# Práctica 1: El almacenamiento HDFS

En este ejercicio nos familiarizaremos con el sistema de almacenamiento distribuido HDFS y algunos comandos para trabajar con él.

La topología de nuestra arquitectura es la siguiente.

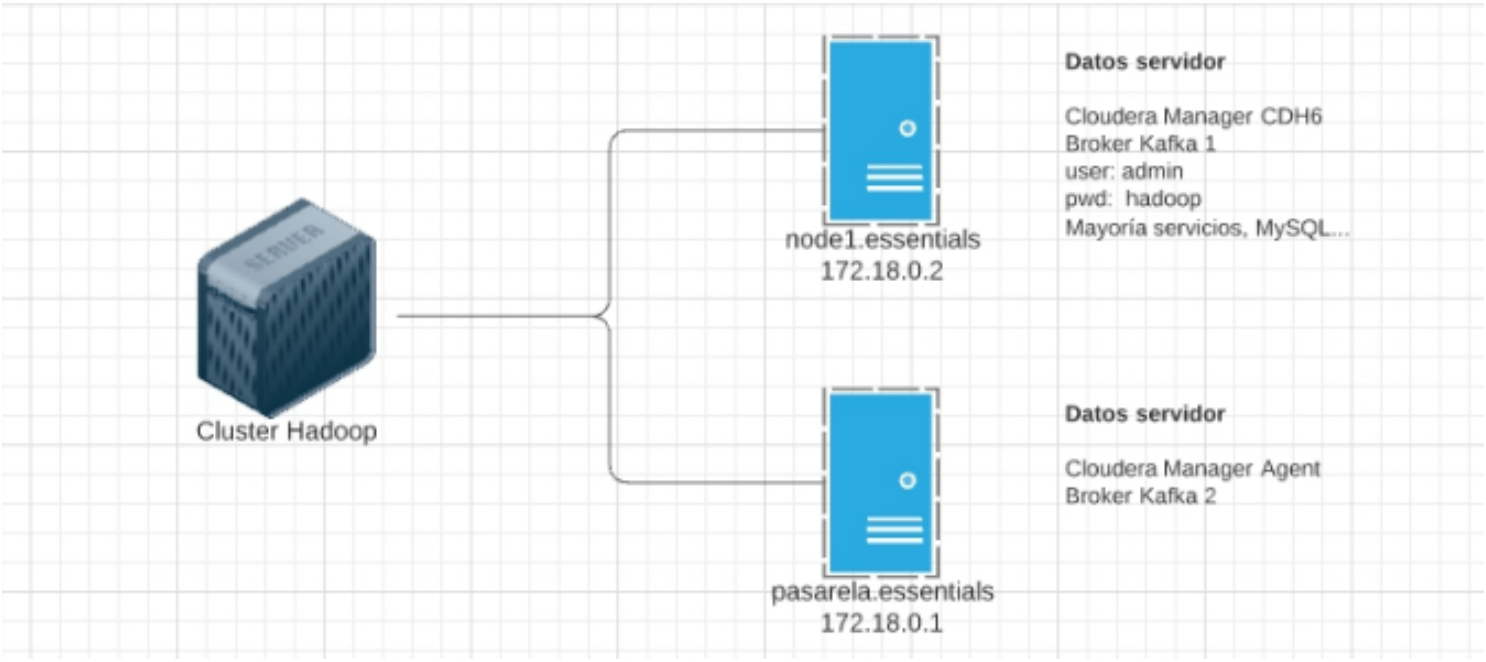


Figure 1: Topología del cluster Hadoop

### Familiarización con el entorno de trabajo

Todos los ejemplos de este ejercicio se ejecutan desde una consola abierta en pasarela.

```
[alumno@pasarela ~]$
```

- Desde Cloudera Manager, seleccionaremos **HDFS** y haremos clic en el enlace *Instances*.
- Confirmaremos que los servicios estan en ejecución y funcionan normalmente.

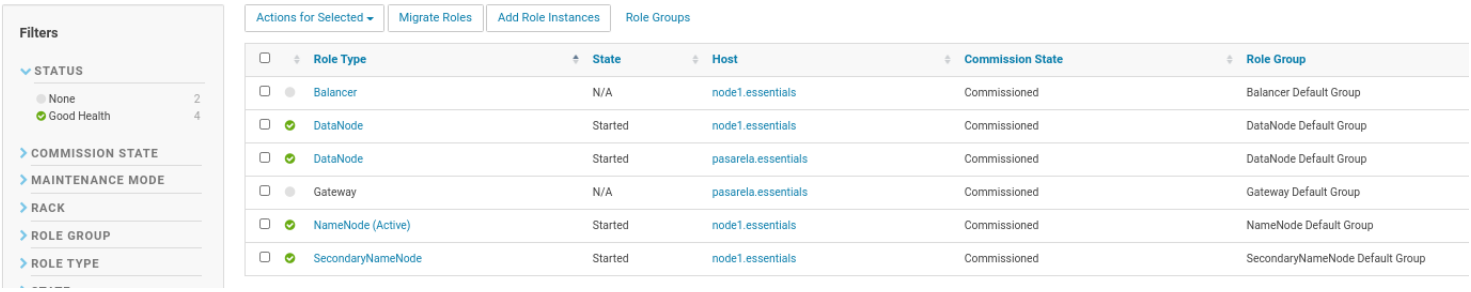


Figure 2: Instancias del cluster Hadoop

Si apareciese un error de **missing blocks** en el servicio HDFS, ejecutaremos los siguientes comandos. El proceso de solución, puede llevar unos minutos hasta que el servicio vuelve a su estado normal en color verde.

```
sudo -u hdfs hdfs dfsadmin -report
sudo -u hdfs hdfs fsck / -delete
```

Comprueba que los HDFS daemons (NameNode, SecondaryNameNode, y varios DataNodes) se están ejecutando correctamente en el cluster. El proceso Gateway no se muestra en color verde (en ejecucion) ya que es un intermediario, un enlace entre el cliente y los daemons.

Desde el servicio **HDFS** seleccionamos el enlace *Configuration*, y en el campo de busqueda que aparece escribimos la propiedad **dfs.blocksize**. El valor debiera ser **128 MB**.

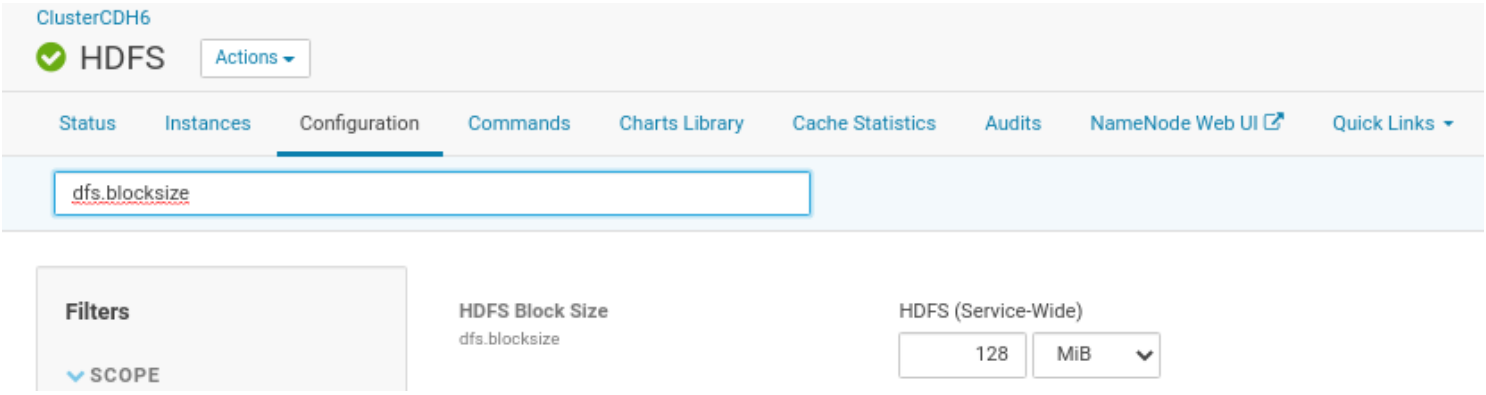


Figure 3: Tamaño del bloque de los archivos hdfs

- Subimos un archivo que supere 1MB de tamaño (*filmoteca.csv*) a HDFS,
- comprobamos que está alojado correctamente
- y obtenemos el valor del tamaño de bloque del archivo

```
[alumno@pasarela ~]$ hdfs dfs -put filmoteca.csv
[alumno@pasarela ~]$ hdfs dfs -ls
-rw-r--r--   2 alumno supergroup    2893226 2022-02-21 00:39 filmoteca.csv
[alumno@pasarela ~]$ hdfs dfs -stat %o filmoteca.csv
134217728
```

## Visualización del sistema HDFS con HUE

Se usa el navegador para abrir la aplicacion **HUE**. Se introduce la siguiente url:

http://node1.essentials:8889/

- Introducimos las credenciales que nos pide HUE:
  - usuario: **root**
  - password: hadoop
- A partir del menú con las tres rayas horizontales vamos a la opción *Browsers -> Files* o hacemos clic directamente sobre el icono con dos hojas de papel.

Por defecto aparece el área del usuario **root**, no *alumno*

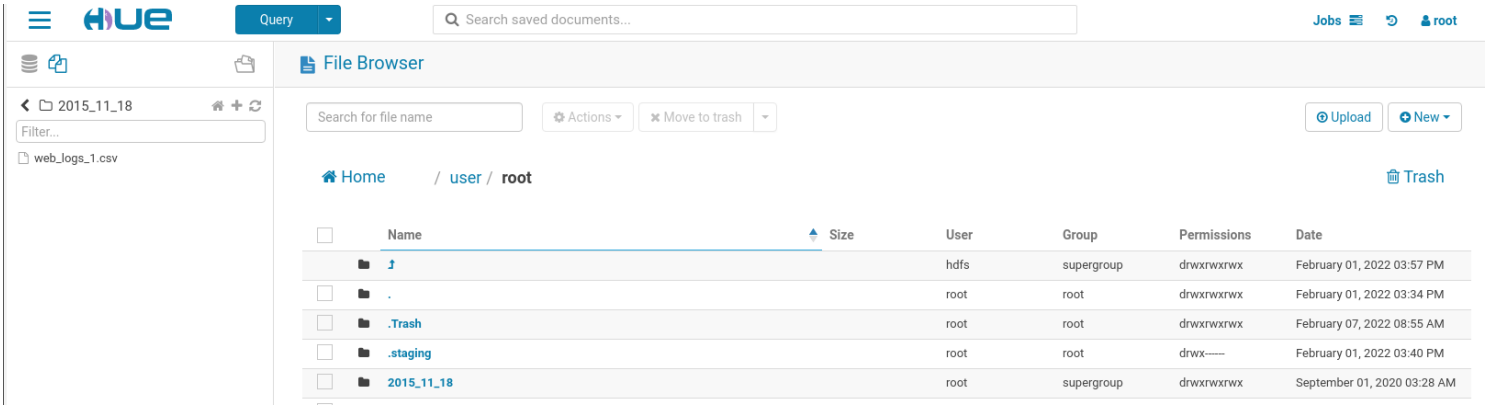


Figure 4: Archivos de root en HDFS vistos en HUE

### Creación de un usuario alumno

Vamos a crear un usuario **alumno** para poder trabajar correctamente.

Para ello, seleccionamos de la lista que aparece al hacer clic sobre el icono del usuario root de la esquina superior derecha, el comando *Manage Users*.

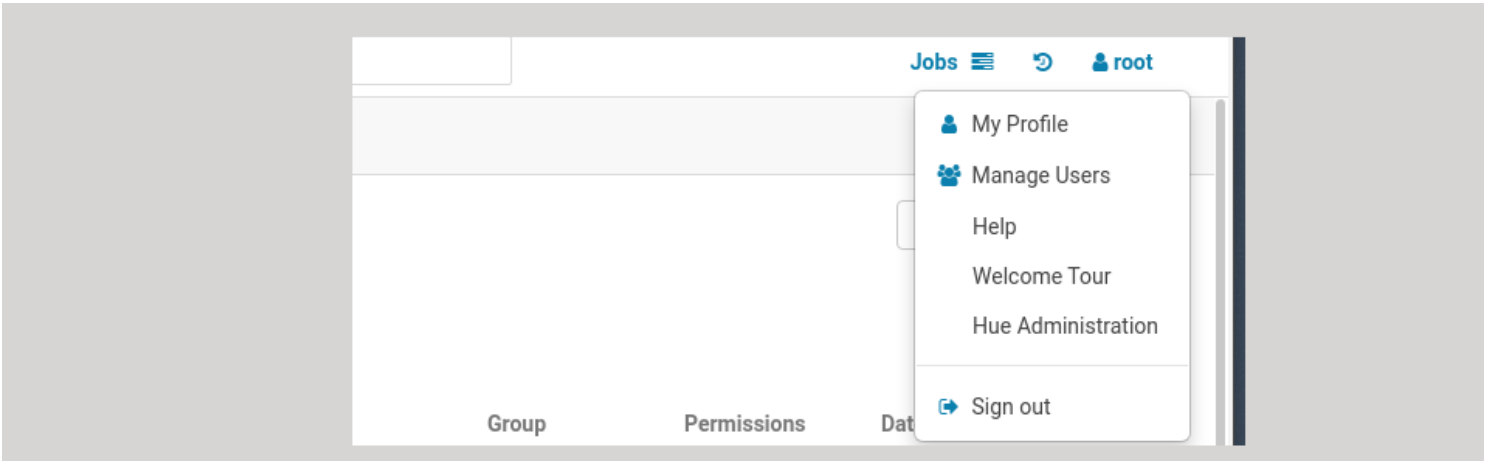


Figure 5: Menú para creación de usuarios

En la pantalla que aparece, hacemos clic sobre el botón *Add user*.

Introducimos “alumno” para el Username, y “@lumn0Clara” para los dos campos de Password y pulsamos el boton Add user.

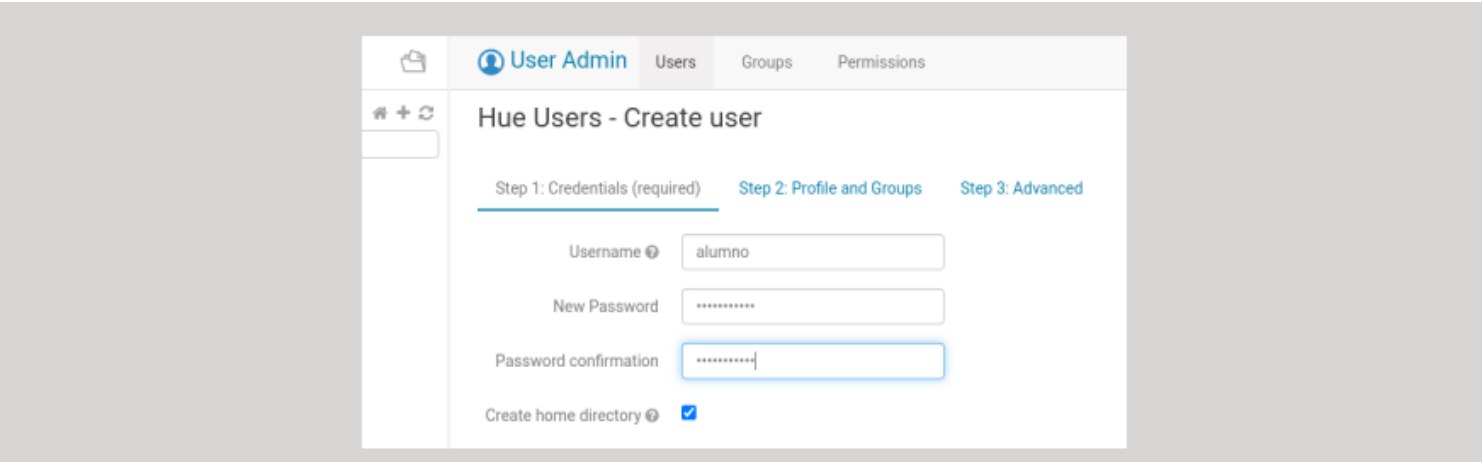


Figure 6: Creación del usuario en HUE

Ahora vamos a comprobar si funciona correctamente. Desde el menú de *root* seleccionamos el comando **Sign out** para salir.

En la pantalla de login introducimos las credenciales para el usuario *alumno* y en la pantalla de HUE, mostramos los archivos en HDFS. En este caso corresponden con el usuario *alumno*.

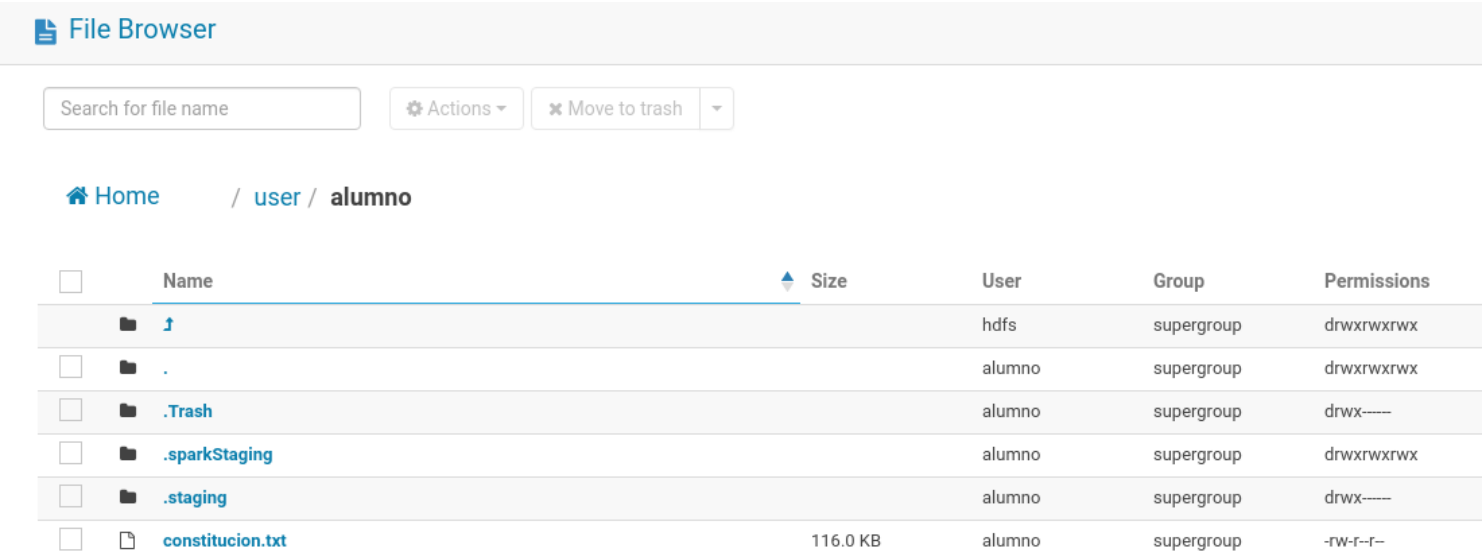


Figure 7: Archivos del usuario alumno en HUE

## Trabajando con archivos en HDFS

Vamos a guardar en HDFS un archivo de unos 250MB para ver el comportamiento del NameNode.

Esta operación podemos realizarla desde **HUE**:

- Primero descomprimimos el archivo *sfpd.tar.gz* en formato *tar.gz* para obtener el fichero *sfpd.json*.
- Situados en HUE en la carpeta */user/alumno*, agregamos el archivo descomprimido *sfpd.json*: Hacemos clic en el botón **Upload** en la esquina superior derecha del interfaz de HUE.
- Hacemos clic en el botón **Select files** del cuadro de diálogo que aparece
- Seleccionamos el archivo descomprimido *sfpd.json*

### Localización del archivo en HDFS

Mostramos la página del NameNode, colocando la siguiente URL en el navegador

http://node1.essentials:9870/

- Seleccionamos el comando **Browse the file system** del menú superior.
- Usando los enlaces de la parte derecha de la columna **Name**, nos situamos en */user/alumno*

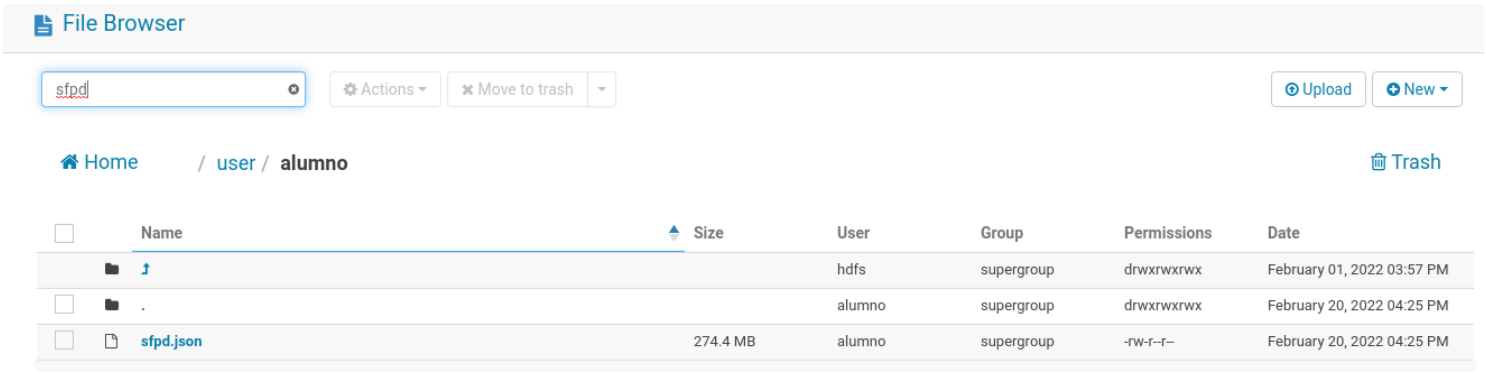


Figure 8: Archivos sfpd.json en la carpeta de alumno

- Hacemos clic sobre el enlace del archivo **sfpd.json** para mostrar la pantalla de metadata del archivo y comprobamos que se han creado 3 bloques para cada réplica del archivo.

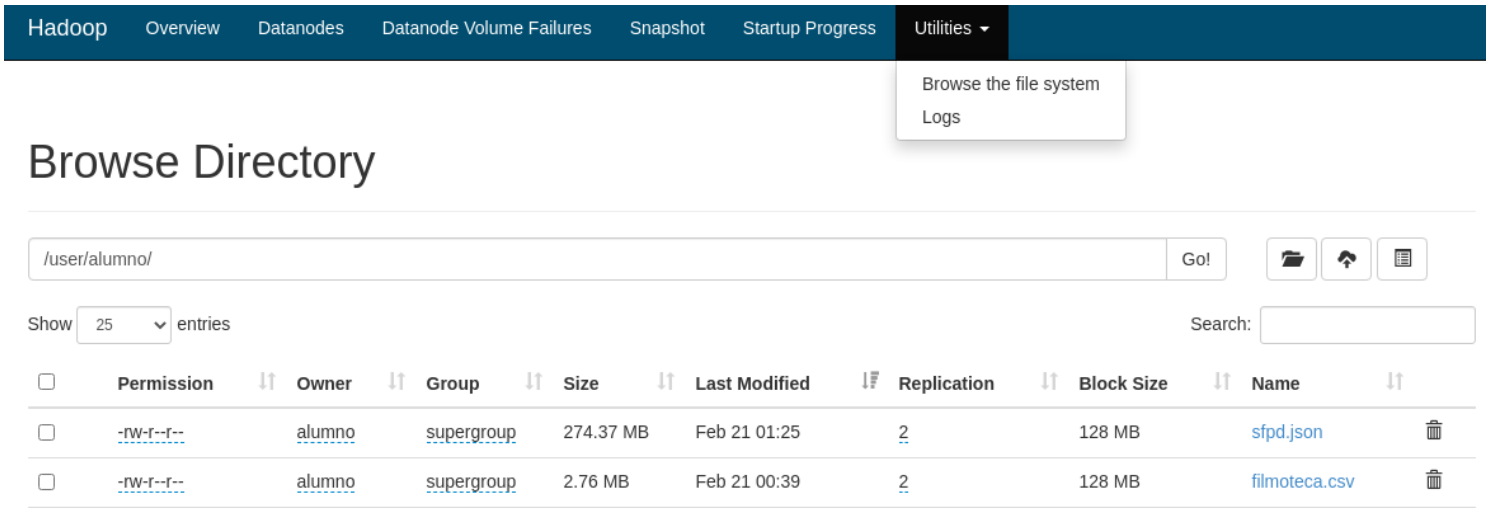


Figure 9: Archivos vistos desde NameNode

Cerramos pulsando sobre el botón **Close**.

**Recoge en un pantallazo la situación del fichero en HDFS** como muestra de que has hecho la práctica. Recuerda que se debe ver tu nombre en la imagen.


Figure 10: Bloques del archivo sfpd.json

Utilizando el id de uno de los bloques del archivo (posiblemente sea otro valor al mostrado aqui), lo localizamos en Linux (No en HDFS)

```
[alumno@pasarela Escritorio]$ sudo find / -name 'blk_1073748134'
[sudo] password for alumno:
find: ‘/run/user/1000/gvfs’: Permiso denegado
/containers/dfs/dn/current/BP-1483051556-172.18.0.2-1598871810253/current/finalized/subdir0/subdir24/
blk_1073748134
```

## Anexo: Ayuda del comando stat

```
stat
Usage: hadoop fs -stat [format] <path> ...
```

Print statistics about the file/directory at in the specified format. Format accepts filesize in blocks (%b), type (%F), group name of owner (%g), name (%n), block size (%o), replication (%r), user name of owner(%u), and modification date (%y, %Y). %y shows UTC date as "yyyy-MM-dd HH:mm:ss" and %Y shows milliseconds since January 1, 1970 UTC. If the format is not specified, %y is used by default.

*Example*:

```
hadoop fs -stat "%F %u:%g %b %y %n" /file
```

Exit Code: Returns 0 on success and -1 on error.