Spark配置

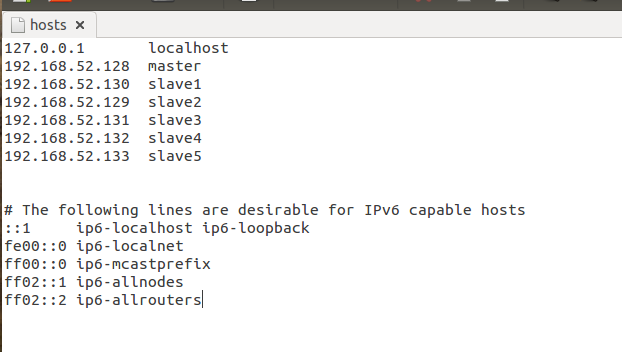
# 1.所需环境：

1. Ubuntu
2. Hadoop
3. Scala 下载地址：http://www.scala-lang.org/download/2.10.4.html
4. JDK
5. ssh

# 2.前期准备工作

确保spark集群中所有机器能够互相访问，都在同一个网段内，直接修改六台机器的hosts文件

sudo gedit /etc/hosts



实际IP根据实际情况修改（第二列要与下面所述的slaves文件相同）

# 3.修改环境变量

修改环境变量打开终端输入命令sudo gedit /etc/profile 添加JDK,HADOOP,SCALA,SPARK环境变量进去

export JAVA\_HOME=/usr/java/jdk

export HADOOP\_HOME=/opt/Hadoop

export SPARK\_HOME=/usr/spark

export SCALA\_HOME=/usr/scala-2.11.4

export CLASSPATH=$JAVA\_HOME/lib:$JRE\_HOME/lib:$CLASSPATH

export PATH=$JAVA\_HOME/bin:$JRE\_HOME/bin:$HADOOP\_HOME/sbin:${SCALA\_HOME}/bin:${SPARK\_HOME}/bin:$PATH

为了能使虚机启动后环境变量就直接可用，我将jdk和hadoop直接配置到~/.bashrc文件中

export JAVA\_HOME=/usr/java/jdk

export SCALA\_HOME=/usr/scala-2.11.4

export SPARK\_HOME=/usr/spark

export CLASSPATH=.:$CLASSPATH:$JAVA\_HOME/lib:$JAVA\_HOME/jre/lib

export PATH=$PATH:$JAVA\_HOME/bin:$JAVA\_HOME/jre/bin:/usr/local/hadoop/hadoop-2.5.2/bin:$SCALA\_HOME/bin:$SPARK\_HOME/bin

# 4.安装和配置scala

先去scala和spark的官网下载scala-2.11.4.tgz、spark-1.2.0-bin-hadoop2.4.tgz

安装scala

创建目录

mkdir /usr/ scala-2.11.4

在新建的目录中解压scala

tar zxvf scala-2.11.4.tgz

将scala添加到环境变量中，

sudo gedit ~/.bashrc

添加如下内容

export SCALA\_HOME=/usr/scala-2.11.4

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

使用source命令是~/.bashrc生效

source ~/.bashrc

测试scala是否安装成功

scala –version

如出现以下提示则便是安装成功



# 5.配置spark

## 5.1 安装spark

将下载好的spark-1.2.0-bin-hadoop2.4.tgz解压到 /usr/spark文件夹中

在环境变量中添加spark路径

export SPARK\_HOME=/usr /spark

export PATH=$PATH:$SCALA\_HOME/bin:$SPARK\_HOME/bin

如果要把jdk和hadoop也一起加入的话，那最后结果为

export JAVA\_HOME=/usr/java/jdk

export SCALA\_HOME=/usr/scala-2.11.4

export SPARK\_HOME=/usr/spark

export CLASSPATH=.:$CLASSPATH:$JAVA\_HOME/lib:$JAVA\_HOME/jre/lib

export PATH=$PATH:$JAVA\_HOME/bin:$JAVA\_HOME/jre/bin:/usr/local/hadoop/hadoop-2.5.2/bin:$SCALA\_HOME/bin:$SPARK\_HOME/bin

使用source命令使文件生效

source ~/.bashrc

## 5.2 配置spark

配置spark-en.sh文件

在spark安装目录下的conf文件夹下修改spark-env.sh文件（如果不存在此文件则复制同目录下的spark-env.sh.template文件并修改文件名称为spark-env.sh）

cd /usr/spark/conf

cp spark-env.sh.template spark-env.sh

修改此文件，在文件末尾添加以下内容

export JAVA\_HOME=/usr/java/jdk

###java安装目录

export SCALA\_HOME=/usr/scala-2.11.4

###scala安装目录

export SPARK\_MASTER\_IP=192.168.52.128

###集群中master机器IP

export SPARK\_WORKER\_MEMORY=2g

###指定的worker节点能够最大分配给Excutors的内存大小

export HADOOP\_CONF\_DIR=/opt/hadoop/etc/Hadoop

###hadoop集群的配置文件目录

配置slaves文件（如果没有复制slaves.template文件并修改文件名称为slaves）

将worker节点的全部添加进去（我的有6个节点）

master

slave1

slave2

slave3

slave4

slave5

保存退出，这样master中spark就安装完成了，所有slave节点和maser一样的设置

# 6.设置master对slave机器免密码SSH登录

## 6.1 SSH的安装

执行sudo apt-get install ssh命令等待SSH安装完成

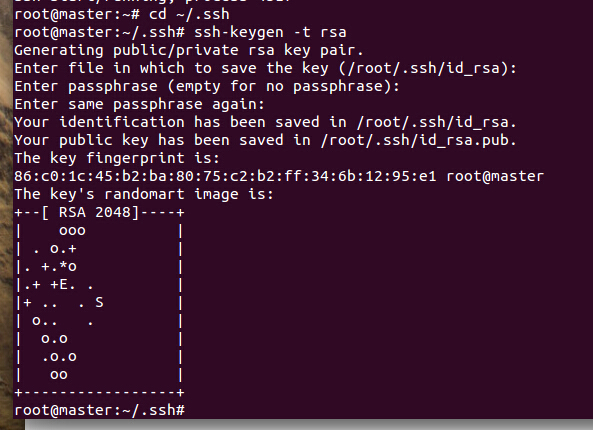
## 6.2 生成公钥

在master机上执行命令

cd ~/.ssh

ssh-keygen –t rsa

因为要配置免密码登录，一直点回车就好

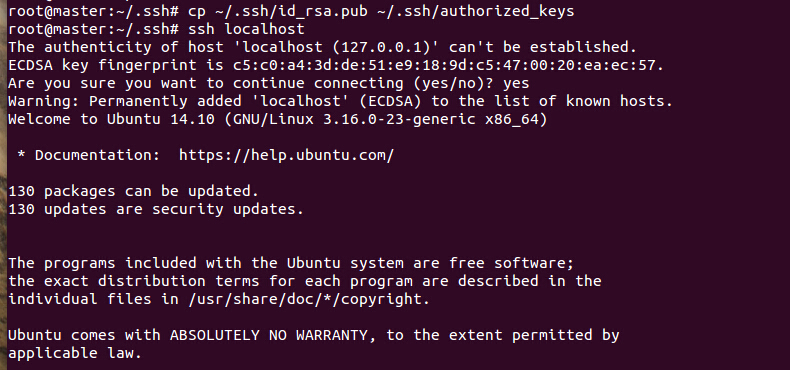


执行cp ~/.ssh/id\_rsa.pub ~/.ssh/authorized\_keys

执行完成后应该就能免密码登录本机了，输入

ssh localhost进行测试

出现以下提示则表示成功

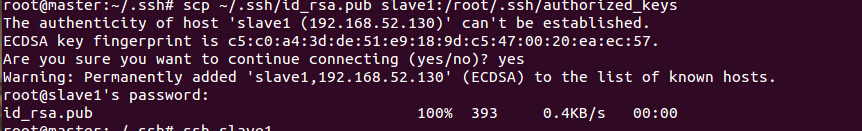


## 6.3 把生成的公钥传到slave机中

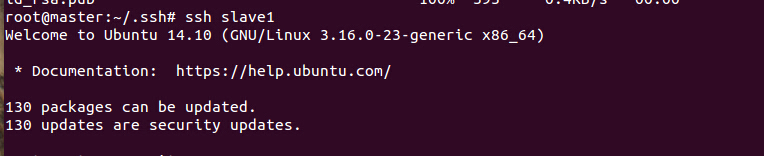
scp ~/.ssh/id\_rsa.pub slave1:/root/.ssh/authorized\_keys

（我的ubuntu系统全部设置的root用户登录所以拷贝到/root/.ssh/authorized\_keys下，如果是其他用户则相应修改为/home/用户名/.ssh/authorized\_keys）

输入slave1的root用户密码，点回车



此时master机已经可以免密码登录slave1机器了



（第一次登录会提示（yes/no）输入yes点回车即可）

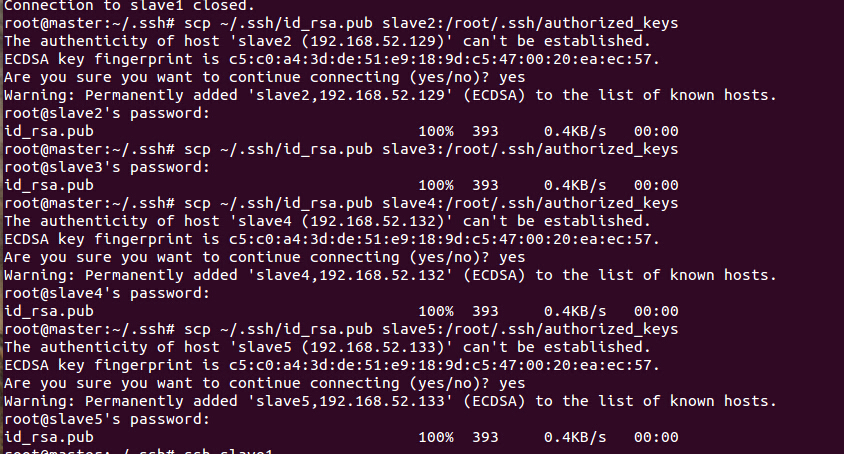
相同的方法把生成的公钥传到其他slave机器中

scp ~/.ssh/id\_rsa.pub slave2:/root/.ssh/authorized\_keys

scp ~/.ssh/id\_rsa.pub slave3:/root/.ssh/authorized\_keys

scp ~/.ssh/id\_rsa.pub slave4:/root/.ssh/authorized\_keys

scp ~/.ssh/id\_rsa.pub slave5:/root/.ssh/authorized\_keys



测试：

ssh slave1

ssh slave2

……

（测试完一台slave机器准备测试下一台时要用exit命令退出登录）

# 7.测试

在master机中执行

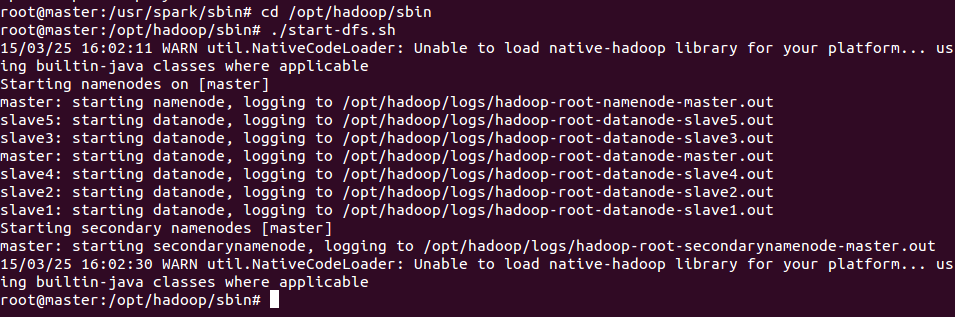
配置完成后即可启动spark集群了

首先CD至hadoop目录下的sbin文件夹中

cd /opt/hadoop/sbin

启动hdfs

./start-dfs.sh

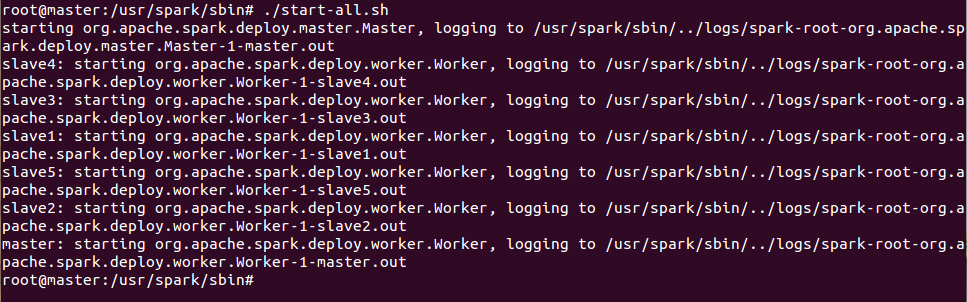


其次CD至spark目录下的sbin文件夹中

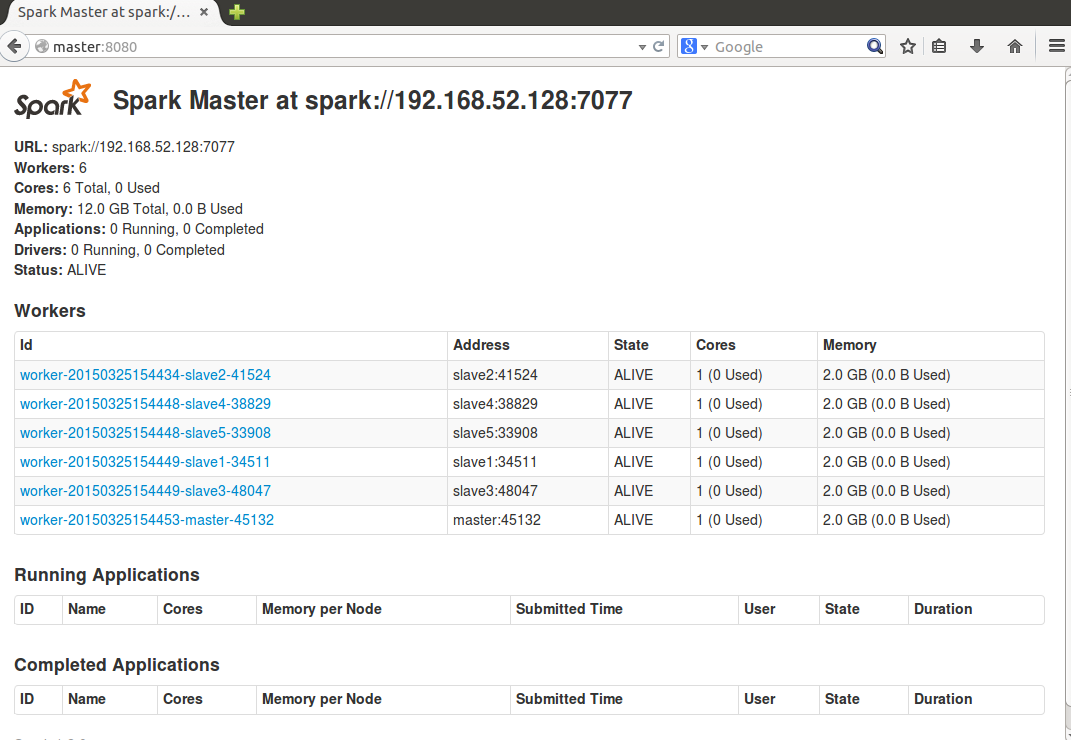
cd /usr/spark/sbin

启动spark集群

./start-all.sh

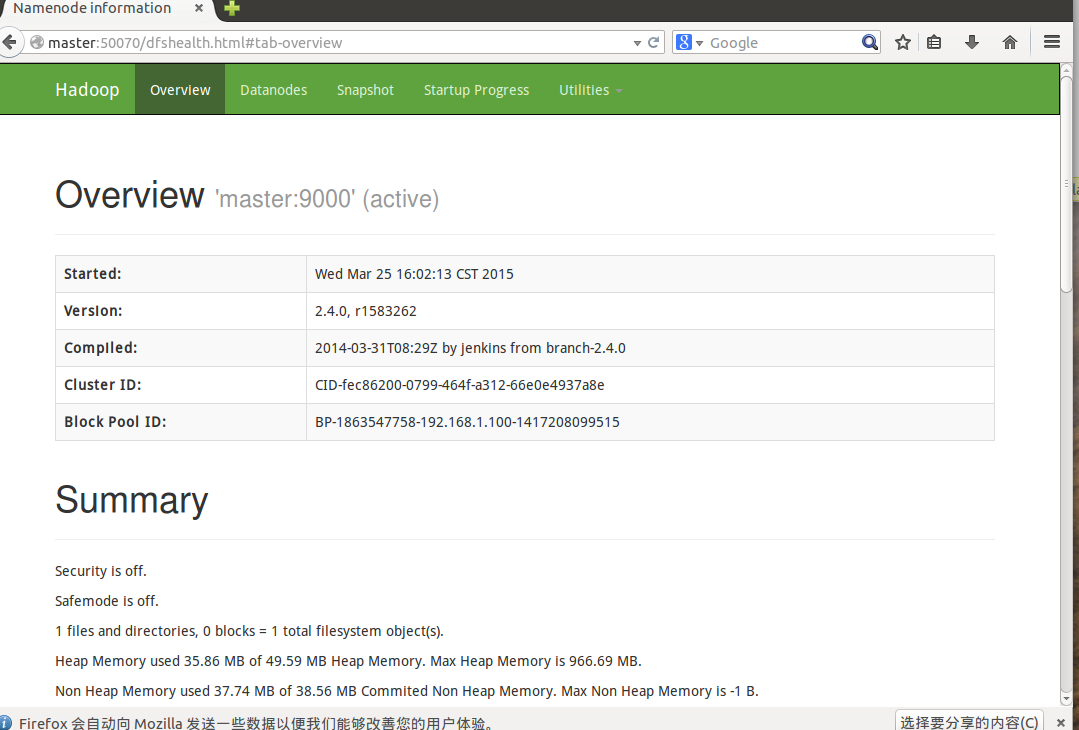


启动完成后可在浏览器内出入地址 master：8080进行查看worker状态

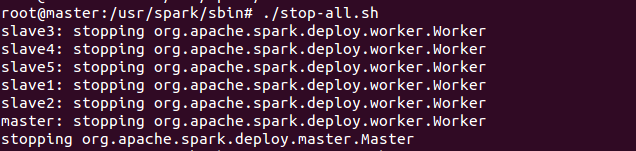


成功后页面会有6个workers的ID

在master：50070查看dfs状态



停止spark集群



# 8.其他注意问题（所遇错误）

## 8.1 设置SSH免密码登录slave机器的root用户是提示Permission denied, please try again.

说明是机器权限问题，在要登录的slave机器上执行vi /etc/ssh/sshd\_config命令修改一下内容

UsePrivilegeSeparation no

PermitRootLogin yes

PermitEmptyPasswords no（如果前边有#则去掉）

重启SSH服务 service ssh restart

## 8.2 启动spark集群后进web查看只有master节点或者缺少节点

考虑是不是某一台机器的hostname问题

在逐台机器上执行命令vi /etc/hostname命令查看机器的hostname，根据spark下conf文件夹下的slaves文件中的内容中修改为相应hostname，重启机器生效