PRÁCTICA Nº 2.1
Hadoop modo Pseudo-Cluster

- Nombre y apellidos: Alvaro Lucio-Villegas de Cea

Índice

Configuración de Hadoop en modo Standalone **3**

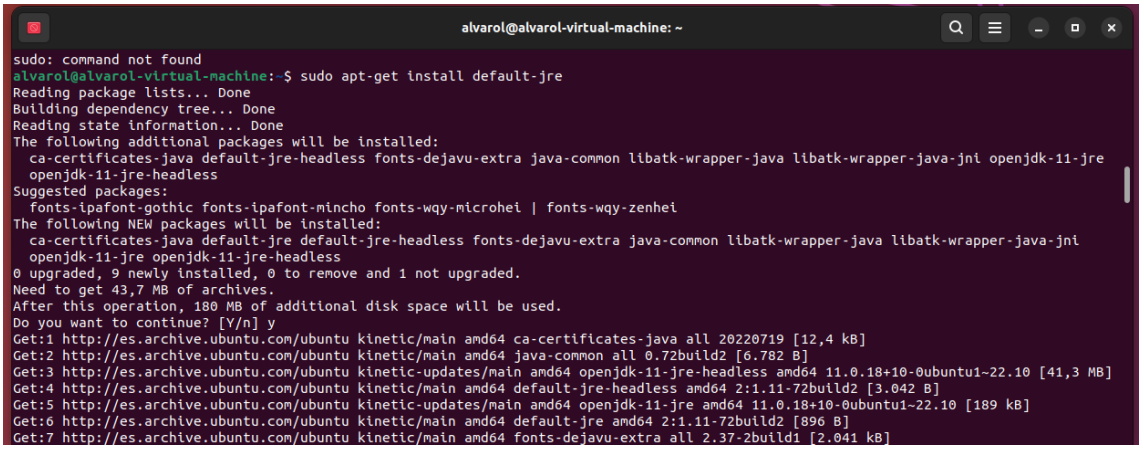
Configuración de Hadoop en modo Pseudo-Cluster **7**

BONUS: Configuración del Resource Manager **15**

Configuración de Hadoop en modo Standalone

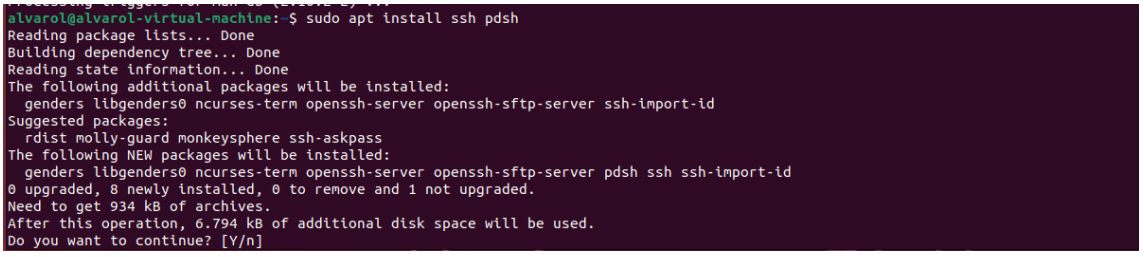
Instala el Java Runtime Environement OpenJDK

```
$ sudo apt-get install default-jre
```



- Instala ssh, pdsh

```
$ sudo apt install ssh pdsh
```



● Descarga y descomprime la última versión de Hadoop (https://downloads.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz):

```
$ tar -xvf hadoop-3.3.4.tar.gz
$ sudo mv hadoop-3.3.4 /usr/share/hadoop
```

```
alvarol@alvarol-virtual-machine:~/Downloads$ ls
hadoop-3.3.4.tar.gz
alvarol@alvarol-virtual-machine:~/Downloads$ tar -xvf hadoop-3.3.4.tar.gz
hadoop-3.3.4/
hadoop-3.3.4/licenses-binary/
hadoop-3.3.4/licenses-binary/LICENSE-dust.txt
hadoop-3.3.4/licenses-binary/LICENSE-re2j.txt
hadoop-3.3.4/licenses-binary/LICENSE-slf4j.txt
hadoop-3.3.4/licenses-binary/LICENSE-jquery.txt
hadoop-3.3.4/licenses-binary/LICENSE-zstd-jni.txt
hadoop-3.3.4/licenses-binary/LICENSE-hsql.txt
hadoop-3.3.4/licenses-binary/LICENSE-datatables.txt
hadoop-3.3.4/licenses-binary/LICENSE-jaf.txt
```

```
alvarol@alvarol-virtual-machine:~/Downloads$ sudo mv hadoop-3.3.4 /usr/share/hadoop
alvarol@alvarol-virtual-machine:~/Downloads$
```

● Añade la variable de entorno HADOOP_HOME al usuario bajo el que se va a ejecutar Hadoop. Para hacer el cambio permanente, añade el comando al fichero .profile del usuario (haz login the nuevo y verifica las variables de entorno ejecutando env)

```
$ export HADOOP_HOME=/usr/share/hadoop
```

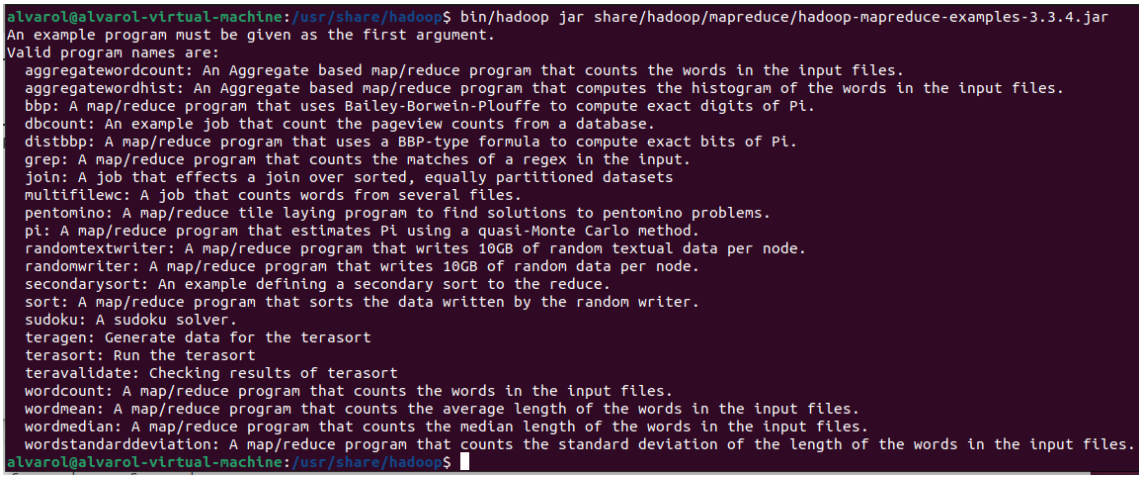
```
alvarol@alvarol-virtual-machine:~/Downloads$ export HADOOP_HOME=/usr/share/hadoop
alvarol@alvarol-virtual-machine:~/Downloads$
```

● Añade la siguiente línea en /usr/share/hadoop/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
```

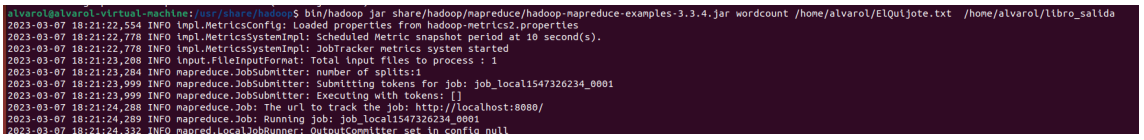
● Para verificar que la instalación funciona correctamente, ejecuta el siguiente comando desde /usr/share/hadoop, que debería mostrar una lista de aplicaciones MapReduce de ejemplo:

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar
```



● Ejecuta el programa de ejemplo WordCount con el fichero el_quijote.txt. Analiza los resultados.

```
$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar wordcount /home/alvarol/ElQuijote.txt  /home/alvarol/libro_salida
```



```
2023-03-07 18:21:27,754 INFO mapred.LocalJobRunner: reduce task executor complete.
2023-03-07 18:21:28,333 INFO mapreduce.Job:  map 100% reduce 100%
2023-03-07 18:21:28,336 INFO mapreduce.Job: Job job_local1547326234_0001 completed successfully
2023-03-07 18:21:28,383 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=6171224
                FILE: Number of bytes written=4115475
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=37861
                Map output records=384260
                Map output bytes=3688608
                Map output materialized bytes=605516
                Input split bytes=97
                Combine input records=384260
                Combine output records=40059
                Reduce input groups=40059
                Reduce shuffle bytes=605516
                Reduce input records=40059
                Reduce output records=40059
                Spilled Records=80118
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=39
                Total committed heap usage (bytes)=480247808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2198936
        File Output Format Counters
                Bytes Written=452417
alvarol@alvarol-virtual-machine:/usr/share/hadoop$
```

Configuración de Hadoop en modo Pseudo-Cluster

- NOTA: Estos pasos deben realizarse a continuación de la configuración inicial realizada para el modo Standalone!
- Añade la siguiente propiedad al fichero de configuración

/usr/share/hadoop/etc/hadoop/core-site.xml, en la sección configuration

Open | *core-site.xml /usr/share/hadoop/etc/hadoop | Save

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4   Licensed under the Apache License, Version 2.0 (the "License");
5   you may not use this file except in compliance with the License.
6   You may obtain a copy of the License at
7
8     http://www.apache.org/licenses/LICENSE-2.0
9
10  Unless required by applicable law or agreed to in writing, software
11  distributed under the License is distributed on an "AS IS" BASIS,
12  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13  See the License for the specific language governing permissions and
14  limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20         <property>
21                 <name>fs.defaultFS</name>
22                 <value>hdfs://localhost:9000</value>
23         </property>
24 </configuration>
25
```

- Añade las siguientes propiedades al fichero de configuración

/usr/share/hadoop/etc/hadoop/hdfs-site.xml, en la sección configuration

Open | hdfs-site.xml /usr/share/hadoop/etc/hadoop | Save

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3 <!--
4   Licensed under the Apache License, Version 2.0 (the "License");
5   you may not use this file except in compliance with the License.
6   You may obtain a copy of the License at
7
8     http://www.apache.org/licenses/LICENSE-2.0
9
10  Unless required by applicable law or agreed to in writing, software
11  distributed under the License is distributed on an "AS IS" BASIS,
12  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13  See the License for the specific language governing permissions and
14  limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18
19 <configuration>
20         <property>
21                 <name>dfs.replication</name>
22                 <value>1</value>
23         </property>
24         <property>
25                 <name>dfs.datanode.data.dir</name>
26                 <value>/mnt/hadoop/data</value>
27         </property>
28         <property>
29                 <name>dfs.datanode.name.dir</name>
30                 <value>/mnt/hadoop/name</value>
31         </property>
32         <property>
33                 <name>dfs.datanode.checkpoint.dir</name>
34                 <value>/mnt/hadoop/namesecondary</value>
35         </property>
36
37 </configuration>
```

XML | Tab Width: 8 | Ln 16, Col 1 | INS

● Ahora crearemos los ficheros donde se almacenarán los datos y metadatos de HDFS:

```
$ sudo mkdir /mnt/hadoop
$ sudo mkdir /mnt/hadoop/data
$ sudo mkdir /mnt/hadoop/name
$ sudo mkdir /mnt/hadoop/namesecondary
$ sudo chown administrador:administrador /mnt/hadoop/data /mnt/hadoop/name / /mnt/hadoop/namesecondary
```

```
alvarol@alvarol-virtual-machine:~$ sudo mkdir /mnt/hadoop
alvarol@alvarol-virtual-machine:~$ sudo mkdir /mnt/hadoop/data
alvarol@alvarol-virtual-machine:~$ sudo mkdir /mnt/hadoop/name
alvarol@alvarol-virtual-machine:~$ sudo mkdir /mnt/hadoop/namesecondary
alvarol@alvarol-virtual-machine:~$ sudo chown administrator:administrator /mnt/hadoop/data /mnt/hadoop/name /mnt/hadoop/namesecondary
chown: invalid user: ‘administrator:administrator’
alvarol@alvarol-virtual-machine:~$ sudo chown alvarol:alvarol /mnt/hadoop/data /mnt/hadoop/name /mnt/hadoop/namesecondary
alvarol@alvarol-virtual-machine:~$
```

● Crea un directorio para el almacenamiento de ficheros temporales:

```
alvarol@alvarol-virtual-machine:~$ sudo mkdir /mnt/hadoop/tmp
alvarol@alvarol-virtual-machine:~$ sudo chown alvarol:alvarol /mnt/hadoop/tmp
alvarol@alvarol-virtual-machine:~$
```

● Añade la siguiente propiedad a core-site.xml para que Hadoop almacene los archivos temporales en el nuevo directorio.

● El siguiente paso es configurar la autenticación SSH sin contraseña. Esto permite a hadoop conectarse a los distintos nodos (en este caso únicamente la maquina local):

```
alvarol@alvarol-virtual-machine:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/alvarol/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/alvarol/.ssh/id_rsa
Your public key has been saved in /home/alvarol/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:1t5CsbGe7G3IMiHofufjNmpK6cJWybbNS2gK1FgCb9Y alvarol@alvarol-virtual-machine
The key's randomart image is:
+---[RSA 3072]----+
|.                |
|.. .             |
| .+.E     o      |
| o=      . =     |
| o o o  S =      |
|.   *o...= o     |
|.. +=+.. o*..    |
| .+=o.= B.oo.    |
| ..o++oBo=...    |
+----[SHA256]-----+
alvarol@alvarol-virtual-machine:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
alvarol@alvarol-virtual-machine:~$ sudo chmod 0600 ~/.ssh/authorized_keys
[sudo] password for alvarol:
alvarol@alvarol-virtual-machine:~$
```

● En Ubuntu es necesario además crear el fichero /etc/pdsh/rcmd_default con el contenido ssh

```
alvarol@alvarol-virtual-machine: ~
  GNU nano 6.4                                  /etc/pdsh/rcmd_default *
ssh
```

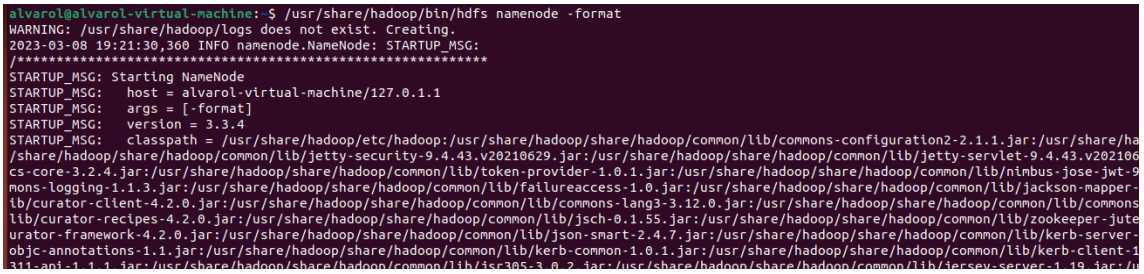
● Verifica que el cambio realizado en el paso anterior permite conexiones SSH sin contraseña:

ssh localhost

```
alvarol@alvarol-virtual-machine:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:bjupt2pSblzyNu5xT0BJrcZKrHHJeXRqVcO+oZHNtXk.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.10 (GNU/Linux 5.19.0-35-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

0 updates can be applied immediately.

*** System restart required ***

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

alvarol@alvarol-virtual-machine:~$ ls -la /mnt/hadoop
```

● El siguiente paso es dar formato al volumen HDFS. Esto se consigue con el siguiente comando:

```
/usr/share/hadoop/bin/hdfs namenode -format
```

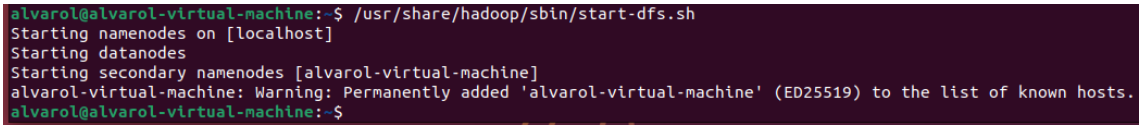


● Por último, vamos a desactivar el sistema de permisos de HDFS para facilitar la configuración. Esto se consigue añadiendo la propiedad siguiente al fichero **/usr/share/hadoop/etc/hadoop/hdfs-site.xml**:

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

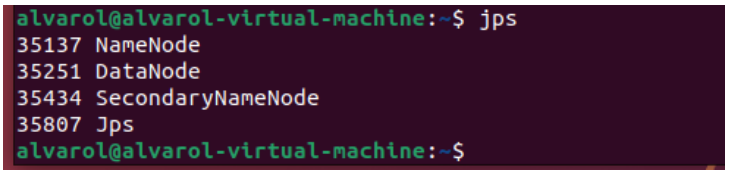
●Una vez finalizada la configuración podemos lanzar los servicios Hadoop relacionados con HDFS ejecutando:

```
$ /usr/share/hadoop/sbin/start-dfs.sh
```



●El comando anterior deberia lanzar los *daemons* **Datanode**, **Namenode** y **SecondaryNamenode**. Podemos verificar que se están ejecutando lanzando el comando **jps**

```
sudo apt install openjdk-11-jdk-headless
jps
```



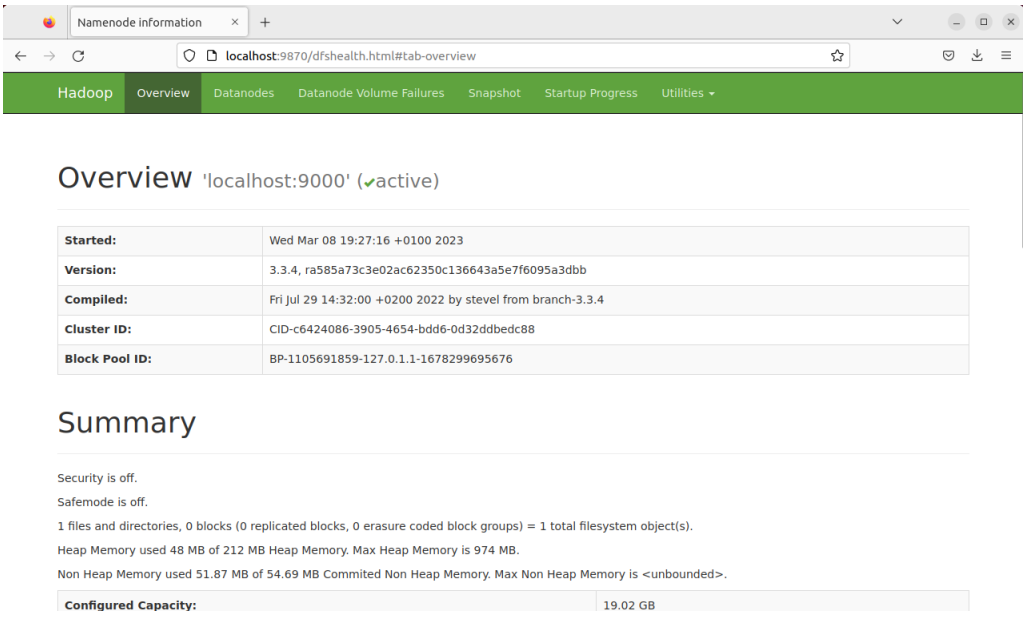
●Una vez que los servicios se están ejecutando es posible manipular el sistema de archivos utilizando el comando

```
/usr/share/hadoop/bin/hadoop fs <subcomandos>
```

```
alvarol@alvarol-virtual-machine:~$ /usr/share/hadoop/bin/hadoop fs -help
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum [-v] <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-concat <target path> <src path> <src path> ...]
        [-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] [-e] [-s] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] [-t <thread count>] [-q <thread pool queue size>] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-v] [-x] <path> ...]
        [-expunge [-immediate] [-fs <path>]]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-head <file>]
        [-help [cmd ...]]
```

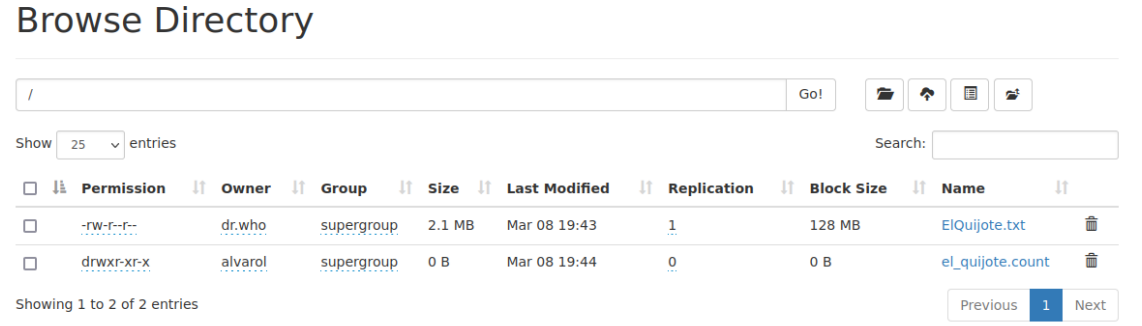
●Hadoop dispone de una interfaz Web para monitorizar los sistemas de archivos HDFS a la que se accede a través del navegador. Permite consultar el estado actual de los nodos asi como consultar los logs y navegar el contenido del sistema de archivos:

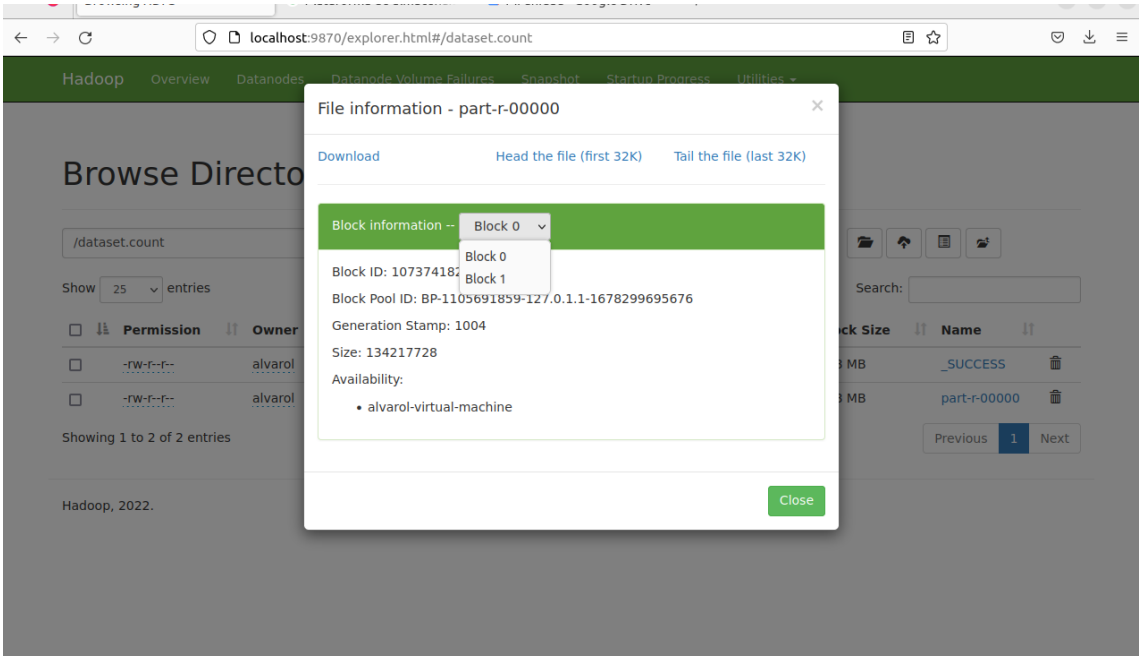
```
http://localhost:9870
```



●Para comprobar que todo funciona correctamente, sube el fichero el_quijote.txt y ejecuta el programa de ejemplo WordCount. Comprueba los resultados:

```
alvarol@alvarol-virtual-machine:/usr/share/hadoop$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar wordcount /ElQuijote.txt /el_quijote.count
2023-03-08 19:44:02,633 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-03-08 19:44:03,100 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-03-08 19:44:03,100 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-03-08 19:44:04,998 INFO input.FileInputFormat: Total input files to process : 1
2023-03-08 19:44:05,753 INFO mapreduce.JobSubmitter: number of splits:1
2023-03-08 19:44:06,899 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local245925236_0001
2023-03-08 19:44:06,900 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-03-08 19:44:07,282 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-03-08 19:44:07,283 INFO mapreduce.Job: Running job: job_local245925236_0001
2023-03-08 19:44:07,316 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2023-03-08 19:44:07,399 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-03-08 19:44:07,399 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2023-03-08 19:44:07,406 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-03-08 19:44:07,649 INFO mapred.LocalJobRunner: Waiting for map tasks
2023-03-08 19:44:07,657 INFO mapred.LocalJobRunner: Starting task: attempt_local245925236_0001_m_000000_0
2023-03-08 19:44:07,839 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-03-08 19:44:07,842 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
```





●Apagar Servicio

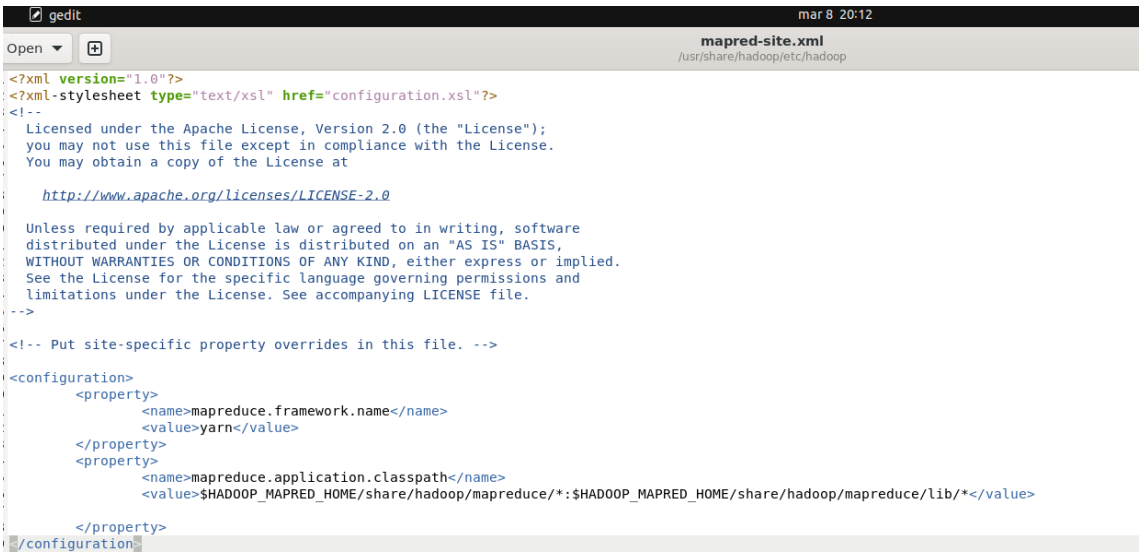
```
$ /usr/share/hadoop/sbin/stop-dfs.sh
```

BONUS: Configuración del Resource Manager

●Un componente clave para gestionar un cluster Hadoop es el Resource Manager, encargado de distribuir la carga de trabajo entre los diferentes nodos. El Resource Manager por defecto de Hadoop desde la versión 2.0 se llama YARN.

●Añade los siguientes parámetros a **/usr/share/hadoop/etc/hadoop/mapred-site.xml**

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

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>

    </property>
```



●Añade los siguientes parámetros a **/usr/share/hadoop/etc/hadoop/yarn-site.xml**

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

<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</value>
    </property>
```

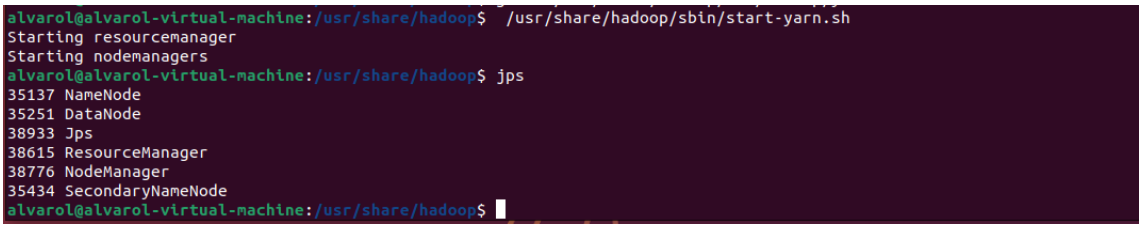
●Por defecto YARN no funcionará correctamente si el volumen HDFS tiene menos del 10% de espacio libre. Los nodos pasarán a modo "Unhealthy" y no aceptarán nuevas

aplicaciones. Este umbral puede modificarse añadiendo el parámetro **yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage**:

```
<property>

<name>yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage</name>
        <value>98.5</value>
</property>
```

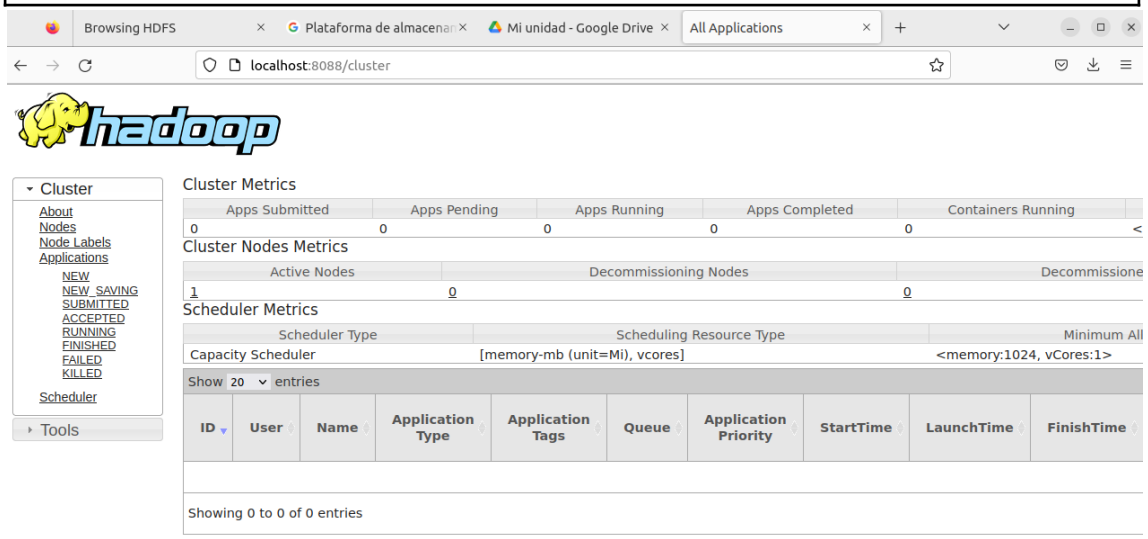
●Lanza YARN ejecutando el comando

```
/usr/share/hadoop/sbin/start-yarn.sh
```

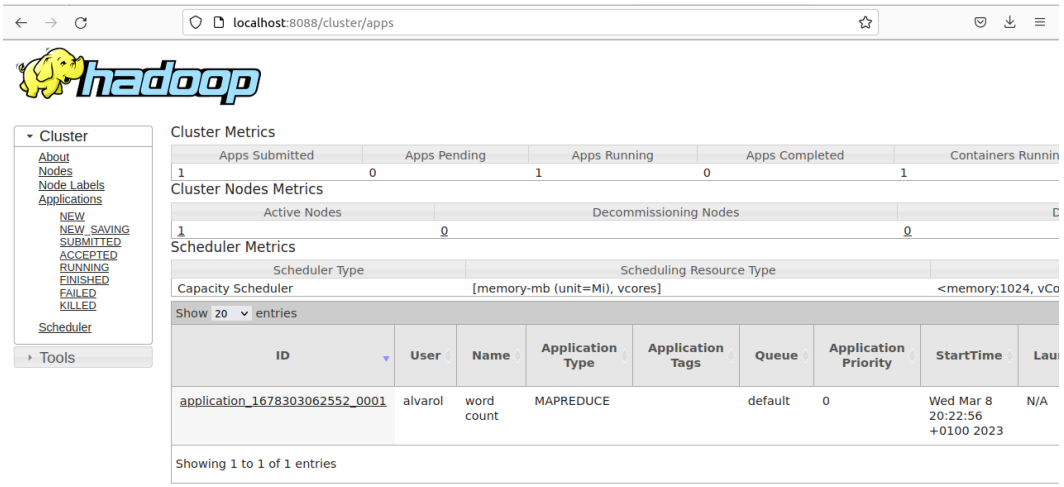


●El ResourceManager ofrece una interfaz Web con información detallada sobre el estado del cluster

http://localhost:8088



●Lanza de nuevo la aplicación Wordcount y monitoriza su estado a través de la interfaz del Resource Manager.



●Podemos cerrar YARN ejecutando el comando

```
/usr/share/hadoop/sbin/stop-yarn.sh
```