Tecnologías de Computación de Datos Masivos

Realización de las prácticas

Máster en Big Data UMU
Alejandro Pérez Belando

Práctica 1

1. Creación de imágenes Docker para los diferentes servicios

1.1.1. / 1.1.2

```
PS C:\Users\Alejandro>
docker container run -ti --name namenode --network=hadoop-cluster --hostname namenode --net-ali
as resourcemanager --expose=8000-10000 -p 9870:9870 -p 8088:8088 dsevilla/hadoop-base /bin/bash

root@namenode:/# mkdir -p /var/data/hdfs/namenode
root@namenode:/# chown hdadmin:hadoop /var/data/hdfs/namenode


root@namenode:/# su - hdadmin
```

```
<configuration>
  <property>
    <!-- Nombre del filesystem por defecto -->
    <!-- Como queremos usar HDFS tenemos que indicarlo con hdfs:// y el servidor y puerto en el que corre el NameNod
    <name>fs.defaultFS</name>
    <value>hdfs://namenode:9000/</value>
    <final>true</final>
  </property>

  <property>
    <!-- Directorio para almacenamiento temporal (debe tener suficiente espacio) -->
    <name>hadoop.tmp.dir</name>
    <value>/var/tmp/hadoop-${user.name}</value>
    <final>true</final>
  </property>
</configuration>
```

Salir: ctrl + X → Y → enter

```
^G Help        ^O Write Out   ^W Where Is    [ Read 47 lines ]   ^K Cut         ^T Execute     ^C Location    M-U Undo
^X Exit        ^R Read File   ^\ Replace                          ^U Paste       ^J Justify     ^/ Go To Line  M-E Redo
```

Pegarlo

→ Abre configuración

Los dobles son para comprobar

```
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop/
hdadmin@namenode:~/hadoop/etc/hadoop$ nano core-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano core-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano hdfs-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano hdfs-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano mapred-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ nano mapred-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$ exit
logout
```

1.1.3.

```
hdadmin@namenode:~$ hdfs namenode -format
2024-11-08 10:31:39,249 INFO snapshot.SnapshotManager: SkipList is disabled
2024-11-08 10:31:39,258 INFO util.GSet: Computing capacity for map cachedBlocks
2024-11-08 10:31:39,258 INFO util.GSet: VM type       = 64-bit
2024-11-08 10:31:39,258 INFO util.GSet: 0.25% max memory 1.9 GB = 4.9 MB
2024-11-08 10:31:39,258 INFO util.GSet: capacity      = 2^19 = 524288 entries
2024-11-08 10:31:39,277 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.bucket
s = 10
2024-11-08 10:31:39,277 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-11-08 10:31:39,277 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes =
 1,5,25
2024-11-08 10:31:39,289 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-11-08 10:31:39,289 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and
 retry cache entry expiry time is 600000 millis
2024-11-08 10:31:39,294 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-11-08 10:31:39,294 INFO util.GSet: VM type       = 64-bit
2024-11-08 10:31:39,298 INFO util.GSet: 0.029999999329447746% max memory 1.9 GB = 607.6 KB
2024-11-08 10:31:39,298 INFO util.GSet: capacity      = 2^16 = 65536 entries
2024-11-08 10:31:39,340 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1656694194-172.18.
0.2-1731058299332
2024-11-08 10:31:39,372 INFO common.Storage: Storage directory /var/data/hdfs/namenode has been
 successfully formatted.
2024-11-08 10:31:39,425 INFO namenode.FSImageFormatProtobuf: Saving image file /var/data/hdfs/n
amenode/current/fsimage.ckpt_0000000000000000000 using no compression
2024-11-08 10:31:39,597 INFO namenode.FSImageFormatProtobuf: Image file /var/data/hdfs/namenode
/current/fsimage.ckpt_0000000000000000000 of size 402 bytes saved in 0 seconds .
2024-11-08 10:31:39,645 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with
txid >= 0
2024-11-08 10:31:39,681 INFO namenode.FSNamesystem: Stopping services started for active state
2024-11-08 10:31:39,681 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-11-08 10:31:39,690 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet
shutdown.
2024-11-08 10:31:39,695 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at namenode/172.18.0.2
************************************************************/
```

Si todo ha ido bien →

Revisar que el directorio se ha creado

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ cd $HADOOP_HOME/logs
hdadmin@namenode:~/hadoop/logs$ ls -l
total 0
-rw-r--r-- 1 hdadmin hadoop 0 nov  8 10:31 SecurityAuth-hdadmin.audit
```

1.1.4.

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ hdfs --daemon start namenode
hdadmin@namenode:~$ cd $HADOOP_HOME/logs
hdadmin@namenode:~/hadoop/logs$ ls
hadoop-hdadmin-namenode-namenode.log  SecurityAuth-hdadmin.audit
hadoop-hdadmin-namenode-namenode.out
hdadmin@namenode:~/hadoop/logs$ jps
323 NameNode
403 Jps
```

→ Todo bien

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ yarn --daemon start resourcemanager
hdadmin@namenode:~$ cd $HADOOP_HOME/logs
hdadmin@namenode:~/hadoop/logs$ jps
323 NameNode
695 Jps
458 ResourceManager
```

→ Todo bien

1.1.5.

No comprendo que debería ocurrir al visitar los dos enlaces

1.1.6.

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ yarn --daemon stop resourcemanager
hdadmin@namenode:~$ hdfs --daemon stop namenode
hdadmin@namenode:~$ exit
logout
```

1.1.7.

```
root@namenode:/# nano /inicio.sh
root@namenode:/# chmod +x /inicio.sh
```

```
#! /bin/sh
export JAVA_HOME=/usr/lib/jvm/default-java
export HADOOP_HOME=/opt/bd/hadoop

# Inicio el NameNode y el ResourceManager
su - hdadmin -c "$HADOOP_HOME/bin/hdfs --daemon start namenode"
su - hdadmin -c "$HADOOP_HOME/bin/yarn --daemon start resourcemanager"

# Lazo para mantener activo el contenedor
while true; do sleep 10000; done
```

Es otra config. como la de antes

1.1.8.

```
root@namenode:/# exit
exit
PS C:\Users\Alejandro> docker container commit namenode namenode-image
sha256:84cb471f2da8dc72da423400db95dd5404cb6d54939d2b86987ad03d5af82376
PS C:\Users\Alejandro> docker images
REPOSITORY                                TAG       IMAGE ID       CREATED          SIZE
namenode-image                            latest    84cb471f2da8   11 seconds ago   3.29GB
datanode-image                            latest    9228bf275c3b   13 days ago      3.29GB
<none>                                    <none>    699e4306340e   13 days ago      3.29GB
dsevilla/hadoop-base                      latest    89f6552f3e1a   7 weeks ago      3.29GB
docker.elastic.co/kibana/kibana           8.12.1    6c8b4a0b5197   9 months ago     1.72GB
docker.elastic.co/elasticsearch/elasticsearch   8.12.1    a476af93763c   9 months ago     2.07GB
PS C:\Users\Alejandro> docker container rm namenode
namenode
```

→ Creo imagen

→ Todo bien

→ Borro contenedor

1.2.1. / 1.2.2 - se repite el proceso pero ahora con datanode

```
PS C:\Users\Alejandro> docker container run -ti --name datanode --network=hadoop-cluster --hostname datanode --exp
ose=8000-10000 --expose=50000-50200 dsevilla/hadoop-base /bin/bash
root@datanode:/# mkdir -p /var/data/hdfs/datanode
root@datanode:/# chown hdadmin:hadoop /var/data/hdfs/datanode
```

```
root@datanode:/# su - hdadmin

hdadmin@datanode:~$ cd $HADOOP_HOME/etc/hadoop/
hdadmin@datanode:~/hadoop/etc/hadoop$ nano core-site.xml
hdadmin@datanode:~/hadoop/etc/hadoop$ nano hdfs-site.xml
hdadmin@datanode:~/hadoop/etc/hadoop$ nano yarn-site.xml
hdadmin@datanode:~/hadoop/etc/hadoop$ nano mapred-site.xml
```

Core y mapred, misma configuración que para el namenode (apartado 1.1.2)

1.2.3.

```
root@datanode:/# su - hdadmin
hdadmin@datanode:~$ hdfs --daemon start datanode
WARNING: /opt/bd/hadoop/logs does not exist. Creating.
hdadmin@datanode:~$ cd $HADOOP_HOME/logs
hdadmin@datanode:~/hadoop/logs$ jps
64 DataNode
126 Jps
```

→ Todo bien

```
root@datanode:/# su - hdadmin
hdadmin@datanode:~$ yarn --daemon start nodemanager
hdadmin@datanode:~$ cd $HADOOP_HOME/logs
hdadmin@datanode:~/hadoop/logs$ jps
64 DataNode
180 NodeManager
286 Jps
```

→ Todo bien

1.2.4.

```
root@datanode:/# su - hdadmin
hdadmin@datanode:~$ yarn --daemon stop nodemanager
WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
hdadmin@datanode:~$ hdfs --daemon stop datanode
hdadmin@datanode:~$ exit
logout
```

¿Es un problema?

1.2.5.

Se hace lo mismo que en 1.1.7, se copia lo que está en el guión

```
root@datanode:/# nano /inicio.sh
root@datanode:/# chmod +x /inicio.sh
```

1.2.6.

```
root@datanode:/# exit
exit
PS C:\Users\Alejandro> docker container commit datanode datanode-image
sha256:82f05154f1b1b7e90fe1e3c4bef1d835e76836969942f987d1b5f4678fa45877
PS C:\Users\Alejandro> docker images
REPOSITORY                                TAG       IMAGE ID       CREATED          SIZE
datanode-image                            latest    82f05154f1b1   9 seconds ago    3.29GB
namenode-image                            latest    84cb471f2da8   2 hours ago      3.29GB
<none>                                    <none>    9228bf275c3b   13 days ago      3.29GB
<none>                                    <none>    699e4306340e   13 days ago      3.29GB
dsevilla/hadoop-base                      latest    89f6552f3e1a   7 weeks ago      3.29GB
docker.elastic.co/kibana/kibana           8.12.1    6c8b4a0b5197   9 months ago     1.72GB
docker.elastic.co/elasticsearch/elasticsearch   8.12.1    a476af93763c   9 months ago     2.07GB
PS C:\Users\Alejandro> docker container rm namenode
Error response from daemon: No such container: namenode
PS C:\Users\Alejandro> docker container rm datanode
datanode
```

→ Todo bien

2. Iniciado de cluster Hadoop con contenedores Docker

2.1. / 2.2.

1 namenode

```
PS C:\Users\Alejandro> docker container run -d --name namenode --network=hadoop-cluster --hostname namenode --net-
alias resourcemanager --cpus=1 --memory=3072m --expose=8000-10000 -p 9870:9870 -p 8088:8088 -p 8888:8888 -p 4040:4
040 namenode-image /inicio.sh
```

4 datanodes

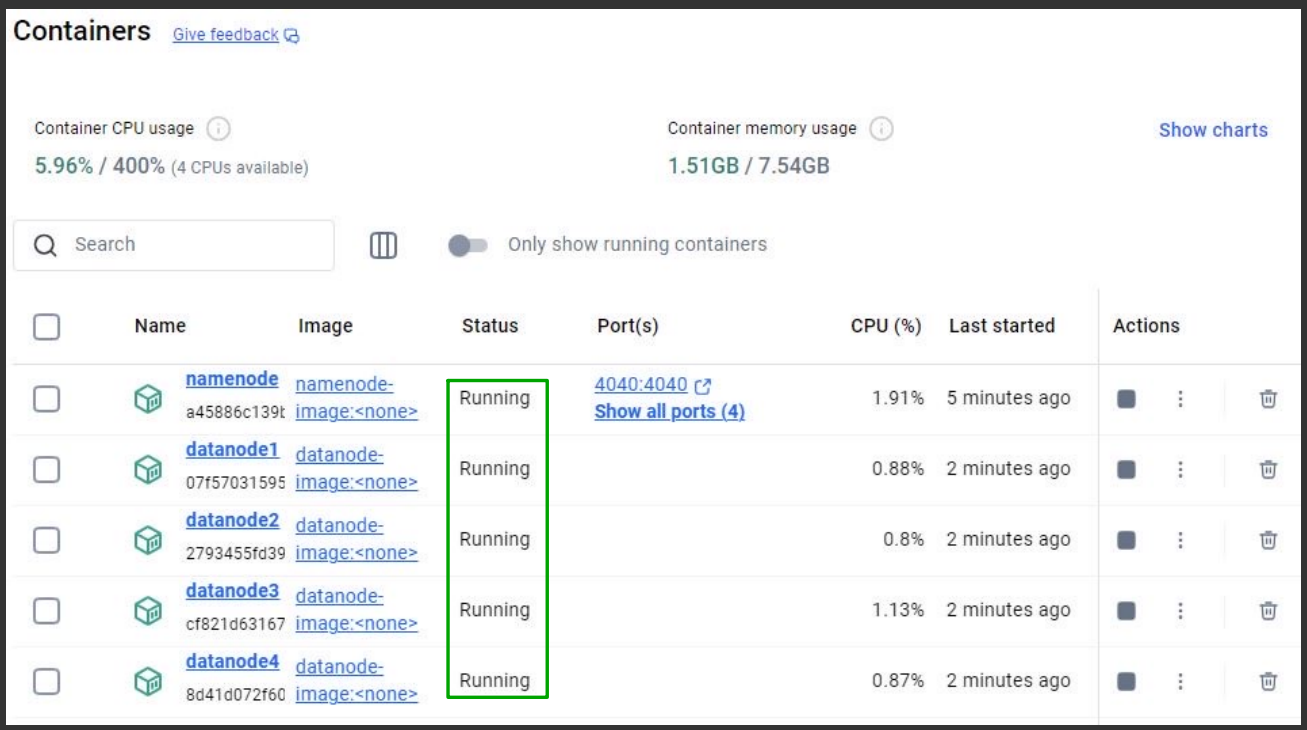
```
PS C:\Users\Alejandro> for ($i=1; $i -le 4; $i++) {    docker container run -d --name "datanode$i" --network hadoo
p-cluster --hostname "datanode$i" --cpus 1 --memory 3072m --expose 8000-10000 --expose 50000-50200 datanode-image
/inicio.sh}
07f570315951eca4970e44f66a035d95e2d16ccab7dfd19bdba573e317611df5
2793455fd3902747368436fffca4702b32ef4128212d18fa97b70a1824f3c666
cf821d63167fb287aff9a1e4197a1142b4057902bac6b2992f489747f1507b14
8d41d072f60706be31114da0de14a310bbc0c93f1c8e4977f6946a324f642667
```

2.3. Compruebo que los contenedores se están ejecutando

Desde la terminal

```
PS C:\Users\Alejandro> docker container ps
CONTAINER ID   IMAGE            COMMAND          CREATED          STATUS          PORTS

                NAMES
8d41d072f607   datanode-image   "/inicio.sh"     50 seconds ago   Up 48 seconds   8000-10000/tcp, 50000-50200/tcp

                datanode4
cf821d63167f   datanode-image   "/inicio.sh"     50 seconds ago   Up 49 seconds   8000-10000/tcp, 50000-50200/tcp

                datanode3
2793455fd390   datanode-image   "/inicio.sh"     51 seconds ago   Up 50 seconds   8000-10000/tcp, 50000-50200/tcp

                datanode2
07f570315951   datanode-image   "/inicio.sh"     52 seconds ago   Up 51 seconds   8000-10000/tcp, 50000-50200/tcp

                datanode1
a45886c139b7   namenode-image   "/inicio.sh"     4 minutes ago    Up 4 minutes    0.0.0.0:4040->4040/tcp, 8000-80
87/tcp, 0.0.0.0:8088->8088/tcp, 8089-8887/tcp, 0.0.0.0:8888->8888/tcp, 8889-9869/tcp, 0.0.0.0:9870->9870/tcp, 9
871-10000/tcp   namenode
```

Desde Docker



2.4. Me conecto a Namenode y compruebo que todo va bien

Conectar

```
PS C:\Users\Alejandro> docker container exec -ti namenode /bin/bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ hdfs dfsadmin -report
```

→ Comprobación 1

```
Configured Capacity: 4324404707328 (3.93 TB)
Present Capacity: 4072034336768 (3.70 TB)
DFS Remaining: 4072034238464 (3.70 TB)
DFS Used: 98304 (96 KB)
DFS Used%: 0.00%
```

→ Todo bien, vacío porque no hay data todavía

```
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
```

↓ Sigue, pero está recortado

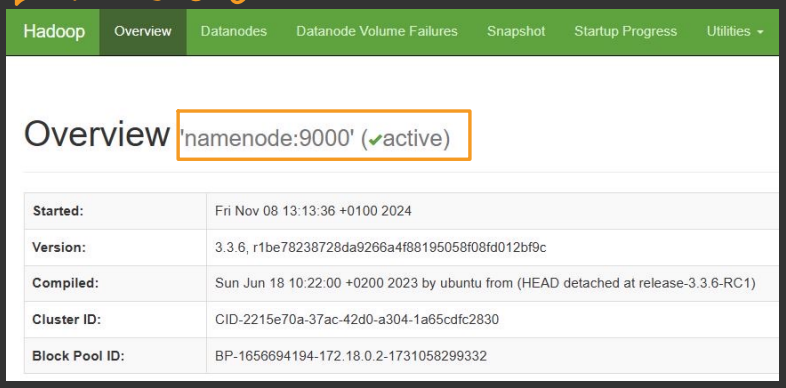
```
hdadmin@namenode:~$ yarn node -list
2024-11-08 13:20:03,678 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at resour
cemanager/172.18.0.2:8032
Total Nodes:4
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
 datanode3:43531                RUNNING       datanode3:8042                             0
 datanode1:46775                RUNNING       datanode1:8042                             0
 datanode2:45077                RUNNING       datanode2:8042                             0
 datanode4:44999                RUNNING       datanode4:8042                             0
```

→ Comprobación 2

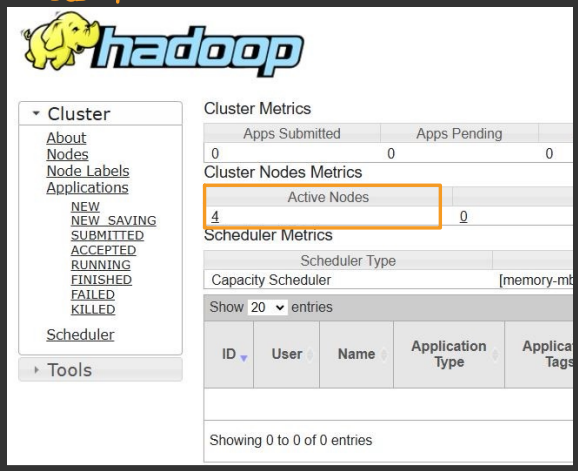
Todo bien

2.5. Compruebo si va bien en las interfaces web puestas en el apartado 1.1.5.

El namenode



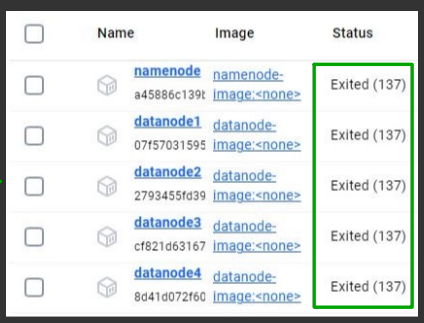
Los 4 datanodes

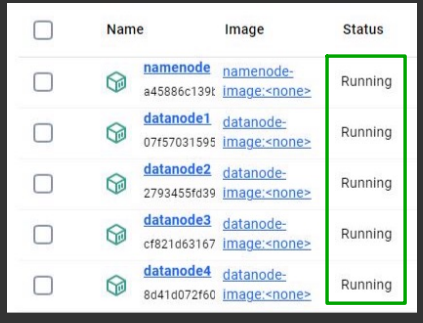


NOTA. A partir de este punto podemos detener y volver a iniciar los contenedores sin problema.

- docker container stop nombres_contenedores
- docker container start nombres_contenedores

```
hdadmin@namenode:~$ exit
logout
root@namenode:/# exit
exit
PS C:\Users\Alejandro> docker container stop namenode datanode1 datanode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
PS C:\Users\Alejandro> docker container start namenode datanode1 datanode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
```





⚠ DETENER LOS CONTENEDORES ANTES DE APAGAR EL ORDENADOR ⚠

3. Creación de directorios en HDFS y copia de datos

NOTA: se asume que esto de ha hecho en otro momento y se han detenido los contenedores antes de apagar el ordenador anteriormente. Ahora hay que iniciar los contenedores.

APUNTES IMPORTANTES DE LOGÍSTICA:

Inicio de los contenedores

```
PS C:\Users\Alejandro> docker container start namenode datanode1 datanode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
```

Entrar en un contenedor, como por ejemplo 'namenode'

```
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode bash
root@namenode:/#
```

Pasar al usuario 'hdadmin' y volver al 'root'

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ exit
logout
root@namenode:/#
```

Idem con el 'luser'

```
root@namenode:/# su - luser
luser@namenode:~$ exit
logout
```

Salir del contenedor 'namenode'

```
root@namenode:/# exit
exit

What's next:
    Try Docker Debug for seamless, persistent debugging tools in any container or image → do
cker debug namenode
    Learn more at https://docs.docker.com/go/debug-cli/
PS D:\Alex\Universidad\MasterBigData>
```

3.1. Creo los directorios de los usuarios de HDFS

```
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ hdfs dfs -mkdir -p /user/hdadmin
hdadmin@namenode:~$ hdfs dfs -mkdir -p /user/luser
hdadmin@namenode:~$ hdfs dfs -chown luser /user/luser
hdadmin@namenode:~$ hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - hdadmin supergroup          0 2024-11-21 11:31 /user/hdadmin
drwxr-xr-x   - luser   supergroup          0 2024-11-21 11:32 /user/luser
hdadmin@namenode:~$ hdfs dfs -mkdir -p /tmp/hadoop-yarn/staging
hdadmin@namenode:~$ hdfs dfs -mkdir -p /tmp/logs
hdadmin@namenode:~$ hdfs dfs -chmod -R 1777 /tmp
hdadmin@namenode:~$ hdfs dfs -chmod -R 1777 /tmp/logs
hdadmin@namenode:~$ exit
logout
```

Todo bien

3.2.1. / 3.2.2. Copia de ficheros de usuario

```
root@namenode:/# su - luser
luser@namenode:~$ hdfs dfs -ls
-ls : Unknown command
Usage: hadoop fs [generic options]
        [-appendToFile [-n] <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum [-v] <src> ...]
```

Por algún motivo, la 1ª vez que lo he puesto no ha funcionado...

```
luser@namenode:~$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - luser supergroup          0 2024-11-22 11:58 libros
```

.. y por algún otro motivo, a la 2ª vez va, pero esto debería estar vacío..

Como no es la 1ª vez que intento esto (llevo 2 días ya :'), compruebo dentro de "/tmp", a ver si están los libros

```
luser@namenode:~$ ls /tmp
hadoop-hdadmin-namenode.pid
hadoop-hdadmin-resourcemanager.pid
hadoop-yarn-hdadmin
hsperfdata_hdadmin
hsperfdata_luser
hsperfdata_root
jetty-namenode-9870-hdfs-_-any-10037146432623863710
jetty-namenode-9870-hdfs-_-any-13971451133253581035
jetty-namenode-9870-hdfs-_-any-5623974998819332113
jetty-namenode-9870-hdfs-_-any-9695388156095628349
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-1201
4261916678849614
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-1271
2035372150415108
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-1410
8366918329738711
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-1795
4952690445978597
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-6768
699447093275076
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-6786
095627193502440
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-6939
096865447265544
jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_-any-7695
698252678225599
luser@namenode:~$ ls /tmp/libros
ls: cannot access '/tmp/libros': No such file or directory
```

Nota: no se por qué, pero tras acabar este apartado (3) apagar ordenador, reiniciar y demás, cuando ejecuto `hdfs dfs -ls` → Sale bien :)

No están, yo no entiendo nada :'

No tengo ni idea de cuál es el problema.

3.2.3. Archivos de libros

```
PS D:\Alex\Universidad\MasterBigData> docker container cp C:\Users
\Alejandro\Downloads\libros.tar namenode:/tmp
Successfully copied 336MB to namenode:/tmp
```

→ Bien

3.2.4. 'Destarear' el fichero y copiar los datos a HDFS

```
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode bash

root@namenode:/# su - luser
luser@namenode:~$ cd /tmp
luser@namenode:/tmp$ tar xvf libros.tar
libros/
libros/pg17013.txt.gz
libros/pg16625.txt.gz
libros/pg25807.txt.gz
libros/pg32315.txt.gz
libros/random_words.txt.bz2
libros/pg1619.txt.gz
libros/pg7109.txt.gz
libros/pg2000.txt.gz
libros/pg14329.txt.gz
libros/pg24536.txt.gz
libros/pg25640.txt.gz
libros/pg8870.txt.gz
libros/pg18005.txt.gz
libros/pg5201.txt.gz
libros/pg9980.txt.gz
libros/pg17073.txt.gz
```

Bien

Poner los libros

```
luser@namenode:/tmp$ hdfs dfs -put libros .
put: `libros/pg8870.txt.gz': File exists
put: `libros/pg14329.txt.gz': File exists
put: `libros/pg2000.txt.gz': File exists
put: `libros/pg17073.txt.gz': File exists
put: `libros/pg5201.txt.gz': File exists
put: `libros/pg9980.txt.gz': File exists
put: `libros/pg18005.txt.gz': File exists
put: `libros/pg32315.txt.gz': File exists
put: `libros/pg1619.txt.gz': File exists
put: `libros/pg24536.txt.gz': File exists
put: `libros/random_words.txt.bz2': File exists
put: `libros/pg25807.txt.gz': File exists
put: `libros/pg7109.txt.gz': File exists
put: `libros/pg17013.txt.gz': File exists
put: `libros/pg16625.txt.gz': File exists
put: `libros/pg25640.txt.gz': File exists
```

(No olvidar el punto)

```
luser@namenode:/tmp$ hdfs dfs -ls libros
Found 16 items
-rw-r--r--   3 luser supergroup     441804 2024-11-22 11:57 libros
/pg14329.txt.gz
-rw-r--r--   3 luser supergroup     264123 2024-11-22 11:57 libros
/pg1619.txt.gz
-rw-r--r--   3 luser supergroup     455129 2024-11-22 11:58 libros/pg16625.txt.gz
-rw-r--r--   3 luser supergroup     939502 2024-11-22 11:58 libros/pg17013.txt.gz
-rw-r--r--   3 luser supergroup     737367 2024-11-22 11:57 libros/pg17073.txt.gz
-rw-r--r--   3 luser supergroup     219304 2024-11-22 11:57 libros/pg18005.txt.gz
-rw-r--r--   3 luser supergroup     813698 2024-11-22 11:57 libros/pg2000.txt.gz
-rw-r--r--   3 luser supergroup     328494 2024-11-22 11:57 libros/pg24536.txt.gz
-rw-r--r--   3 luser supergroup     504188 2024-11-22 11:58 libros/pg25640.txt.gz
-rw-r--r--   3 luser supergroup      38194 2024-11-22 11:58 libros/pg25807.txt.gz
-rw-r--r--   3 luser supergroup     103986 2024-11-22 11:57 libros/pg32315.txt.gz
-rw-r--r--   3 luser supergroup     125693 2024-11-22 11:57 libros/pg5201.txt.gz
-rw-r--r--   3 luser supergroup      82099 2024-11-22 11:58 libros/pg7109.txt.gz
-rw-r--r--   3 luser supergroup      99685 2024-11-22 11:57 libros/pg8870.txt.gz
-rw-r--r--   3 luser supergroup      85187 2024-11-22 11:57 libros/pg9980.txt.gz
-rw-r--r--   3 luser supergroup  330326458 2024-11-22 11:58 libros/random_words.txt.b
z2
```

3 réplicas

Borro el directorio "libros" del disco local

```
luser@namenode:/tmp$ rm -rf /tmp/libros
luser@namenode:/tmp$ exit
logout
```

Borro el fichero "libros.tar"

```
root@namenode:/# rm -f /tmp/libros.tar
root@namenode:/# ls -l /tmp
total 84
-rw-r--r-- 1 hdadmin hadoop    3 Nov 22 11:37 hadoop-hdadmin-namenode.pid
-rw-r--r-- 1 hdadmin hadoop    3 Nov 22 11:37 hadoop-hdadmin-resourcemanager.pid
drwxr-xr-x 1 hdadmin hadoop 4096 Nov  8 10:43 hadoop-yarn-hdadmin
drwxr-xr-x 1 hdadmin hadoop 4096 Nov 22 11:55 hsperfdata_hdadmin
drwxr-xr-x 2 luser   hadoop 4096 Nov 22 11:58 hsperfdata_luser
drwxr-xr-x 2 root    root   4096 Sep 19 01:12 hsperfdata_root
drwx------ 2 hdadmin hadoop 4096 Nov 22 11:25 jetty-namenode-9870-hdfs-_-any-10037146432623863710
drwx------ 2 hdadmin hadoop 4096 Nov 22 11:37 jetty-namenode-9870-hdfs-_-any-13971451133253581035
drwx------ 2 hdadmin hadoop 4096 Nov 21 11:08 jetty-namenode-9870-hdfs-_-any-14510574086946344691
drwx------ 2 hdadmin hadoop 4096 Nov 21 12:28 jetty-namenode-9870-hdfs-_-any-5623974998819332113
drwx------ 2 hdadmin hadoop 4096 Nov  8 13:13 jetty-namenode-9870-hdfs-_-any-5999307908324819038
drwx------ 2 hdadmin hadoop 4096 Nov 21 18:58 jetty-namenode-9870-hdfs-_-any-6592058203630617326
drwx------ 2 hdadmin hadoop 4096 Nov  8 13:36 jetty-namenode-9870-hdfs-_-any-9695388156095628349
drwx------ 3 hdadmin hadoop 4096 Nov 21 18:58 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-12712035372150415108
drwx------ 3 hdadmin hadoop 4096 Nov 21 12:28 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-14108366918329738711
drwx------ 3 hdadmin hadoop 4096 Nov 22 11:38 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-17954952690445978597
drwx------ 3 hdadmin hadoop 4096 Nov  8 13:36 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-6768699447093275076
drwx------ 3 hdadmin hadoop 4096 Nov  8 13:13 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-6786095627193502440
drwx------ 3 hdadmin hadoop 4096 Nov 21 11:08 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-6939096865447265544
drwx------ 3 hdadmin hadoop 4096 Nov 22 11:25 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_jar-_
-any-7695698252678225599
root@namenode:/#
```

→ Para comprobar si está o se ha borrado

No aparece → Borrado ✓

Nota: si apareciera ...

```
root@namenode:/# ls -l /tmp
total 327804
-rw-r--r-- 1 hdadmin hadoop         3 Nov 22 11:37 hadoop-hdadmin-namenode.pid
-rw-r--r-- 1 hdadmin hadoop         3 Nov 22 11:37 hadoop-hdadmin-resourcemanager.pid
drwxr-xr-x 1 hdadmin hadoop      4096 Nov  8 10:43 hadoop-yarn-hdadmin
drwxr-xr-x 1 hdadmin hadoop      4096 Nov 22 11:40 hsperfdata_hdadmin
drwxr-xr-x 2 luser   hadoop      4096 Nov 22 11:41 hsperfdata_luser
drwxr-xr-x 2 root    root        4096 Sep 19 01:12 hsperfdata_root
drwx------ 2 hdadmin hadoop      4096 Nov 22 11:25 jetty-namenode-9870-hdfs-_-any-10037146432623863710
drwx------ 2 hdadmin hadoop      4096 Nov 22 11:37 jetty-namenode-9870-hdfs-_-any-13971451133253581035
drwx------ 2 hdadmin hadoop      4096 Nov 21 11:08 jetty-namenode-9870-hdfs-_-any-14510574086946344691
drwx------ 2 hdadmin hadoop      4096 Nov 21 12:28 jetty-namenode-9870-hdfs-_-any-5623974998819332113
drwx------ 2 hdadmin hadoop      4096 Nov  8 13:13 jetty-namenode-9870-hdfs-_-any-5999307908324819038
drwx------ 2 hdadmin hadoop      4096 Nov 21 18:58 jetty-namenode-9870-hdfs-_-any-6592058203630617326
drwx------ 2 hdadmin hadoop      4096 Nov  8 13:36 jetty-namenode-9870-hdfs-_-any-9695388156095628349
drwx------ 3 hdadmin hadoop      4096 Nov 21 18:58 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-12712035372150415108
drwx------ 3 hdadmin hadoop      4096 Nov 21 12:28 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-14108366918329738711
drwx------ 3 hdadmin hadoop      4096 Nov 22 11:38 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-17954952690445978597
drwx------ 3 hdadmin hadoop      4096 Nov  8 13:36 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-6768699447093275076
drwx------ 3 hdadmin hadoop      4096 Nov  8 13:13 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-6786095627193502440
drwx------ 3 hdadmin hadoop      4096 Nov 21 11:08 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-6939096865447265544
drwx------ 3 hdadmin hadoop      4096 Nov 22 11:25 jetty-resourcemanager-8088-hadoop-yarn-common-3_3_6_
jar-_-any-7695698252678225599
-rwxr-xr-x 1 root    root    335585280 Nov 22 11:43 libros.tar
```

... aparecería así.

4. Prueba de aplicaciones MapReduce simples

4.1. Aplicación MapReduce para el cálculo de π (pi)

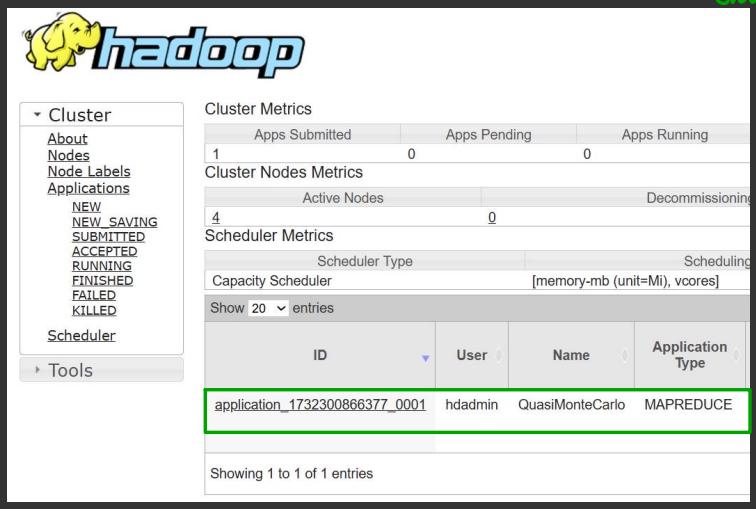
```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ export MAPRED_EXAMPLES=$HADOOP_HOME/share/hadoop/mapred
uce
hdadmin@namenode:~$ yarn jar $MAPRED_EXAMPLES/hadoop-mapreduce-examples-*.j
ar pi 16 1000
Number of Maps  = 16
Samples per Map = 1000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Starting Job
```

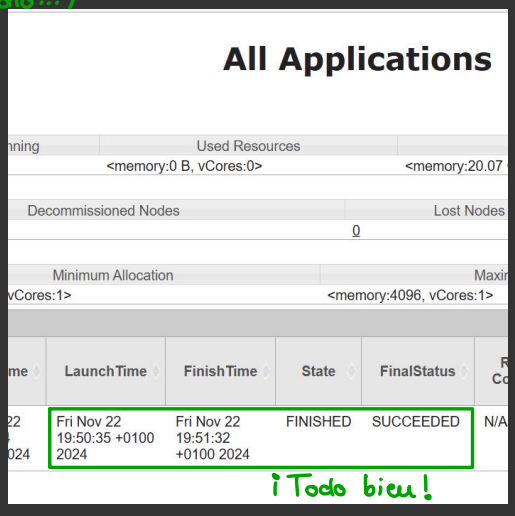
Más respuesta, hasta llegar a...

Resultado: ¡todo bien!

```
Job Finished in 62.41 seconds
Estimated value of Pi is 3.14250000000000000000
```

Comprueba ahora la interfaz de YARN (Más cosas en medio...)





4.2. Aplicación wordcount

4.2.1. Copio en NameNode

```
PS D:\Alex\Universidad\MasterBigData> docker cp C:\Users\Alejandro\Downloads\wordcount.tgz namenode:/home/luser
```

4.2.2. Entro en NameNode → luser y ejecuto una a una las filas:

```
cd
tar xvzf wordcount.tgz
cd wordcount
mvn package
```

Tiene, al final y tras muchísima respuesta, que dar esto:

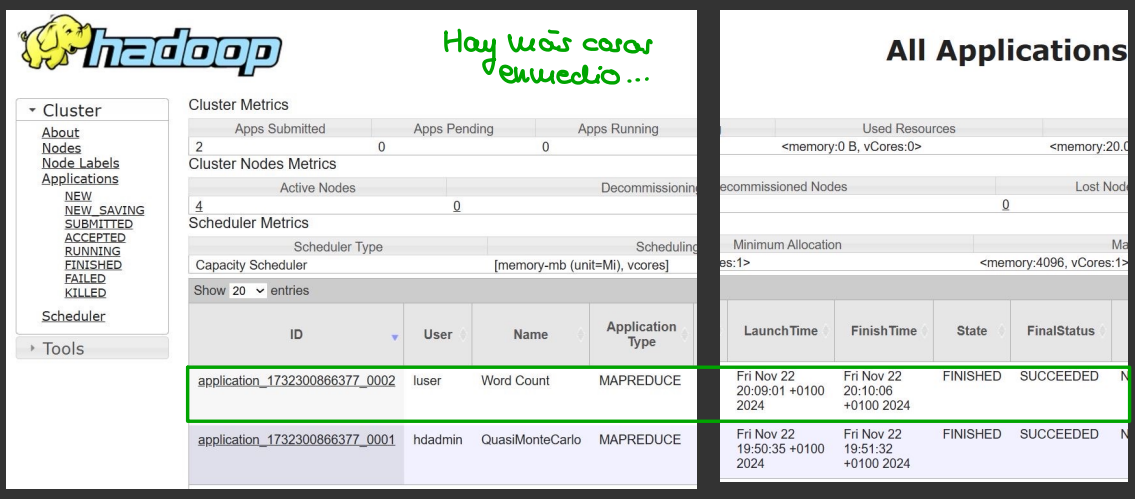
```
[INFO] Building jar: /home/luser/wordcount/target/wordcount-0.0.1-SNAPSHOT.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  45.414 s
[INFO] Finished at: 2024-11-22T20:08:32+01:00
[INFO] ------------------------------------------------------------------------
```

4.2.3.

```
luser@namenode:~/wordcount$ yarn jar target/wordcount*.jar libros/p* wordcount-out
2024-11-22 20:08:58,985 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at resourcemanager/172.18.0.2:8032
2024-11-22 20:08:59,787 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/luser/.staging/job_1732300866377_0002
2024-11-22 20:09:00,404 INFO input.FileInputFormat: Total input files to process : 15
```

Hay más respuesta...

Comprobando en la interfaz de YARN



4.2.4. / 4.2.5

Llevo los ficheros al disco local de Namenode y compruebo

```
luser@namenode:~/wordcount$ hdfs dfs -get wordcount-out
luser@namenode:~/wordcount$ ls -l wordcount-out
total 1072
-rw-r--r-- 1 luser hadoop 363947 nov 22 20:17 part-r-00000
-rw-r--r-- 1 luser hadoop 365976 nov 22 20:17 part-r-00001
-rw-r--r-- 1 luser hadoop 361370 nov 22 20:17 part-r-00002
-rw-r--r-- 1 luser hadoop      0 nov 22 20:17 _SUCCESS
```

Asumo que está bien pero no estoy seguro

Tarea 1

T1.1.0. Antes de nada, comprobar los metadatos del NameNode dentro de current

```
PS D:\Alex\Universidad\MasterBigData> docker container start na
menode datanode1 datanode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode
bash
```

Inicio de todo y entrada en NameNode

Vamos a current

```
root@namenode:/# cd /var/data/hdfs/namenode/current
```

```
root@namenode:/var/data/hdfs/namenode/current# ls -l
total 12320
-rw-r--r-- 1 hdadmin hadoop     214 Dec  4 09:22 VERSION
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 10:40 edits_0000000000000000001-0000000000000000001
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:13 edits_0000000000000000002-0000000000000000002
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:36 edits_0000000000000000003-0000000000000000003
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:07 edits_0000000000000000004-0000000000000000035
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:28 edits_0000000000000000036-0000000000000000036
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 18:58 edits_0000000000000000037-0000000000000000037
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:26 edits_0000000000000000038-0000000000000000039
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:58 edits_0000000000000000040-0000000000000000150
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 17:26 edits_0000000000000000151-0000000000000000151
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 20:10 edits_0000000000000000152-0000000000000000530
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  2 16:57 edits_0000000000000000531-0000000000000000531
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 09:22 edits_inprogress_0000000000000000532
-rw-r--r-- 1 hdadmin hadoop    5329 Dec  2 16:57 fsimage_0000000000000000530
-rw-r--r-- 1 hdadmin hadoop      62 Dec  2 16:57 fsimage_0000000000000000530.md5
-rw-r--r-- 1 hdadmin hadoop    5295 Dec  4 09:22 fsimage_0000000000000000531
-rw-r--r-- 1 hdadmin hadoop      62 Dec  4 09:22 fsimage_0000000000000000531.md5
-rw-r--r-- 1 hdadmin hadoop       4 Dec  4 09:22 seen_txid
```

Dejamos esto aquí de momento, más adelante se usará

T1.1.1. Nuevo Docker: backupnode

```
PS D:\Alex\Universidad\MasterBigData> docker container run -ti --name b
ackupnode --network=hadoop-cluster --hostname backupnode --cpus=1 --mem
ory=3072m --expose=50100 -p 50105:50105 dsevilla/hadoop-base /bin/bash
root@backupnode:/#
```

T1.1.2. Directorio donde se guardarán los backups. hdadmin será el propietario y se crea dentro la carpeta dfs/name

```
root@backupnode:/# mkdir -p /var/data/hdfs/dfs/name
root@backupnode:/# chown hdadmin:hadoop /var/data/hdfs/dfs/name
root@backupnode:/# ls -l /var/data/hdfs/dfs
total 4
drwxr-xr-x 2 hdadmin hadoop 4096 Dec  4 09:50 name
```

Veo qué hay en 'dfs'

→ Está 'name', todo bien

T1.1.3. Como hdadmin añado propiedades al fichero 'core-site.xml':

```
root@backupnode:/# su - hdadmin
hdadmin@backupnode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@backupnode:~/hadoop/etc/hadoop$ nano core-site.xml
```

```
<configuration>
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://namenode:9000/</value>
    <final>true</final>
</property>
<property>
    <name>hadoop.tmp.dir</name>
    <value>/var/data/hdfs</value>
    <final>true</final>
</property>
</configuration>
```

T1.1.4. Como hdadmin añado propiedades al fichero 'hdfs-site.xml':

```
hdadmin@backupnode:~/hadoop/etc/hadoop$ nano hsfs-site.xml
```

```
<configuration>
<property>
    <name>dfs.namenode.backup.address</name>
    <value>backupnode:50100</value>
    <final>true</final>
</property>
<property>
    <name>dfs.namenode.backup.http-address</name>
    <value>backupnode:50105</value>
    <final>true</final>
</property>
</configuration>
```

T1.1.5. Iniciar el servidor de backup

```
2024-12-04 17:39:12,647 INFO namenode.FSImage: Reading /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001 expecting start txid #1
2024-12-04 17:39:12,648 INFO namenode.FSImage: Start loading edits file /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001 maxTxnsToRead = 9223372036854775807
2024-12-04 17:39:12,698 INFO namenode.FSImage: Loaded 1 edits file(s) (the last named /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001) of total size 1048576.0, total edits 1.0, total load time 33.0 ms
2024-12-04 17:39:13,063 INFO namenode.NameCache: initialized with 0 entries 0 lookups
2024-12-04 17:39:13,063 INFO namenode.LeaseManager: Number of blocks under construction: 0
2024-12-04 17:39:13,093 INFO namenode.FSImageFormatProtobuf: Saving image file /var/data/hdfs/dfs/name/current/fsimage.ckpt_0000000000000000545 using no compression
2024-12-04 17:39:13,280 INFO namenode.FSImageFormatProtobuf: Image file /var/data/hdfs/dfs/name/current/fsimage.ckpt_0000000000000000545 of size 5329 bytes saved in 0 seconds .
2024-12-04 17:39:13,289 INFO namenode.FSImageTransactionalStorageInspector: No version file in /var/data/hdfs/dfs/name
2024-12-04 17:39:13,300 INFO namenode.FSImageTransactionalStorageInspector: No version file in /var/data/hdfs/dfs/name
2024-12-04 17:39:13,388 INFO namenode.TransferFsImage: Image Transfer timeout configured to 60000 milliseconds
2024-12-04 17:39:13,390 INFO namenode.TransferFsImage: Sending fileName: /var/data/hdfs/dfs/name/current/fsimage_0000000000000000545, fileSize: 5329. Sent total: 5329 bytes. Size of last segment intended to send: -1 bytes.
2024-12-04 17:39:13,414 INFO namenode.TransferFsImage: Uploaded image with txid 545 to namenode at http://namenode:9870 in 0.051 seconds
2024-12-04 17:39:13,454 INFO namenode.FSImage: Going to finish converging with remaining 1 txns from in-progress stream org.apache.hadoop.hdfs.server.namenode.RedundantEditLogInputStream@357d4a7d
2024-12-04 17:39:13,455 INFO namenode.RedundantEditLogInputStream: Fast-forwarding stream '/var/data/hdfs/dfs/name/current/edits_inprogress_0000000000000000546' to transaction ID 546
2024-12-04 17:39:13,462 INFO namenode.FSImage: Successfully synced BackupNode with NameNode at txnid 546
2024-12-04 17:39:13,463 INFO namenode.Checkpointer: Checkpoint completed in 1 seconds. New Image Size: 5329
```

T1.1.6. Comprobar los metadatatos del backupnode y compararlos con los del NameNode

Del backupnode

```
hdadmin@backupnode:~$ ls -l /var/data/hdfs/dfs/name/current
total 14376
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000001-0000000000000000001
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000002-0000000000000000002
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000003-0000000000000000003
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000004-0000000000000000035
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000036-0000000000000000036
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000037-0000000000000000037
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000038-0000000000000000039
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000040-0000000000000000150
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000151-0000000000000000151
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000152-0000000000000000530
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000531-0000000000000000531
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000532-0000000000000000533
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000534-0000000000000000535
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000536-0000000000000000537
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000538-0000000000000000539
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000540-0000000000000000541
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000542-0000000000000000542
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_0000000000000000543-0000000000000000543
-rw-r--r-- 1 hdadmin hadoop      42 dic  4 17:39 edits_0000000000000000544-0000000000000000545
-rw-r--r-- 1 hdadmin hadoop 1048576 dic  4 17:39 edits_inprogress_0000000000000000546
-rw-r--r-- 1 hdadmin hadoop    5329 dic  4 17:39 fsimage_0000000000000000545
-rw-r--r-- 1 hdadmin hadoop      62 dic  4 17:39 fsimage_0000000000000000545.md5
-rw-r--r-- 1 hdadmin hadoop     214 dic  4 17:39 VERSION
```

Del NameNode una vez ejecutado el servicio de backup

```
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode bash
root@namenode:/# cd /var/data/hdfs/namenode/current
root@namenode:/var/data/hdfs/namenode/current# ls -l
total 14392
-rw-r--r-- 1 hdadmin hadoop     214 Dec  4 16:46 VERSION
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 10:40 edits_0000000000000000001-0000000000000000001
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:13 edits_0000000000000000002-0000000000000000002
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:36 edits_0000000000000000003-0000000000000000003
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:07 edits_0000000000000000004-0000000000000000035
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:28 edits_0000000000000000036-0000000000000000036
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 18:58 edits_0000000000000000037-0000000000000000037
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:26 edits_0000000000000000038-0000000000000000039
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:58 edits_0000000000000000040-0000000000000000150
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 17:26 edits_0000000000000000151-0000000000000000151
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 20:10 edits_0000000000000000152-0000000000000000530
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  2 16:57 edits_0000000000000000531-0000000000000000531
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 10:21 edits_0000000000000000532-0000000000000000533
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:16 edits_0000000000000000534-0000000000000000535
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:19 edits_0000000000000000536-0000000000000000537
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:50 edits_0000000000000000538-0000000000000000539
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 13:22 edits_0000000000000000540-0000000000000000541
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 13:22 edits_0000000000000000542-0000000000000000542
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 16:46 edits_0000000000000000543-0000000000000000543
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 17:39 edits_0000000000000000544-0000000000000000545
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 17:39 edits_inprogress_0000000000000000546
-rw-r--r-- 1 hdadmin hadoop    5295 Dec  4 16:46 fsimage_0000000000000000542
-rw-r--r-- 1 hdadmin hadoop      62 Dec  4 16:46 fsimage_0000000000000000542.md5
-rw-r--r-- 1 hdadmin hadoop    5329 Dec  4 17:39 fsimage_0000000000000000545
-rw-r--r-- 1 hdadmin hadoop      62 Dec  4 17:39 fsimage_0000000000000000545.md5
-rw-r--r-- 1 hdadmin hadoop       4 Dec  4 17:39 seen_txid
```

Del NameNode antes de crear el backupnode

```
root@namenode:/var/data/hdfs/namenode/current# ls -l
total 13364
-rw-r--r-- 1 hdadmin hadoop     214 Dec  4 16:46 VERSION
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 10:40 edits_0000000000000000001-0000000000000000001
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:13 edits_0000000000000000002-0000000000000000002
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov  8 13:36 edits_0000000000000000003-0000000000000000003
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:07 edits_0000000000000000004-0000000000000000035
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 12:28 edits_0000000000000000036-0000000000000000036
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 21 18:58 edits_0000000000000000037-0000000000000000037
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:26 edits_0000000000000000038-0000000000000000039
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 11:58 edits_0000000000000000040-0000000000000000150
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 17:26 edits_0000000000000000151-0000000000000000151
-rw-r--r-- 1 hdadmin hadoop 1048576 Nov 22 20:10 edits_0000000000000000152-0000000000000000530
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  2 16:57 edits_0000000000000000531-0000000000000000531
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 10:21 edits_0000000000000000532-0000000000000000533
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:16 edits_0000000000000000534-0000000000000000535
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:19 edits_0000000000000000536-0000000000000000537
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 11:50 edits_0000000000000000538-0000000000000000539
-rw-r--r-- 1 hdadmin hadoop      42 Dec  4 13:22 edits_0000000000000000540-0000000000000000541
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 13:22 edits_0000000000000000542-0000000000000000542
-rw-r--r-- 1 hdadmin hadoop 1048576 Dec  4 16:46 edits_inprogress_0000000000000000543
-rw-r--r-- 1 hdadmin hadoop    5329 Dec  4 13:22 fsimage_0000000000000000541
-rw-r--r-- 1 hdadmin hadoop      62 Dec  4 13:22 fsimage_0000000000000000541.md5
-rw-r--r-- 1 hdadmin hadoop    5295 Dec  4 16:46 fsimage_0000000000000000542
-rw-r--r-- 1 hdadmin hadoop      62 Dec  4 16:46 fsimage_0000000000000000542.md5
-rw-r--r-- 1 hdadmin hadoop       4 Dec  4 16:46 seen_txid
```

T1.1.7. Mira en los mensajes del servicio de backup información que indique que se ha realizado un checkpoint

Esto lo he resuelto en el apartado T.1.5, lo vuelvo a poner

```
2024-12-04 17:39:12,647 INFO namenode.FSImage: Reading /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001 expecting start txid #1
2024-12-04 17:39:12,648 INFO namenode.FSImage: Start loading edits file /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001 maxTxnsToRead = 9223372036854775807
2024-12-04 17:39:12,698 INFO namenode.FSImage: Loaded 1 edits file(s) (the last named /var/data/hdfs/dfs/name/current/edits_0000000000000000001-0000000000000000001) of total size 1048576.0, total edits 1.0, total load time 33.0 ms
2024-12-04 17:39:13,063 INFO namenode.NameCache: initialized with 0 entries 0 lookups
2024-12-04 17:39:13,063 INFO namenode.LeaseManager: Number of blocks under construction: 0
2024-12-04 17:39:13,093 INFO namenode.FSImageFormatProtobuf: Saving image file /var/data/hdfs/dfs/name/current/fsimage.ckpt_0000000000000000545 using no compression
2024-12-04 17:39:13,280 INFO namenode.FSImageFormatProtobuf: Image file /var/data/hdfs/dfs/name/current/fsimage.ckpt_0000000000000000545 of size 5329 bytes saved in 0 seconds .
2024-12-04 17:39:13,289 INFO namenode.FSImageTransactionalStorageInspector: No version file in /var/data/hdfs/dfs/name
2024-12-04 17:39:13,300 INFO namenode.FSImageTransactionalStorageInspector: No version file in /var/data/hdfs/dfs/name
2024-12-04 17:39:13,388 INFO namenode.TransferFsImage: Image Transfer timeout configured to 60000 milliseconds
2024-12-04 17:39:13,390 INFO namenode.TransferFsImage: Sending fileName: /var/data/hdfs/dfs/name/current/fsimage_0000000000000000545, fileSize: 5329. Sent total: 5329 bytes. Size of last segment intended to send: -1 bytes.
2024-12-04 17:39:13,414 INFO namenode.TransferFsImage: Uploaded image with txid 545 to namenode at http://namenode:9870 in 0.051 seconds
2024-12-04 17:39:13,454 INFO namenode.FSImage: Going to finish converging with remaining 1 txns from in-progress stream org.apache.hadoop.hdfs.server.namenode.RedundantEditLogInputStream@357d4a7d
2024-12-04 17:39:13,455 INFO namenode.RedundantEditLogInputStream: Fast-forwarding stream '/var/data/hdfs/dfs/name/current/edits_inprogress_0000000000000000546' to transaction ID 546
2024-12-04 17:39:13,462 INFO namenode.FSImage: Successfully synced BackupNode with NameNode at txnid 546
2024-12-04 17:39:13,463 INFO namenode.Checkpointer: Checkpoint completed in 1 seconds. New Image Size: 5329
```

T1.1.Nota. Para reiniciar el servicio de forma fácil, salgo del contenedor backupnode, lo guardo como imagen, lo inicio de nuevo de otra forma y compruebo que el servicio de backup se está ejecutando bien.

Guardar como imagen y parar

```
PS D:\Alex\Universidad\MasterBigData> docker container commit backupnode
backupnode-image → Guardar como imagen
sha256:ac363c84a536e15c65b682b5477d47e8874a1c6a64be2e8ce1ebe47983250854
PS D:\Alex\Universidad\MasterBigData> docker images → Compruebo
backupnode-image                                  latest    ac363c84a536
12 seconds ago    3.3GB                                                    → Todo bien
datanode-image                                    latest    82f05154f1b1
3 weeks ago       3.29GB
namenode-image                                    latest    84cb471f2da8
<none>                                            <none>    9228bf275c3b
5 weeks ago       3.29GB
<none>                                            <none>    699e4306340e
5 weeks ago       3.29GB
docker.elastic.co/kibana/kibana                   8.12.1    6c8b4a0b5197
docker.elastic.co/elasticsearch/elasticsearch     8.12.1    a476af93763c
10 months ago     2.07GB
PS D:\Alex\Universidad\MasterBigData> docker container stop backupnode
backupnode →Paro el backupnode
```

Elimino (es necesario eliminar) el backupnode y ejecuto para iniciarlo:

```
PS D:\Alex\Universidad\MasterBigData> docker container rm backupnode
backupnode
PS D:\Alex\Universidad\MasterBigData> docker container run -d --name back
upnode --network=hadoop-cluster --hostname backupnode --cpus=1 --memory=3
072m --expose=50100 -p 50105:50105 backupnode-image su - hdadmin -c "JAVA
_HOME=/usr/lib/jvm/default-java /opt/bd/hadoop/bin/hdfs namenode -backup"

e07dfff3132f44c557c47c8eedae2657deabeccee9c2f74c4d9b9148486705eb
PS D:\Alex\Universidad\MasterBigData> docker container logs backupnode →Para ver que el servicio se está ejecutando bien
2024-12-04 11:16:15,506 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = backupnode/172.18.0.7
STARTUP_MSG:   args = [-backup]
STARTUP_MSG:   version = 3.3.6
STARTUP_MSG:   classpath = /opt/bd/hadoop/etc/hadoop:/opt/bd/hadoop/share
```

(Muchos respuesta enmedio…)

```
2024-12-04 17:50:19,947 INFO namenode.NameCache: initialized with 0 entries 0 lookups
2024-12-04 17:50:19,948 INFO namenode.LeaseManager: Number of blocks under construction: 0
2024-12-04 17:50:20,045 INFO namenode.FSImageFormatProtobuf: Saving image file /var/data/hdfs/dfs/name/current
/fsimage.ckpt_0000000000000000547 using no compression
2024-12-04 17:50:20,334 INFO namenode.FSImageFormatProtobuf: Image file /var/data/hdfs/dfs/name/current/fsimag
e.ckpt_0000000000000000547 of size 5329 bytes saved in 0 seconds .
2024-12-04 17:50:20,345 INFO namenode.NNStorageRetentionManager: Going to retain 2 images with txid >= 545
2024-12-04 17:50:20,437 INFO namenode.TransferFsImage: Image Transfer timeout configured to 60000 milliseconds
2024-12-04 17:50:20,438 INFO namenode.TransferFsImage: Sending fileName: /var/data/hdfs/dfs/name/current/fsima
ge_0000000000000000547, fileSize: 5329. Sent total: 5329 bytes. Size of last segment intended to send: -1 byte
s.
2024-12-04 17:50:20,451 INFO namenode.FSImage: Going to finish converging with remaining 1 txns from in-progre
ss stream org.apache.hadoop.hdfs.server.namenode.RedundantEditLogInputStream@5334d993
2024-12-04 17:50:20,452 INFO namenode.RedundantEditLogInputStream: Fast-forwarding stream '/var/data/hdfs/dfs/
name/current/edits_inprogress_0000000000000000548' to transaction ID 548
2024-12-04 17:50:20,453 INFO namenode.FSImage: Successfully synced BackupNode with NameNode at txnid 548
2024-12-04 17:50:20,453 INFO namenode.Checkpointer: Checkpoint completed in 1 seconds. New Image Size: 5329
```

¡Todo bien!

T1.2. - TimeLineServer

T1.2.1. En el NameNode, detener el servicio ResourceManager

```
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ yarn --daemon stop resourcemanager
hdadmin@namenode:~$
```

T1.2.2. En el NameNode, editar el fichero 'yarn-site.xml' y añadir propiedades:

```
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn-site.xml
hdadmin@namenode:~/hadoop/etc/hadoop$
```

Esta, y más propiedades arriba, estaban ya y se mantienen

Propiedades que se añade

```
<property>
  <!-- Memoria maxima (MB) que un ApplicationMaster puede solicitar al RM (por defecto: 8192 MB) -->
  <!-- Peticiones mayores lanzan una InvalidResourceRequestException -->
  <!-- Puedes aumentar o reducir este valor en funcion de la memoria de la que dispongas -->
  <name>yarn.scheduler.maximum-allocation-mb</name>
  <value>4096</value>
  <final>true</final>
</property>

  <!-- Nombre del host que ejecutará el Timeline Server -->
  <property>
      <name>yarn.timeline-service.hostname</name>
      <value>timelineserver</value>
      <final>true</final>
  </property>

  <!-- Habilitar el Timeline Server -->
  <property>
      <name>yarn.timeline-service.enabled</name>
      <value>true</value>
      <final>true</final>
  </property>

  <!-- Habilitar la publicación de métricas del sistema en el Timeline Server -->
  <property>
      <name>yarn.system-metrics-publisher.enabled</name>
      <value>true</value>
      <final>true</final>
  </property>
</configuration>
```

T1.2.3. En el NameNode, reiniciar el servicio ResourceManager:

```
hdadmin@namenode:~$ yarn --daemon start resourcemanager
```

T1.2.4. Nuevo docker: TimeLineServer

¡Ojo! En la práctica está este comando que da error.

```
PS D:\Alex\Universidad\MasterBigData> docker container run -ti --name timelineserver --network=hadoop-cluster
--hostname timelineserver --cpus=1 --memory=3072m --expose=10200 -p 8188:8188 hadoop-base /bin/bash
Unable to find image 'hadoop-base:latest' locally
docker: Error response from daemon: pull access denied for hadoop-base, repository does not exist or may requi
re 'docker login'.
See 'docker run --help'.
```

Error

Para que funcione, hay que añadir:

```
PS D:\Alex\Universidad\MasterBigData> docker container run -ti --name timelineserver --network=hadoop-cluster
--hostname timelineserver --cpus=1 --memory=3072m --expose=10200 -p 8188:8188 dsevilla/hadoop-base /bin/bash
root@timelineserver:/#
```

¡Esto!

T1.2.5. Levantar el servicio TimeLineServer

```
root@timelineserver:/# su - hdadmin
hdadmin@timelineserver:~$ yarn --daemon start timelineserver
WARNING: /opt/bd/hadoop/logs does not exist. Creating.
```

→ Todo bien, lo crea porque es la 1ª vez que se hace

T1.2.6. En NameNode, ejecutar una aplicación con yarn (o bien el cálculo de pi o el wordcount)

En mi caso voy a hacer la de pi

```
PS D:\Alex\Universidad\MasterBigData> docker exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ export MAPRED_EXAMPLES=$HADOOP_HOME/share/hadoop/mapreduce
hdadmin@namenode:~$ yarn jar $MAPRED_EXAMPLES/hadoop-mapreduce-examples-*.jar pi 16 1000
Number of Maps  = 16
Samples per Map = 1000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Starting Job
2024-12-04 18:42:37,025 INFO service.AbstractService: Service org.apache.hadoop.yarn.client.api.impl.YarnClien
tImpl failed in state INITED
```

Aplicación mapreduce para el cálculo de π

Ha fallado → ¿Por qué? → No tengo ni idea

Solución. - Si el contenedor backupnode de la anterior parte se está ejecutando, lo paro.
- Si el contenedor timelineserver está parado, lo ejecuto.

Ahora bien:

1. → Abro una nueva terminal donde entro al timelineserver como hdadmin e inicio en esta nueva terminal el servicio timelineserver como se ha hecho en el apartado T.1.2.5
2. → Una vez hecho el paso 1., se repite el procedimiento de la aplicación mapreduce en la terminal original y por alguna razón todo va bien :)

2. →

```
hdadmin@namenode:~$ export MAPRED_EXAMPLES=$HADOOP_HOME/share/hadoop/mapreduce
hdadmin@namenode:~$ yarn jar $MAPRED_EXAMPLES/hadoop-mapreduce-examples-*.jar pi 16 1000
Number of Maps  = 16
Samples per Map = 1000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Wrote input for Map #10
Wrote input for Map #11
Wrote input for Map #12
Wrote input for Map #13
Wrote input for Map #14
Wrote input for Map #15
Starting Job
```

→ Después de más output dará:

```
Job Finished in 101.377 seconds
Estimated value of Pi is 3.14250000000000000000
```

→ Todo bien :)

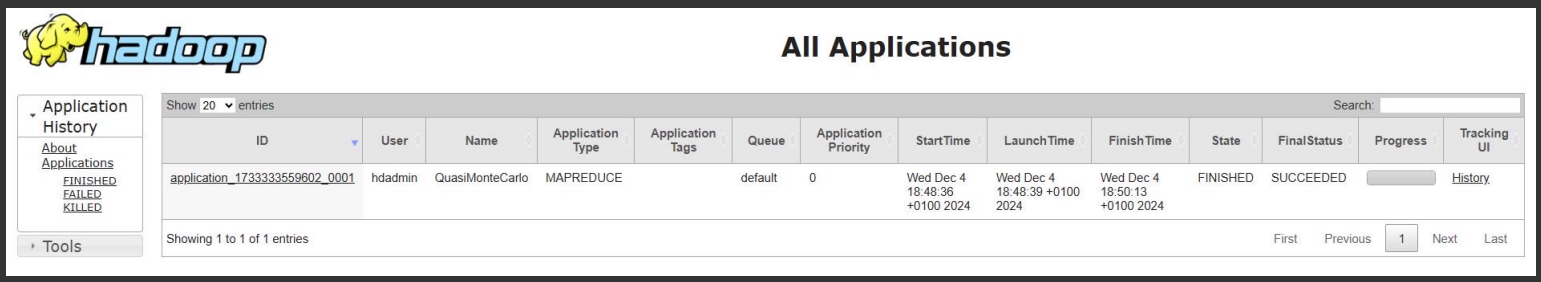
Terminal original

Terminal nueva

1 →

```
PS D:\Alex\Universidad\MasterBigData> docker start timelineserver
timelineserver
PS D:\Alex\Universidad\MasterBigData> docker exec -ti timelineserver bash
root@timelineserver:/# su - hdadmin
hdadmin@timelineserver:~$ yarn --daemon start timelineserver
hdadmin@timelineserver:~$ jps
83 Jps
59 ApplicationHistoryServer
```

T1.2.7. Comprobación en el servidor web del TimeLineServer, accediendo a través de: http://localhost:8188



Tarea 2

T2.1. Creación de ficheros de nodos incluidos y excluidos

T2.1.1. Entro en NameNode como hdadmin y paro los demonios del NameNode y ResourceManager

```
PS D:\Alex\Universidad\MasterBigData> docker container start namenode datanode1 datan
ode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ hdfs --daemon stop namenode
hdadmin@namenode:~$ yarn --daemon stop resourcemanager
hdadmin@namenode:~$ jps → Comprobar
470 Jps → Perfecto
```

T2.1.2. Creo 4 ficheros dentro de '$HADOOP_HOME/etc/hadoop':

- dfs.include - dentro pongo los 4 datanodes
- dfs.exclude - dentro no pongo nada
- yarn.include - dentro pongo los 4 datanodes
- yarn.exclude - dentro no pongo nada

Cuando se abra, simplemente escribo:

datanode1
datanode2
datanode3
datanode4

```
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.include
hdadmin@namenode:~/hadoop/etc/hadoop$ cat dfs.include → para ver lo que hay dentro
datanode1
datanode2
datanode3
datanode4
→ Perfecto

hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.exclude → No escribo nada
hdadmin@namenode:~/hadoop/etc/hadoop$ cat dfs.exclude → No devuelve nada → perfecto
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.include
hdadmin@namenode:~/hadoop/etc/hadoop$ cat yarn.include
datanode1
datanode2
datanode3
datanode4
→ Perfecto

hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$ cat yarn.exclude → Perfecto
```

Comprobación extra de que los archivos están

```
hdadmin@namenode:~/hadoop/etc/hadoop$ ls
capacity-scheduler.xml            kms-log4j.properties
configuration.xsl                 kms-site.xml
container-executor.cfg            log4j.properties
core-site.xml                     mapred-env.cmd
dfs.exclude                       mapred-env.sh
dfs.include                       mapred-queues.xml.template
hadoop-env.cmd                    mapred-site.xml
hadoop-env.sh                     shellprofile.d
hadoop-metrics2.properties        ssl-client.xml.example
hadoop-policy.xml                 ssl-server.xml.example
hadoop-user-functions.sh.example  user_ec_policies.xml.template
hdfs-rbf-site.xml                 workers
hdfs-site.xml                     yarn-env.cmd
httpfs-env.sh                     yarn-env.sh
httpfs-log4j.properties           yarn.exclude
httpfs-site.xml                   yarn.include
kms-acls.xml                      yarnservice-log4j.properties
kms-env.sh                        yarn-site.xml
```

T2.1.3. Añadir propiedades a los ficheros 'hugs-site.xml' y 'yarn-site.xml'

```
hdadmin@namenode:~/hadoop/etc/hadoop$ nano hdfs-site.xml   (1)
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn-site.xml   (2)
```

(1)

```
  <property>
    <!-- Dirección y puerto del interfaz web del namenode -->
    <name>dfs.namenode.http-address</name>
    <value>namenode:9870</value>
    <final>true</final>
  </property>
```

ya están antes

```
<!-- Lista de hosts que pueden actuar como DataNodes -->
<property>
  <name>dfs.hosts</name>
  <value>/opt/bd/hadoop/etc/hadoop/dfs.include</value>
</property>

<!-- Lista de hosts que NO pueden actuar como DataNodes -->
<property>
  <name>dfs.hosts.exclude</name>
  <value>/opt/bd/hadoop/etc/hadoop/dfs.exclude</value>
</property>
</configuration>
```

Lo que se añade al final

Explicación: poner '$HADOOP_HOME/etc/hadoop' es otra forma de poner la ruta: '/opt/bd/hadoop/etc/hadoop'

(2)

```
<!-- Habilitar la publicación de métricas del sistema en el Timeline Server -->
<property>
    <name>yarn.system-metrics-publisher.enabled</name>
    <value>true</value>
    <final>true</final>
</property>
<!-- Lista de hosts que pueden actuar como NodeManagers -->
<property>
  <name>yarn.resourcemanager.nodes.include-path</name>
  <value>/opt/bd/hadoop/etc/hadoop/yarn.include</value>
</property>

<!-- Lista de hosts que NO pueden actuar como NodeManagers -->
<property>
  <name>yarn.resourcemanager.nodes.exclude-path</name>
  <value>/opt/bd/hadoop/etc/hadoop/yarn.exclude</value>
</property>
</configuration>
```

T2.1.4. Reinicio los demonios del NameNode y ResourceManager

```
hdadmin@namenode:~$ hdfs --daemon start namenode
hdadmin@namenode:~$ yarn --daemon start resourcemanager
hdadmin@namenode:~$ jps   -> Comprobar
660 Jps
634 ResourceManager   -> Perfecto
526 NameNode
```

T2.1.5. Compruebo en los ficheros de log del NameNode y del ResourceManager que se han incluido los nodos.

```
hdadmin@namenode:~$ cd $HADOOP_HOME/logs
hdadmin@namenode:~/hadoop/logs$ ls
hadoop-hdadmin-namenode-namenode.log
hadoop-hdadmin-namenode-namenode.out
hadoop-hdadmin-namenode-namenode.out.1
hadoop-hdadmin-namenode-namenode.out.2
hadoop-hdadmin-namenode-namenode.out.3
hadoop-hdadmin-namenode-namenode.out.4
hadoop-hdadmin-namenode-namenode.out.5
hadoop-hdadmin-resourcemanager-namenode.log
hadoop-hdadmin-resourcemanager-namenode.out
hadoop-hdadmin-resourcemanager-namenode.out.1
hadoop-hdadmin-resourcemanager-namenode.out.2
hadoop-hdadmin-resourcemanager-namenode.out.3
hadoop-hdadmin-resourcemanager-namenode.out.4
hadoop-hdadmin-resourcemanager-namenode.out.5
SecurityAuth-hdadmin.audit
```

Me interesan estos dos

Para el NameNode

```
hdadmin@namenode:~/hadoop/logs$ cat hadoop-hdadmin-namenode-namenode.log
```

Respuesta:

```
2024-12-08 13:39:08,936 INFO org.apache.hadoop.net.NetworkTopology: Removing a node: /default-rack/172.18.0.5:9866
2024-12-08 13:39:08,937 INFO org.apache.hadoop.net.NetworkTopology: Adding a new node: /default-rack/172.18.0.5:9866
2024-12-08 13:39:08,938 INFO org.apache.hadoop.hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(172.18.0.4:9866, datanodeUuid=e776500d-a1da-4519-9f3f-431ecacb4950, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-2215e70a-37ac-42d0-a304-1a65cdfc2830;nsid=625498134;c=1731058299332) storage e776500d-a1da-4519-9f3f-431ecacb4950
2024-12-08 13:39:08,938 INFO org.apache.hadoop.net.NetworkTopology: Removing a node: /default-rack/172.18.0.4:9866
2024-12-08 13:39:08,939 INFO org.apache.hadoop.net.NetworkTopology: Adding a new node: /default-rack/172.18.0.4:9866
2024-12-08 13:39:08,940 INFO org.apache.hadoop.hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(172.18.0.3:9866, datanodeUuid=17096ff4-1594-41f6-9dba-dbbb835d94bc, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-2215e70a-37ac-42d0-a304-1a65cdfc2830;nsid=625498134;c=1731058299332) storage 17096ff4-1594-41f6-9dba-dbbb835d94bc
2024-12-08 13:39:08,940 INFO org.apache.hadoop.net.NetworkTopology: Removing a node: /default-rack/172.18.0.3:9866
2024-12-08 13:39:08,941 INFO org.apache.hadoop.net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:9866
2024-12-08 13:39:08,943 INFO org.apache.hadoop.hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(172.18.0.6:9866, datanodeUuid=7e103f42-91ea-4262-a46e-ff4eeaea6445, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-2215e70a-37ac-42d0-a304-1a65cdfc2830;nsid=625498134;c=1731058299332) storage 7e103f42-91ea-4262-a46e-ff4eeaea6445
2024-12-08 13:39:08,943 INFO org.apache.hadoop.net.NetworkTopology: Removing a node: /default-rack/172.18.0.6:9866
2024-12-08 13:39:08,943 INFO org.apache.hadoop.net.NetworkTopology: Adding a new node: /default-rack/172.18.0.6:9866
2024-12-08 13:39:11,931 INFO org.apache.hadoop.hdfs.server.blockmanagement.DatanodeDescriptor: Adding new storage ID DS-ab317081-645b-4096-9e68-df24e9c684b8 for DN 172.18.0.6:9866
2024-12-08 13:39:11,931 INFO org.apache.hadoop.hdfs.server.blockmanagement.DatanodeDescriptor: Adding new storage ID DS-ad3abc70-0898-4e2f-83c5-73cd077bed12 for DN 172.18.0.5:9866
2024-12-08 13:39:11,932 INFO org.apache.hadoop.hdfs.server.blockmanagement.DatanodeDescriptor: Adding new storage ID DS-9224175b-e9a9-4fd7-92a1-eef97be9416d for DN 172.18.0.4:9866
2024-12-08 13:39:11,932 INFO org.apache.hadoop.hdfs.server.blockmanagement.DatanodeDescriptor: Adding new storage ID DS-056fa5ae-96ab-40b5-bc52-5af4cb7fab1e for DN 172.18.0.3:9866
```

(Más respuesta enmedio…)

```
2024-12-08 13:39:12,056 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Total number of blocks            = 49
2024-12-08 13:39:12,056 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Number of invalid blocks          = 0
2024-12-08 13:39:12,056 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Number of under-replicated blocks = 0
2024-12-08 13:39:12,056 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Number of  over-replicated blocks = 0
2024-12-08 13:39:12,056 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Number of blocks being written    = 0
2024-12-08 13:39:12,057 INFO org.apache.hadoop.hdfs.StateChange: STATE* Replication Queue initialization scan for invalid, over- and under-replicated blocks completed in 13 msec
2024-12-08 13:39:12,057 INFO BlockStateChange: BLOCK* processReport 0x6d60051195014e0 with lease ID 0x226a7ee41109121: Processing first storage report for DS-9224175b-e9a9-4fd7-92a1-eef97be9416d from datanode DatanodeRegistration(172.18.0.4:9866, datanodeUuid=e776500d-a1da-4519-9f3f-431ecacb4950, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-2215e70a-37ac-42d0-a304-1a65cdfc2830;nsid=625498134;c=1731058299332)
2024-12-08 13:39:12,059 INFO BlockStateChange: BLOCK* processReport 0x6d60051195014e0 with lease ID 0x226a7ee41109121: from storage DS-9224175b-e9a9-4fd7-92a1-eef97be9416d node DatanodeRegistration(172.18.0.4:9866, datanodeUuid=e776500d-a1da-4519-9f3f-431ecacb4950, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-2215e70a-37ac-42d0-a304-1a65cdfc2830;nsid=625498134;c=1731058299332), blocks: 49, hasStaleStorage: false, processing time: 3 msecs, invalidatedBlocks: 0
2024-12-08 13:39:14,132 INFO org.apache.hadoop.hdfs.server.blockmanagement.BlockManager: Rescan of postponedMisreplicatedBlocks completed in 64 msecs. 0 blocks are left. 39 blocks were removed.
2024-12-08 13:39:32,048 INFO org.apache.hadoop.hdfs.StateChange: STATE* Safe mode ON, in safe mode extension.
The reported blocks 49 has reached the threshold 0,9990 of total blocks 49. The minimum number of live datanodes is not required. In safe mode extension. Safe mode will be turned off automatically in 9 seconds.
2024-12-08 13:39:42,050 INFO org.apache.hadoop.hdfs.StateChange: STATE* Safe mode is OFF
2024-12-08 13:39:42,050 INFO org.apache.hadoop.hdfs.StateChange: STATE* Leaving safe mode after 33 secs
2024-12-08 13:39:42,050 INFO org.apache.hadoop.hdfs.StateChange: STATE* Network topology has 1 racks and 4 datanodes
2024-12-08 13:39:42,050 INFO org.apache.hadoop.hdfs.StateChange: STATE* UnderReplicatedBlocks has 0 blocks
```

Para el Resource Manager

```
hdadmin@namenode:~/hadoop/logs$ cat hadoop-hdadmin-resourcemanager-namenode.log
```

Respuesta:

```
2024-12-08 13:39:21,869 INFO org.apache.hadoop.ipc.Server: IPC Server Responder: starting
2024-12-08 13:39:21,870 INFO org.apache.hadoop.ipc.Server: IPC Server listener on 8032: starting
2024-12-08 13:39:22,771 INFO org.apache.hadoop.yarn.server.webproxy.ProxyCA: Created Certificate for OU=YARN-24978702-7462-4ad6-9c79-fb6ca8e149be
2024-12-08 13:39:22,943 INFO org.apache.hadoop.yarn.server.resourcemanager.recovery.RMStateStore: Storing CA Certificate and Private Key
2024-12-08 13:39:22,943 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceManager: Transitioned to active state
2024-12-08 13:39:35,849 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: Node not found resyncing datanode2:35445
2024-12-08 13:39:35,849 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: Node not found resyncing datanode1:36547
2024-12-08 13:39:35,849 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: Node not found resyncing datanode4:36587
2024-12-08 13:39:35,850 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: Node not found resyncing datanode3:34253
2024-12-08 13:39:36,034 INFO org.apache.hadoop.yarn.server.resourcemanager.rmnode.RMNodeImpl: datanode2:35445 Node Transitioned from NEW to RUNNING
2024-12-08 13:39:36,034 INFO org.apache.hadoop.yarn.server.resourcemanager.rmnode.RMNodeImpl: datanode1:36547 Node Transitioned from NEW to RUNNING
2024-12-08 13:39:36,034 INFO org.apache.hadoop.yarn.server.resourcemanager.rmnode.RMNodeImpl: datanode4:36587 Node Transitioned from NEW to RUNNING
2024-12-08 13:39:36,035 INFO org.apache.hadoop.yarn.server.resourcemanager.rmnode.RMNodeImpl: datanode3:34253 Node Transitioned from NEW to RUNNING
2024-12-08 13:39:36,036 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: NodeManager from node datanode1(cmPort: 36547 httpPort: 8042) registered with capability: <memory:5137, vCores:4>, assigned nodeId datanode1:36547
2024-12-08 13:39:36,036 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: NodeManager from node datanode4(cmPort: 36587 httpPort: 8042) registered with capability: <memory:5137, vCores:4>, assigned nodeId datanode4:36587
2024-12-08 13:39:36,036 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: NodeManager from node datanode2(cmPort: 35445 httpPort: 8042) registered with capability: <memory:5137, vCores:4>, assigned nodeId datanode2:35445
2024-12-08 13:39:36,036 INFO org.apache.hadoop.yarn.server.resourcemanager.ResourceTrackerService: NodeManager from node datanode3(cmPort: 34253 httpPort: 8042) registered with capability: <memory:5137, vCores:4>, assigned nodeId datanode3:34253
2024-12-08 13:39:36,078 INFO org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler: Added node datanode2:35445 clusterResource: <memory:5137, vCores:4>
2024-12-08 13:39:36,130 INFO org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler: Added node datanode1:36547 clusterResource: <memory:10274, vCores:8>
2024-12-08 13:39:36,131 INFO org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler: Added node datanode4:36587 clusterResource: <memory:15411, vCores:12>
2024-12-08 13:39:36,133 INFO org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.CapacityScheduler: Added node datanode3:34253 clusterResource: <memory:20548, vCores:16>
2024-12-08 13:49:21,535 INFO org.apache.hadoop.yarn.server.resourcemanager.scheduler.AbstractYarnScheduler: Release request cache is cleaned up
```

T2.2. Añadir un datanode/nodemanager

T2.2.1. En el NameNode, añadir 'datanode5' al archivo 'yarn.include' (solo a ese de momento)

```
PS D:\Alex\Universidad\MasterBigData> docker container start namenode datanode1
datanode2 datanode3 datanode4
namenode
datanode1
datanode2
datanode3
datanode4
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.include
hdadmin@namenode:~/hadoop/etc/hadoop$
```

```
 GNU nano 6.2                    yarn.include *
datanode1
datanode2
datanode3
datanode4
datanode5
```

T2.2.2. Actualizo el ResourceManager con el nuevo NodeManager

```
hdadmin@namenode:~$ yarn rmadmin -refreshNodes
2024-12-08 15:57:56,300 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at resourcemanager/172.18.0.2:8033
```

→ Perfecto

T2.2.3. Nuevo contenedor para hacer de DataNode -> datanode5

```
PS D:\Alex\Universidad\MasterBigData> docker container run -d --name datanode5 -
-network=hadoop-cluster --hostname datanode5 --cpus=1 --memory=3072m --expose=80
00-10000 --expose=50000-50200 datanode-image /inicio.sh
f71a43ed90d7f78463a4289d57bb90c105aec80e394fb6bd9abdad647181e715
```

T2.2.4. Compruebo, dentro de NameNode como hdadmin que el nuevo contenedor de ha añadido al yarn pero no al hdfs

```
PS D:\Alex\Universidad\MasterBigData> docker exec container -ti namenode bash
Error response from daemon: No such container: container
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
```

```
hdadmin@namenode:~$ hdfs dfsadmin -report
Configured Capacity: 4324404707328 (3.93 TB)
Present Capacity: 4065395834880 (3.70 TB)
DFS Remaining: 4063367430144 (3.70 TB)
DFS Used: 2028404736 (1.89 GB)
DFS Used%: 0.05%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
```

```
-------------------------------------------------
Live datanodes (4):

Name: 172.18.0.3:9866 (datanode1.hadoop-cluster)
Hostname: datanode1
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9590845440 (8.93 GB)
DFS Remaining: 1015841857536 (946.08 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:03:40 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.4:9866 (datanode2.hadoop-cluster)
Hostname: datanode2
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9590845440 (8.93 GB)
DFS Remaining: 1015841857536 (946.08 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:03:40 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.5:9866 (datanode3.hadoop-cluster)
Hostname: datanode3
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 549797888 (524.33 MB)
Non DFS Used: 9717166080 (9.05 GB)
DFS Remaining: 1015841857536 (946.08 GB)
DFS Used%: 0.05%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:03:40 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 39
```

```
Name: 172.18.0.6:9866 (datanode4.hadoop-cluster)
Hostname: datanode4
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 126369792 (120.52 MB)
Non DFS Used: 10140594176 (9.44 GB)
DFS Remaining: 1015841857536 (946.08 GB)
DFS Used%: 0.01%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:03:41 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 10
```

datanode 5 no añadido al HDFS

Para ver el YARN, el contenedor timelineserver tiene que estar ejecutándose

```
hdadmin@namenode:~$ exit
logout
root@namenode:/# exit
exit
PS D:\Alex\Universidad\MasterBigData> docker container start timelineserver
timelineserver
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
```

```
hdadmin@namenode:~$ yarn node -list
2024-12-08 16:11:10,477 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceM
anager at resourcemanager/172.18.0.2:8032
2024-12-08 16:11:10,882 INFO client.AHSProxy: Connecting to Application History server at timel
ineserver/172.18.0.8:10200
Total Nodes:5
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
 datanode3:39687                RUNNING    datanode3:8042                                  0
 datanode2:40511                RUNNING    datanode2:8042                                  0
 datanode1:39127                RUNNING    datanode1:8042                                  0
 datanode4:33267                RUNNING    datanode4:8042                                  0
 datanode5:38187                RUNNING    datanode5:8042                                  0
```

Perfecto

T2.2.5. Ahora, añadir 'datanode5' al archivo 'dfs.include'

```
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.include
```

```
 GNU nano 6.2                                  dfs.include *
datanode1
datanode2
datanode3
datanode4
datanode5
```

T2.2.6. Actualizo el NameNode con el nuevo DataNode ejecutado

```
hdadmin@namenode:~$ hdfs dfsadmin -refreshNodes
Refresh nodes successful
```

→ perfecto

T2.2.7. Compruebo que ahora el nuevo contenedor está incluido en el HDFS

```
hdadmin@namenode:~$ hdfs dfsadmin -report
Configured Capacity: 5405505884160 (4.92 TB)
Present Capacity: 5081232805888 (4.62 TB)
DFS Remaining: 5079204372480 (4.62 TB)
DFS Used: 2028433408 (1.89 GB)
DFS Used%: 0.04%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
```

```
Live datanodes (5):

Name: 172.18.0.3:9866 (datanode1.hadoop-cluster)
Hostname: datanode1
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9591828480 (8.93 GB)
DFS Remaining: 1015840874496 (946.08 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:22:19 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.4:9866 (datanode2.hadoop-cluster)
Hostname: datanode2
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9591828480 (8.93 GB)
DFS Remaining: 1015840874496 (946.08 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:22:20 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.5:9866 (datanode3.hadoop-cluster)
Hostname: datanode3
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 549797888 (524.33 MB)
Non DFS Used: 9718149120 (9.05 GB)
DFS Remaining: 1015840874496 (946.08 GB)
DFS Used%: 0.05%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:22:19 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 39
```

```
Name: 172.18.0.6:9866 (datanode4.hadoop-cluster)
Hostname: datanode4
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 126369792 (120.52 MB)
Non DFS Used: 10141577216 (9.45 GB)
DFS Remaining: 1015840874496 (946.08 GB)
DFS Used%: 0.01%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:22:20 CET 2024
Last Block Report: Sun Dec 08 15:53:20 CET 2024
Num of Blocks: 10
```

```
Name: 172.18.0.7:9866 (datanode5.hadoop-cluster)
Hostname: datanode5
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 10267918336 (9.56 GB)
DFS Remaining: 1015840874496 (946.08 GB)
DFS Used%: 0.00%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:22:19 CET 2024
Last Block Report: Sun Dec 08 16:17:19 CET 2024
Num of Blocks: 0
```

Eu la interfață web de HDFS

Overview 'namenode:9000' (✔active)

Started:	Sun Dec 08 15:53:12 +0100 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 10:22:00 +0200 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-2215e70a-37ac-42d0-a304-1a65cdfc2830
Block Pool ID:	BP-1656694194-172.18.0.2-1731058299332

Summary

Security is off.

Safemode is off.

71 files and directories, 49 blocks (49 replicated blocks, 0 erasure coded block groups) = 120 total filesystem object(s).

Heap Memory used 40.48 MB of 77.84 MB Heap Memory. Max Heap Memory is 742.44 MB.

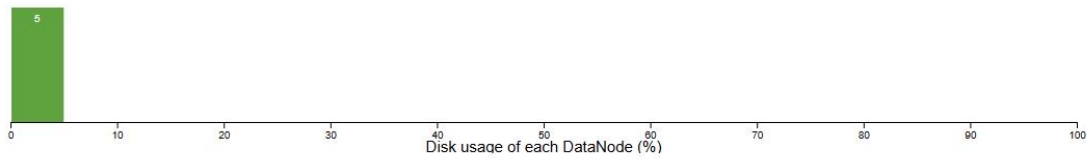
Non Heap Memory used 62.84 MB of 65.19 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	4.92 TB
Configured Remote Capacity:	0 B
DFS Used:	1.89 GB (0.04%)
Non DFS Used:	45.92 GB
DFS Remaining:	4.62 TB (93.96%)
Block Pool Used:	1.89 GB (0.04%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.05% / 0.06% / 0.03%
Live Nodes	5 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)

Datanode Information

✔ In service ❗ Down ⊘ Decommissioning ⊘ Decommissioned ⊘ Decommissioned & dead
Entering Maintenance In Maintenance In Maintenance & dead

Datanode usage histogram



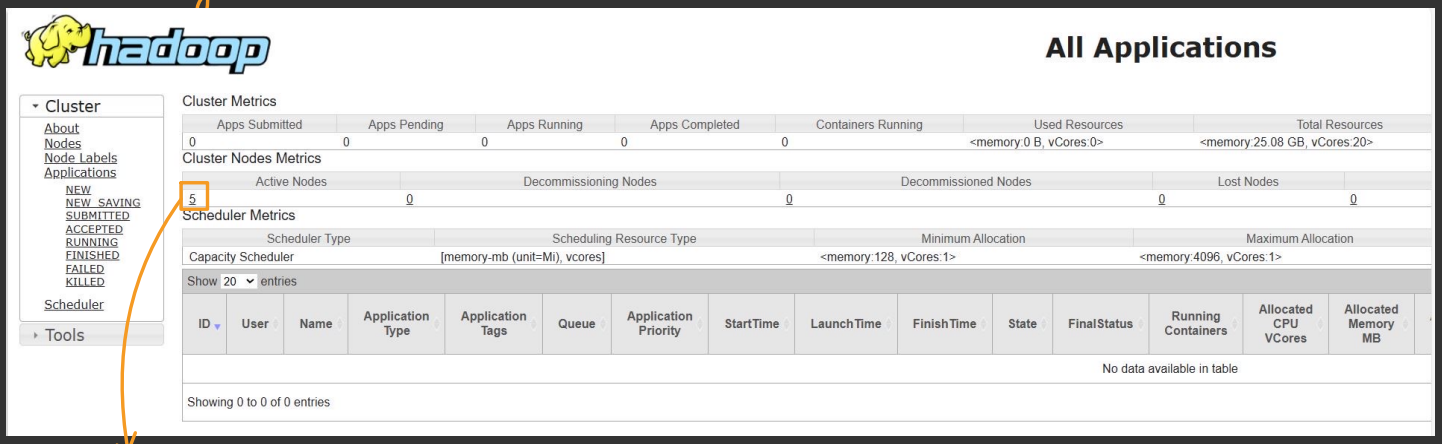
In operation

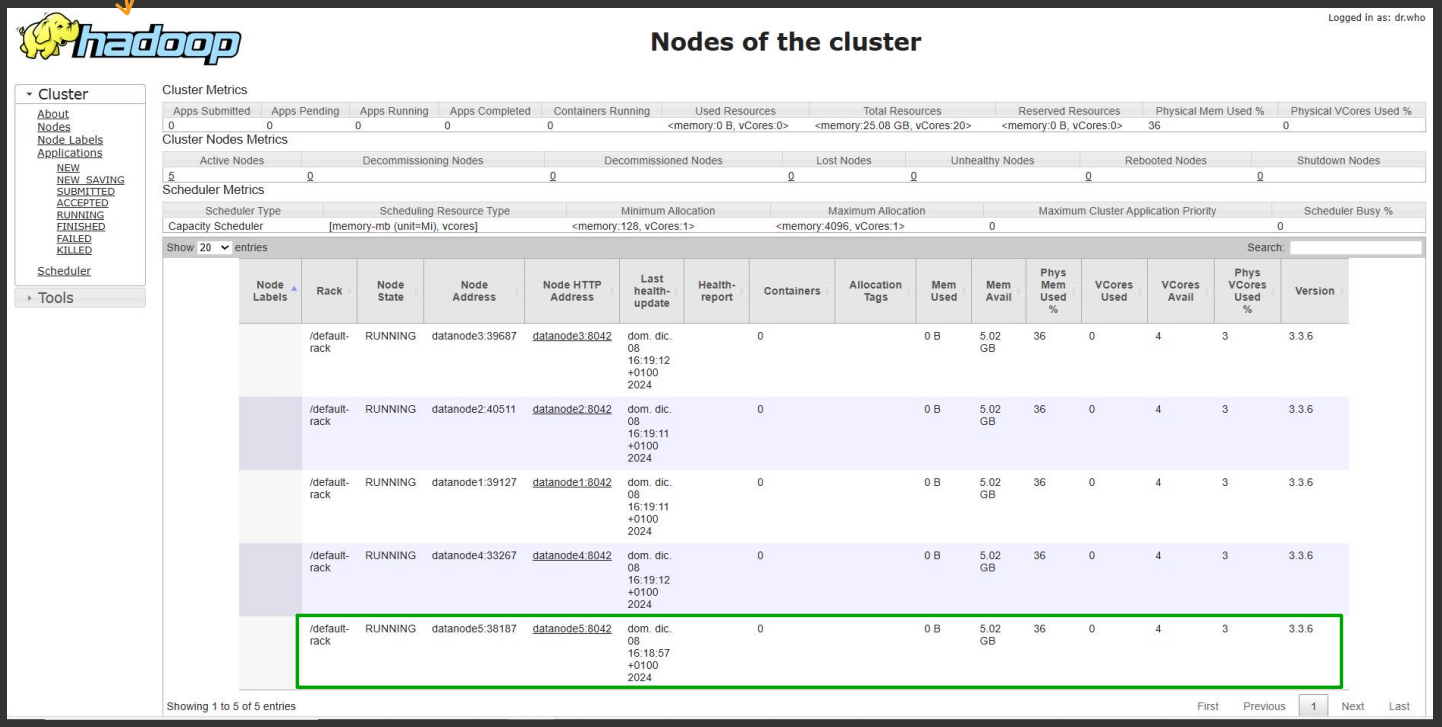
DataNode State: All Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✔/default-rack/datanode2:9866 (172.18.0.4:9866)	http://datanode2:9864	2s	95m	644.8 MB	8.93 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6
✔/default-rack/datanode5:9866 (172.18.0.7:9866)	http://datanode5:9864	0s	71m	44 KB	9.56 GB	1006.85 GB	0	44 KB (0%)	3.3.6
✔/default-rack/datanode3:9866 (172.18.0.5:9866)	http://datanode3:9864	2s	95m	524.33 MB	9.05 GB	1006.85 GB	39	524.33 MB (0.05%)	3.3.6
✔/default-rack/datanode1:9866 (172.18.0.3:9866)	http://datanode1:9864	2s	95m	644.8 MB	8.93 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6
✔/default-rack/datanode4:9866 (172.18.0.6:9866)	http://datanode4:9864	2s	95m	120.52 MB	9.45 GB	1006.85 GB	10	120.52 MB (0.01%)	3.3.6

Showing 1 to 5 of 5 entries Previous 1 Next

En la interfaz web de YARN





T2.2.8. Balanceo el cluster

```
hdadmin@namenode:~$ hdfs balancer
2024-12-08 16:26:03,039 INFO balancer.Balancer: namenodes  = [hdfs://namenode:9000]
2024-12-08 16:26:03,042 INFO balancer.Balancer: parameters = Balancer.BalancerParameters [Bal
ancingPolicy.Node, threshold = 10.0, max idle iteration = 5, #excluded nodes = 0, #included n
odes = 0, #source nodes = 0, #blockpools = 0, run during upgrade = false]
2024-12-08 16:26:03,042 INFO balancer.Balancer: included nodes = []
2024-12-08 16:26:03,043 INFO balancer.Balancer: excluded nodes = []
2024-12-08 16:26:03,043 INFO balancer.Balancer: source nodes = []
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Mov
ed  NameNode
2024-12-08 16:26:03,098 INFO balancer.NameNodeConnector: getBlocks calls for hdfs://namenode:
9000 will be rate-limited to 20 per second
2024-12-08 16:26:05,069 INFO balancer.Balancer: dfs.namenode.get-blocks.max-qps = 20 (default
=20)
2024-12-08 16:26:05,069 INFO balancer.Balancer: dfs.balancer.movedWinWidth = 5400000 (default=5400000)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.moverThreads = 1000 (default=1000)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.dispatcherThreads = 200 (default=200)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.getBlocks.size = 2147483648 (default=2147483648)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.getBlocks.min-block-size = 10485760 (default=10485760)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.datanode.balance.max.concurrent.moves = 100 (default=100)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.datanode.balance.bandwidthPerSec = 104857600 (default=104857600)
2024-12-08 16:26:05,077 INFO balancer.Balancer: dfs.balancer.max-size-to-move = 10737418240 (default=10737418240)
2024-12-08 16:26:05,078 INFO balancer.Balancer: dfs.blocksize = 67108864 (default=134217728)
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.6:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.5:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.7:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.4:9866
2024-12-08 16:26:05,112 INFO balancer.Balancer: 0 over-utilized: []
2024-12-08 16:26:05,112 INFO balancer.Balancer: 0 underutilized: []
8 dic. 2024 16:26:05               0                 0 B                 0 B                 0 B                  0  hdfs://namenode:9000
The cluster is balanced. Exiting...
8 dic. 2024 16:26:05     Balancing took 2.989 seconds
```

Nota. Creo que el balanceo del cluster no se ha hecho bien, porque dice que no se ha movido datos y tras comprobar los datanodes en el hdfs dice que el 'datanode5' tiene 0 bloques.

```
hdadmin@namenode:~$ hdfs balancer
2024-12-08 16:26:03,039 INFO balancer.Balancer: namenodes  = [hdfs://namenode:9000]
2024-12-08 16:26:03,042 INFO balancer.Balancer: parameters = Balancer.BalancerParameters [Bal
ancingPolicy.Node, threshold = 10.0, max idle iteration = 5, #excluded nodes = 0, #included n
odes = 0, #source nodes = 0, #blockpools = 0, run during upgrade = false]
2024-12-08 16:26:03,042 INFO balancer.Balancer: included nodes = []
2024-12-08 16:26:03,043 INFO balancer.Balancer: excluded nodes = []
2024-12-08 16:26:03,043 INFO balancer.Balancer: source nodes = []
Time Stamp               Iteration#  Bytes Already Moved  Bytes Left To Move  Bytes Being Mov
ed  NameNode
2024-12-08 16:26:03,098 INFO balancer.NameNodeConnector: getBlocks calls for hdfs://namenode:
9000 will be rate-limited to 20 per second
2024-12-08 16:26:05,069 INFO balancer.Balancer: dfs.namenode.get-blocks.max-qps = 20 (default
=20)
2024-12-08 16:26:05,069 INFO balancer.Balancer: dfs.balancer.movedWinWidth = 5400000 (default=5400000)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.moverThreads = 1000 (default=1000)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.dispatcherThreads = 200 (default=200)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.getBlocks.size = 2147483648 (default=2147483648)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.balancer.getBlocks.min-block-size = 10485760 (default=10485760)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.datanode.balance.max.concurrent.moves = 100 (default=100)
2024-12-08 16:26:05,070 INFO balancer.Balancer: dfs.datanode.balance.bandwidthPerSec = 104857600 (default=104857600)
2024-12-08 16:26:05,077 INFO balancer.Balancer: dfs.balancer.max-size-to-move = 10737418240 (default=10737418240)
2024-12-08 16:26:05,078 INFO balancer.Balancer: dfs.blocksize = 67108864 (default=134217728)
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.6:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.5:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.7:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.3:9866
2024-12-08 16:26:05,104 INFO net.NetworkTopology: Adding a new node: /default-rack/172.18.0.4:9866
2024-12-08 16:26:05,112 INFO balancer.Balancer: 0 over-utilized: []
2024-12-08 16:26:05,112 INFO balancer.Balancer: 0 underutilized: []
8 dic. 2024 16:26:05               0                0 B             0 B             0 B             0 hdfs://namenode:9000
The cluster is balanced. Exiting...
8 dic. 2024 16:26:05     Balancing took 2.989 seconds
```

No se han movido datos

Ejecutando tras el balanceo:

```
hdadmin@namenode:~$ hdfs dfsadmin -report
```

```
Name: 172.18.0.7:9866 (datanode5.hadoop-cluster)
Hostname: datanode5
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 45056 (44 KB)
Non DFS Used: 10267959296 (9.56 GB)
DFS Remaining: 1015840817152 (946.08 GB)
DFS Used%: 0.00%
DFS Remaining%: 93.96%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Sun Dec 08 16:51:22 CET 2024
Last Block Report: Sun Dec 08 16:17:19 CET 2024
Num of Blocks: 0
```

→ Tras el balanceo no hay bloques en el datanode 5.

Le he mandado un correo al profesor y dice que es normal es que a veces no se balancee nada porque hay pocos datos

Tarea 3

T3. Retirar un DataNode/NodeManager. Se va a eliminar el datanode4

T3.1. Pongo el nombre del nodo o nodos que queremos retirar en los ficheros 'dfs.exclude' y 'yarn.exclude'. Después refresco los nodos

```
PS D:\Alex\Universidad\MasterBigData> docker container start namenode datanode1 datanode2 datano
de3 datanode4 datanode5 backupnode timelineserver
namenode
datanode1
datanode2
datanode3
datanode4
datanode5
backupnode
timelineserver
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.exclude
```

```
 GNU nano 6.2                                  dfs.exclude *
datanode4
```

```
hdadmin@namenode:~/hadoop/etc/hadoop$ cat dfs.exclude
datanode4
hdadmin@namenode:~/hadoop/etc/hadoop$ cat yarn.exclude
datanode4
```

Comprobar → Perfecto → Perfecto

Refresco nodos

```
hdadmin@namenode:~$ hdfs dfsadmin -refreshNodes
Refresh nodes successful
hdadmin@namenode:~$ yarn rmadmin -refreshNodes
2024-12-09 18:05:12,516 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMa
nager at resourcemanager/172.18.0.2:8033
```

T3.2. Compruebo que el nodo espxcluido (datanode4) está *Decomissioned* en el HDFS y no aparece en el YARN de dos formas:

- Aplicando los comandos 'hdfs dfsadmin -report' y 'yarn node -list'.
- Viendo en las interfaces web de HDFS y YARN.

```
hdadmin@namenode:~$ hdfs dfsadmin -report
Configured Capacity: 432440407328 (3.93 TB)
Present Capacity: 406487941367 (3.70 TB)
DFS Remaining: 4062851055616 (3.70 TB)
DFS Used: 2028355751 (1.89 GB)
DFS Used%: 0.05%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
```

```
Live datanodes (5):

Name: 172.18.0.3:9866 (datanode1.hadoop-cluster)
Hostname: datanode1
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9719939072 (9.05 GB)
DFS Remaining: 1015712763904 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:05:58 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.4:9866 (datanode2.hadoop-cluster)
Hostname: datanode2
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9719939072 (9.05 GB)
DFS Remaining: 1015712763904 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:05:58 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 49
```

```
Name: 172.18.0.5:9866 (datanode3.hadoop-cluster)
Hostname: datanode3
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 613171965 (584.77 MB)
Non DFS Used: 9782885635 (9.11 GB)
DFS Remaining: 1015712763904 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:05:58 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 46
```

```
Name: 172.18.0.6:9866 (datanode4.hadoop-cluster)
Hostname: datanode4
Decommission Status : Decommissioned
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 126369792 (120.52 MB)
Non DFS Used: 10269687808 (9.56 GB)
DFS Remaining: 1015712763904 (945.96 GB)
DFS Used%: 0.01%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:05:58 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 10
```

```
Name: 172.18.0.7:9866 (datanode5.hadoop-cluster)
Hostname: datanode5
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 62946730 (60.03 MB)
Non DFS Used: 10333110870 (9.62 GB)
DFS Remaining: 1015712763904 (945.96 GB)
DFS Used%: 0.01%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:05:58 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 3
```

```
hdadmin@namenode:~$ yarn node -list
2024-12-09 18:08:29,999 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMa
nager at resourcemanager/172.18.0.2:8032
2024-12-09 18:08:30,484 INFO client.AHSProxy: Connecting to Application History server at timeli
neserver/172.18.0.9:10200
Total Nodes:4
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
 datanode3:34727                RUNNING    datanode3:8042                                  0
 datanode2:40737                RUNNING    datanode2:8042                                  0
 datanode1:44069                RUNNING    datanode1:8042                                  0
 datanode5:45679                RUNNING    datanode5:8042                                  0
```

No aparece datanode4: perfecto

Interfaz web de HDFS

Overview 'namenode:9000' (✔active)

Started:	Mon Dec 09 18:01:35 +0100 2024
Version:	3.3.6, r1be78238728da9266a4f88195058f08fd012bf9c
Compiled:	Sun Jun 18 10:22:00 +0200 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-2215e70a-37ac-42d0-a304-1a65cdfc2830
Block Pool ID:	BP-1656694194-172.18.0.2-1731058299332

Summary

Security is off.

Safemode is off.

71 files and directories, 49 blocks (49 replicated blocks, 0 erasure coded block groups) = 120 total filesystem object(s).

Heap Memory used 42.63 MB of 77.84 MB Heap Memory. Max Heap Memory is 742.44 MB.

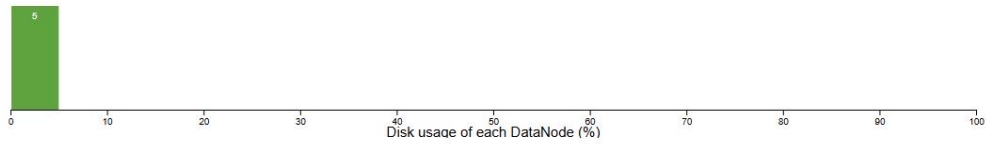
Non Heap Memory used 56.16 MB of 59.31 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	3.93 TB
Configured Remote Capacity:	0 B
DFS Used:	1.89 GB (0.05%)
Non DFS Used:	36.84 GB
DFS Remaining:	3.7 TB (93.95%)
Block Pool Used:	1.89 GB (0.05%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.06% / 0.06% / 0.02%
Live Nodes	5 (Decommissioned: 1, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)

Datanode Information

✔ In service | Down | Decommissioning | Decommissioned | Decommissioned & dead | Entering Maintenance | In Maintenance | In Maintenance & dead

Datanode usage histogram



In operation

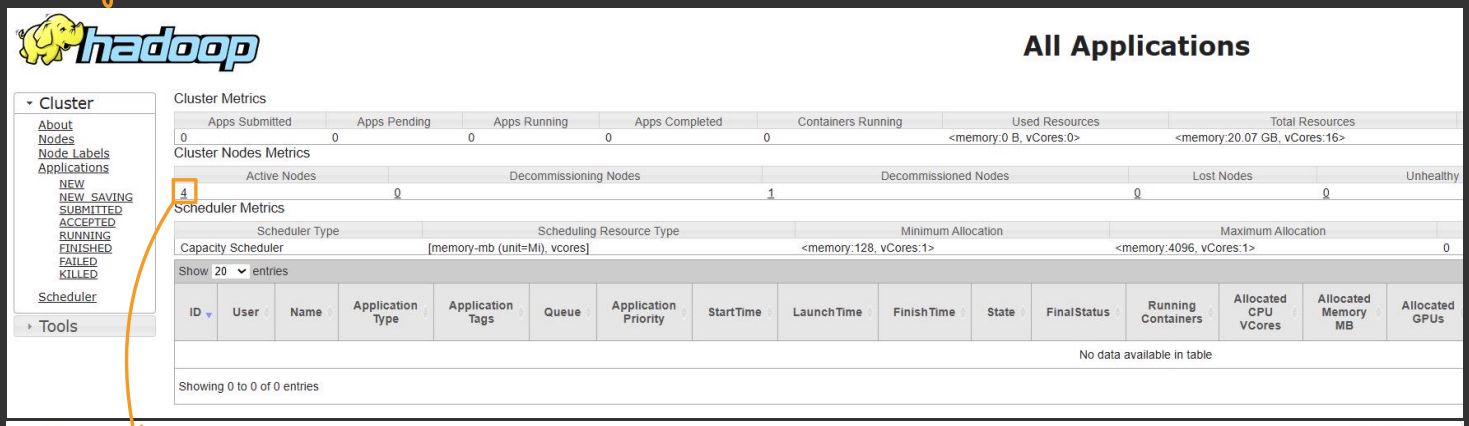
DataNode State: All | Show 25 entries | Search:

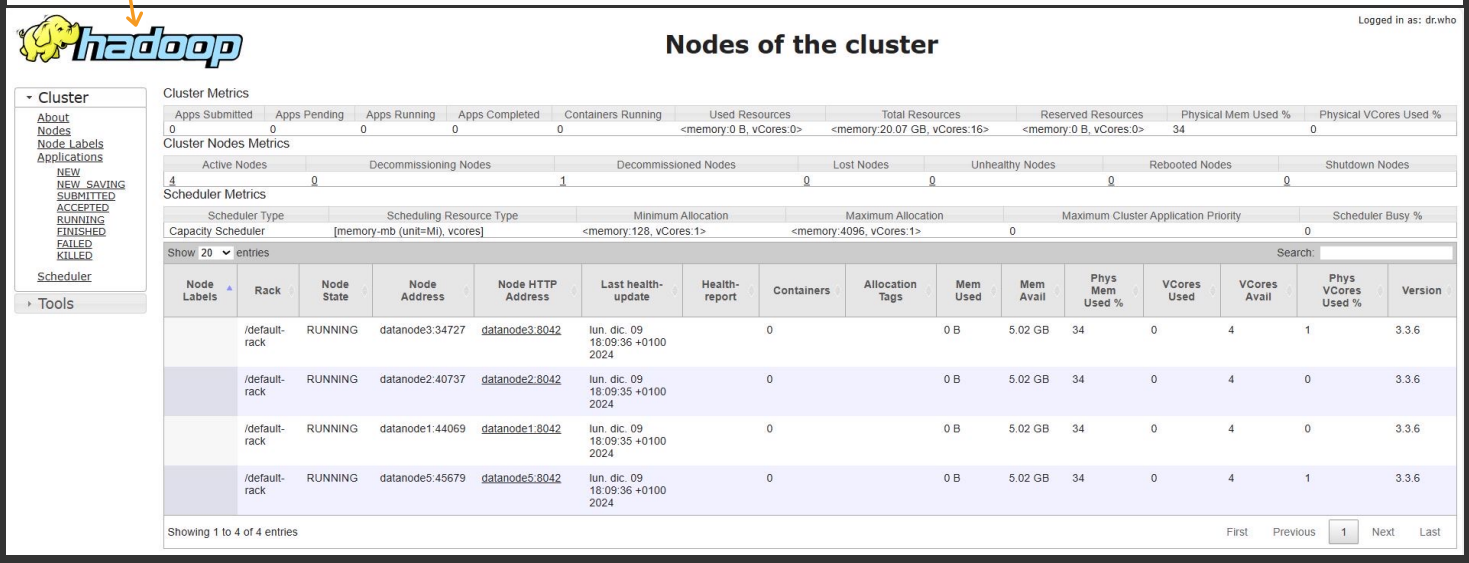
Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✔/default-rack/datanode2:9866 (172.18.0.4:9866)	http://datanode2:9864	1s	8m	644.8 MB	9.05 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6
✔/default-rack/datanode5:9866 (172.18.0.7:9866)	http://datanode5:9864	1s	8m	60.03 MB	9.62 GB	1006.85 GB	3	60.03 MB (0.01%)	3.3.6
✔/default-rack/datanode3:9866 (172.18.0.5:9866)	http://datanode3:9864	1s	8m	584.77 MB	9.11 GB	1006.85 GB	46	584.77 MB (0.06%)	3.3.6
✔/default-rack/datanode1:9866 (172.18.0.3:9866)	http://datanode1:9864	1s	8m	644.8 MB	9.05 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6
⊘/default-rack/datanode4:9866 (172.18.0.6:9866)	http://datanode4:9864	1s	8m	120.52 MB	9.56 GB	1006.85 GB	10	120.52 MB (0.01%)	3.3.6

Showing 1 to 5 of 5 entries | Previous 1 Next

datanode4: Decomissionesd -> perfecto

Interfaz web de YARN





No aparece datanode4 -> perfecto

T3.3.

- Paro el datandoe4.
- Elimino el nodo ‘datanode4’ de los ficheros include y exclude.
- Refresco los nodos.
- Compruebo con los comandos ‘hdfs dfsadmin -report’ y ‘yarn node -list’ que ahora el datanode4 no aparece ni en el YARN ni en el HDFS.

```
PS D:\Alex\Universidad\MasterBigData> docker container stop datanode4
datanode4
PS D:\Alex\Universidad\MasterBigData> docker container exec -ti namenode bash
root@namenode:/# su - hdadmin
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.include
hdadmin@namenode:~/hadoop/etc/hadoop$ cat dfs.include
datanode1
datanode2
datanode3
datanode5
hdadmin@namenode:~/hadoop/etc/hadoop$ nano dfs.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$ cat dfs.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.include
hdadmin@namenode:~/hadoop/etc/hadoop$ cat yarn.include
datanode1
datanode2
datanode3
datanode5
hdadmin@namenode:~/hadoop/etc/hadoop$ nano yarn.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$ cat yarn.exclude
hdadmin@namenode:~/hadoop/etc/hadoop$
```

GNU nano 6.2
datanode1
datanode2
datanode3
datanode5

Comprobar

→ Perfecto

→ Ahora queda vacío

→ No hay salida → perfecto

GNU nano 6.2
datanode1
datanode2
datanode3
datanode5

→ Perfecto

→ debe quedar vacío

→ No hay salida → perfecto

Refresco nodos

```
hdadmin@namenode:~$ hdfs dfsadmin -refreshNodes
Refresh nodes successful
hdadmin@namenode:~$ yarn rmadmin -refreshNodes
2024-12-09 18:20:50,474 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMa
nager at resourcemanager/172.18.0.2:8033
```

Compruebo para HDFS

```
hdadmin@namenode:~$ hdfs dfsadmin -report
Configured Capacity: 4324404707328 (3.93 TB)
Present Capacity: 4064879624192 (3.70 TB)
DFS Remaining: 4062851203072 (3.70 TB)
DFS Used: 2028421120 (1.89 GB)
DFS Used%: 0.05%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
```

→

```
Live datanodes (4):
```

→ Ahora no hay datanode 4

```
Name: 172.18.0.3:9866 (datanode1.hadoop-cluster)
Hostname: datanode1
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9719902208 (9.05 GB)
DFS Remaining: 1015712800768 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:21:43 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 49
```

→

→

```
Name: 172.18.0.4:9866 (datanode2.hadoop-cluster)
Hostname: datanode2
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 676118528 (644.80 MB)
Non DFS Used: 9719902208 (9.05 GB)
DFS Remaining: 1015712800768 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:21:43 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 49
```

→

```
Name: 172.18.0.5:9866 (datanode3.hadoop-cluster)
Hostname: datanode3
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 613216256 (584.81 MB)
Non DFS Used: 9782804480 (9.11 GB)
DFS Remaining: 1015712800768 (945.96 GB)
DFS Used%: 0.06%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:21:43 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 46
```

→

```
Name: 172.18.0.7:9866 (datanode5.hadoop-cluster)
Hostname: datanode5
Decommission Status : Normal
Configured Capacity: 1081101176832 (1006.85 GB)
DFS Used: 62967808 (60.05 MB)
Non DFS Used: 10333052928 (9.62 GB)
DFS Remaining: 1015712800768 (945.96 GB)
DFS Used%: 0.01%
DFS Remaining%: 93.95%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Mon Dec 09 18:21:43 CET 2024
Last Block Report: Mon Dec 09 18:01:46 CET 2024
Num of Blocks: 3
```

Compruebo para el YARN

```
hdadmin@namenode:~$ yarn node -list
2024-12-09 18:23:59,068 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMa
nager at resourcemanager/172.18.0.2:8032
2024-12-09 18:23:59,403 INFO client.AHSProxy: Connecting to Application History server at timeli
neserver/172.18.0.9:10200
Total Nodes:4
         Node-Id             Node-State Node-Http-Address       Number-of-Running-Containers
 datanode3:34727                RUNNING    datanode3:8042                                  0
 datanode2:40737                RUNNING    datanode2:8042                                  0
 datanode1:44069                RUNNING    datanode1:8042                                  0
 datanode5:45679                RUNNING    datanode5:8042                                  0
```

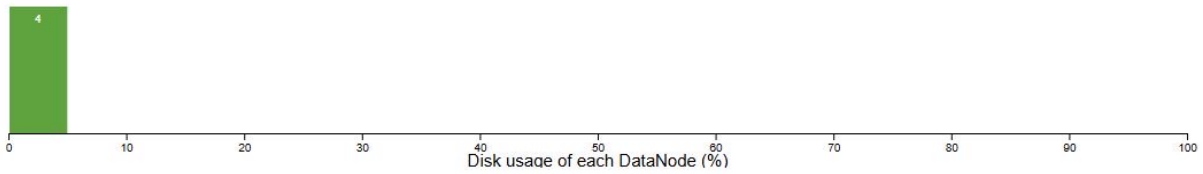
No aparece datanode4
(Igual que antes)

Interfaz web de HDFS

Datanode Information

In service | Down | Decommissioning | Decommissioned | Decommissioned & dead
Entering Maintenance | In Maintenance | In Maintenance & dead

Datanode usage histogram



In operation

DataNode State: All | Show 25 entries | Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
/default-rack/datanode2:9866 (172.18.0.4:9866)	http://datanode2:9864	2s	23m	644.8 MB	9.05 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6
/default-rack/datanode5:9866 (172.18.0.7:9866)	http://datanode5:9864	2s	23m	60.05 MB	9.62 GB	1006.85 GB	3	60.05 MB (0.01%)	3.3.6
/default-rack/datanode3:9866 (172.18.0.5:9866)	http://datanode3:9864	2s	23m	584.81 MB	9.11 GB	1006.85 GB	46	584.81 MB (0.06%)	3.3.6
/default-rack/datanode1:9866 (172.18.0.3:9866)	http://datanode1:9864	2s	23m	644.8 MB	9.05 GB	1006.85 GB	49	644.8 MB (0.06%)	3.3.6

Showing 1 to 4 of 4 entries

Previous 1 Next

Ya no aparece datanode4

Interfaz web de YARN

hadoop

Logged in as: dr.who

Nodes of the cluster

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:20.07 GB, vCores:16>	<memory:0 B, vCores:0>	34	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
4	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:128, vCores:1>	<memory:4096, vCores:1>	0	0

Show 20 entries | Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %	VCores Used	VCores Avail	Phys VCores Used %	Version
	/default-rack	RUNNING	datanode3:34727	datanode3:8042	lun. dic. 09 18:25:36 +0100 2024		0		0 B	5.02 GB	34	0	4	0	3.3.6
	/default-rack	RUNNING	datanode2:40737	datanode2:8042	lun. dic. 09 18:25:35 +0100 2024		0		0 B	5.02 GB	34	0	4	1	3.3.6
	/default-rack	RUNNING	datanode1:44069	datanode1:8042	lun. dic. 09 18:25:35 +0100 2024		0		0 B	5.02 GB	34	0	4	1	3.3.6
	/default-rack	RUNNING	datanode5:45679	datanode5:8042	lun. dic. 09 18:25:36 +0100 2024		0		0 B	5.02 GB	34	0	4	0	3.3.6

Showing 1 to 4 of 4 entries

First Previous 1 Next Last

No aparece datanode4 (Igual que antes)

Aviso: se podrá ver como el datanode4 no lo he eliminado.
No es que no haya que eliminarlo, es que simplemente como con lo que se ha hecho aquí ya no se usa, me da pánico eliminar un datanode por si hace falta usarlo mas adelante.
Yo creo que se puede eliminar sin problemas en realidad.

Tarea 4

T4. Rack awareness

Aviso: no he iniciado, obviamente, el datanode4. Quizás debería eliminarlo

T4.1. Ejecutar 'hdfs dfsadmin -printTopology' para ver la topología actual y apuntar las IPs de los datanodes

```
hdadmin@namenode:~$ hdfs dfsadmin -printTopology

Rack: /default-rack
   172.18.0.4:9866 (datanode2.hadoop-cluster) In Service
   172.18.0.5:9866 (datanode3.hadoop-cluster) In Service
   172.18.0.3:9866 (datanode1.hadoop-cluster) In Service
   172.18.0.7:9866 (datanode5.hadoop-cluster) In Service
```

IPs

T4.2. Apagar los demonios del NameNode

```
hdadmin@namenode:~$ hdfs --daemon stop namenode
hdadmin@namenode:~$ yarn --daemon stop resourcemanager
hdadmin@namenode:~$ jps → Comprobar
1429 Jps → Perfecto
```

T4.3. Crear fichero '$HADOOP_HOME/etc/hadoop/topology.data' que tenga en cada línea la IP de uno de los DataNodes y el rach donde está.

```
hdadmin@namenode:~$ cd $HADOOP_HOME/etc/hadoop
hdadmin@namenode:~/hadoop/etc/hadoop$ nano topology.data
hdadmin@namenode:~/hadoop/etc/hadoop$ cat topology.data
172.18.0.3      /rack1
172.18.0.4      /rack1
172.18.0.5      /rack2
172.18.0.7      /rack2
```

Comprobar

Perfecto

```
 GNU nano 6.2
172.18.0.3      /rack1
172.18.0.4      /rack1
172.18.0.5      /rack2
172.18.0.7      /rack2
```

T4.4. Creo un script de bash '$HADOOP_HOME/etc/hadoop/topology.script'. Le doy permisos de ejecución

```
hdadmin@namenode:~/hadoop/etc/hadoop$ nano topology.script
hdadmin@namenode:~/hadoop/etc/hadoop$ chmod +x topology.script
hdadmin@namenode:~/hadoop/etc/hadoop$ cat topology.script
#! /bin/bash

HADOOP_CONF=$HADOOP_HOME/etc/hadoop
while [ $# -gt 0 ] ; do
  nodeArg=$1
  exec< ${HADOOP_CONF}/topology.data
  result=""
  while read line ; do
    ar=( $line )
    if [ "${ar[0]}" = "$nodeArg" ] ; then
      result="${ar[1]}"
    fi
  done
  shift
  if [ -z "$result" ] ; then
    echo -n "/default-rack "
  else
    echo -n "$result "
  fi
done
```

Permisos

```
 GNU nano 6.2                                    topology.script
#! /bin/bash

HADOOP_CONF=$HADOOP_HOME/etc/hadoop
while [ $# -gt 0 ] ; do
  nodeArg=$1
  exec< ${HADOOP_CONF}/topology.data
  result=""
  while read line ; do
    ar=( $line )
    if [ "${ar[0]}" = "$nodeArg" ] ; then
      result="${ar[1]}"
    fi
  done
  shift
  if [ -z "$result" ] ; then
    echo -n "/default-rack "
  else
    echo -n "$result "
  fi
done
```

T4.5. Defino en el fichero ‘core-site.xml’ la propiedad ‘net.topology.script.file.name’ y le doy como valor el path completo al script

```
hdadmin@namenode:~/hadoop/etc/hadoop$ nano core-site.xml
```

```
  <property>
    <!-- Directorio para almacenamiento temporal (debe tener suficiente espacio) -->
    <name>hadoop.tmp.dir</name>
    <value>/var/tmp/hadoop-${user.name}</value>
    <final>true</final>
  </property>

  <property>
    <!-- Nueva propiedad de la tarea 4 -->
    <name>net.topology.script.file.name</name>
    <value>/opt/bd/hadoop/etc/hadoop/topology.script</value>
  </property>
</configuration>
```

añado esta nueva

T4.6. Inicio los demonios y compruebo que se han identificado los racks

```
hdadmin@namenode:~$ hdfs --daemon start namenode
hdadmin@namenode:~$ yarn --daemon start resourcemanager
hdadmin@namenode:~$ jps
1588 ResourceManager
1476 NameNode
1612 Jps
```

→ Compruebo

→ Perfecto

```
hdadmin@namenode:~$ hdfs dfsadmin -printTopology

Rack: /rack1
   172.18.0.4:9866 (datanode2.hadoop-cluster) In Service
   172.18.0.3:9866 (datanode1.hadoop-cluster) In Service

Rack: /rack2
   172.18.0.5:9866 (datanode3.hadoop-cluster) In Service
   172.18.0.7:9866 (datanode5.hadoop-cluster) In Service
```

Práctica 3

1.

C: > Users > Alejandro > Downloads > citingpatents1.py > ...

Código del ej 1

```python
from mrjob.job import MRJob
from mrjob.protocol import TextProtocol, TextValueProtocol
from typing import Generator, Any

class MRCitingPatents(MRJob):

    # El protocolo de entrada sólo tiene en cuenta el valor (la línea de entrada)
    INPUT_PROTOCOL = TextValueProtocol

    # El protocolo de salida por defecto separa clave y valor por tabulador
    OUTPUT_PROTOCOL = TextProtocol

    def mapper(self, key, value) -> Generator[tuple, Any, None]:
        citing, cited = value.split(',')
        yield cited, citing


    def reducer(self, key, values) -> Generator[tuple, Any, None]:
        yield key, ','.join(values)

if __name__ == '__main__':
    MRCitingPatents.run()
```

Cambiar el directorio

```
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla\tcdm-public\practicas\p3> cd D:\Al
ex\Universidad\Master en Big Data\TCDM\Dsevilla\tcdm-public\practicas\p3
```

Me ha dado un error "no module 'distutils'"

```
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla\tcdm-public\practicas\p3
 pip install setuptools
```

"Abro" el archivo de las patentes que tiene que estar en la misma carpeta

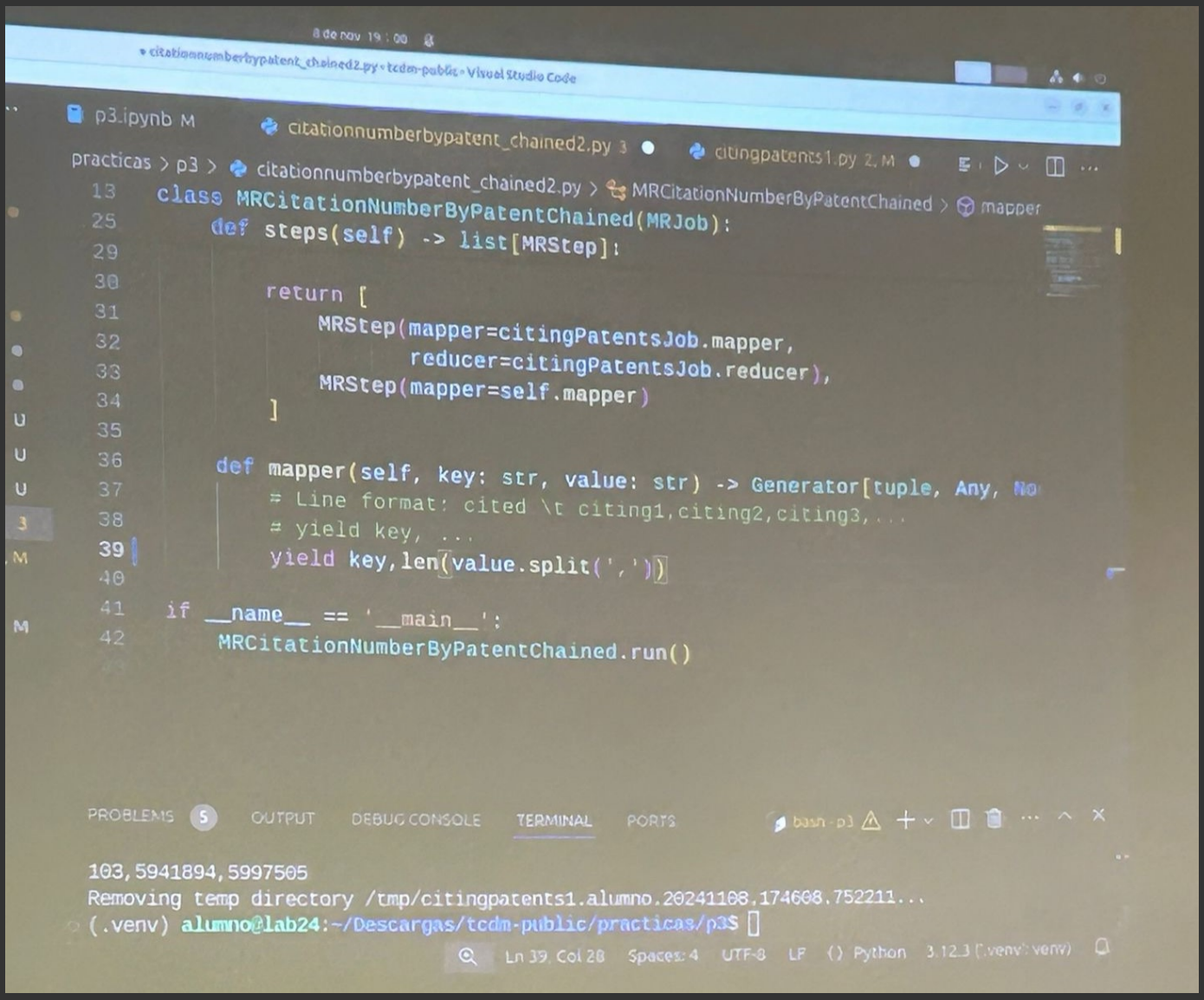
```
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla\tcdm-public\practicas\p3>
 python citingpatents1.py cite75_99.txt
```

Resultado final

```
956872  4959920
963882  5490550
971818  4020975
973582  4646437
976640  4828244
978550  4646807
9799    4087887
980158  4765458
981736  4591350
983468  5301456
986241  3896636
991785  3927583
"CITED" "CITING"
Removing temp directory C:\Users\ALEJAN~1\AppData\Local\Temp\citingpatents1.
Alejandro.20241108.180650.875971...
```

↑ muchas más datos arriba

2.


p3.ipynb M
citationnumberbypatent_chained2.py 3
citingpatents1.py 2, M
practicas > p3 > citationnumberbypatent_chained2.py > MRCitationNumberByPatentChained > mapper
class MRCitationNumberByPatentChained(MRJob):
def steps(self) -> list[MRStep]:
return [
MRStep(mapper=citingPatentsJob.mapper,
reducer=citingPatentsJob.reducer),
MRStep(mapper=self.mapper)
]
def mapper(self, key: str, value: str) -> Generator[tuple, Any, No
Line format: cited \t citing1,citing2,citing3,...
yield key, ...
yield key,len(value.split(','))
if __name__ == '__main__':
MRCitationNumberByPatentChained.run()
PROBLEMS 5
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
bash - p3
103,5941894,5997505
Removing temp directory /tmp/citingpatents1.alumno.20241108.174608.752211...
(.venv) alumno@lab24:~/Descargas/tcdm-public/practicas/p3$
Ln 39, Col 28
Spaces: 4
UTF-8
LF
Python
3.12.3 ('.venv': venv)

3.

```
from mrjob.job import MRJob
from mrjob.protocol import TextProtocol, TextValueProtocol
from typing import Generator, Any

class MRCountryPatents(MRJob):

    # El fichero country_codes.txt se incluirá en el trabajo
    FILES: list[str] = ['patentes-mini/country_codes.txt']

    # El protocolo de entrada sólo tiene en cuenta el valor (la línea de entrada)
    INPUT_PROTOCOL = TextValueProtocol
    # El protocolo de salida por defecto separa clave y valor por tabulador
    OUTPUT_PROTOCOL = TextProtocol

    # Mapa de códigos de país a nombres de país
    country_map: dict[str,str] = {}

    def mapper(self, key: str, value: str) -> Generator[tuple, Any, None]:
        # Line format:
"PATENT","GYEAR","GDATE","APPYEAR","COUNTRY","POSTATE","ASSIGNEE","ASSCODE","
CLAIMS","NCLASS","CAT","SUBCAT","CMADE","CRECEIVE","RATIOCIT","GENERAL","ORIGI
NAL","FWDAPLAG","BCKGTLAG","SELFCTUB","SELFCTLB","SECDUPBD","SECDLWBD"
        # Se puede acceder al mapa de países MRCountryPatents.country_map
        if not value.startswith('"PATENT"'):
            year = value.split(",")[1]
            country = (value.split(",")[4]).replace('"', '')
            patent = value.split(",")[0]
            yield MRCountryPatents.country_map[country], patent + ',' + year

if __name__ == '__main__':
    # Cargar el mapa de países
    with open('patentes-mini/country_codes.txt') as f:
        for line in f:
            (code, name) = line.strip().split('\t')
            MRCountryPatents.country_map[code] = name

    job = MRCountryPatents()
    job.run()
```

Práctica 4

Esta es una guía de cómo empezar a hacer ka práctica 4 (Spark)

1. Crear entorno virtual e instalar Spark

Nota: si falla, probar con

```
apt update
apt install python3-venv
```

```
root@namenode:/# su - luser
luser@namenode:~$ python3 -m venv .venv
luser@namenode:~$ ls -a
.               .bashrc   Python-3.10.12      .wget-hsts
..              .m2       Python-3.10.12.tgz  wordcount
.bash_history   PATH...]  specify             wordcount.tgz
.bash_logout    .profile  .venv
luser@namenode:~$ source .venv/bin/activate
(.venv) luser@namenode:~$ pip install pyspark
```

Crear entorno virtual

Entrar en el entorno virtual

Instalar pyspark

```
  Downloading pyspark-3.5.3.tar.gz (317.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.3/317.3 MB 2.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.7
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 200.5/200.5 KB 2.7 MB/s eta 0:00:00
Using legacy 'setup.py install' for pyspark, since package 'wheel' is no
t installed.
Installing collected packages: py4j, pyspark
  Running setup.py install for pyspark ... done
Successfully installed py4j-0.10.9.7 pyspark-3.5.3
(.venv) luser@namenode:~$
```

Perfecto

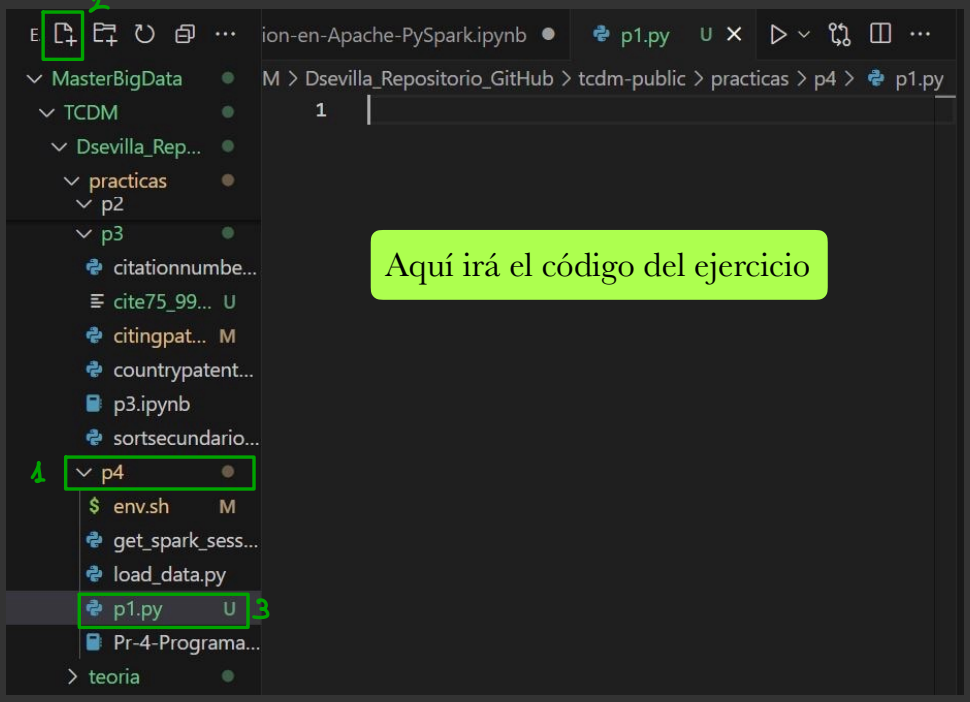
```
(.venv) luser@namenode:~$ pyspark
Python 3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use se
tLogLevel(newLevel).
24/12/13 19:01:05 WARN NativeCodeLoader: Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.3
      /_/

Using Python version 3.10.12 (main, Nov  6 2024 20:22:13)
Spark context Web UI available at http://namenode:4040
Spark context available as 'sc' (master = local[*], app id = local-17341
12866935).
SparkSession available as 'spark'.
>>> exit()
(.venv) luser@namenode:~$
```

Todo bien

2. En el repositorio dsevilla-> tcdm-public -> practicas -> p4, hay un archivo que se llama ‘env.sh’. Hay que ejecutar las dos líneas de código que aparecen

```
(.venv) luser@namenode:~$ export PATH=~/.local/bin:$PATH
(.venv) luser@namenode:~$ export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
```

3. Creamos un nuevo archivo de Python que se va a llamar 'p1.py', que va a ser donde se encuentre el primer ejercicio de la práctica.



4. Esto todavía no es el ejercicio 1, vamos a hacer un ejercicio de prueba para ver que todo se ejecuta correctamente, tanto en local como en HDFS (que es online, en el cluster que se ha hecho en la práctica 1)

```python
#! /usr/bin/env python3
from pyspark.sql import SparkSession, DataFrame
import sys

def load_data(spark: SparkSession, path_cite: str, path_apat: str) -> tuple[DataFrame, DataFrame]:
    cites: DataFrame = (spark
        .read
        .option("inferSchema", "true")
        .option("header", "true")
        .csv(path_cite))
    cites.printSchema()
    cites.show()
    print(cites.count())
    apat: DataFrame = (spark
        .read
        .option("inferSchema", "true")
        .option("header", "true")
        .csv(path_apat))
    apat.printSchema()
    apat.show()
    print(apat.count())
    return cites, apat

def main():
    # Comprueba el número de argumentos
    # sys.argv[1] es el primer argumento, sys.argv[2] el segundo, etc.
    if len(sys.argv) != 5:
        print(f"Uso: {sys.argv[0]} cite75_99.txt apat63_99.txt dfCitas.parquet dfInfo.parquet")
        exit(-1)

    spark: SparkSession = SparkSession\
        .builder\
        .appName("Practica 1 de Tomás")\
        .getOrCreate()

    # Cambio la verbosidad para reducir el número de
    # mensajes por pantalla
    spark.sparkContext.setLogLevel("FATAL")
    # Código del programa

    path_cite = sys.argv[1]
    path_apat = sys.argv[2]

    load_data(spark, path_cite, path_apat)

if __name__ == "__main__":
    main()
```

Como podemos ver, es el ejem0lo que aparece al principio de la tarea, pero con algunos cambios

5. Ejecución en local

Ahora abrimos una nueva terminal. Vamos a estar trabajando con dos terminales

- T1: La que está dentro del NameNode con el usuario luser y el entorno virtual (la que se ha usado antes)
- T2: Una nueva terminal vacía

T2. Me muevo al directorio donde tengo la práctica 4 y copio el archivo p4.py en el namenode/home/luser

```
PS D:\Alex\Universidad\MasterBigData> cd "D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla_Repositorio_GitHub\tcdm-public\practicas\p4"
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla_Repositorio_GitHub\tcdm-public\practicas\p4> docker cp p1.py namenode:/home/luser
Successfully copied 3.07kB to namenode:/home/luser
```

→ Perfecto

T1: Comprueba que se ha copiado bien

```
(.venv) luser@namenode:~$ ls
p1.py     Python-3.10.12      specify     wordcount.tgz
PATH...]  Python-3.10.12.tgz  wordcount
```

→ Perfecto

T2: Me muevo ahora al directorio de dentro del repositorio donde están los datos y copio el archivo patentes-mini.tar.gz en el namenode/home/luser

```
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla_Repositorio_GitHub\tcdm-public\practicas\p4> cd D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla_Repositorio_GitHub\tcdm-public\datos
PS D:\Alex\Universidad\MasterBigData\TCDM\Dsevilla_Repositorio_GitHub\tcdm-public\datos> docker cp patentes-mini.tar.gz namenode:/home/luser
Successfully copied 614kB to namenode:/home/luser
```

→ Perfecto

T1: Comprueba que se ha copiado bien → descomprimo → comprueba que se ha hecho

```
(.venv) luser@namenode:~$ ls
p1.py              PATH...]          Python-3.10.12.tgz  wordcount
patentes-mini.tar.gz  Python-3.10.12  specify             wordcount.tgz
(.venv) luser@namenode:~$ tar xzf patentes-mini.tar.gz
(.venv) luser@namenode:~$ ls
p1.py              PATH...]          specify
patentes-mini      Python-3.10.12    wordcount
patentes-mini.tar.gz  Python-3.10.12.tgz  wordcount.tgz
(.venv) luser@namenode:~$
```

→ Descomprimo

Bien copiado

Perfecto

T1 Ejecutando en local

```
(.venv) luser@namenode:~$ python3 p1.py patantes-mini/apat63_99.tct patentes-mini/cite75_99.txt dfCitas.parquet dfInfo.packet
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/12/13 20:48:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Aparecerá algo así

```
+-------+-------+
| CITING|  CITED|
+-------+-------+
|5497295|4943137|
|5653004|4839947|
|5758472|4113100|
|5490848|4610684|
|4615147|2370792|
|3979372|3857795|
|4815425|4610225|
|5121794|4029160|
|4953371|1490453|
|5815261|5434666|
|4806338|4268499|
|5834327|5436744|
|5574255|4263472|
|4981144|3405706|
|4224146|2959285|
|4679838|1088471|
|5629577|3902084|
|5432544|5022397|
|5672266|5185488|
|5229709|4723108|
+-------+-------+
```

No me acuerdo ezpxactamente de los comandos de aquí en adelante para ponerlo en HDFS y ejecutarlo con el 'spark-summit' >:]