阿里Hadoop集群运维介绍

柯旻（大舞）

阿里巴巴技术保障部门-云计算运维

大纲

- 阿里hadoop集群发展现状
- 监控报警
- 自动化运维
- 数据化运维
- 大规模集群下遇到的运维问题和新挑战

集群容量与负载

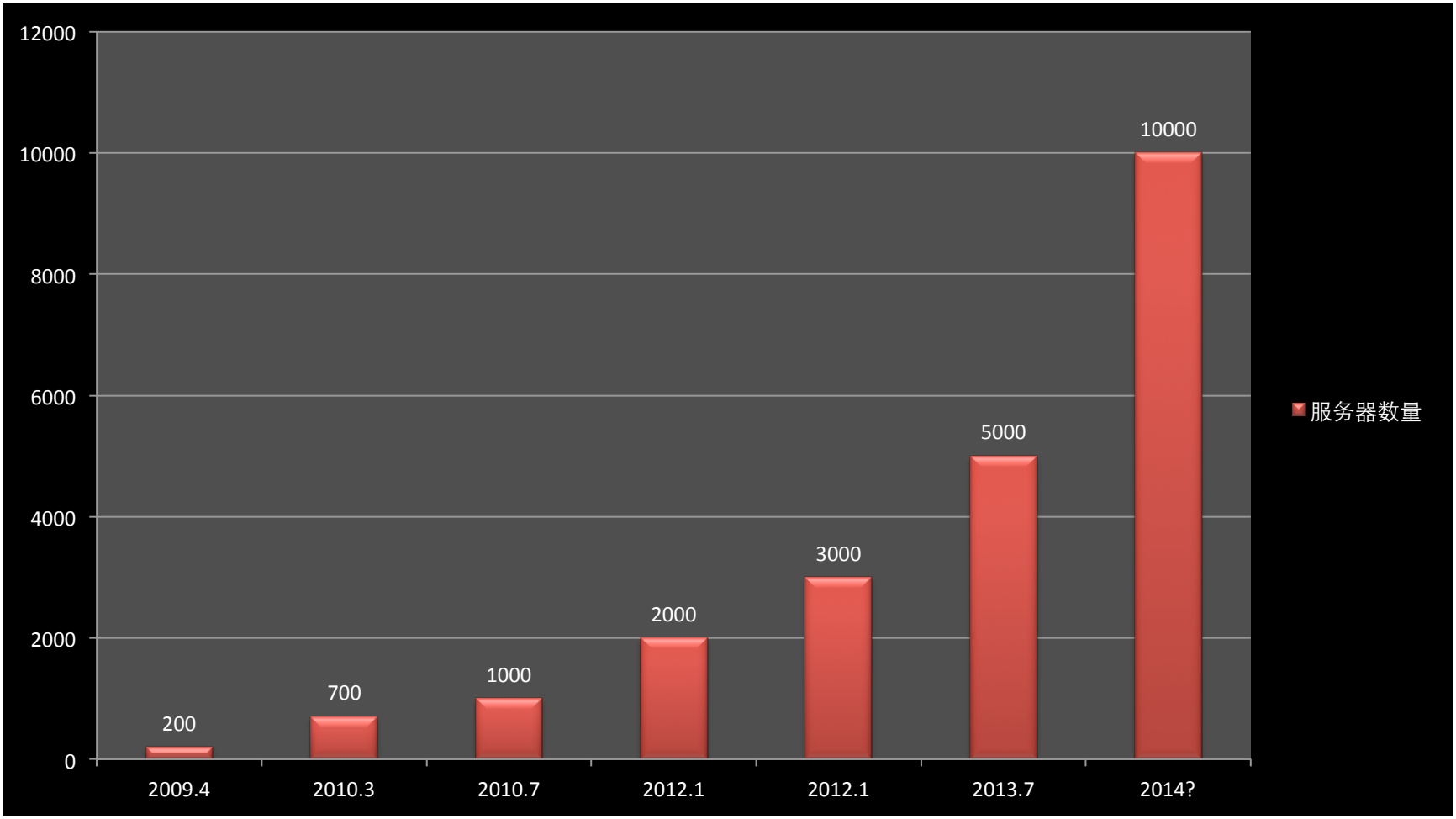
集群容量

- –约~ 5000台服务器
- –CPU core ~50000核
- –内存 ~260TB
- –磁盘 ~120000块
- –存储容量 ~110PB

集群负载（每天）

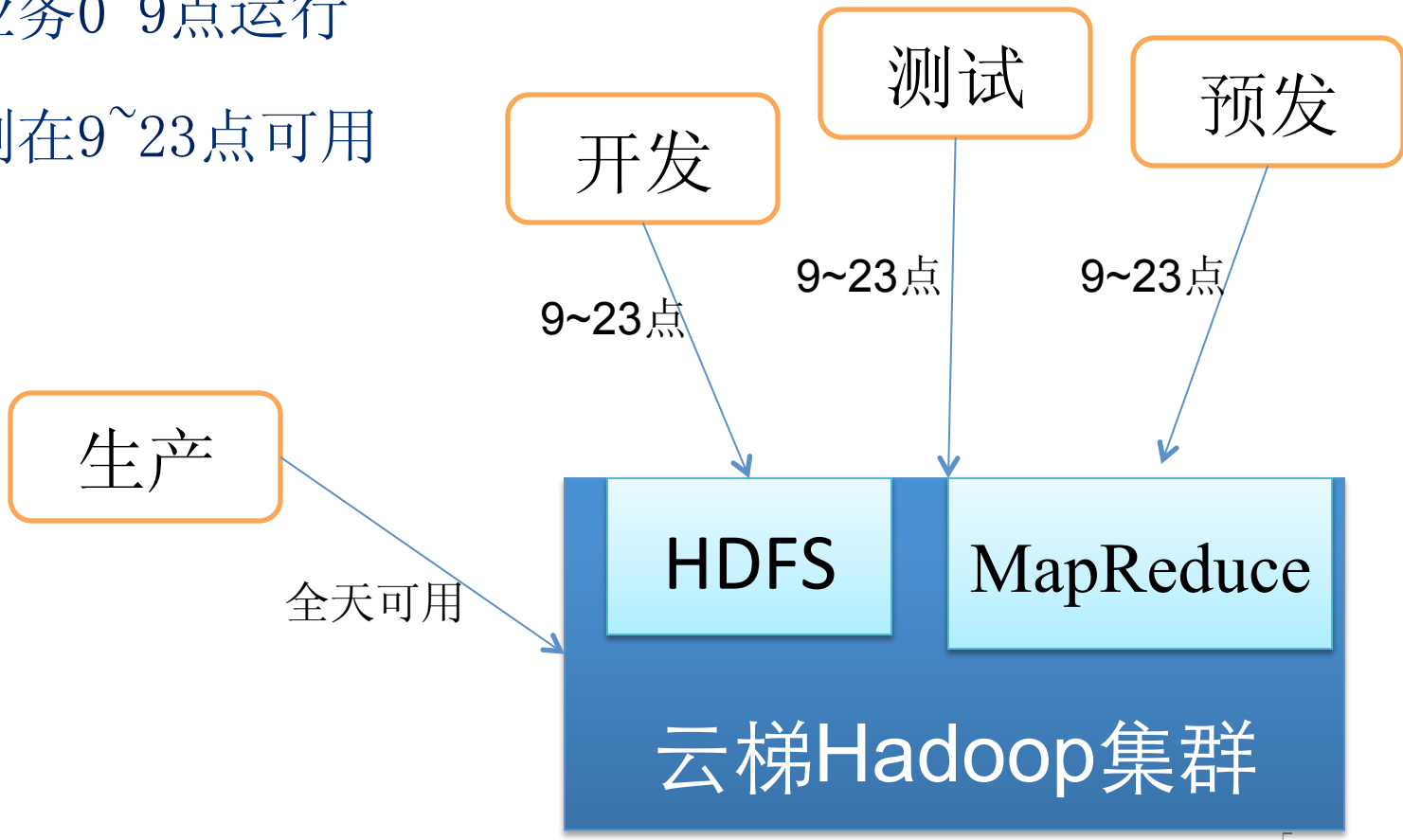
- Job 200,000+
- 扫描数据量~10PB
- 扫描文件数 ~4亿
- 存储利用率 ~75-80%
- –CPU利用率~70% 峰值85%

服务器数量增长

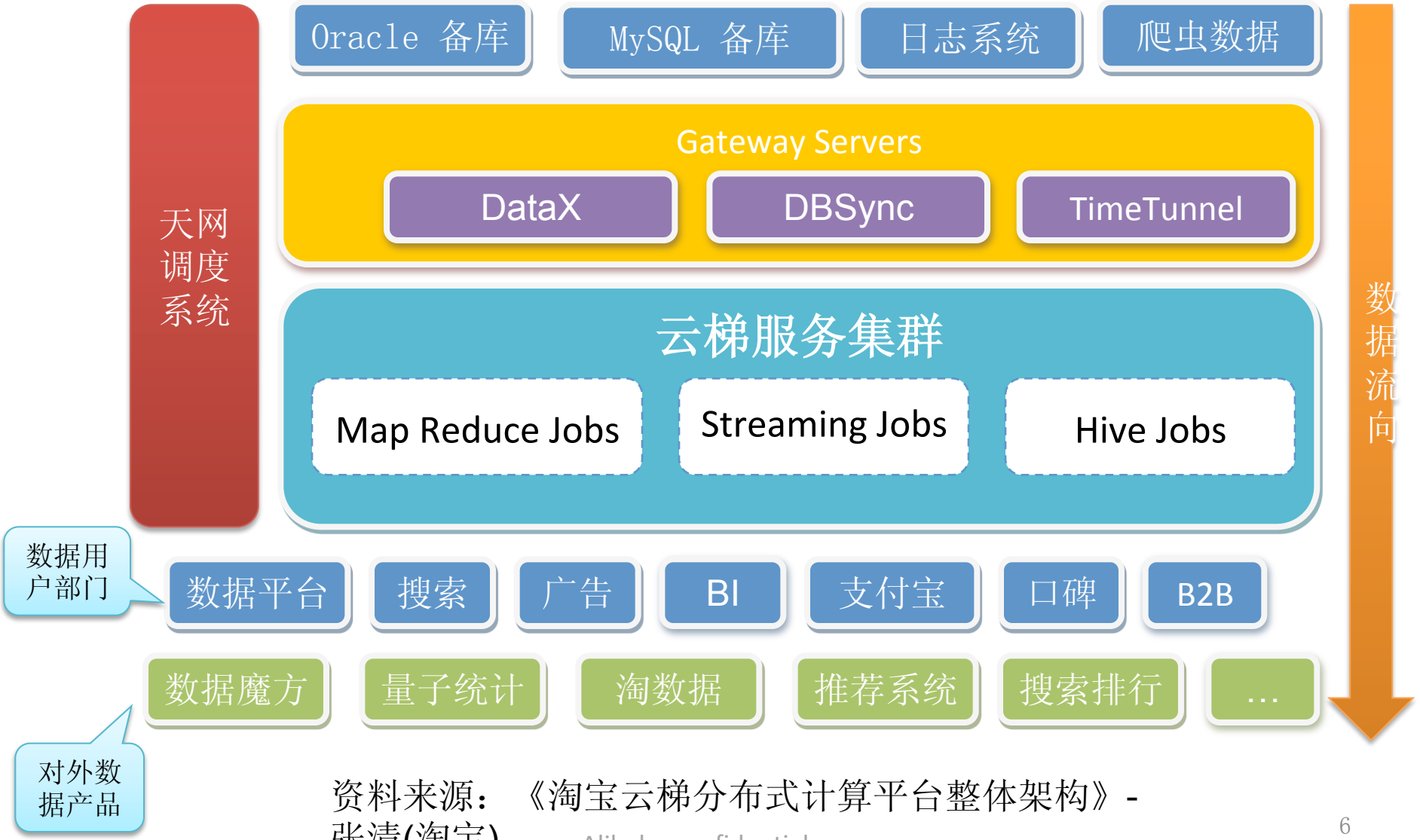


集群服务模式

- 生产\开发\测试\预发共享一个集群
- 重点生产业务0~9点运行
- 非生产限制在9~23点可用



集群核心业务平台架构

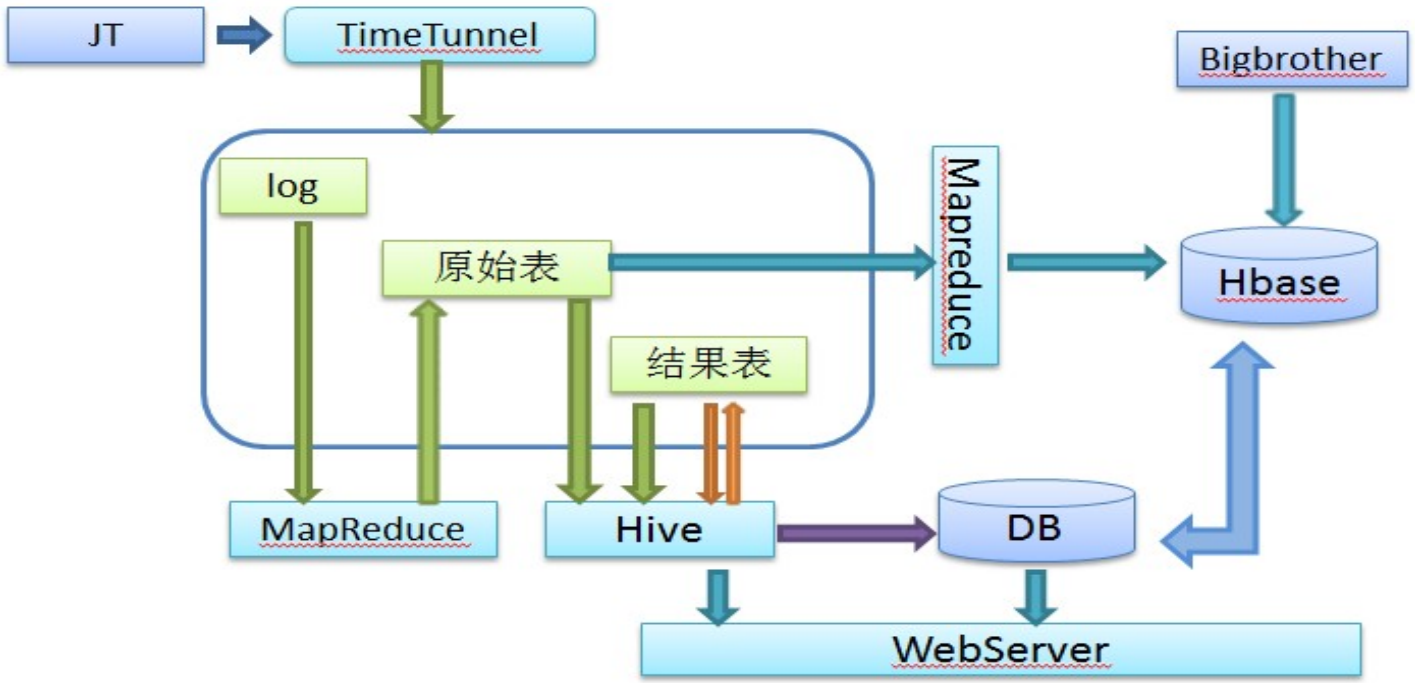


资料来源：《淘宝云梯分布式计算平台整体架构》-张清(淘宝)

监控报警

- 监控Hadoop关键进程，磁盘运行状况等短信、旺旺、邮件报警等
- 监控集群整体运行状态和Hadoop运行参数数据
  - Job的Counter数限制
  - 创建HDFS文件数目的监控
  - 本地文件系统数据读写量监控
- 异常作业监控
- End to end 监控
- 云梯医生监控各类用户态数据

云梯医生



- 展示一些应用组件基本信息

（setup、map、reduce、cleanup, split、map、copy、sort、reduce、output HDFS读写数据量、本地读写数据量、使用slots、调度等待时间、task失败的比例、task失败原因分类、task失败的机器）

- 针对用户提供体检服务
- 支持定制服务和实时体检

自动化运维

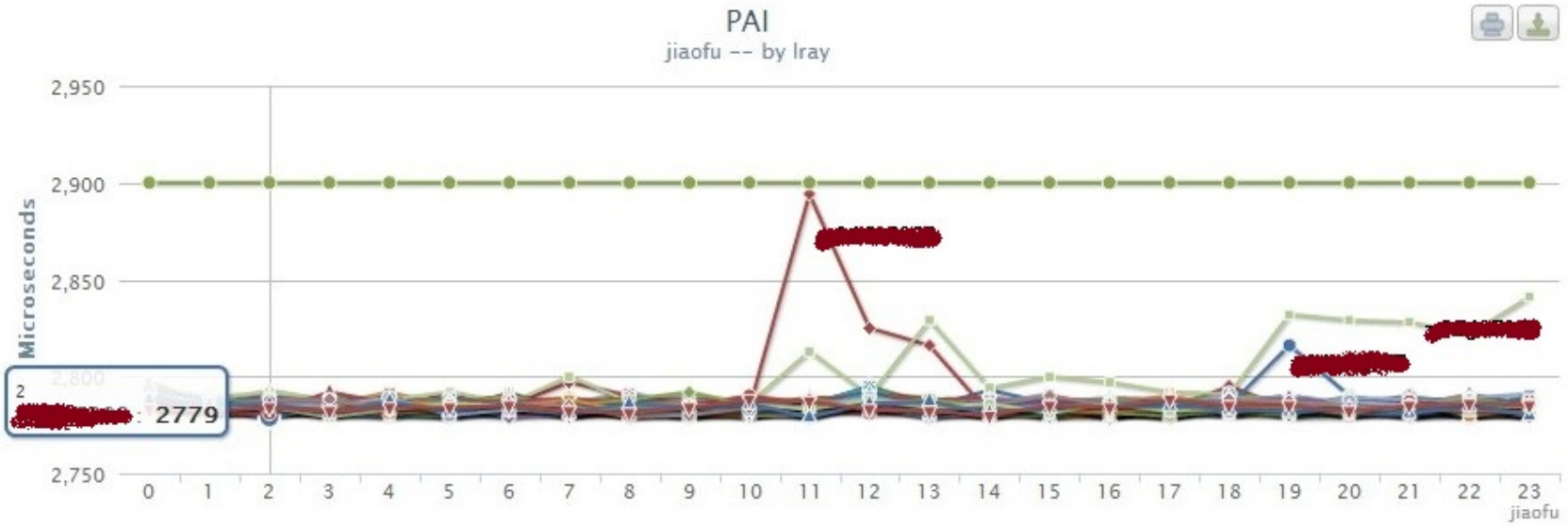
1. 服务器上线前自动化检查
2. 硬盘异常自动化处理
3. 集群用户一站式portal
4. 日常各类自动化运行报表

……

服务器上线前自动化检查

硬件上线前监测（fw版本，bios配置，驱动版本以及性能情况）

CPU_PAI



1. [redacted] 值有点异常

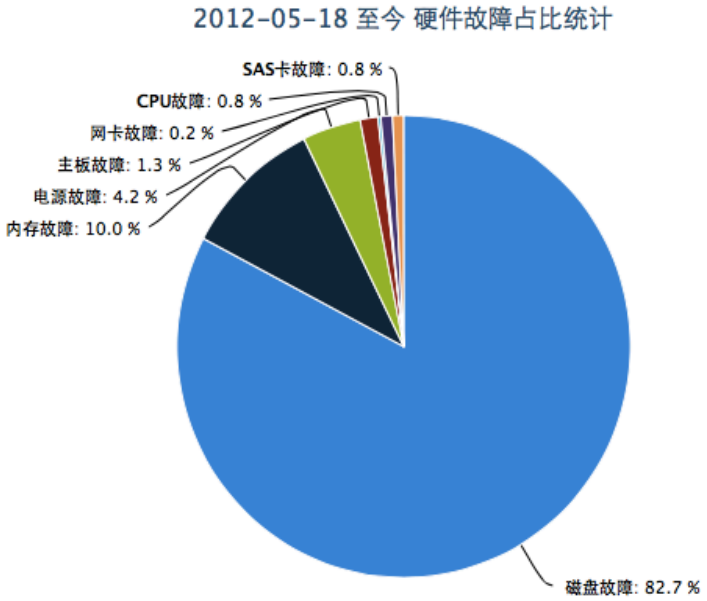
硬盘异常自动化处理

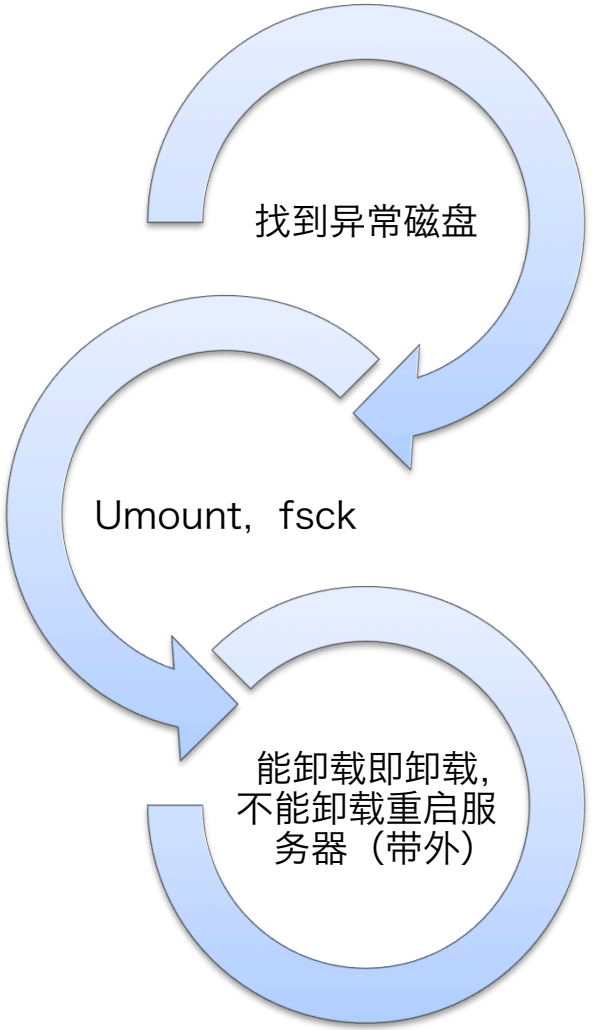
硬盘异常自动处理

1.廉价、大容量的硬盘

2.磁盘繁忙度和利用率很高

3.硬盘故障率远高于其他硬件





集群用户一站式Portal

▽用户服务

- 申请用户
- 申请用户组
- 申请Gateway
- 进度查询
- 查询用户
- 查询用户组
- 查询Gateway
- 用户手册

▽组管理员服务

- 申请审批
- 申请Slots
- 申请存储

▽集群管理员服务

- 申请审批
- 管理用户
- 管理用户组
- 管理Gateway
- Queue管理
- Slots管理

- 集群用户一站式完成各类申请
- 组管理员负责申请计算\存储资源
- 集群管理员通过web控制调整集群配置

日常各类自动化运行报表

AY10 集群监控报警开关情况日报

集群主机总数	报警关闭总数	报警开启总数
[redacted]	12	[redacted]

监控报警关闭主机如下：

IP	服务器状态	关键监控项状态	监控开启
	UP	Tasktracker OK Datanode OK	SUCCESS
	DOWN		no need
	DOWN		no need
	UP		no need
	UP		no need
	UP		no need
	UP		no need
	UP		no need
	UP		no need
	UP		no need

数据化运维

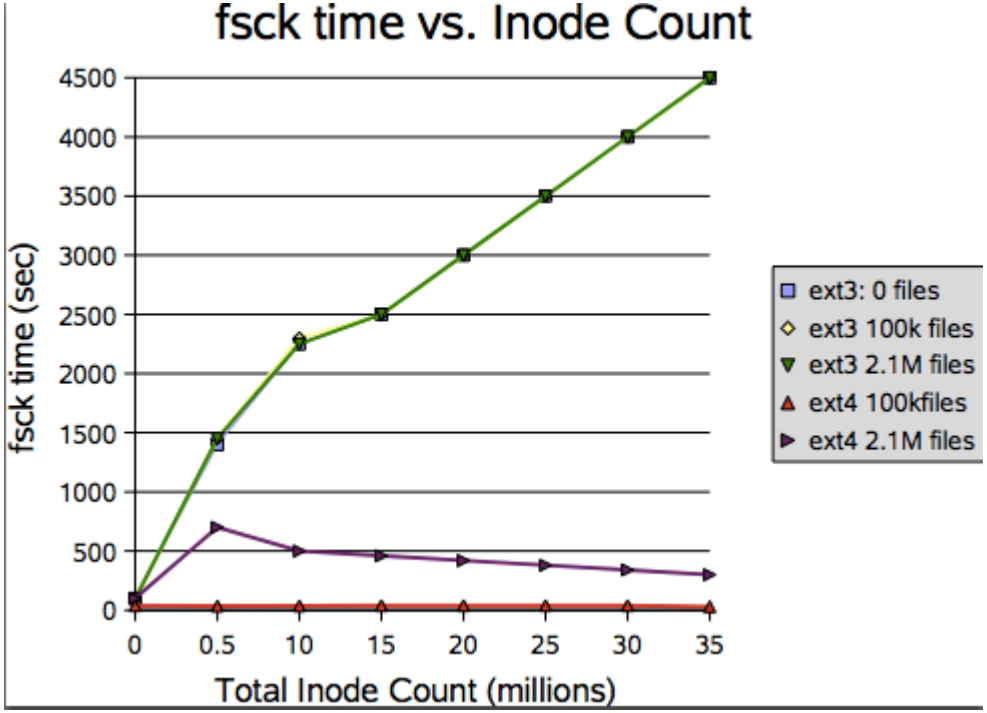
- 自动化后是不是就够了？
- 1000台，1万台我们还有经验可以借鉴，10万台，50万台，100万台后我们借鉴什么？
- 拍脑袋的决定不一定靠谱了，随着规模的扩大也许一拍下去会跟公司造成巨大损失

数据才是唯一真实可靠的！

磁盘

- EXT3文件系统，当时的数据量，做一次fsck需要至少半个小时

- 每次修复需要停掉一台节点的应用
  H云计算平台的133天中：
  625次文件系统的fsck。
  平均每天需进行4.7次的修复。
  保守按耗费半小时来计算
  造成每天有141分钟单台不可用时间



有数据就有底气：

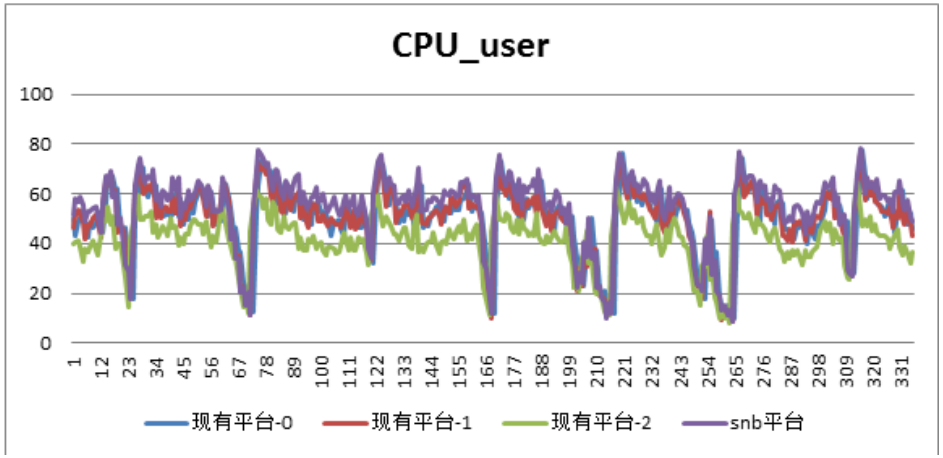
- 应用程序改造能允许在线卸载目录
- 在线挂载目录后应用程序能够识别
- 选用更可靠的EXT4文件系统

→ 141分钟 downtime

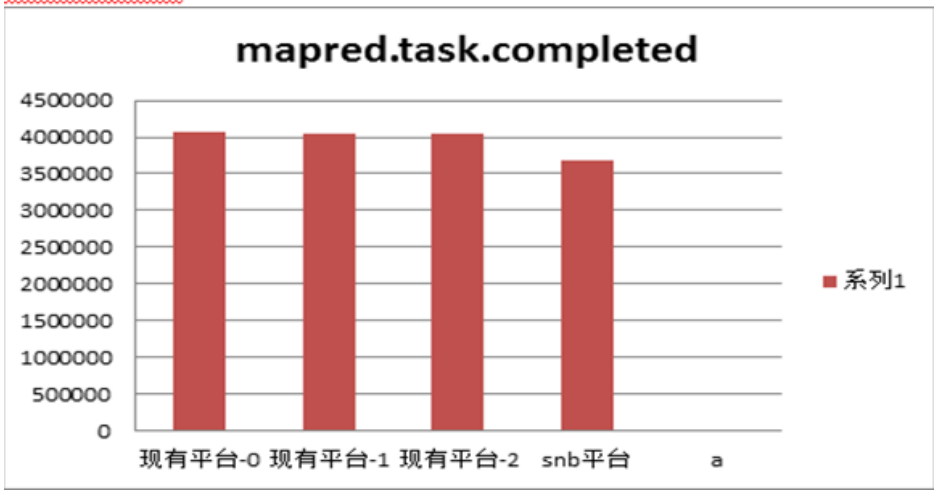
→ 减少恢复时间

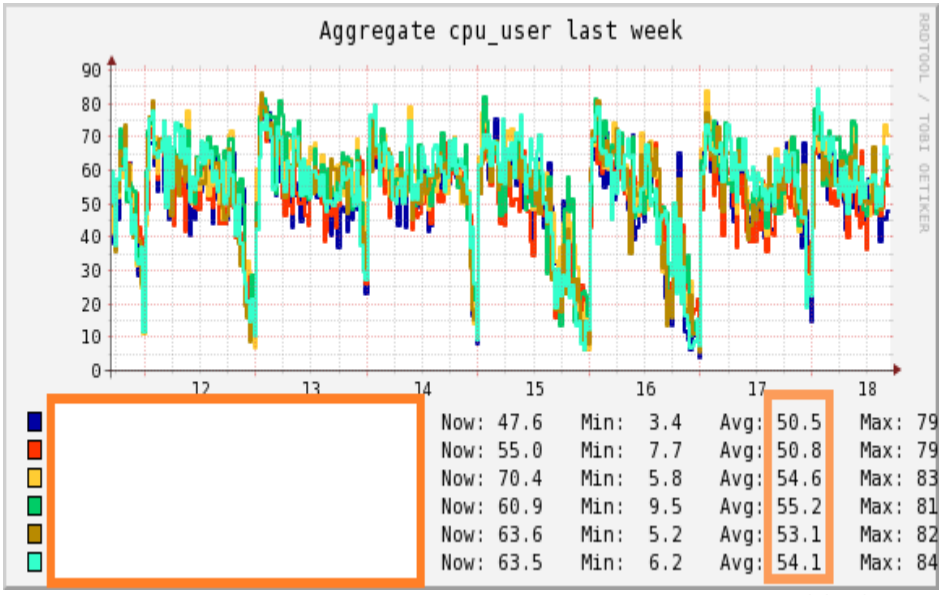
服务器

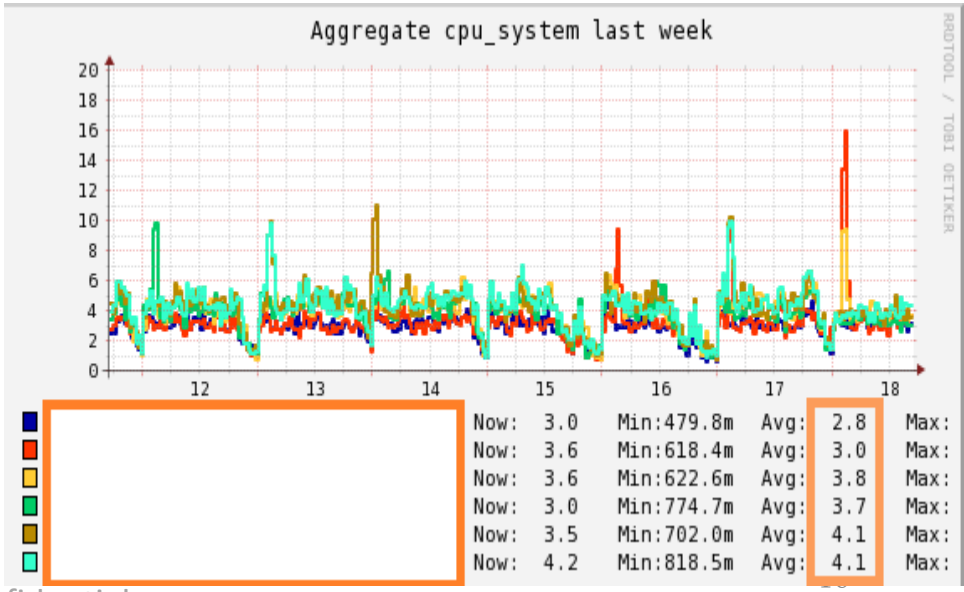
CPU_USER 的利用率，低谷基本和现有平台一致，高峰要高于现有平台 2%-5%



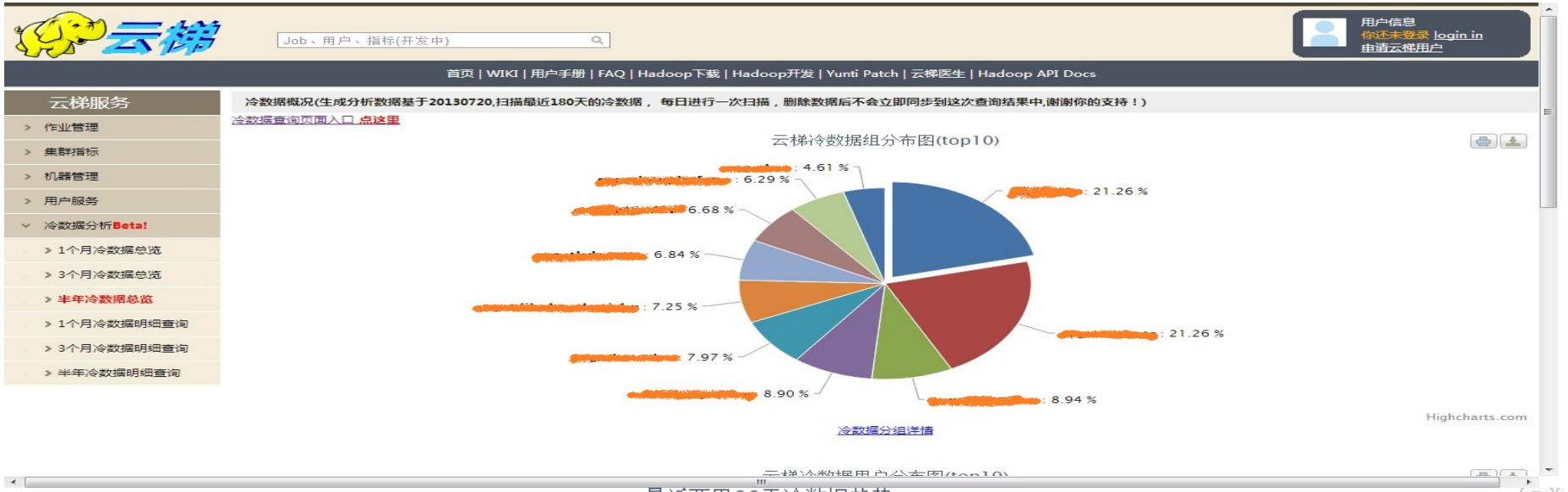
Task_completed 的数，采样了 3 组现有平台的数据对比，SNB 平台低 8.9%

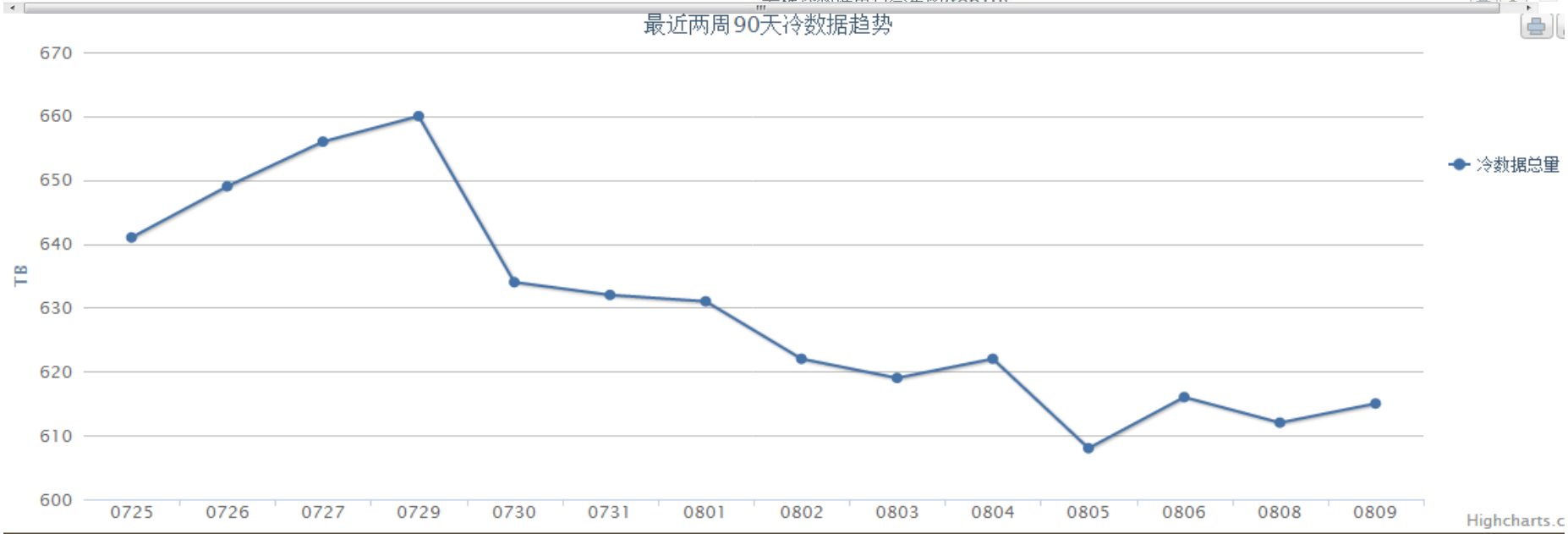




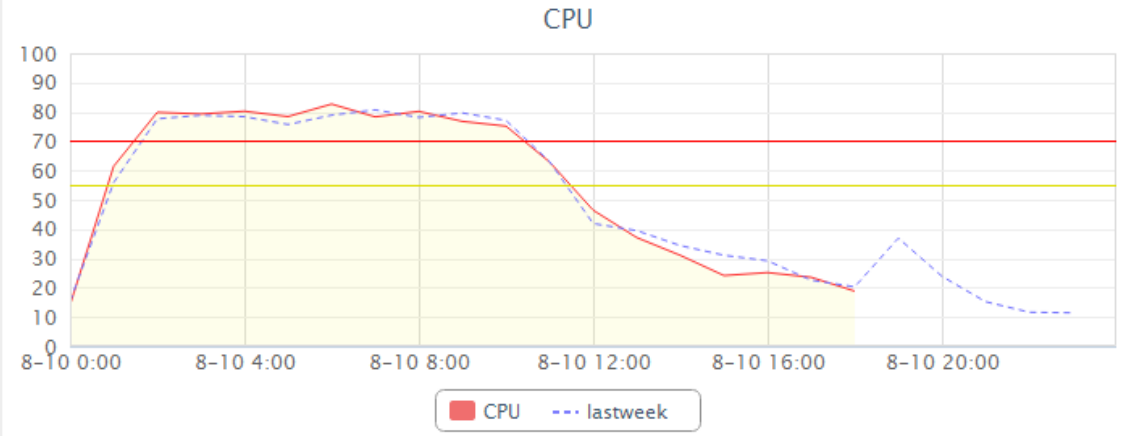


冷数据


云梯
Job、用户、指标(开发中)
用户信息
你还未登录 login in
申请云梯用户
首页 | WIKI | 用户手册 | FAQ | Hadoop下载 | Hadoop开发 | Yunti Patch | 云梯医生 | Hadoop API Docs
云梯服务
作业管理
集群指标
机器管理
用户服务
冷数据分析Beta!
1个月冷数据总览
3个月冷数据总览
半年冷数据总览
1个月冷数据明细查询
3个月冷数据明细查询
半年冷数据明细查询
冷数据概况(生成分析数据基于20130720,扫描最近180天的冷数据， 每日进行一次扫描，删除数据后不会立即同步到这次查询结果中,谢谢你的支持！)
冷数据查询页面入口 点这里
云梯冷数据组分布图(top10)
4.61 %
6.29 %
21.26 %
6.68 %
6.84 %
7.25 %
21.26 %
7.97 %
8.90 %
8.94 %
Highcharts.com
冷数据分组详情

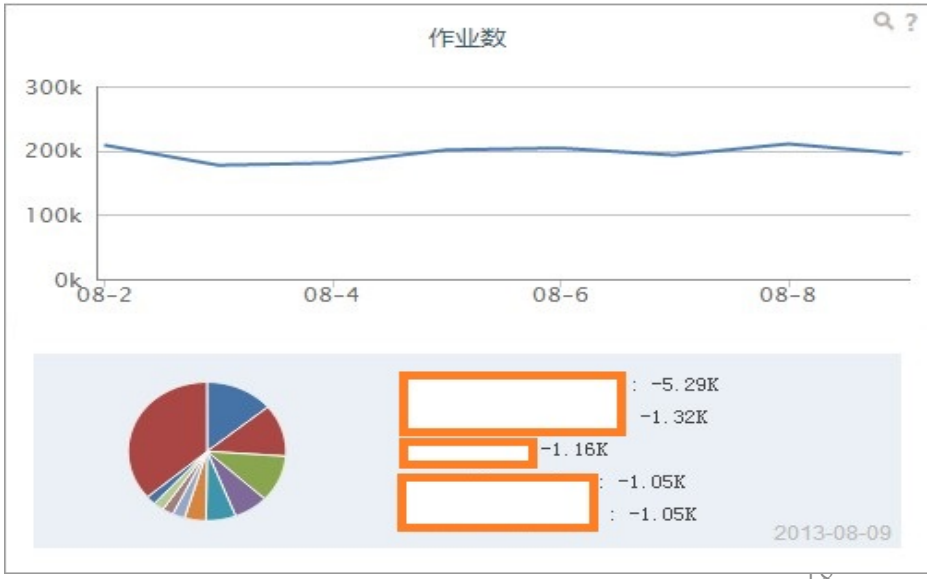

最近两周90天冷数据趋势
670
660
650
640
630
620
610
600
TB
0725
0726
0727
0729
0730
0731
0801
0802
0803
0804
0805
0806
0808
0809
冷数据总量
Highcharts.c

集群数据



- 集群全局指标
  - 存储\计算利用率趋势
- 用户\组资源使用趋势分析
  - Slots*Sec
  - HDFS/Local r/w
- 机器\机器组视图

- 业务作业对比(前一天\前一周)
  - 数据量增长趋势
  - 不同优先级作业资源消耗
- Master节点关键指标
  - JobTracker心跳频率\时间
  - NameNode RPC各项性能指标

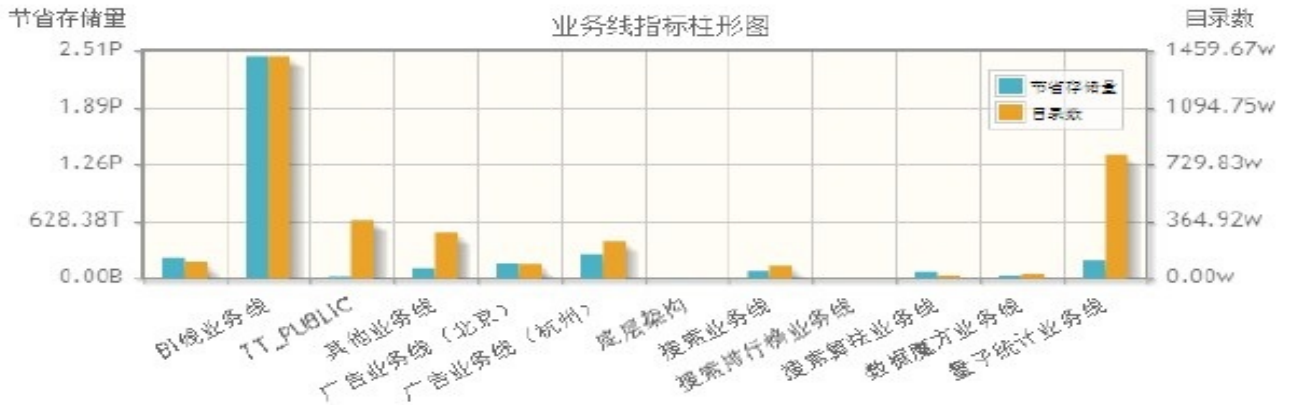


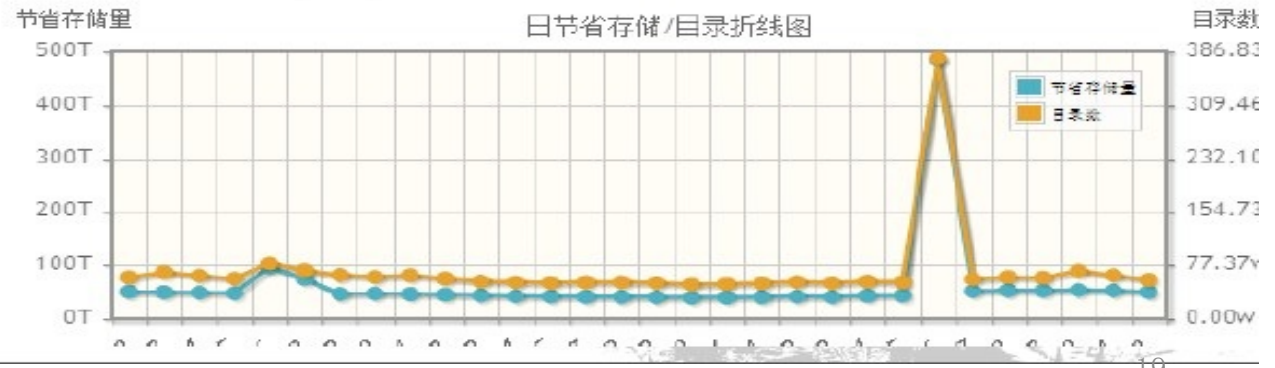
用户数据

云存储 > 仪表盘

策略名	节省存储	目录数
彻底删除	1.29T	1.55w
周期性删除	3.65P	3442.65w

业务线	节省存储	目录数
	230.75T	100.41w
	2.45P	1424.07w
	27.42T	373.69w
	108.15T	291.69w
	156.94T	88.23w
	266.44T	241.57w
	8.11T	0.32w
	81.28T	77.98w
	2.06T	1.77w
	74.67T	19.96w
	35.64T	31.31w
	205.79T	793.21w
总计	3.65P	3444.20w





集群数量快速膨胀遇到的运维压力

1. 服务器硬件配置情况一直在发生变化
2. 大批机器上线某些机器性能不一致
3. Kernel bug
4. 用户数，分组，业务急剧膨胀
5. 突发状况变多，集群突然变慢了？某个组新上线大规模作业？
6. 大压力情况下出现边界效应，小概率事件触发成为常态
7. 目前规模单机房已经无法满足我们需求，跨机房集群该如何运维？
8. 成本，成本，如何控制成本

……

加入我们

欢迎加入阿里巴巴技术保障部门-云计算运维

http://job.alibaba.com/zhaopin/job_detail.htm?refNo=JI002985

我们还在路上，一起改变世界！！！

Q&A