Hadoop 3.0以及未来

刘 轶

自我简介

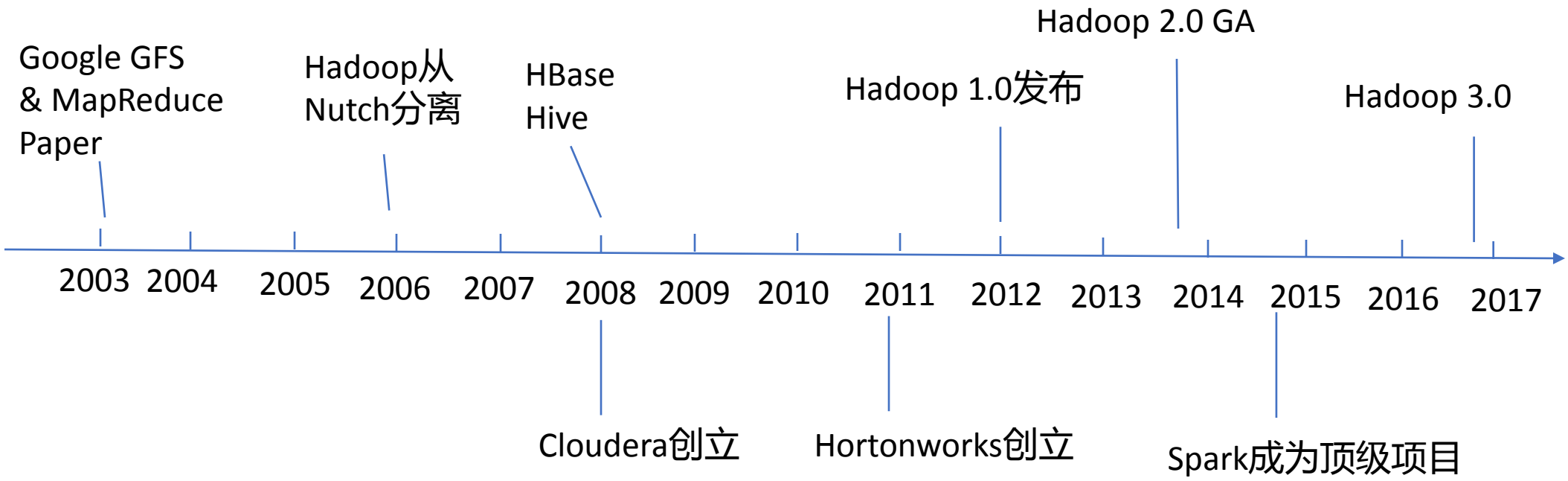
- Apache Hadoop的committer和项目管理委员会成员。
- ebay的Paid IM(互联网市场)部门架构师，领导ebay产品广告、互联网市场数据和实验平台的架构设计。负责领导使用Hadoop、Spark、 Kafka、 Cassandra等开源大数据项目建立ebay的广告和数据平台。
- 加入ebay前，在intel工作6年，大数据架构师，负责领导大数据的开源贡献、基于Intel平台的开源项目优化以及一些基于Spark的大规模机器 / 深度学习项目。
- 超过9年的互联网、云计算、大数据的工作经验。


Google
google.com/datacenters

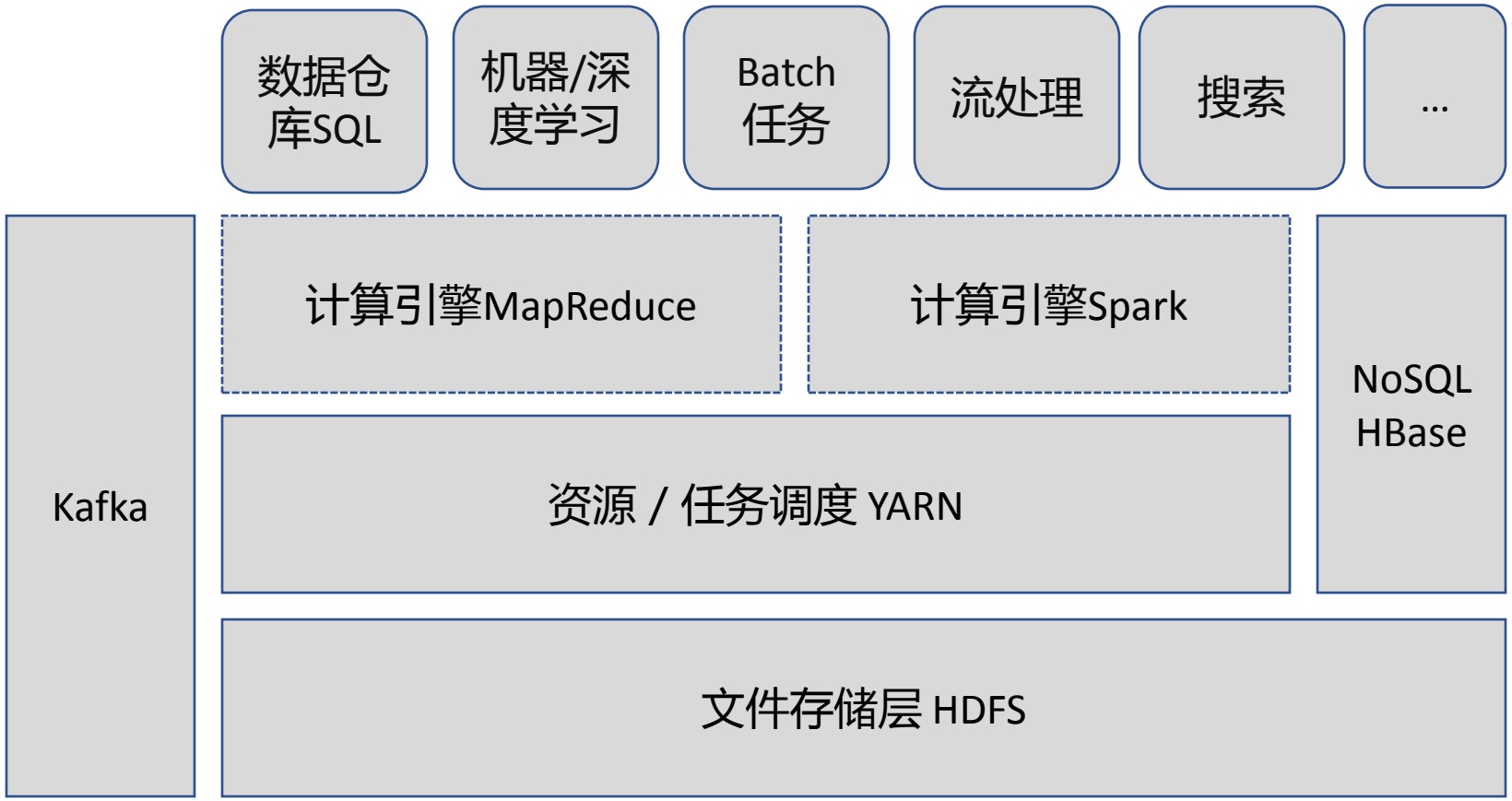
概要

- Hadoop的历史
- Hadoop 3介绍
  - Common
  - HDFS
  - YARN
  - MapReduce
- Hadoop的未来发展方向

Hadoop的历史


Google GFS
& MapReduce
Paper
Hadoop从
Nutch分离
HBase
Hive
Hadoop 1.0发布
Hadoop 2.0 GA
Hadoop 3.0
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Cloudera创立
Hortonworks创立
Spark成为顶级项目

Hadoop生态系统


数据仓库SQL
机器/深度学习
Batch任务
流处理
搜索
...
计算引擎MapReduce
计算引擎Spark
Kafka
资源 / 任务调度 YARN
NoSQL HBase
文件存储层 HDFS

Hadoop 3介绍

- Common
  - JDK 8+ 升级
  - Classpath隔离
  - Shell脚本的重构

- HDFS
- YARN
- MapReduce

Classpath隔离

- HADOOP-11656, HDFS-6200

问题：依赖性地狱(Dependency Hell)，版本冲突
解决方案：客户端(client-side)和服务器端(server-side)的隔离

Shell脚本的重构 - HADOOP-9902

- 脚本重构，提升可维护性和易用性
- 修正一些长期存在的bugs
- 加入一些改进
- 加入一些新功能
- 带来一些不兼容性
- Shell脚本现在更易于调试: --debug

Hadoop 3介绍

- Common
- HDFS
  - 纠错码(Erasure Coding)
  - 多个Standby Namenode
  - Datanode内部balance工具
  - 云计算平台的支持
- YARN
- MapReduce

HDFS纠错码(Erasure Coding)

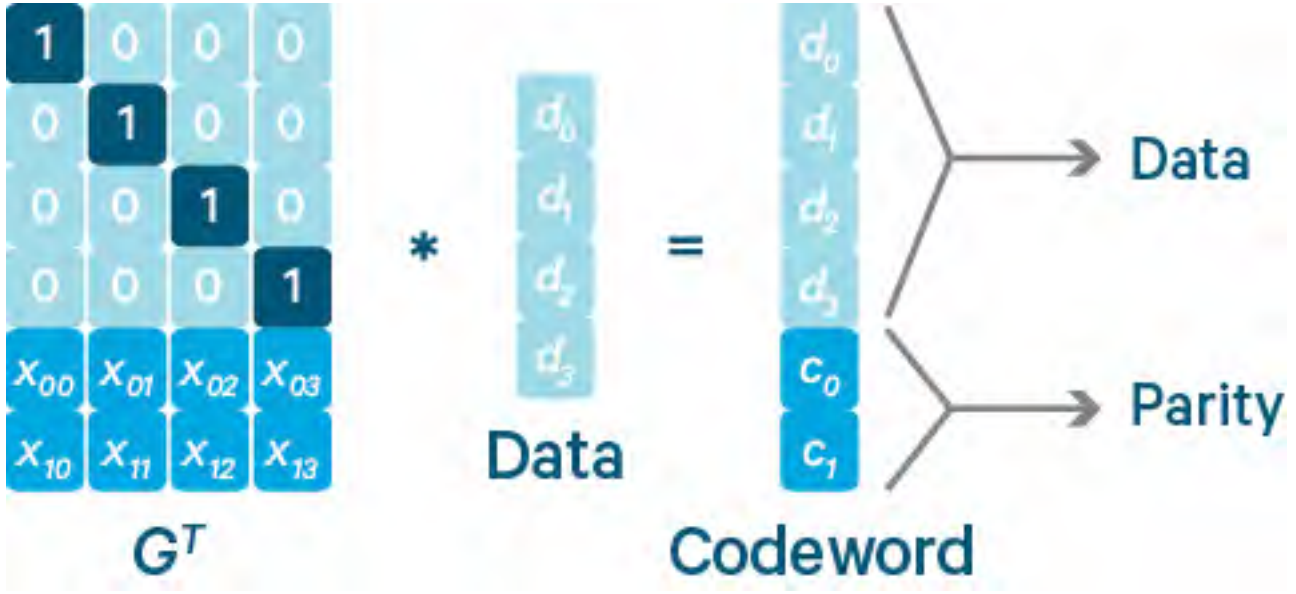
- 一个简单的例子

X	Y	X ⊕ Y
0	0	0
0	1	1
1	0	1
1	1	0

1备份: 1,0 需要额外的2位
XOR编码: 1,0 需要额外的1位

HDFS纠错码(Erasure Coding)

- Reed-Solomon (RS) 编码

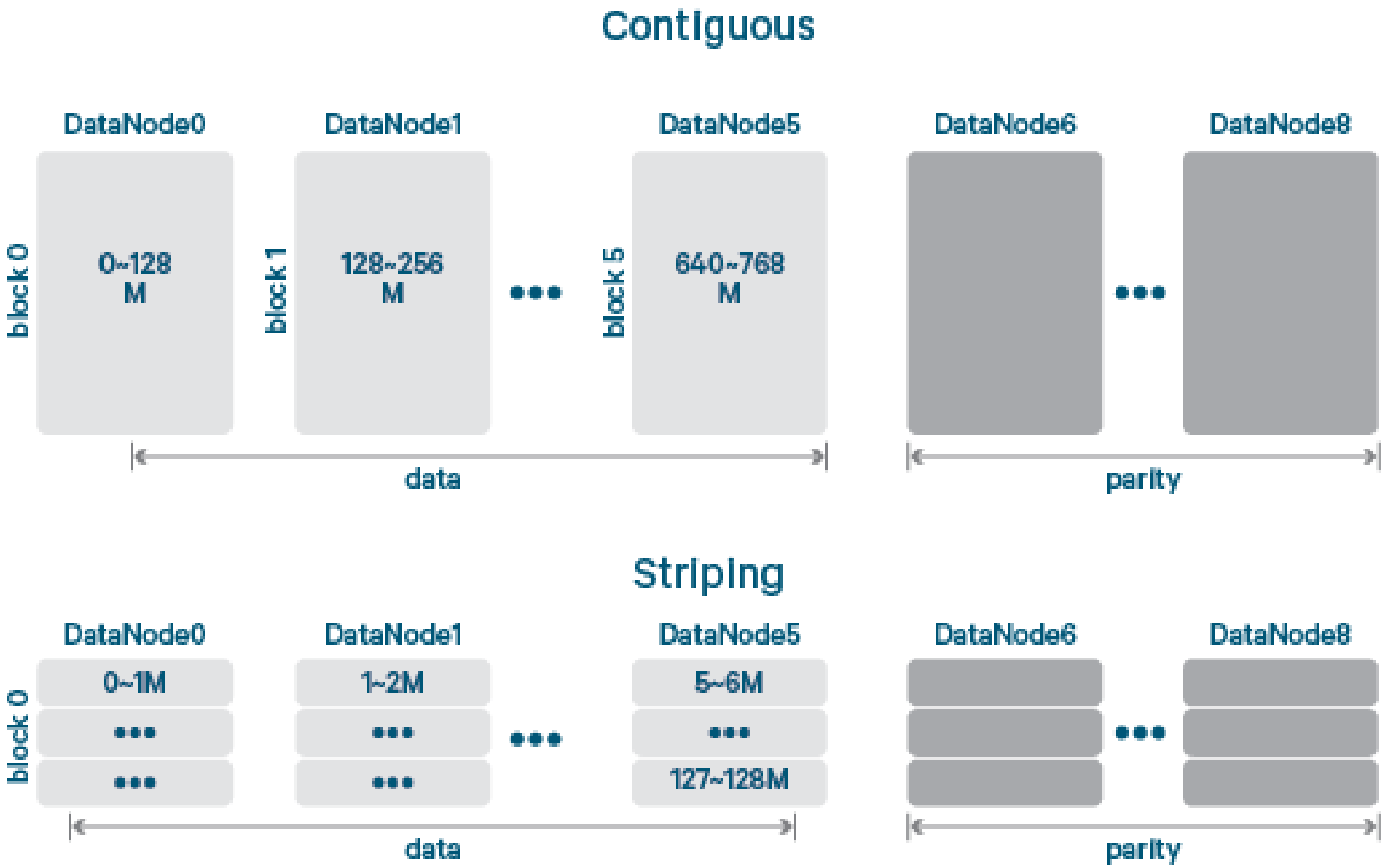


数据可靠性和存储效率

- 数据可靠性：可以最多几个节点故障
- 存储效率：k/(k+m)

	可靠性	存储效率
单副本	0	100%
3副本	2	33%
XOR(6个数据单元)	1	86%
RS(6,3)	3	67%
RS(10,4)	4	71%

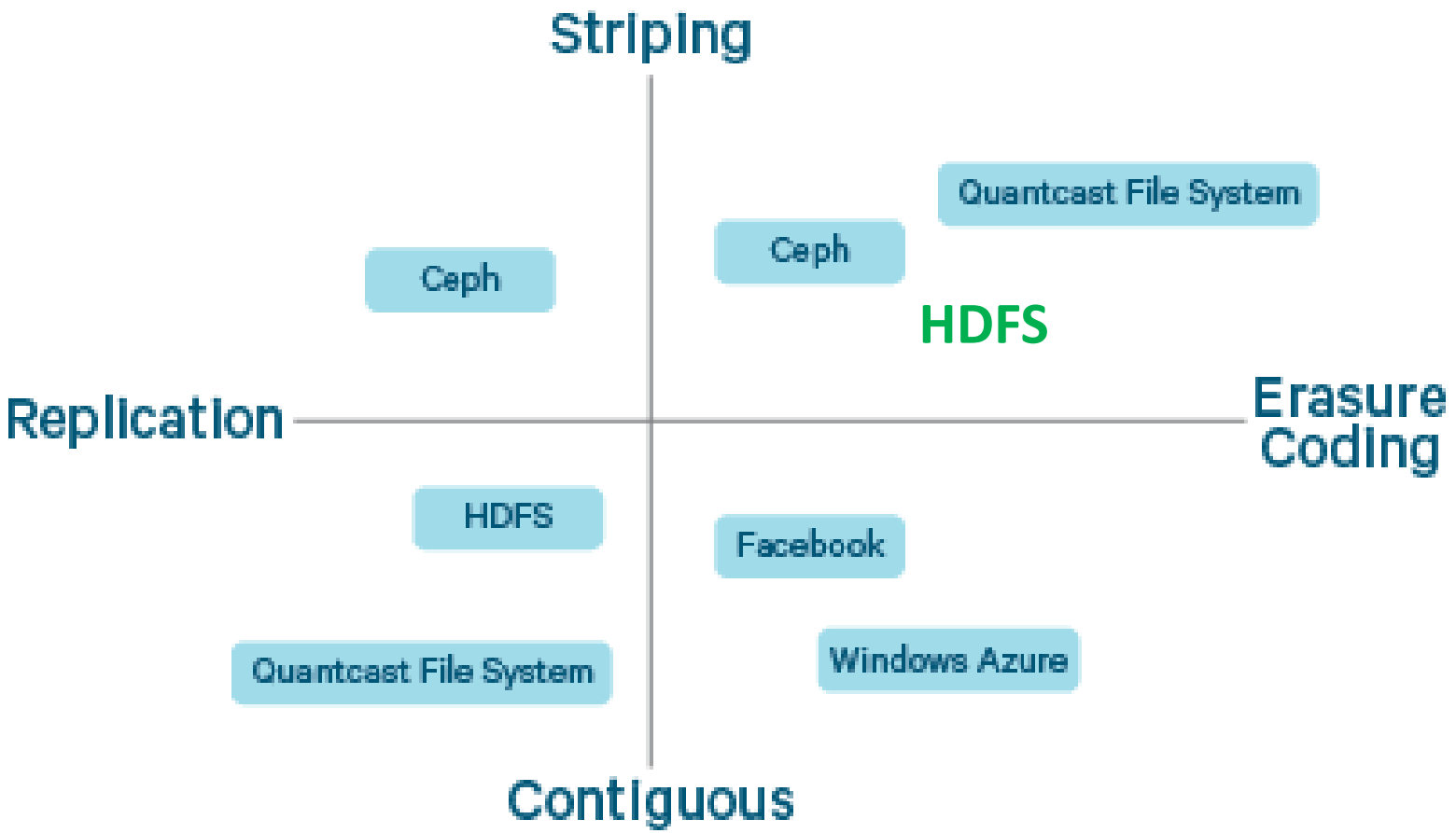
存储布局 - 连续和条状



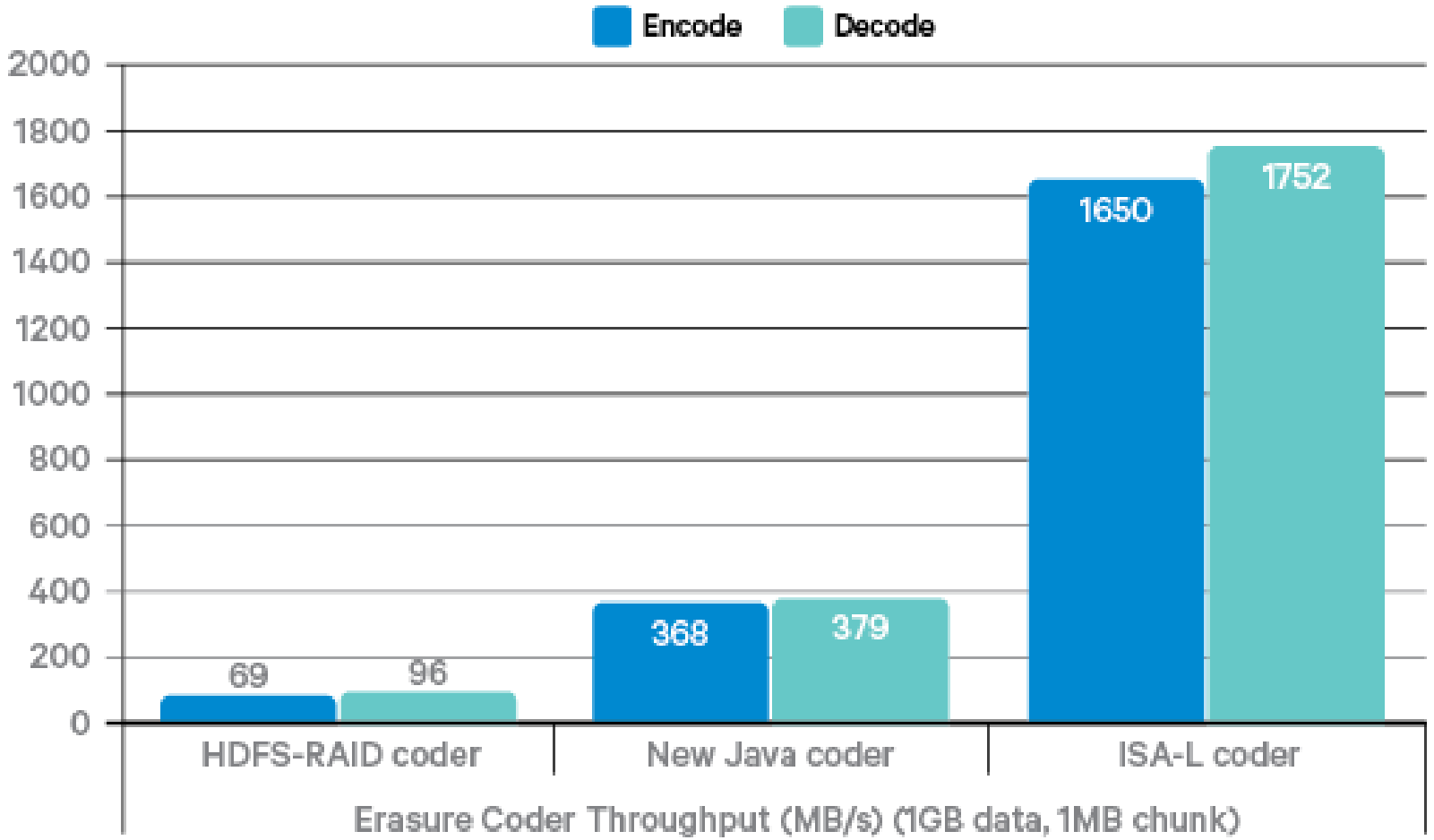
数据本地性
小文件处理

小文件处理
并行IO
数据本地性

纠错码在分布式存储系统中

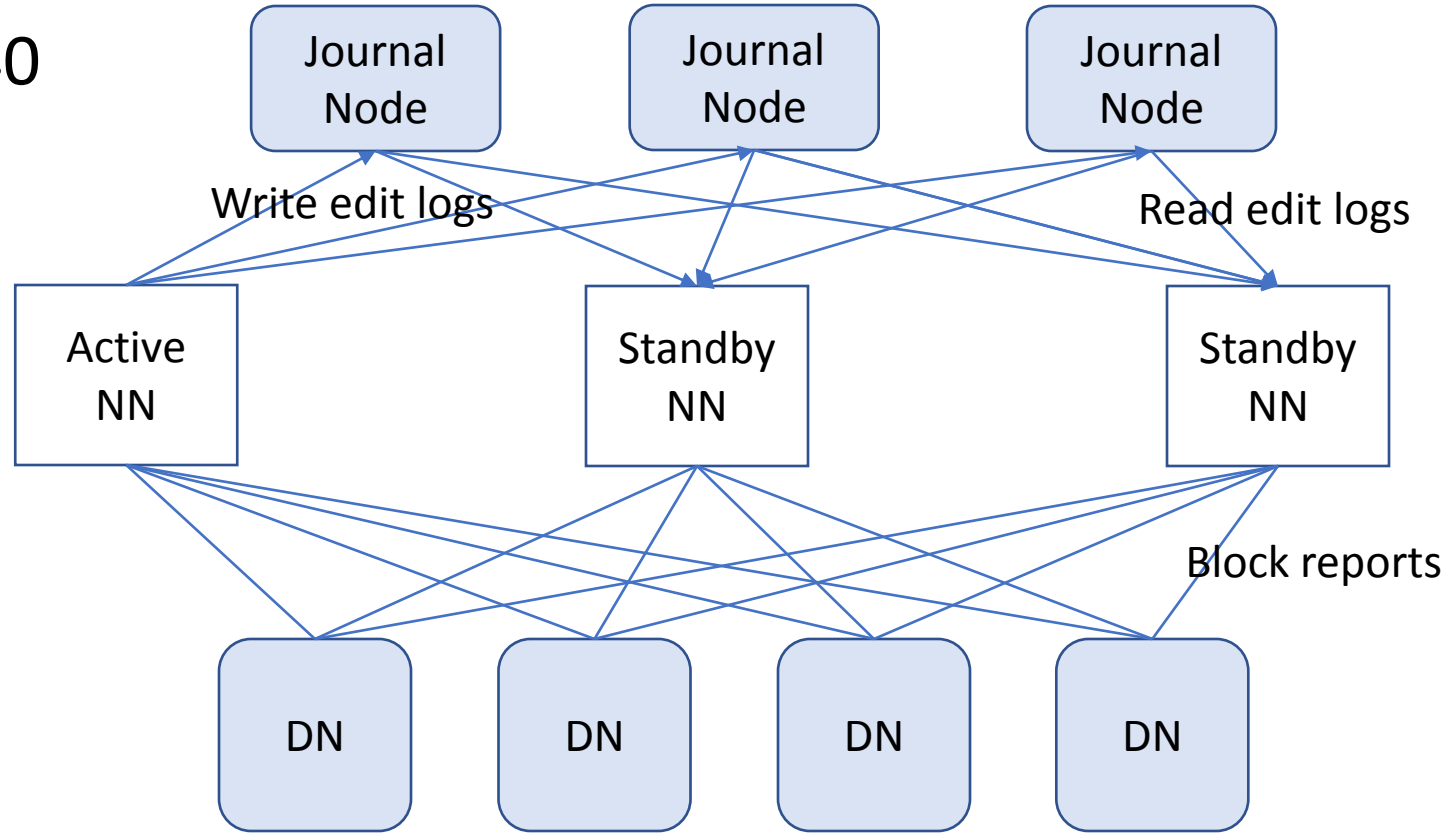


性能

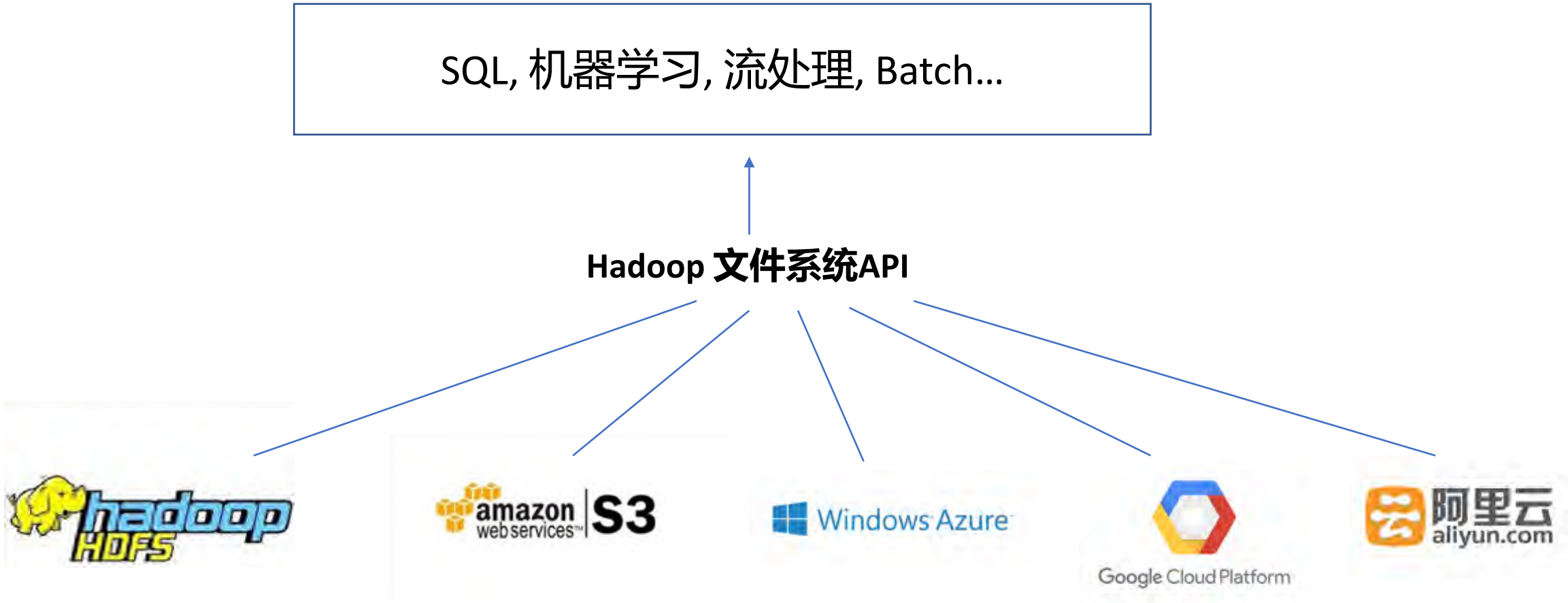

Encode
Decode
2000
1800
1600
1400
1200
1000
800
600
400
200
0
69
96
368
379
1650
1752
HDFS-RAID coder
New Java coder
ISA-L coder
Erasure Coder Throughput (MB/s) (1GB data, 1MB chunk)

多个Standby Namenode

HDFS-6440



云计算 - 存储虚拟化

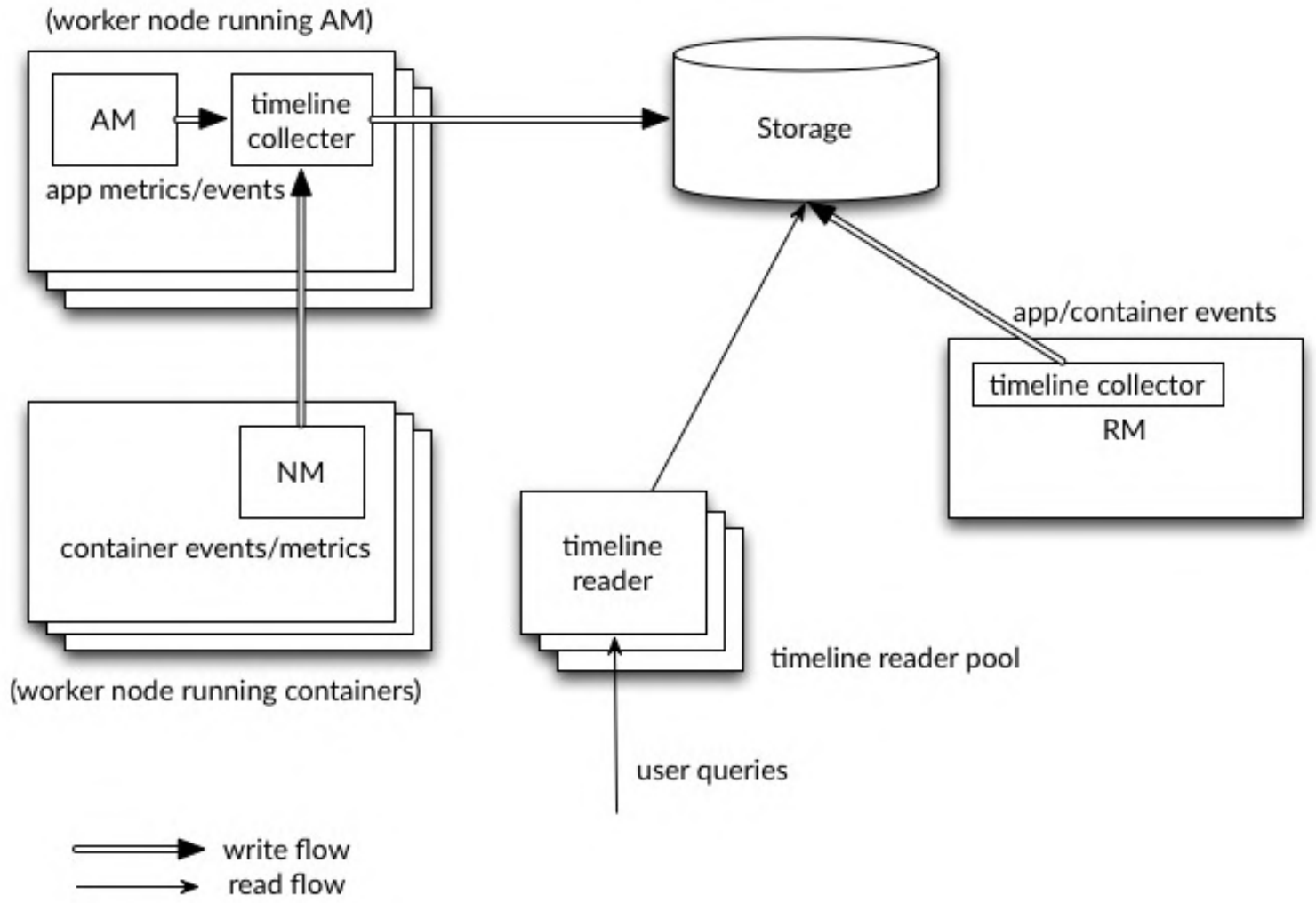

SQL, 机器学习, 流处理, Batch…
Hadoop 文件系统API
hadoop HDFS
amazon web services S3
Windows Azure
Google Cloud Platform
阿里云 aliyun.com

Hadoop 3介绍

- Common
- HDFS
- YARN
  - YARN Timeline Service v.2
  - YARN Federation
  - 动态资源配置
  - 容器资源的动态调整
  - 资源隔离
  - 调度的增强
  - YARN的Web页面的增强
- MapReduce

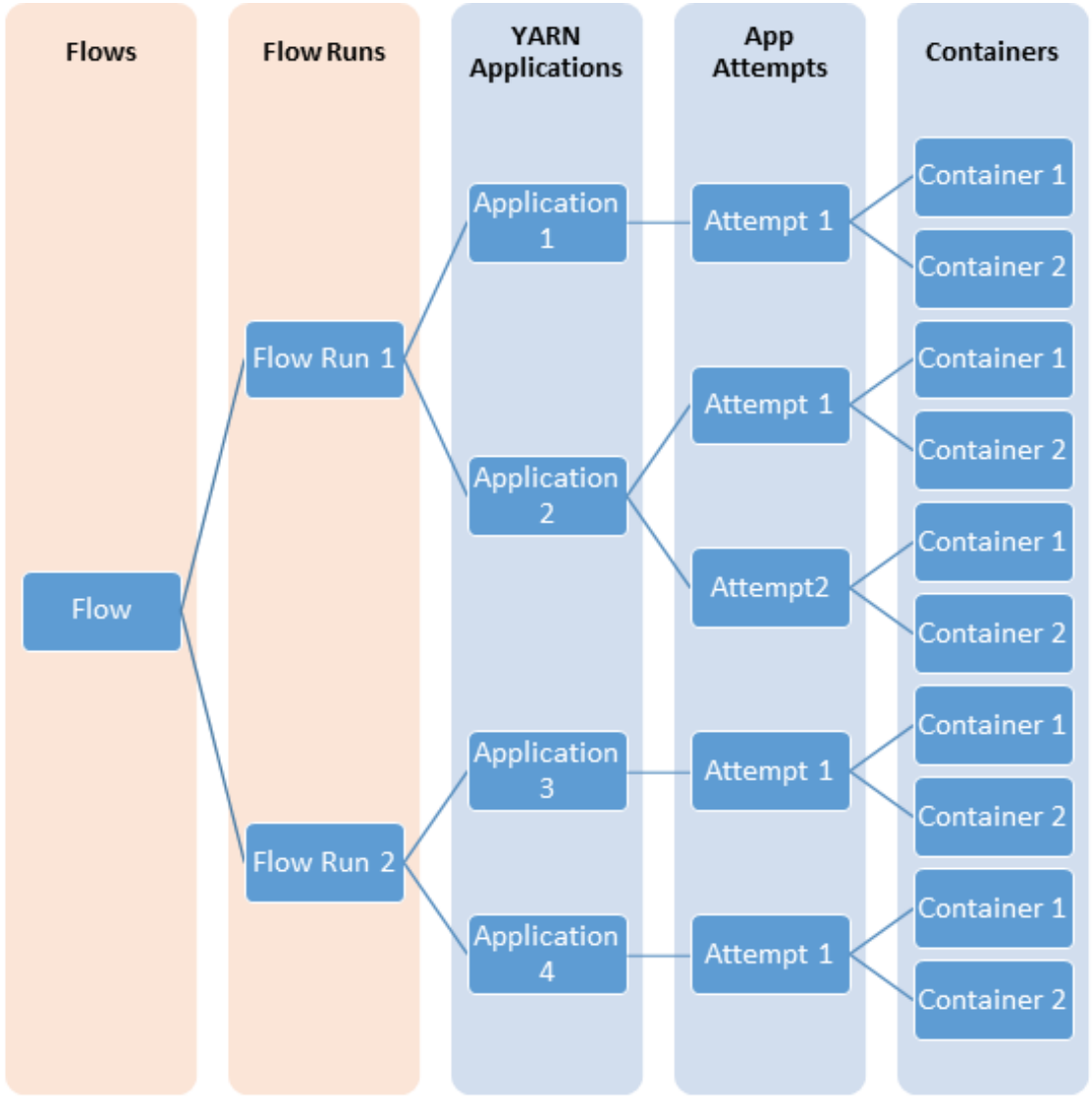
YARN Timeline Service v.2

- 扩展性
  - 分布式读写
  - 读写分离
  - HBase存储



YARN Timeline Service v.2

- 可用性
  - 流(flow)
  - 聚合(aggregation)



YARN Federation

- YARN-2915

  允许YARN的集群扩展到一万个或更多个节点

  YARN的集群的集群对用户来说是一个整体的集群

动态资源配置

- YARN-291

  允许动态的改变NM的资源配置

容器资源的动态调整

- YARN-1197

  允许运行时动态的调整分配给容器的资源

资源隔离

- 磁盘资源的隔离 - YARN-2619
- 网络IO的隔离 - YARN-2140
- Docker Container - YARN-3611

调度的增强

- 在同一个队列(queue)的优先级 - YARN-1963

YARN的Web页面的增强

- YARN-3368

Hadoop 3介绍

- Common
- HDFS
- YARN

- MapReduce
  - Task层次的Native优化

MapReduce Task层次Native优化

- 对map output collector的Native实现，对于shuffle密集型的task能带来30%的性能提升。

Hadoop 的未来

HDFS的未来

- 对象存储 - HDFS-7240
- 更高性能的Namenode：更高效的内存使用，锁的改进等
- Erasure Coding的完善

YARN的未来

- 更大规模的集群支持
- 更好的资源调度，隔离和多租户
- 支持更多的应用，包括long running的service

谢谢

Q&A