



Doctor **Fleming**
Instituto de Educación Secundaria

BIG DATA

2025/26

Ciclo	Especialización IA & BIG DATA
Nombre	Rodrigo Medina
Correo	YMQ06518@educastur.es
Nº Unidad Didáctica	02

PR_02.3

1.	HDFS	2
----	------------	---

1. HDFS

ssh root@localhost -p 2222

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

Lista de subcomandos con el comando hdfs desde terminal

```
hdfs
```

```
[maria_dev@sandbox-hdp ~]$ hdfs
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
```

Lista de opciones con el comando con hdfs dfs.

```
hdfs dfs
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs
Usage: hadoop fs [generic options]
      [-appendToFile <localsrc> ... <dst>]
      [-cat [-ignoreCrc] <src> ...]
      [-checksum <src> ...]
      [-chgrp [-R] GROUP PATH...]
      [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
      [-chown [-R] [OWNER][:[GROUP]] PATH...]
      [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
      [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
      [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] <path> ...]
      [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
      [-createSnapshot <snapshotDir> [<snapshotName>]]
      [-deleteSnapshot <snapshotDir> <snapshotName>]
      [-df [-h] [<path> ...]]
      [-du [-s] [-h] <path> ...]
      [-expunge]
      [-find <path> ... <expression> ...]
      [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
      [-getfacl [-R] <path>]
      [-getfattr [-R] {-n name | -d} [-e en] <path>]
      [-getmerge [-nl] <src> <localdst>]
      [-help [cmd ...]]
      [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
      [-mkdir [-p] <path> ...]
      [-moveFromLocal <localsrc> ... <dst>]
      [-moveToLocal <src> <localdst>]
      [-mv <src> ... <dst>]
      [-put [-f] [-p] [-l] <localsrc> ... <dst>]
```

2. Muestra el contenido de la raíz de HDFS.

```
hdfs dfs -ls /
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /
Found 11 items
drwxrwxrwx  - yarn   hadoop          0 2018-06-18 15:18 /app-logs
drwxr-xr-x  - hdfs   hdfs           0 2018-06-18 16:13 /apps
drwxr-xr-x  - yarn   hadoop          0 2018-06-18 14:52 /ats
drwxr-xr-x  - hdfs   hdfs           0 2018-06-18 14:52 /hdp
drwx-----  - livy   hdfs           0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x  - mapred hdfs          0 2018-06-18 14:52 /mapred
drwxrwxrwx  - mapred hadoop        0 2018-06-18 14:52 /mr-history
drwxr-xr-x  - hdfs   hdfs           0 2018-06-18 15:59 /ranger
drwxrwxrwx  - spark   hadoop        0 2025-11-07 18:25 /spark2-history
drwxrwxrwx  - hdfs   hdfs           0 2018-06-18 16:06 /tmp
drwxr-xr-x  - hdfs   hdfs           0 2018-06-18 16:08 /user
```

3. Visualiza dicha carpeta raíz desde el gestor de archivos del navegador.

¿En qué puerto se localiza?

Tras acceder desde navegador:

```
localhost:1080 → HDFS → Quick Links → NameNode UI
```

The screenshot shows the Ambari interface for the HDFS service. On the left sidebar, 'HDFS' is selected. In the main area, the 'Summary' tab is active. The 'Quick Links' menu is open, showing options like 'NameNode JMX', 'NameNode Logs', and 'NameNode UI'. The 'NameNode UI' link is highlighted with a red box and an arrow pointing to it. Other links in the 'Quick Links' menu are also highlighted with red boxes.

En la nueva ventana:

```
Utilities → Browse the file system
```

El gestor de archivo de la interfaz gráfica se localiza en el puerto **50070**

BIG DATA

The screenshot shows a web browser window titled "Browsing HDFS". The address bar displays "No es seguro sandbox-hdp.hortonworks.com 50070/explorer.html#/". The top navigation bar includes links for "Hadoop", "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot" (which is highlighted with a red box and arrow), "Startup Progress", and "Utilities". The main content area is titled "Browse Directory" and shows a table of file and directory listings under the root directory "/". The table columns are: Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. The data includes entries for app-logs, apps, ats, hdp, ivy2-recovery, mapred, mr-history, ranger, spark2-history, tmp, and user.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	yarn	hadoop	0 B	18/6/2018, 17:18:48	0	0 B	app-logs
drwxr-xr-x	hdfs	hdfs	0 B	18/6/2018, 18:13:31	0	0 B	apps
drwxr-xr-x	yarn	hadoop	0 B	18/6/2018, 16:52:05	0	0 B	ats
drwxr-xr-x	hdfs	hdfs	0 B	18/6/2018, 16:52:12	0	0 B	hdp
drwx-----	ivy	hdfs	0 B	18/6/2018, 17:11:42	0	0 B	ivy2-recovery
drwxr-xr-x	mapred	hdfs	0 B	18/6/2018, 16:52:10	0	0 B	mapred
drwxrwxrwx	mapred	hadoop	0 B	18/6/2018, 16:52:18	0	0 B	mr-history
drwxr-xr-X	hdfs	hdfs	0 B	18/6/2018, 17:59:35	0	0 B	ranger
drwxrwxrwx	spark	hadoop	0 B	7/11/2025, 19:26:34	0	0 B	spark2-history
drwxrwxrwx	hdfs	hdfs	0 B	18/6/2018, 18:06:11	0	0 B	tmp
drwxr-xr-x	hdfs	hdfs	0 B	18/6/2018, 18:08:07	0	0 B	user

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

```
hdfs dfs -mkdir /user/maria_dev/datos
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir /user/maria_dev/datos
```

5. Comprobar que existe.

```
hdfs dfs -ls /user/maria_dev
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls /user/maria_dev
Found 1 items
drwxr-xr-x  - maria_dev hdfs          0 2025-11-07 18:33 /user/maria_dev/datos
```

6. Mostrarlo desde el navegador.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	maria_dev	hdfs	0 B	7/11/2025, 19:33:13	0	0 B	datos

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

```
echo "texto de prueba" > practicas.txt
```

```
[maria_dev@sandbox-hdp ~]$ pwd
/home/maria_dev
[maria_dev@sandbox-hdp ~]$ echo "texto de prueba" > practicas.txt
[maria_dev@sandbox-hdp ~]$ cat practicas.txt
texto de prueba
```

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

```
hdfs dfs -put practicas.txt /user/maria_dev/datos
```

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

9. Comprueba su existencia desde la utilidad del navegador.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	maria_dev	hdfs	16 B	7/11/2025, 19:50:58	1	128 MB	practicas.txt

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

Tamaño fichero en HDFS : 16 (bytes)

Tamaño del bloque: 128 MB6 (bytes)

Tamaño real en HDFS: 16 (bytes)

Veces replicado: 1

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r-r-	maria_dev	hdfs	16 B	7/11/2025, 19:50:58	1	128 MB	practicas.txt

11. Visualizar su contenido en HDFS.

```
hdfs dfs -cat /user/maria_dev/datos/practicas.txt
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat /user/maria_dev/datos/practicas.txt
texto de prueba
```

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?

El archivo de configuración de Hadoop se encuentra dentro de Linux. Para acceder a él, hacemos un cat en la siguiente ruta.

```
cat /etc/hadoop/conf/hdfs-site.xml
```

En el archivo de configuración `hdfs-site.xml`, podemos comprobar el subdirectorio de Linux con la ruta donde se encuentran los datos (*DataNode*) con la división de los bloques de los archivos en HDFS

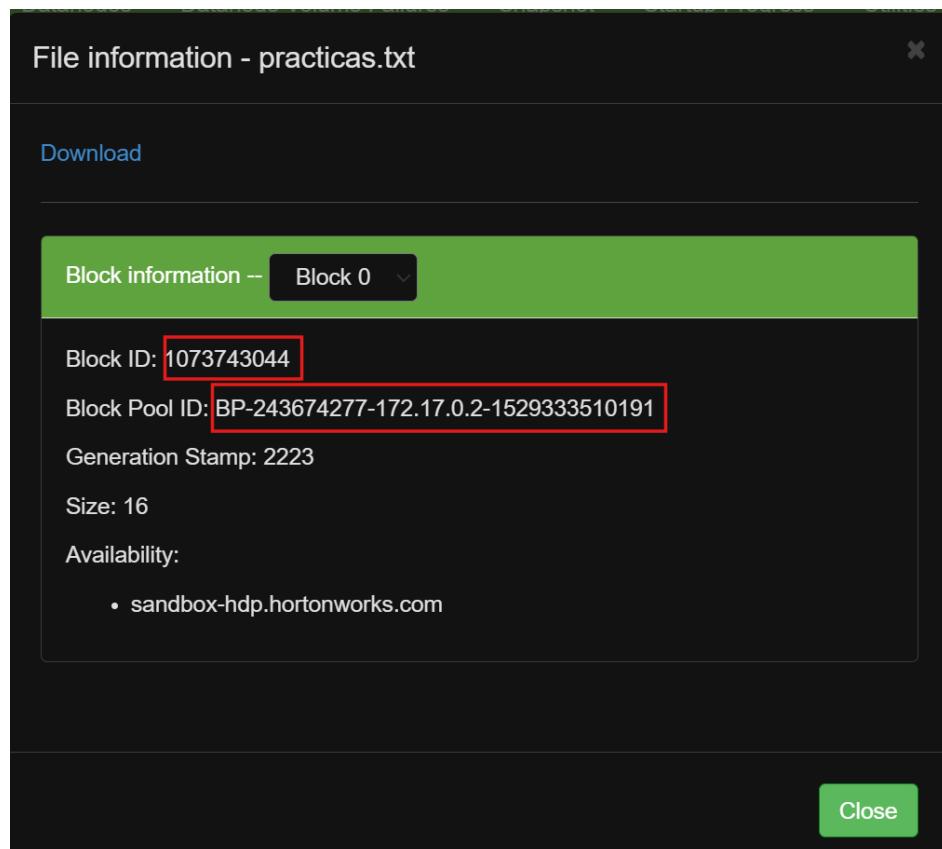
```
<property>
  <name>dfs.datanode.data.dir</name>
  <value>/hadoop/hdfs/data</value>
  <final>true</final>
</property>
```

/hadoop/hdfs/data

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?

Comprobamos el ID del bloque del archivo para su búsqueda.

El Block Pool ID corresponde el directorio donde se almacenan todos los bloques de un archivo.



BIG DATA

Una vez conseguido el ID del bloque del archivo, una opción sería su búsqueda a través de find. Para tener acceso a los directorios de hdfs, necesitamos permisos de administrador con *sudo su*

```
sudo su  
find / -type f -name *1073743044*
```

```
[root@sandbox-hdp maria_dev]# find / -type f -name *1073743044*  
/hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/finalized/subdir0/subdir4/blk_1073743044 ←  
/hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/finalized/subdir0/subdir4/blk_1073743044_2223.meta
```

Una vez encontrada la ruta, podemos visualizar el contenido a través de cat

```
cat /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-  
1529333510191/current/finalized/subdir0/subdir4/blk_1073743  
044
```

```
[root@sandbox-hdp maria_dev]# cat /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/finalized/subdir0/sub  
dir4/blk_1073743044  
texto de prueba
```

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.

Crear un archivo de 1GB en Linux

```
fallocate -l 1G ~/archivo_1GB
```

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

Copiar el archivo de Linux a HDFS

```
hdfs dfs -put archivo_1GB /user/maria_dev/datos
```

```
[maria_dev@sandbox-hdp ~]$ fallocate -l 1G ~/archivo_1GB  
[maria_dev@sandbox-hdp ~]$ ls  
archivo_1GB  practicas.txt  
[maria_dev@sandbox-hdp ~]$ hdfs dfs -put archivo_1GB /user/maria_dev/datos
```

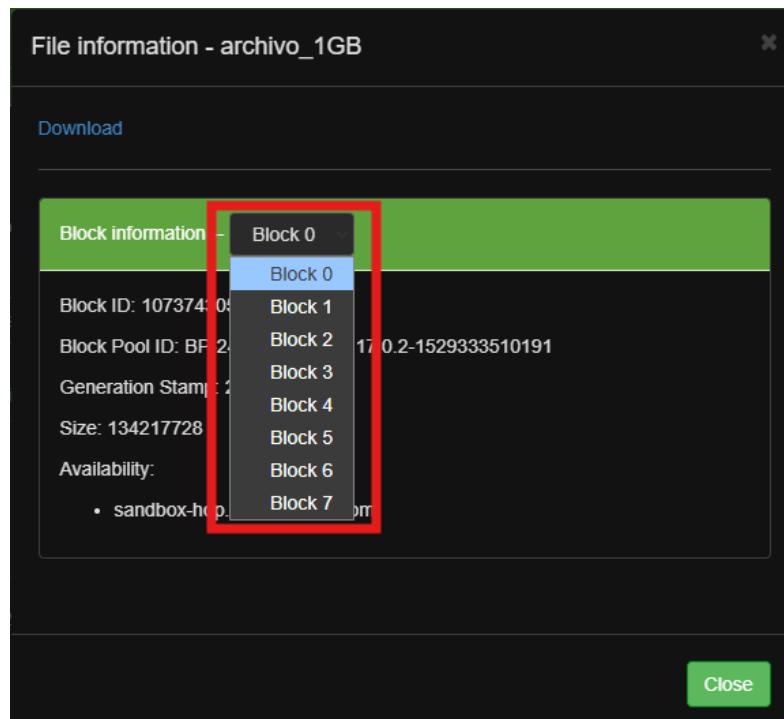
Comprobación desde terminal

```
hdfs dfs -ls /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 18:23 /user/maria_dev/datos/archivo_1GB
-rw-r--r-- 1 maria_dev hdfs          16 2025-11-07 18:50 /user/maria_dev/datos/practicas.txt
```

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.

El archivo está dividido en 8 bloques (de bloque 0 al 7) cada uno de ellos con un tamaño de 128 MB ($128\text{MB} * 8 = 1024\text{MB} = 1\text{GB}$)



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

Conociendo el ID de un bloque o del Block Pool, buscamos la ruta exacta, con sudo para tener permisos de administrador.

Si buscamos sobre el bloque 0, el ID del resto serán los 7 siguientes, incrementando el ID en +1. Por lo tanto, en este caso del 1073743053 al 1073743060

```
sudo find / -name *1073743053*
```

```
[maria_dev@sandbox-hdp ~]$ sudo find / -name *1073743053*
/hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/finalized/subdir0/subdir4/blk_1073743053_2233.meta
```

Obtenida la ruta, copiamos y listamos el directorio

```
sudo ls -l /hadoop/hdfs/data/current/BP-243674277-172.17.0.2-1529333510191/current/finalized/subdir0/subdir4/
```

```
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743053
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743053_2233.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743054
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743054_2234.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743055
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743055_2235.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743056
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743056_2236.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743057
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743057_2237.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743058
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743058_2238.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:22 blk_1073743059
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:22 blk_1073743059_2239.meta
-rw-r--r-- 1 hdfs hadoop 134217728 Nov 10 18:23 blk_1073743060
-rw-r--r-- 1 hdfs hadoop 1048583 Nov 10 18:23 blk_1073743060_2240.meta
[maria_dev@sandbox-hdp ~]$
```

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

```
hdfs dfs -mkdir practicas
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mkdir practicas
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x  - maria_dev hdfs          0 2025-11-10 18:23 datos
drwxr-xr-x  - maria_dev hdfs          0 2025-11-10 19:01 practicas
```

19.. Copiamos prueba.txt (*practicas.txt*) desde datos a practicas.

*Renombrar para el ejercicio el archivo *practicas.txt* por *prueba.txt*

```
hdfs dfs -mv datos/practicas.txt datos/prueba.txt
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -mv datos/practicas.txt datos/prueba.txt
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls -R
drwxr-xr-x  - maria_dev hdfs          0 2025-11-10 19:17 datos
-rw-r--r--  1 maria_dev hdfs 1073741824 2025-11-10 18:23 datos/archivo_1GB
-rw-r--r--  1 maria_dev hdfs         16 2025-11-07 18:50 datos/prueba.txt
drwxr-xr-x  - maria_dev hdfs          0 2025-11-10 19:01 practicas
```

Copiar prueba.txt

```
hdfs dfs -cp datos/prueba.txt practicas
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cp datos/prueba.txt practicas
```

20.Comprobamos el contenido de practicas.

```
hdfs dfs -ls practicas
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls practicas
Found 1 items
-rw-r--r--  1 maria_dev hdfs         16 2025-11-10 19:18 practicas/prueba.txt
```

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

```
hdfs dfs -cat practicas/prueba.txt
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -cat practicas/prueba.txt
texto de prueba
```

22.Borramos el fichero prueba.txt.

```
hdfs dfs -rm practicas/prueba.txt
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -rm practicas/prueba.txt
25/11/10 19:26:28 INFO fs.TrashPolicyDefault: Moved: 'hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/practicas/prueba.txt' to trash at: hdfs://sandbox-hdp.hortonworks.com:8020/user/maria_dev/.Trash/Current/user/maria_dev/practicas/prueba.txt
```

23. Borra el directorio practicas.

```
hdfs dfs -rmr practicas
```

```
[maria_dev@sandbox-hdp ~]$ hdfs dfs -rmr practicas
[maria_dev@sandbox-hdp ~]$ hdfs dfs -ls -R
drwx-----  - maria_dev hdfs          0 2025-11-10 19:26 .Trash
drwx-----  - maria_dev hdfs          0 2025-11-10 19:26 .Trash/Current
drwx-----  - maria_dev hdfs          0 2025-11-10 19:26 .Trash/Current/user
drwx-----  - maria_dev hdfs          0 2025-11-10 19:26 .Trash/Current/user/maria_dev
drwx-----  - maria_dev hdfs          0 2025-11-10 19:26 .Trash/Current/user/maria_dev/practicas
-rw-r--r--  1 maria_dev hdfs         16 2025-11-10 19:18 .Trash/Current/user/maria_dev/practicas/prueba.txt
drwxr-xr-x  - maria_dev hdfs          0 2025-11-10 19:17 datos
-rw-r--r--  1 maria_dev hdfs 1073741824 2025-11-10 18:23 datos/archivo_1GB
-rw-r--r--  1 maria_dev hdfs         16 2025-11-07 18:50 datos/prueba.txt
```