Universidad EAFIT
ST0263: Tópicos Especiales en Telemática, 2016-2
Profesor: Edwin Montoya M - emontoya@eafit.edu.co

Montaje de hadoop 2.7.3

MÁQUINAS

```
10.131.137.X1hadoop-master
10.131.137.X2 hadoop-slave1
10.131.137.X3 hadoop-slave2
```

//Reemplazar X1, X2, X3 por sus respectivos números de máquinas asignadas

verificar actualizar el archivo /etc/hosts en cada una de las máquinas del cluster, ej:

```
hadoop@hadoop-master$vi /etc/hosts
10.131.137.X1 hadoop-master
10.131.137.X2 hadoop-slave1
10.131.137.X3 hadoop-slave2
```

// crear un usuario así en cada nodo del cluster.

user: hadoop
pass: ****

DESCARGAR EL HADOOP 2.7.3

```
$ wget http://www.eu.apache.org/dist/hadoop/common/hadoop-2.7.3/hadoop-2.7.3.tar.gz
```

tener en cada uno de las máquinas un usuario hadoop.

```
hadoop-master# adduser hadoop
hadoop-master# passwd hadoop
                    # Enter password: ****
```

en cada máquina realizar esto.

actualizar el archivo .bashrc: (actualice los valores de acuerdo a su instalación), probablemente deban revisar el directorio de Java para garantizar que apunta correctamente al lugar indicado.

```
# User specific aliases and functions
export JAVA_HOME=/usr/java/latest
export HADOOP_HOME=/home/hadoop/hadoop-2.7.3
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
```

Generar las claves ssh y distribuirlas:

```
hadoop@hadoop-master $ ssh-keygen -t rsa
hadoop@hadoop-master $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@hadoop-slave1
hadoop@hadoop-master $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@hadoop-slave2
hadoop@hadoop-master $ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@hadoop-slave3
```

archivos de configuración de hadoop:

en el directorio: $HADOOP_HOME/etc/hadoop

core-site.xml

```
<configuration>
<property>
  <name>fs.default.name</name>
              <value>hdfs://hadoop-master:9000</value>
</property>
    <property>
              <name>hadoop.tmp.dir</name>
              <value>/tmp</value>
              <description>A base for other temporary directories.</description>
    </property>
</configuration>
```

hdfs-core.xml

```
<configuration>
<property>
 <name>dfs.replication</name>
 <value>1</value>
</property>

<property>
```

```
  <name>dfs.name.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
</property>

<property>
  <name>dfs.data.dir</name>
        <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
</property>
<property>
        <name>hadoop.tmp.dir</name>
        <value>/tmp</value>
        <description>A base for other temporary directories.</description>
    </property>
</configuration>
```

yarn-site.xml

```
<configuration>

<!-- Site specific YARN configuration properties -->

<property>
  <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
 </property>

<property>
        <name>yarn.resourcemanager.resource-tracker.address</name>
        <value>hadoop-master:8025</value>
</property>
<property>
        <name>yarn.resourcemanager.scheduler.address</name>
        <value>hadoop-master:8035</value>
</property>
<property>
        <name>yarn.resourcemanager.address</name>
        <value>hadoop-master:8050</value>
</property>

<property>
        <name>yarn.scheduler.minimum-allocation-mb</name>
        <value>128</value>
        <description>Minimum limit of memory to allocate to each container request at the Resource Manager.</description>
    </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
        <value>2048</value>
        <description>Maximum limit of memory to allocate to each container request at the Resource Manager.</description>
  </property>
  <property>
    <name>yarn.scheduler.minimum-allocation-vcores</name>
        <value>1</value>
        <description>The minimum allocation for every container request at the RM, in terms of virtual CPU cores. Requests lower than this won't take effect, and the specified value will get allocated the minimum.</description>
  </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-vcores</name>
        <value>2</value>
        <description>The maximum allocation for every container request at the RM, in terms of virtual CPU cores. Requests higher than this won't take effect, and will get capped to this value.</description>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
        <value>4096</value>
        <description>Physical memory, in MB, to be made available to running containers</description>
  </property>
  <property>
    <name>yarn.nodemanager.resource.cpu-vcores</name>
        <value>4</value>
        <description>Number of CPU cores that can be allocated for containers.</description>
  </property>
</configuration>
```

mapred-site.xml

```
<configuration>
<property>
        <name>mapreduce.job.tracker</name>
        <value>hadoop-master:5431</value>
</property>

<property>
```

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

</configuration>
```

Crear el archivo “slaves” en /home/hadoop/hadoop-2.7.3/etc/hadoop/ donde se indican los nodos slaves

```
//En caso de tener problema con los hostnames, reemplazar con las ips, pero no mezclar
//slaves contiene:
hadoop-slave1
hadoop-slave2
```

Crear el directorio de datos:

```
hadoop@hadoop-master$ mkdir /home/hadoop/hadoopdata
hadoop@hadoop-master$ mkdir /home/hadoop/hadoopdata/hdfs
```

Copiar el directorio de hadoop ya configurado, a cada uno de los slaves:

```
hadoop@hadoop-master$ scp –rp /home/hadoop/hadoop-2.7.3/ hadoop@hadoop-slave1:/home/hadoop/
hadoop@hadoop-master$ scp –rp /home/hadoop/hadoop-2.7.3/ hadoop@hadoop-slave2:/home/hadoop/
hadoop@hadoop-master$ scp –rp /home/hadoop/hadoop-2.7.3/ hadoop@hadoop-slave3:/home/hadoop/
```

Asegurarse que los firewalls de las máquinas del cluster permitan (abrir) los puertos, son muchos los puertos: 9000, 50070, 50075, 5421, etc. la alternativa es bajar el firewall:

```
# systemctl stop firewalld
```

Formatear por primera vez el FS de hadoop:

```
hadoop@hadoop-master$ hdfs namenode -format
```

IMPORTANTE: En caso de necesitar formatear de nuevo debe asegurarse que todo lo que está adentro de hadoopdata/hdfs/ fue borrado en todos los nodos, de lo contrario podría ocurrir que al iniciar HDFS el nodo master no detecte los slaves.

Subir y bajar el cluster hadoop:

```
hadoop@hadoop-master$ cd hadoop-2.7.3/sbin

//Subir, por favor respetar el orden
hadoop@hadoop-master$ ./start-dfs.sh

hadoop@hadoop-master$ ./start-yarn.sh

// Bajar
hadoop@hadoop-master$ ./stop-dfs.sh

hadoop@hadoop-master$ ./stop-yarn.sh
```

un ejemplo:

el WordCount

se compila y queda en wc.jar

y se ejecuta:

```
$ hadoop dfs -mkdir /datos_in
$ hadoop dfs -copyFromLocal file1.txt /datos_in
$ hadoop dfs -copyFromLocal file2.txt /datos_in

$ hadoop jar wc.jar WordCount /datos_in /datos_out
```

Para correr programas python-mrjob en hadoop

```
$ pythonenv/bin/python2.6 wordcount.py hdfs:///data_in/file.txt -r hadoop > salida.txt
```

o

```
$ pythonenv/bin/python2.6 wordcount.py hdfs:///data_in/file.txt -r hadoop --output-dir hdfs:///data_out
```

(el directorio data_out, no debe existir antes de ejecutar:

```
$ hadoop dfs –rmdir /data_out
```

una vez lo ejecute:

```
$ hadoop dfs –copyToLocal /data_out/part-00000 part-0000
```

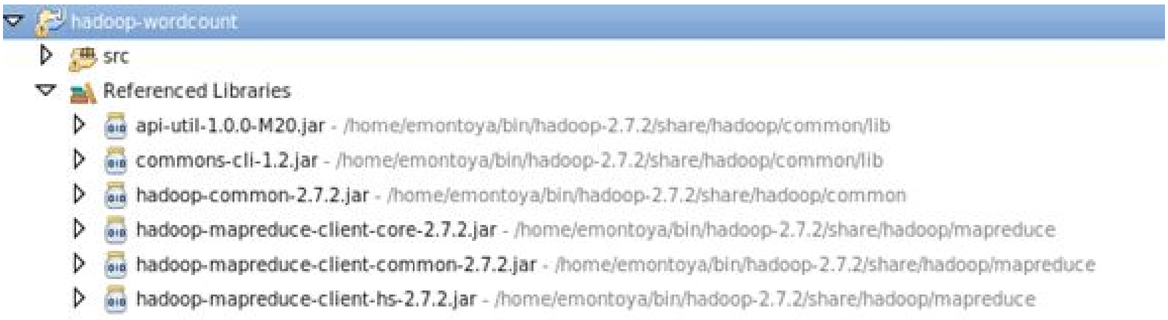
Para monitorear el cluster en un browser:

//Revisar HDFS

http://hadoop-master:50070 (info general)

//Revisar Yarn
http://hadoop-master:8088 (info de los jobs mapreduce del cluster)

Bibliotecas para compilar proyectos en Eclipse o Netbeans para Apache Hadoop



```
hadoop-wordcount
  src
  Referenced Libraries
    api-util-1.0.0-M20.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/common/lib
    commons-cli-1.2.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/common/lib
    hadoop-common-2.7.2.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/common
    hadoop-mapreduce-client-core-2.7.2.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/mapreduce
    hadoop-mapreduce-client-common-2.7.2.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/mapreduce
    hadoop-mapreduce-client-hs-2.7.2.jar - /home/emontoya/bin/hadoop-2.7.2/share/hadoop/mapreduce
```