



Alibaba

technology
Association

基于Hadoop的内部 海量数据服务平台

罗李（吴威）

阿里数据平台-海量数据



- 阿里大数据发展历程
- 云梯Hadoop服务集群
- 问题和挑战
- 我们的对策
- 案例介绍 – 淘宝数据平台
- 未来展望



- 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
- 为何选择 Hadoop ?
 - MapReduce 和 HDFS 能满足大部分离线业务的需求
 - 商业公司 Yahoo / Facebook 支持，工业级应用
 - 可扩展，大规模
 - 开源软件，社区活跃



Alibaba
technology Association

云梯Hadoop服务集群



- **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
 - 资源按需申请，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

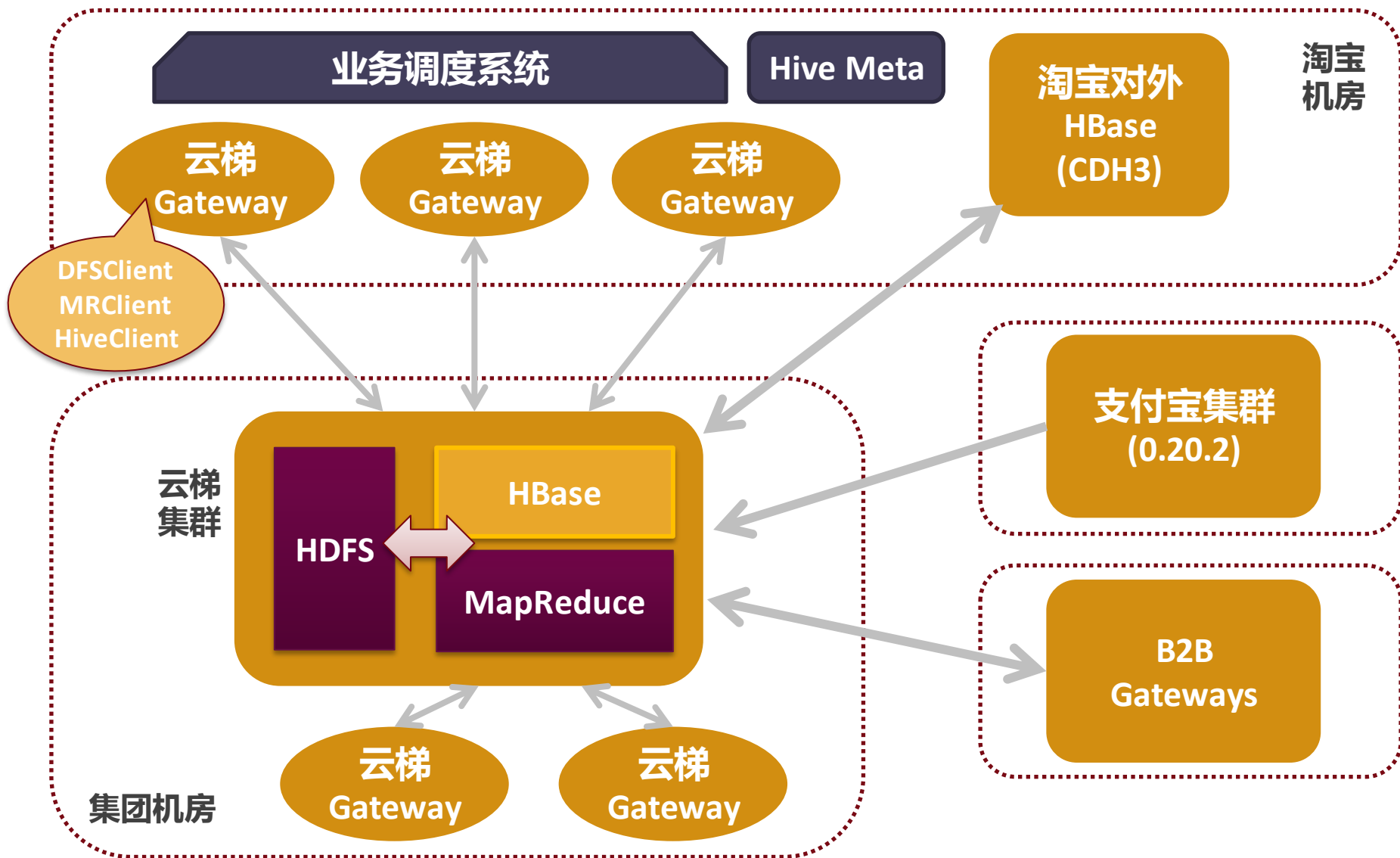
类似于Hive的大规模数据分析平台

HBase

离线和在线存储服务

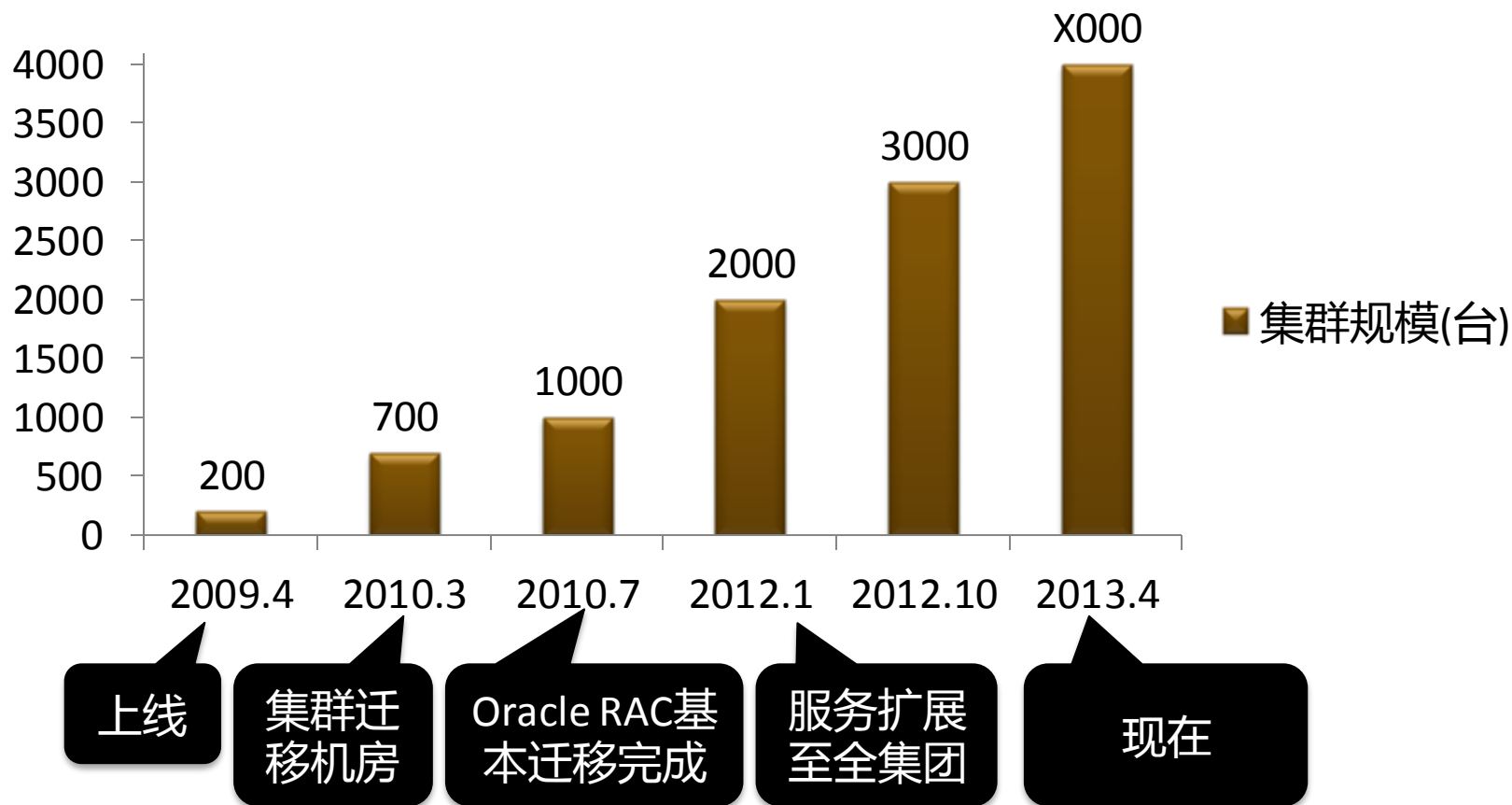


服务基本架构





云梯集群发展历程





Alibaba
technology Association

云梯集群现状

20多个事业
群

150多用户组

3000多用户



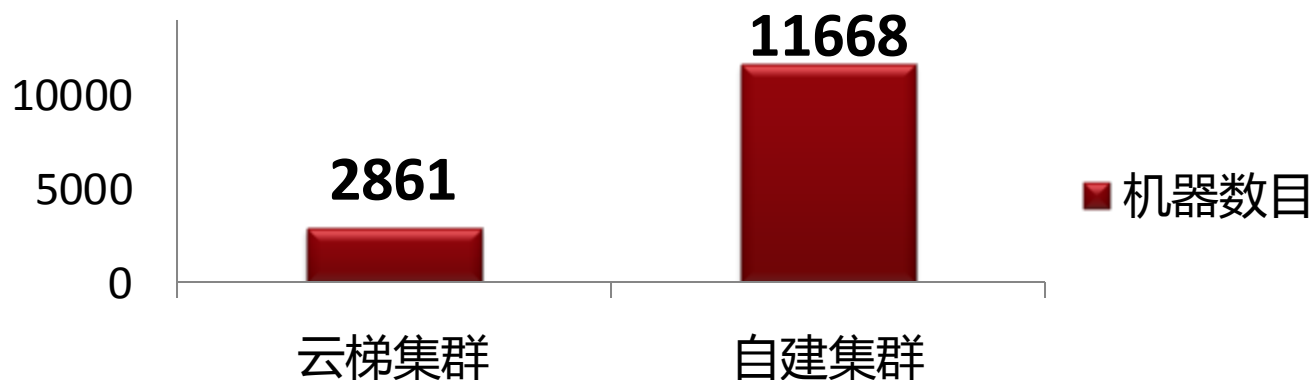
- 从用户角度出发

	自建Hadoop集群	使用平台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	网页申请，审批通过即可生效
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
 - 客户端版本：
 - Hadoop 0.19.0
 - Hadoop 0.20.2
 - CDH3



- 性能和扩展性
 - Hadoop Master节点是单点
 - NameNode 压力：几亿文件 + 几亿 Block，RPC 日请求量超过几十亿次
 - JobTracker 调度压力：日调度运行超过几十万个 Job，几千万个 Task，高并发 (1000+ jobs, 55000 tasks)，多用户 (3000+)
 - JVM的极限，超过 150G 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调度器
 - 资源组的划分: 消除某些组的大作业对其他组的影响 (Min slots vs. Max slots)
 - Slot 资源动态管理 (create/delete/increase/decrease)
 - 完整的作业优先级支持: 支持业务优先级调度
 - 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
 - Apache Hadoop 0.17 已经实现
 - User/Group/Other, rwxr-x---, 750
 - 数据组内可读, 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/hive/auctions: **alipay:+R:tbclient:+RW**
 - 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
 -



- 现状

- Hadoop Server : 云梯 Hadoop (基于Apache Hadoop 0.19.1)
- Hadoop Client :
 - 0.19.x : 公司内大规模部署 , 几百个Gateway
 - 0.20.x/1.0.x : 社区主流版本 , Hadoop生态圈支持

- 方案

- 实现0.20上的新增重要功能
 - HDFS Append
 - MapReduce new API
- Hack Hadoop协议 , 服务可以同时支持多个客户端
 - 0.19.x , 0.20.2 , CDH3uX



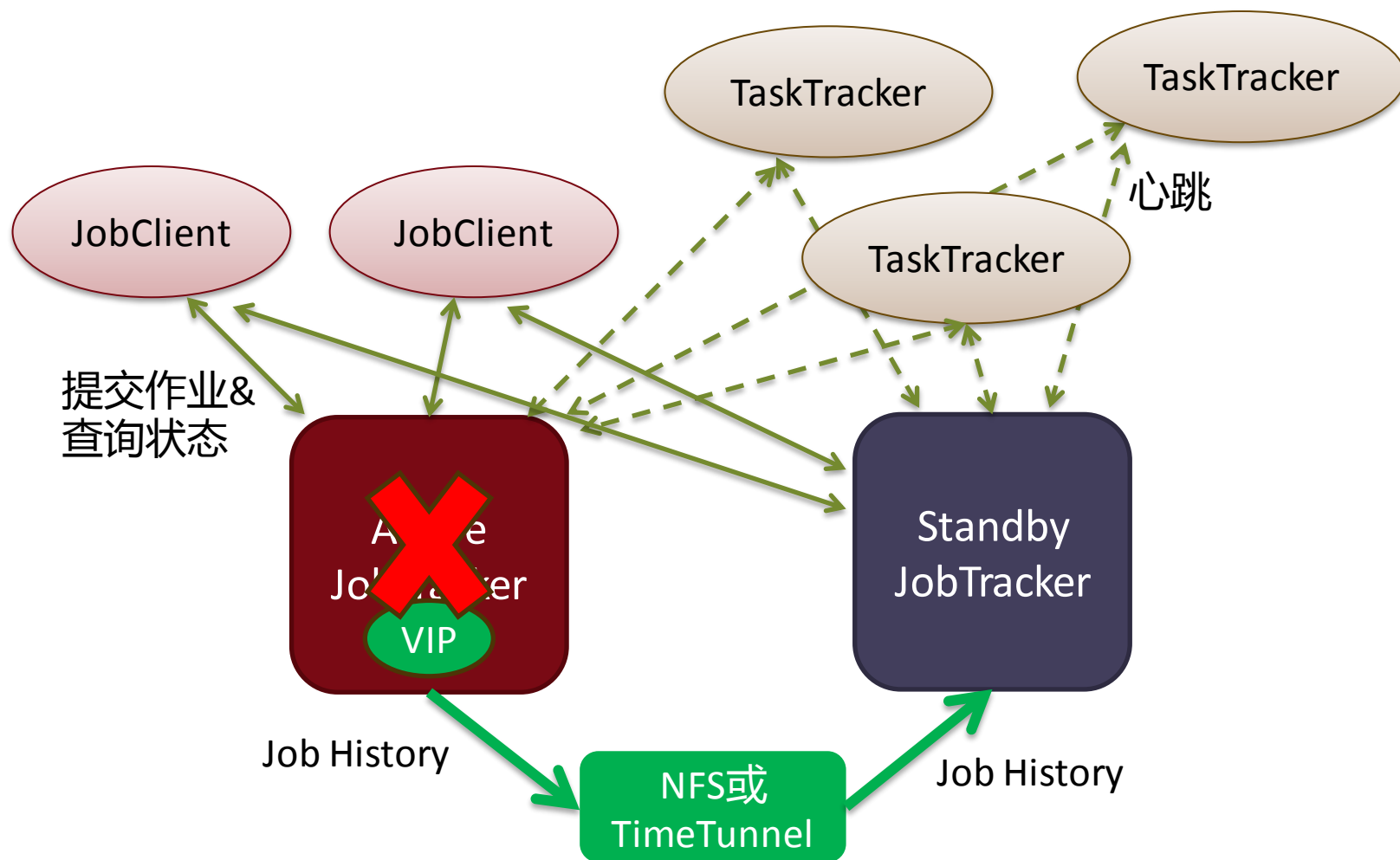
- 性能：解决Master节点的单点性能压力
 - **NameNode 改进**
 - RPC 改造，Listener 拆分出多个 Reader
 - 使用读写锁，尽可能的提高NameNode内部的并发
 - 写操作在等待 edit log commit 阶段时释放 handler
 - **JobTracker 改进**
 - Scheduler调度算法重写，从 $O(n^2)$ 降低到 $O(1)$
 - 一次心跳分配多个Task
 - Job History log 改造成异步写
 - Out-of-bound heartbeat提高调度的效率



- 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 fix)
- 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)

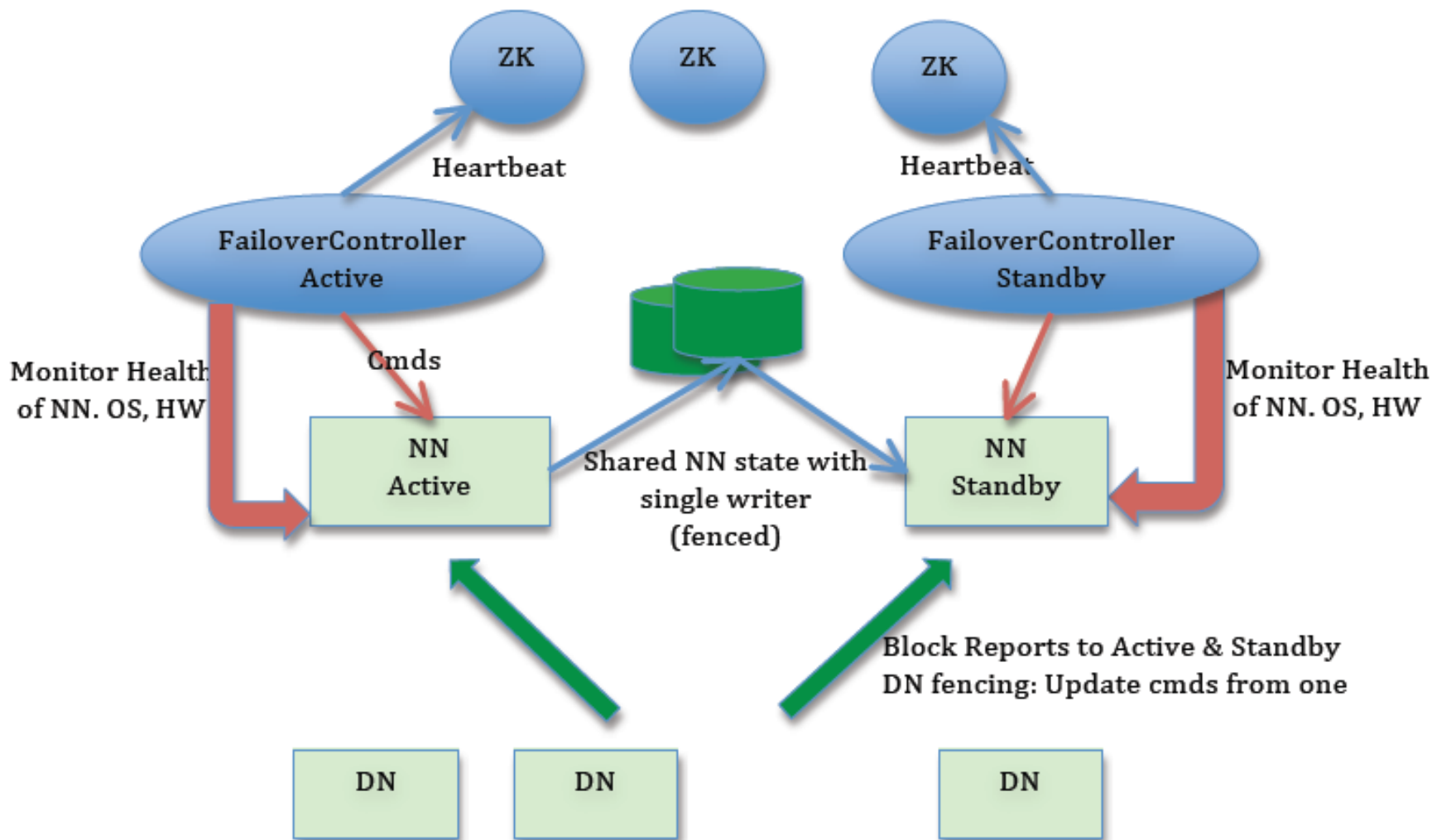


JobTracker HA实现





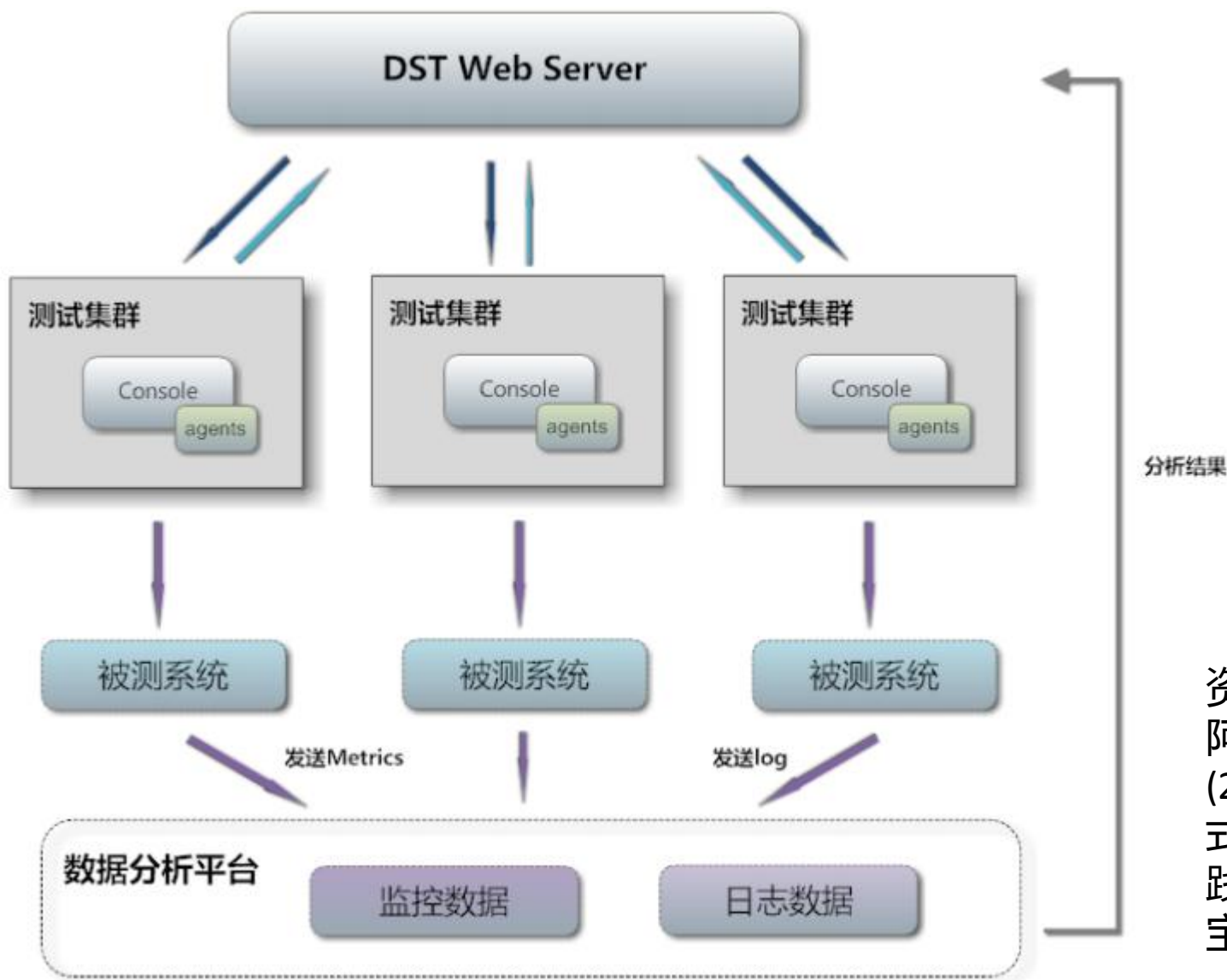
NameNode HA实现





Alibaba
technology Association

DST: 分布式系统测试工具

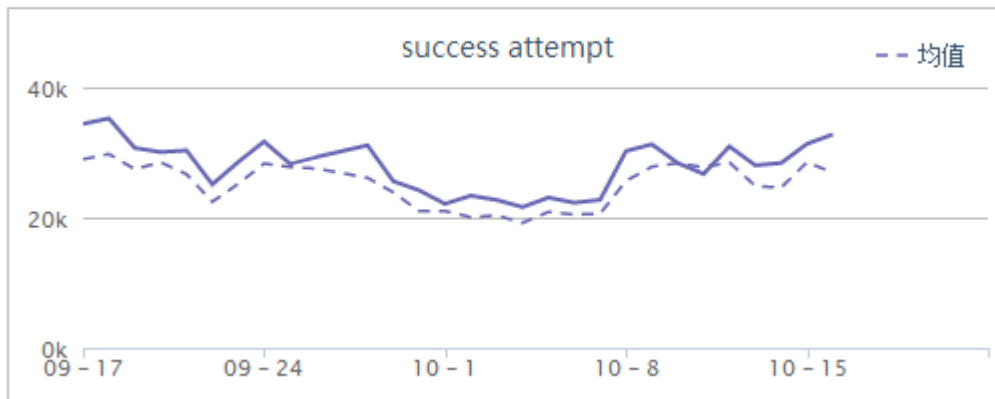


资料来源：
阿里技术嘉年华
(2012) - 《分布
式系统测试实
践》 - 神秀 (淘
宝网)



云梯医生: 集群诊断系统

r01b05028.aliyun.com 指标图



- 业务作业对比(vs. 前一天/前一周)

- 数据量增长趋势
- 不同优先级作业消耗的资源

- Master节点关键指标

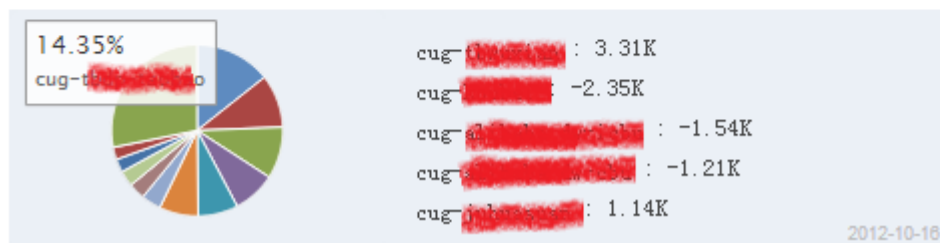
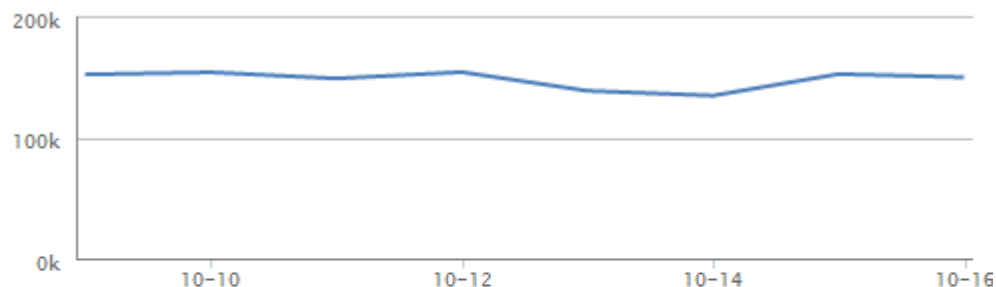
- JobTracker每秒心跳频率/时间
- NameNode RPC process time, queue time, queue len, OPS

- 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

作业数

Q ?



2012-10-16



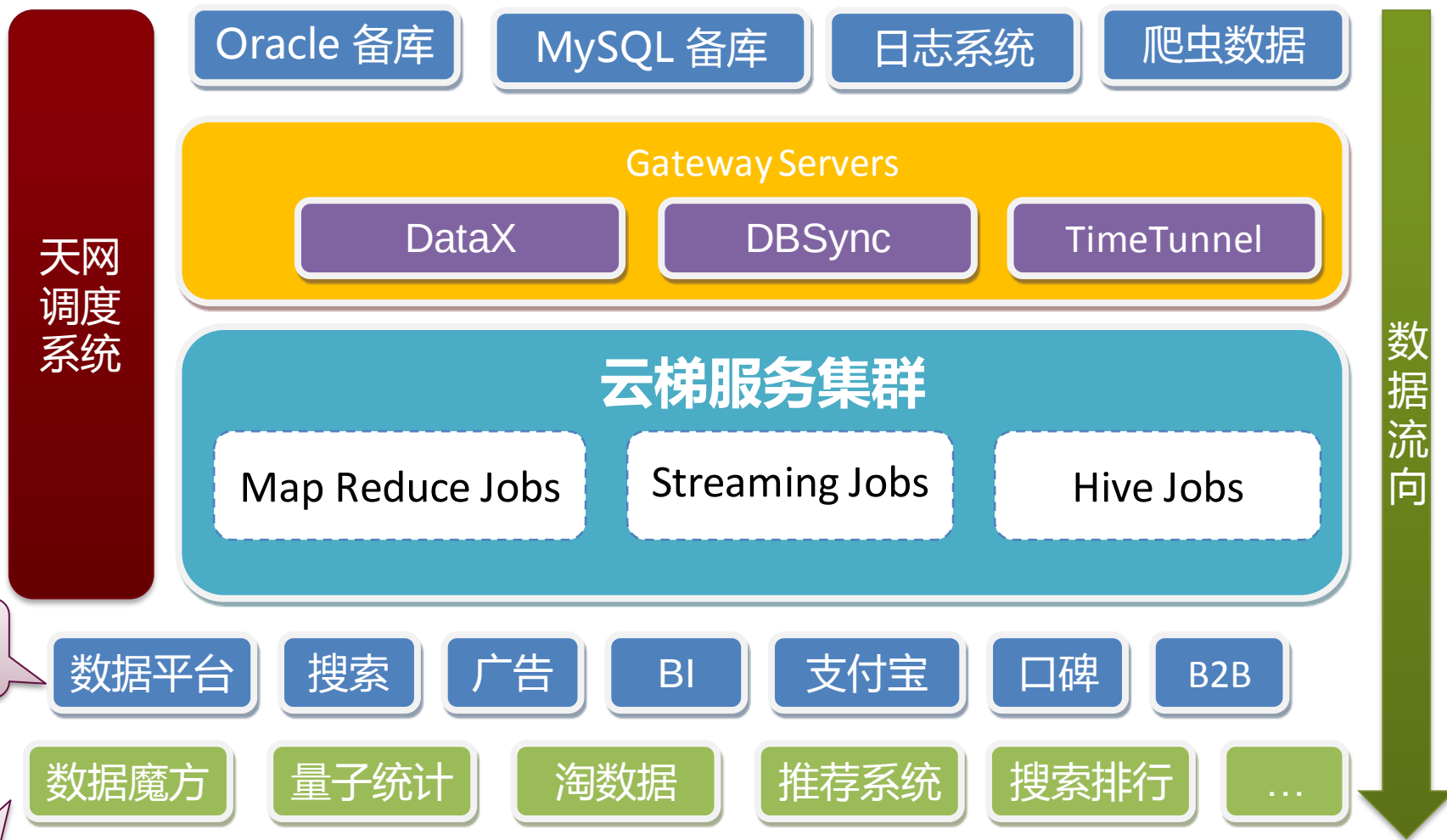
Alibaba
technology Association

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)



- 数据流入

- 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



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	使用Mahout



- Gateway管理
 - 提交Hadoop Job
 - 运行数据导入导出任务
- 作业优先级管理
 - `hadoop.job.level`: 利用云梯作业调度器开发的接口，完整的优先级支持
 - 云梯作业调度器的特点：
 - 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



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，隔离和计费，利用cgroup等(基于Hadoop 2.0 Yarn)
- 期望和开源社区结合更加紧密



- 服务质量提升
 - 继续推进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
 - 安全性
 - Hive表的权限控制，对MR/Pig程序的等访问控制
 - Hive表字段级别的访问控制



- 性能
 - 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)
- 扩展性
 - NameNode扩展性：Federation
 - MapReduce扩展性：Yarn
 - 支持跨机房 (当集群规模渐渐到达机房上限...)



Alibaba
technology Association

Q&A