Guía reto número 2

Por: Sebastián Urrego García

Prerrequisitos

Antes de comenzar a realizar el copiado de archivos con HDFS y S3, usted debe completar los requisitos que a continuación se indican:

1. Tener una cuenta de AWS
2. Haber realizado la guía del reto número 1 posteriormente

Paso a paso

En esta guía nos vamos a enfocar donde lo dejamos en la guía anterior y nos vamos a enfocar sobre todo en la parte de realizar la copia de archivos en S3 y en HDFS por dos medios en cada uno:

1. El primer medio es SSH y realizaremos la copia de archivos que tengamos almacenados localmente
2. El segundo medio es por la interfaz gráfica de Hue que nos facilita la subida de los archivos por medio de su UI.

HDFS

Ya que estamos utilizando una cuenta de AWS academy o cada vez que terminemos el cluster que estemos utilizando se borraran los datos sin embargo es necesario el uso de los mismos para entender todo el entorno de Hadoop.

Copiar archivos por medio de ssh

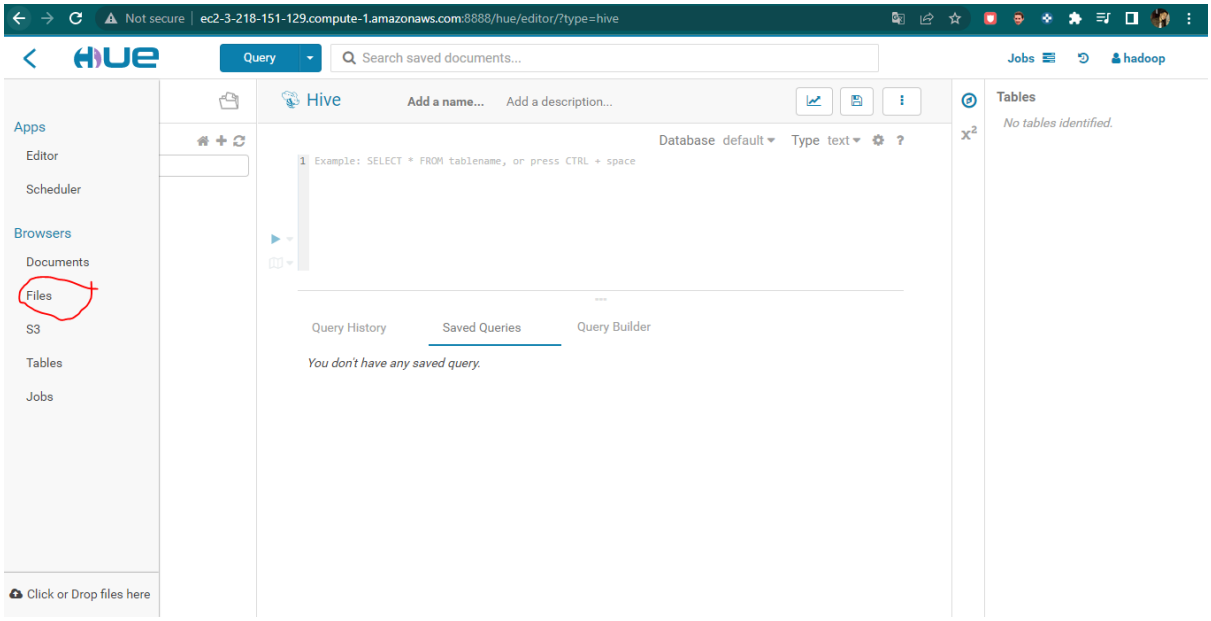
Debemos de dirigirnos a la terminal y seguir los pasos de la guía 1 para entrar al EMR e ingresar con nuestro usuario que creamos en Hue, de aquí en adelante, debemos de tener en nuestro equipo local muy cerca nuestra carpeta de datasets y ahora ejecutamos los siguientes comandos para montar nuestros datasets:

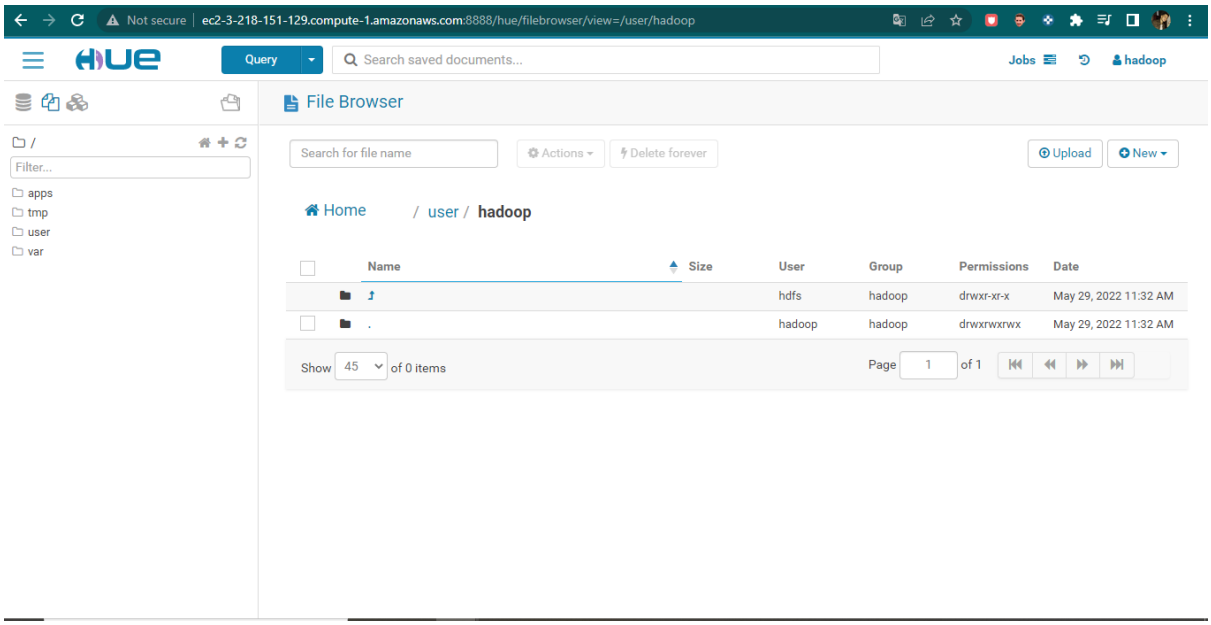
```
hdfs dfs -mkdir /user/<username>/datasets

hdfs dfs -mkdir /user/<username>/datasets/gutenberg-small

hdfs dfs -put /datasets/gutenberg/gutenberg-small.zip /user//datasets/

hdfs dfs -ls /user/<username>/datasets
```

Con estos comandos en este orden y remplazándolos con los elementos que necesitamos, creamos la carpeta de datasets dentro de nuestro usuario que creamos en nuestro caso es hadoop, ya después aclaramos que vamos a poner y donde lo vamos poner los datasets para su uso y ya el ultimo comando nos ayuda a visualizar como están los archivos y cuantos de ellos se encuentran (Es como ejecutar el comando ls en la computadora)

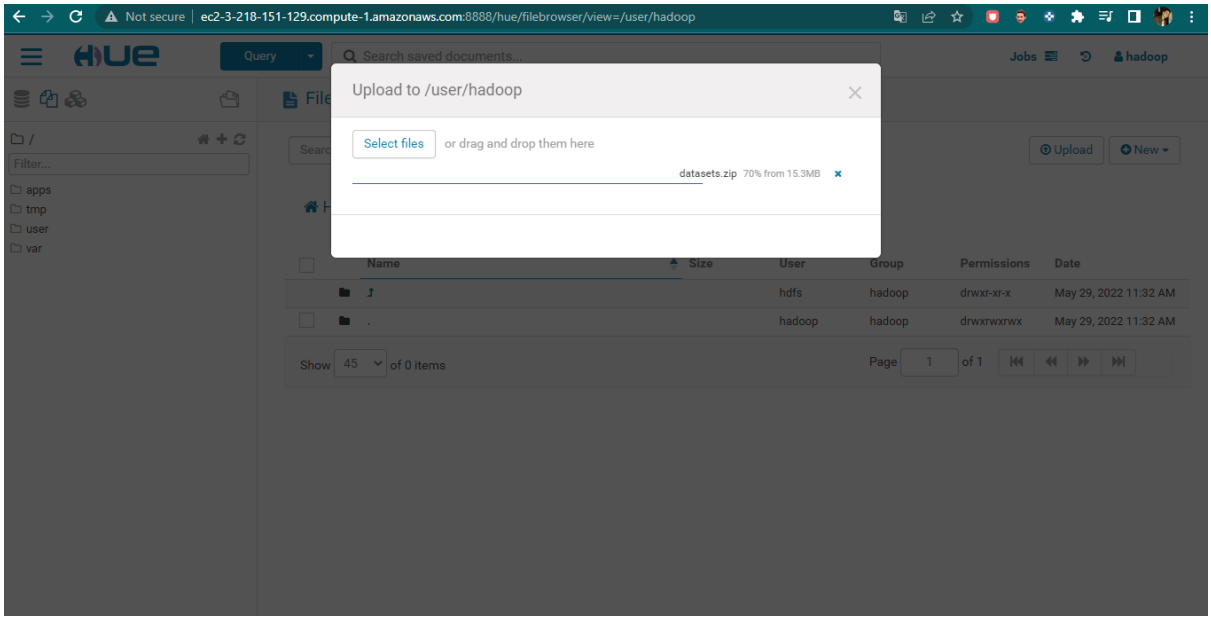
Copiar archivos por medio de Hue

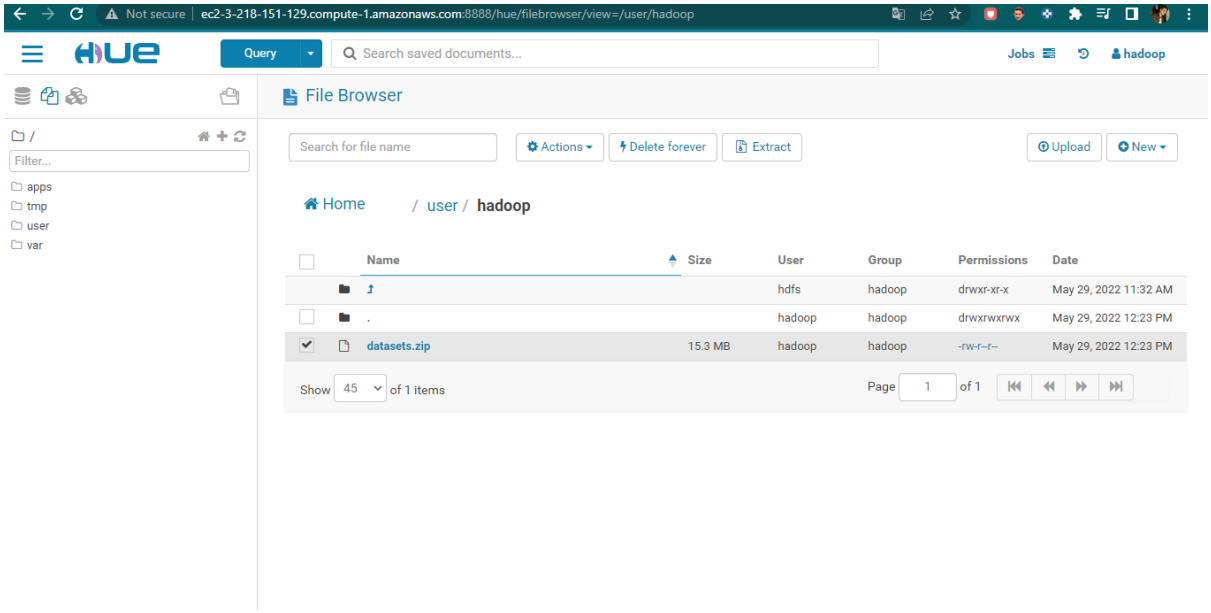
Para comenzar debemos de seguir las instrucciones de la guía numero 1 para entrar al portal de Hue, registrarnos con nuestro usuario, en este caso es hadoop y estar en la consola de ahí nos dirigimos a la barra lateral de opciones y nos dirigimos a la parte que nos dice files y entramos ahí, nos debería de llevar a la ruta de /user/<nombre_usuario_para_hue>, en nuestro caso tenemos que estamos en user/hadoop/





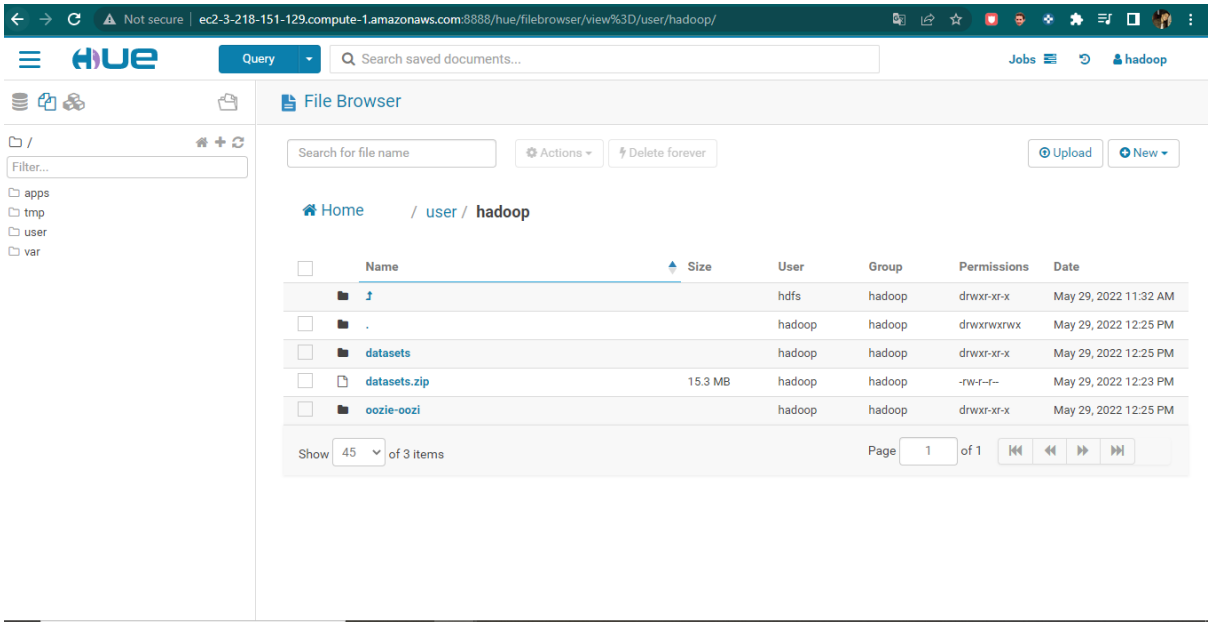
Por el momento aquí no tenemos nada, pero vamos a subir el dataset que se nos dío para montar en este laboratorio, antes de montarlo, lo debemos de comprimir para solo subir un archivo y en Hue le damos a la opción de upload, allí seleccionamos el archivo .zip que tengamos en nuestro equipo y esperamos a que se monte, nos debío de quedar así:





Ahora con el .zip seleccionado, le damos a la opción que nos aparece de Actions y buscamos la que nos dice extract, esto hará que nos extraiga todos los archivos que tenemos en el .zip y los coloque

dentro del HDFS así debería de quedar:



Ya podemos navegar los archivos de datasets para lo que necesitemos, con esto completaríamos la subida de archivos por medio de Hue a HDFS

S3

La ventaja que nos ofrece S3 es que los archivos quedan guardados permanentemente en AWS e incluso quedan después utilizables a futuro para cualquiera que quiera acceder a ellos por medio de la URL, en la ultima sección de este tutorial, se explicará como poner algunos de los elementos del bucket públicos para el mundo.

Copiar archivos por medio de ssh

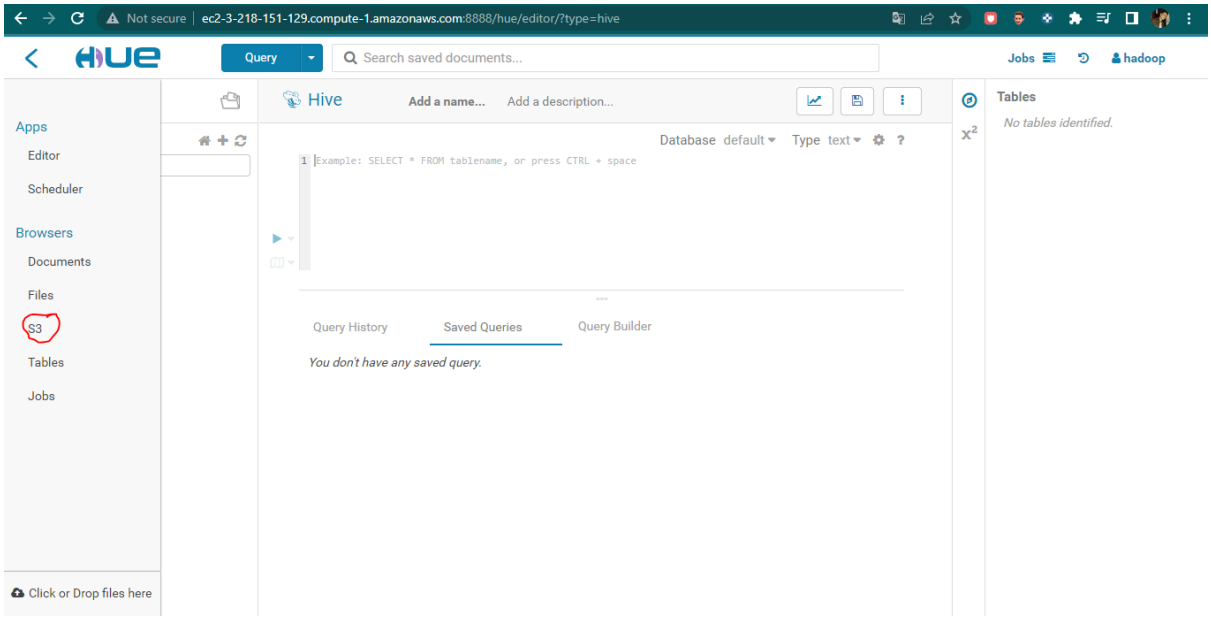
Debemos de dirigirnos a la terminal y seguir los pasos de la guía 1 para entrar al EMR e ingresar con nuestro usuario que creamos en Hue, de aquí en adelante, debemos de tener en nuestro equipo local muy cerca nuestra carpeta de datasets y ahora ejecutamos los siguientes comandos para montar nuestros datasets, recurde que para completar esta parte usted ya debio haber montado los archivos por ssh al sistema de hdfs:

```
hadoop distcp s3://<Nombre_bucket>/datasets_ssh/ /tmp/
```

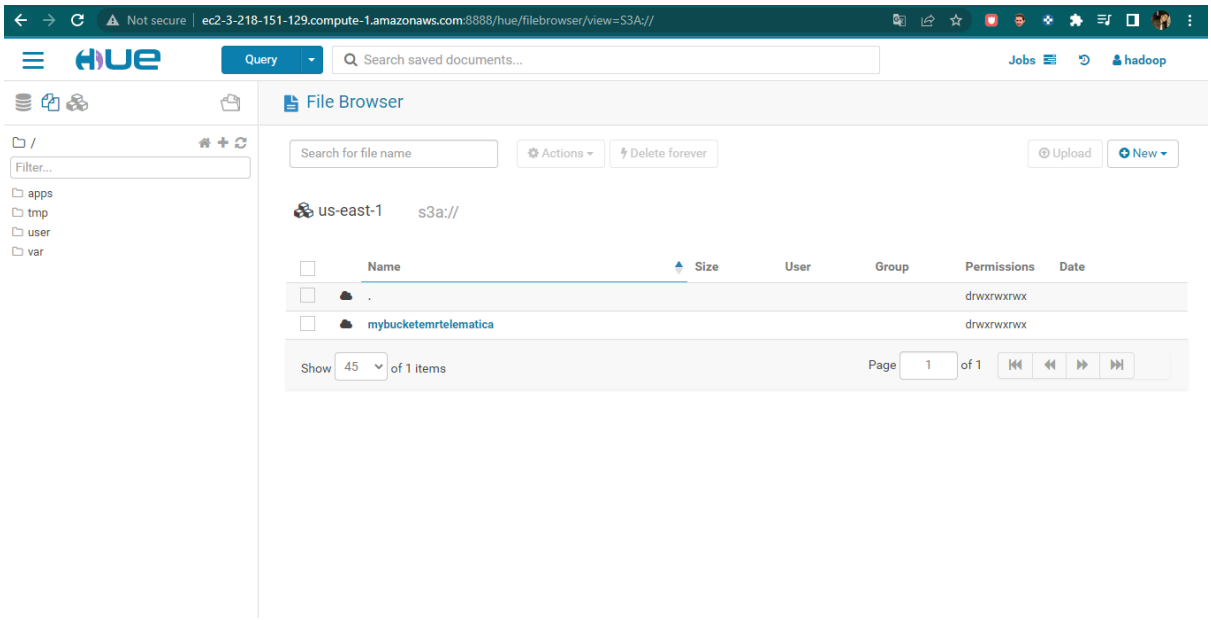
Lo que hace esto es copiar la información que queramos tener en la carpeta de HDFS de preferencia y lo debemos montar en nuestro bucket de S3 en este caso esta copiando la información que esta contenido en la carpeta de tmp hacia nuestros datasets por ssh, de esta forma colocamos nuestros datasets en S3 por ssh.

Copiar archivos por medio de Hue

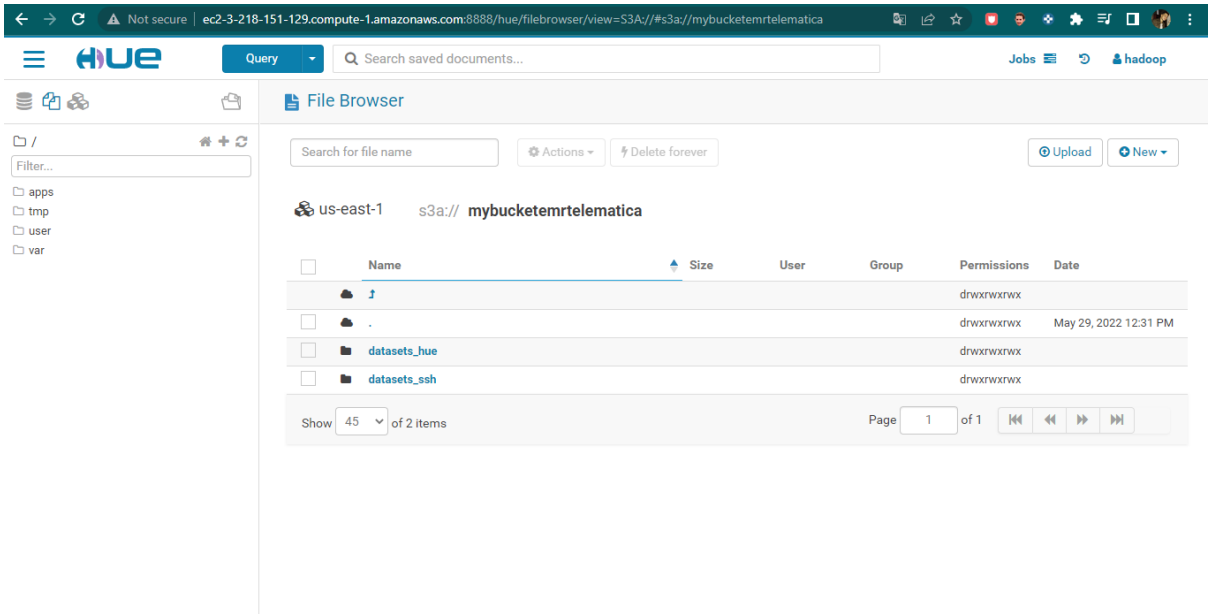
De la misma forma que se utilizo en HDFS nos vamos a dirigir a Hue y vamos a seleccionar en la barra lateral de opciones S3, esto nos redirige donde realizaremos la montada de archivos por medio de Hue a S3:



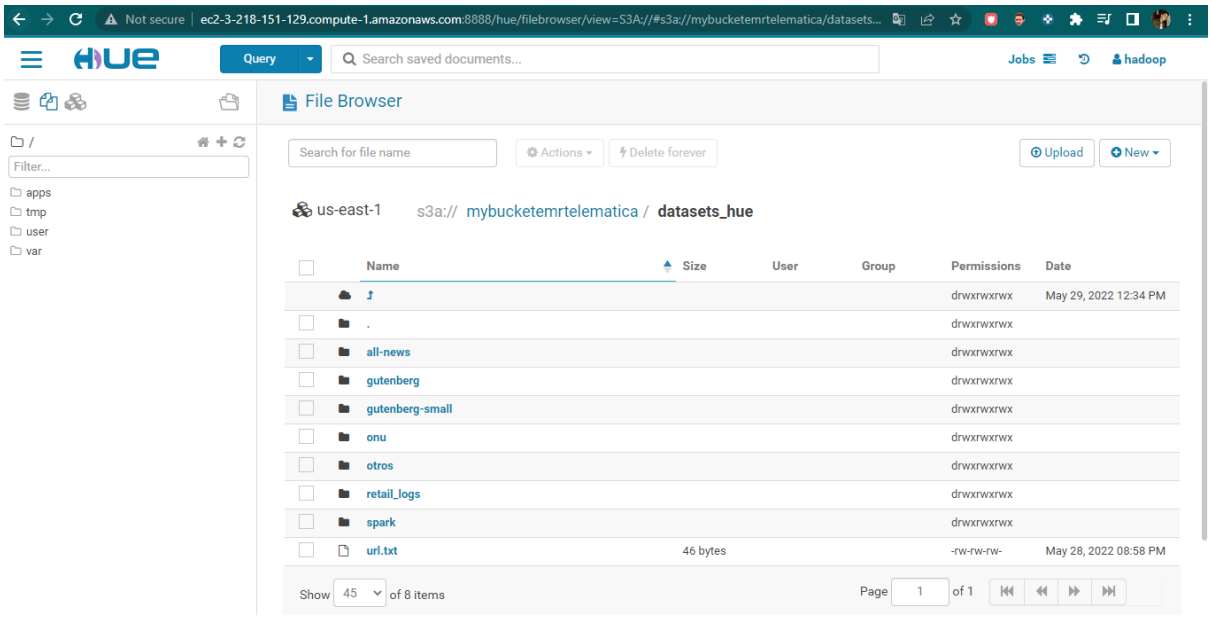
Aquí podremos ver el bucket que en la guía anterior creamos, ingresamos a el:



Ya dentro de este como los archivos se guardan creamos dos jerarquías de carpetas una para ssh y otra para hue, en este caso nos dirigimos a la de Hue y ahí al igual que en el proceso anterior le damos en la opción de upload y montamos los archivos del dataset que se nos dio para este proyecto



Ya al montarlos uno por uno, tendríamos una vista de esta forma de todas las carpetas y sus respectivos archivos:

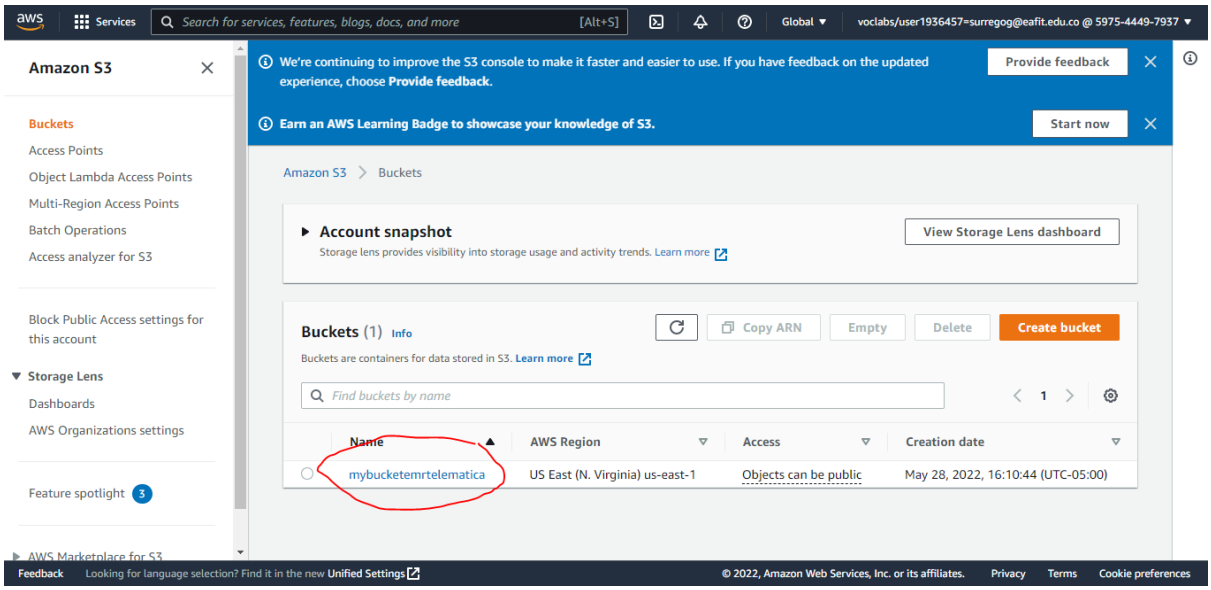


Con esto finalizamos la montada de archivos en S3, ahora en el siguiente capitulo vamos a ponerlos públicos para que cualquiera con una URL de internet pueda acceder a ellos de forma más fácil

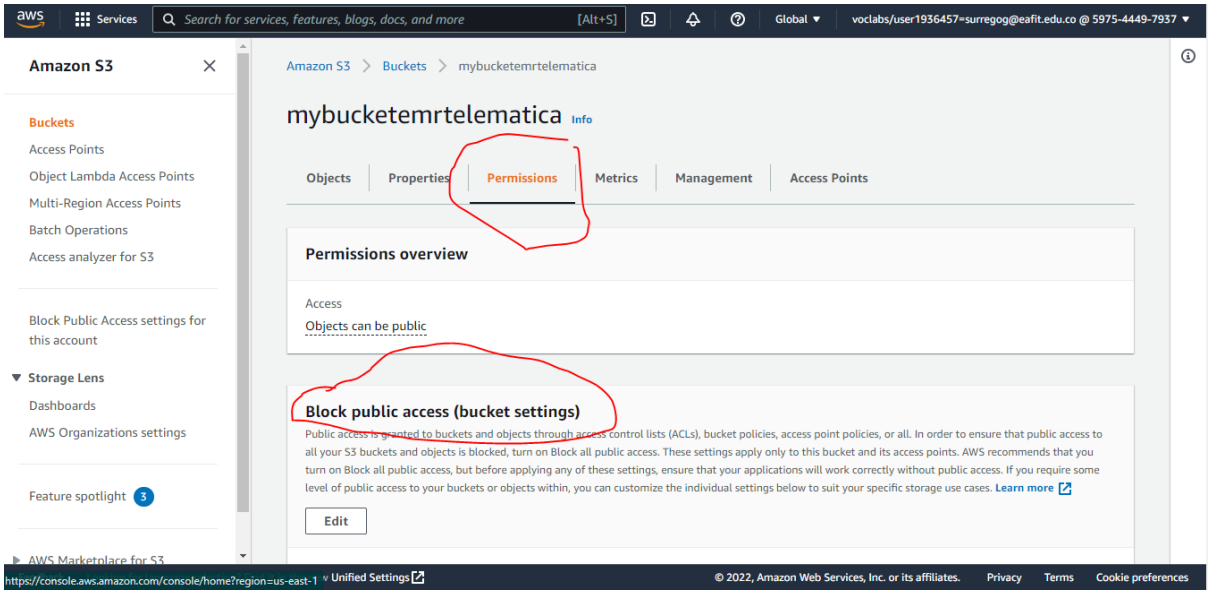
Bucket para lectura publica

Ya con todos los archivos montados en S3, ahora queremos compartirlo al público, pero como lo hacemos ya que en la creación del bucket en el primer momento, le negamos los permisos de acceso público.

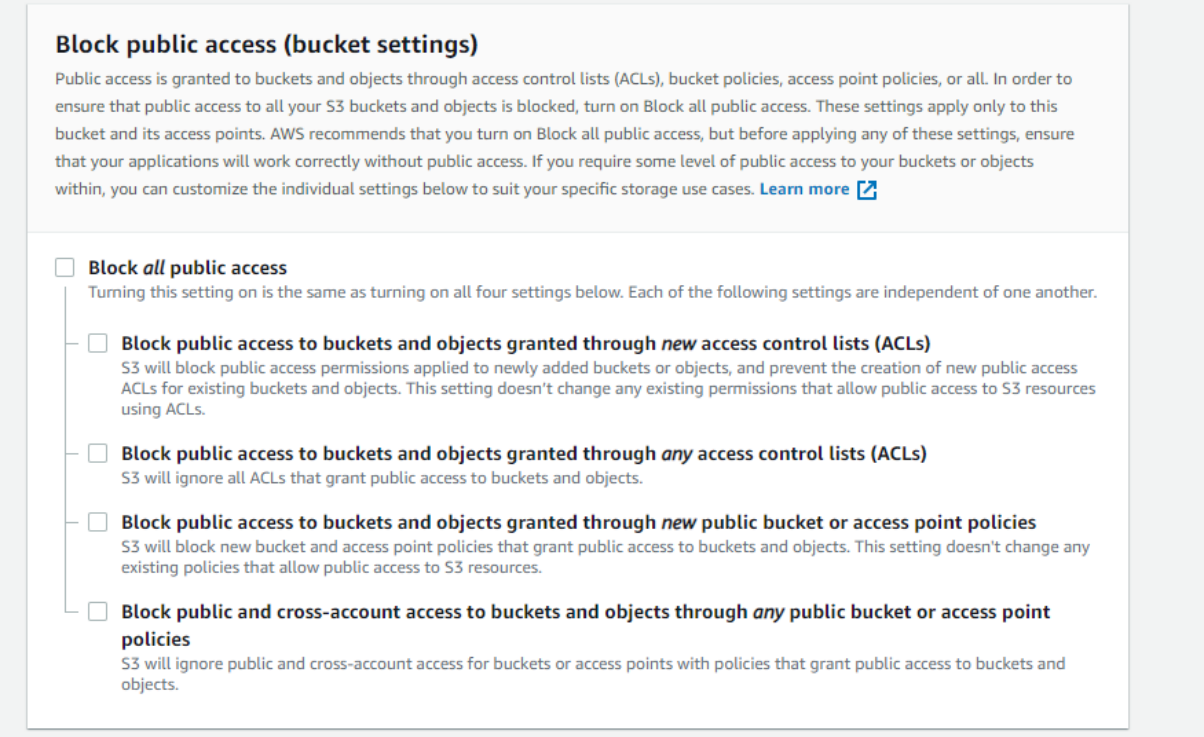
Pero no hay que temer, nos dirigimos en este caso en AWS a nuestra consola de S3 y localizamos nuestro bucket en este caso a “mybucketemrtelematica”



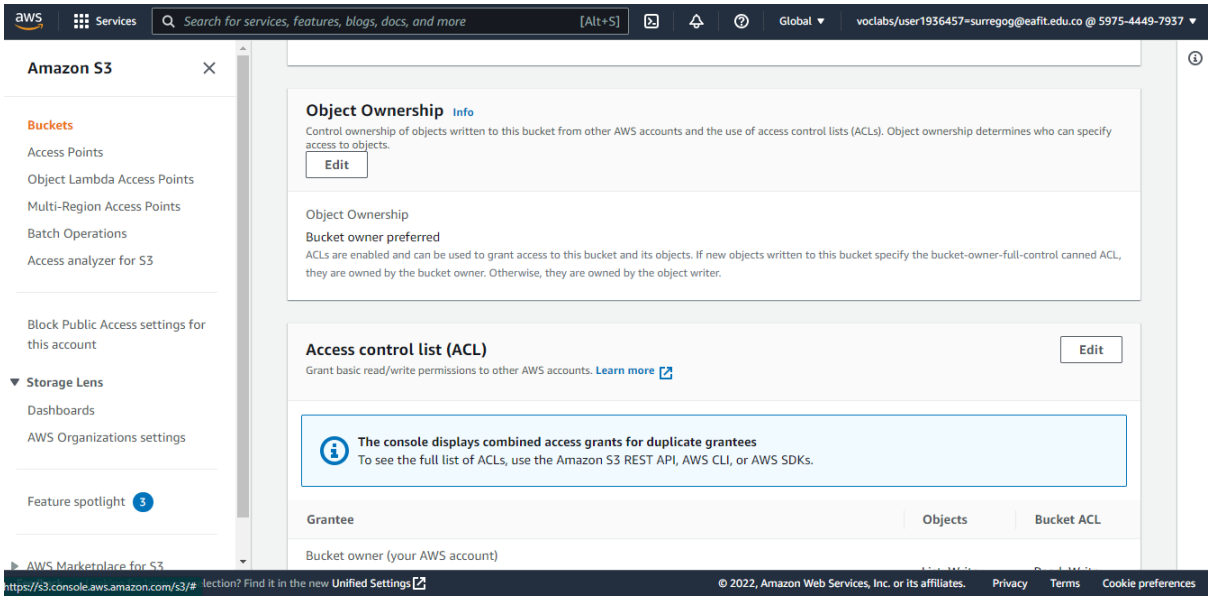
Cuando entramos allí nos dirigimos a la sección de permissions y luego buscamos la zona que nos diga “block public Access” y lo editamos:

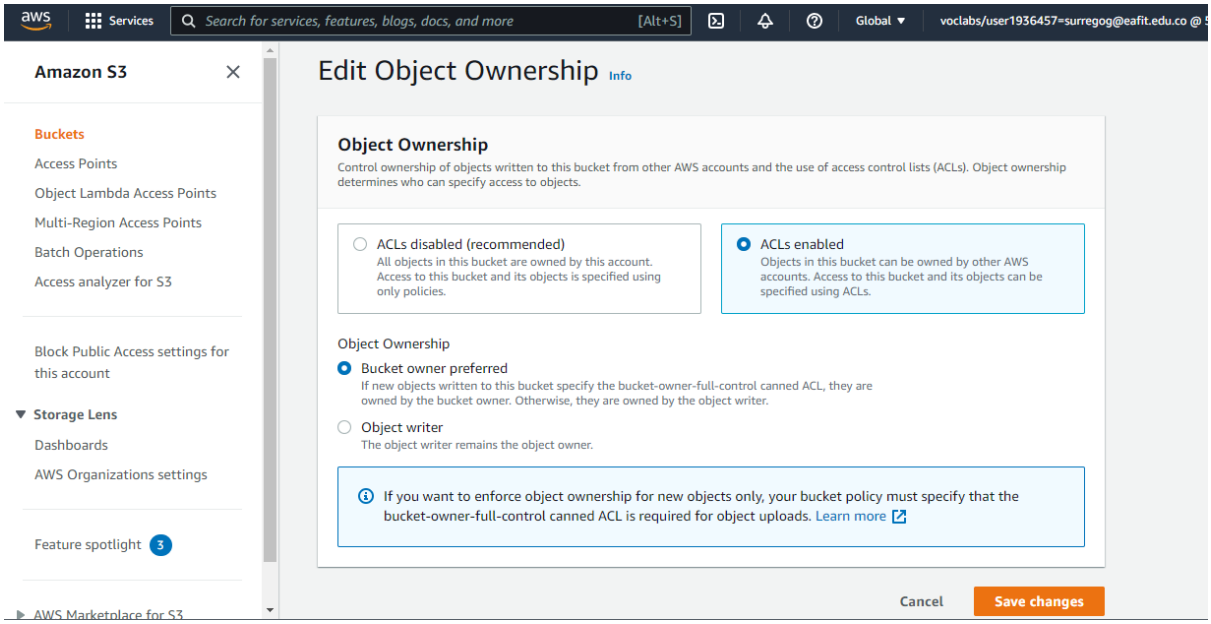


Aquí deshabilitamos todos los protocolos de seguridad y le damos a guardar los cambios, te debe de salir algo así:

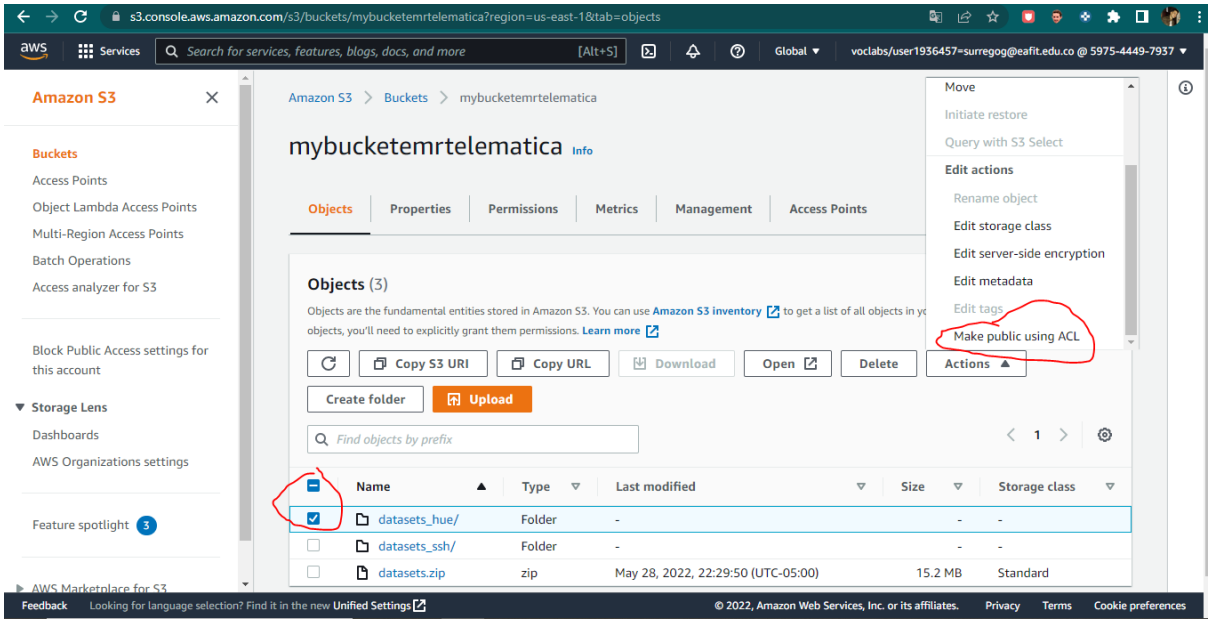


Luego volvemos a la sección de permission, pero ahora buscamos la parte que nos diga Object Ownership y le damos para editarlo, ahora seleccionamos la parte que nos dice habilitar ACL y lo habilitamos para poder usarlo:





Ya con estos cambios nos volvemos a dirigir a lo que tenemos en nuestro bucket y buscamos las carpetas que deseamos volver publicas para el mundo, en este caso tenemos las carpetas de datasets_hue y datasets_ssh, nos paramos sobre ellas y le damos al botón de actions en la parte final nos aparece la opción de volverlas publicas



Y ahora al volverlas publicas solo nos paramos sobre ellas o el archivo que queremos ver por medio de una URL que se encuentre dentro de una carpeta y colocamos la URL en el navegador, de esta forma tendríamos por ejemplo esto:

Miren la URL los endpoints que tienen, el endpoint final es el archivo que se desea abrir el inicio de la URL es un bucket y la carpeta a la que entramos es a la de datasets_hue.