HADOOP 3.0新特性与展望

董西成 @ HULU

自我介绍

- 董西成
- 就职于HULU北京研发中心
- 《Hadoop技术内幕》书籍作者
- 专注于Hadoop与Spark等大数据技术

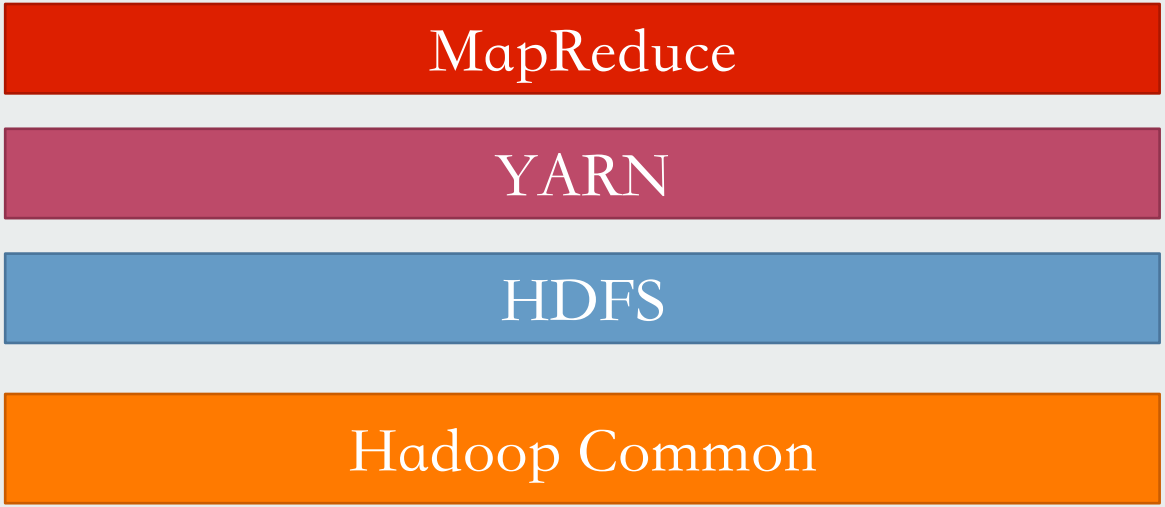
AGENDA

- Hadoop 3.0概况
- Hadoop Common主要改进
- HDFS新功能与特性
- YARN新功能与特性
- MapReduce主要改进
- Q & A

HADOOP 3.0概况

- 基于JDK 1.8发布一个新的Hadoop版本
  - JDK 1.7在2015年4月已停止更新
  - 影响力不如hadoop 1.0 → 2.0
- 发布时间
  - Alpha版预计今年夏天发布(brach-3.0.0-alpha)
  - GA版本11月或12月发布

HADOOP模块构成



AGENDA

- Hadoop 3.0概况
- Hadoop Common主要改进
- HDFS新功能与特性
- YARN新功能与特性
- MapReduce主要改进
- Q & A

HADOOP COMMON主要改进

- 精简Hadoop内核
  - 剔除过期的API和实现
  - 将默认组件实现替换成最高效的实现
    - 将FileOutputCommitter缺省实现换为v2版本(MAPREDUCE-4815, ~30%+)
    - 废除hftp(read-only)转由webhdfs替代(HDFS-2316)
    - 移除Hadoop子实现序列化库org.apache.hadoop.Records (protobuf)
- Classpath isolation (HADOOP-11656)
  - 防止不同版本jar包冲突(guava, protbuf)

HADOOP COMMON主要改进

- Shell脚本重构(HADOOP-9920)
  - 修复了大量bug
  - 增加了新特性
  - 支持动态命令

AGENDA

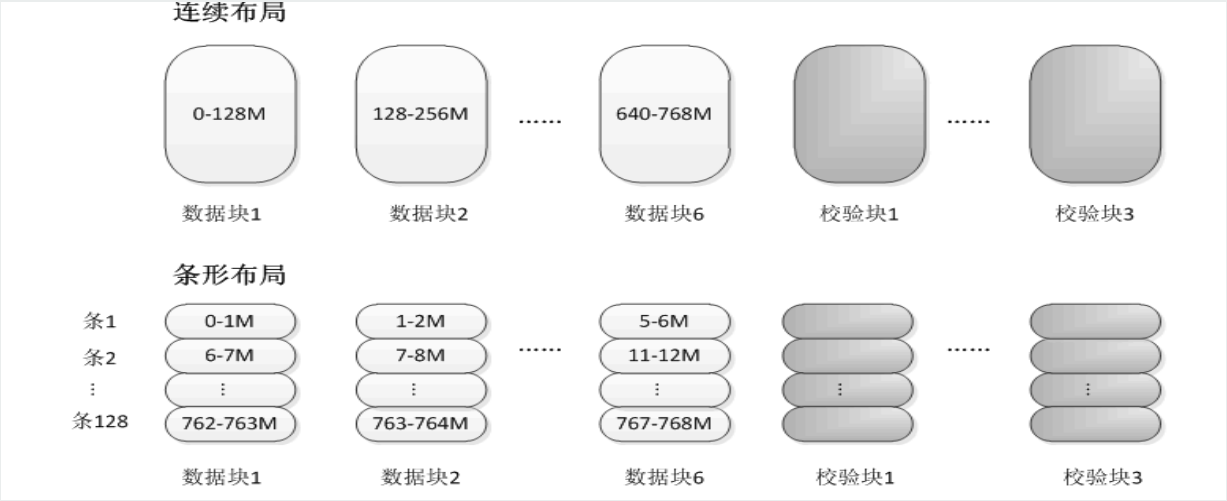
- Hadoop 3.0概况
- Hadoop Common主要改进
- HDFS新功能与特性
- YARN新功能与特性
- MapReduce主要改进
- Q & A

HDFS新功能与特性

- HDFS纠删码(Erasure Coding, HDFS-7285)
  - 引入动机:在不降低可靠性的前提下,节省一半存储空间
  - 基本原理:对数据分块,计算产生冗余的校验块,当部分数据块丢失时,通过剩余数据块和校验块计算出丢失的数据块。
- 实现方案
  - 方案1: 引入新的服务对数据编码和恢复,代表:facebook的HDFS-RAID
  - 方案2:将纠删码融入HDFS内部,代表:Hadoop 3.0,由Intel和Cloudera主导

HDFS新功能与特性

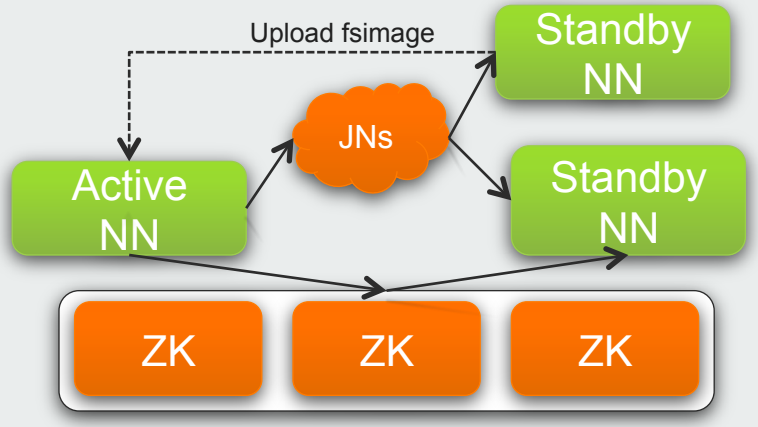
- HDFS纠删码(Erasure Coding)
  - 编码**方式**:Reed-Solomon(RS)码，RS(k,m)
  - 连续**布局**(Contiguous Layout) VS **条形布局**(Striping Layout)



《程序员》杂志，李波：
“HDFS EC：
将纠删码技术融入HDFS”

HDFS新功能与特性

- 多NameNode(3~5)支持(HDFS-6440)
  - 当前HA方案:一个active namenode，一个standby namenode
  - Hadoop 3.0：一个active namenode，多个standby namenode(3~5个)
  - 注:YARN在hadoop 2.x版本中已经支持多ResourceManager



AGENDA

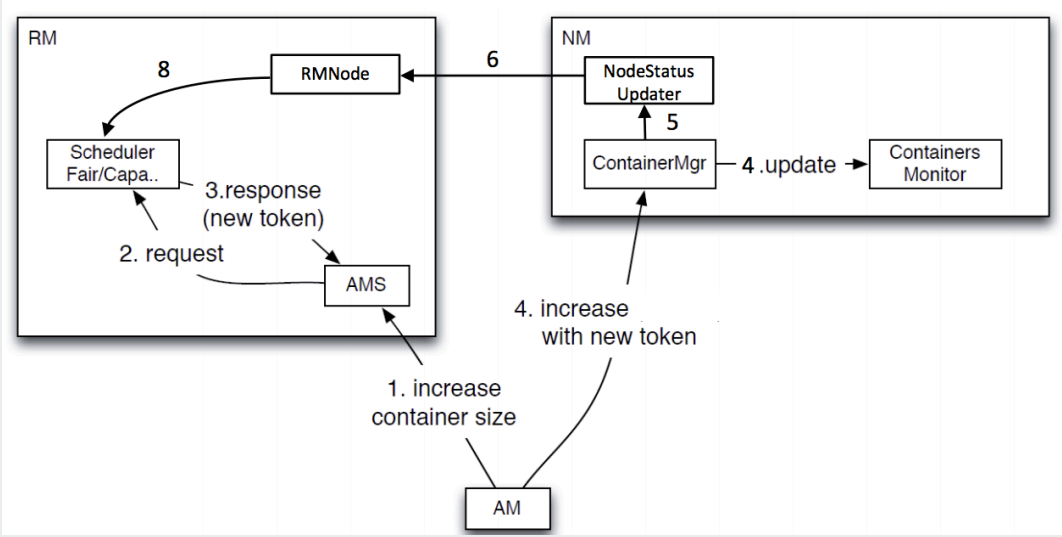
- Hadoop 3.0概况
- Hadoop Common主要改进
- HDFS新功能与特性
- YARN新功能与特性
- MapReduce主要改进
- Q & A

YARN新功能与特性

- 更细粒度资源隔离
  - 当前实现
    - Default （process）
    - Cgroup（only cpu）
    - Docker（alpha）
  - Hadoop 3.0（YARN-2619）
    - Memory 隔离
    - IO 隔离 (Support for Disk as a Resource in YARN)

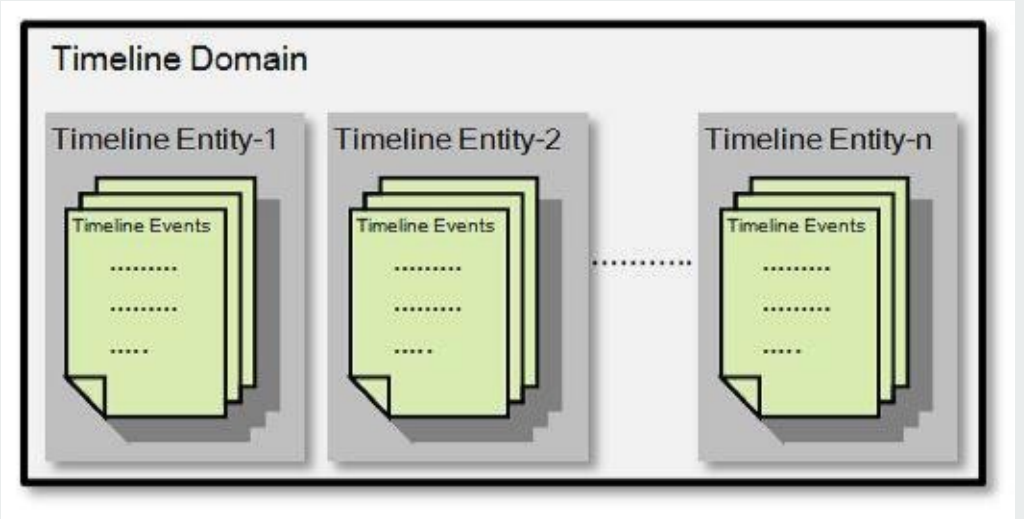
YARN新功能与特性

- Container Resizing(YARN-1197)
  - 动态调整Container资源
  - JVM-based container
    - NOT easy to do



YARN新功能与特性

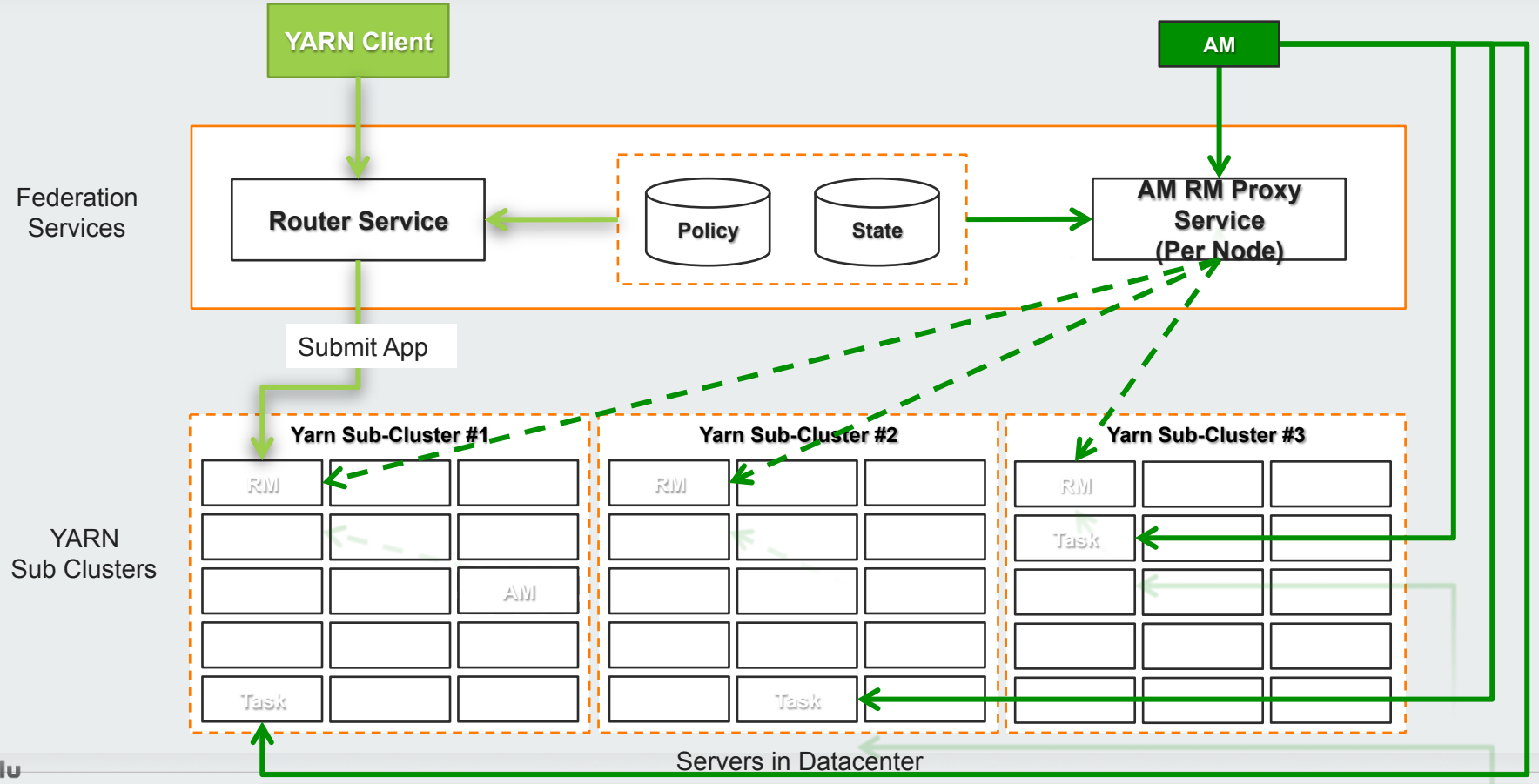
- Timeline Server v2(YARN-2928)
  - Timeline Server:应用程序信息存储和检索系统
    - 正在运行信息和历史信息(timeline domain/entity/event)
    - MR和Tez
  - Hadoop 3.0
    - 提高系统扩展性与可靠性
    - 增加聚集功能



YARN FEDERATION(3.X)

- 动机：使YARN支持100K级别的节点规模
- YARN RM是单实例，其扩展性受一下因素影响
  - 基数：|nodes|, |apps|, |tasks|
  - 频率：NM和AM的心跳间隔，task运行时间
- 目前已知最大YARN集群（apache原生）规模为4~8K
- 微软提出YARN FEDERATION(YARN-2915)，> 50K节点

YARN FEDERATION(3.X)



AGENDA

- Hadoop 3.0概况
- Hadoop Common主要改进
- HDFS新功能与特性
- YARN新功能与特性
- MapReduce主要改进
- Q & A

MAPREDUCE主要改进

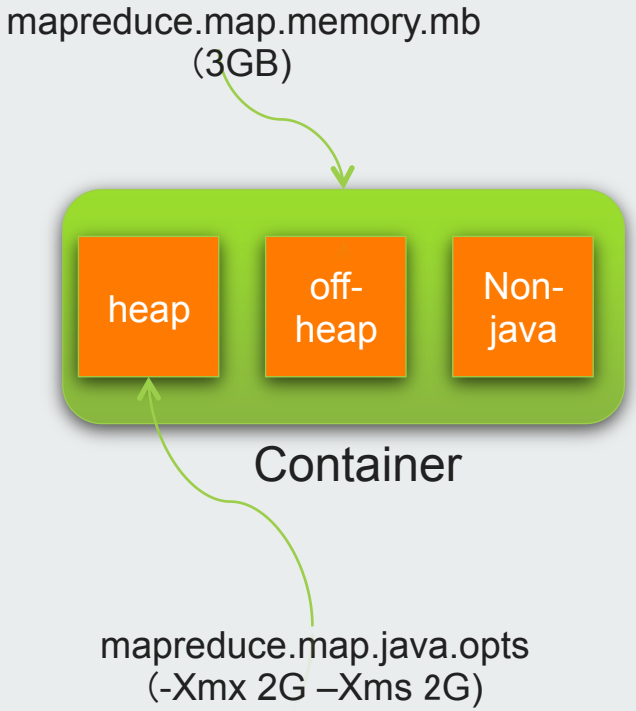
- MapReduce目前仍是主流的计算引擎
  - Hive是使用最广泛的hadoop组件之一，其底层引擎仍以MR为主
- MapReduce vs Spark
  - Spark在处理大数据，Hive支持方面，仍有很长的路要走
  - Spark Shuffle实现不够高效，尤其处理大数据（SPARK-2926）

MAPREDUCE主要改进

- 引入Native Task(MAPREDUCE-2841)
  - C++版 MR runtime，实现了map output collector（包括Spill，Sort和IFile等）
  - 对于shuffle密集型应用，其性能可提高约30%
- Native Task性能（github）
  - High performance
  - Support no sort
  - Binary based, no serialization/deserialization overhead

MAPREDUCE主要改进

- 内存自动推断(MAPREDUCE-5785)
  - 当前MR:
    - mapreduce.map/reduce.memory.mb
    - mapreduce.map/reduce.java.opts
  - Hadoop 3.0
    - mapreduce.job.heap.memory-mb.ratio
    - mapreduce.map/reduce.java.opts



我的微信公众号

Q & A