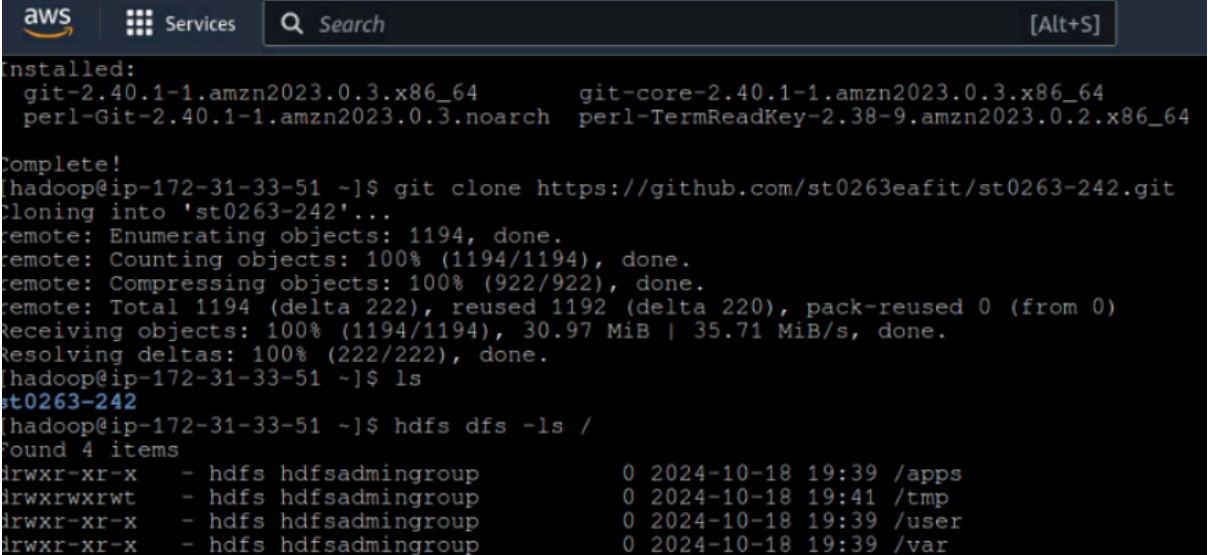


1. **Descripción:** Se utiliza el comando `hdfs dfs -ls /` para listar el contenido de la raíz de HDFS y verificar la estructura de directorios existente.

2. **Comando utilizado:** `hdfs dfs -ls /`
3. **Imagen de referencia:**

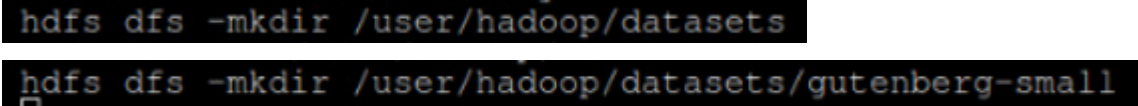


```
aws Services Search [Alt+S]
Installed:
git-2.40.1-1.amzn2023.0.3.x86_64      git-core-2.40.1-1.amzn2023.0.3.x86_64
perl-Git-2.40.1-1.amzn2023.0.3.noarch  perl-TermReadKey-2.38-9.amzn2023.0.2.x86_64

Complete!
[hadoop@ip-172-31-33-51 ~]$ git clone https://github.com/st0263eafit/st0263-242.git
Cloning into 'st0263-242'...
remote: Enumerating objects: 1194, done.
remote: Counting objects: 100% (1194/1194), done.
remote: Compressing objects: 100% (922/922), done.
remote: Total 1194 (delta 222), reused 1192 (delta 220), pack-reused 0 (from 0)
Receiving objects: 100% (1194/1194), 30.97 MiB | 35.71 MiB/s, done.
Resolving deltas: 100% (222/222), done.
[hadoop@ip-172-31-33-51 ~]$ ls
st0263-242
[hadoop@ip-172-31-33-51 ~]$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x  - hdfs hdfsadmin  group  0 2024-10-18 19:39 /apps
drwxrwxrwt  - hdfs hdfsadmin  group  0 2024-10-18 19:41 /tmp
drwxr-xr-x  - hdfs hdfsadmin  group  0 2024-10-18 19:39 /user
drwxr-xr-x  - hdfs hdfsadmin  group  0 2024-10-18 19:39 /var
```

Paso 3: Crear un Directorio en HDFS

1. **Descripción:** Creamos un directorio en HDFS donde se almacenarán los datasets subidos desde el repositorio del curso.
2. **Comando utilizado:** `hdfs dfs -mkdir /user/hadoop/datasets`
3. **Imagen de referencia:**



```
hdfs dfs -mkdir /user/hadoop/datasets

hdfs dfs -mkdir /user/hadoop/datasets/gutenberg-small
```

Paso 4: Subir Archivos a HDFS

1. **Descripción:** Subimos los archivos desde la máquina local o desde una ubicación en el nodo master hacia el nuevo directorio en HDFS.
2. **Comando utilizado:** `hdfs dfs -put <ruta_del_dataset> /user/hadoop/datasets`
3. **Imagen de referencia:**



```
hdfs dfs -put /home/hadoop/st0263-242/bigdata/datasets/gutenberg-small/*.txt /user/hadoop/datasets/gutenberg-small
```

Paso 5: Verificar Archivos Subidos

1. **Descripción:** Verificamos que los archivos se hayan subido correctamente listando el contenido del directorio `datasets`.
2. **Comando utilizado:** `hdfs dfs -ls /user/hadoop/datasets`
3. **Imagen de referencia:**

```

Found 16 items
-r--r-- 1 hadoop hdfsadmingroup 5878 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/LincolnLetters.txt
-r--r-- 1 hadoop hdfsadmingroup 21586 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/LincolnsFirstInauguralAddress.txt
-r--r-- 1 hadoop hdfsadmingroup 1653 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/LincolnsGettysburgAddressGivenNovember
-r--r-- 1 hadoop hdfsadmingroup 262083 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/LincolnsInauguralsAddressesandLettersSe
-r--r-- 1 hadoop hdfsadmingroup 4093 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/LincolnsSecondInauguralAddress.txt
-r--r-- 1 hadoop hdfsadmingroup 516298 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/SpeechesandLettersofAbrahamLincoln1832-
-r--r-- 1 hadoop hdfsadmingroup 167895 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/StateoftheUnionAddresses.txt
-r--r-- 1 hadoop hdfsadmingroup 3928 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheEmancipationProclamation.txt
-r--r-- 1 hadoop hdfsadmingroup 45664 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheLifeandPublicServiceofGeneralZachary
-r--r-- 1 hadoop hdfsadmingroup 459006 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume1.txt
-r--r-- 1 hadoop hdfsadmingroup 505150 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume2.txt
-r--r-- 1 hadoop hdfsadmingroup 254941 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume3.txt
-r--r-- 1 hadoop hdfsadmingroup 209643 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume4.txt
-r--r-- 1 hadoop hdfsadmingroup 692051 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume5.txt
-r--r-- 1 hadoop hdfsadmingroup 601102 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume6.txt
-r--r-- 1 hadoop hdfsadmingroup 478689 2024-10-18 19:51 /user/hadoop/datasets/gutenberg-small/AbrahamLincoln/TheWritingsofAbrahamLincolnVolume7.txt

```

Paso 6: Copia de Archivos a HDFS

- Descripción:** Esta imagen muestra la salida del comando `hdfs dfs -copyFromLocal /home/hadoop/st0263-242/bigdata/datasets/* /user/hadoop/datasets/`, el cual copia los archivos y directorios de la ruta local `/home/hadoop/st0263-242/bigdata/datasets/` al sistema de archivos distribuido Hadoop (HDFS) en la ruta `/user/hadoop/datasets/`. Esto permite asegurarse de que los archivos se han transferido correctamente y que tienen los permisos y propiedades adecuados en HDFS.
- Comando utilizado:** `hdfs dfs -copyFromLocal /home/hadoop/st0263-242/bigdata/datasets/* /user/hadoop/datasets/` Imagen de referencia: La imagen muestra los nombres de los archivos que se han copiado de la ruta local a HDFS, incluyendo archivos relacionados con los escritos de Abraham Lincoln.

```

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
[hadoop@ip-172-31-33-51 ~]$ hdfs dfs -ls /user/hadoop/datasets/
Found 11 items

```

Hue

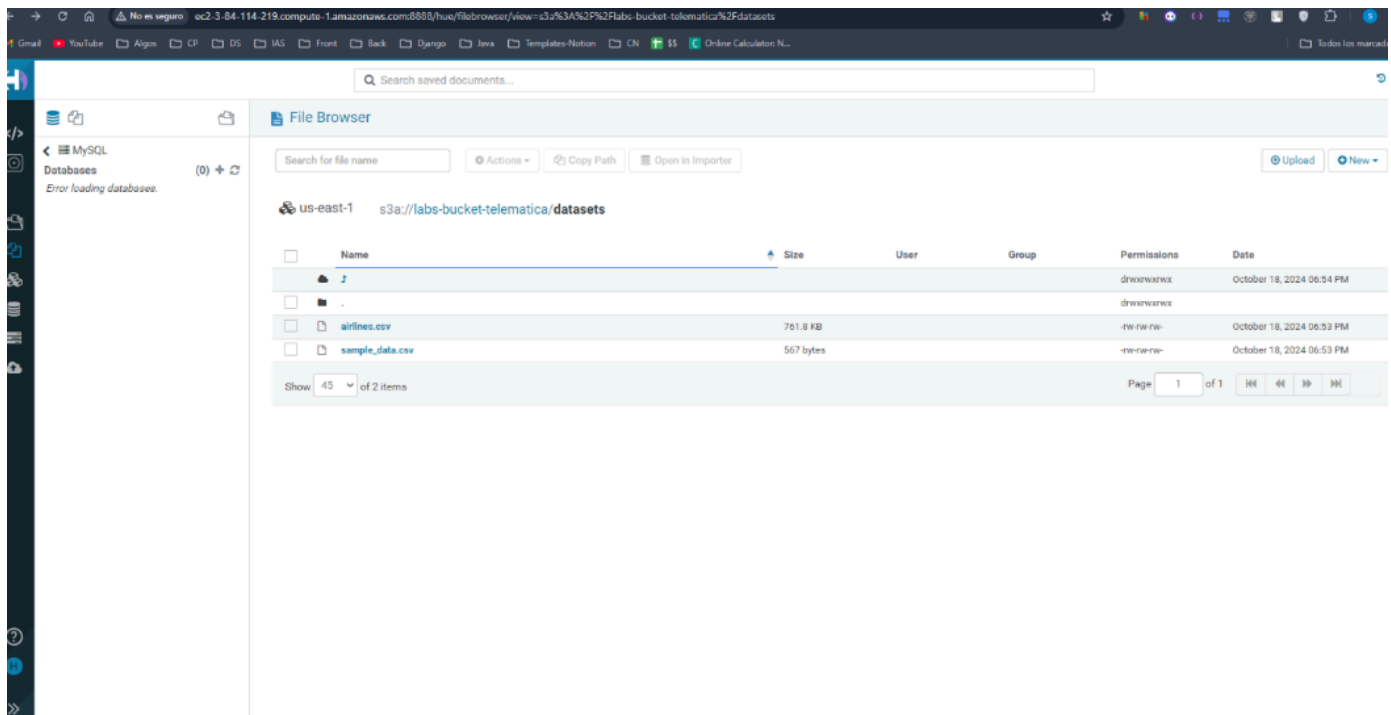
1. Acceso a Hue en el Cluster EMR de AWS

Primero, se accedió a la interfaz gráfica de Hue a través del cluster EMR configurado en AWS. Hue proporciona una interfaz amigable para interactuar con HDFS y otros componentes del ecosistema Hadoop.

2. Navegación en Hue

Una vez dentro de Hue, navegamos al módulo de "File Browser" o explorador de archivos, donde podemos ver la estructura de carpetas y archivos de HDFS.

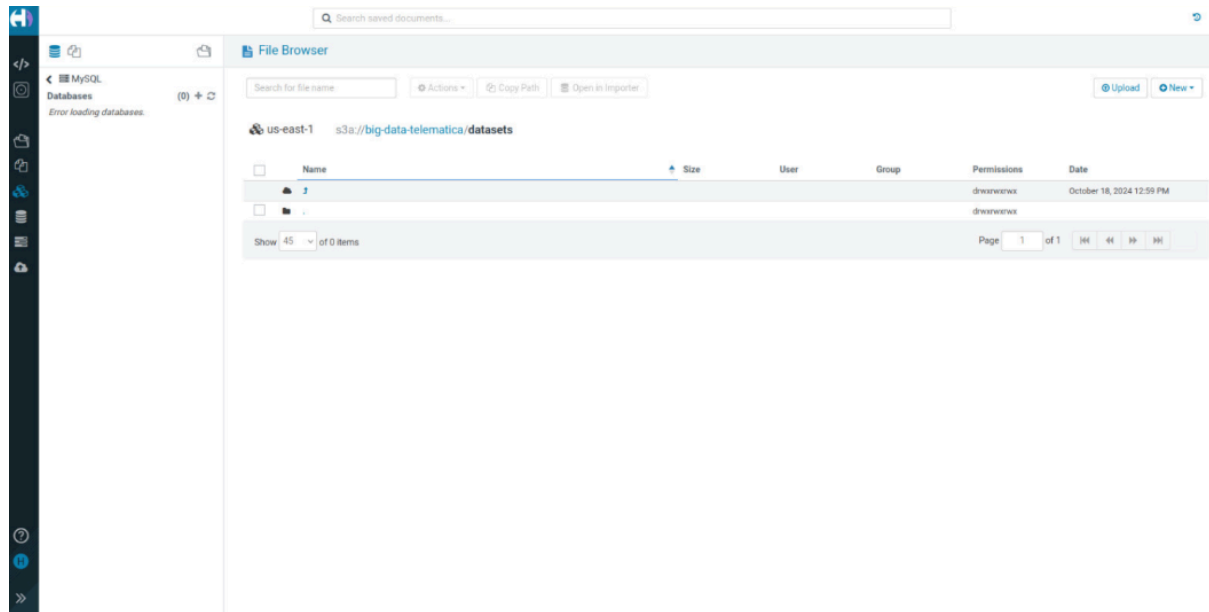
Imagen:



3. Creación de una Carpeta en HDFS

En el siguiente paso, se creó una carpeta en HDFS para almacenar los datasets. Esta carpeta será el directorio donde subiremos los archivos necesarios.

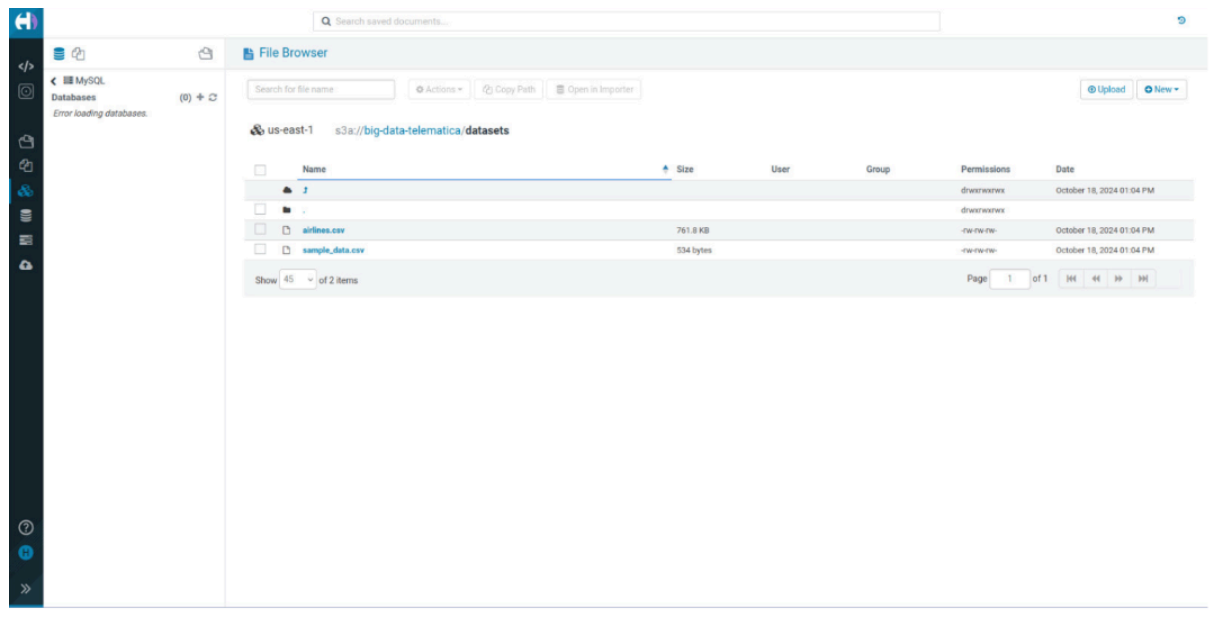
Imagen:



4. Subida de Datasets a HDFS

Con la carpeta creada, procedimos a subir los datasets descargados desde el repositorio del curso en GitHub: [Repositorio del curso](#). Estos archivos se subieron a la carpeta creada previamente en HDFS.

Imagen:



5. Verificación de Archivos

Una vez completada la subida de archivos, se verificó que todos los archivos estuvieran correctamente cargados en la carpeta de HDFS y que fueran accesibles desde la interfaz de Hue.

Imagen:

