UNIVERSIDAD EAFIT
Laboratorio de AWS S3-Glue-Athena
Datalake y Motor de consulta SQL

Todos estos laboratorios se realizan en AWS ACADEMY:

Parte 1: ALMACENAMIENTO DE DATOS EN AWS S3

Fuentes de datos: datasets de:
https://github.com/st0263eafit/st0263-252.git

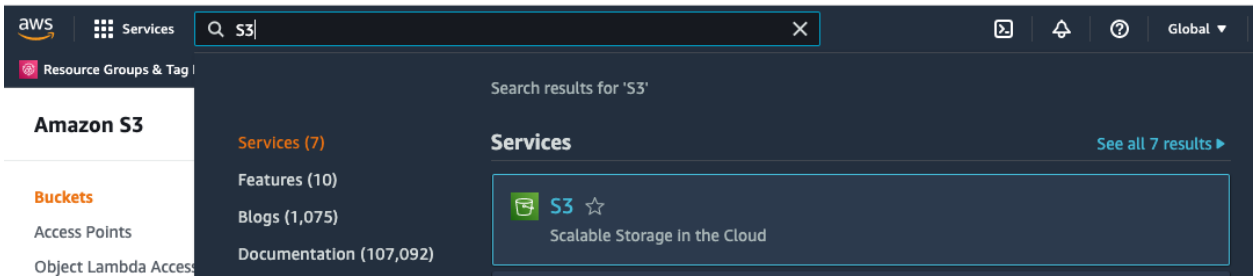
Realizar la ingesta manual y almacenamiento de los datos en AWS S3.

Aprenderá a gestionar los datos desde la página web de AWS.

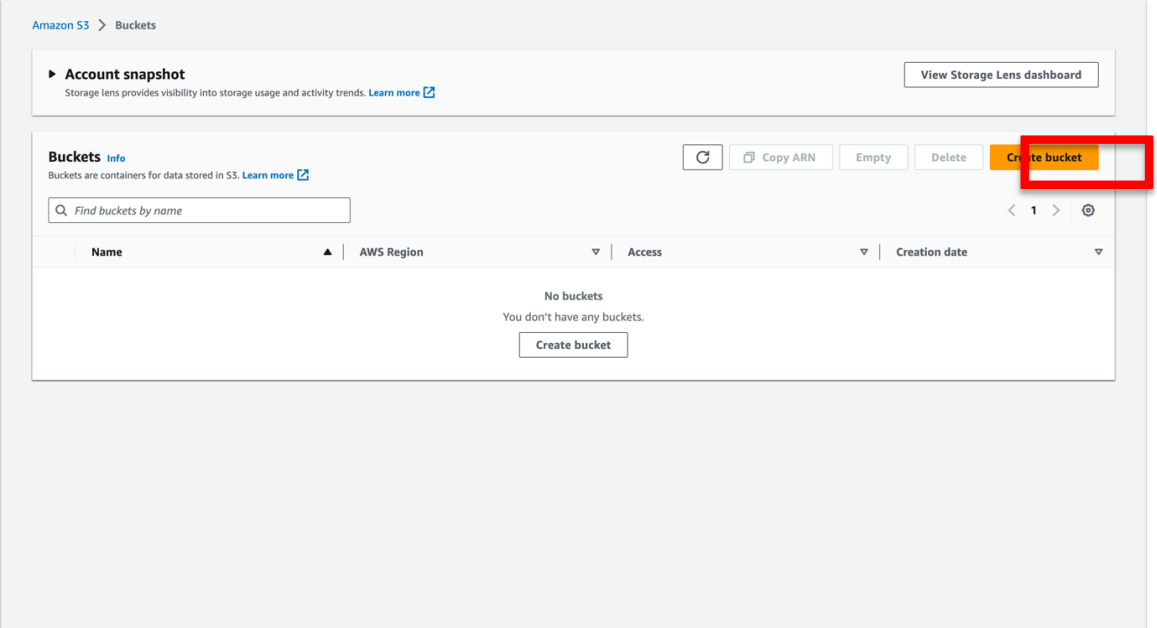
Copiará los datos de: https://github.com/st0263eafit/st0263-252/tree/main/datasets

Paso 1: crear el nuevo bucket

1. Buscar el servicio S3 en el cuadro de búsqueda de servicios AWS

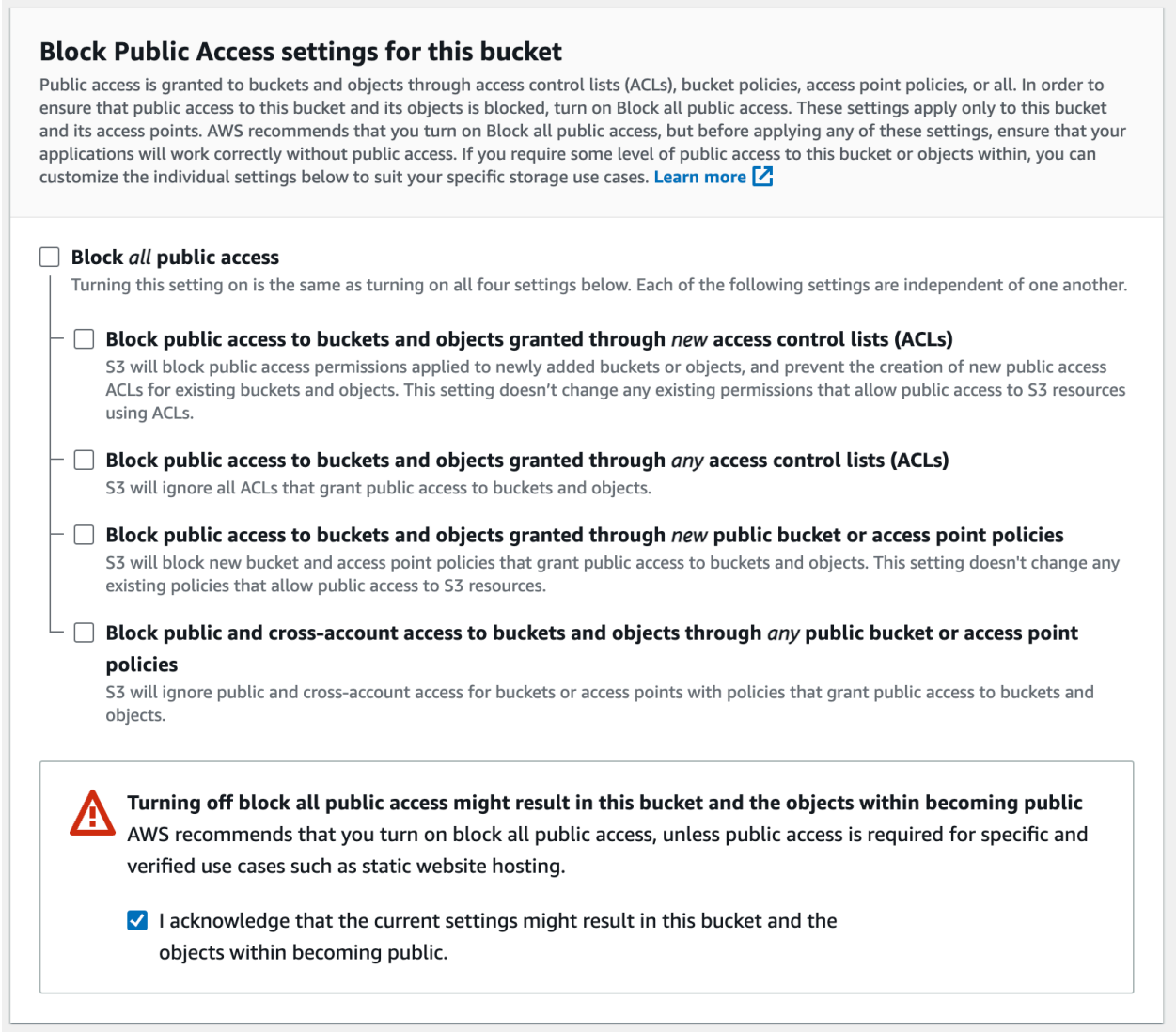


2. Click en **Create Bucket**



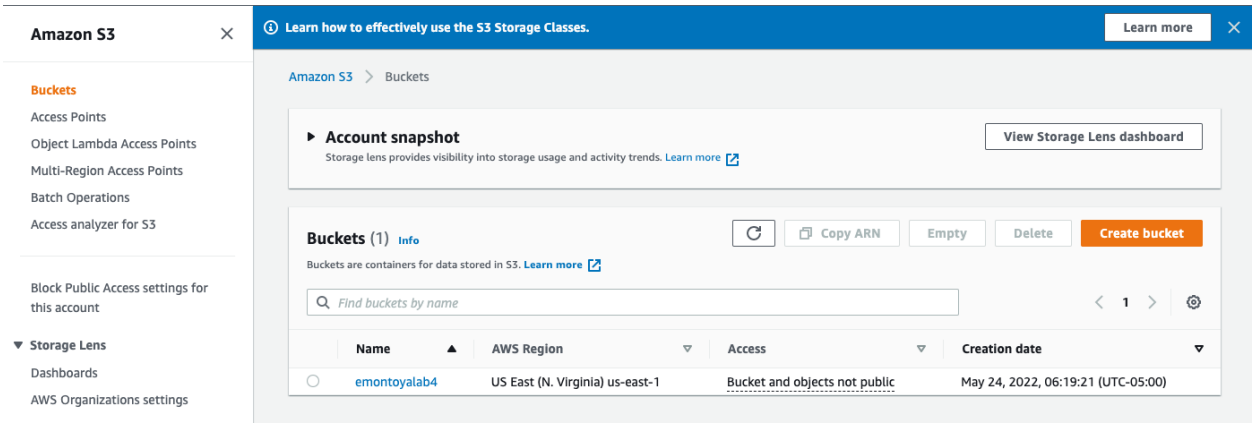
Parámetros:

- **Bucket name:** <username>lab1 (ojo: por favor reemplace el *username* por su propio usuario, ejemplo el profesor lo llamaría: 'your-username-datalake'. Estos nombres son únicos a nivel mundial, no se pueden repetir.
- **AWS Region**: us-east-1
- **Deshabilitar** '*Configuración de bloqueo de acceso público para este bucket' Debe verse asi:*



NOTA: Notese el ultimo check de la imagen, deben marcarlo para que los deje crear el bucket.

3. En el panel izquierdo, le da **Buckets** y debe aparecer algo así:



Paso 2: copiar los archivos al bucket

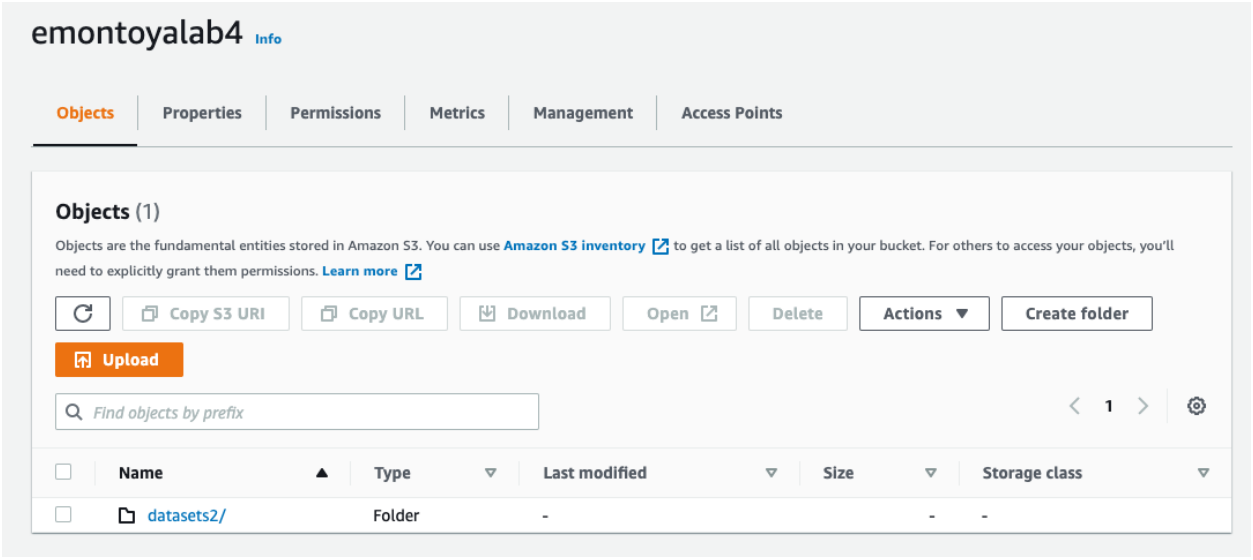
1. Click en el bucket nuevo creado, ej: your-username-datalake
2. crear directorios manualmente

   Click en **Create folder**:

   Folder name: datasets2

   Todo lo demás por defecto.

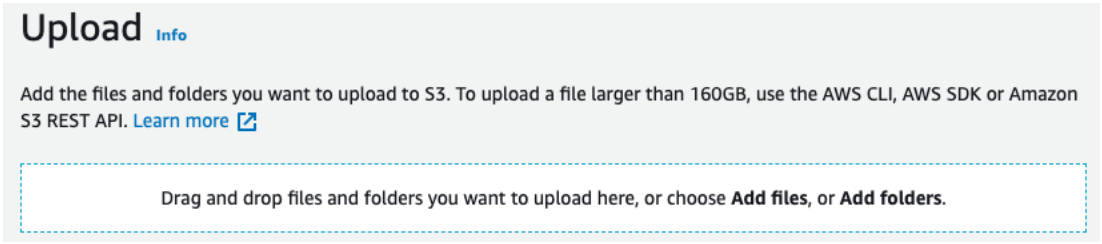
   Click en **Create folder**



3. subir archivos a datasets2:

   Seleccione ‘datasets2’
   Click en **Upload**

   Puede ‘Drag and drop’ (seleccionar y soltar) archivos o directorios

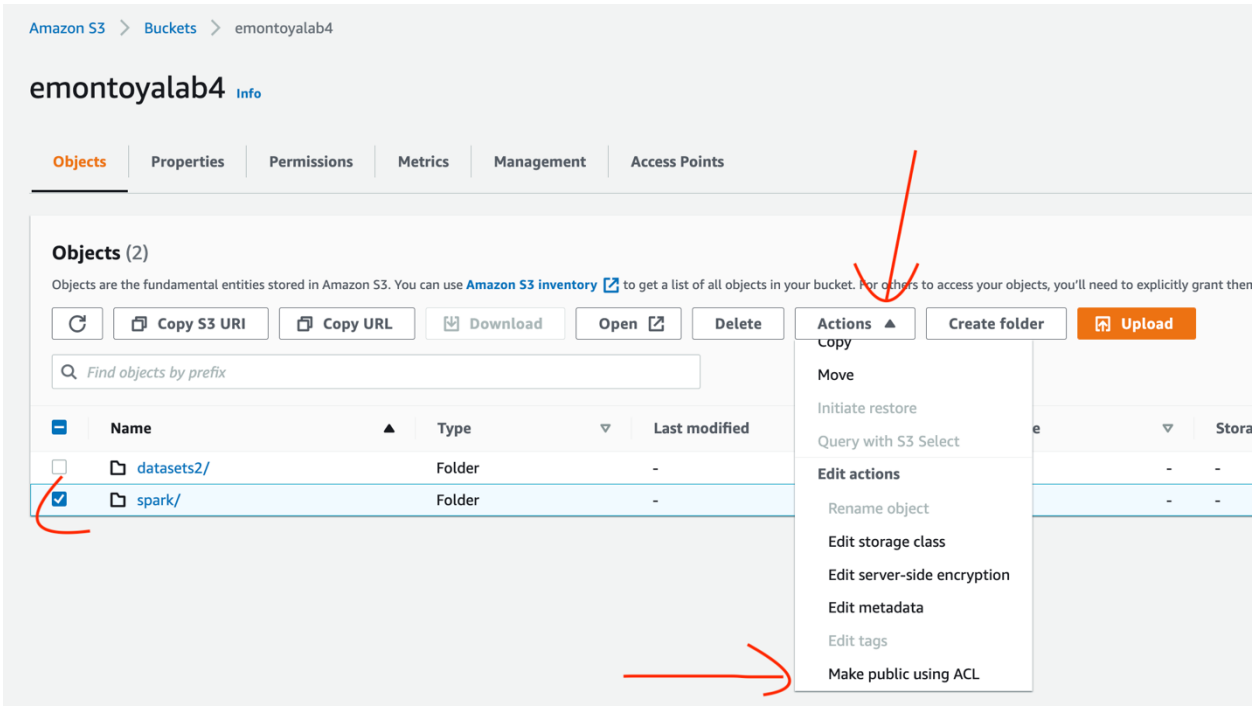


   Puede **Add files**:

   Puede **Add folder**:

   Explore las 3 opciones, luego navegue por el bucket y sus nuevos directorios.

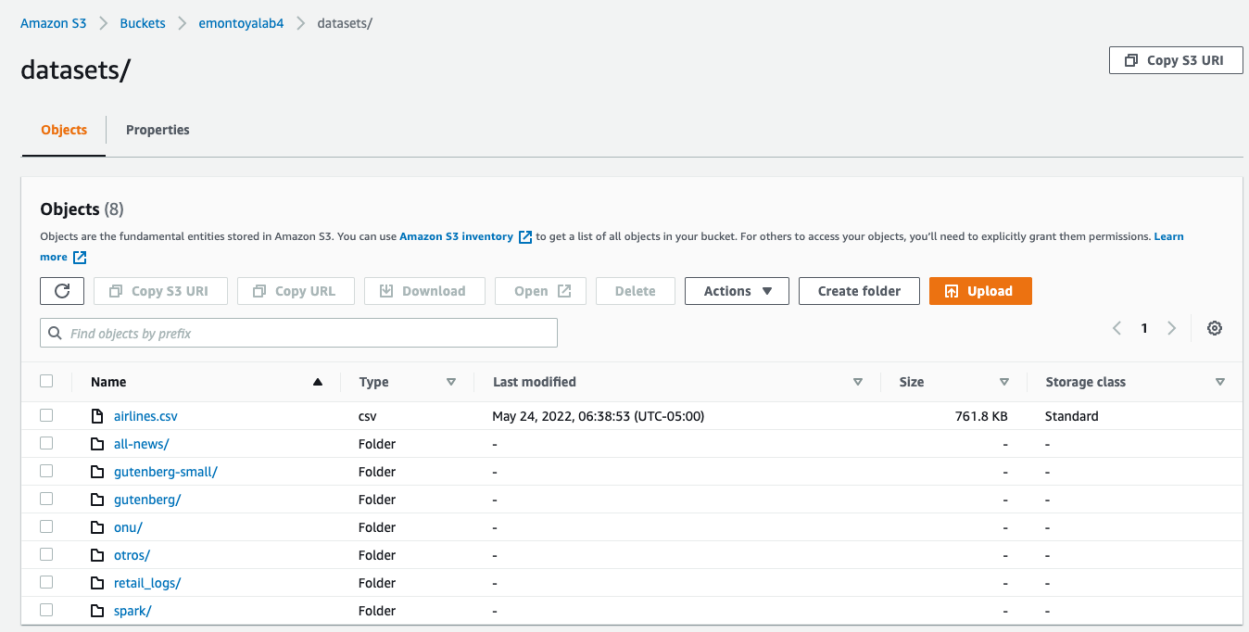
   Hacerlo público luego de subir el archivo/carpeta.



Seguir con ‘datasets’

1. subir nuevamente los archivos a un folder nuevo a llamarse ‘datasets’:

2. verificar que ha creado un folder con todos los archivos y quedará así:



Dele click en: **Copy S3 URI**:

Y saldrá algo así: s3://your-username-datalake/datasets/

Esta URL es la que utilizará el profesor para verificar que completo bien esta parte del laboratorio

Parte 2: IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS ATHENA.

Ver video:

- Lab-AWS-S3-Glue-Athena: https://youtu.be/VbyVaAMF9EA

Fuentes de datos: datasets de:

https://github.com/st0263eafit/st0263-252.git

datos específicos de: onu y tickit

Ingesta de datos: manual a AWS S3

Almacenamiento: en datalake con AWS S3

Catalogación: con AWS Glue, creación de las 2 bases de datos (onudb y tickitdb) y las respectivas tablas.

Consultas: Con AWS Athena SQL realizar diferentes consultas a diferentes tablas.