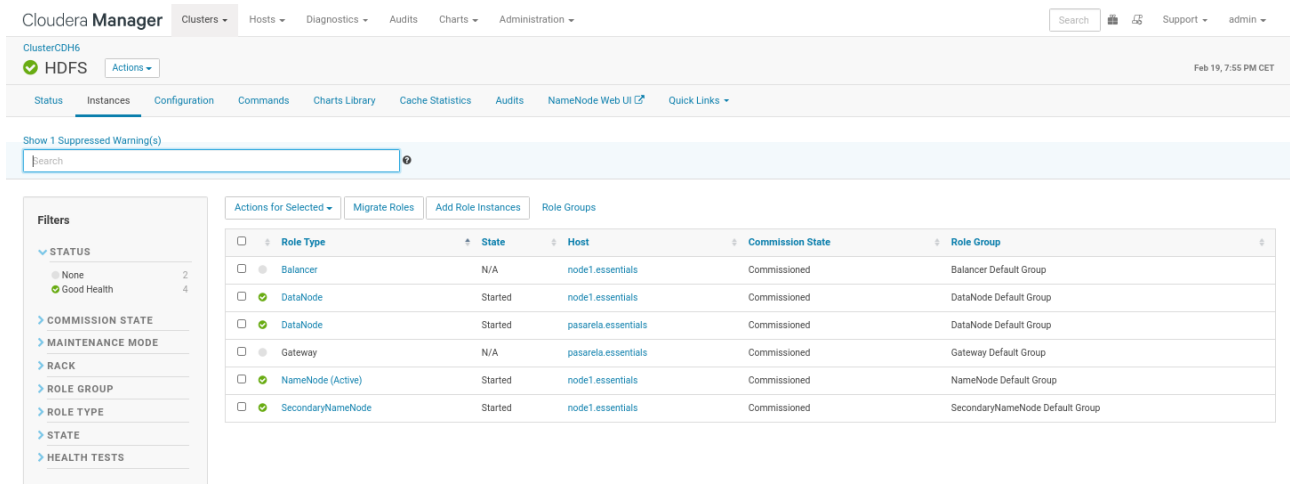


Práctica 1: El almacenamiento HDFS

Familiarización con el entorno de trabajo

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



The screenshot shows the Cloudera Manager interface for the 'ClusterCDH6' environment. The 'HDFS' service is selected, and the 'Instances' tab is active. A search bar at the top shows 'Show 1 Suppressed Warning(s)'. On the left, a 'Filters' sidebar shows 'STATUS' with 'None' (2) and 'Good Health' (4) options. The main table displays the following data:

Role Type	State	Host	Commission State	Role Group
Balancer	N/A	node1.essentials	Commissioned	Balancer Default Group
DataNode	Started	node1.essentials	Commissioned	DataNode Default Group
DataNode	Started	pasarela.essentials	Commissioned	DataNode Default Group
Gateway	N/A	pasarela.essentials	Commissioned	Gateway Default Group
NameNode (Active)	Started	node1.essentials	Commissioned	NameNode Default Group
SecondaryNameNode	Started	node1.essentials	Commissioned	SecondaryNameNode Default Group

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.

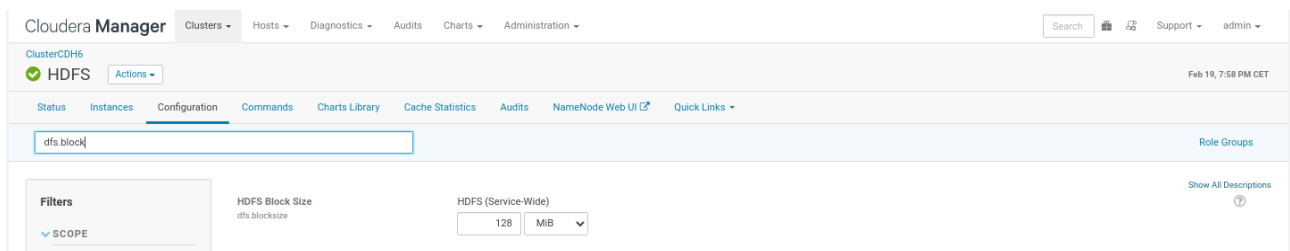
En mi caso no me aparece el error “missing blocks”, en el caso de tenerlo, tenemos que ejecutar los siguientes comandos:

- `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.

Efectivamente, el NameNode, SecondaryNameNode y los dos DataNodes están en verde (Good Health).

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



- 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

Para subir un archivo utilizamos el comando “hdfs dfs -put <archivo>”

Para listar un archivo utilizamos el comando “hdfs dfs -ls”

Para ver el tamaño del bloque donde está alojado el archivo utilizamos el comando “hdfs dfs -stat %O <archivo>”

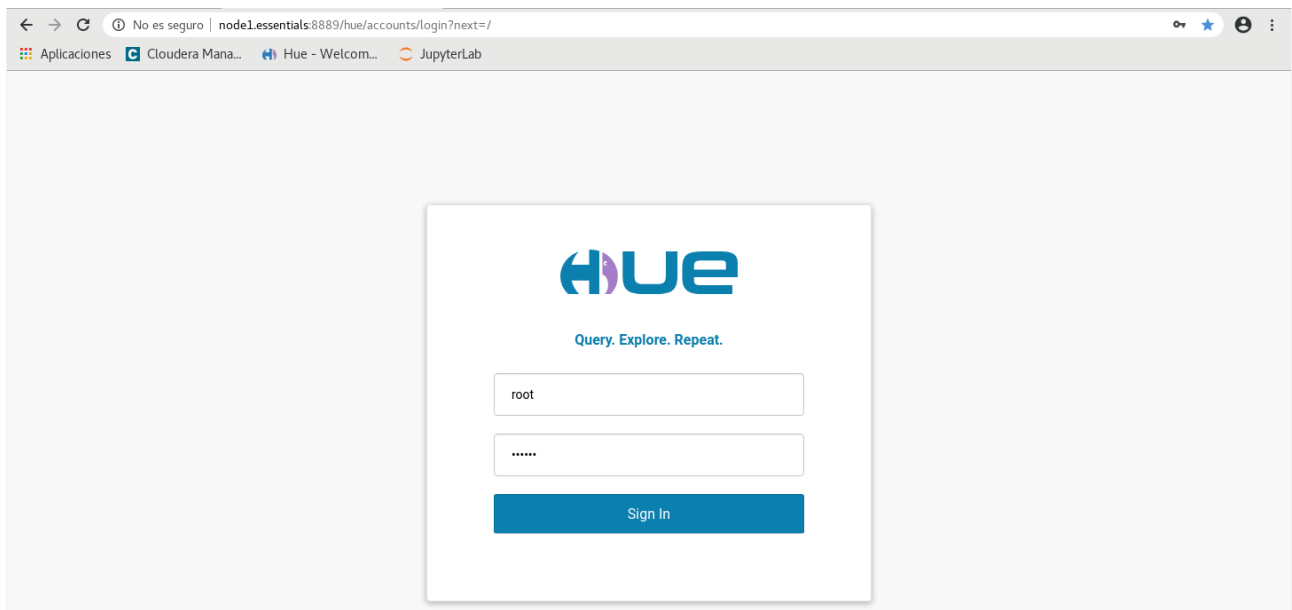
```
[alumno@pasarela datos_pr1]$ hdfs dfs -put filmoteca.csv
[alumno@pasarela datos_pr1]$ hdfs dfs -ls
Found 3 items
drwxrwxrwx - alumno supergroup 0 2021-09-15 10:25 .sparkStaging
-rwxrwxrwx 2 alumno supergroup 2221 2021-09-09 17:54 devices.csv
-rw-r--r-- 2 alumno supergroup 2893226 2024-02-19 20:05 filmoteca.csv
[alumno@pasarela datos_pr1]$ hdfs dfs -stat %o filmoteca.csv
134217728
```

Visualización del sistema HDFS con HUE

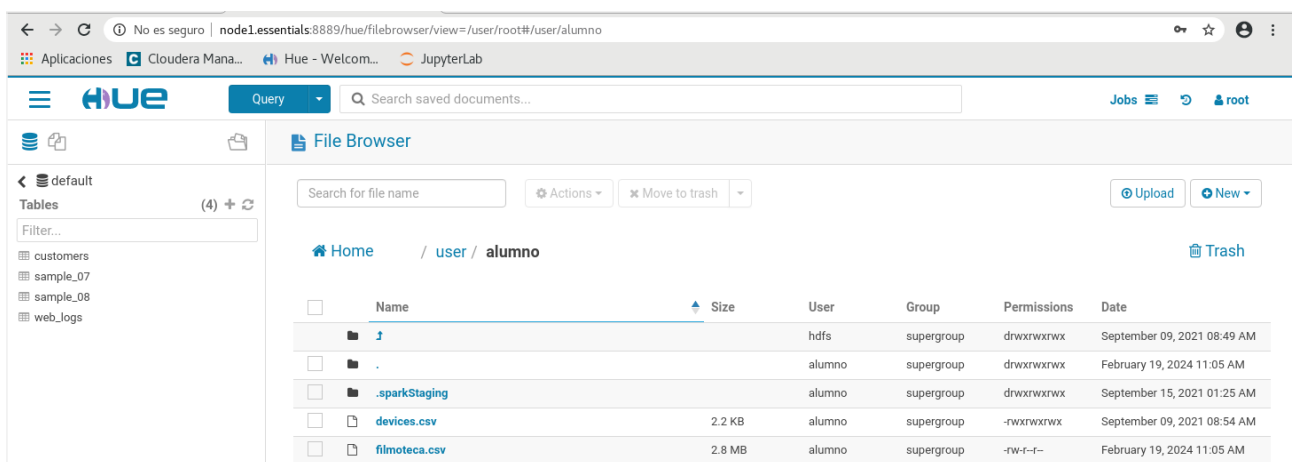
Se usa el navegador para abrir la aplicación HUE. Se introduce la siguiente url:

<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

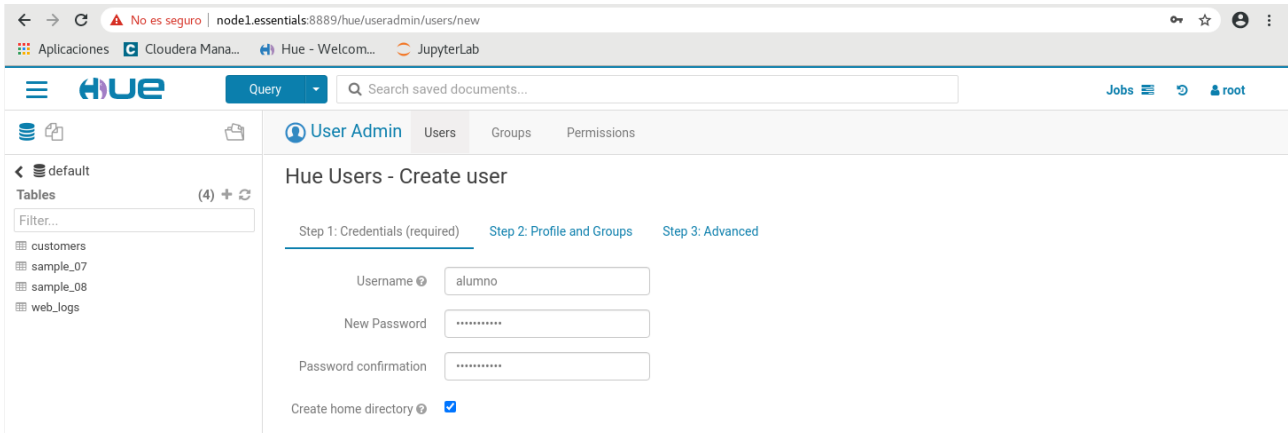


En este caso he navegado a la ruta del usuario alumno.

Creación de un usuario alumno

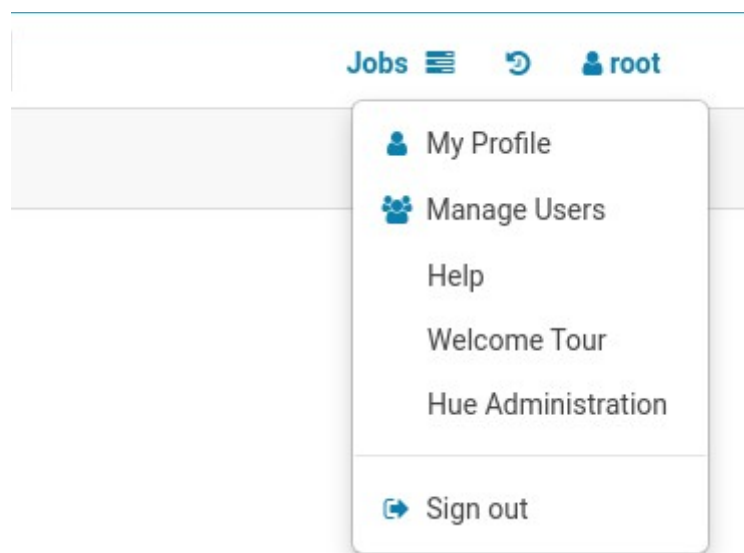
Vamos a crear un usuario alumno para poder trabajar correctamente. Para ello, seleccionamos de la lista que aparece al hacer clic sobre el icono del usuario root de la esquina superior derecha, el comando Manage Users.

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.

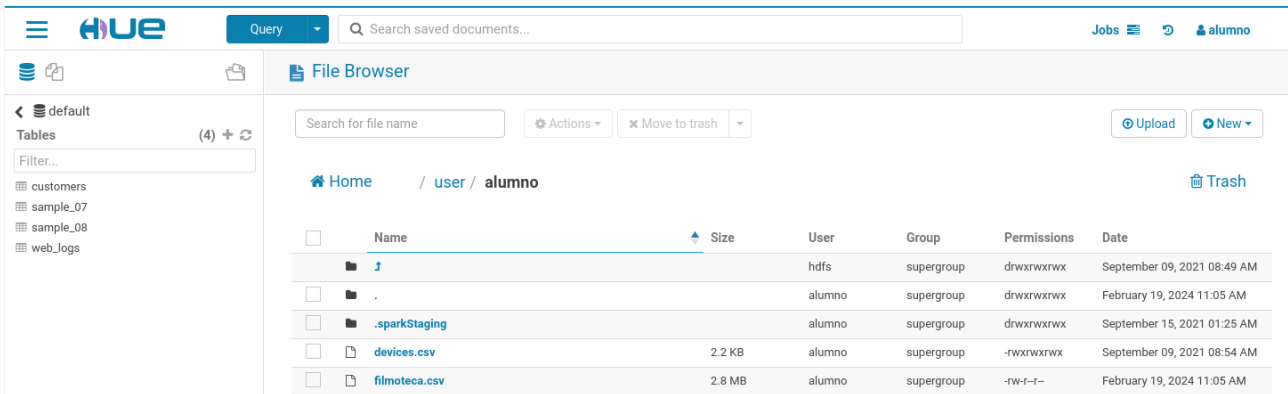


The screenshot shows the Hue web interface. The browser address bar indicates the URL: node1.essentials.8889/hue/useradmin/users/new. The interface has a top navigation bar with 'Jobs', a refresh icon, and a user profile icon labeled 'root'. Below this is a 'User Admin' section with tabs for 'Users', 'Groups', and 'Permissions'. The 'Users' tab is active, showing 'Hue Users - Create user'. The page is divided into three steps: 'Step 1: Credentials (required)', 'Step 2: Profile and Groups', and 'Step 3: Advanced'. In Step 1, there are input fields for 'Username' (containing 'alumno'), 'New Password' (masked with dots), and 'Password confirmation' (masked with dots). There is also a checkbox for 'Create home directory' which is checked. On the left side, there is a sidebar with a 'Tables' section showing a list of tables: 'customers', 'sample_07', 'sample_08', and 'web_logs'.

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.



Name	Size	User	Group	Permissions	Date
hdfs		hdfs	supergroup	drwxrwxrwx	September 09, 2021 08:49 AM
.		alumno	supergroup	drwxrwxrwx	February 19, 2024 11:05 AM
.sparkStaging		alumno	supergroup	drwxrwxrwx	September 15, 2021 01:25 AM
devices.csv	2.2 KB	alumno	supergroup	-rwxrwxrwx	September 09, 2021 08:54 AM
filmoteca.csv	2.8 MB	alumno	supergroup	-rw-r--r--	February 19, 2024 11:05 AM

En este caso me aparece directamente el directorio del alumno sin necesidad de tener que navegar hacia el.

Trabajando con archivos en HDFS

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

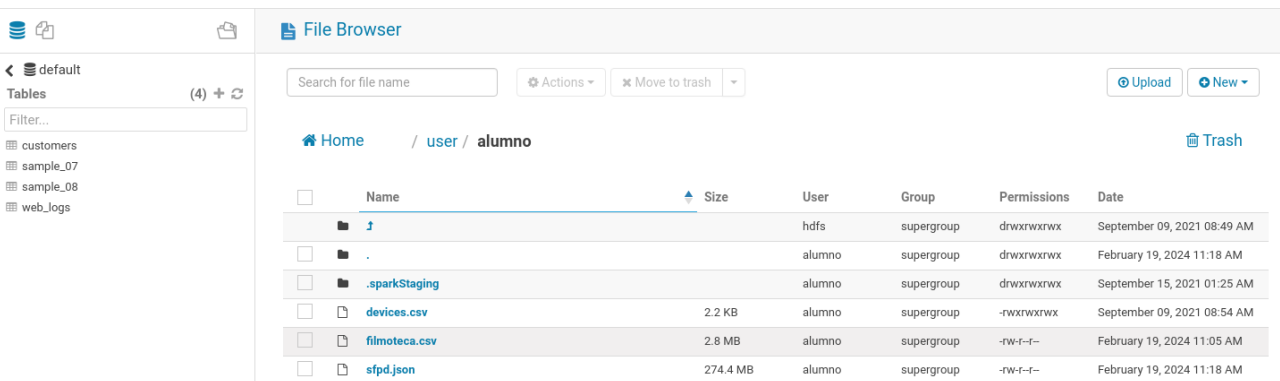
Esta operación podemos realizarla desde HUE:

- Primero descomprimos el archivo sfpd.tar.gz en formato tar.gz para obtener el fichero sfpd.json.

Para descomprimir “tar -xvzf <archivo>”

```
[alumno@pasarela datos_pr1]$ tar -xvzf sfpd.tar.gz
sfpd.json
[alumno@pasarela datos_pr1]$ ls
constitucion.txt demo.parquet distribuidores.parquet filmoteca.csv paises.avro sfpd.json sfpd.tar.gz
[alumno@pasarela datos_pr1]$
```

- 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



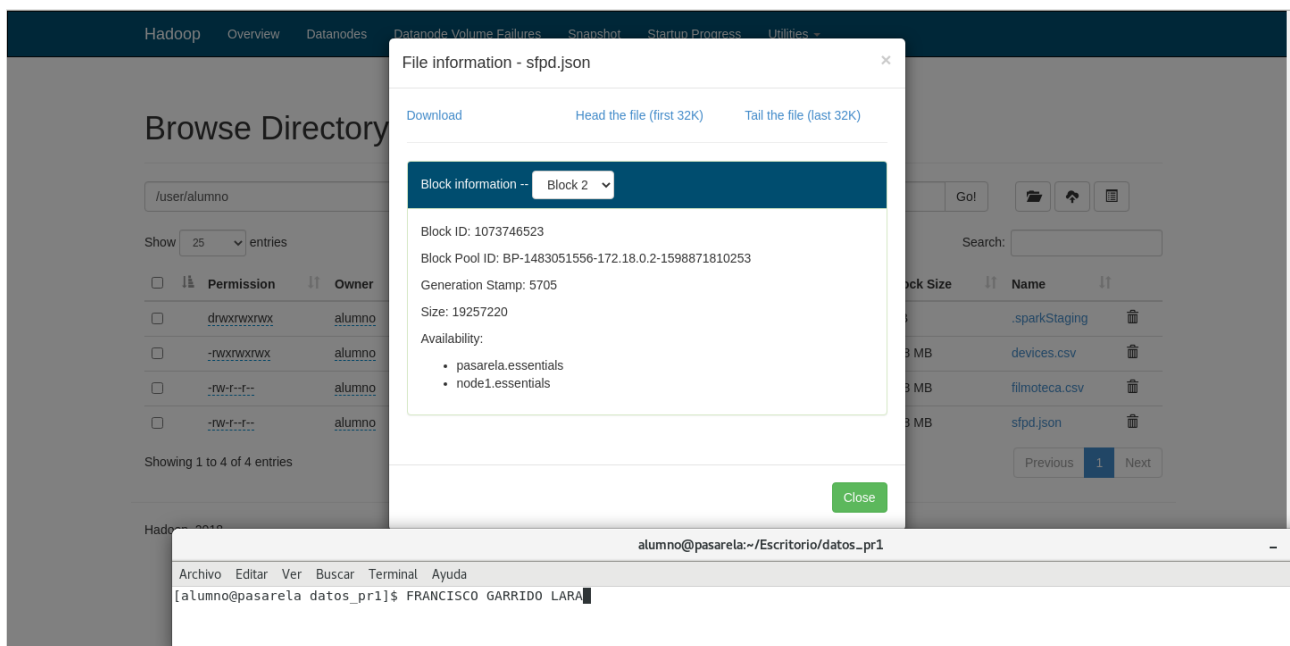
Name	Size	User	Group	Permissions	Date
hdfs		hdfs	supergroup	drwxrwxrwx	September 09, 2021 08:49 AM
.		alumno	supergroup	drwxrwxrwx	February 19, 2024 11:18 AM
.sparkStaging		alumno	supergroup	drwxrwxrwx	September 15, 2021 01:25 AM
devices.csv	2.2 KB	alumno	supergroup	-rwxrwxrwx	September 09, 2021 08:54 AM
filmoteca.csv	2.8 MB	alumno	supergroup	-rw-r--r--	February 19, 2024 11:05 AM
sfpd.json	274.4 MB	alumno	supergroup	-rw-r--r--	February 19, 2024 11:18 AM

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

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.



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 datos_pr1]$ sudo find / -name 'blk_1073746523'
[sudo] password for alumno:
find: '/run/user/1000/gvfs': Permiso denegado
/var/lib/docker/containers/dfs/dn/current/BP-1483051556-172.18.0.2-1598871810253/current/finalized/subdir0/subdir18/blk_1073746523
/containers/dfs/dn/current/BP-1483051556-172.18.0.2-1598871810253/current/finalized/subdir0/subdir18/blk_1073746523
[alumno@pasarela datos_pr1]$
```