

Prácticas con HDFS

1. Muestra la ayuda del comando para manejar el sistema de archivos de HDFS.

```
C:\WINDOWS\system32\cmd.exe X maria_dev@sandbox-hdp:~ X + v
[maria_dev@sandbox-hdp ~]$ hdfs --help
Usage: hdfs [--config confdir] [--loglevel loglevel] COMMAND
    where COMMAND is one of:
    dfs                run a filesystem command on the file systems supported in Hadoop.
    classpath          prints the classpath
    namenode -format    format the DFS filesystem
    secondarynamenode  run the DFS secondary namenode
    namenode            run the DFS namenode
    journalnode         run the DFS journalnode
    zkfc               run the ZK Failover Controller daemon
    datanode            run a DFS datanode
    dfsadmin           run a DFS admin client
    envvars            display computed Hadoop environment variables
    haadmin            run a DFS HA admin client
    fsck              run a DFS filesystem checking utility
    balancer          run a cluster balancing utility
    jmxget            get JMX exported values from NameNode or DataNode.
    mover            run a utility to move block replicas across
                    storage types
    oiv              apply the offline fsimage viewer to an fsimage
    oiv_legacy       apply the offline fsimage viewer to an legacy fsimage
    oev             apply the offline edits viewer to an edits file
    fetchdt         fetch a delegation token from the NameNode
    getconf         get config values from configuration
    groups          get the groups which users belong to
    snapshotDiff    diff two snapshots of a directory or diff the
                    current directory contents with a snapshot
    lsSnapshottableDir list all snapshottable dirs owned by the current user
                    Use -help to see options
    portmap         run a portmap service
    nfs3            run an NFS version 3 gateway
    cacheadmin      configure the HDFS cache
    crypto          configure HDFS encryption zones
    storagepolicies list/get/set block storage policies
    version         print the version

Most commands print help when invoked w/o parameters.
[maria_dev@sandbox-hdp ~]$ |
```

2. Muestra el contenido de la raíz de HDFS.

```

Most commands print help when invoked w/o parameters.
[maria_dev@sandbox-hdp ~]$ cd /
[maria_dev@sandbox-hdp /]$ pwd
/
[maria_dev@sandbox-hdp /]$ ls -l
total 5336
-rw-r--r-- 1 root root 11976 Apr 2 2018 anaconda-post.log
drwxr-xr-x 3 root root 4096 Jun 18 2018 apps
lrwxrwxrwx 1 root root 7 Apr 2 2018 bin -> usr/bin
drwxr-xr-x 3 root root 4096 Jun 18 2018 boot
drwxr-xr-x 1 root root 4096 Jun 18 2018 cgroups_test
-rw----- 1 root root 2068488 Nov 6 08:01 core.30149
-rw----- 1 root root 2068488 Nov 6 08:05 core.31168
-rw----- 1 root root 2068488 Nov 6 08:05 core.31308
drwxr-xr-x 19 root root 2988 Nov 6 07:32 dev
drwxr-xr-x 1 root root 4096 Nov 5 10:09 etc
drwxr-xr-x 1 root root 4096 Jun 18 2018 hadoop
drwxr-xr-x 1 root root 4096 Jun 18 2018 home
drwxr-xr-x 1 kafka hadoop 4096 Nov 6 09:05 kafka-logs
lrwxrwxrwx 1 root root 7 Apr 2 2018 lib -> usr/lib
lrwxrwxrwx 1 root root 9 Apr 2 2018 lib64 -> usr/lib64
drwxr-xr-x 2 root root 4096 Nov 5 2016 media
drwxr-xr-x 2 root root 4096 Nov 5 2016 mnt
drwxr-xr-x 1 root root 4096 Jun 18 2018 mysql-connector-java-5.1.45
drwxr-xr-x 1 root root 4096 Jun 18 2018 opt
drwxr-xr-x 2 root root 4096 Jun 18 2018 packer-files
dr-xr-xr-x 265 root root 0 Nov 6 07:32 proc
dr-xr-xr-x 1 root root 4096 Jun 18 2018 root
drwxr-xr-x 34 root root 788 Nov 6 07:53 run
drwxr-xr-x 1 root root 4096 Jun 18 2018 sandbox
-rw-r--r-- 1 root root 4 Jun 18 2018 sandbox-flavour
lrwxrwxrwx 1 root root 8 Apr 2 2018 sbin -> usr/sbin
drwxr-xr-x 2 root root 4096 Nov 5 2016 srv
dr-xr-xr-x 13 root root 0 Nov 6 07:32 sys
drwxrwxrwt 1 root root 4096 Nov 6 09:03 tmp
drwxr-xr-x 1 root root 4096 Jun 18 2018 usr
drwxr-xr-x 1 root root 4096 Jun 18 2018 var
[maria_dev@sandbox-hdp /]$ ls
anaconda-post.log bin cgroups_test core.31168 dev hadoop kafka-logs lib64 mnt
apps boot core.30149 core.31308 etc home lib media mysql-connector-java-5.1.45
opt packer-files proc run sandbox sandbox-flavour srv tmp var
[maria_dev@sandbox-hdp /]$

```

3. Visualiza dicha carpeta raíz desde el gestor de archivos del navegador. ¿En qué puerto se localiza?

Ambari - Sandbox - Google Chrome

localhost:8080/#/main/view/FILES/auto_files_instance

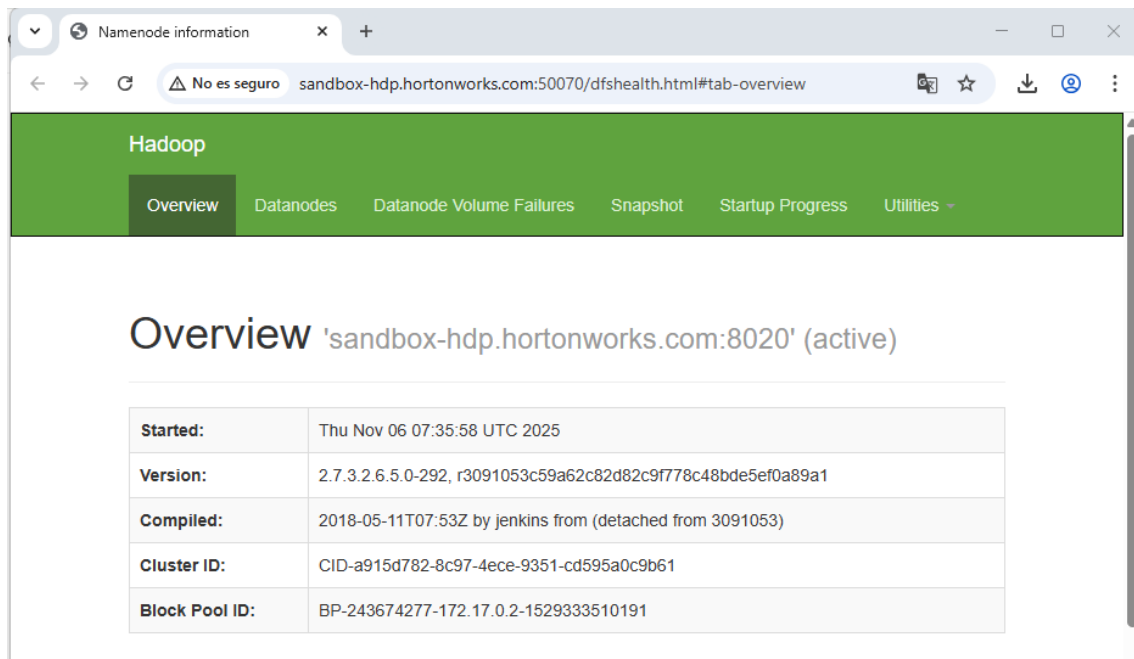
Ambari Sandbox 0 ops 0 alerts Dashboard Services Hosts Alerts Admin maria_dev

Total: 11 files or folders

+ Select All New Folder Upload

Search in current directory...

| Name > | Size > | Last Modified > | Owner > | Group > | Permission |
|----------------|--------|------------------|---------|---------|------------|
| app-logs | -- | 2018-06-18 17:18 | yarn | hadoop | drwxrwxrwx |
| apps | -- | 2018-06-18 18:13 | hdfs | hdfs | drwxr-xr-x |
| ats | -- | 2018-06-18 16:52 | yarn | hadoop | drwxr-xr-x |
| hdp | -- | 2018-06-18 16:52 | hdfs | hdfs | drwxr-xr-x |
| livy2-recovery | -- | 2018-06-18 17:11 | livy | hdfs | drwx----- |
| mapred | -- | 2018-06-18 16:52 | mapred | hdfs | drwxr-xr-x |
| mr-history | -- | 2018-06-18 16:52 | mapred | hadoop | drwxrwxrwx |
| ranger | -- | 2018-06-18 17:59 | hdfs | hdfs | drwxr-xr-x |
| spark2-history | -- | 2025-11-06 10:11 | spark | hadoop | drwxrwxrwx |
| tmp | -- | 2018-06-18 18:06 | hdfs | hdfs | drwxrwxrwx |
| user | -- | 2018-06-18 18:08 | hdfs | hdfs | drwxr-xr-x |



En el puerto 500700

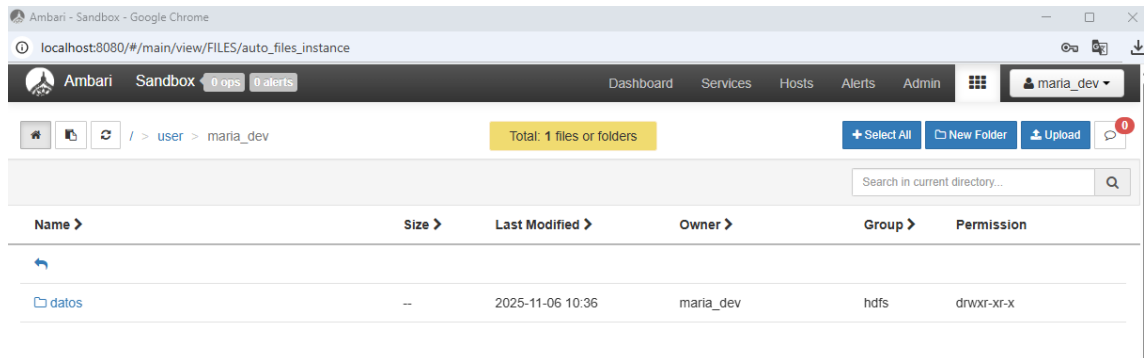
4. Crea en HDFS un nuevo directorio llamado datos dentro de la carpeta `/user/maria_dev`.

```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user
Found 15 items
drwxr-xr-x  - admin      hdfs          0 2018-06-18 14:52 /user/admin
drwxrwx---  - ambari-qa hdfs          0 2018-06-18 14:52 /user/ambari-qa
drwxr-xr-x  - amy_ds    hdfs          0 2018-06-18 14:53 /user/amy_ds
drwxr-xr-x  - root      hdfs          0 2018-06-18 14:52 /user/anonymous
drwxr-xr-x  - druid      hadoop        0 2018-06-18 16:06 /user/druid
drwxr-xr-x  - hbase     hdfs          0 2018-06-18 15:08 /user/hbase
drwxr-xr-x  - hcat      hdfs          0 2018-06-18 15:12 /user/hcat
drwxr-xr-x  - hive      hdfs          0 2018-06-18 15:18 /user/hive
drwxrwxr-x  - livy      hdfs          0 2018-06-18 15:11 /user/livy
drwxr-xr-x  - maria_dev hdfs          0 2025-11-06 09:36 /user/maria_dev
drwxrwxr-x  - oozie     hdfs          0 2018-06-18 16:08 /user/oozie
drwxr-xr-x  - raj_ops   hdfs          0 2018-06-18 14:53 /user/raj_ops
drwxr-xr-x  - root      hdfs          0 2018-06-18 14:52 /user/root
drwxrwxr-x  - spark     hdfs          0 2018-06-18 15:10 /user/spark
drwxr-xr-x  - zeppelin  hdfs          0 2018-06-18 15:10 /user/zeppelin
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 1 items
drwxr-xr-x  - maria_dev hdfs          0 2025-11-06 09:36 /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$
```

5. Comprobar que existe.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 1 items
drwxr-xr-x - maria_dev hdfs          0 2025-11-06 09:36 /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$
```

6. Mostrarlo desde el navegador.



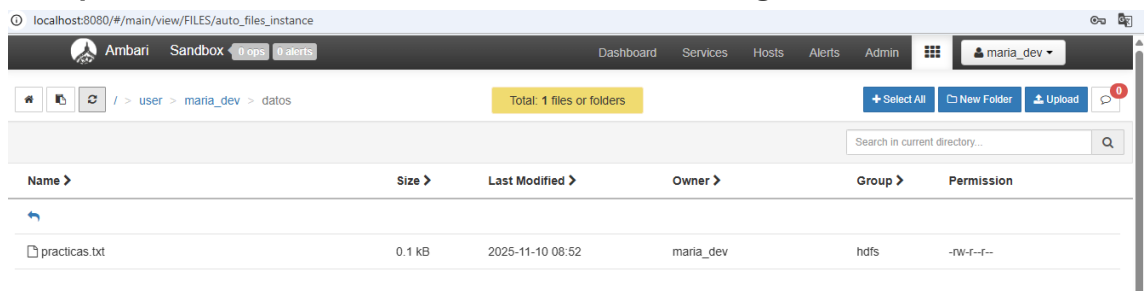
7. Crea un fichero llamado practicas.txt en tu directorio home de tu usuario en Linux con alguna frase dentro.

```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ echo "Esta es una práctica de HDFS." > practicas.txt
[maria_dev@sandbox-hdp ~]$ ls
practicas.txt
[maria_dev@sandbox-hdp ~]$ cat practicas.txt
Esta es una práctica de HDFS.
[maria_dev@sandbox-hdp ~]$ |
```

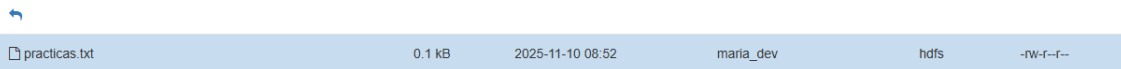
8. Copiarlo en HDFS, en concreto al directorio datos anterior.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 2 items
drwxr-xr-x - maria_dev hdfs          0 2025-11-10 07:50 /user/maria_dev/.Trash
drwxr-xr-x - maria_dev hdfs          0 2025-11-10 07:50 /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put ~/practicas.txt /user/maria_dev/datos/
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/datos
Found 1 items
-rw-r--r--  1 maria_dev hdfs          31 2025-11-10 07:52 /user/maria_dev/datos/practicas.txt
[maria_dev@sandbox-hdp ~]$
```

9. Comprueba su existencia desde la utilidad del navegador.



10. Haz clic sobre el fichero ¿Cuál es el tamaño del fichero en HDFS? ¿Cuánto ocupa realmente en HDFS? ¿Cuántas veces está replicado?



| | | | | | |
|---------------|--------|------------------|-----------|------|------------|
| practicas.txt | 0.1 kB | 2025-11-10 08:52 | maria_dev | hdfs | -rw-r--r-- |
|---------------|--------|------------------|-----------|------|------------|

El tamaño es de 0.1 kB en el navegador

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -du -h /user/maria_dev/datos/practicas.txt
31 /user/maria_dev/datos/practicas.txt
[maria_dev@sandbox-hdp ~]$
```

Desde ssh nos da un tamaño de 31B y esta replicado 1 vez

```
The filesystem under path '/user/maria_dev/datos/practicas.txt' is HEALTHY
[maria_dev@sandbox-hdp ~]$ hdfs fsck /user/maria_dev/datos/practicas.txt -files -blocks -locations
```

```
2.18.0.2:50010,DS-ab75b94d-c6f2-4415-8639-1aaec2609e13,DISK]]

Status: HEALTHY
Total size:      31 B
Total dirs:      0
Total files:     1
Total symlinks:  0
Total blocks (validated): 1 (avg. block size 31 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Nov 10 07:57:27 UTC 2025 in 1 milliseconds
```

11. Visualizar su contenido en HDFS.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/datos/practicas.txt
Esta es una práctica de HDFS.
[maria_dev@sandbox-hdp ~]$ |
```

12. HDFS es un sistema de archivos distribuido que está dentro de una carpeta local de nuestro Linux. En su archivo de configuración está su ubicación. ¿Podrías localizarla? Explora su contenido desde Linux ¿Qué carpetas tiene dentro? ¿Dónde están los datos?

```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ cd /etc/hadoop
[maria_dev@sandbox-hdp hadoop]$ ls
2.6.5.0-292  conf  conf.backup
[maria_dev@sandbox-hdp hadoop]$
```

Se encuentra en la carpeta 2.6.5.0-292/0/

```
[maria_dev@sandbox-hdp ~]$ grep -R "dfs.datanode.data.dir" $HADOOP_CONF_DIR -n || grep -R "dfs.datanode.data.dir" /etc/hadoop -n
/etc/hadoop/2.6.5.0-292/0/hdfs-site.xml:54:      <name>dfs.datanode.data.dir</name>
/etc/hadoop/2.6.5.0-292/0/hdfs-site.xml:60:      <name>dfs.datanode.data.dir.perm</name>
/etc/hadoop/conf/hdfs-site.xml:54:      <name>dfs.datanode.data.dir</name>
/etc/hadoop/conf/hdfs-site.xml:60:      <name>dfs.datanode.data.dir.perm</name>
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
```

13. Si damos clic sobre el fichero en el entorno del navegador aparece el Block Id del fichero que nos indica el nombre del subdirectorio dentro del sistema de archivos local del punto anterior ¿Podrías mostrar su contenido?

```
[maria_dev@sandbox-hdp 0]$ sudo find / -type d -name "BP-*" 2>/dev/null
/hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191
[maria_dev@sandbox-hdp 0]$ sudo ls -lh /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191
total 16K
drwxr-xr-x 1 hdfs hadoop 4.0K Nov  6 09:45 current
-rw-r--r-- 1 hdfs hadoop 166 Nov  5 10:48 scanner.cursor
drwxr-xr-x 2 hdfs hadoop 4.0K Nov 10 07:44 tmp

[maria_dev@sandbox-hdp 0]$ sudo ls -lh /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current
total 24K
-rw-r--r-- 1 hdfs hadoop 24 Nov  6 09:45 dfsUsed
drwxr-xr-x 1 hdfs hadoop 4.0K Jun 18 2018 finalized
drwxr-xr-x 1 hdfs hadoop 4.0K Nov 10 07:52 rbw
-rw-r--r-- 1 hdfs hadoop 129 Nov 10 07:44 VERSION
[maria_dev@sandbox-hdp 0]$ sudo hexdump -C /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/dfsUsed | head
00000000 32 32 31 39 39 39 35 31 33 36 20 31 37 36 32 34 |2219995136 17624|
00000010 32 32 33 30 31 32 32 31 |22301221|
00000018
[maria_dev@sandbox-hdp 0]$
```

14. Vamos a crear otro ejemplo con un fichero grande. Investiga como crear automáticamente desde Linux con un comando un archivo de 1GB en la carpeta home de tu usuario en Linux.

```
[maria_dev@sandbox-hdp /]$ dd if=/dev/zero of=~/.datos/archivo_1GB.bin bs=1M count=1024
dd: failed to open ‘/home/maria_dev/.datos/archivo_1GB.bin’: No such file or directory
[maria_dev@sandbox-hdp /]$ dd if=/dev/zero of=~/.archivo_1GB.bin bs=1M count=1024
1024+0 records in
1024+0 records out
1073741824 bytes (1.1 GB) copied, 15.826 s, 67.8 MB/s
[maria_dev@sandbox-hdp /]$ pwd
/
```

```
[maria_dev@sandbox-hdp /]$ ls -lh ~/archivo_1GB.bin ; du -h ~/archivo_1GB.bin
-rw-rw-r-- 1 maria_dev maria_dev 1.0G Nov 10 08:33 /home/maria_dev/archivo_1GB.bin
1.1G    /home/maria_dev/archivo_1GB.bin
[maria_dev@sandbox-hdp /]$
```

15. Copia el archivo anterior al directorio /datos de nuestro HDFS.

```
[maria_dev@sandbox-hdp /]$ hdfs dfs -put ~/archivo_1GB.bin /user/maria_dev/datos/
[maria_dev@sandbox-hdp /]$ hdfs dfs -ls /user/maria_dev/datos/
Found 2 items
-rw-r--r-- 1 maria_dev hdfs 1073741824 2025-11-10 08:39 /user/maria_dev/datos/archivo_1GB.bin
-rw-r--r-- 1 maria_dev hdfs 31 2025-11-10 07:52 /user/maria_dev/datos/practicas.txt
[maria_dev@sandbox-hdp /]$
```

| Ambari Sandbox 0 ops 0 alerts | | | | | |
|---------------------------------------|-----------|------------------|-----------|---------|------------|
| Dashboard Services Hosts Alerts Admin | | | | | |
| / > user > maria_dev > datos | | | | | |
| Total: 2 files or folders | | | | | |
| + Select All New Folder Upload | | | | | |
| Search in current directory... | | | | | |
| Name > | Size > | Last Modified > | Owner > | Group > | Permission |
| archivo_1GB.bin | 1024.0 MB | 2025-11-10 09:39 | maria_dev | hdfs | -rw-r--r-- |
| practicas.txt | 0.1 KB | 2025-11-10 08:52 | maria_dev | hdfs | -rw-r--r-- |

16. Comprueba en la página web que ha creado múltiples bloques ¿Cuántos ha creado? ¿De qué tamaño son? Como solamente tenemos un nodo aparecen todos los bloques en el mismo, pero en un clúster real cada bloque estaría en un nodo distinto.

```
Status: HEALTHY
Total size: 1073741824 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 8 (avg. block size 134217728 B)
Minimally replicated blocks: 8 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Mon Nov 10 08:41:23 UTC 2025 in 1 milliseconds

The filesystem under path '/user/maria_dev/datos/archivo_1GB.bin' is HEALTHY
[maria_dev@sandbox-hdp /]$ hdfs fsck /user/maria_dev/datos/archivo_1GB.bin -files -blocks -locations
```

Ha creado 8 Bloques de más o menos 134 Mb

17. ¿Puedes localizar en el sistema de archivos local dichos bloques?

```
[maria_dev@sandbox-hdp ~]$ sudo find /var -type f -name "blk_*" 2>/dev/null | grep <parte_del_blockid
>
-bash: syntax error near unexpected token `newline'
[maria_dev@sandbox-hdp ~]$ ls -lh
total 3.1G
-rw-rw-r-- 1 maria_dev maria_dev 1.0G Nov 10 08:33 archivo_1GB.bin
-rw-r--r-- 1 maria_dev maria_dev 1.0G Nov 10 08:25 archivo_de_1GB_usuario.bin
-rw-r--r-- 1 maria_dev maria_dev 1.0G Nov 10 08:32 archivo_de_1GB_usuario.bin
rm -rf /var/clear
```

18. Vamos a crear otro directorio llamado practicas dentro de la carpeta /user/maria_dev. Big Data

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir -p /user/maria_dev/practicas
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x - maria_dev hdfs 0 2025-11-10 08:39 datos
drwxr-xr-x - maria_dev hdfs 0 2025-11-10 09:06 practicas
[maria_dev@sandbox-hdp ~]$ |
```

19. Copiamos prueba.txt desde datos a prácticas.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cp /user/maria_dev/datos/prueba.txt /user/maria_dev/practicas/
cp: '/user/maria_dev/datos/prueba.txt': No such file or directory
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cp /user/maria_dev/datos/practicas.txt /user/maria_dev/practicas/
[maria_dev@sandbox-hdp ~]$
```

20. Comprobamos el contenido de prácticas.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/practicas
Found 1 items
-rw-r--r-- 1 maria_dev hdfs 31 2025-11-10 09:09 /user/maria_dev/practicas/practicas.txt
[maria_dev@sandbox-hdp ~]$
```

21. Comprobamos el contenido de prueba.txt con un comando de HDFS.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/practicas/practicas.txt
Esta es una práctica de HDFS.
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev/practicas
Found 1 items
-rw-r--r-- 1 maria_dev hdfs 31 2025-11-10 09:09 /user/maria_dev/practicas/practicas.txt
[maria_dev@sandbox-hdp ~]$ |
```

22. Borramos el fichero prueba.txt.

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -rm /user/maria_dev/practicas/practicas.txt
25/11/10 09:10:52 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/practicas/practicas.txt' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/.Trash/Current/user/maria_dev/practicas/practicas.txt
[maria_dev@sandbox-hdp ~]$
```

23. Borra el directorio prácticas.


```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -rmdir /user/maria_dev/practicas
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 2 items
drwx----- - maria_dev hdfs          0 2025-11-10 09:10 /user/maria_dev/.Trash
drwxr-xr-x - maria_dev hdfs          0 2025-11-10 08:39 /user/maria_dev/datos
[maria_dev@sandbox-hdp ~]$
```