

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 borrarán 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
```

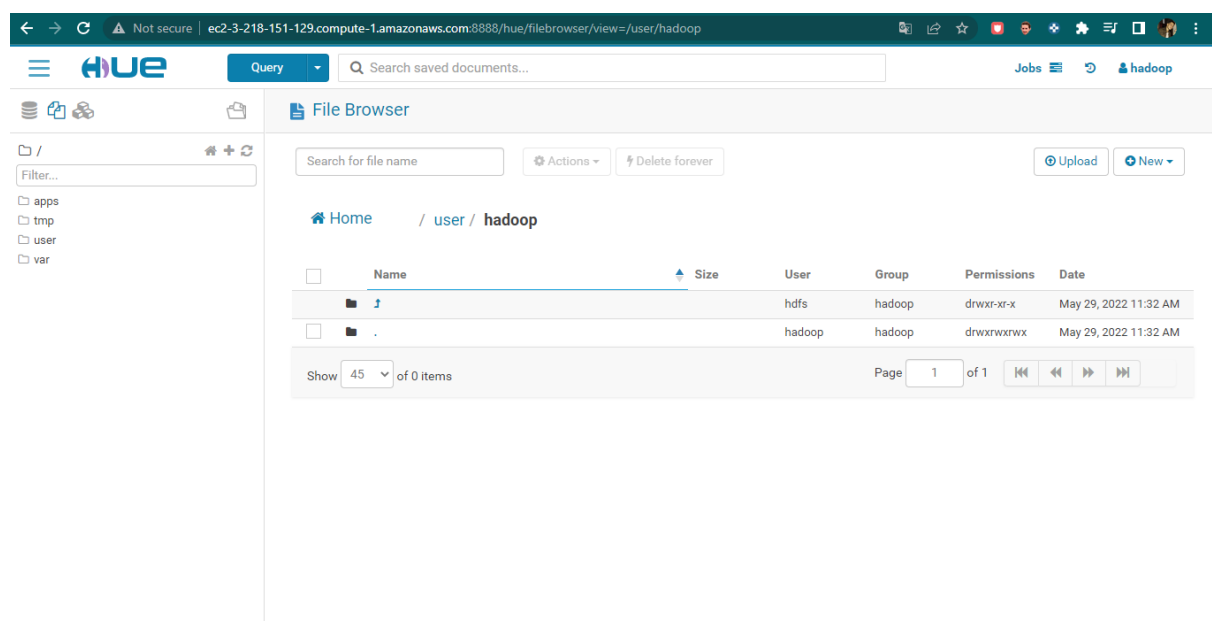
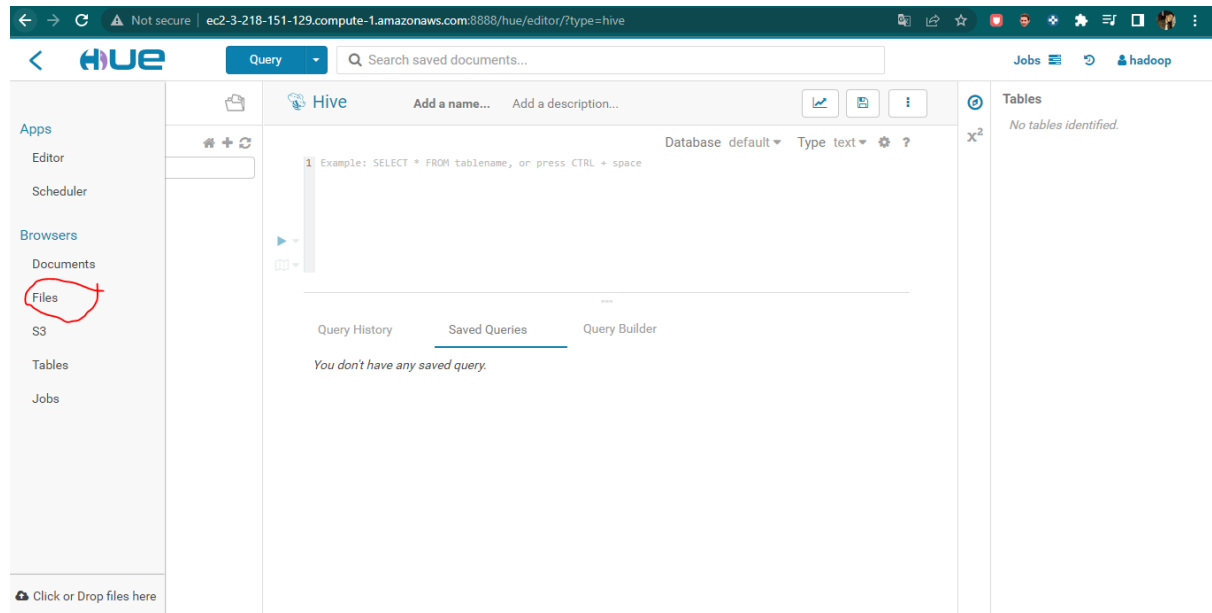
```
hdfs dfs -put /datasets/gutenberg/gutenberg-small.zip /user//datasets/
```

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

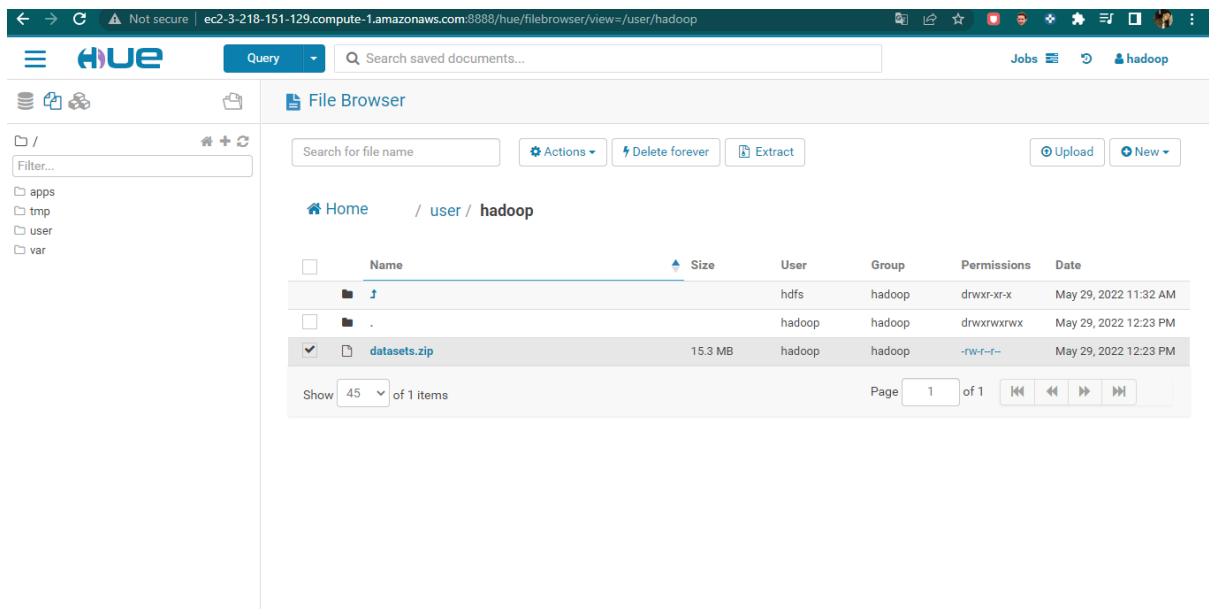
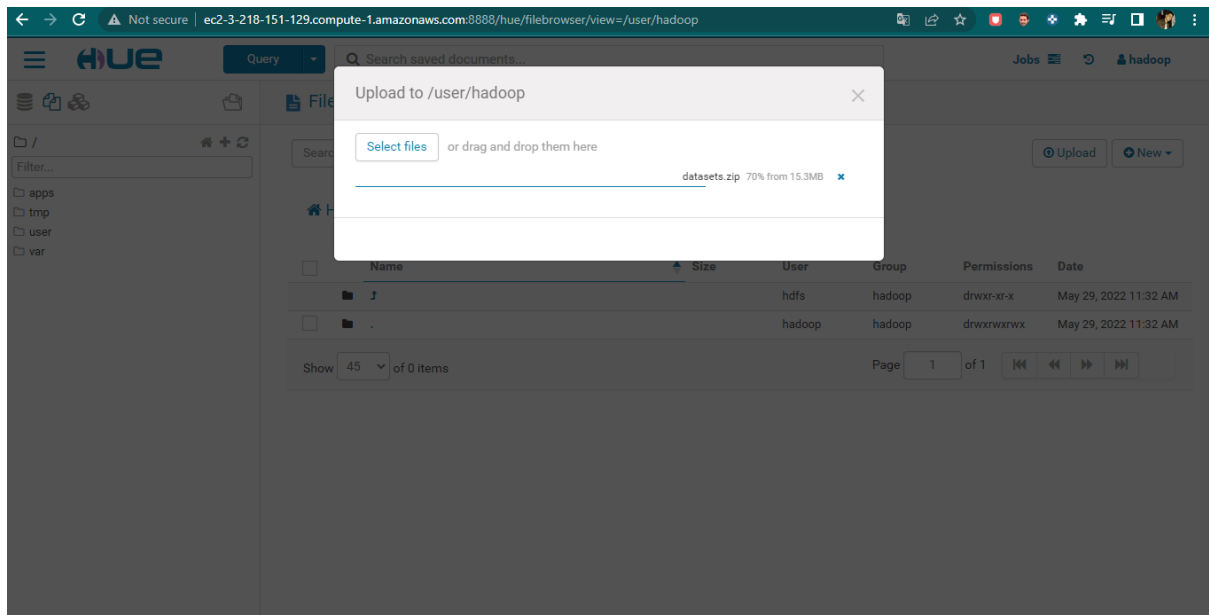
Con estos comandos en este orden y remplazándolos con los elementos que necesitamos, creamos la carpeta de datasets dentro de nuestro usuario que creamos en nuestro caso es hadoop, ya después aclaramos que vamos a poner y donde lo vamos poner los datasets para su uso y ya el ultimo comando nos ayuda 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/

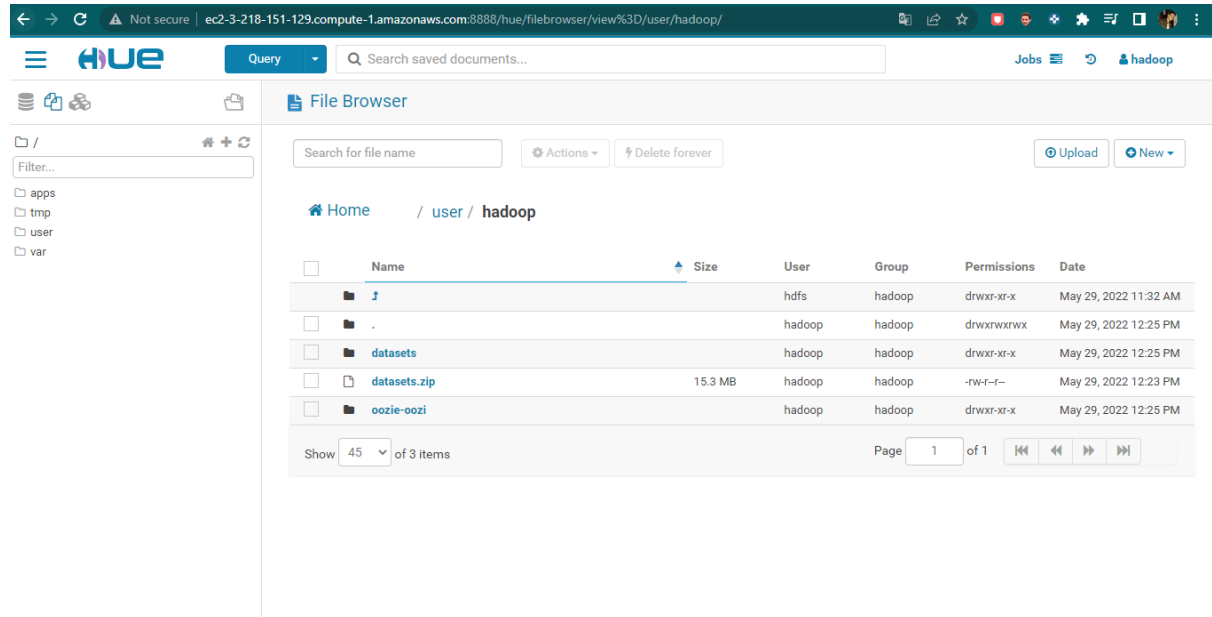


Por el momento aquí no tenemos nada, pero vamos a subir el dataset que se nos dio para montar en este laboratorio, antes de montarlo, lo debemos de comprimir para solo subir un archivo y en Hue le damos a la opción de upload, allí seleccionamos el archivo .zip que tengamos en nuestro equipo y esperamos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

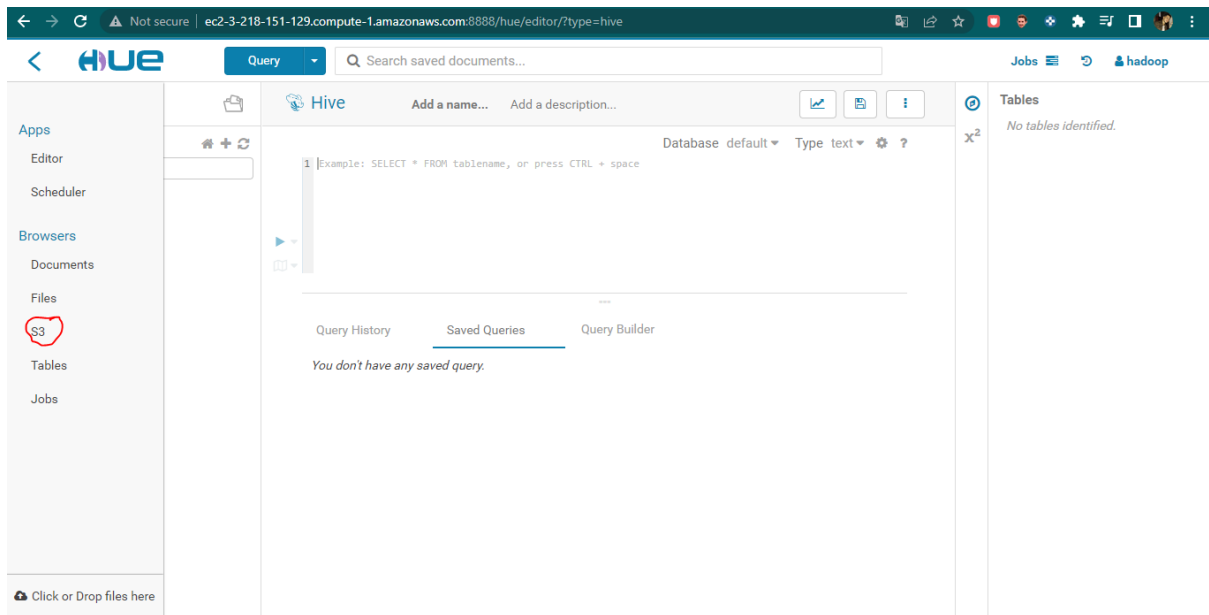
Debemos de dirigirnos a la terminal y seguir los pasos de la guía 1 para entrar al EMR e ingresar con nuestro usuario que creamos en Hue, de aquí en adelante, debemos de tener en nuestro equipo local muy cerca nuestra carpeta de datasets y ahora ejecutamos los siguientes comandos para montar nuestros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

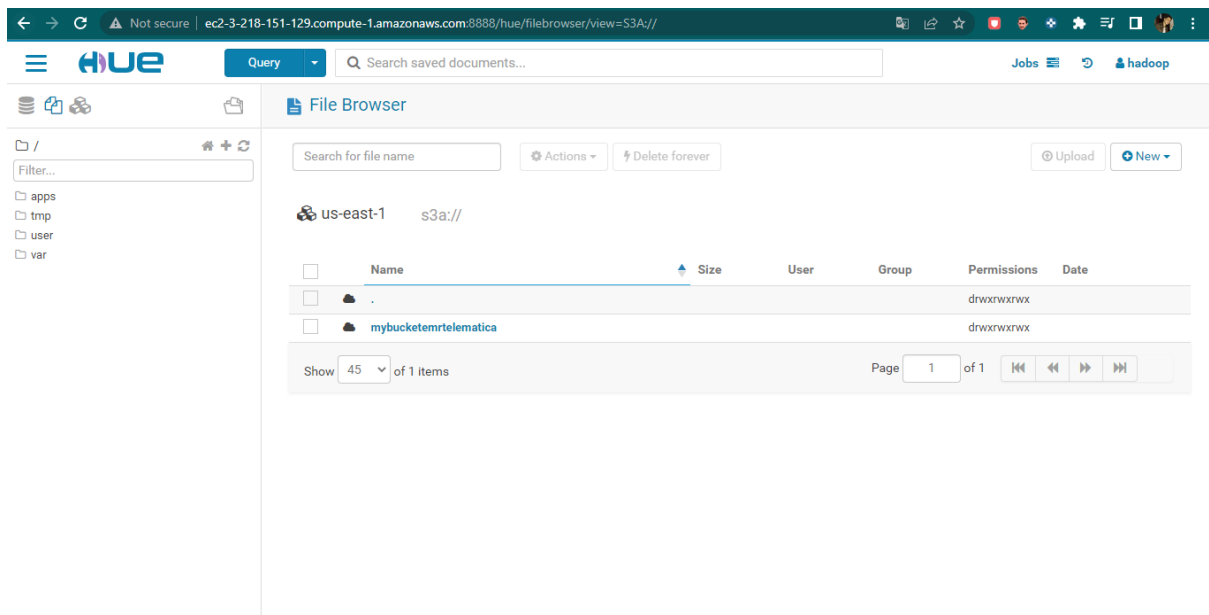
Lo que hace esto es copiar la información que queramos tener en la carpeta de HDFS de preferencia y lo debemos montar en nuestro bucket de S3 en este caso esta copiando la información que esta contenido en la carpeta de tmp hacia nuestros datasets por ssh, de esta forma colocamos nuestros datasets en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

Copiar archivos por medio de Hue

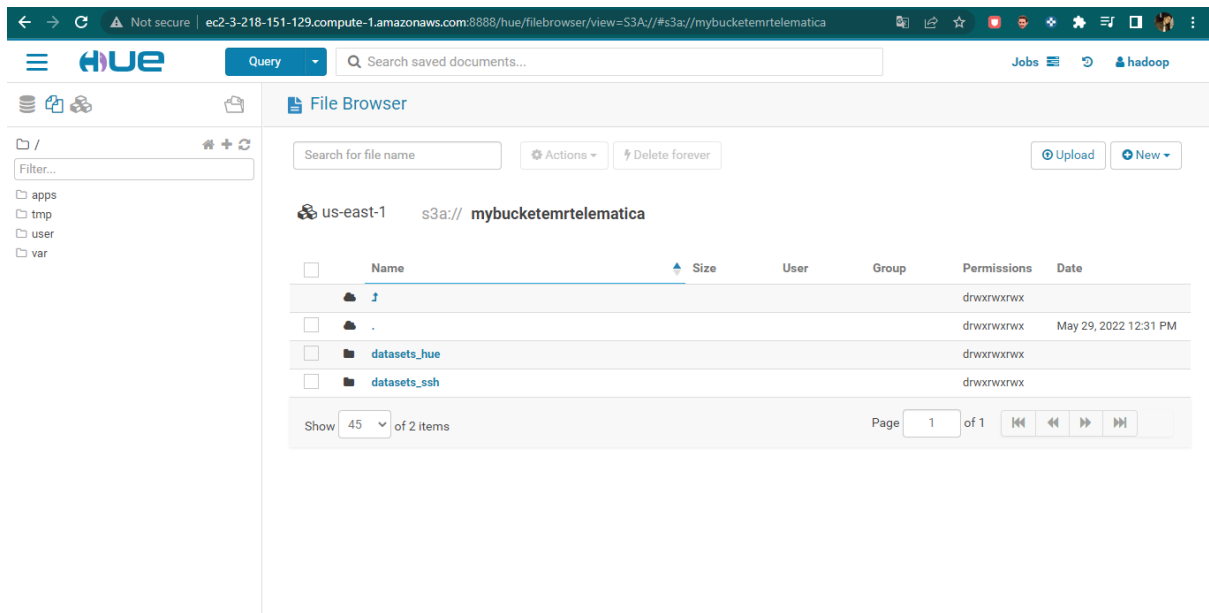
De la misma forma que se utilizo en HDFS nos vamos a dirigir a Hue y vamos a seleccionar en la barra lateral de opciones S3, esto nos redirige donde realizaremos la montada de archivos por medio de Hue a S3:



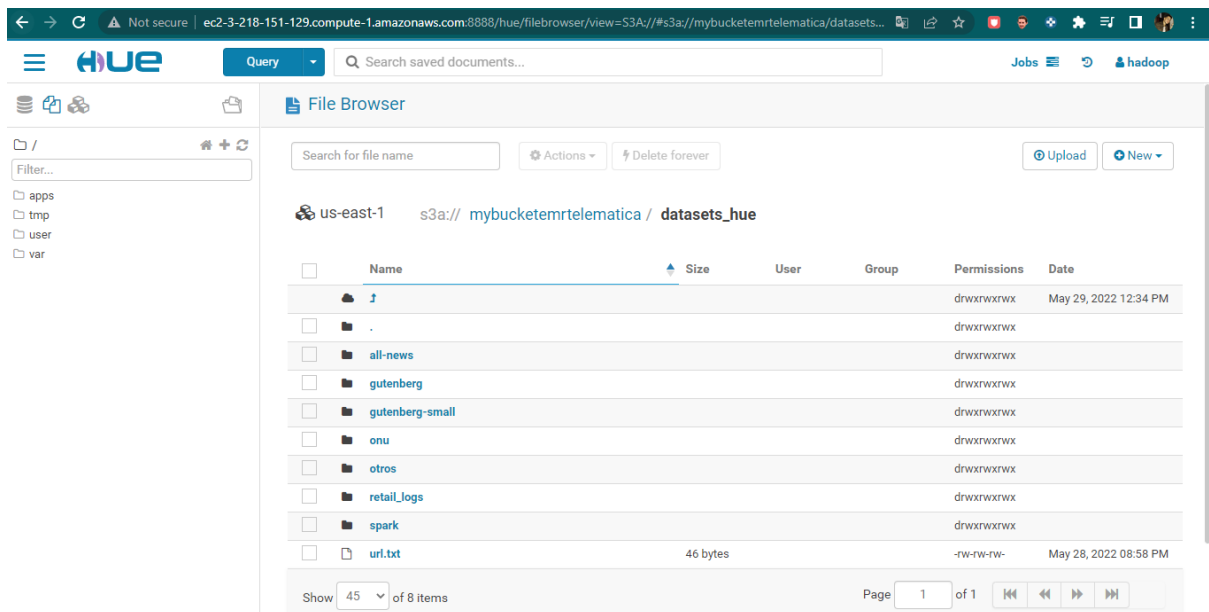
Aquí podremos ver el bucket que en la guía anterior creamos, ingresamos a el:



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:

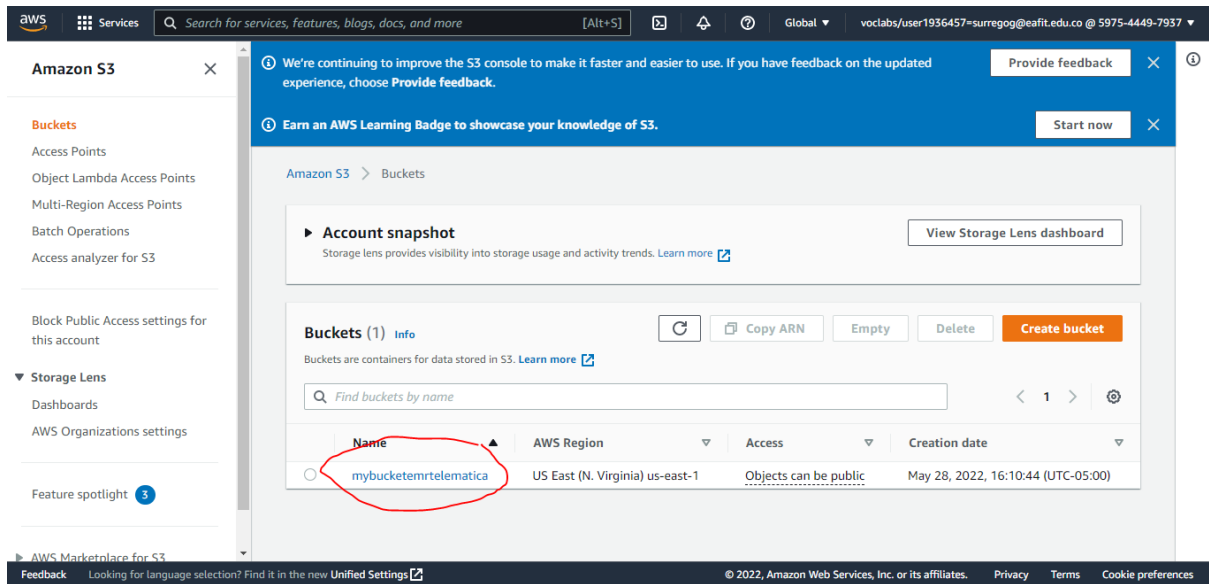


Con esto finalizamos la montada de archivos en S3, ahora en el siguiente capitulo vamos a ponerlos públicos para que cualquiera con una URL de internet pueda acceder a ellos de forma más fácil

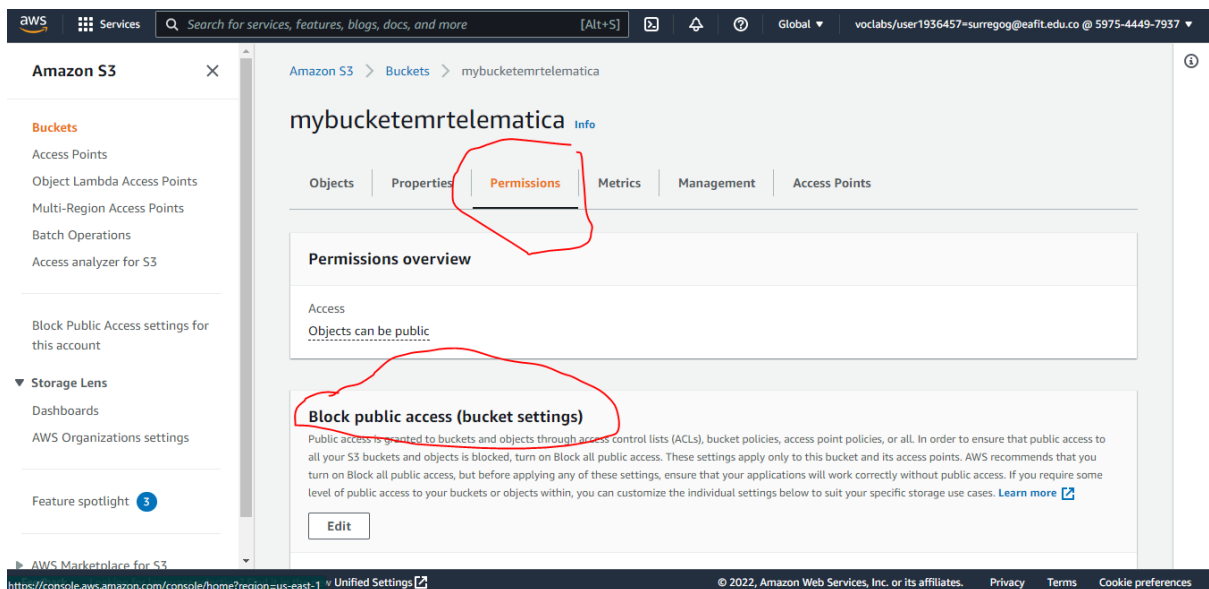
Bucket para lectura publica

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:

Block public access (bucket settings)

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. [Learn more](#)

☐ Block *all* public access

Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

☐ Block public access to buckets and objects granted through *new* access control lists (ACLs)

S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.

☐ Block public access to buckets and objects granted through *any* access control lists (ACLs)

S3 will ignore all ACLs that grant public access to buckets and objects.

☐ Block public access to buckets and objects granted through *new* public bucket or access point policies

S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.

☐ Block public and cross-account access to buckets and objects through *any* public bucket or access point policies

S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

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:

The screenshot shows the AWS Management Console interface for an S3 bucket. The left sidebar contains navigation links for Buckets, Access Points, Object Lambda Access Points, Multi-Region Access Points, Batch Operations, Access analyzer for S3, Block Public Access settings for this account, Storage Lens, Dashboards, AWS Organizations settings, and Feature spotlight. The main content area is titled 'Object Ownership' and 'Access control list (ACL)'. Under 'Object Ownership', there is an 'Edit' button and a section titled 'Object Ownership' with the text 'Bucket owner preferred'. Below this, it states 'ACLs are enabled and can be used to grant access to this bucket and its objects. If new objects written to this bucket specify the bucket-owner-full-control canned ACL, they are owned by the bucket owner. Otherwise, they are owned by the object writer.' Under 'Access control list (ACL)', there is an 'Edit' button and a section titled 'The console displays combined access grants for duplicate grantees'. At the bottom, there is a table with columns 'Grantee', 'Objects', and 'Bucket ACL'. The first row shows 'Bucket owner (your AWS account)'.

Object Ownership [Info](#)

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

[Edit](#)

Object Ownership

Bucket owner preferred

ACLs are enabled and can be used to grant access to this bucket and its objects. If new objects written to this bucket specify the bucket-owner-full-control canned ACL, they are owned by the bucket owner. Otherwise, they are owned by the object writer.

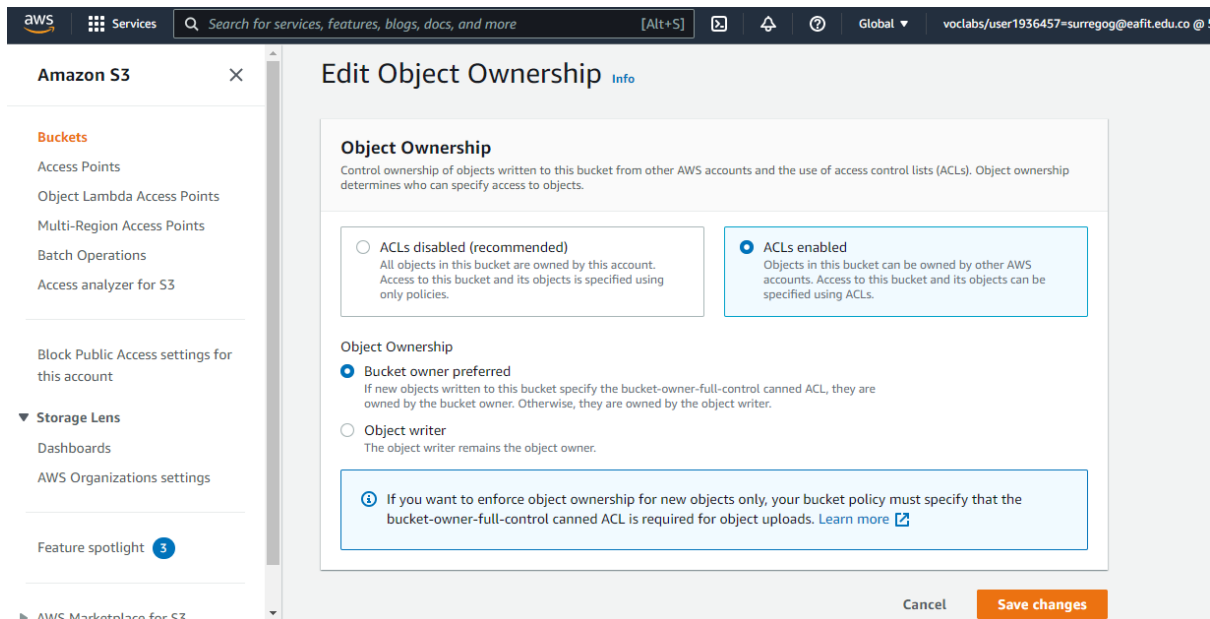
Access control list (ACL) [Learn more](#) [Edit](#)

Grant basic read/write permissions to other AWS accounts.

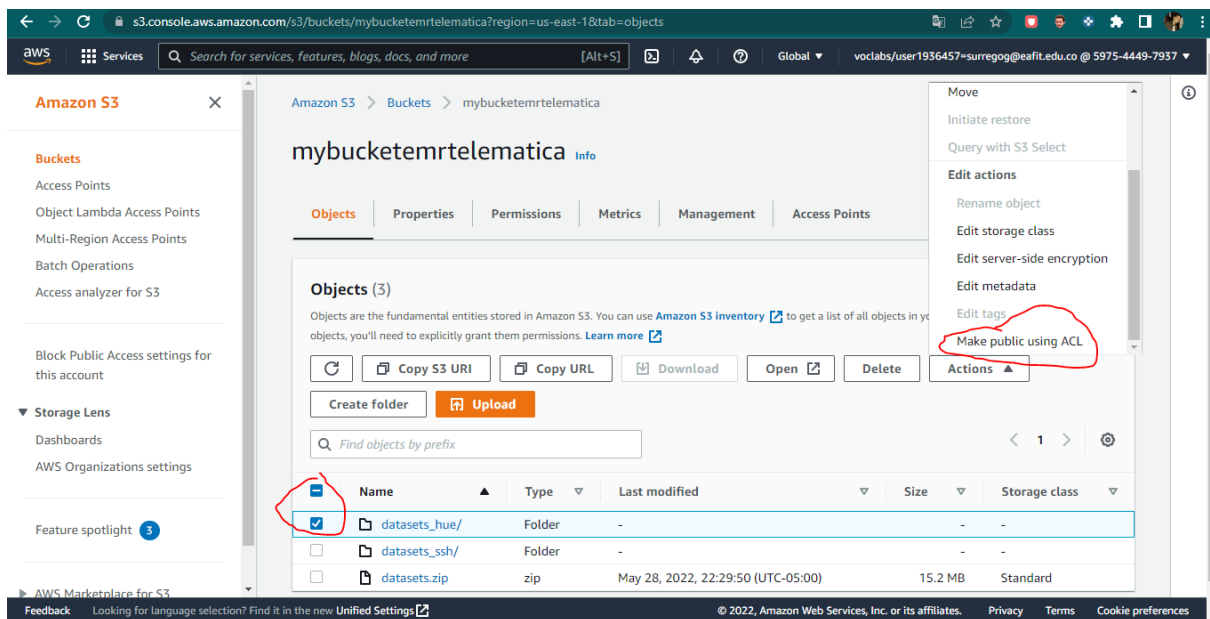
The console displays combined access grants for duplicate grantees

To see the full list of ACLs, use the Amazon S3 REST API, AWS CLI, or AWS SDKs.

Grantee	Objects	Bucket ACL
Bucket owner (your AWS account)		



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.