MAPREDUCE实验

MAPREDUCE

- 即Map&Reduce
  - Map: 将作业分成若干份，分配到不同的机器上去执行
  - Reduce: 同样也是分配到不同的机器去执行，目标是将Map任务生成的中间文件汇总
- 例子:
  - 某省高考作弊严重，教育部要核查所有考生成绩，并重新计算该省的高考成绩情况。

如何解决？

- 全省的考卷非常多，需要将考卷划分再，然后分配到不同的地区，地区按学校分，然后学校……最后，他们同时进行核查
- 核查完成后，在一步一步把成绩报告给上一层，最终得到整个省的考试情况。

- 实际上map/reduce就是这个思想

MAPREDUCE特点

- 大规模数据
- 并行运算
- 节点可信

MAPREDUCE术语

- 作业
  - 用户的计算请求
- 作业服务器
  - 用户提交作业的服务器，同时，它还负责各个作业任务的分配，管理所有的任务服务器
- 任务服务器
  - 负责执行具体的任务
- 任务
  - 由作业拆分出来的执行单位
- 备份任务
  - 某些任务执行失败或者执行效率低下，可能需要再另外的任务服务器上面执行同样一个任务

HADOOP

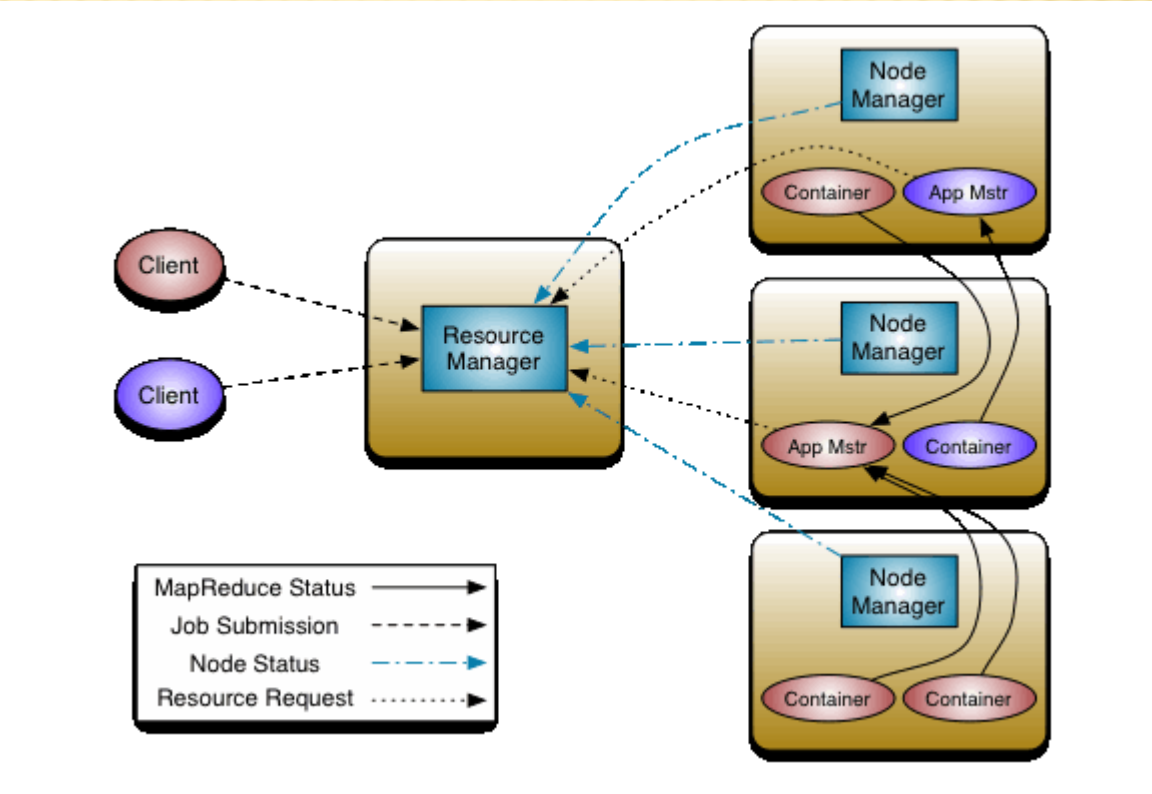
- Apache Hadoop是一款支持数据密集型分布式应用并以Apache 2.0许可协议发布的开源软件框架。
- 整个Apache Hadoop“平台”包括：
  - Hadoop内核
  - MapReduce
  - Hadoop分布式文件系统(HDFS)
  - 相关项目（Hbase, Hive等）

HADOOP重要概念

- HDFS（Hadoop分布式文件系统）
  - NameNode：保存所有文件的目录树，并不存放升级数据
  - DataNode：存放实际数据
- Map/Reduce
  - 对mapreduce的实现，最新版本升级为yarn
- JobTracker
  - 分发mapreduce任务给集群中的节点
- TaskTracker
  - 接收任务的节点，进行实际的任务

YARN

- Hadoop NextGen MapReduce(MRv2)
  - 新一代的mapreduce（version2）

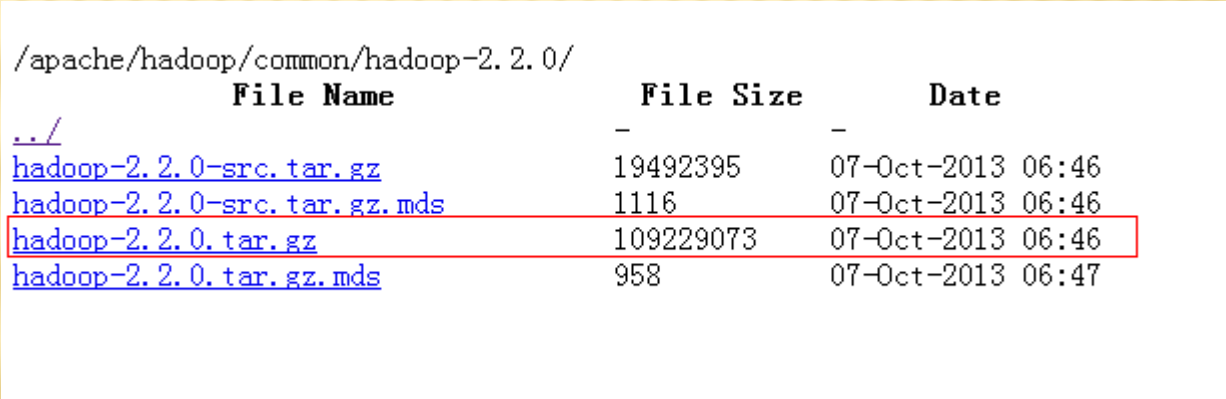


安装配置HADOOP

- 安装jdk，这个相对容易，请大家参考网上教程
- 主要步骤为
  - 下载对应于自己系统的包，如果为源码包，就解压编译，如果是编译过的包就直接解压，如果为rpm，bin结尾的直接用相关命令运行安装
  - 设置环境变量

安装配置HADOOP（单节点安装）

- Linux系统（自己选择虚拟机）
- 下载好编译的包
  - http://mirrors.hust.edu.cn/apache/hadoop/common/hadoop-2.2.0/



/apache/hadoop/common/hadoop-2.2.0/

File Name	File Size	Date
../	-	-
hadoop-2.2.0-src.tar.gz	19492395	07-Oct-2013 06:46
hadoop-2.2.0-src.tar.gz.mds	1116	07-Oct-2013 06:46
hadoop-2.2.0.tar.gz	109229073	07-Oct-2013 06:46
hadoop-2.2.0.tar.gz.mds	958	07-Oct-2013 06:47

安装配置HADOOP

- 将下载好的文件复制到linux系统上，解压并进入到解压目录
  - tar -zxvf hadoop-2.2.0.tar.gz
  - cd hadoop-2.2.0
- 目录是这个样子的

```
[root@localhost hadoop-2.2.0]# ls -la
总用量 64
drwxr-xr-x. 10 67974 users  4096 12月  9 15:57 .
drwxr-xr-x.  3 root  root   4096 12月  9 14:44 ..
drwxr-xr-x.  2 67974 users  4096 10月  7 14:38 bin
drwxr-xr-x.  3 67974 users  4096 10月  7 14:38 etc
drwxr-xr-x.  2 67974 users  4096 10月  7 14:38 include
drwxr-xr-x.  3 67974 users  4096 10月  7 14:38 lib
drwxr-xr-x.  2 67974 users  4096 10月  7 14:38 libexec
-rw-r--r--.  1 67974 users 15164 10月  7 14:46 LICENSE.txt
drwxr-xr-x.  3 root  root   4096 12月  9 18:35 logs
-rw-r--r--.  1 67974 users   101 10月  7 14:46 NOTICE.txt
-rw-r--r--.  1 67974 users  1366 10月  7 14:46 README.txt
drwxr-xr-x.  2 67974 users  4096 10月  7 14:38 sbin
drwxr-xr-x.  4 67974 users  4096 10月  7 14:38 share
```

HADOOP目录结构

- 目录中的文件简要说明:
  - bin：命令文件，命令使用参考
  - sbin：脚本文件，用于启动、关闭hadoop进程
  - etc：主要存放配置文件
  - share：文档及jar库
  - include & lib: 用于c/c++ 调用hadoop库函数
  - libexec: 主要存放生成系统变量及配置文件属性脚本

配置HADOOP

- 从上个目录进入配置文件目录
  - cd etc/hadoop/
- 主要配置core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml, hadoop-env.sh
  - 使用文本编辑器打开对应的文件，参照附件中修改即可

启动HADOOP

- 进入到hadoop主目录
- 格式化namenode
  - bin/hadoop namenode –format
- 启动hadoop所需进程
  - sbin/start-all.sh

- 查看进程是否启动，
  - jps
  - 结果：

```
5673 Jps
4986 NodeManager
3447 NameNode
3687 SecondaryNameNode
4888 ResourceManager
3528 DataNode
```

执行任务测试（单词计数）

- 进入hadoop主目录，随便拷贝一个文本文件到hdfs
  - bin/hdfs dfs -copyFromLocal /home/youlc/java/HelloWorld.java /in
- 提交单词计数作业（运行结果见备注）
  - bin/hadoop jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples-2.2.0.jar wordcount /in ./output

实验内容

- 理解hadoop框架
- 自己实现单词计数源码，在装有hadoop环境的主机上编译后并运行。
  - （可参考http://wiki.apache.org/hadoop/）