大数据编程模型和使用技巧

Hadoop/YARN

陈一帅

yschen@bjtu.edu.cn

北京交通大学电子信息工程学院

内容

- Hadoop HDFS
- YARN

Hadoop

- 第一个主要的云数据分析工具
- 发展为了 Apache YARN

核心：HDFS

- Hadoop Distributed File System
  - Hadoop 的核心
  - 用 Java 编写，完全可移植并且基于标准的网络 TCP 套接字
- 不是 POSIX 文件系统
  - 一次写入，多次读取，仅保持"最终一致"

核心：HDFS

- 可以被直接 Mount 为用户空间文件系统（FUSE）
  - 从而可以用命令行，以类似于标准 Unix 文件系统的方式进行操作

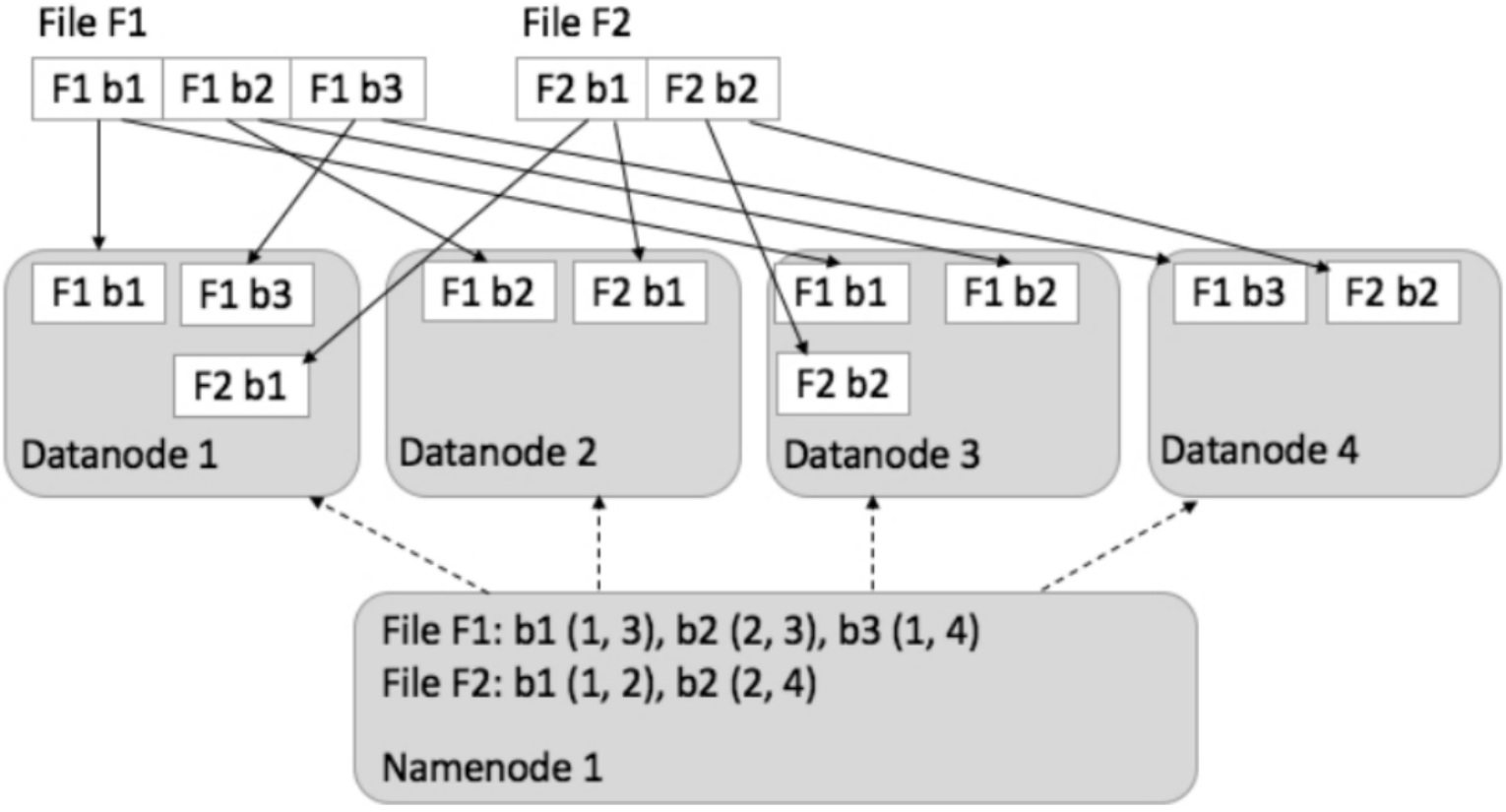
```
$hadoop fs -mkdir /user/wiki
$curl -s -L http://dumps.wikimedia.org/enwiki/...multisream.xml.bz2\
   | bzip2 -cd |hadoop fs -put - /user/wiki/wikidump-en.xml
$hadoop fs -ls /user/wiki
  Found 1 items
  -rw-r--r-- hadoop 59189269956 21:29 /user/wiki/wikidump-en.xml
```

HDFS

- NameNode，跟踪数据位置
  - 管理名称空间，确定块到 DataNode 的映射
- DataNode 集群，保存分布式数据
  - 单个文件分为 64 MB 的块
  - 这些块分布在各个 DataNode 上
  - 被复制到多个节点上，实现容错

HDFS

- 1 个 NameNode 跟踪块和副本
- 4 个 DataNodes，2 个文件，文件块分布式存储



辅助 NameNode

- HDFS 文件系统包括一个所谓的辅助 NameNode
- 该名称具有误导性，可能会被认为是在主 NameNode 脱机时备用的 NameNode
- 实际上，辅助 NameNode 的作用是定期与主 NameNode 连接，为主 NameNode 的目录信息构建快照，将其保存到本地或远程目录中
- 这些快照可用于主 NameNode 失败时的重新启动，不必通过重现文件系统操作的整个日志来重建目录

HDFS 文件复制

- HDFS 在多台计算机上存储 GB 或 TB 的大文件
- 它通过在多个主机之间复制数据来实现可靠性，因此理论上讲不需要主机上的独立磁盘冗余阵列（RAID）存储（但是要提高 I/O 性能，某些 RAID 配置仍然有用）
- 例如，复制值为 3，则数据存储在三个节点上：两个存储在同一机架上，一个存储在不同机架上
- 数据节点可以相互通信以重新平衡数据，四处移动副本，保持数据复制

HDFS 文件读取流程

- 用户向 NameNode 发送“打开”请求，获取文件块的位置
- 对于每个文件块，NameNode 返回一组数据节点的地址，该数据节点包含所请求文件的副本信息，数量取决于块副本的数量
- 收到此类信息后，用户调用“读取”以连接到包含文件第一个块的最近的 DataNode
- 在将第一个块从相应的 DataNode 流传输到用户之后，终止已建立的连接
- 对请求的文件的所有块重复相同的过程，直到将整个文件传输到用户

HDFS 周期消息

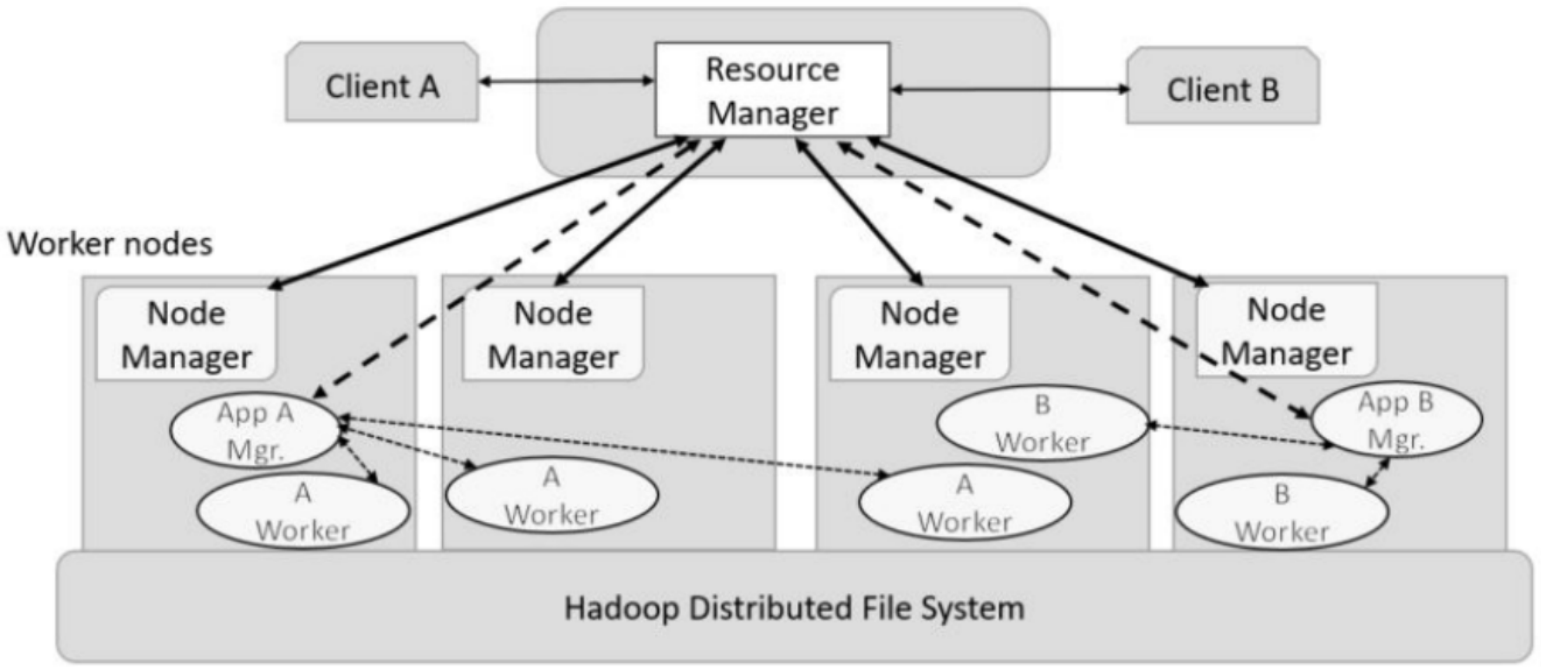
- 心跳和 Blockreport 消息
- 每个 DataNode 周期性地发送到 NameNode 的消息
- 接收到心跳信号表示 DataNode 正常运行
- 每个 Blockreport 都包含 DataNode 上所有块的列表
- NameNode 基于该消息，调度系统中所有副本

内容

- Hadoop HDFS
- YARN

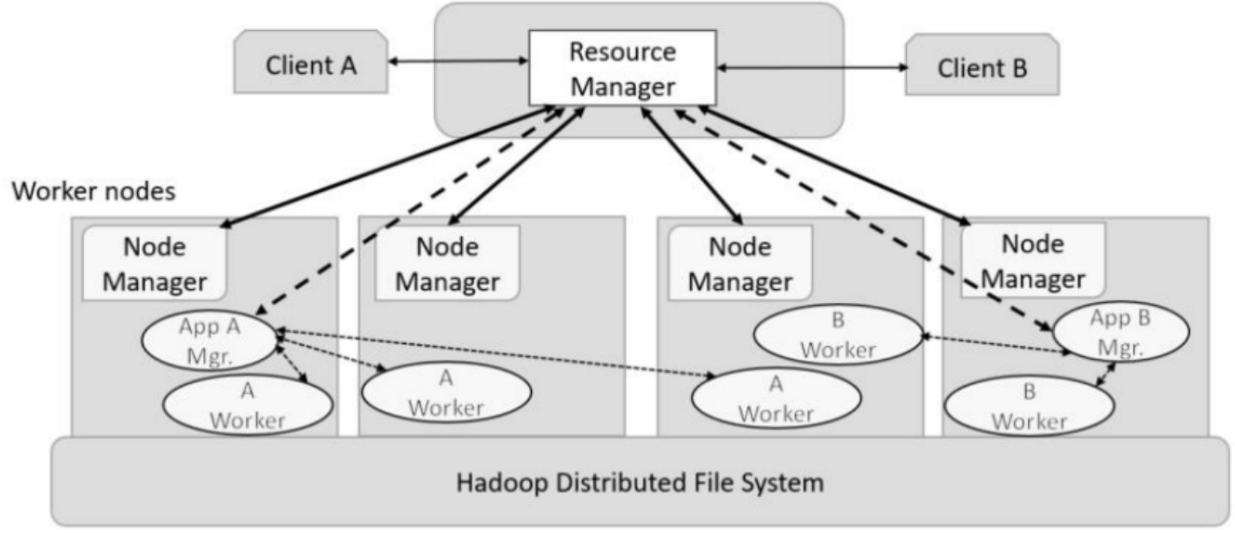
YARN 分布式资源管理器

- Yet Another Resource Negotiator
- 由 Hadoop 演进而来的完整分布式作业管理系统
- 体系结构



YARN 分布式资源管理器

- 应用程序连接资源管理器
  - 资源管理器启动该程序的程序管理器
  - 允许多个应用程序同时在系统上运行
- 资源管理器负责调度
  - 可与每个工作节点中的节点管理器进行通信
  - 程序管理器与资源管理器交互，获取其工作节点“容器”



YARN vs. Mesos

- YARN 在许多方面与 Mesos 系统相似
- 主要区别
  - YARN 旨在安排 MapReduce 样式的作业
  - Mesos 旨在支持更通用的计算类别，包括容器和微服务
- 两种系统都被广泛使用。

YARN 资源管理

- YARN 是一种资源调度程序，它允许使用简单的编程模型跨计算机集群对大型数据集进行分布式处理
- 它旨在从单个服务器扩展到数千个服务器，每个服务器都提供本地计算和存储
- 它通过在应用层进行智能故障检测和管理来支持高可用性
- 支持 MPI，Hadoop 和 Spark 等多种应用程序

YARN 三级资源管理

- 节点管理器（NM）
  - 管理 VM 和容器
- 应用程序管理器（AM）
  - 将应用程序容器的集合作为业务流程组进行处理
- 资源管理器（RM）
  - 监督 NM 和 AM，监视最高级别的全局资源

YARN 调度机制

- 默认：先进先出（FIFO）调度
- Fair Scheduler
  - Facebook 开发
  - 将作业分组到池中
  - 为每个池分配保证的最低份额
  - 过剩能力在工作之间分配
  - 未分类的作业进入默认池
  - 池必须指定 Map Slot，Reduce Slot 的最小数量，和正在运行的作业数量限制

Capacity Scheduler

- Yahoo!开发
- 与 Fair Scheduler 相似
- 队列
  - 队列被分配了总资源容量的一小部分
  - 空闲资源分配给容量被超出了的队列
  - 在队列中，具有高优先级的作业可以优先访问队列资源
- 一旦作业在 Hadoop 中运行，就不会被 preemption（先发制人）

商业系统中的 YARN

- AWS，Azure 都集成了 YARN
  - Amazon 版的 YARN，就是 E-MR
  - Azure 版的 YARN，就是 HDInsight

练习

- 调研Haddop、YARN在你们组选定的云平台的应用情况