

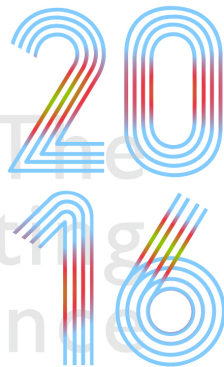



2016 杭州·云栖大会
THE COMPUTING CONFERENCE

云栖社区
yq.aliyun.com

Hadoop的过去现在和未来

——从阿里云梯到E-MapReduce



主办单位： 杭州

 Alibaba Group
阿里巴巴集团

战略合作伙伴：

无谓（高级技术专家）
阿里云-数据库技术组-EMR



扫码观看大会视频

关于我（吴威，花名无谓）

- 2008年加入阿里，搜索技术中心分布式计算团队
- 2009年，阿里云数据平台，云梯Hadoop集群开发和维护
- 2014年，ODPS（MaxCompute），性能和稳定性
- 2016年，阿里云E-MapReduce



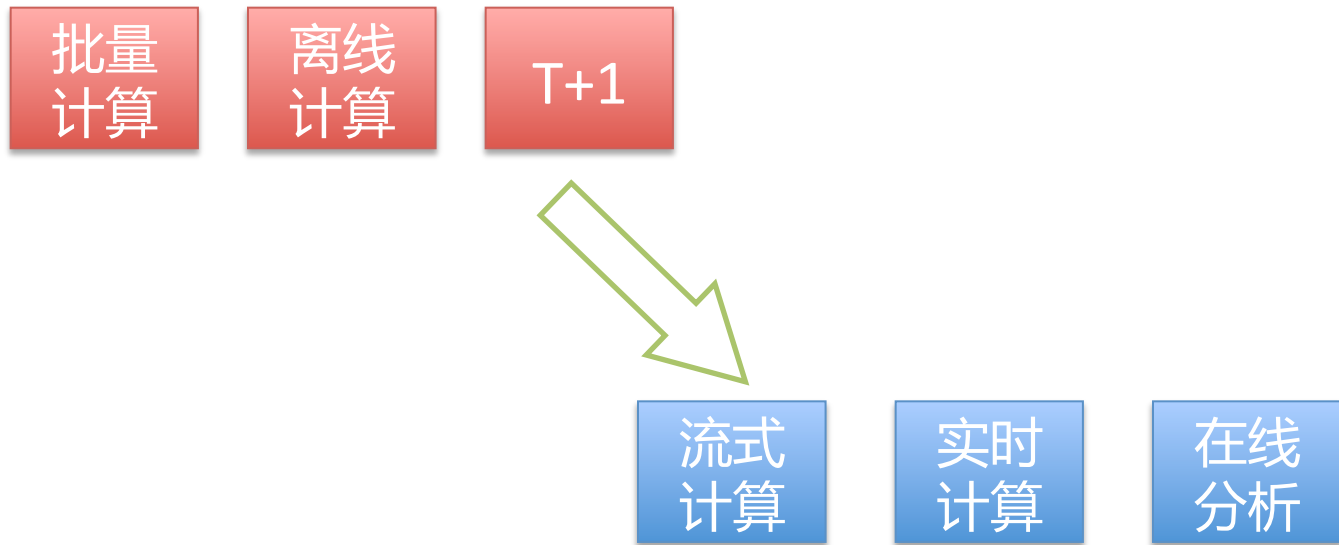
目 录
content

1. Hadoop 10年发展历程
2. 阿里集团的Hadoop之路
3. 阿里云E-MapReduce：云上Hadoop服务

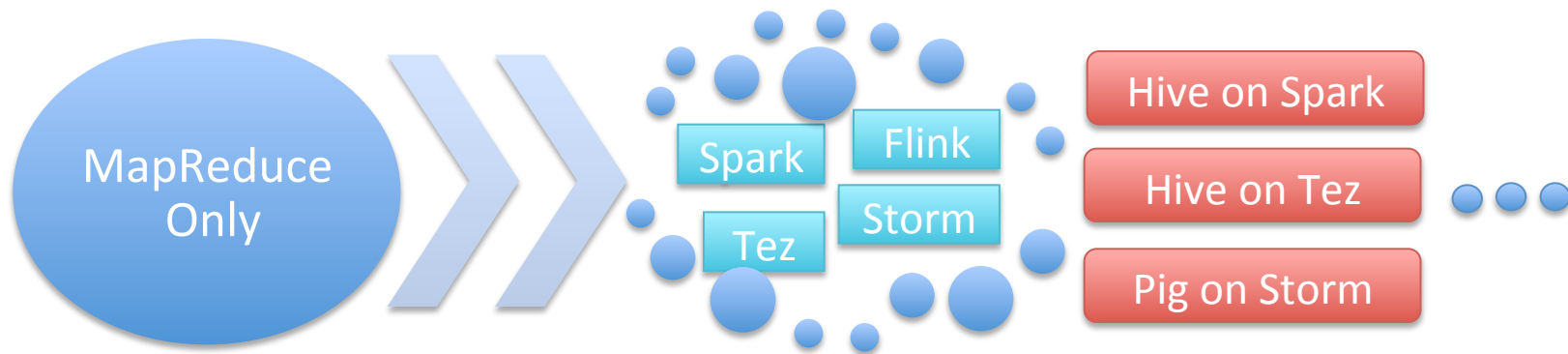




离线平台到在线平台



YARN 成为大数据操作系统



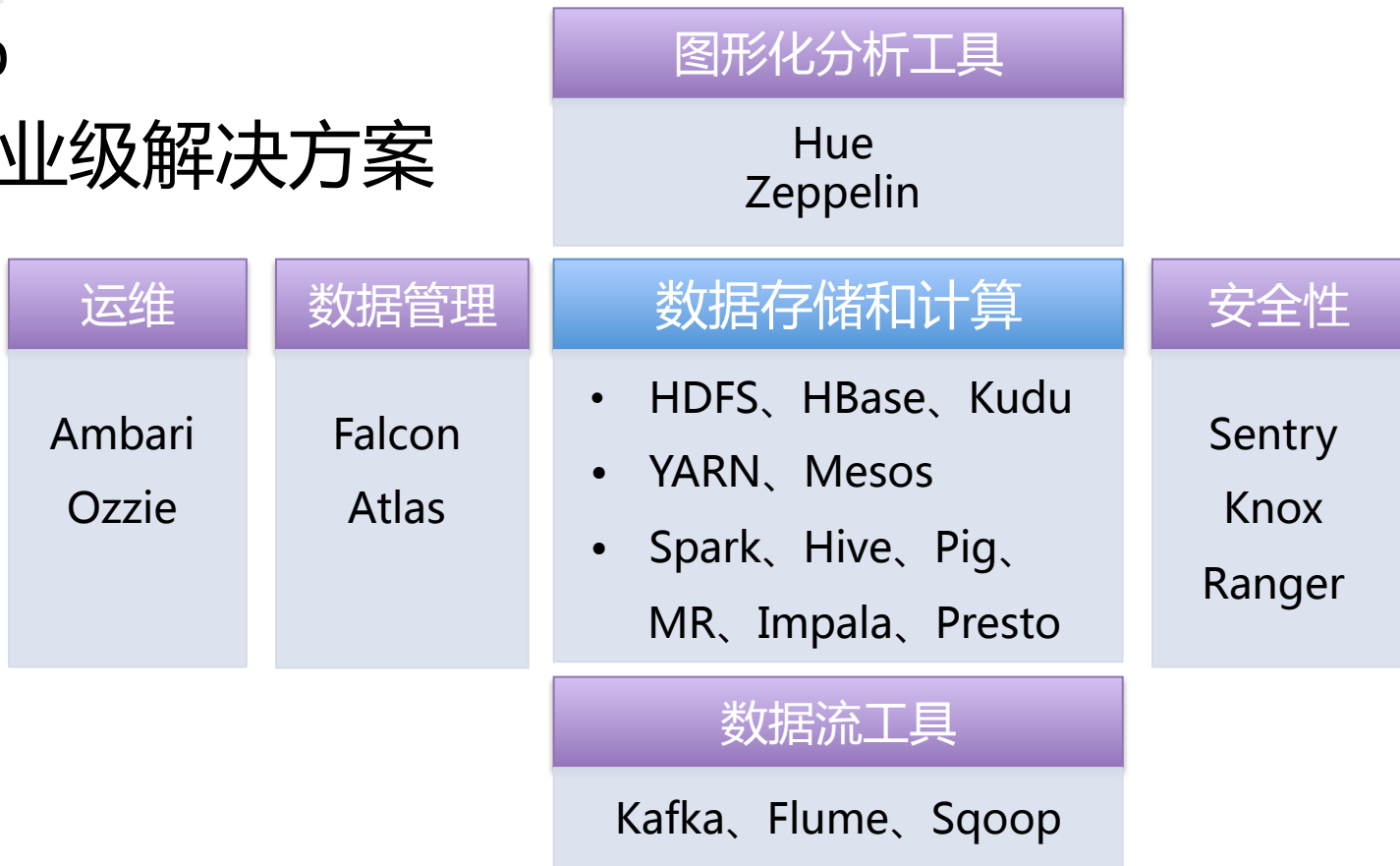
YARN之前

YARN之后



Hadoop

成为企业级解决方案



机器学习和人工智能

- Mahout -> Oryx：批处理模式到实时模式的机器学习工具
- 分布式编程框架都有机器学习的库并且扩展到更多的语言
 - Spark MLlib、FlinkML
 - SparkR、Python
- 深度学习和 Spark、Hadoop 结合更加紧密：
 - CaffeOnSpark、Deeplearning4j
 - TensorFlow：和HDFS、Spark的结合



阿里集团的 Hadoop 之路

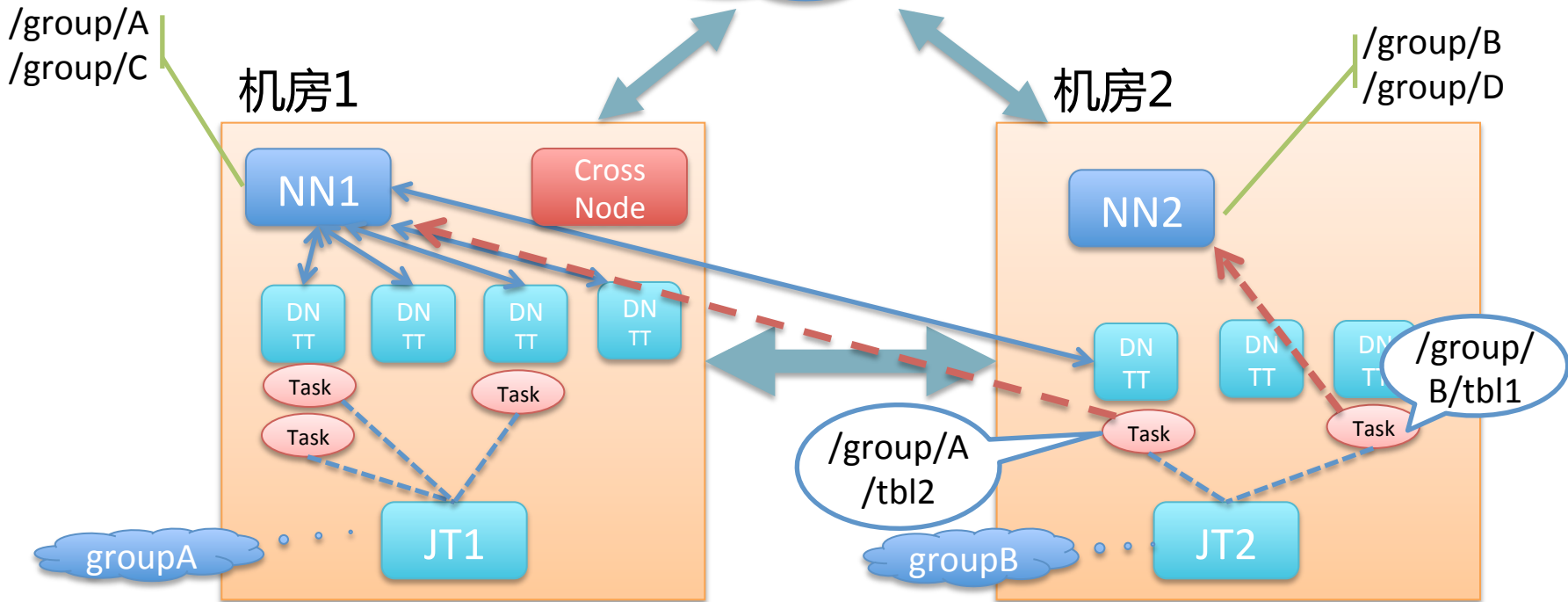
- 2008年 - 2009年：多部门独立的 Hadoop 集群
- 2009年 - 2015年：云梯集群和服务
 - 集群统一运维，专业的开发团队
 - 数据统一管理，集团层面的全局视图
 - 资源错峰分配，整体成本最优
- 2015年 - 至今：阿里云 E-MapReduce
 - 阿里云对外的 Hadoop 基础服务



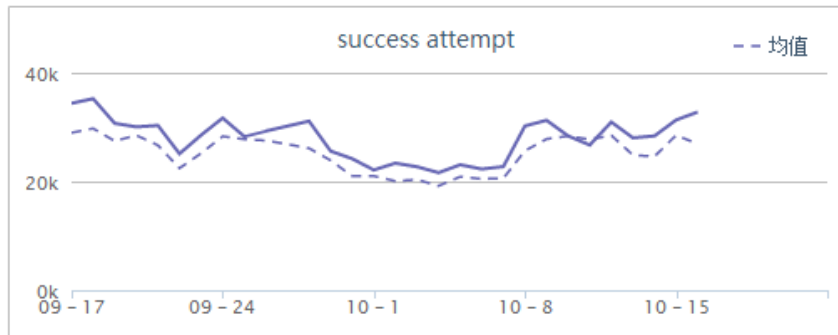
云梯：阿里内部的Hadoop服务

- 全局资源调度：支持业务优先级（基于Fair Scheduler）
- 安全性：HDFS上的扩展ACL，Hive安全认证和授权
- 稳定性：消除异常作业对全局的影响；Master HA
- 扩展性：Master节点的单点性能压力；跨机房部署架构
- 云梯医生：集群诊断系统

云梯跨机房部署架构

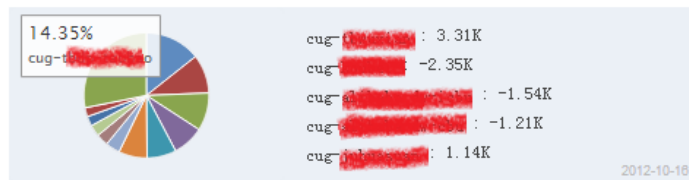


r01b05028@aliyun.com 指标图



- 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
 - Resource*sec, HDFS/local r/w
- 机器/机器组视图

- 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



阿里云E-MapReduce：云上Hadoop服务

- 定制化Hadoop版本
- Hadoop生态系统的完整支持：
 - Hive、Spark、HBase、Pig、Presto等等
- 云产品深度整合：OSS、LogService、MaxCompute...
- 运维自动化：一键部署、一键扩容、监控报警...
- 专家服务

Hadoop 生态圈未来展望

- 云存储成为和HDFS并列的分布式存储方案
 - AWS S3、Azure Datalake、阿里云OSS进入 Hadoop 核心版本
- 离线系统和流式系统的整合
 - Apache Beam
- 内存计算

2016 The
Computing
Conference
THANKS