# Hadoop 集群从无到有

## 1、下载

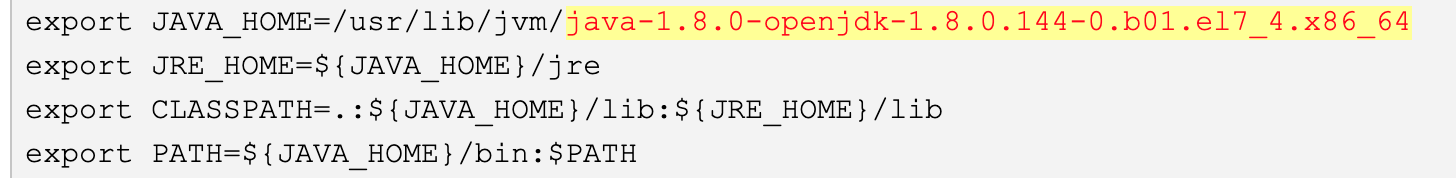
Cenos7、jdk8、mysql5.7，hadoop-2.7.2

## 2、安装

1. 虚拟机安装Linux
2. 创建路径/soft
3. 修改路径所属，chown -R vova:vova /soft
4. 创建文件夹/soft/zip
5. 导入外部安装包，jdk，mysql，hadoop-2.7.2
6. 解压jdk，mysql，hadoop到/soft

## 3、环境

1. javahome：/etc/profile



export JAVA\_HOME=/soft/jdk1.8.0\_121/

export PATH=$JAVA\_HOME/bin:$PATH

export CLASSPATH=.:$JAVA\_HOME/lib/dt.jar:$JAVA\_HOME/lib/tools.jar

## 4、Hadoop组成

1. HADOOP核心进程
   1. Namenode（1）
      1. 职责：负责接收客户端的所有请求
      2. 负责接收DN上报的块信息
      3. 负责向DN分配任务，例如维护文件的副本数
   2. Datenode（N）
   3. secondaryNamenode（N）
2. YARN中的核心进程
   1. ResourceManager（1）负责整个集群所有资源的管理和调度！
      1. 负责接收客户端的所有请求
      2. 负责接收NM上报的块信息
      3. 负责向NM分配任务，例如检查NM是否健康，是否在线等。
   2. NodeManager（N）负责当前机器所有资源的管理和调度
3. MapReduce核心进程，是一个编程模型，分为map阶段和reduce阶段

## **Hadoop**安装

1. 环境要求
   1. 必须安装JDK以及JAVA\_HOME
2. 安装
   1. 加压在linux下进行编译
3. 配置hadoop\_home
   1. 配置HADOOP\_HOME
   2. 配置hadoop中的bin，sbin
4. 目录结构
   1. bin:常用的hadoop工具
   2. sbin:集群管理，启动和停止
   3. etc:默认的配置文件目录

## Hadoop配置文件

core-site.xml------core-default.xml

hdfs-site.xml------hdfs-default.xml

yarn-site.xml------yarn-default.xml

mapred-site.xml------mapred-default.xml

配置文件位置：hadoop --confdir

默认配置文件目录：$HADOOP\_HOME/etc/hadoop

## HDFS运行模式

1. 本地模式 fs.defaultFS=file://
2. 分布式文件系统 fs.defaultFS=hdfs://NN所在的主机名:9000

## 启动一个分布式文件系统

1. 在$HADOOP\_HOME/etc/hadoop,配置core-site.xml
   * 1. fs.defaultFS=hdfs://NN主机名:9000
2. 配置Hadoop默认的工作目录，在$HADOOP\_HOME/etc/hadoop
   * 1. 配置core-site.xml
     2. hadoop.tmp.dir=配置一个当前用户有写权限的非tmp目录
3. 格式化NN
   * 1. hadoop namenode -format
     2. 目的：
     3. 生成NN的工作目录
     4. 在工作目录下生成NN所要使用的特殊的文件，例如version，fsimage0000
     5. 注意：一个集群搭建完成后，只需要格式化一次。
4. 启动
   * 1. hadoop-daemon.sh start namenode |datanode
5. MR的运行模式
   * 1. 本地模式：在本机使用多线程方式模拟多个task的运行
     2. 分布式模式：在yarn上运行，mapreduce.framework.name=yarn
     3. 配置yarn
        1. yarn-site.xml
        2. yarn.resourcemanager.hostname=RM运行的主机名
        3. yarn.xxx-auxservice=mapreduce\_shuffle
     4. 启动yarn，yarn-daemon.sh start resourcemanager | nodemanager
     5. 查看，jps，http://rm所运行的主机名：8088
     6. 提交作业 hadoop jar xxx.jar 主类名 输入目录。。 输出目录
        1. 要求
           1. 输出目录必须不存在
           2. 输入目录必须全部是文件

## 完全分布式集群

1. 准备集群，安装软件
   * 1. scp（安全拷贝）
     2. rsync（远程同步）rsync -rvlt path1 目标文件path
     3. 免密登录（evernote）
2. 编辑/.bashrc
   * 1. 增加source /etc/profile
3. 修改HADOOP\_HOME/etc/hadoop/slaves
4. 修改core-site.xml并分发

|  |
| --- |
| <!-- 指定HDFS中NameNode的地址 -->  <property>  <name>fs.defaultFS</name>  <value>hdfs://h11:9000</value>  </property>  <!-- 指定Hadoop运行时产生文件的存储目录 -->  <property>  <name>hadoop.tmp.dir</name>  <value>/soft/hadoop-2.7.2/data/tmp</value>  </property> |

1. 修改hdfs-site.xml

|  |
| --- |
| <configuration>  <!-- 指定Hadoop辅助名称节点主机配置 -->  <property>  <name>dfs.namenode.secondary.http-address</name>  <value>h13:50090</value>  </property>  </configuration> |

1. 格式化NN
   * 1. hadoop namenode -format
     2. 目的：
     3. 生成NN的工作目录
     4. 在工作目录下生成NN所要使用的特殊的文件，例如version，fsimage0000
     5. 注意：一个集群搭建完成后，只需要格式化一次。
2. 配置yarn-site.xml

|  |
| --- |
| <!-- Site specific YARN configuration properties -->  <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property>  <!-- 指定YARN的ResourceManager的地址 -->  <property>  <name>yarn.resourcemanager.hostname</name>  <value>h12</value>  </property>  <!-- 日志聚集功能使能 -->  <property>  <name>yarn.log-aggregation-enable</name>  <value>true</value>  </property>  <!-- 日志保留时间设置7天 -->  <property>  <name>yarn.log-aggregation.retain-seconds</name>  <value>604800</value>  </property> |

1. 启动
   * 1. start-dfs.sh start-yarn.sh 或者start-all.sh
2. 验证
   * 1. hadoop fs -mkdir /vova
     2. 访问50070
     3. hadoop fs -put hello /vova/（需要检查各个DN的防火墙）
     4. MR：/soft/hadoop-2.7.2/share/hadoop/mapreduce
     5. hadoop jar hadoop-mapreduce-examples-2.7.2.jar wordcount /vova /vovaout
     6. hadoop fs -cat /vovaout/part-r-00000（拿到结果）
3. 日志配置

|  |
| --- |
| mapred-site.xml  <configuration>  <property>  <name>mapreduce.framework.name</name>  <value>yarn</value>  </property>  <property>  <name>mapreduce.jobhistory.address</name>  <value>h11:10020</value>  </property>  <property>  <name>mapreduce.jobhistory.webapp.address</name>  <value>h11:19888</value>  </property>  <!--第三方框架使用yarn计算的日志聚集功能 -->  <property>  <name>yarn.log.server.url</name>  <value>http://h11:19888/jobhistory/logs</value>  </property>  </configuration> |
| yarn-site.xml  <!-- 日志聚集功能使能 -->  <property>  <name>yarn.log-aggregation-enable</name>  <value>true</value>  </property>  <!-- 日志保留时间设置7天 -->  <property>  <name>yarn.log-aggregation.retain-seconds</name>  <value>604800</value>  </property> |

1. 启动日志服务器
   * 1. ./mr-jobhistory-daemon.sh start historyserver