Universidad Eafit

Tutorial: Manejo de HDFS en Amazon EMR

1. Conexión al Clúster de Amazon EMR:

Inicia una conexión segura utilizando SSH con tu llave de acceso:

`

ssh -i [ruta-de-tu-llave.pem] hadoop@[dirección-IP-del-master-node]

Ejemplo de conexión:

ssh -i bigdata-key-pair.pem hadoop@ec2-100-26-198-124.compute-1.amazonaws.com

2. Operaciones con Archivos en HDFS Mediante la Terminal:

Preparación de Datos:

Prepara un espacio en HDFS para tus datasets creando un directorio llamado 'datasets' dentro de tu directorio home (/user/hadoop).

Asegúrate de que tus datos ya estén descomprimidos y listos para su procesamiento.

Ejemplo de Datasets y Comandos Básicos:

Para obtener los datasets, clona el repositorio con el siguiente comando:

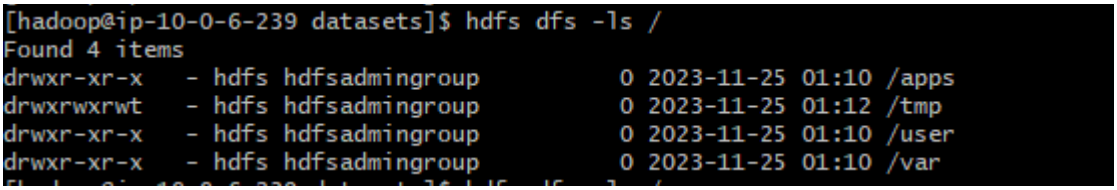
git clone https://github.com/hfbanilatq/hfbanilatq-st0263

Navega al directorio de datasets:

cd hfbanilatq-st0263/reto3.1/

Lista los archivos presentes en el directorio raíz de HDFS:

hdfs dfs -ls /

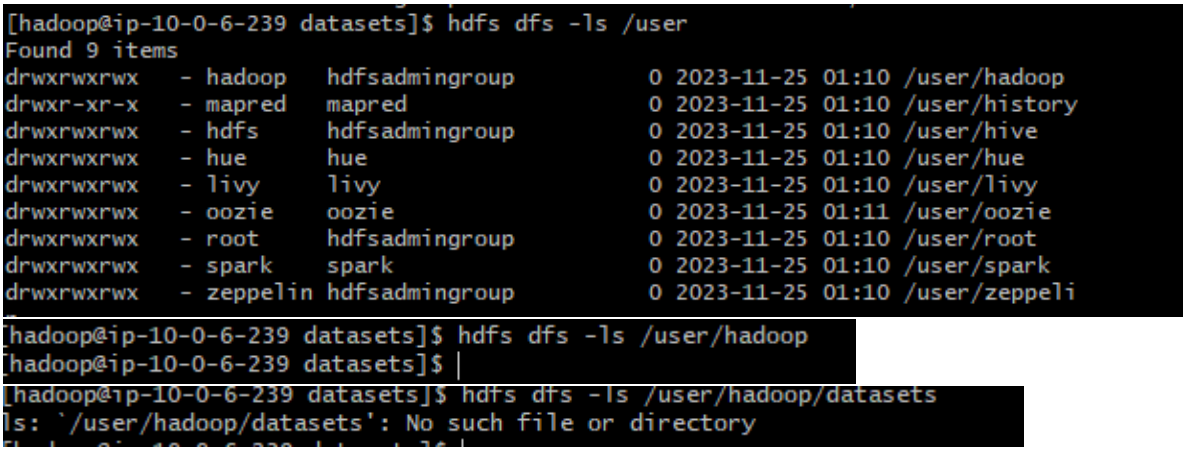


Inspecciona los contenidos de otros directorios relevantes con comandos similares:

hdfs dfs -ls /user

hdfs dfs -ls /user/hadoop

hdfs dfs -ls /user/hadoop/datasets



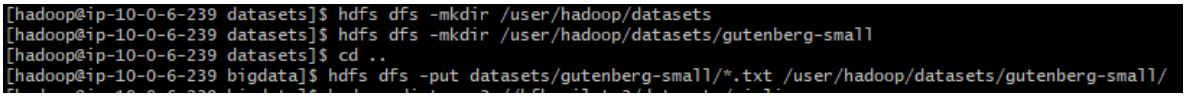
Gestión de tu Directorio de Datasets:

Crea tu directorio 'datasets' en HDFS con:

hdfs dfs -mkdir /user/hadoop/datasets

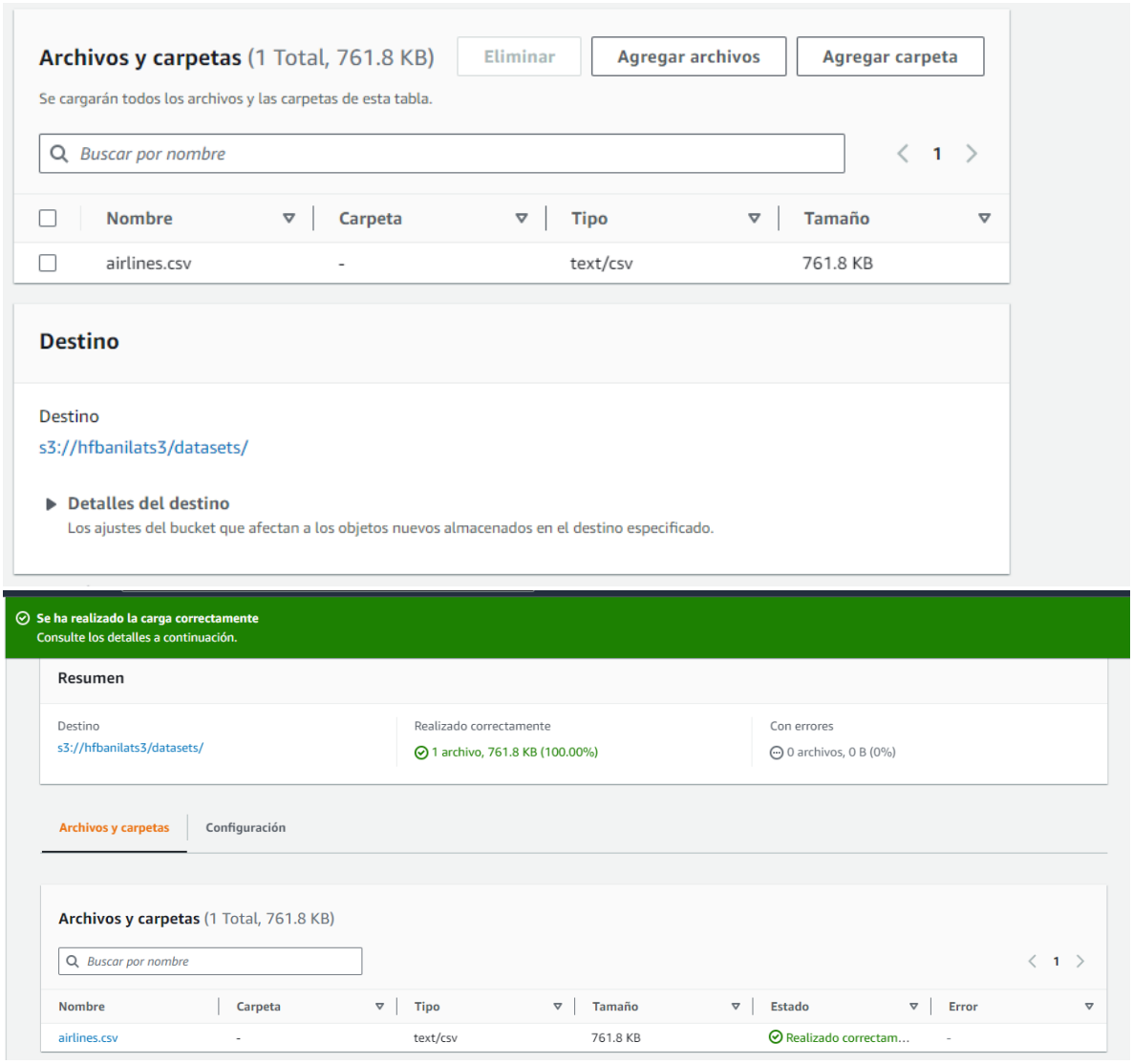
Transfiere tus archivos desde el servidor local al directorio HDFS que acabas de crear. Por ejemplo:

hdfs dfs -put [ruta-local] /user/hadoop/datasets/



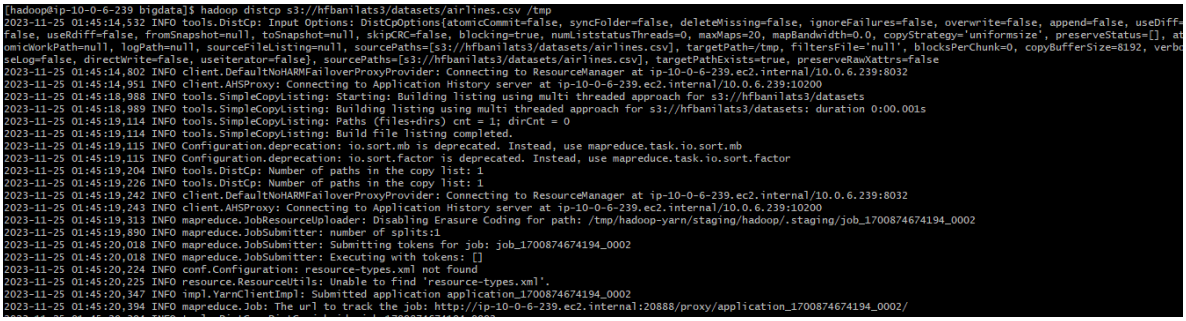
Si tus datos están en Amazon S3, puedes transferirlos directamente con:

Primero debes subir los archivos a S3 para eso emplea la consola de aws, para crear un carpeta en el bucket de s3 llamado datasets, luego sube el archivo airline.csv



Luego emplea:

hadoop distcp s3://[ruta-en-s3] /destino-en-hdfs



Para una copia de datos recursiva utiliza:

hdfs dfs -copyFromLocal [ruta-local] /user/hadoop/datasets/

```
^C[hadoop@ip-10-0-6-239 bigdata]$ hdfs dfs -copyFromLocal datasets/* /user/hadoop/datasets/
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnLetters.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsFirstInauguralAddress.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsGettysburgAddressGivenNovember-19-1863.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsInauguralsAddressesandLettersSelections.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsSecondInauguralAddress.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___SpeechesandLettersofAbrahamLincoln1832-1865.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___StateoftheUnionAddresses.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheEmancipationProclamation.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume1.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume2.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume3.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume4.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume5.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume6.txt': File exists
copyFromLocal: `/user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume7.txt': File exists
```

Lista los archivos en tu directorio 'datasets' para verificar la transferencia:

hdfs dfs -ls /user/hadoop/datasets

```
-rw-r--r--   1 hadoop hdfsadmingroup     780058 2023-11-25 01:49 /user/hadoop/datasets/airlines.csv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/all-news
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/covid19
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/flights
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/gutenberg
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/onu
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/otros
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/retail_logs
-rw-r--r--   1 hadoop hdfsadmingroup        534 2023-11-25 01:49 /user/hadoop/datasets/sample_data.csv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-25 01:49 /user/hadoop/datasets/spark
[hadoop@ip-10-0-6-239 bigdata]$ hdfs dfs -ls /user/hadoop/datasets/gutenberg-small/
Found 16 items
-rw-r--r--   1 hadoop hdfsadmingroup       5878 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnLetters.txt
-rw-r--r--   1 hadoop hdfsadmingroup      21586 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsFirstInauguralAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup       1653 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsGettysburgAddressGivenNovember-19-1863.txt
-rw-r--r--   1 hadoop hdfsadmingroup     262083 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsInauguralsAddressesandLettersSelections.txt
-rw-r--r--   1 hadoop hdfsadmingroup       4093 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___LincolnsSecondInauguralAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup     516298 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___SpeechesandLettersofAbrahamLincoln1832-1865.txt
-rw-r--r--   1 hadoop hdfsadmingroup     167895 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___StateoftheUnionAddresses.txt
-rw-r--r--   1 hadoop hdfsadmingroup       3928 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheEmancipationProclamation.txt
-rw-r--r--   1 hadoop hdfsadmingroup      45664 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup     459006 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume1.txt
-rw-r--r--   1 hadoop hdfsadmingroup     505150 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume2.txt
-rw-r--r--   1 hadoop hdfsadmingroup     254941 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume3.txt
-rw-r--r--   1 hadoop hdfsadmingroup     209643 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume4.txt
-rw-r--r--   1 hadoop hdfsadmingroup     692051 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume5.txt
-rw-r--r--   1 hadoop hdfsadmingroup     601102 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume6.txt
-rw-r--r--   1 hadoop hdfsadmingroup     478689 2023-11-25 01:32 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume7.txt
```

Recuperación de Archivos desde HDFS:

Para copiar archivos de vuelta al servidor local, usa:

hdfs dfs -get /user/hadoop/datasets/* ~/[tu-username]/mis_datasets/

Otra opción para recuperar archivos es:

hdfs dfs -copyToLocal /user/hadoop/datasets/[nombre-del-archivo] ~/[tu-username]/mis_datasets/

Verifica la transferencia con:

ls -l ~/[tu-username]/mis_datasets

```
[hadoop@ip-10-0-6-239 ~]$ hdfs dfs -get /user/hadoop/datasets/gutenberg-small/* ~/hfbanilatq/mis_dataset/
[hadoop@ip-10-0-6-239 ~]$ hdfs dfs -copyToLocal /user/hadoop/datasets/gutenberg/gutenberg-small.zip ~/hfbanilatq/mis_dataset/
copyToLocal: `/user/hadoop/datasets/gutenberg/gutenberg-small.zip': No such file or directory
[hadoop@ip-10-0-6-239 ~]$ cd hfbanilatq/
[hadoop@ip-10-0-6-239 hfbanilatq]$ ls -l mis_datasets
ls: cannot access mis_datasets: No such file or directory
[hadoop@ip-10-0-6-239 hfbanilatq]$ ls -l mis_dataset
total 4160
-rw-r--r-- 1 hadoop hadoop   5878 Nov 25 01:52 AbrahamLincoln___LincolnLetters.txt
-rw-r--r-- 1 hadoop hadoop  21586 Nov 25 01:52 AbrahamLincoln___LincolnsFirstInauguralAddress.txt
-rw-r--r-- 1 hadoop hadoop   1653 Nov 25 01:52 AbrahamLincoln___LincolnsGettysburgAddressGivenNovember-19-1863.txt
-rw-r--r-- 1 hadoop hadoop 262083 Nov 25 01:52 AbrahamLincoln___LincolnsInauguralsAddressesandLettersSelections.txt
-rw-r--r-- 1 hadoop hadoop   4093 Nov 25 01:52 AbrahamLincoln___LincolnsSecondInauguralAddress.txt
-rw-r--r-- 1 hadoop hadoop 516298 Nov 25 01:52 AbrahamLincoln___SpeechesandLettersofAbrahamLincoln1832-1865.txt
-rw-r--r-- 1 hadoop hadoop 167895 Nov 25 01:52 AbrahamLincoln___StateoftheUnionAddresses.txt
-rw-r--r-- 1 hadoop hadoop   3928 Nov 25 01:52 AbrahamLincoln___TheEmancipationProclamation.txt
-rw-r--r-- 1 hadoop hadoop  45664 Nov 25 01:52 AbrahamLincoln___TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt
-rw-r--r-- 1 hadoop hadoop 459006 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume1.txt
-rw-r--r-- 1 hadoop hadoop 505150 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume2.txt
-rw-r--r-- 1 hadoop hadoop 254941 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume3.txt
-rw-r--r-- 1 hadoop hadoop 209643 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume4.txt
-rw-r--r-- 1 hadoop hadoop 692051 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume5.txt
-rw-r--r-- 1 hadoop hadoop 601102 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume6.txt
-rw-r--r-- 1 hadoop hadoop 478689 Nov 25 01:52 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume7.txt
[hadoop@ip-10-0-6-239 hfbanilatq]$
```

Experimenta con otros comandos:

Aplica diversos comandos de HDFS para gestionar tus archivos, como:

hdfs dfs -du [ruta] # Muestra el uso de disco

hdfs dfs -mv [origen] [destino] # Mueve archivos

hdfs dfs -cp [origen] [destino] # Copia archivos

hdfs dfs -rm [ruta] # Elimina archivos

hdfs dfs -put [src-local] [dest-hdfs] # Copia local a HDFS

hdfs dfs -cat [nombre-de-archivo] # Muestra contenido de archivo

```
[hadoop@ip-10-0-6-239 hfbanilatq]$ hdfs dfs -cat /user/hadoop/datasets/airline.csv
cat: `/user/hadoop/datasets/airline.csv': No such file or directory
[hadoop@ip-10-0-6-239 hfbanilatq]$ /user/hadoop/datasets/airlines.csv
-bash: /user/hadoop/datasets/airlines.csv: No such file or directory
[hadoop@ip-10-0-6-239 hfbanilatq]$ hdfs dfs -cat /user/hadoop/datasets/airlines.csv
id,airline,date,location,rating,cabin,value,recommended,review
10001,Delta Air Lines,21-Jun-14,Thailand,7,Economy,4,YES,Flew Mar 30 NRT to BKK. All flights were great. Flight was on-time and the in-flight entertainment was great. Apart from the meals - so
me Thai passengers cannot eat beef so the flight crews tried to ask other passengers who could eat beef and changed the meals around. We feel disappointed with their food services.
10002,Delta Air Lines,19-Jun-14,USA,0,Economy,2,NO,Flight 2463 leaving West Palm Beach (PBI) at 2.42pm on June 15 arriving at New York LaGuardia (LGA) at 5.30pm. I was slated to take flight 39
97 from LaGuardia to Pittsburgh (PIT) leaving at 6.59pm arriving in Pittsburgh at 8.35pm. Every time I fly Delta I end up getting stuck in whatever airport my layover is for 12+ hours due to c
ancellations or delays. There is no attempt to accommodate me the customer with a hotel for the night a partial refund or even a blanket or pillow for my extended stay in the airport.
10003,Delta Air Lines,18-Jun-14,USA,0,Economy,1,NO,Delta Website froze 4 times trying to set up flight to three different locations. Had to call and set up over the phone only to find out at t
```

Además, gestiona los permisos y propiedad de los archivos con:

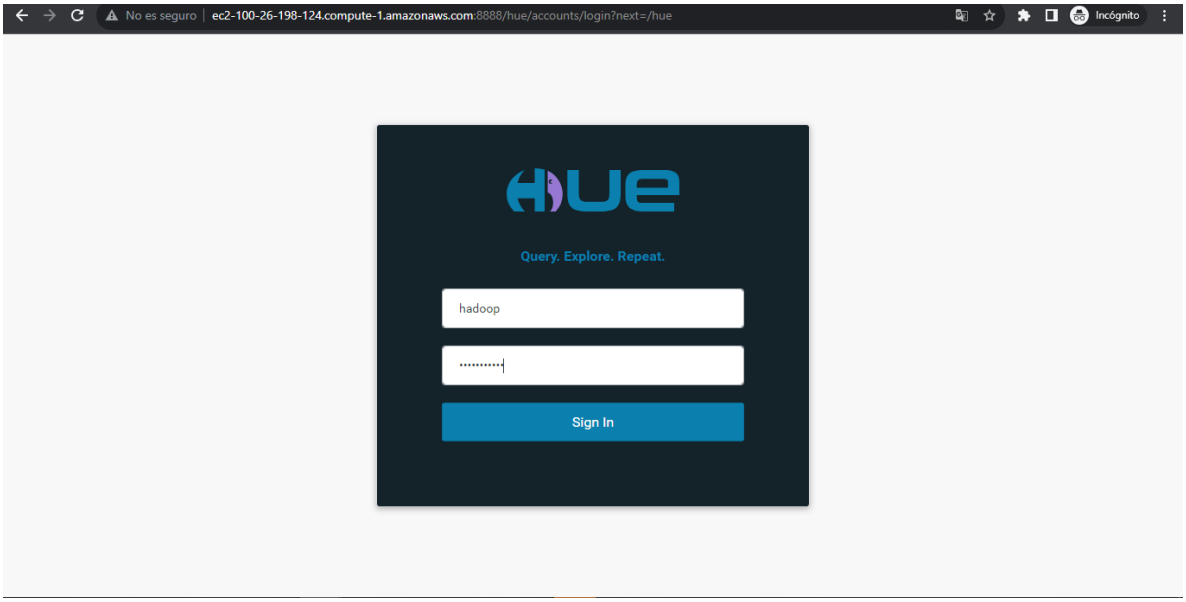
hdfs dfs -chmod [-R] [modo] [archivo]

hdfs dfs -chown [username] [archivo]

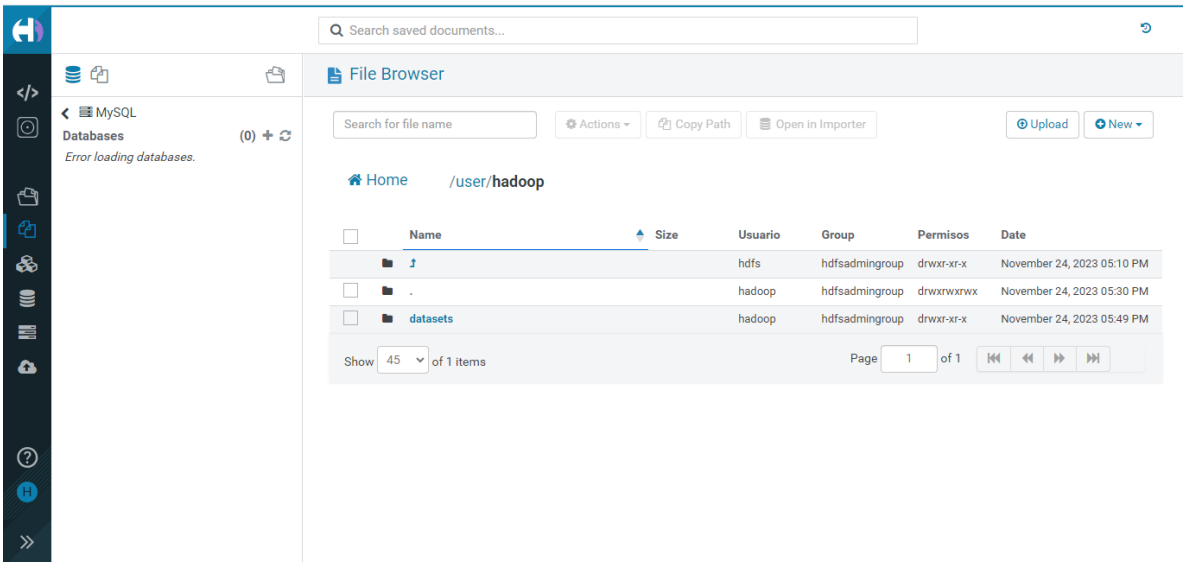
hdfs dfs -chgrp [grupo] [archivo]

3. Administración de Archivos a través de HUE en Amazon EMR:

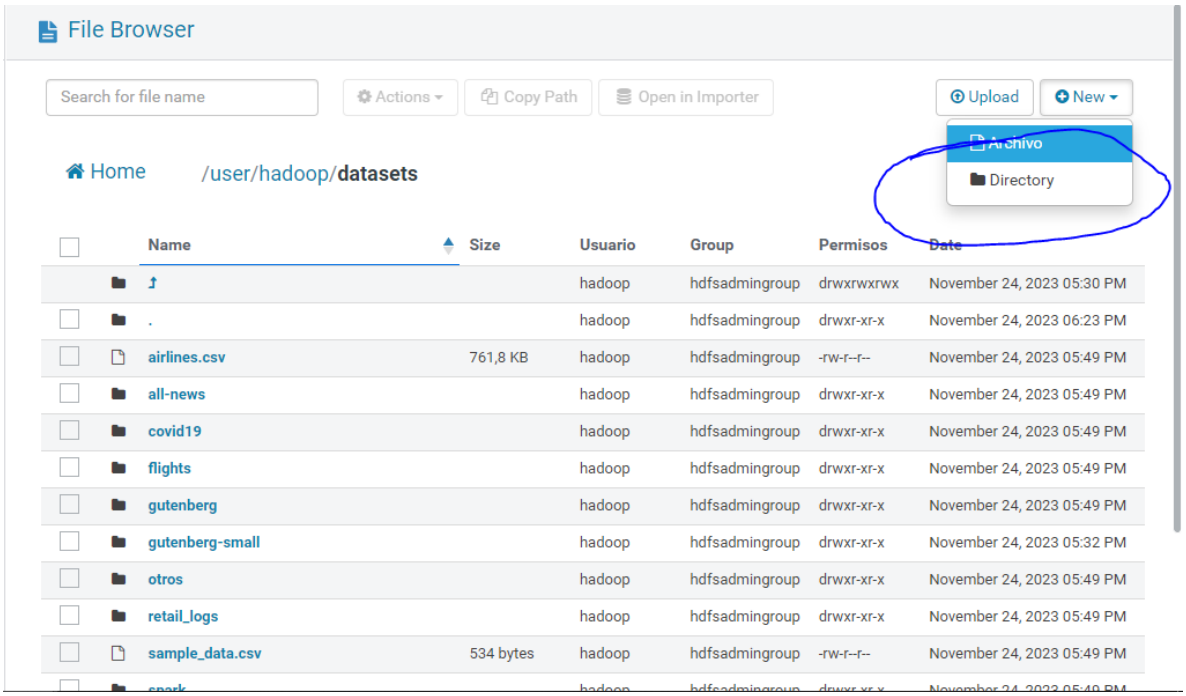
Inicia sesión en la interfaz de HUE para una gestión más visual de los archivos.



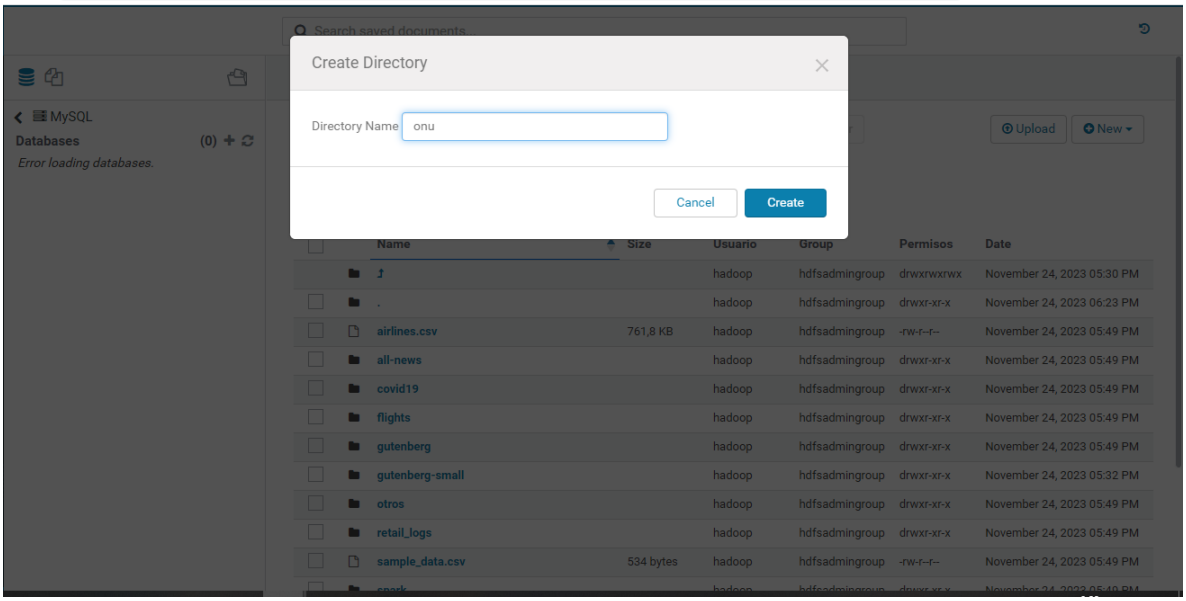
Navega por el sistema de archivos, crea directorios y sube archivos directamente desde la interfaz de HUE.



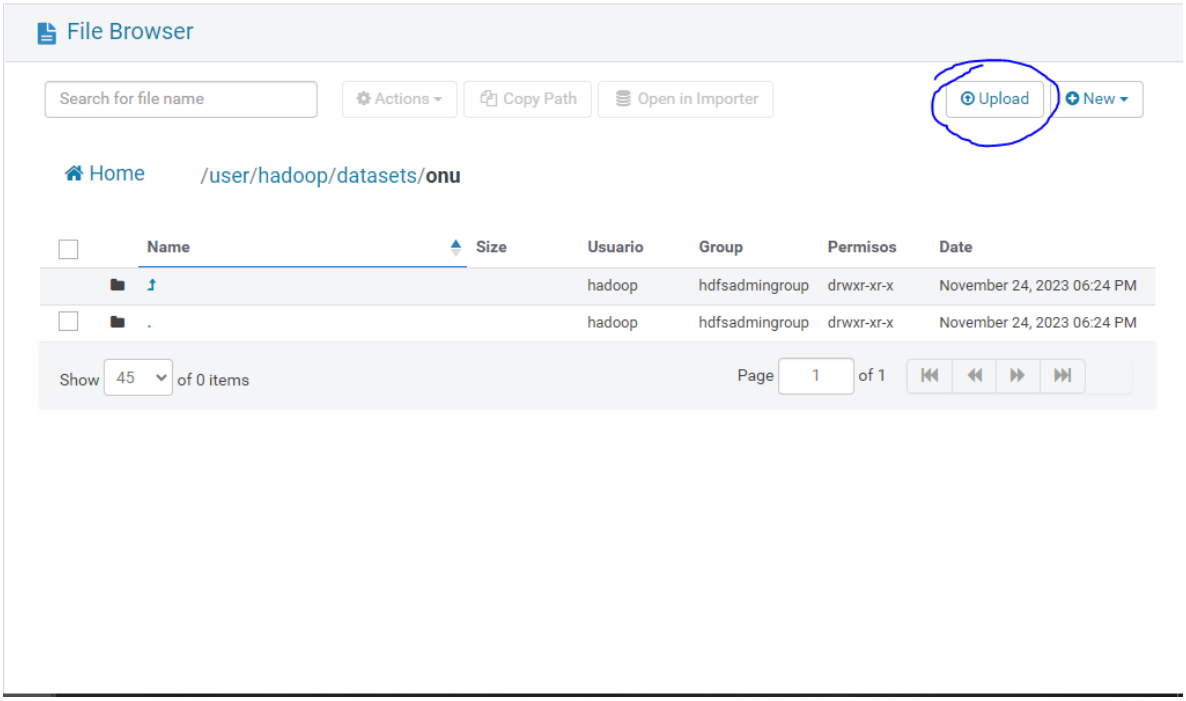
Como ya subimos todos los archivos usando la terminal, voy a eliminar la carpeta ONU para crearla nuevamente y subir los archivos en esta.

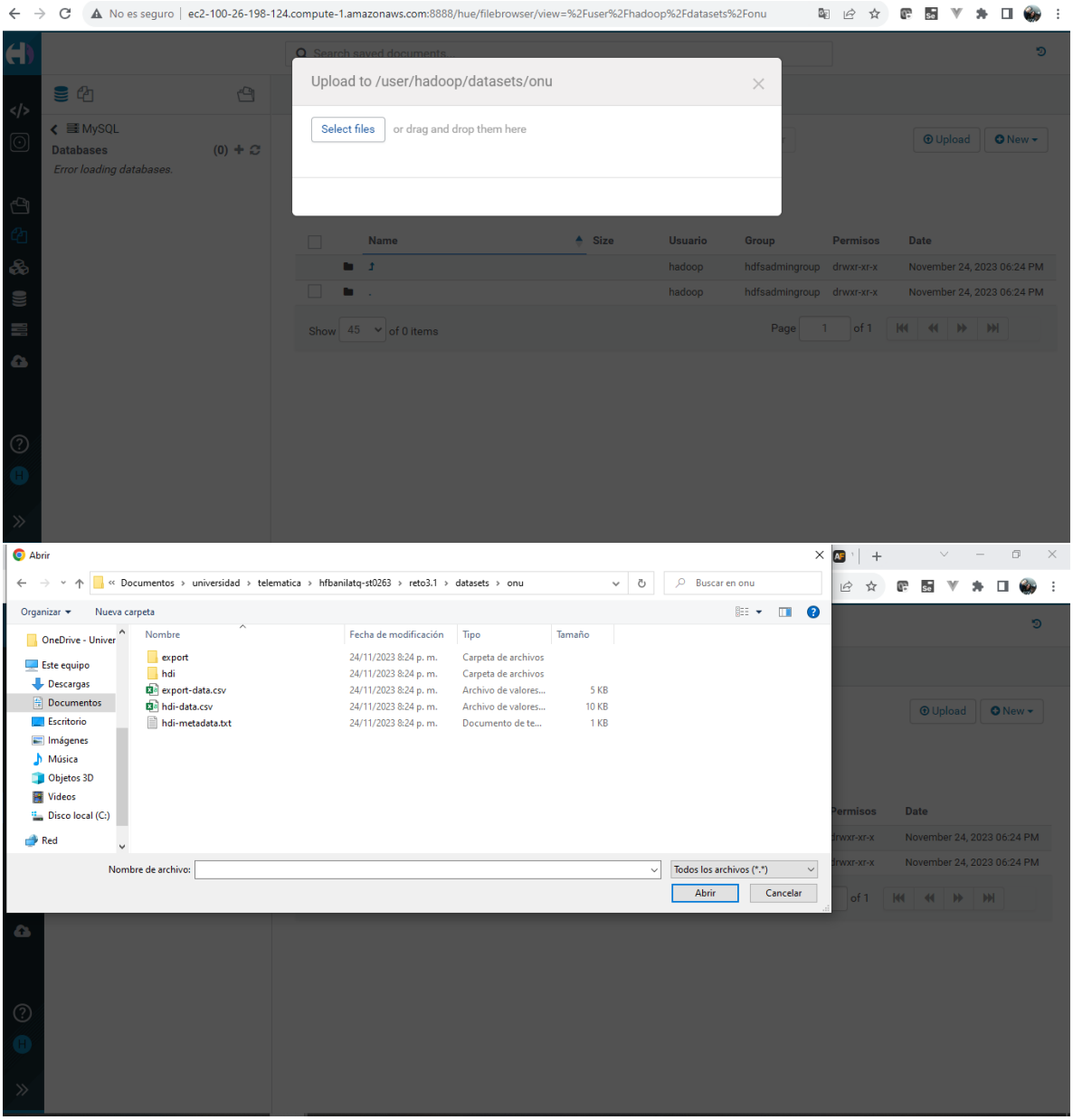


Damo nombre a la carpeta:

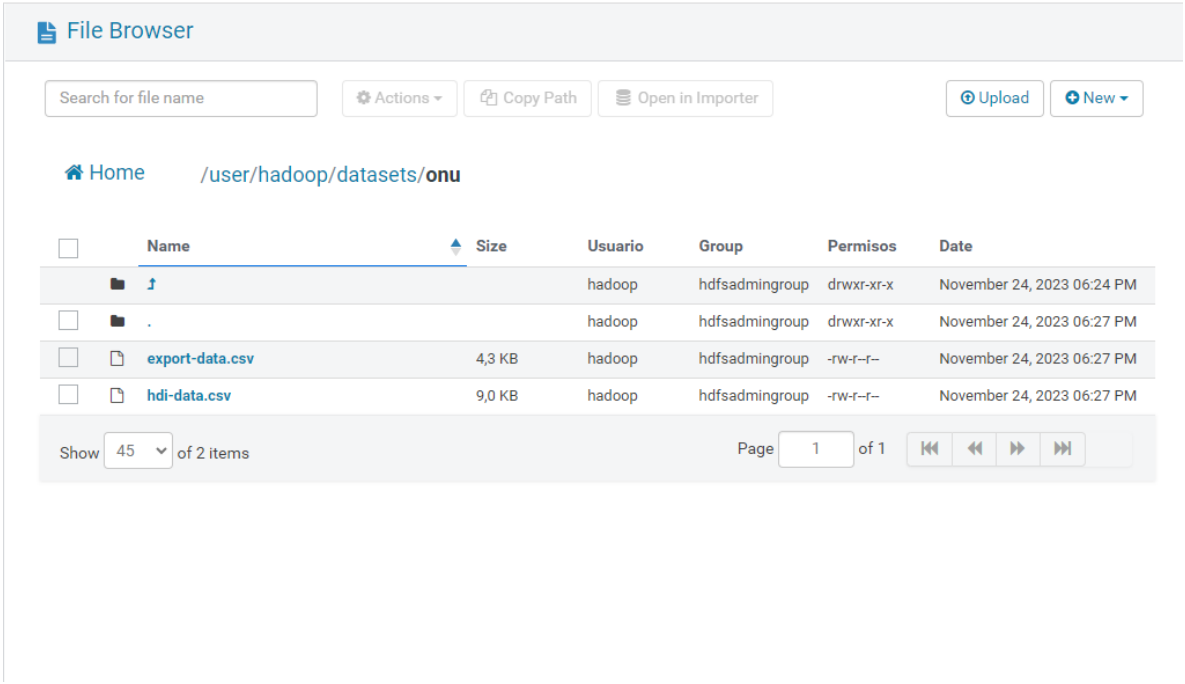


La seleccionamos y damos clic en upload para subir archivos

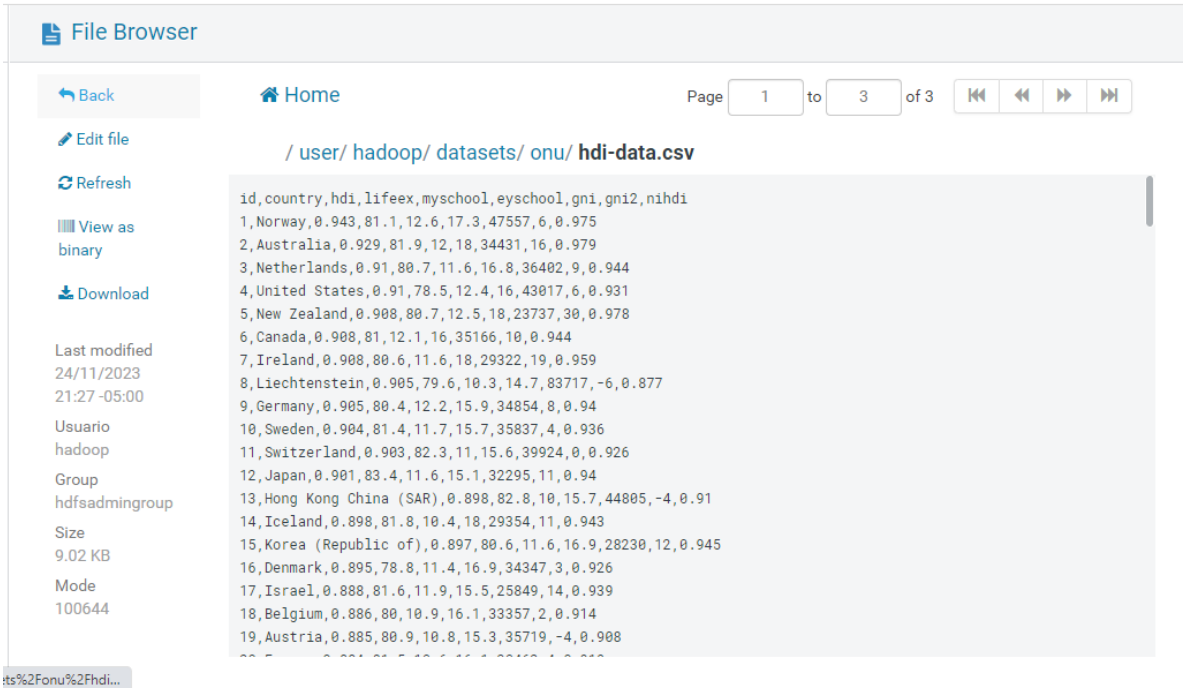

File Browser
Search for file name
Actions
Copy Path
Open in Importer
Upload
New
Home
/user/hadoop/datasets/onu
Name
Size
Usuario
Group
Permisos
Date
hadoop
hdfsadmingroup
drwxr-xr-x
November 24, 2023 06:24 PM
hadoop
hdfsadmingroup
drwxr-xr-x
November 24, 2023 06:24 PM
Show
45
of 0 items
Page
1
of 1


No es seguro | ec2-100-26-198-124.compute-1.amazonaws.com:8888/hue/filebrowser/view=%2Fuser%2Fhadoop%2Fdatasets%2Fonu
Upload to /user/hadoop/datasets/onu
Select files
or drag and drop them here
Upload
New
MySQL
Databases
(0)
Error loading databases.
Name
Size
Usuario
Group
Permisos
Date
hadoop
hdfsadmingroup
drwxr-xr-x
November 24, 2023 06:24 PM
Show
45
of 0 items
Page
1
of 1
Abrir
Documentos › universidad › telematica › hfbanilatq-st0263 › reto3.1 › datasets › onu
Buscar en onu
Organizar
Nueva carpeta
OneDrive - Univer
Este equipo
Descargas
Documentos
Escritorio
Imágenes
Música
Objetos 3D
Videos
Disco local (C:)
Red
Nombre
Fecha de modificación
Tipo
Tamaño
export
24/11/2023 8:24 p. m.
Carpeta de archivos
hdi
24/11/2023 8:24 p. m.
Carpeta de archivos
export-data.csv
24/11/2023 8:24 p. m.
Archivo de valores...
5 KB
hdi-data.csv
24/11/2023 8:24 p. m.
Archivo de valores...
10 KB
hdi-metadata.txt
24/11/2023 8:24 p. m.
Documento de te...
1 KB
Nombre de archivo:
Todos los archivos (*.*)
Abrir
Cancelar

Con lo anterior quedan los archivos subidos:



Ahora para ver el contenido de los archivos simplemente damos clic en el archivo que queremos ver



Y listo, estos son los métodos con los cuales podemos subir archivos en HDFS.