

# Prácticas BigData

## 1. Instalación pseudo-distribuida

- Accedemos como usuario ROOT
- Creamos un directorio denominado /datos, donde crearemos nuestros directorios de datos

```
mkdir /datos
```

- Le damos permisos para que pueda ser usado por el usuario hadoop

```
chown hadoop:hadoop /datos
```

- Salimos de la sesión de ROOT y nos logamos como HADOOP
- Acceder al directorio “/opt/hadoop/etc/hadoop” de hadoop
- Modificamos el fichero “core-site.xml para que quede de la siguiente manera. Debemos asegurarnos de no poner localhost, o de lo contrario no funcionará cuando pongamos un cluster real

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://nodo1:9000</value>
  </property>
</configuration>
```

- Modificamos el hdfs-site.xml para que quede de la siguiente forma. Como solo tenemos un nodo, debemos dejarlo con un valor de replicación de 1.
- También le indicamos como se va a llamar el directorio para los metadatos y el directorio para los datos.

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/datos/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
```

```
<value>/datos/datanode</value>
</property>
</configuration>
```

- Vamos a crear los directorios para el sistema de ficheros HDFS, tanto para el namenode como para el datanode (no se tienen porque llamar así, pero para el curso queda más claro). Debemos crearlos con el mismo nombre que pusimos en el fichero de configuración.
- No es necesario crearlos porque los crea automáticamente, pero de ese modo nos aseguramos de que no tenemos problemas de permisos

```
mkdir /datos/namenode
mkdir /datos/datanode
```

- Formateamos el sistema de ficheros que acabamos de crear

```
dfs namenode -format
18/01/06 16:29:46 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = nodo1/192.168.56.101
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.9.0
STARTUP_MSG:                                classpath                                =
/opt/hadoop/etc/hadoop:/opt/hadoop/share/hadoop/common/lib/nimbus-jose-jwt-
3.9.jar:/opt/hadoop/share/hadoop/common/lib/java-xmlbuilder-
0.4.jar:/opt/hadoop/share/hadoop/common/lib/commons-configuration-
1.6.jar:/opt/hadoop/share/hadoop/common/lib/commons-cli-
1.2.jar:/opt/hadoop/share/hadoop/common/lib/commons-net-
3.1.jar:/opt/hadoop/share/hadoop/common/lib/jersey-core-
1.9.jar:/opt/hadoop/share/hadoop/common/lib/guava-
11.0.2.jar:/opt/hadoop/share/hadoop/common/lib/gson-
2.2.4.jar:/opt/hadoop/share/hadoop/common/lib/jackson-core-asl-
1.9.13.jar:/opt/hadoop/share/hadoop/common/lib/log4j-
1.2.17.jar:/opt/hadoop/share/hadoop/common/lib/woodstox-core-
5.0.3.jar:/opt/hadoop/share/hadoop/
....
...
..
```

- Si todos es correcto debemos tener el directorio /datos/namenode que debe tener otro subdirectorio denominado “current”

```
ls -l /datos/namenode/
total 0
drwxrwxr-x. 2 hadoop hadoop 112 ene 6 16:57 current
```

- Arrancamos los procesos de HDFS. Debe arrancar el NAMENODE, el SECONDARY NAMENODE y el DATANODE

#### start-dfs.sh

Starting namenodes on [localhost]

localhost: starting namenode, logging to /opt/hadoop/logs/hadoop-hadoop-namenode-nodo1.out

nodo1: starting datanode, logging to /opt/hadoop/logs/hadoop-hadoop-datanode-nodo1.out

Starting secondary namenodes [0.0.0.0]

0.0.0.0: starting secondarynamenode, logging to /opt/hadoop/logs/hadoop-hadoop-secondarynamenode-nodo1.out

- Comprobamos con el comando jps los procesos que se están ejecutando

#### jps

22962 DataNode

23187 SecondaryNameNode

23395 Jps

22837 NameNode

- También podemos comprobarlo con ps, para ver los procesos java

#### ps -ef | grep java

```
hadoop 20694 1 2 16:46 ? 00:00:06 /usr/java/jdk1.8.0_151/bin/java -
Dproc_datanode -Xmx1000m -Djava.net.preferIPv4Stack=true -
Dhadoop.log.dir=/opt/hadoop/logs -Dhadoop.log.file=hadoop.log -
Dhadoop.home.dir=/opt/hadoop -Dhadoop.id.str=hadoop -
Dhadoop.root.logger=INFO,console -Djava.library.path=/opt/hadoop/lib/native -
Dhadoop.policy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -
Djava.net.preferIPv4Stack=true -Djava.net.preferIPv4Stack=true -
Dhadoop.log.dir=/opt/hadoop/logs -Dhadoop.log.file=hadoop-hadoop-datanode-nodo1.log -
Dhadoop.home.dir=/opt/hadoop -Dhadoop.id.str=hadoop -Dhadoop.root.logger=INFO,RFA -
Djava.library.path=/opt/hadoop/lib/native -Dhadoop.policy.file=hadoop-policy.xml -
Djava.net.preferIPv4Stack=true -server -Dhadoop.security.logger=ERROR,RFAS -
Dhadoop.security.logger=ERROR,RFAS -Dhadoop.security.logger=ERROR,RFAS -
Dhadoop.security.logger=INFO,RFAS org.apache.hadoop.hdfs.server.datanode.DataNode

hadoop 20885 1 2 16:46 ? 00:00:05 /usr/java/jdk1.8.0_151/bin/java -
Dproc_secondarynamenode -Xmx1000m -Djava.net.preferIPv4Stack=true -
Dhadoop.log.dir=/opt/hadoop/logs -Dhadoop.log.file=hadoop.log -
Dhadoop.home.dir=/opt/hadoop -Dhadoop.id.str=hadoop -
Dhadoop.root.logger=INFO,console -Djava.library.path=/opt/hadoop/lib/native -
Dhadoop.policy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -
Djava.net.preferIPv4Stack=true -Djava.net.preferIPv4Stack=true -
Dhadoop.log.dir=/opt/hadoop/logs -Dhadoop.log.file=hadoop-hadoop-secondarynamenode-
nodo1.log -Dhadoop.home.dir=/opt/hadoop -Dhadoop.id.str=hadoop -
Dhadoop.root.logger=INFO,RFA -Djava.library.path=/opt/hadoop/lib/native -
Dhadoop.policy.file=hadoop-policy.xml -Djava.net.preferIPv4Stack=true -
Dhadoop.security.logger=INFO,RFAS -Dhdfs.audit.logger=INFO,NullAppender -
Dhadoop.security.logger=INFO,RFAS -Dhdfs.audit.logger=INFO,NullAppender -
Dhadoop.security.logger=INFO,RFAS -Dhdfs.audit.logger=INFO,NullAppender -
Dhadoop.security.logger=INFO,RFAS
org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode

hadoop 21140 26575 0 16:51 pts/0 00:00:00 grep --color=auto java
....
```

...

...

- Debe haber creado el directorio /datos/datanode que también tiene que tener un directorio llamado current.

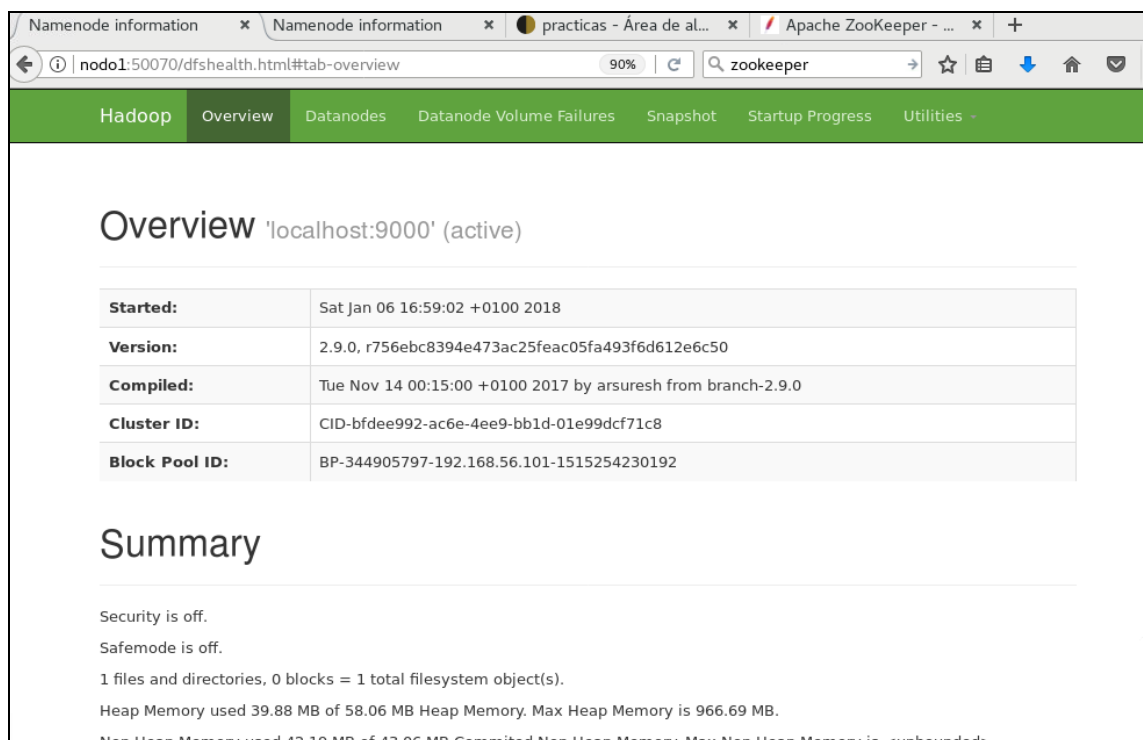
**ls -l /datos/datanode**

total 4

drwxrwxr-x. 3 hadoop hadoop 70 ene 6 16:59 current

-rw-rw-r--. 1 hadoop hadoop 11 ene 6 16:59 in\_use.lock

- Podemos acceder a la web de Administración para ver el resultado



Overview 'localhost:9000' (active)

<b>Started:</b>	Sat Jan 06 16:59:02 +0100 2018
<b>Version:</b>	2.9.0, r756ebc8394e473ac25feac05fa493f6d612e6c50
<b>Compiled:</b>	Tue Nov 14 00:15:00 +0100 2017 by arsureh from branch-2.9.0
<b>Cluster ID:</b>	CID-bfdee992-ac6e-4ee9-bb1d-01e99dcf71c8
<b>Block Pool ID:</b>	BP-344905797-192.168.56.101-1515254230192

**Summary**

Security is off.  
Safemode is off.  
1 files and directories, 0 blocks = 1 total filesystem object(s).  
Heap Memory used 39.88 MB of 58.06 MB Heap Memory. Max Heap Memory is 966.69 MB.  
Non Heap Memory used 42.19 MB of 43.06 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>

- Podemos comprobar el datanode

✓ In service

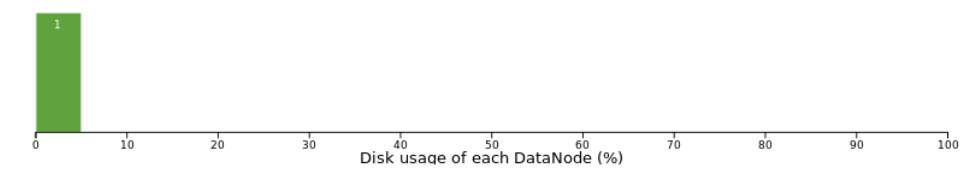
⚠ Down

⚡ Decommissioned

⚡ Decommissioned & dead

🔧 In Maintenance & dead

Datanode usage histogram



Disk usage of each DataNode (%)

In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ nodol:50010 (127.0.0.1:50010)	<a href="http://nodol:50075">http://nodol:50075</a>	0s	6m	49.98 GB	<div></div>	0	8 KB (0%) 2.9.0

Showing 1 to 1 of 1 entries

Previous

1

Next

●