C60020%2C1369742789675.1369746397210	0	hbas
└─ db135022.sqa.cm4,60020,1369742789721	0	hbas
└─ db135023.sqa.cm4,60020,1369742789765	0	hbas
└─ .META.	0	hbas
└─ .oldlogs	0	hbas
└─ ECRM_BUYER_SELLER	0	hbas
└─ .tmp	0	hbas
└─ 0ac02ed72a1005d13f87f6982a795d32	0	hbas
└─ .oldlogs	0	hbas
└─ .tmp	0	hbas
└─ BASE	0	hbas
CF Name	0	hbas
72c1de9f62424fb6946605d478b386f3	70.9 MB	hbas
recovered.edits	0	hbas
.regioninfo	756 bytes	hbas
.tableinfo.0000000001	497 bytes	hbas
└─ ECRM_MEMBER_GROUP	0	hbas
└─ ECRM_MEMBER_INFO	0	hbas
└─ ECRM_MEMBER_OFFLINE_INFO	0	hbas
└─ ECRM_SYS_SELLER	0	hbas
└─ ECRM_TRADE_INFO	0	hbas

HLog存储所在目录,按照RS实例分组(host+port+startcode),对感知和认定挂掉的RS, HMaster会分割这些HLog文件,分配给其他RS处理 HLog文件数量和大小均可配置,超过配置数量,会flush,从而使HLog不再有效,就可以滚动删除了

一张表就是一个目录(表名要符合路径规范)

RegionEncodeName(RegionName hash处理)是表下一级目录,存储一个Region的数据

BASE CF Name是下一级目录,底层就是StoreFile文件,文件名字是UUID处理的(90是long)

这个和.META.中的Region信息对应,94主要就是记住startkey(90包含tableinfo部分)

这个是表DDL存储部分(92开始从.regioninfo中独立出来)

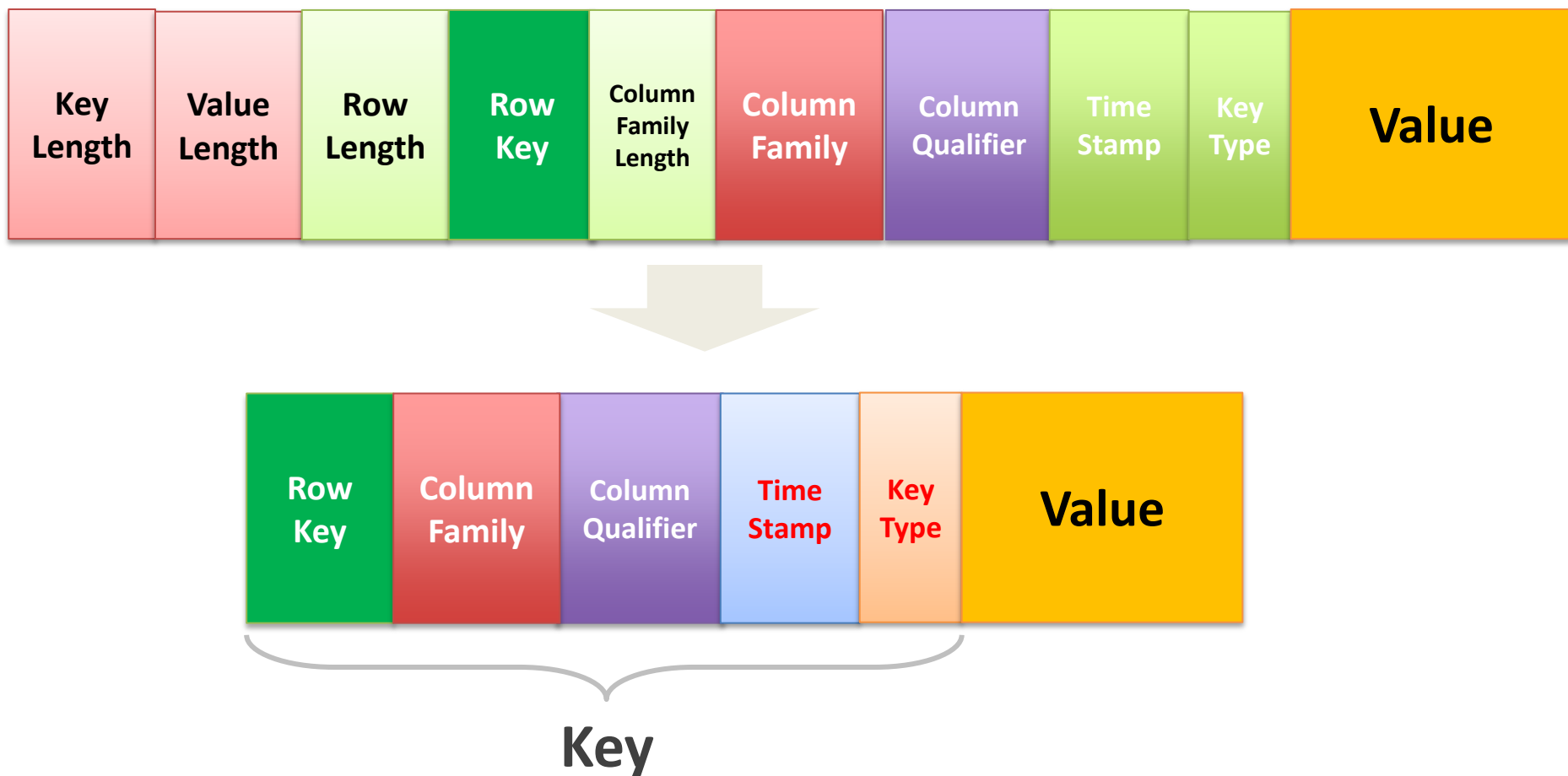


## HBase数据模型

- KeyValue
- HFileBlock
- HFile
- HLog

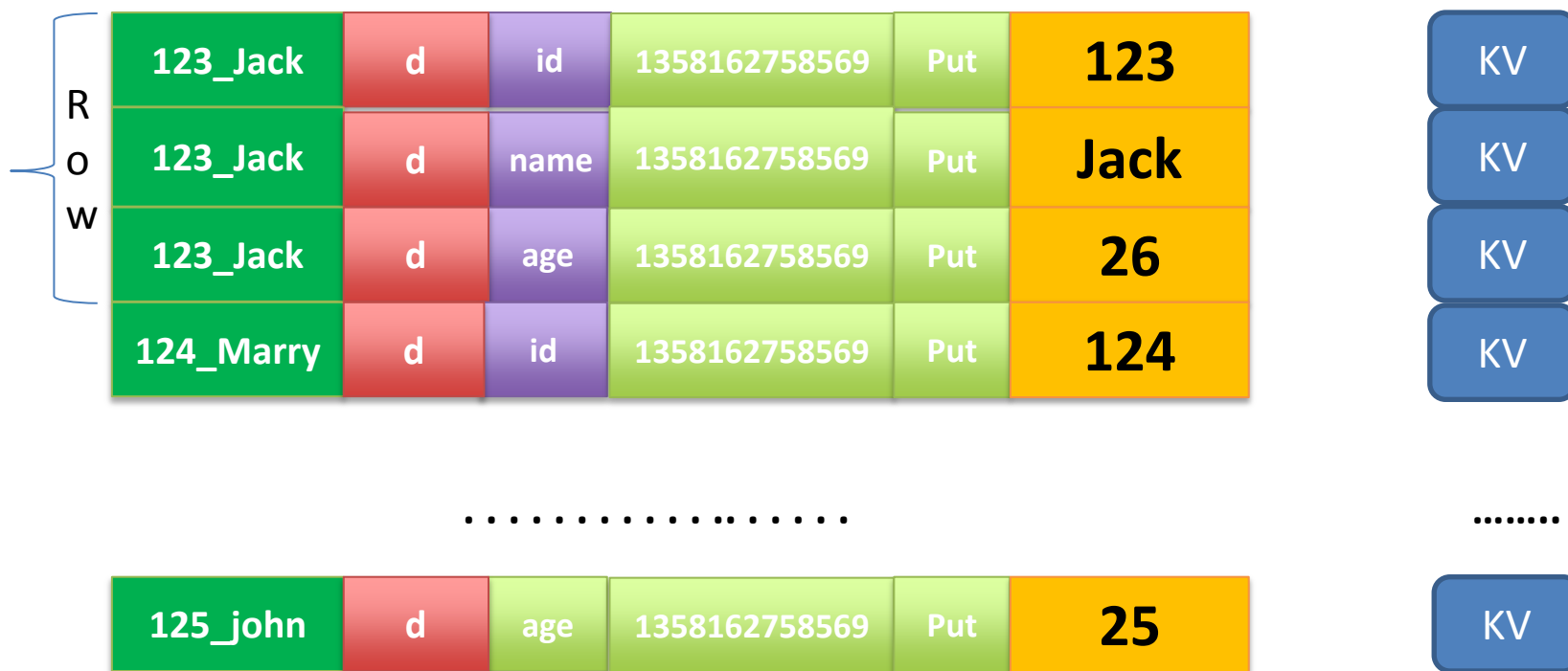


# HBase数据模型之KeyValue



id	name	age
123	Jack	26
124	Marry	24
125	John	25

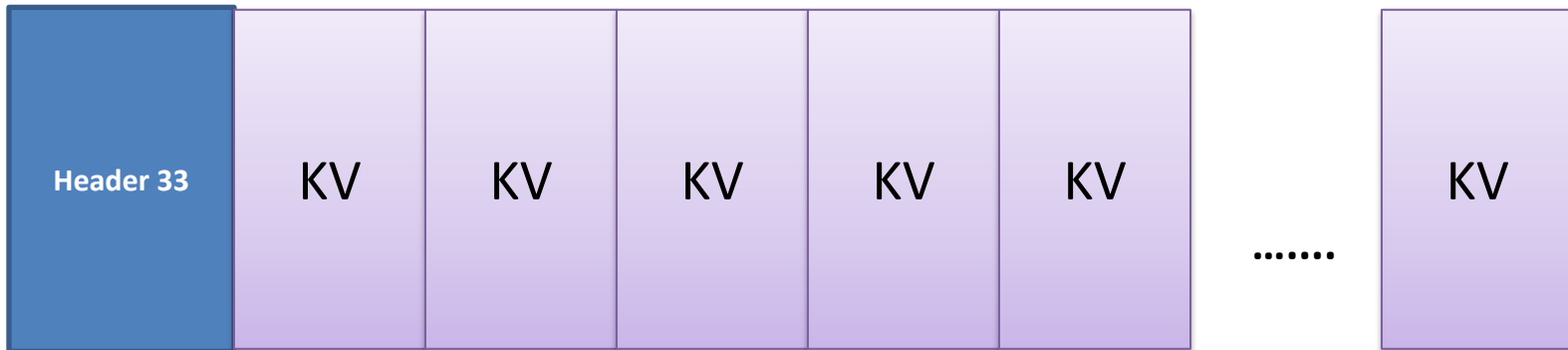
一张列簇为d的Hbase表 Rowkey就是id\_name



# HBase数据模型之HFileBlock

## HFileBlock

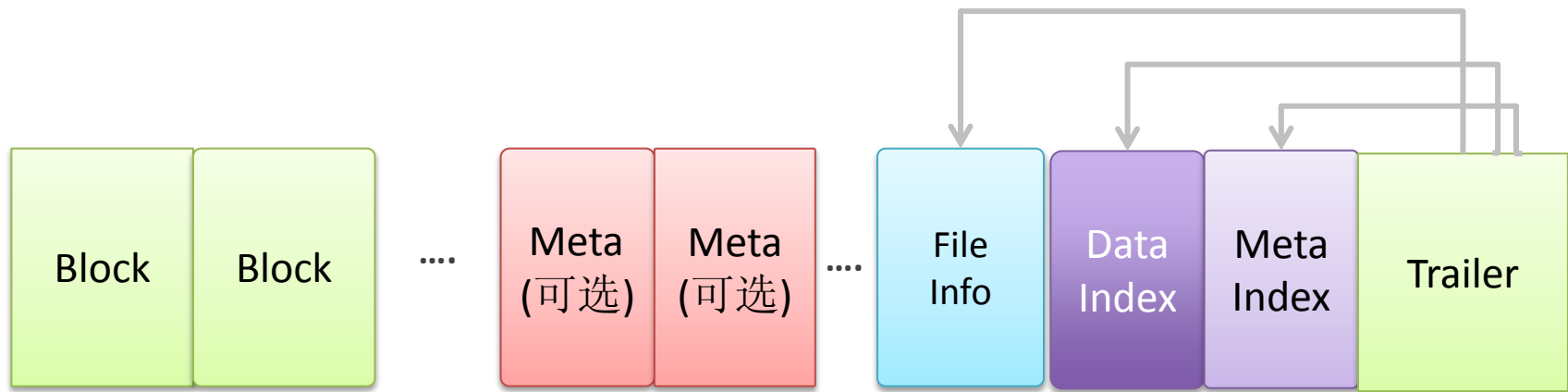
- header=HEADER\_SIZE\_NO\_CHECKSUM+ 1(byte)+2\*4(int)      后面可能还有checksum  
=8(magic\_length)+2\*4(int)+8(long) + 1(byte)+2\*4(int)



这就是HFileBlock 大小就是CF定义下的Blocksize，默认64K，就是HBase底层读写的基本单位，也是Cache的主要对象，也是StoreFile的主要结构;在这个以上的级别会有根据Rowkey的bloomFilter和StoreFileIndex;

较小的blockSize 对HBase的随机读改进有较大的帮助；但会使bloomFilter和StoreFileIndex大小膨胀的比较厉害

# HBase数据模型之StoreFile



将各个HFileBlock叠加 再加上BloomFilter StoreFileIndex FileInfo, Trailer 等信息就构成了StoreFile;

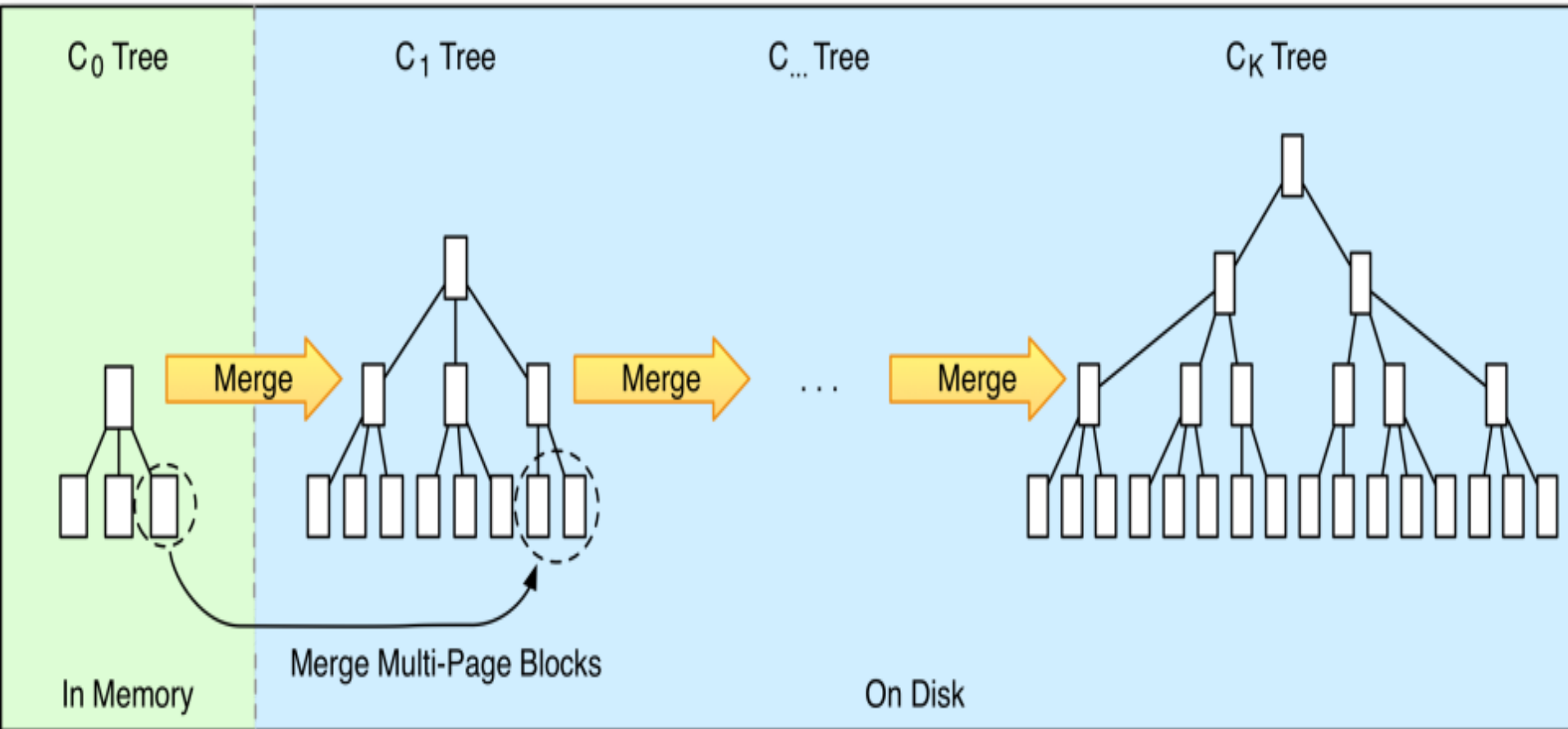
这里常说底层存储是HFile, 实际真正的文件级别的是StoreFile;

HFile只是定义读写StoreFile的接口及相关的工厂方法:

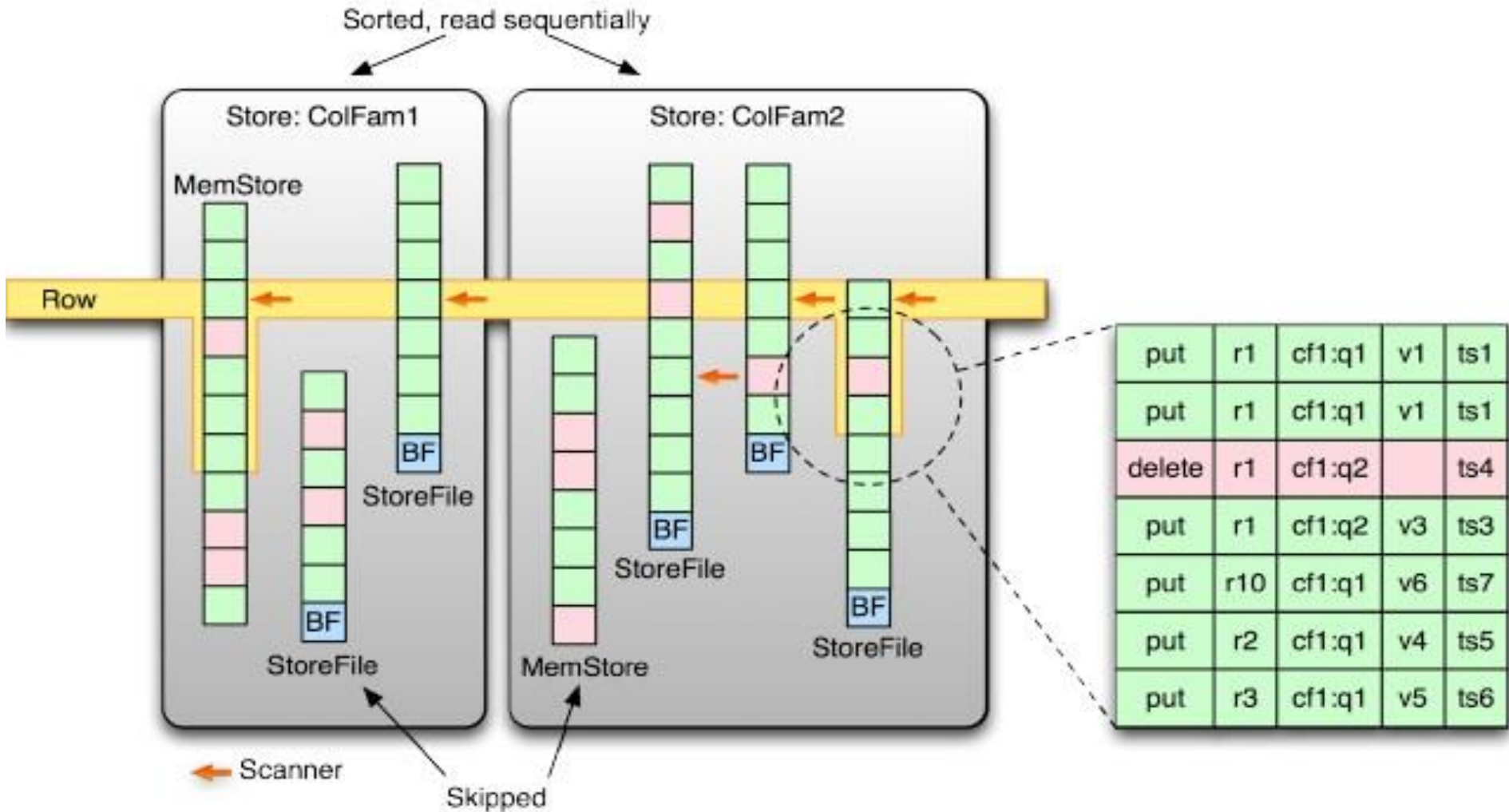
具体一点 HFile.Reader, HFile.Writer 和HFile.WriterFactory

HFileScanner 接口

# LSM-Tree(Log-Structured Merge-Tree)



# Merge Read



# Seek vs Transfer (B+ vs LSM)

- LSM追加写 合并读

读和写是相互独立的 所以不会有读写竞争的问题(准确说是读写竞争被控制在了很小的范围, 目前主要是IO资源的竞争), 写的响应时间可以预测, LSM通过定期的数据合并消除Delete和其他无效数据, 这个操作是可控的, 数据合并速率也直接受益于硬件性能的提升

- B+树

B+树的一些特性使得对于通过Key标识的记录可以进行有效的插入, 查找和删除操作, 也可以进行有效的范围查找。读写容易造成不连续的写和磁盘碎片, 影响效率 (ptcfree)



# HBase优缺点

- 优点

1. 强一致性
2. 自动扩容
3. 超高的写入性能
4. 低成本

- 缺点

1. 可用性相对较低(CAP)
2. 不支持跨行的事务
3. 没有通用的二级索引方案(要么性能低下)
4. 读性能非常的一般
5. 数据冗余严重

# 大纲

- HBase简介
- HBase在阿里的使用场景
- HBase 运维
- HBase阿里改进
- 其他

# HBase在阿里的应用

- HBase 规模和团队
- HBase 的主要使用场景
- HBase一些特有特性的使用
- HBase表设计

# HBase在阿里

- 目前阿里3大在线存储MySQL, OceanBase和HBase
- 目前机器规模在线1000+, 离线3000+
- 目前有超过200个集群, 数据总量已经无法统计了
- 目前有专门的HBase开发团队(沈春辉 committer)
- 两个HBase运维团队(目前已经合并)

# HBase的主要使用场景

- 海量历史数据
  1. 结合Mysql等传统数据库，降低存储成本(混合存储)
  2. 历史轨迹跟踪类(安全审计，行为分析等)
- 消息类
  1. 在线消息存储
  2. 通知消息推送
- 实时计算类
- 日志类
- 离线分析类

# HBase 一些特有的特性的使用

- 动态列
- TTL 数据自动过期删除
- 多版本
- ○ ○ ○ ○

# HBase表设计

- Rowkey设计
- ColumnFamily特性选择
- 表预切分
- 客户端使用调优



# HBase 运维

- HBase的运维挑战
- HBase 自动化运维
- HBase 容灾

# HBase日常运维

- 集群运维

1. 独立小集群还是混合大集群？如何有效运维大规模机器？
2. HBase依赖HDFS和Zookeeper,出现问题很难排查
3. HBase监控指标非常多，提取有效信息比较困难
4. 热点的检测和处理。。。。

- 项目接入

1. 什么样的项目适合使用HBase？
2. HBase如何正确评估扩容？

# HBase 自动化运维

- HBSqlplus 面向业务方的HBase自助查询平台
  - Ork 提供HBase项目接入和审核平台
  - HFree HBase自动化运维平台(有点类似HUE)
  - 改进的HBase监控(使用HBase存储Ganglia采集的监控数据)
- .....

# Hbsqlplus

The screenshot displays the HBase GUI interface. The top bar contains tabs for '查询窗口' (Query Window), '监控' (Monitoring), '建表工具' (Table Creation Tool), and '日常90性能测试' (Daily 90 Performance Test). The left sidebar shows a tree view of database clusters, with 'Pu\_(0.90.2RC5)' selected. The main area is divided into two panes. The top pane, titled '查询' (Query), shows a SQL query: `select * from ".META."`. The bottom pane, titled '查询结果' (Query Result), displays a message: '服务器返回消息' (Server return message) with a timestamp '2014-07-21 23:44:02'. The bottom window, titled '类型映射信息' (Type Mapping Information), shows a table with columns: '表名' (Table Name), 'ColumnF...', 'Column', '类型' (Type), '释义' (Description), '修改人员' (Modified by), and '修改时间' (Modified time). The table contains one row with the following data: 'info', 'rowkey', 'STRING\_UTF-8', '瑛阳(xuyang.xff)', and '2014-07-08 16:37:22'.

# Ork HBase项目接入评审平台

新建 打开 删除 刷新 高级查询

	创建时间	创建人	状态	最终存储方案	备注	审核人	DBA	项目名称	产品线
1	2013-04-09 16:05:14	王书	待TL审核			穆公	穆公	风险审核平台	集团内 rcp
2	2013-04-08 20:41:21	孙金	待TL审核			穆公	穆公	MTCC模型平台	集团内
3	2013-04-08 11:55:21	王书	HBase资源已分配	HBase	日志型业务 append型 后续写入量增加几十倍-100倍 ====	穆公	穆公	ASD公有云	阿里云
4	2013-04-08 11:44:24	王书	HBase资源已分配	HBase	写多读少 缓存走 tair	穆公	穆公	店铺动态	集团内
5	2013-04-08 10:12:29	王书	推荐非HBase资源	MySQL		穆公	穆公	标签化导购	淘宝内
6	2013-03-29 14:22:2								
7	2013-03-26 15:41:4								
8	2013-03-25 19:04:4								
9	2013-03-25 11:40:2								
10	2013-03-19 15:34:1								
11	2013-03-19 13:38:2								
12	2013-03-19 10:35:2								
13	2013-03-14 15:09:3								

查看资源申请工单

项目信息 ① 数据量规划表 ② 项目建表 ③ 审批 ④ 项目分配资源详情 ⑤ 意见反馈 ⑥

申请人: 必成 项目故障级别: p4 项目名称: 店铺动态

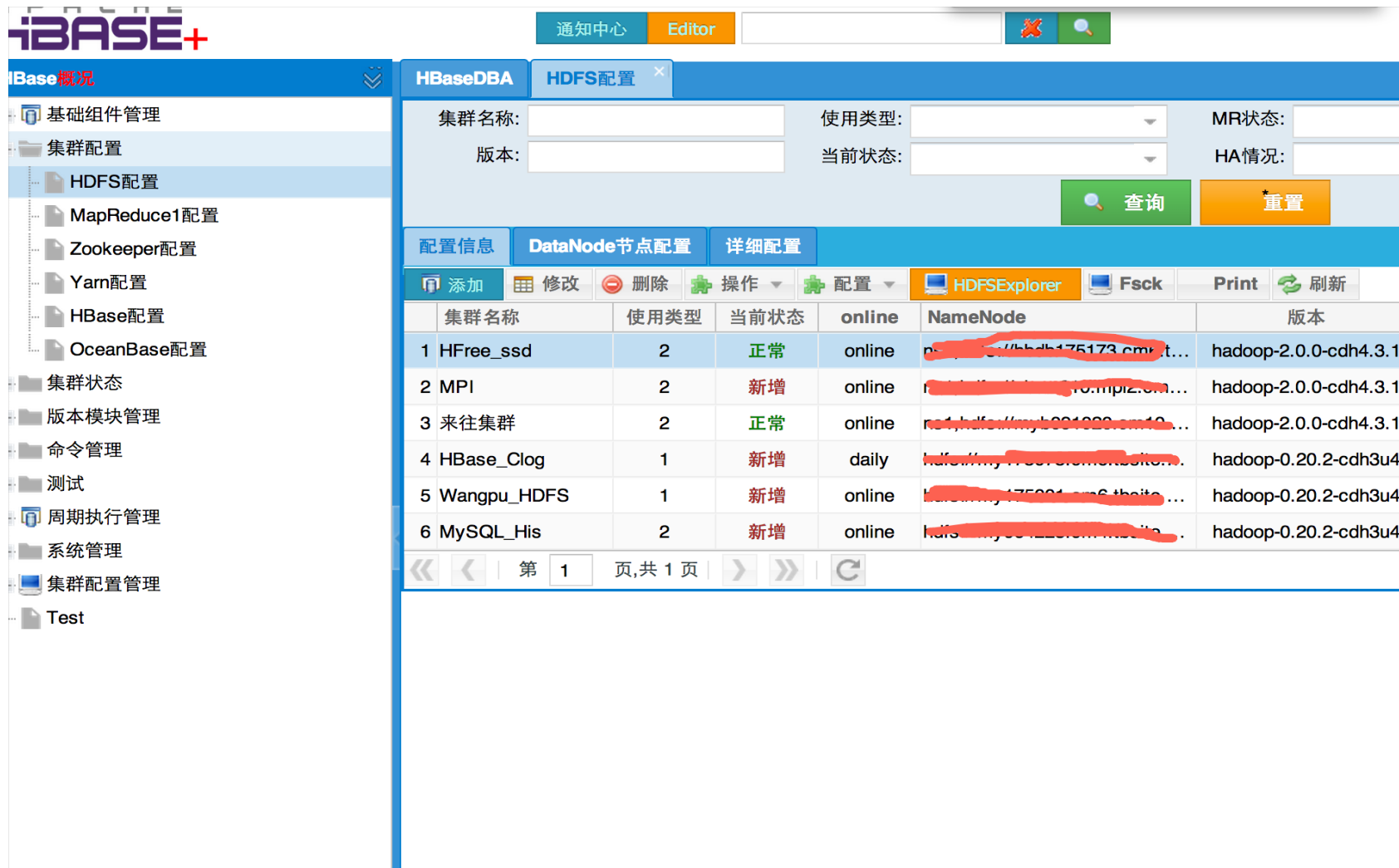
产品线: sns>卖家服务>ishopbook TL: 穆公

项目背景: 店铺动态改版项目。

资源申请理由: 现有mysql无法支撑 对应DBA: 袁斌 项目上线日期: 2013-04-30

需求详细描述: 1、买家访问店铺动态时，页面显示店铺列表，其中用户特别关注的店铺会排在前面。  
2、卖家需要知道自己店铺的特别关注的买家列表。店铺有时会推送消息给特别关注的买家。

# HFree HBase 自动化运维平台



通知中心 Editor

HBaseDBA HDFS配置

集群名称: 使用类型: MR状态: 版本: 当前状态: HA情况:

查询 重置

配置信息 DataNode节点配置 详细配置

添加 修改 删除 操作 配置 HDFSExplorer Fscck Print 刷新

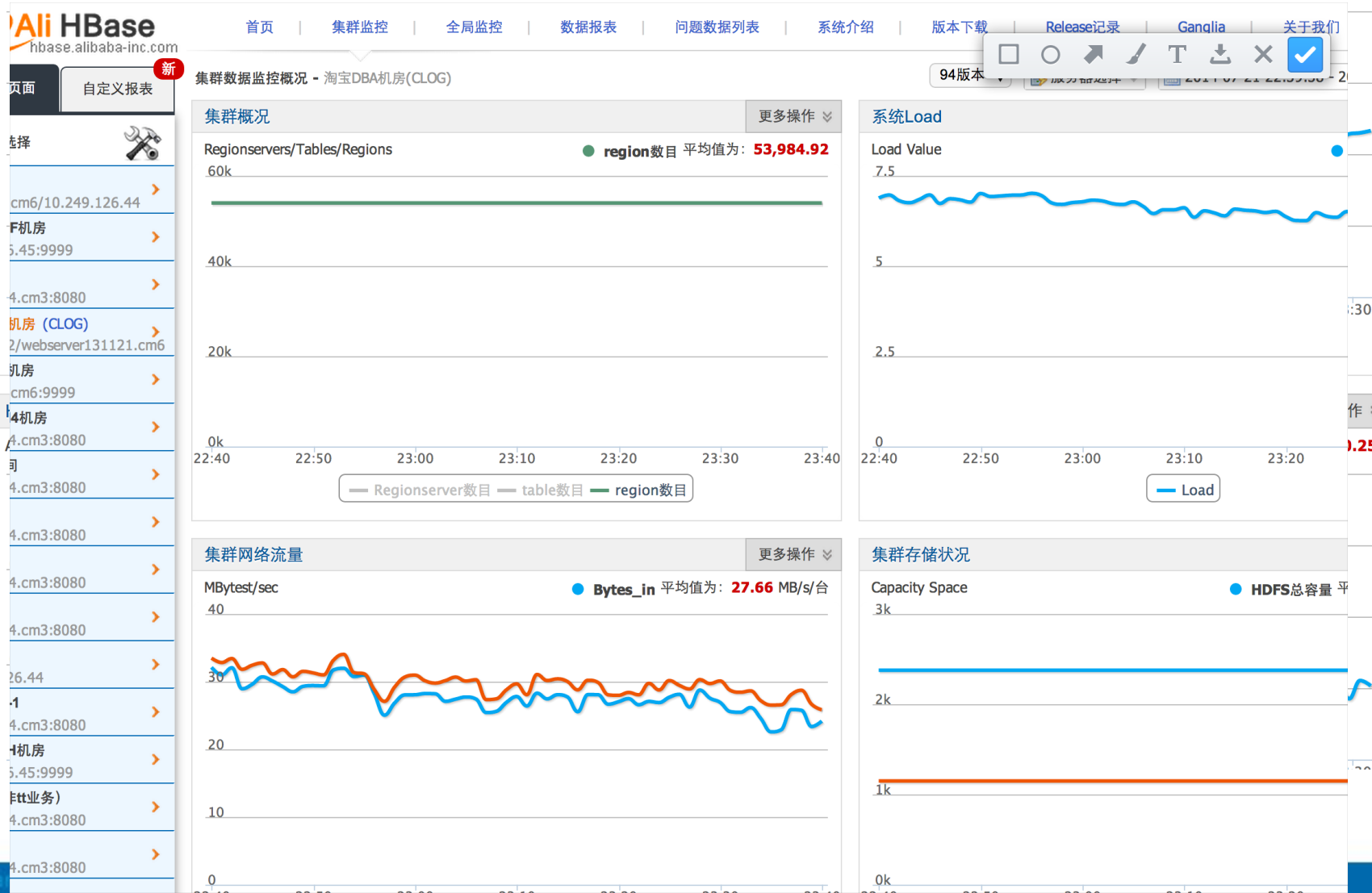
	集群名称	使用类型	当前状态	online	NameNode	版本
1	HFree_ssd	2	正常	online	hdfs://hbase175173.com:...	hadoop-2.0.0-cdh4.3.1
2	MPI	2	新增	online	hdfs://hbase175173.com:...	hadoop-2.0.0-cdh4.3.1
3	来往集群	2	正常	online	hdfs://hbase175173.com:...	hadoop-2.0.0-cdh4.3.1
4	HBase_Clog	1	新增	daily	hdfs://hbase175173.com:...	hadoop-0.20.2-cdh3u4
5	Wangpu_HDFS	1	新增	online	hdfs://hbase175173.com:...	hadoop-0.20.2-cdh3u4
6	MySQL_His	2	新增	online	hdfs://hbase175173.com:...	hadoop-0.20.2-cdh3u4

第 1 页, 共 1 页

节点配置		详细配置			
loadFile		载入	修改	删除	刷新
		类别:			配置:
配置文件	配置	值		配置信息	
		当前值	默认值	级别	来源
default	BASE_HOME	/u01/hbase/			
hadoop-env.sh	JAVA_HOME	/opt/taobao/java	/opt/taobao/java	10	public
core-site.xml	fs.trash.interval	1440			
core-site.xml	ipc.server.tcpnodelay	true			
core-site.xml	ipc.client.tcpnodelay	true			
core-site.xml	hadoop.tmp.dir	\${BASE_HOME}/temp			
hdfs-site.xml	dfs.datanode.handler.count	500			
hdfs-site.xml	dfs.datanode.data.dir	file:/data/1			
hdfs-site.xml	dfs.namenode.handler.count	300			
hdfs-site.xml	dfs.namenode.name.dir	\${BASE_HOME}/hdfs/...			
hdfs-site.xml	dfs.journalnode.edits.dir	\${BASE_HOME}/hdfs/...			



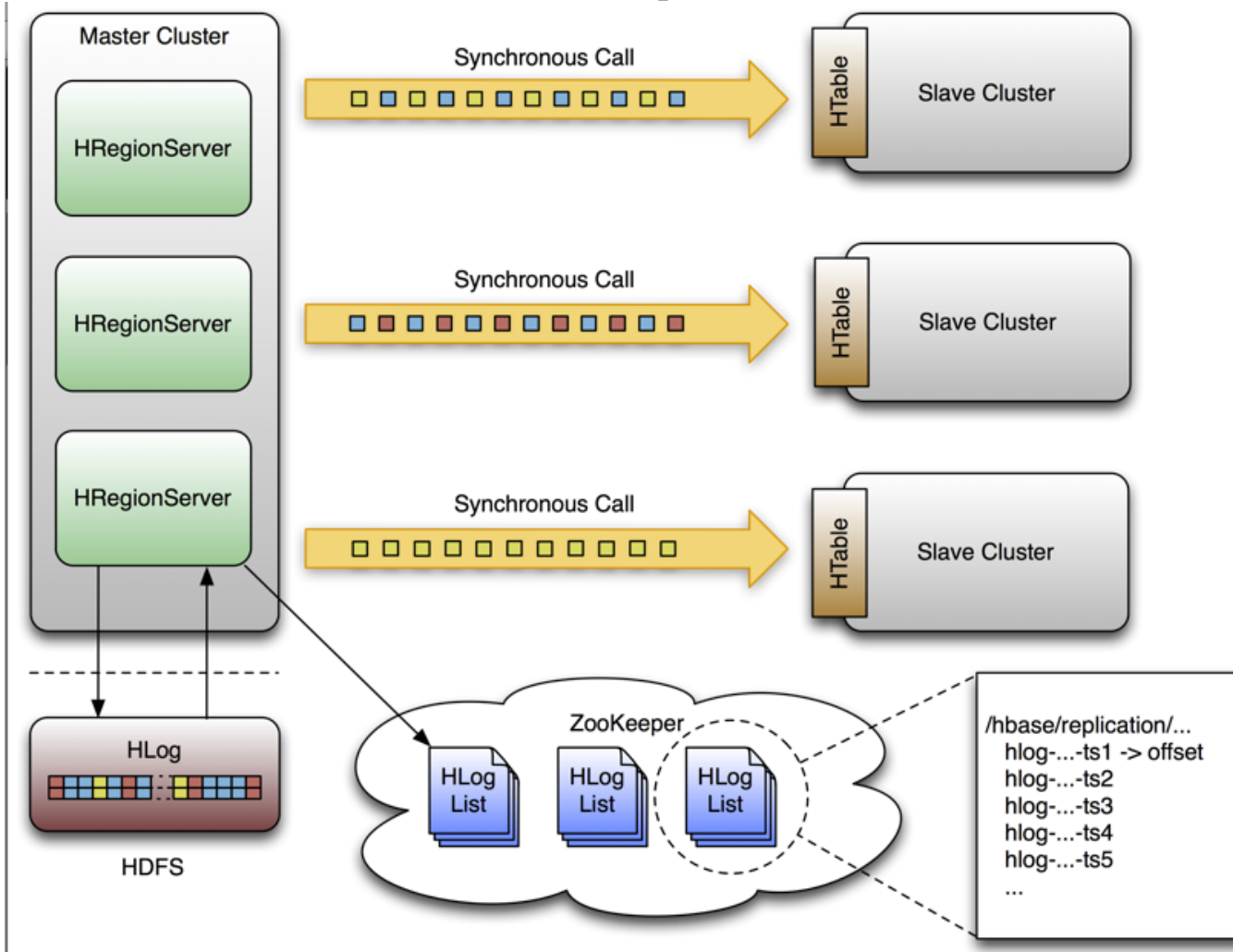
# HBase 监控



# HBase容灾(HBase主备方案)

- Replication方案
- IBack方案
- 自动化切换客户端

# HBase Replication

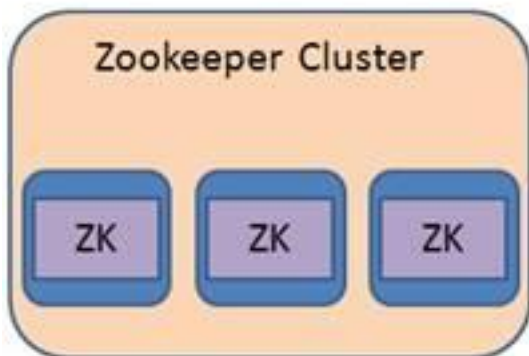


# HBase Replication的问题

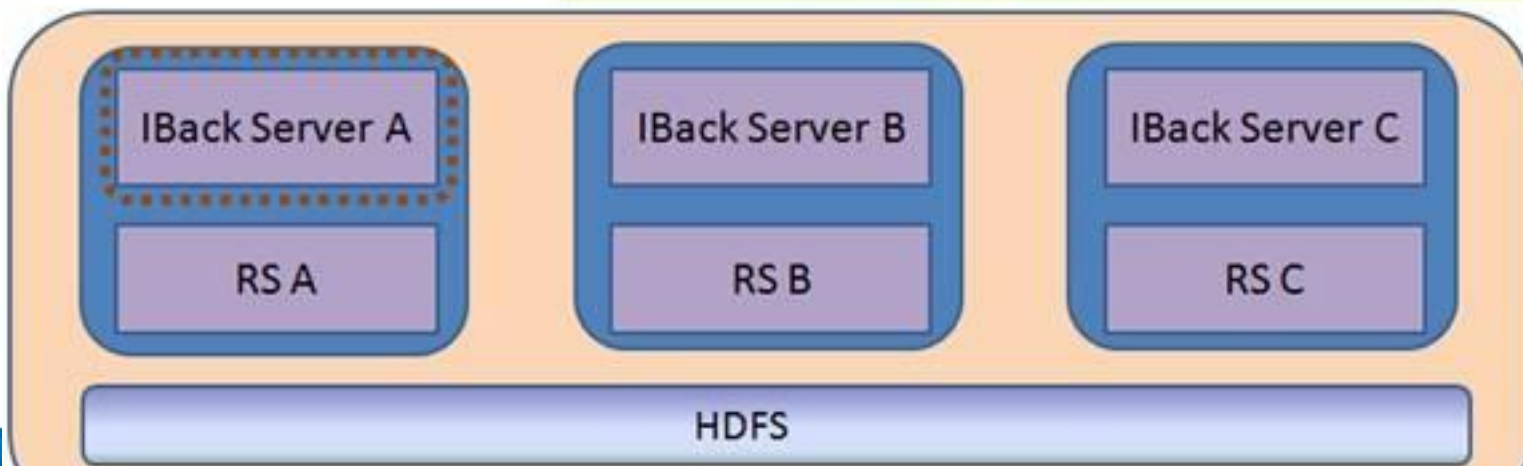
- 1. HBase的replication只支持从Master Push数据到Slave集群，也就是replication程序的处理过程需要在Master集群上进行，会影响Master集群的性能。
- 2. replication必须依赖于HBase进程，只支持持续的数据迁移这一种模式，不支持指定时间段的数据迁移模式，这将无法满足指定时间段的数据补齐,当主集群由于某种原因无法提供服务，例如机房断电等时，Master集群的部分数据还没有来得及写入Slave，会造成Slave集群部分数据不可用，对于核心的业务，这点是无法接受的。
- 3. HBase的replication无法支持扩展的需求，比如增量表；

# IBack方案

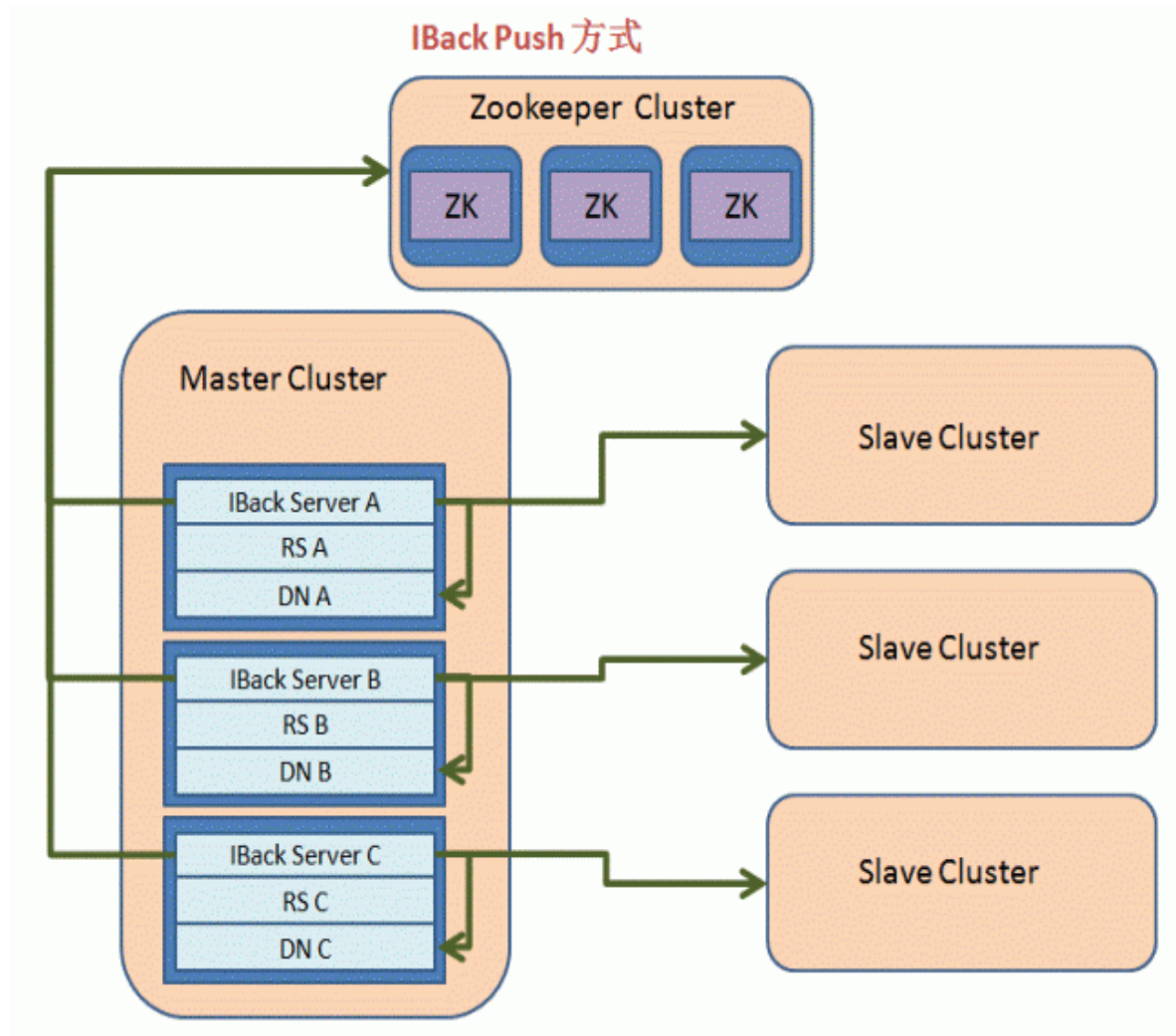
## IBack 架构



```
/iback
  /config/tables      qianzhen_iback_test,qianzhen_iback_test2
  /rs/A/hlog1         path and position
    / hlog2           path and position
  /lock               A
  /B .....           B...
  /C .....           C...
  /ib/A               true
  /B                  true
  /C                  true
  /master/findhlogib
                        /starttimefornext 1368102808953
                        /lock               A
  /herald             A
```

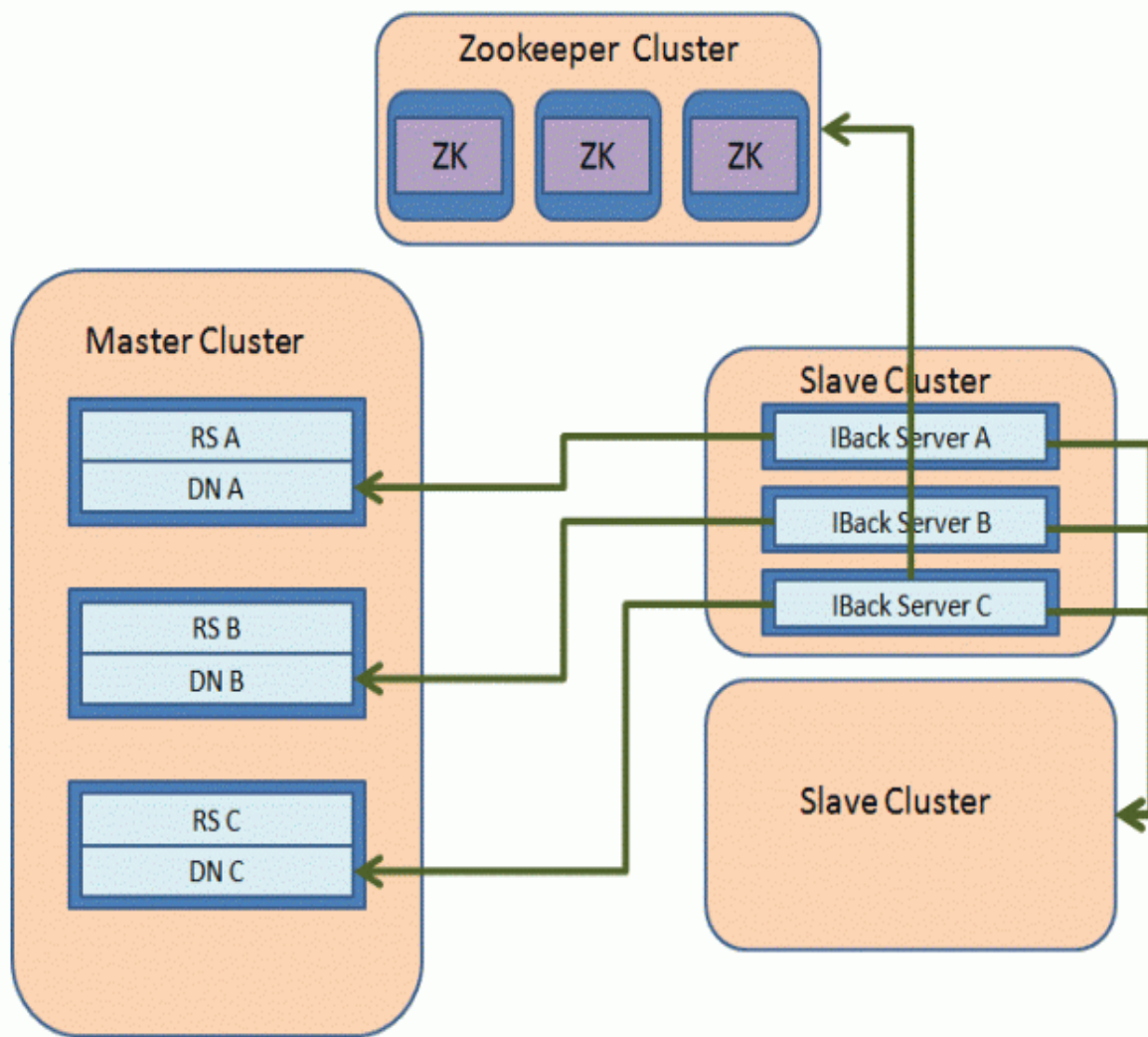


# IBack的两种运行模式





## IBack Pull 方式





# HBase 自动切换客户端

- 主备切换 客户端同步切换集群
- 主备直接流量调节
- 基于HBaseClient改造完成
- 目前只能使用手动切换

# HBase改进

- BucketCache(目前已经合并到HBase主干)
- 辅助存储(一种折中的二级存储方案)
- 混合存储的尝试(SATA SSD)
- RegionServer级别的配置改到表级别，更加灵活支持混合集群

## 其他一些场景

- HBase数据分析
- 全量 Hfile or HBase
- T+1 增量 HLog
- Mapreduce Spark ODPS

- 可以关注技术保障部的微博@阿里技术保障
- 有兴趣来阿里试试的可以发邮件到xuyang.xff@alibaba-inc.com

谢谢！

