Anexo 7: Manual de instalación Hadoop

Fuentes:
https://www.youtube.com/watch?v=phHVGOwvhKA
http://www.highlyscalablesystems.com/3597/hadoop-installation-tutorial-hadoop-2-x/#core-site.xml
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html
http://disi.unitn.it/~lissandrini/notes/installing-hadoop-on-ubuntu-14.html

Para editar el nombre del equipo
bigtexts@bigtexts-1:~$ sudo nano /etc/hostname
bigtexts@bigtexts-1:~$ sudo nano /etc/hosts
Cerrar la consola y volverla a abrir

Actualizar el repositorio de Ubuntu
bigtexts@bigtexts-1:~$ sudo apt-get update

Instalar Java
bigtexts@bigtexts-1:~$ sudo apt-get install openjdk-7-jdk
bigtexts@bigtexts-1:~$ java -version
Instalar SSH
bigtexts@bigtexts-1:~$ sudo apt-get install openssh-server
bigtexts@bigtexts-1:~$ sudo addgroup hadoop
bigtexts@bigtexts-1:~$ sudo adduser --ingroup hadoop hduser

Nota: contraseña -> 123456

bigtexts@bigtexts-1:~$ sudo adduser hduser sudo
bigtexts@bigtexts-1:~$ su -l hduser

hduser@bigtexts-1:~$ ssh-keygen -t rsa -P ""
hduser@bigtexts-1:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hduser@bigtexts-1:~$ ssh localhost
hduser@bigtexts-1:~$ ssh localhost

Deshabilitar IPv6
http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
http://askubuntu.com/questions/346126/how-to-disable-ipv6-on-ubuntu

wilson@wilson-VirtualBox:~$ sudo su -
root@wilson-VirtualBox:~# nano /etc/sysctl.conf

Al final del archivo pegar lo siguiente:

#disable ipv6
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1

root@bigtexts-1:~# sudo sysctl -p
root@bigtexts-1:~# cat /proc/sys/net/ipv6/conf/all/disable_ipv6
si el resultado es 1 entonces ha quedado deshabilitado

Instalar Hadoop
root@bigtexts-1:~# su -l hduser

Descargamos Hadoop:
hduser@bigtexts-1:~$ wget
http://mirror.sdunix.com/apache/hadoop/common/hadoop-2.5.0/hadoop-2.5.0.tar.gz

hduser@bigtexts-1:~$ sudo tar -xzvf hadoop-2.5.0.tar.gz -C /usr/local/
hduser@bigtexts-1:~$ cd /usr/local/
hduser@bigtexts-1:/usr/local$ sudo ln -s hadoop-2.5.0/ hadoop
hduser@bigtexts-1:/usr/local$ sudo chown -R hduser:hadoop hadoop
hduser@bigtexts-1:/usr/local$ sudo chown -R hduser:hadoop hadoop-2.5.0/

hduser@bigtexts-1:/usr/local$ cd ~

hduser@bigtexts-1:~$ nano .bashrc
Pegar al final del archivo

Set Hadoop-related environment variables
#Java path
export JAVA_HOME='/usr/lib/jvm/java-1.7.0-openjdk-amd64'
export HADOOP_INSTALL=/usr/local/hadoop
Add Hadoop bin/ directory to PATH
export
PATH=$PATH:$HADOOP_INSTALL/bin:$JAVA_HOME/bin:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=${HADOOP_INSTALL}
export HADOOP_COMMON_HOME=${HADOOP_INSTALL}
export HADOOP_HDFS_HOME=${HADOOP_INSTALL}

export YARN_HOME=${HADOOP_INSTALL}

hduser@bigtexts-1:~$ cd /usr/local/hadoop/etc/hadoop/
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano hadoop-env.sh

cambiar la linea
export JAVA_HOME=${JAVA_HOME}
a
export JAVA_HOME=/usr/lib/jvm/java-1.7.0-openjdk-amd64

cerrar y abrir nuevamente la consola
bigtexts@bigtexts-1:~$ su -l hduser
hduser@bigtexts-1:~$ hadoop version
se debe ver algo asi

Hadoop 2.5.0
Subversion http://svn.apache.org/repos/asf/hadoop/common -r 1616291
Compiled by jenkins on 2014-08-06T17:31Z
Compiled with protoc 2.5.0
From source with checksum 423dcd5a752eddd8e45ead6fd5ff9a24
This command was run using
/usr/local/hadoop/share/hadoop/common/hadoop-common-2.5.0.jar

Configuración Hadoop
hduser@bigtexts-1:~$ cd /usr/local/hadoop/etc/hadoop/
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano core-site.xml

```
<configuration>
        <property>
        <name>fs.default.name</name>
        <value>hdfs://localhost:9000</value>
        </property>
</configuration>
```

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano yarn-site.xml

```
<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.hostname</name>
                <!-- hostname that is accessible from all NMs -->
                <value>master</value>
        </property>
</configuration>

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ cp mapred-site.xml.template
mapred-site.xml
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano mapred-site.xml

<configuration>
        <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
        </property>
</configuration>
```

Crear las carpetas para HDFS

```
hduser@bigtexts-1:~$ mkdir -p /usr/local/hadoop/yarn_data/hdfs/namenode
hduser@bigtexts-1:~$ mkdir -p /usr/local/hadoop/yarn_data/hdfs/datanode
hduser@bigtexts-1:~$ cd /usr/local/hadoop/etc/hadoop/
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano hdfs-site.xml

<configuration>
        <property>
        <name>dfs.replication</name>
        <value>1</value>
        </property>
        <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:/usr/local/hadoop/yarn_data/hdfs/namenode</value>
        </property>
        <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:/usr/local/hadoop/yarn_data/hdfs/datanode</value>
        </property>
</configuration>
```

Formatear el name node

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ hdfs namenode -format

Iniciar los servicios

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ start-dfs.sh
Para ver los servicios:
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ jps

Se debe ver algo así
11547 NameNode
11968 Jps
11663 DataNode
11860 SecondaryNameNode

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ start-yarn.sh

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ jps
se debe ver algo asi

5948 SecondaryNameNode
5752 DataNode
6096 ResourceManager
5631 NameNode
6522 Jps
6218 NodeManager

Interfaces Web

- HDFS Namenode and check health using http://localhost:50070
- HDFS Secondary Namenode status using http://localhost:50090

Make the HDFS directories required to execute MapReduce jobs:
hduser@bigtexts-1:/usr/local/hadoop$ hdfs dfs -mkdir /user
hduser@bigtexts-1:/usr/local/hadoop$ hdfs dfs -mkdir /user/hduser

Copy the input files into the distributed filesystem:
hduser@bigtexts-1:/usr/local/hadoop$ hdfs dfs -put etc/hadoop input

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html

Instalación de Hadoop en cluster

Fuentes:
http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/
http://raseshmori.wordpress.com/2012/10/14/install-hadoop-nextgen-yarn-multi-node-cluster/
http://www.elcct.com/installing-hadoop-2-3-0-on-ubuntu-13-10/
http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html

1. Parar todo en las dos máquinas

wilson@wilson-VirtualBox:~$ su -l hduser
hduser@wilson-VirtualBox:~$ stop-all.sh

2. Darle conectividad entre ellas estableciendo ips fijas

Fuente configuración VirtualBox:
http://coding4streetcred.com/blog/post/VirtualBox-Configuring-Static-IPs-for-VMs
Fuente configuración IP fija en Ubuntu:
http://askubuntu.com/questions/338442/how-to-set-static-ip-address

Nombre	Ip Address	Netmask	Gateway	DNS
master	192.168.0.1	255.255.255.0	192.168.0.254	192.168.0.254
slave	192.168.0.2	255.255.255.0	192.168.0.254	192.168.0.254

En las todas máquinas
wilson@wilson-VirtualBox:~$ sudo nano /etc/hosts

192.168.0.1 master
192.168.0.2 slave

192.168.0.101 master
192.168.0.102 slave-2
192.168.0.103 slave-3
192.168.0.104 slave-4
192.168.0.105 slave-5

Para conectarse sin contraseña desde el master al slave y viceversa

Desde el master
hduser@master:~$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hduser@slave

Desde el slave
hduser@slave:~$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hduser@master

3. Configuración (en el master)
hduser@bigtexts-1:~$ cd /usr/local/hadoop/etc/hadoop/
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano slaves

master
slave

master
slave-2
slave-3
slave-4
slave-5

conf/*-site.xml (all machines)

hduser@wilson-VirtualBox:/usr/local/hadoop/etc/hadoop$ nano core-site.xml
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano core-site.xml

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://master:9000</value>
    </property>
</configuration>
```

hdfs-site.xml(all machines)
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ nano hdfs-site.xml
hduser@wilson-VirtualBox:/usr/local/hadoop/etc/hadoop$ nano hdfs-site.xml

```
    <property>
        <name>dfs.replication</name>
        <value>2</value>
    </property>
    <property>
        <name>dfs.permissions</name>
        <value>false</value>
```

</property>

Formatear namenode (all machines)
hduser@wilson-VirtualBox:/usr/local/hadoop/etc/hadoop$ hadoop namenode -format
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ hadoop namenode -format

5. Iniciar el cluster

En el master
hduser@wilson-VirtualBox:/usr/local/hadoop$ start-dfs.sh
hduser@wilson-VirtualBox:/usr/local/hadoop$ start-yarn.sh
hduser@wilson-VirtualBox:/usr/local/hadoop$ jps

8279 DataNode
5782 ResourceManager
8485 SecondaryNameNode
9475 Jps
8850 NodeManager
8157 NameNode
En los slaves
hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ start-dfs.sh
5301 DataNode
5482 SecondaryNameNode
5823 Jps

6. Prueba de funcionamiento
hduser@wilson-VirtualBox:~$ mkdir in
hduser@wilson-VirtualBox:~$ cat > in/file
This is one line
This is another one

hduser@wilson-VirtualBox:~$ hadoop dfs -copyFromLocal in /in
hduser@wilson-VirtualBox:~$ cd /usr/local/hadoop
hduser@wilson-VirtualBox:/usr/local/hadoop$ hadoop jar
share/hadoop/mapreduce/hadoop-mapreduce-examples-2.*.jar wordcount /in /out
hduser@wilson-VirtualBox:/usr/local/hadoop$ hadoop dfs -cat /out/*

192.168.0.101 bigtexts-1
192.168.0.102 bigtexts-2
192.168.0.103 bigtexts-3
192.168.0.104 bigtexts-4
192.168.0.105 bigtexts-5

y en el slave

hduser@bigtexts-1:/usr/local/hadoop/etc/hadoop$ hadoop dfs -cat /out/*

7. Troubleshooting
si no llega a subir un name node o un data node

hduser@wilson-VirtualBox:~$ stop-all.sh
--Por Cada nodo
hduser@wilson-VirtualBox:~$ rm -Rf /usr/local/hadoop/yarn_data/hdfs/datanode/*
hduser@wilson-VirtualBox:~$ rm -Rf /usr/local/hadoop/yarn_data/hdfs/namenode/*
--Fin Por cada nodo
hduser@wilson-VirtualBox:~$ hadoop namenode -format
hduser@wilson-VirtualBox:~$ start-all.sh

Comandos

hadoop dfsadmin -report	List the namenode and datanodes of a cluster

```
<?xml version="1.0"?>
<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.resource-tracker.address</name>
    <value>master:8025</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>master:8030</value>
  </property>
```

```
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>master:8050</value>
  </property>
</configuration>
```

Parar especificamente
Por cada nodo
cd /usr/local/hadoop/sbin/
hduser@bigtexts-2:/usr/local/hadoop/sbin$ hadoop-daemon.sh --script hdfs stop datanode
hduser@bigtexts-2:/usr/local/hadoop/sbin$ yarn-daemon.sh stop nodemanager

En el master
hadoop dfsadmin -refreshNodes
yarn rmadmin -refreshNodes

Para listar los nodos activos en yarn
yarn node -list

Para consultar los nodos yarn activos
yarn node -list