

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 Environment OpenJDK

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

```
alvarol@alvarol-virtual-machine: ~
sudo: command not found
alvarol@alvarol-virtual-machine:~$ sudo apt-get install default-jre
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java default-jre-headless fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni openjdk-11-jre
  openjdk-11-jre-headless
Suggested packages:
  fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei | fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java default-jre default-jre-headless fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni
  openjdk-11-jre openjdk-11-jre-headless
0 upgraded, 9 newly installed, 0 to remove and 1 not upgraded.
Need to get 43,7 MB of archives.
After this operation, 180 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://es.archive.ubuntu.com/ubuntu kinetic/main amd64 ca-certificates-java all 20220719 [12,4 kB]
Get:2 http://es.archive.ubuntu.com/ubuntu kinetic/main amd64 java-common all 0.72build2 [6.782 B]
Get:3 http://es.archive.ubuntu.com/ubuntu kinetic-updates/main amd64 openjdk-11-jre-headless amd64 11.0.18+10-0ubuntu1-22.10 [41,3 MB]
Get:4 http://es.archive.ubuntu.com/ubuntu kinetic/main amd64 default-jre-headless amd64 2:1.11-72build2 [3.042 B]
Get:5 http://es.archive.ubuntu.com/ubuntu kinetic-updates/main amd64 openjdk-11-jre amd64 11.0.18+10-0ubuntu1-22.10 [189 kB]
Get:6 http://es.archive.ubuntu.com/ubuntu kinetic/main amd64 default-jre amd64 2:1.11-72build2 [896 B]
Get:7 http://es.archive.ubuntu.com/ubuntu kinetic/main amd64 fonts-dejavu-extra all 2.37-2build1 [2.041 kB]
```

- Instala ssh, pdsh

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

```
alvarol@alvarol-virtual-machine:~$ sudo apt install ssh pdsh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  genders libgenders0 ncurses-term openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  rdist molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  genders libgenders0 ncurses-term openssh-server openssh-sftp-server pdsh ssh ssh-import-id
0 upgraded, 8 newly installed, 0 to remove and 1 not upgraded.
Need to get 934 kB of archives.
After this operation, 6.794 kB of additional disk space will be used.
Do you want to continue? [Y/n]
```

- 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/
```



```
hadoop-env.sh
/usr/share/hadoop/etc/hadoop
430 # export HADOOP_REGISTRYDNS_SECURE_EXTRA_OPTS="-jvm server"
431
432
433
434 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
```

```
alvarol@alvarol-virtual-machine: /usr/share/hadoop$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar
An example program must be given as the first argument.
Valid program names are:
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in the input files.
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.
  dbcount: An example job that count the pageview counts from a database.
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.
  grep: A map/reduce program that counts the matches of a regex in the input.
  join: A job that effects a join over sorted, equally partitioned datasets
  multifielwc: A job that counts words from several files.
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.
  randomwriter: A map/reduce program that writes 10GB of random data per node.
  secondarysort: An example defining a secondary sort to the reduce.
  sort: A map/reduce program that sorts the data written by the random writer.
  sudoku: A sudoku solver.
  teragen: Generate data for the terasort
  terasort: Run the terasort
  teravalidate: Checking results of terasort
  wordcount: A map/reduce program that counts the words in the input files.
  wordmean: A map/reduce program that counts the average length of the words in the input files.
  wordmedian: A map/reduce program that counts the median length of the words in the input files.
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the words in the input files.
alvarol@alvarol-virtual-machine: /usr/share/hadoop$
```

- 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
```

```
alvarol@alvarol-virtual-machine: /usr/share/hadoop$ 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 18:21:22,554 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-03-07 18:21:22,778 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2023-03-07 18:21:22,778 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2023-03-07 18:21:23,280 INFO InputFileInputFormat: Total input files to process : 1
2023-03-07 18:21:23,284 INFO mapreduce.JobSubmitter: number of splits:1
2023-03-07 18:21:23,999 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1547326234_0001
2023-03-07 18:21:23,999 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-03-07 18:21:24,288 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2023-03-07 18:21:24,289 INFO mapreduce.Job: Running job: job_local1547326234_0001
2023-03-07 18:21:24,332 INFO mapred.local.JobRunner: OutputCommitter set in config null
```

```
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  [icon] *core-site.xml /usr/share/hadoop/etc/hadoop Save [menu] [minus] [square] [x]
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

```

hdfs-site.xml
/usr/share/hadoop/etc/hadoop

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>
  
```


- 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 administrador:administrador /mnt/hadoop/data /mnt/hadoop/name /mnt/hadoop/namesecondary
chown: invalid user: 'administrador:administrador'
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.



```
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   <property>
25     <name>hadoop.tmp.dir</name>
26     <value>/mnt/hadoop/tmp</value>
27   </property>
28 </configuration>
29
```

- 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: ~
/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:bjupt2psblzyNu5xT0BJrcZKRHHJexRqVc0+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
```

[illegible]

- 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>
```

```
hdfs-site.xml
/usr/share/hadoop/etc/hadoop

33         <name>dfs.datanode.checkpoint.dir</name>
34         <value>/mnt/hadoop/namesecondary</value>
35     </property>
36 </property>
37     <name>dfs.permissions</name>
38     <value>false</value>
39 </property>
40
41
42
```

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

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

```
alvarol@alvarol-virtual-machine:~$ /usr/share/hadoop/sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [alvarol-virtual-machine]
alvarol-virtual-machine: Warning: Permanently added 'alvarol-virtual-machine' (ED25519) to the list of known hosts.
alvarol@alvarol-virtual-machine:~$
```

- El comando anterior debería 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
```

```
alvarol@alvarol-virtual-machine:~$ jps
35137 NameNode
35251 DataNode
35434 SecondaryNameNode
35807 Jps
alvarol@alvarol-virtual-machine:~$
```

- 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 así como consultar los logs y navegar el contenido del sistema de archivos:

`http://localhost:9870`

The screenshot shows the 'Overview' tab of the Hadoop NameNode Information page. The page title is 'Overview 'localhost:9000' (active)'. Below the title is a table with the following information:

Started:	Wed Mar 08 19:27:16 +0100 2023
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 14:32:00 +0200 2022 by stebel from branch-3.3.4
Cluster ID:	CID-c6424086-3905-4654-bdd6-0d32ddbedc88
Block Pool ID:	BP-1105691859-127.0.1.1-1678299695676

Below the table is a 'Summary' section with the following text:

Security is off.
 Safemode is off.
 1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
 Heap Memory used 48 MB of 212 MB Heap Memory. Max Heap Memory is 974 MB.
 Non Heap Memory used 51.87 MB of 54.69 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

At the bottom, there is a table with the following information:

Configured Capacity:	19.02 GB
-----------------------------	----------

- 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 19:44:07,406 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
```

Browse Directory

Show entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rwxr-xr-x	dr.who	supergroup	2.1 MB	Mar 08 19:43	1	128 MB	ElQuijote.txt	<input type="checkbox"/>
<input type="checkbox"/>	-rwxr-xr-x	alvarol	supergroup	0 B	Mar 08 19:44	0	0 B	el_quijote.count	<input type="checkbox"/>

Showing 1 to 2 of 2 entries

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rwxr-xr-x	alvarol	supergroup	0 B	Mar 08 19:44	0	0 B	part-r-00000	<input type="checkbox"/>

Showing 1 to 2 of 2 entries

File information - part-r-00000

Download
Head the file (first 32K)
Tail the file (last 32K)

Block information -- Block 0

Block ID: 10737418

Block Pool ID: BP-1105691859-127.0.1.1-1678299695676

Generation Stamp: 1004

Size: 134217728

Availability:

- alvarol-virtual-machine

Close

- 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_MAP
    RED_HOME/share/hadoop/mapreduce/lib/*</value>

</property>
```



```
gedit mar 8 20:12
mapred-site.xml
/usr/share/hadoop/etc/hadoop

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <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>
</configuration>
```

● 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,HADO
  OP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,H
  ADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</value>
</property>
```

```
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>

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

- 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>
```

```
Open  yarn-site.xml  Save
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<!-- Site specific YARN configuration properties -->
  <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>
  <property>
    <name>yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage</name>
    <value>98.5</value>
  </property>
</configuration>
```

- Lanza YARN ejecutando el comando

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

```
alvarol@alvarol-virtual-machine:/usr/share/hadoop$ /usr/share/hadoop/sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
alvarol@alvarol-virtual-machine:/usr/share/hadoop$ jps
35137 NameNode
35251 DataNode
38933 Jps
38615 ResourceManager
38776 NodeManager
35434 SecondaryNameNode
alvarol@alvarol-virtual-machine:/usr/share/hadoop$
```

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

http://localhost:8088

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum All
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime
Showing 0 to 0 of 0 entries									

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

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
1	0	1	0	1

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum All
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime
application_1678303062552_0001	alvarol	word count	MAPREDUCE		default	0	Wed Mar 8 20:22:56 +0100 2023		N/A

Showing 1 to 1 of 1 entries

- Podemos cerrar YARN ejecutando el comando

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