

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>

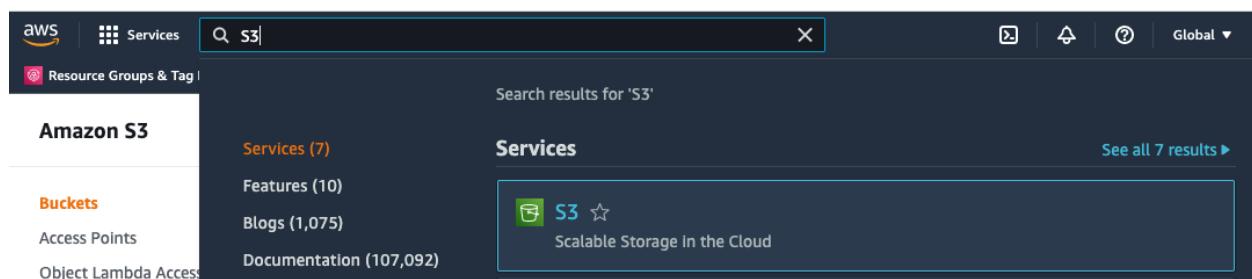
Realizar la ingestión manual y almacenamiento de los datos en AWS S3.

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

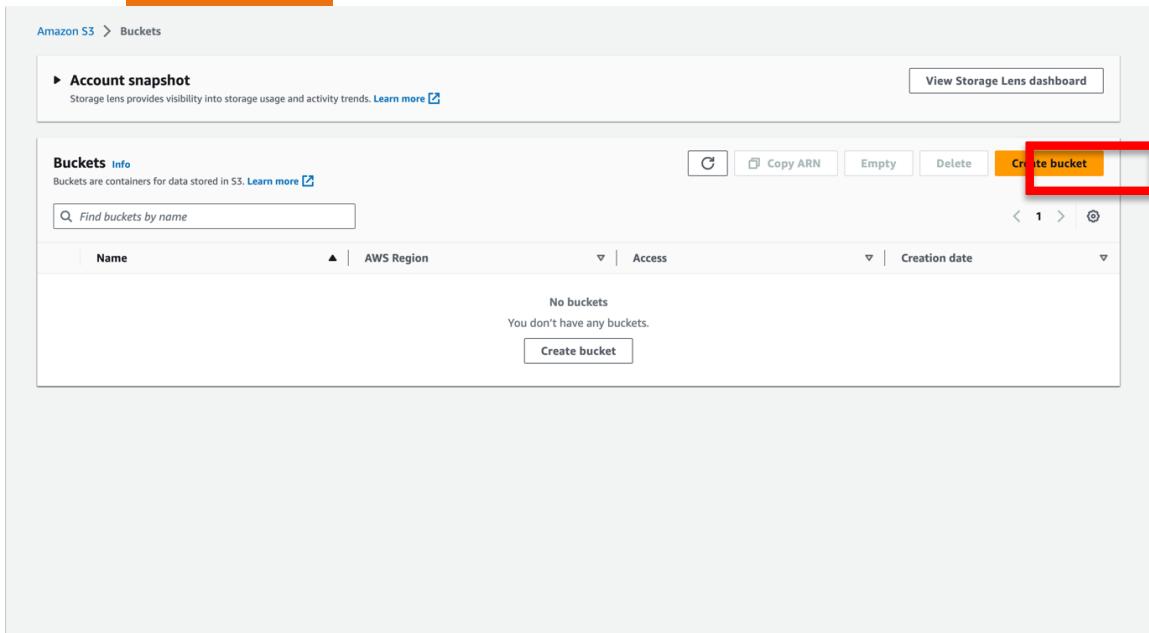
Copiará los datos de: <https://github.com/st0263eafit/st0263-252/tree/main/datasets>

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 así:

Block Public Access settings for this bucket

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 this bucket and its 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 this bucket 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.



Turning off block all public access might result in this bucket and the objects within becoming public

AWS recommends that you turn on block all public access, unless public access is required for specific and verified use cases such as static website hosting.

- I acknowledge that the current settings might result in this bucket and the objects within becoming public.

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:

The screenshot shows the Amazon S3 Buckets page. On the left sidebar, under the 'Buckets' section, there is a link to 'Storage Lens'. The main content area displays an 'Account snapshot' with a 'View Storage Lens dashboard' button. Below it, a table lists the bucket 'emontoyalab4' with details: Name (emontoyalab4), AWS Region (US East (N. Virginia) us-east-1), Access (Bucket and objects not public), and Creation date (May 24, 2022, 06:19:21 (UTC-05:00)).

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**

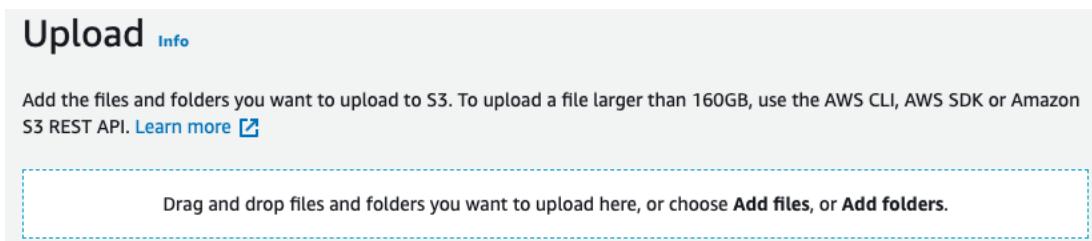
The screenshot shows the Amazon S3 Objects page for the bucket 'emontoyalab4'. At the top, there are tabs for 'Objects', 'Properties', 'Permissions', 'Metrics', 'Management', and 'Access Points'. The 'Objects' tab is selected. The main content area shows a table with one item: 'datasets2/' (Type: Folder). There are buttons for 'Upload', 'Actions', and 'Create folder' at the top of the list. A search bar at the bottom allows filtering by prefix.

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.

Name	Type	Last modified
datasets2/	Folder	-
<input checked="" type="checkbox"/> spark/	Folder	-

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:

Amazon S3 > Buckets > emontoyalab4 > datasets/

datasets/

Copy S3 URI

Objects (8)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	airlines.csv	CSV	May 24, 2022, 06:38:53 (UTC-05:00)	761.8 KB	Standard
<input type="checkbox"/>	all-news/	Folder	-	-	-
<input type="checkbox"/>	gutenberg-small/	Folder	-	-	-
<input type="checkbox"/>	gutenberg/	Folder	-	-	-
<input type="checkbox"/>	onu/	Folder	-	-	-
<input type="checkbox"/>	otros/	Folder	-	-	-
<input type="checkbox"/>	retail_logs/	Folder	-	-	-
<input type="checkbox"/>	spark/	Folder	-	-	-

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.