1. 环境准备

集群有三台机器：

* master：W118PC01VM01/192.168.0.112
* slave1：W118PC02VM01/192.168.0.113
* slave2：W118PC03VM01/192.168.0.114

首先配置/etc/hosts中ip和主机名的映射关系：

* 192.168.0.112 W118PC01VM01
* 192.168.0.113 W118PC02VM01
* 192.168.0.114 W118PC03VM01

其次配置3台机器互相免密码ssh连接，参考《在Win7虚拟机下搭建Hadoop2.6.0伪分布式环境》。

2. 基本安装步骤

（1） 安装Java（本例使用jdk1.7.0\_55）和Scala（使用scala2.10.4）。

（2） 安装Hadoop2.6.0集群。

（3） 安装Spark1.1.0集群。

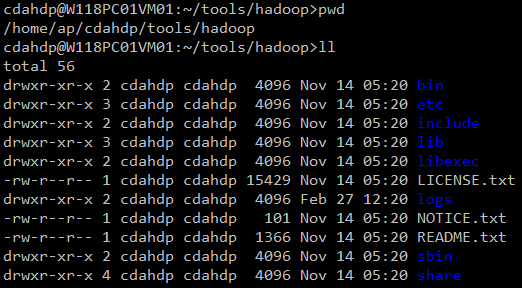
3. Jdk和Scala安装

在master和slave机器的安装路径和环境变量配置保持一致。安装过程参考《在Win7虚拟机下搭建Hadoop2.6.0+Spark1.4.0单机环境》。

4. Hadoop集群安装

4.1. 安装Hadoop并配置环境变量

安装Hadoop2.6.0版本，安装目录如下。在~/.bash\_profile中配置环境变量，参考《在Win7虚拟机下搭建Hadoop2.6.0伪分布式环境》。



4.2. 修改Hadoop配置文件

涉及到的hadoop配置文件主要有以下7个：

1. /home/ap/cdahdp/tools/hadoop/etc/hadoop/hadoop-env.sh
2. /home/ap/cdahdp/tools/hadoop/etc/hadoop/yarn-env.sh
3. /home/ap/cdahdp/tools/Hadoop/etc/hadoop/slaves
4. /home/ap/cdahdp/tools/hadoop/etc/hadoop/core-site.xml
5. /home/ap/cdahdp/tools/hadoop/etc/hadoop/hdfs-site.xml
6. /home/ap/cdahdp/tools/hadoop/etc/hadoop/mapred-site.xml
7. /home/ap/cdahdp/tools/hadoop/etc/hadoop/yarn-site.xml

配置 hadoop-env.sh（修改JAVA\_HOME）

# The java implementation to use.

export JAVA\_HOME=/home/ap/cdahdp/tools/jdk1.7.0\_55

配置 yarn-env.sh （修改JAVA\_HOME）

# some Java parameters

export JAVA\_HOME=/home/ap/cdahdp/tools/jdk1.7.0\_55

配置slaves（增加slave节点）

W118PC02VM01

W118PC03VM01

配置 core-site.xml（增加hadoop核心配置）

<configuration>  
    <property>  
        <name>fs.defaultFS</name>  
        <value>hdfs://192.168.0.112:9000</value>  
    </property>  
    <property>  
        <name>io.file.buffer.size</name>  
        <value>131072</value>  
    </property>  
    <property>  
        <name>hadoop.tmp.dir</name>  
        <value>file:/home/ap/cdahdp/app/tmp</value>  
        <description>Abasefor other temporary directories.</description>  
    </property>  
    <property>  
        <name>hadoop.proxyuser.spark.hosts</name>  
        <value>\*</value>  
    </property>  
    <property>  
        <name>hadoop.proxyuser.spark.groups</name>  
        <value>\*</value>  
    </property>  
</configuration>

配置hdfs-site.xml（增加hdfs配置信息，namenode、datanode端口和目录位置）

<configuration>  
    <property>  
        <name>dfs.namenode.secondary.http-address</name>  
        <value>192.168.0.112:9001</value>  
    </property>  
    <property>  
        <name>dfs.namenode.name.dir</name>  
        <value>file:/home/ap/cdahdp/app/name</value>  
    </property>  
    <property>  
        <name>dfs.datanode.data.dir</name>  
        <value>file:/home/ap/cdahdp/app/data</value>  
    </property>  
    <property>  
        <name>dfs.replication</name>  
        <value>3</value>  
    </property>  
    <property>  
        <name>dfs.webhdfs.enabled</name>  
        <value>true</value>  
    </property>  
    <property>  
        <name>dfs.datanode.du.reserved</name>【配置磁盘中保留不用作hdfs集群的空间大小，单位是Byte】  
        <value>10240000000</value>  
    </property>  
</configuration>

配置mapred-site.xml（增加mapreduce配置，使用yarn框架、jobhistory地址以及web地址）

<configuration>  
    <property>  
        <name>mapreduce.framework.name</name>  
        <value>yarn</value>  
    </property>  
    <property>  
        <name>mapreduce.jobhistory.address</name>  
        <value>192.168.0.112:10020</value>  
    </property>  
    <property>  
        <name>mapreduce.jobhistory.webapp.address</name>  
        <value>192.168.0.112:19888</value>  
    </property>  
</configuration>

配置 yarn-site.xml（增加yarn功能）

<configuration>  
    <property>  
        <name>yarn.nodemanager.aux-services</name>  
        <value>mapreduce\_shuffle</value>  
    </property>  
    <property>  
        <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>  
        <value>org.apache.hadoop.mapred.ShuffleHandler</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.address</name>  
        <value>192.168.0.112:8032</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.scheduler.address</name>  
        <value>192.168.0.112:8030</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.resource-tracker.address</name>  
        <value>192.168.0.112:8035</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.admin.address</name>  
        <value>192.168.0.112:8033</value>  
    </property>  
    <property>  
        <name>yarn.resourcemanager.webapp.address</name>  
        <value>192.168.0.112:8088</value>  
    </property>  
</configuration>

将配置好的hadoop文件copy到两台slave机器上，配置和路径和master一模一样。

4.3. 格式化namenode

在master和slave机器上分别操作：

cd ~/tools/hadoop/bin

./hdfs namenode -format

4.4. 启停hdfs和yarn

cd ~/tools/hadoop/sbin

./start-hdfs.sh

./stop-hdfs.sh

./start-yarn.sh

./stop-yarn.sh

启动后可以用jps查看进程，通常有这几个：

NameNode、SecondaryNameNode、ResourceManager、DataNode

如果启动异常，可以查看日志，在master机器的/home/ap/cdahdp/tools/hadoop/logs目录。

4.5. 查看集群状态

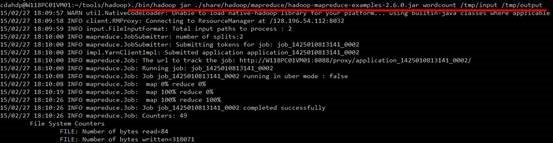
查看hdfs：http://192.168.0.112:50070/

查看RM：http://192.168.0.112:8088/

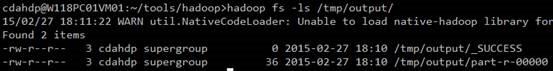
4.6. 运行wordcount示例程序

上传几个文本文件到hdfs，路径为/tmp/input/

之后运行：



查看执行结果：

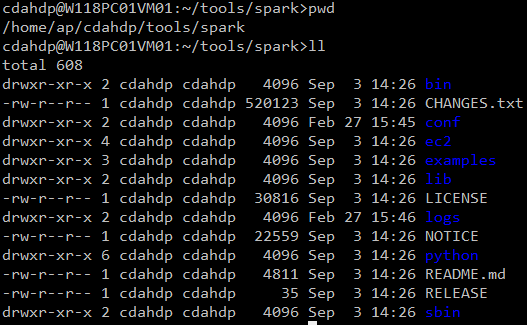


正常运行，表示hadoop集群安装成功。

5. Spark集群部署

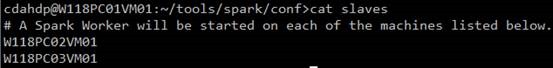
5.1. 安装Spark并配置环境变量

安装Spark1.1.0版本，安装目录如下。在~/.bash\_profile中配置环境变量。



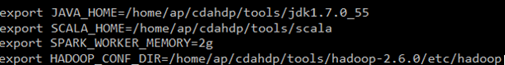
5.2. 修改Hadoop配置文件

配置slaves（增加slave节点）



配置spark-env.sh（设置spark运行的环境变量）

把spark-env.sh.template复制为spark-env.sh



将配置好的spark文件copy到两台slave机器上，配置和路径和master一模一样。

5.3. Spark的启停

cd ~/tools/spark/sbin

./start-all.sh

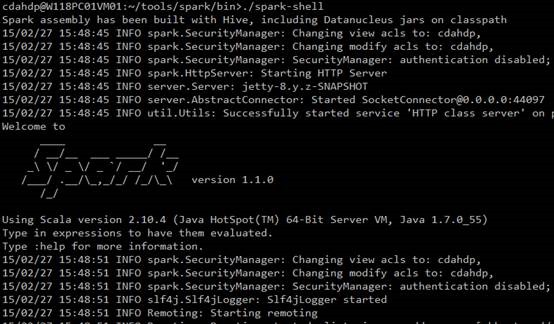
./stop-all.sh

5.4. 查看集群状态

spark集群的web管理页面：http://192.168.0.112:8080/

spark WEBUI页面：http://192.168.0.112:4040/

启动spark-shell控制台：



5.5. 运行示例程序

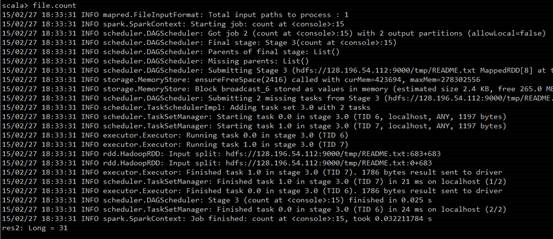
往hdfs上上传一个文本文件README.txt：



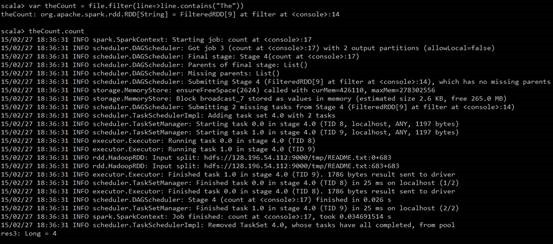
在spark-shell控制台执行：

统计README.txt中有多少单词：

https://www.linuxidc.com/upload/2015_08/150802174682589.jpg



过滤README.txt包括The单词有多少行：



正常运行，表示Spark集群安装成功。