



QCon全球软件开发大会（北京站）2013

即将启动敬请关注

www.qconbeijing.com

全球企业开发大会

INTERNATIONAL
SOFTWARE DEVELOPMENT
CONFERENCE



Alibaba

technology
Association

基于Hadoop的内部 海量数据服务平台

无谓 (吴威)

阿里巴巴集团-海量数据



- 吴威(淘宝花名 无谓)
- 高级技术专家
- 2008年加入阿里：中国雅虎、B2B、阿里云
- 参与Hadoop官方中文文档翻译
- 阿里集团内部Hadoop技术推广者
- 阿里云梯Hadoop服务负责人

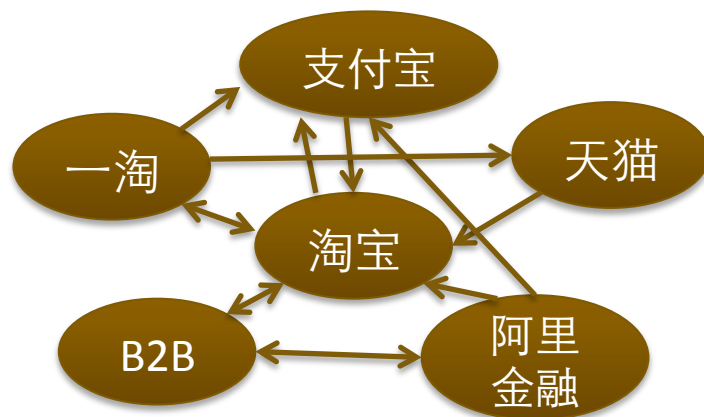
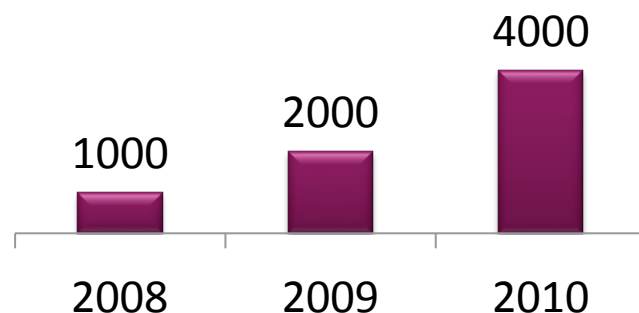


- 大数据...
- Hadoop as a Service
- 问题和挑战
- 我们的对策
- 案例介绍 – 淘宝数据平台
- 未来展望



- 数据的价值
 - 阿里的三个发展阶段: 平台、金融、数据
- 数据增长趋势
 - 用户、商品、交易
- 数据的复杂度
 - 子公司众多
 - 业务逻辑复杂并相互依赖

淘宝交易额(亿)





- 单机
 - 容量，性能
- 分布式数据库
 - Oracle RAC / Greenplum / ...
 - 商业软件(价格)，容量，稳定性
- 分散的Hadoop集群
 - 数据互操作，稳定性，成本和效率



- 云梯
 - 一个项目
 - 一个集群
 - 一项服务
- 为阿里集团提供海量数据的存储和计算服务
- Hadoop集群合并
- GP和Oracle的迁移



Alibaba
technology Association

Hadoop as a Service



- **HDFS - 海量数据存储服务**
 - 分组，通过quota(空间/文件数)限制： /group/taobao
 - 数据共享： 淘宝/天猫/一淘/B2B/支付宝
- **MapReduce - 大规模分布式计算服务**
 - 分组，slot限制，按需申请，集中分配和调度
 - 生产/开发/测试共享集群，白天开发，晚上生产
- **服务特色**
 - 单一大集群
 - 多用户共享
 - 计算分时
 - 按需申请，按使用量计费



Hive

基于MapReduce的SQL引擎

Streaming

可以用任意可执行程序或脚本运行MapReduce

Mahout

机器学习算法库

Pig

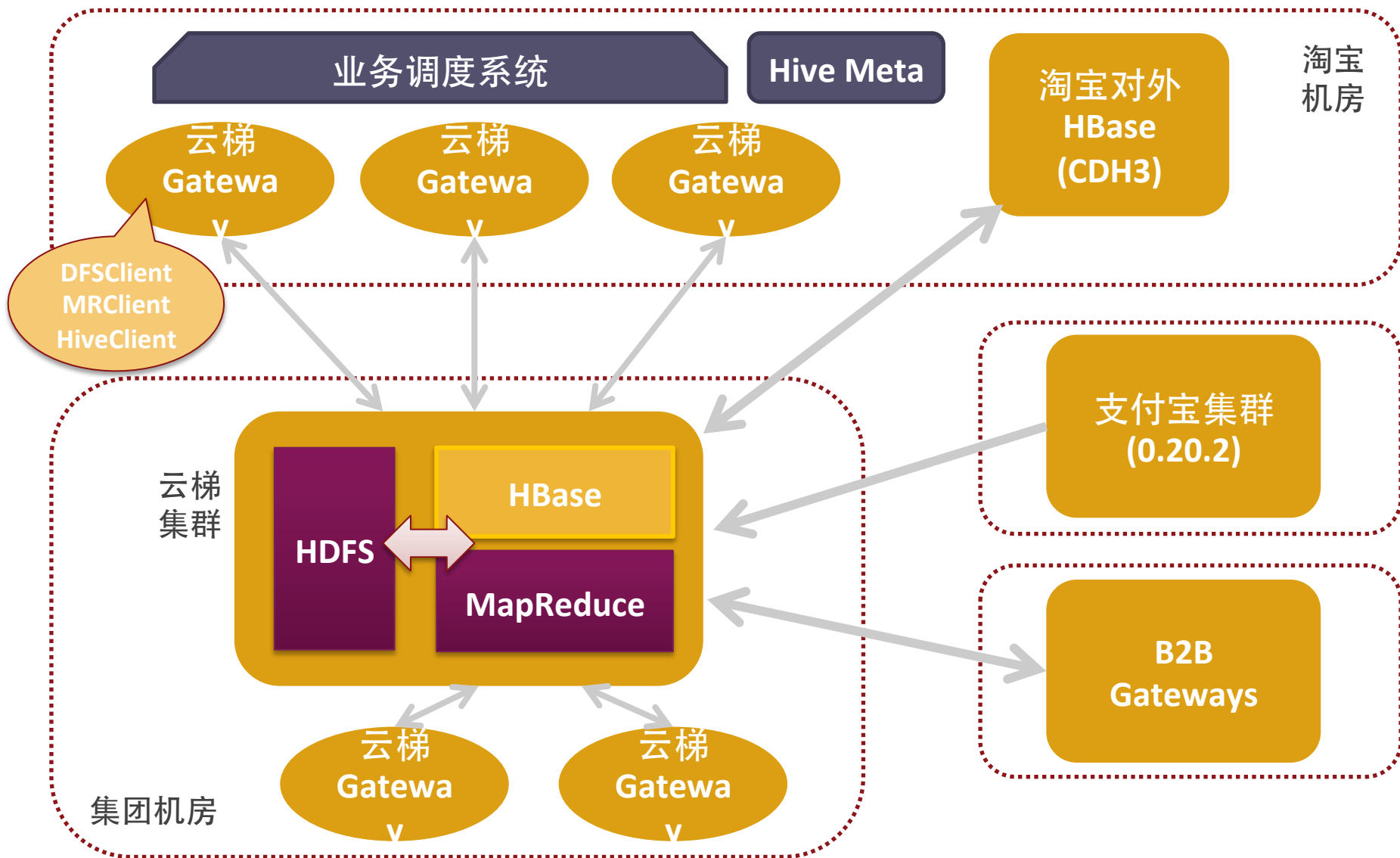
类似于Hive的大规模数据分析平台

HBase

离线和在线存储服务

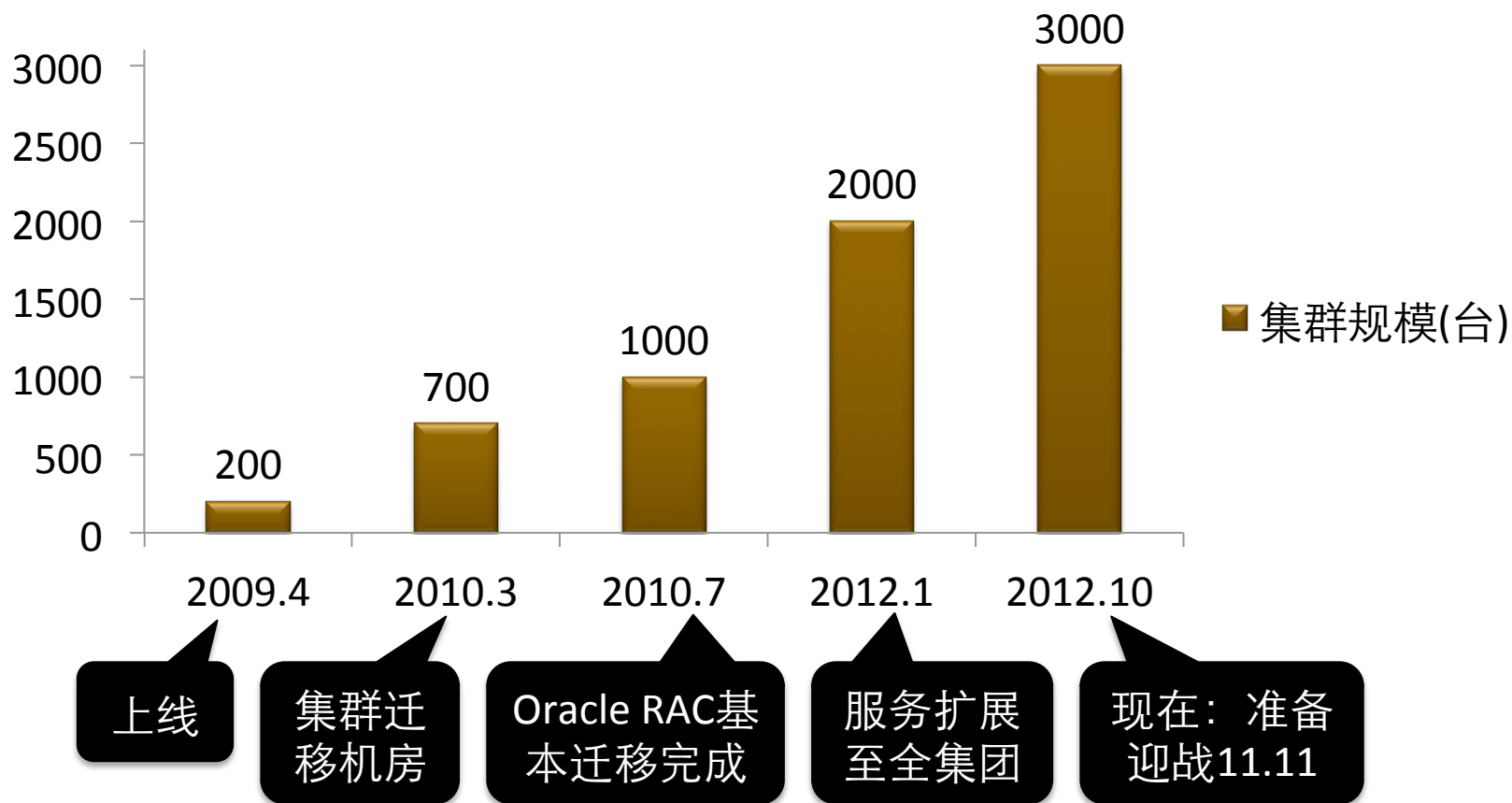


服务基本架构





集群发展历程





集群现状

3000台服务
器

36000核
物理CPU

100TB内存

36000块磁盘

60PB存储容
量 (利用率
80%)

10家子公司

150多用户组

3000多用户



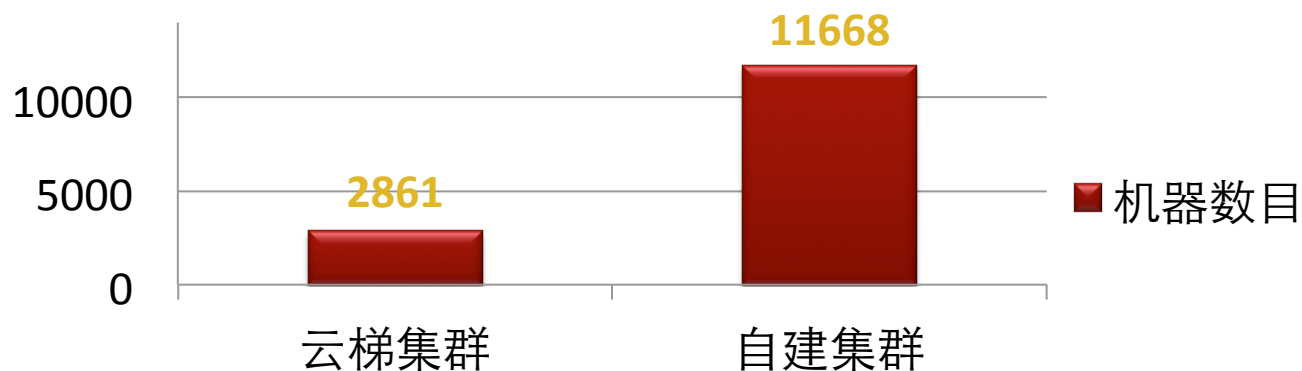
- 从用户角度出发

	自建Hadoop集群	使用云梯Hadoop服务
集群搭建	机器采购，机房布局	不用考虑
集群运维	节点宕机后需要立即介入	不用考虑
集群扩容	计算资源不足，存储空间不足，需要扩容，采购新的机器	网页上申请，即日生效
代码维护	Hadoop代码维护，专业的Hadoop开发人员	不用考虑
数据复制	使用其他团队数据，需要从另外集群复制过来	阿里集团大部分数据能在云梯上找到



平台服务的成本优势

资源组	高峰时段slot申请量	自建集群需要机器数
cug-groupA	31000	1550
cug-groupB	7500	375
cug-groupC	5500	275
cug-groupD	4500	225
cug-groupE	4000	200
cug-groupF	4000	200
其他100多个组	176860	8843
总计	233360	11668





Alibaba
technology Association

问题和挑战



- 稳定性和安全性
 - 大作业占用集群的所有slot(计算资源)
 - 某些机器网卡打满
 - NameNode被某个用户的作业拖慢
- 共享
 - 计算资源共享: A组在白天用, B组晚上用
 - 数据共享: 支付宝读取淘宝的某张表数据, 怎么开放?



- 兼容性
 - 上千个客户端/Gateway, 上百个部门
 - 客户端全量升级代价大
 - 服务器端升级要尽量保持向下兼容
 - 有些客户端是Hadoop 0.19.0, 有些是Hadoop 0.20.2或者CDH3版



- 性能和扩展性
 - Hadoop Master节点是单点
 - NameNode RPC请求压力：2亿文件 + 2.6亿block，日请求量超过15亿次
 - JobTracker调度压力：日调度运行超过15万个Job，7500万个Task，高并发(1000+ jobs, 55000 tasks)，多用户(3000+)
 - JVM的极限, 超过100G的JVM Heap
 - 单点故障



- 可观测和可测试
 - 上千台机器，多个master
 - 上百个指标：系统，Java GC，Hadoop metrics...
 - 集群突然变慢了？某个组新上线大规模作业？
 - 大压力情况下出现bug了！
 - 每个季度都有新版本发布，版本性能是否有提升？



Alibaba
technology Association

我们的对策



- 重构Task Scheduler (基于Fair Scheduler)
 - 资源组的划分: 消除某些组的大作业对其他组的影响
 - 计算资源动态管理(create/delete/increase/decrease)
 - 完整的作业优先级支持: 支持天网调度
 - 对异构操作系统或硬件的兼容性: 比如支持跨OS版本调度



- 传统Unix文件系统权限
 - User/Group/Other, rwxr-x---, 750
 - Hadoop 0.17已经实现
 - 数据组内(同一子公司的用户)可读, 但外部用户不可读
- 跨组, 跨部门, 跨公司文件共享
 - 新功能: 扩展ACL
 - ACL条目: /group/taobao/auctions:b2b_dw:+R:tbclient:
+RW
 - PermissionChecker: 如果UGO检查失败, 则调用扩展ACL
 - 外部系统:
 - 资源注册, 权限申请, 权限审批, ACL条目同步



- 消除异常Job的影响
 - 内存监控: 单个Task内存限制, 计算节点内存上限控制
 - 磁盘IO监控: 单个Job shuffle线程对单块磁盘的读取限制
 - 限制单个Job map/reduce task数目
 - 限制单个Job counter数目
 - Job本地文件系统数据读写量监控
 - Job创建HDFS文件数目的监控



- 服务可持续发展

1. Hadoop Server: 云梯Hadoop (基于Apache Hadoop 0.19.1)
2. Hadoop Client:
 - 0.19.x: 公司内大规模部署, 几百个Gateway
 - 0.20.x: 社区主流版本, Hadoop生态圈支持
3. 社区Hadoop版本(0.23之前)不支持跨版本的兼容
4. Hadoop 0.19和0.20代码库差别不大
5. 0.20重要的新增feature都在云梯Hadoop中实现:
 - A. HDFS Append
 - B. MapReduce new API

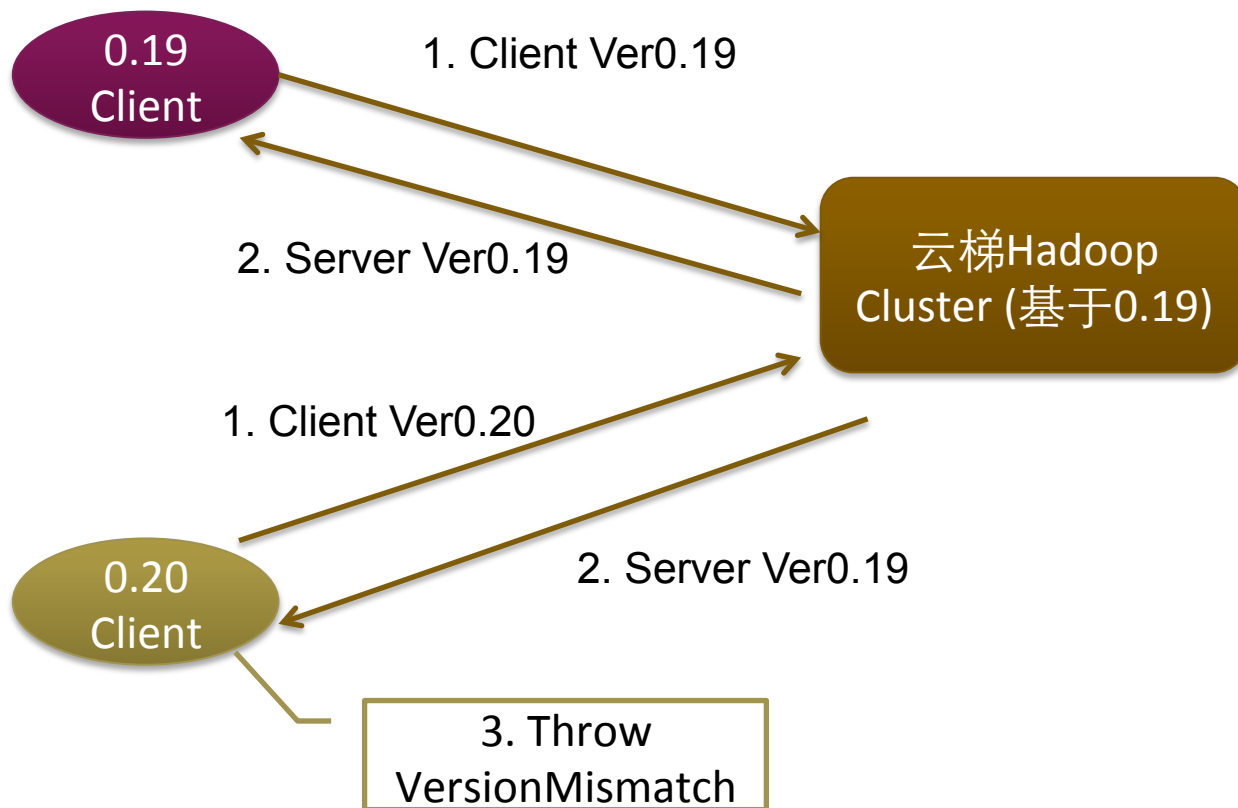


- Hack Hadoop protocols (Hadoop 0.20 vs. 0.19)

协议	0.19版本号	0.20版本号	改变
RPC protocol	2	3	Change RPC connection header (Request带版本信息)
HDFS ClientProtocol	40	41	NEW FEATURE saveNamespace(DFSAdmin)
HDFS DataTransferProtocol	14	14	No change
MapReduce JobSubmissionProtocol	16	20	NEW FEATURE getClusterStatus(boolean)

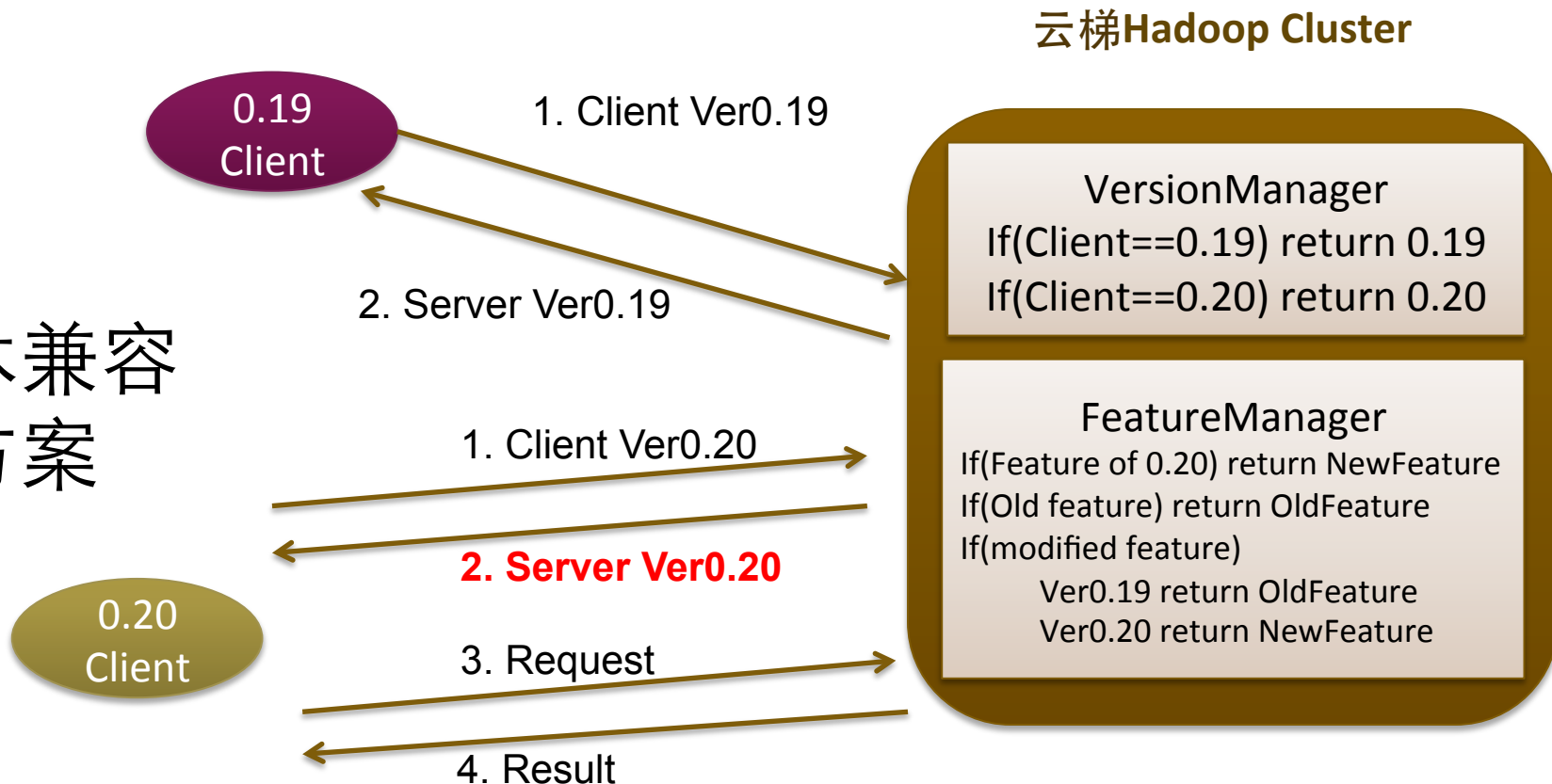


之前





版本兼容 方案

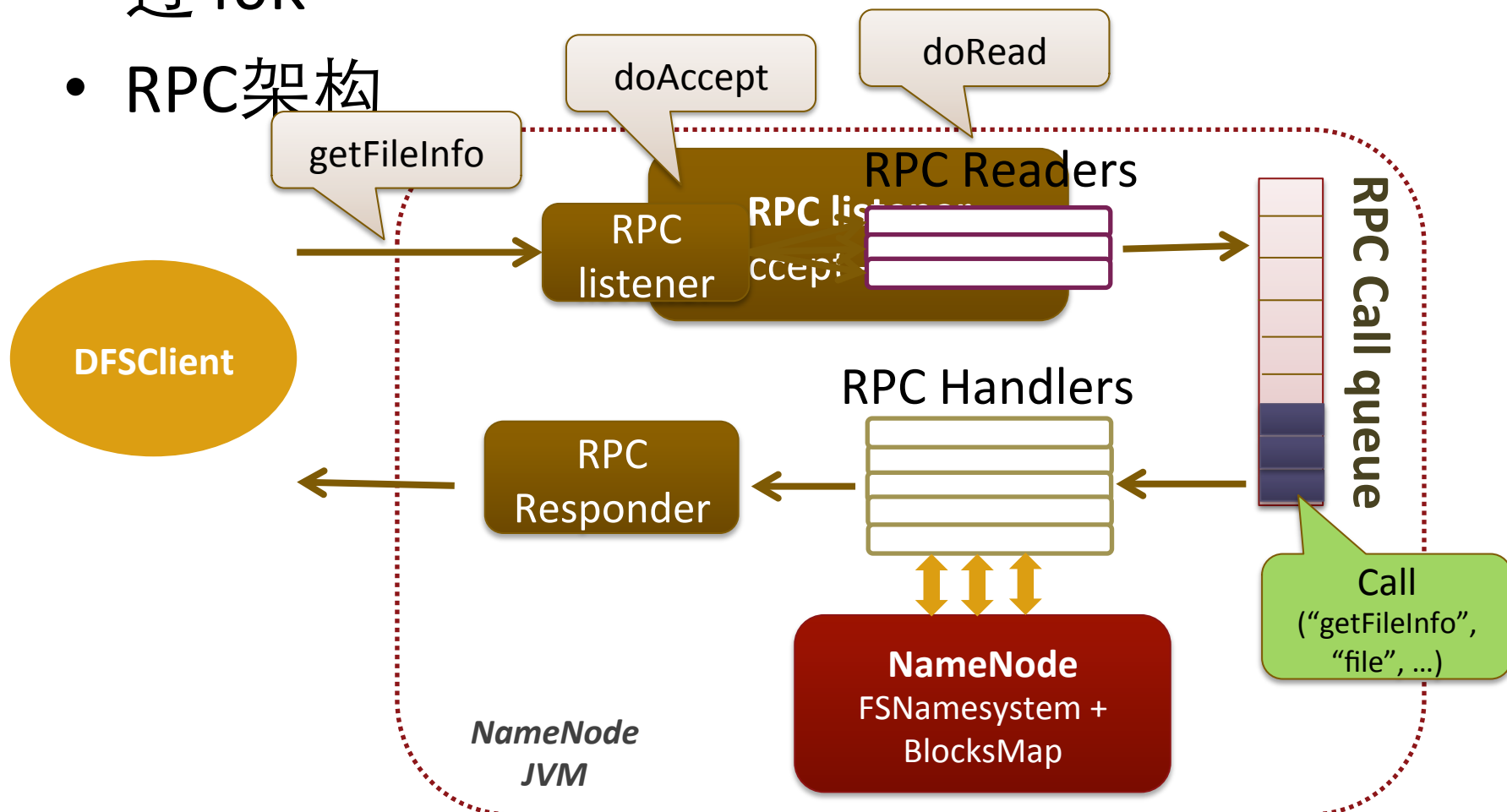




- 性能：解决Master节点的单点性能压力
 - **NameNode** 改进
 - RPC reader改造
 - 使用读写锁，尽可能的提高NameNode内部的并发
 - **JobTracker** 改进
 - Scheduler调度算法重写
 - 一次心跳分配多个Task
 - Out-of-bound heartbeat提高调度的效率



- NameNode RPC平均每秒请求15K，高峰时超过40K
- RPC架构





- 全局锁（FSNamesystem 对象）
 - RPC handler 多线程串行化
- 读操作：
 - getFileInfo, getBlockLocations, getListing 等
 - 调用次数占75%以上
- 写操作：
 - create, delete, mkdir, rename, addBlock等
- 读写锁改造 ReentrantReadWriteLock
 - RPC handler 并行度提高



- Heartbeat RPC 调用返回的 <Task List>:

Hadoop 0.19版本

一个Map Task 或 一个Reduce Task, Map优先

问题: 调度大量 Map 时, Reduce 饿死

优化 1

一个Map + 一个Reduce

对策: 每个 job
配置reduce
slow start 参数

问题: Reduce提前启动, 占用大量 Slot

优化 2

多个Map + 多个Reduce

对策: 设置非
local task 调度
等待时间

问题: task locality下降, data locality/rack
locality



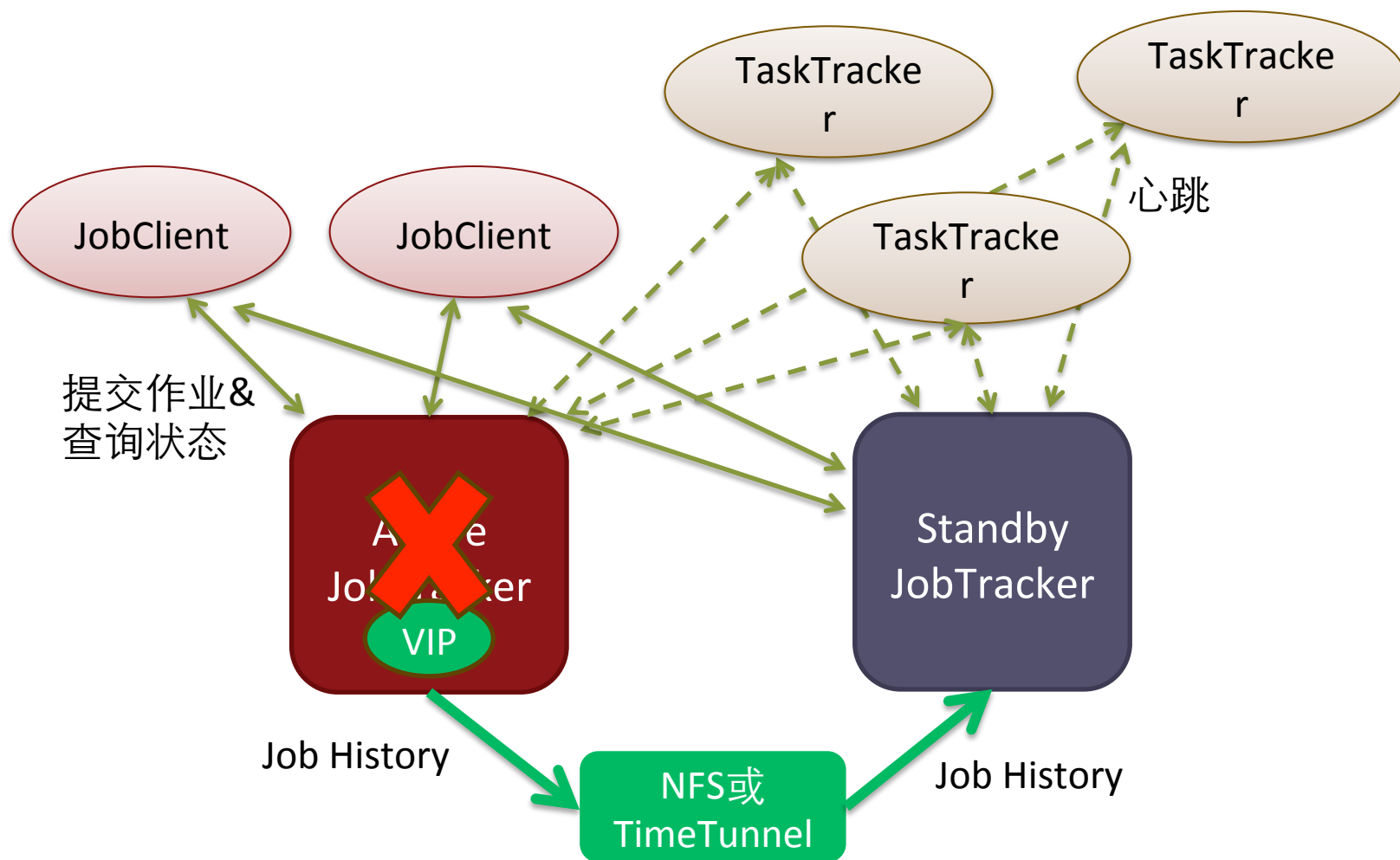
- Scheduler 算法
 - 从 job 队列中选取一个job的task运行
 - 多个 Job queue; Job优先级
- 远古时代: $O(n^2)$
- 优化 1: $O(n)$
- 优化 2: $O(1)$
 - 异步线程更新 job的状态和优先级, 得到一个排序后的列表
 - 每次 heartbeat 调用, 选取列表中的第一个 Job

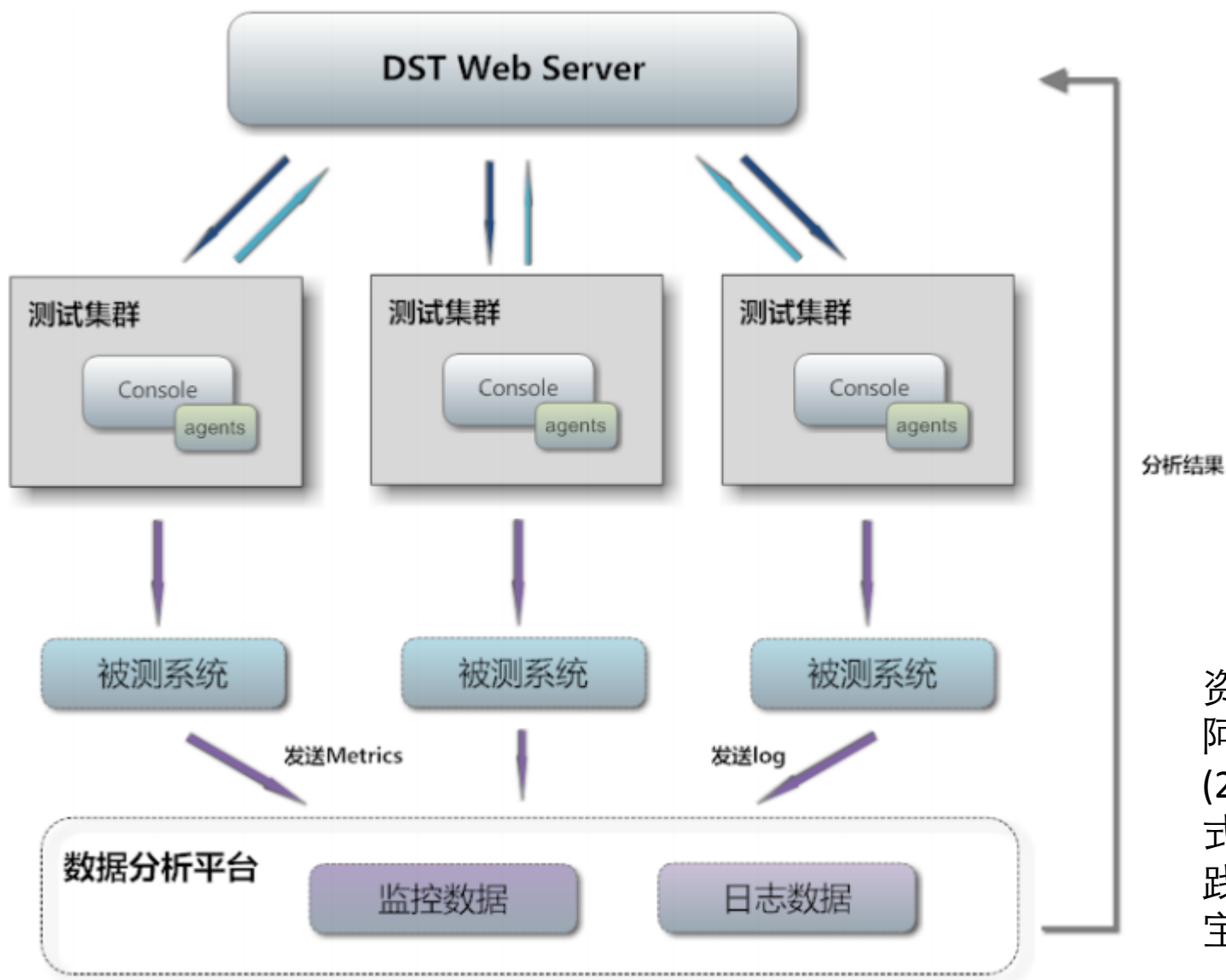


- NameNode内存泄露
 - NameNode高并发RPC
 - Java nio SocketAdapter创建的SocksSocketImpl对象需要finalize，但在CMS gc回收不及时
 - Oracle JDK bug ID: [7115586](#) (Oracle JDK 6u32)
- CMS gc使用135G的Heap后JVM crash
 - NameNode大内存
 - $1 \ll 32$ 移位操作溢出
 - Oracle JDK bug ID: [7197906](#) (OpenJDK)



JobTracker HA实现

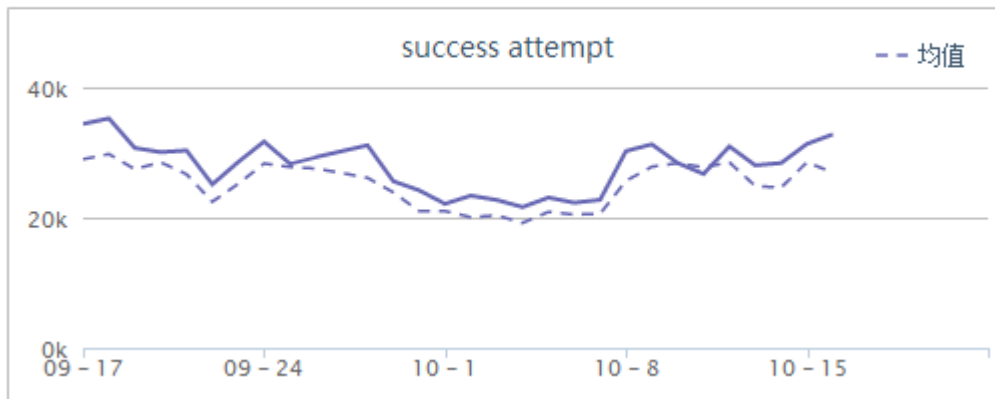




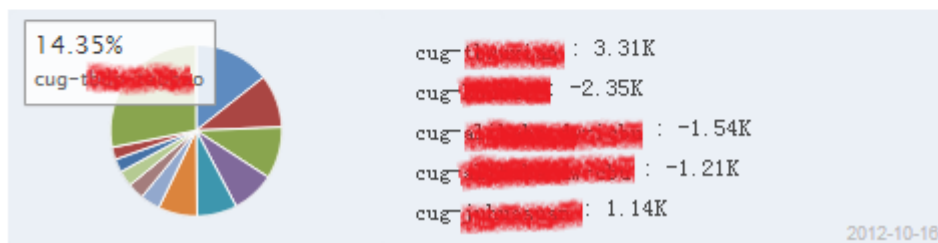
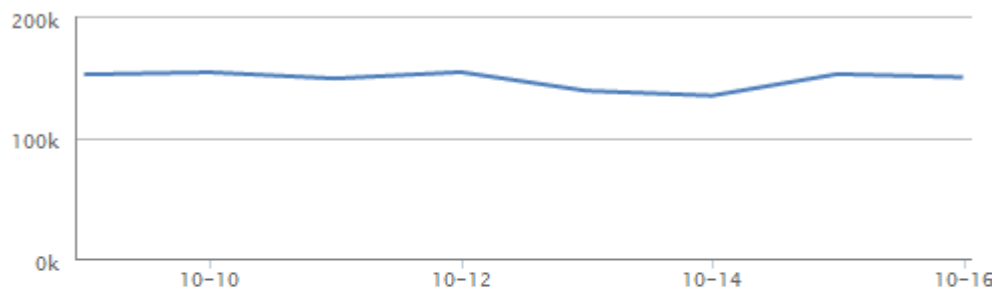
资料来源：
阿里技术嘉年华
(2012) - 《分布
式系统测试实
践》 - 神秀 (淘
宝网)



r01b05028.aliyun.com 指标图



作业数





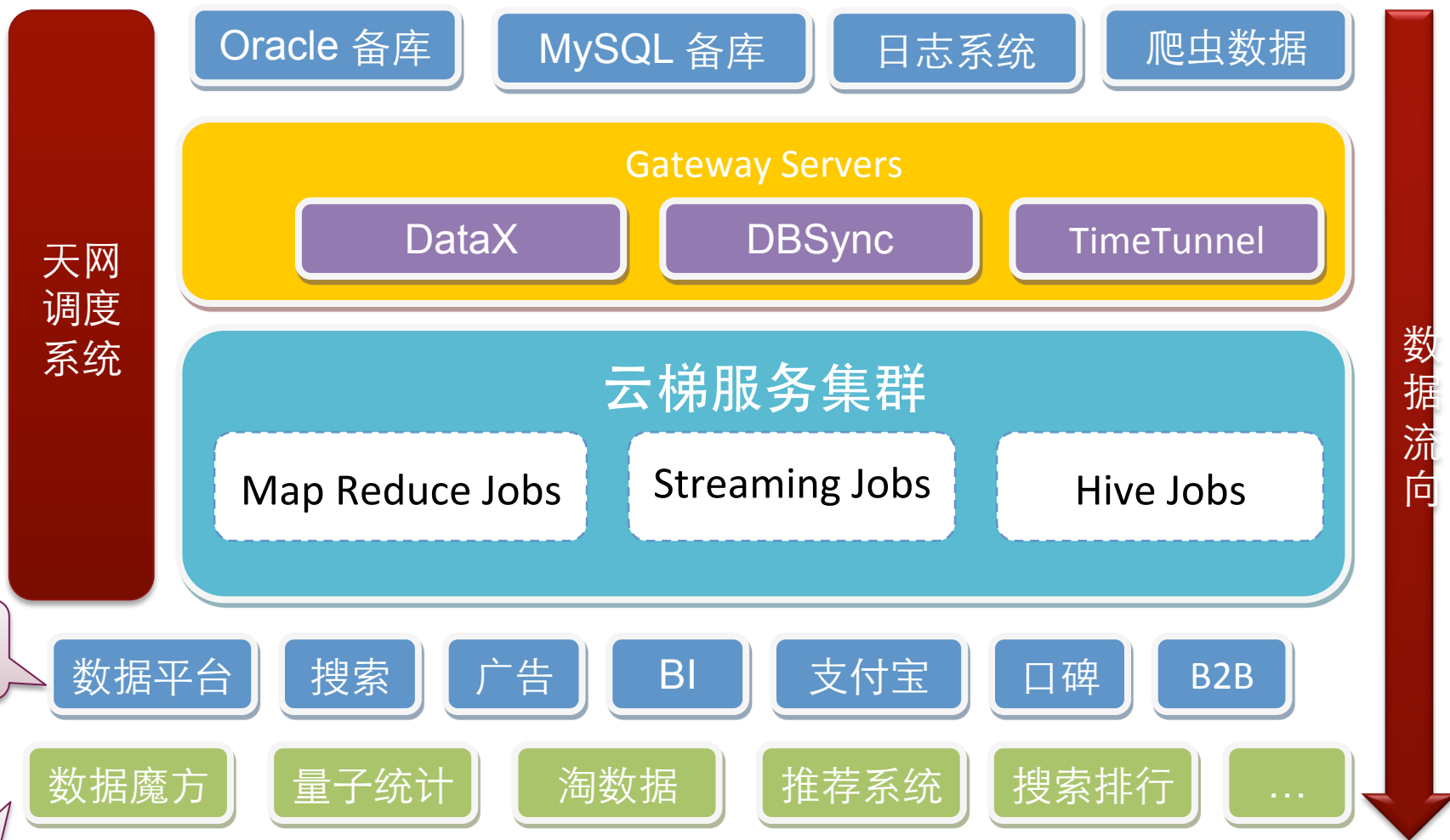
Alibaba
technology Association

案例分享



Alibaba
technology Association

基于云梯的淘宝数据 平台架构



资料来源: Velocity China 2010 - 《淘宝云梯分布式计算平台整体架构》 - 张清(淘宝)



- 数据如何流入云梯?

- 日志数据:

TimeTunnel, 分布式日志收集工具

- 数据库表:

DataX, 前台数据库<=>云梯 (双向同步)

DBsync, 增量, 大表的快速同步

- 其他数据来源

- 来自其他团队和公司的数据, 比如支付宝数据, 广告反作弊数据, 通过云梯共享

- 数据流出

- 前台业务系统, 如传统数据库或NoSQL (主要是HBase): **DataX ...**

- 在云梯上共享给其他团队和公司, 做进一步分析



数据处理使用 MapReduce和Hive

计算内容	处理方式
ETL数据分析处理，OLAP 大数据量分析场景	主要使用Hive
点击流日志分析	MapReduce批量处理
搜索排行榜和其他搜索 相关业务	大量使用C/C++算法库，分词库， 利用MapReduce Streaming或Pipes
机器学习	可以使用Mahout



- Gateway管理
 - 提交Hadoop Job
 - 运行数据导入导出任务
- Job优先级管理
 - hadoop.job.level: 天网利用云梯Job Scheduler开发的接口，完整的优先级支持
 - 云梯scheduler的特点：
 - 资源空闲时，低优先级作业可以运行
 - 后提交的高优先级作业立即占用低优先级作业释放的资源
- 监控报警管理



- 数据分析
 - Hive SQL Web IDE
 - 帐号和云梯服务集成
- 知识管理
 - 元数据/数据字典/数据订阅/表字段血缘分析
- 存储管理
 - 数据生命周期管理
 - 数据保留策略： 周期性删除/极限存储/压缩/HDFS Raid



项目 公共表 个人表 函数

建表 上传 下载 删除

搜索

表名	大小(M)	建表时间
s_audit_log_ww	0.0000	2012-01-21

1-1 共1

列信息 分区信息 数据预览

分区

pt=20120221

pt=20120120

新建 保存 运行 运行并保存 停止 格式化 撤销

auditlog* x run-auditlog* x

```
1 --alter table s_audit_log_ww add partition (pt='20120221')
2 add jar /home/hive/hive/lib/hive_contrib.jar;
3 select * from s_audit_log_ww where pt = '20120221' and s:
```

Log selec...log%'

	time1	level	class	ugi	ip	cmd	src
02-21	01:50:26,104	INFO		taobao,cug-...	/119.42.229.18	create	/group
02-21	01:50:26,158	INFO		taobao,cug-...	/119.42.229.18	rename	/group
02-21	01:51:15,044	INFO		taobao,cug-...	/119.42.229.14	listStatus	/group
02-21	02:04:04,080	INFO		taobao,cug-...	/172.24.208.63	rename	/group
02-21	02:04:46,317	INFO		taobao,cug-...	/10.249.64.13	open	/group
02-21	06:05:54,075	INFO		taobao,cug-...	/10.249.38.17	open	/group

1-1 共1

下载全部结果 复制选中行 复制选中列 复



Alibaba
technology Association

未来展望



- 服务类型扩展
 - 支持多种计算模型，比如MPI/Storm等，超越Hadoop MapReduce (Hadoop 2.0 Yarn)
 - 更好的资源控制和计费，利用cgroup等 (基于Hadoop 2.0 Yarn)
 - 云梯成为阿里集团大规模计算/存储资源管理中心
- 期望和开源社区结合更加紧密



- 服务质量提升
 - Master节点HA
 - NameNode HA (Hadoop 2.0)
 - 做到不停机升级，加快软件的进化速度
 - Hive实时化
 - M/R调度性能的深度优化
 - 结合HBase或索引等相关技术



- 性能和扩展性
 - M/R Shuffle性能优化
 - 利用操作系统的底层性能优化 (Linux内核团队)
 - 利用JVM的性能优化 (淘宝JVM团队)
 - NameNode扩展性
 - Federation
 - MapReduce扩展性
 - Yarn
 - 支持跨机房 (当前机房上限是4000 - 6000台)



Alibaba
technology Association

Q&A



QCon

QCon全球软件开发大会（杭州站）2012

讲师资料下载