

PROYECTO 1 – PARTE 2: MANUAL DE REPRODUCIBILIDAD

ALEJANDRO ARANGO GIRALDO

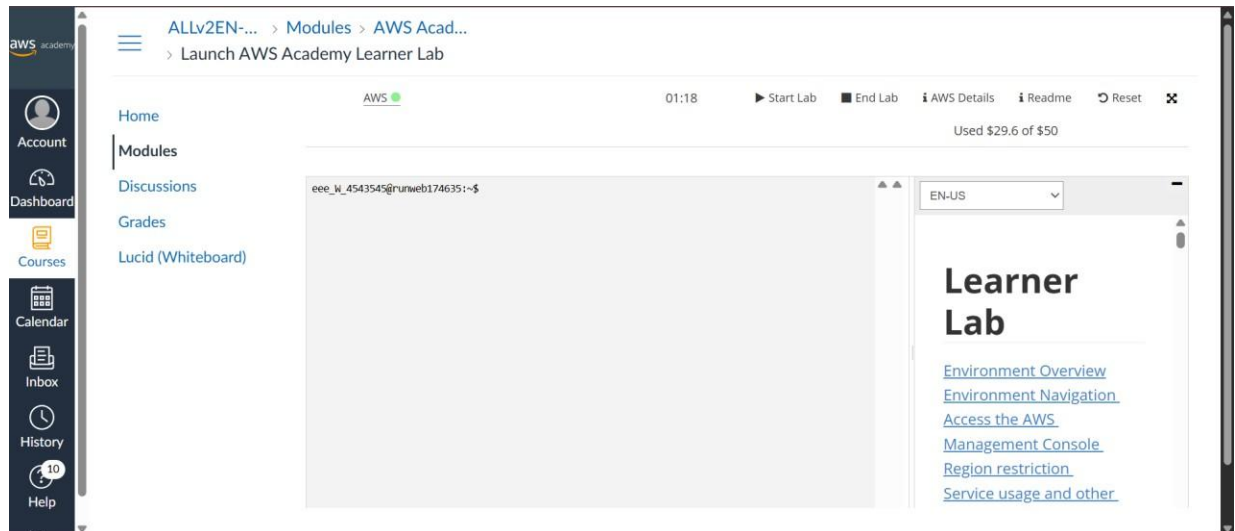
Línea de énfasis en Ciencias de los Datos

ST1801 – Almacenamiento y recuperación de la información

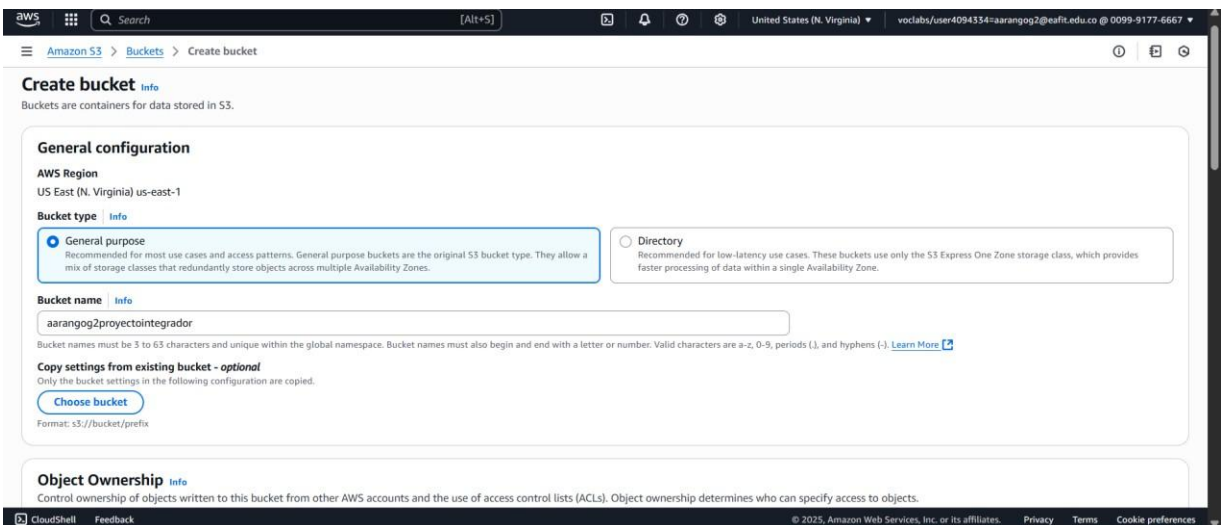
**UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
CARRERA DE INGENIERÍA MECÁNICA
MEDELLÍN
2025 - 1**

Para reproducir este proyecto, se deben seguir los siguientes pasos de manera secuencial:

1. Activar una sesión de AWS Academy Learner Lab, o, si se cuenta con otras credenciales, activar la sesión.



2. Entrar a la interfaz web de AWS y buscar el servicio de AWS S3. Oprimir en la opción “Create bucket” y mantener la siguiente configuración:



Si se cambia el nombre del bucket, se deben cambiar las rutas definidas en los códigos para leer y guardar los archivos allí almacenados.

aws [Search] [Alt+S] United States (N. Virginia) voclabs/user4094334-aarangog2@eafit.edu.co @ 0099-9177-6667

Amazon S3 Buckets Create bucket

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.

☒ **ACLs disabled (recommended)**
All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

☐ **ACLs enabled**
Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

Object Ownership
 Bucket owner enforced

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.

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

aws [Search] [Alt+S] United States (N. Virginia) voclabs/user4094334-aarangog2@eafit.edu.co @ 0099-9177-6667

Amazon S3 Buckets Create bucket

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. [Learn more](#)

Bucket Versioning

☒ **Disable**

☐ **Enable**

Tags - optional (0)

You can use bucket tags to track storage costs and organize buckets. [Learn more](#)

No tags associated with this bucket.

[Add tag](#)

Default encryption info

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type info

☒ **Server-side encryption with Amazon S3 managed keys (SSE-S3)**
Server-side encryption with AWS Key Management Service keys (SSE-KMS)

☐ **Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)**
Secure your objects with two separate layers of encryption. For details on pricing, see DSSE-KMS pricing on the Storage tab of the [Amazon S3 pricing page](#).

Bucket Key
Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. [Learn more](#)

☐ **Disable**

☒ **Enable**

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

3. Una vez creado el bucket “aarangog2proyectointegrador”, se crean las carpetas: zona raw, zona trusted y zona refined.

aws [Search] [Alt+S] United States (N. Virginia) voclabs/user4094334-aarangog2@eafit.edu.co @ 0099-9177-66

Amazon S3

Amazon S3

- General purpose buckets
- Directory buckets
- Table buckets
- Access Grants
- Access Points
- Object Lambda Access Points
- Multi-Region Access Points
- Batch Operations
- IAM Access Analyzer for S3

Block Public Access settings for this account

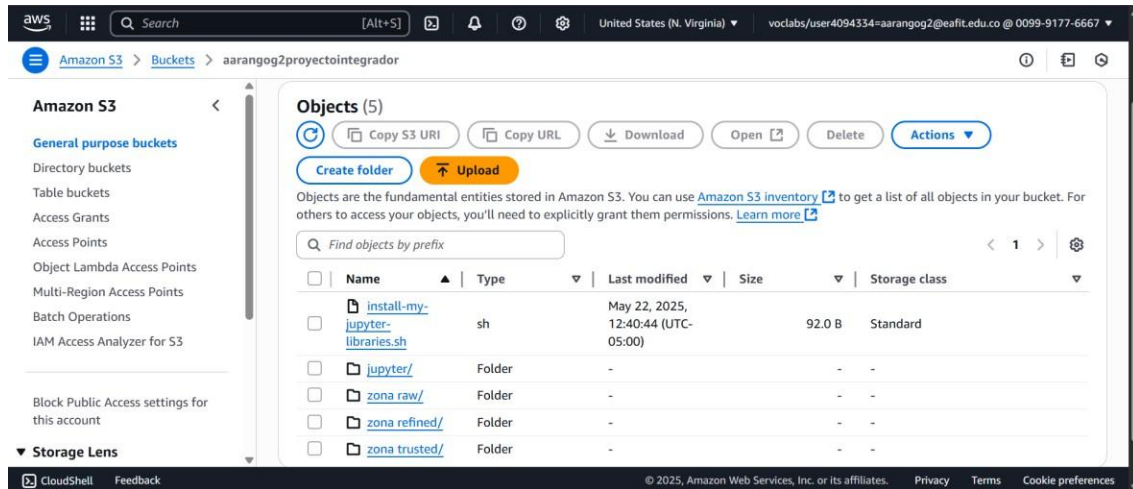
General purpose buckets (3) Info All AWS Regions

[Refresh](#) [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

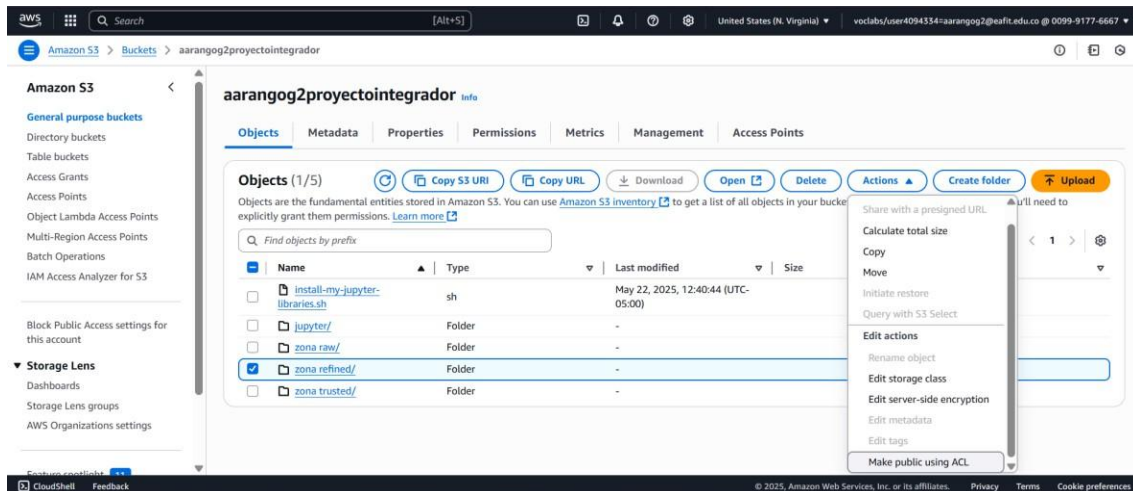
Buckets are containers for data stored in S3.

	Name	AWS Region	IAM Access Analyzer	Creation date
<input type="radio"/>	aarangog2lab1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 17, 2025, 16:04:58 (UTC-05:00)
<input type="radio"/>	aarangog2proyectointegrador	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 20, 2025, 18:08:42 (UTC-05:00)
<input type="radio"/>	aws-logs-009991776667-us-east-1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 20, 2025, 18:21:36 (UTC-05:00)

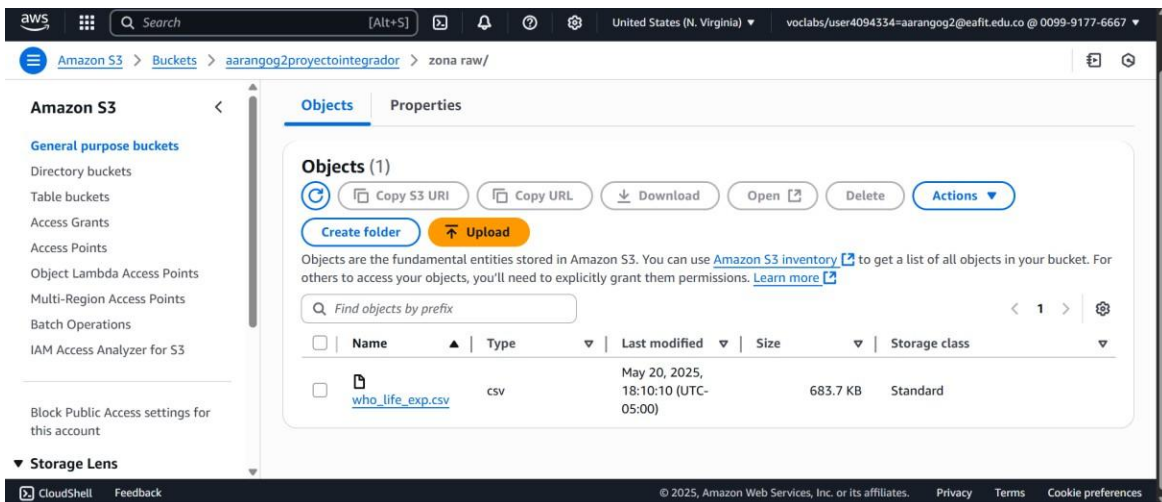
CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences



Las zonas “trusted” y “refined” deben volverse públicas para su futura consulta. Esto se realiza por medio de la acción “Make public using ACL”.



4. En la carpeta “zoned raw”, se deben cargar manualmente los datos crudos. En nuestro caso, sería el archivo “who_life_exp.csv”.



5. Buscar en la interfaz web el servicio EMR y crear un cluster con la siguiente configuración:

This screenshot shows the 'Create cluster' page in the AWS EMR console, specifically the 'Name and applications' step. The cluster name is 'aarangog2 Proyecto Integrador'. The Amazon EMR release label is 'emr-7.8.0'. Under 'Application bundle', the 'Custom' option is selected. A list of applications is shown with checkboxes: AmazonCloudWatchAgent, HCatalog 3.1.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Oozie 5.2.1, Phoenix 5.2.1, Pig 0.17.0, Presto 0.287, Spark 3.5.4, Tez 0.10.2, Trino 467, Zeppelin 0.11.1, and ZooKeeper 3.9.3. Under 'AWS Glue Data Catalog settings', 'Use for Hive table metadata' and 'Use for Spark table metadata' are checked. The 'Summary' panel on the right shows the configuration details. At the bottom, there are 'Cancel' and 'Clone cluster' buttons.

This screenshot shows the 'Create cluster' page in the AWS EMR console, specifically the 'Cluster configuration' step. The 'Operating system options' section has 'Amazon Linux release' selected. The 'Cluster configuration' section has 'Uniform instance groups' selected. Under 'Uniform instance groups', the 'Primary' group is configured with 'm5.xlarge' instance type. The 'Node configuration - optional' section is currently empty. The 'Summary' panel on the right shows the configuration details. At the bottom, there are 'Cancel' and 'Clone cluster' buttons.

This screenshot shows the 'Create cluster' page in the AWS EMR console, specifically the 'Core and task node configuration' step. The 'Core' section shows the 'm5.xlarge' instance type. The 'Task 1 of 1' section shows the 'm5.xlarge' instance type. The 'EBS root volume' section shows the 'Size (GiB)' as 15, 'IOPS' as 3000, and 'Throughput (MiB/s)' as 125. The 'Summary' panel on the right shows the configuration details. At the bottom, there are 'Cancel' and 'Clone cluster' buttons.

EBS root volume
EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#)

Size (GiB) IOPS Throughput (MiB/s)
 15 - 100 GiB per volume. General Purpose SSD (gp3)
 3000 - 16000 IOPS per volume. Choose a maximum ratio of 500:1 between IOPS and volume size.
 125 - 1000 MiB/s per volume. Choose a maximum ratio of 0.25:1 between throughput and IOPS.

Cluster scaling and provisioning - required
Choose how Amazon EMR should size your cluster.
Choose an option

☒ Set cluster size manually
Use this option if you know your workload patterns in advance.

☐ Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

☐ Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.

Provisioning configuration
Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

Name	Instance type	Instance(s) size	Use Spot purchasing option
Task - 1	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>
Core	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>

Networking - required
Choose the network settings that determine how you and other entities communicate with your cluster.

Summary
Name and applications - required
 Name: aarangog2 Projecto Integrador
 Amazon EMR release: emr-7.8.0
 Application bundle: Custom (HCatalog 3.1.3, Hadoop 3.4.1, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2...)
Cluster configuration - required
 Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)
Cluster scaling and provisioning - required
 Provisioning configuration: Core size: 1 instance, Task size: 1 instance

Networking - required
Choose the network settings that determine how you and other entities communicate with your cluster.

Virtual private cloud (VPC)
vpc-05fd9d2bcd64089b6 [Browse](#) [Create VPC](#)

Subnet
subnet-004789ee942853556 [Browse](#) [Create subnet](#)

EC2 security groups (firewall)

Steps (0)
Use commands and scripts to tell your cluster where to find and how to process your data. Steps run consecutively unless you enable the Concurrency option. [Remove](#) [Edit](#) [Add](#)

Cluster termination and node replacement
Choose termination settings and protect your cluster from accidental shutdown. [Remove](#) [Edit](#) [Add](#)

Bootstrap actions (1)
Use bootstrap actions to install software or customize your instance configuration. [Remove](#) [Edit](#) [Add](#)

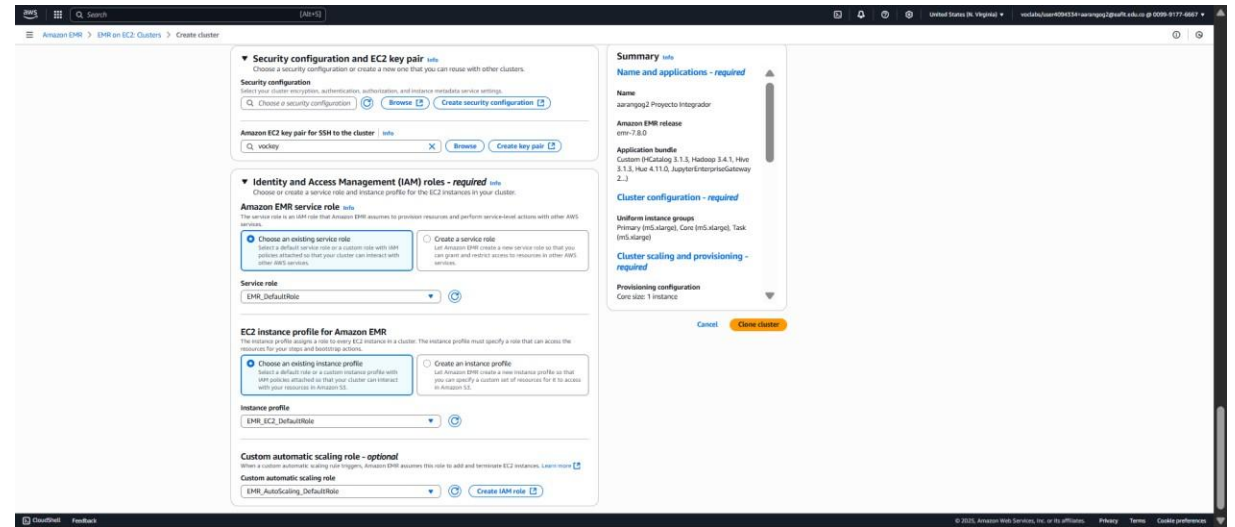
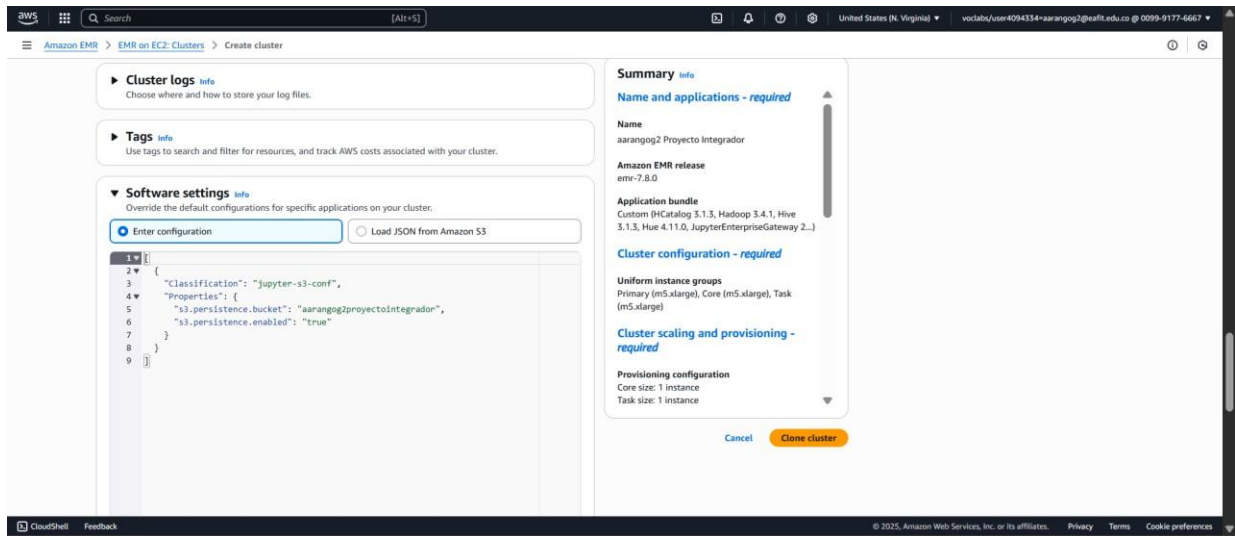
Name	Amazon S3 location	Argument
install-my-jupyter-libraries	s3://aarangog2projectointegrator/install-my-jupyter-libraries.sh	

Cluster logs

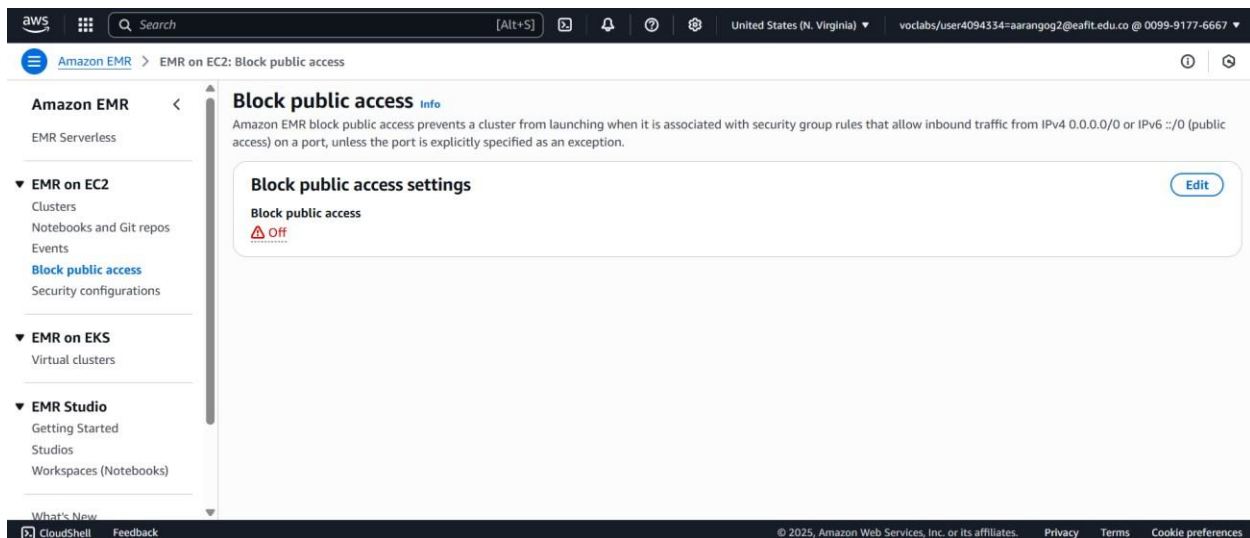
Summary
Name and applications - required
 Name: aarangog2 Projecto Integrador
 Amazon EMR release: emr-7.8.0
 Application bundle: Custom (HCatalog 3.1.3, Hadoop 3.4.1, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2...)
Cluster configuration - required
 Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)
Cluster scaling and provisioning - required
 Provisioning configuration: Core size: 1 instance, Task size: 1 instance

Adicionar “Bootstrap actions” es opcional en este caso de uso, puesto que las librerías utilizadas en las diferentes fases son SparkML y SparkSQL, las cuales forman parte de Apache Spark. Sin embargo, de necesitarse otras librería, se puede crear un código como el siguiente, guardarlo con la extensión “sh”, y añadirlo a la sección “Bootstrap actions” en la configuración del cluster.

```
File Edit Selection View Go Run ... Search
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More
install-my-jupyter-libraries.sh x Release Notes: 1.100.2
C:\> Users > aleje > OneDrive - Universidad EAFIT > Escritorio > EAFIT > 2025-1 > Proyecto Integrador > Entregas > $ install-my-jupyter-libraries.sh
1 #!/bin/bash
2
3 sudo python3 -m pip install shap seaborn matplotlib scipy scikit-learn pandas
```

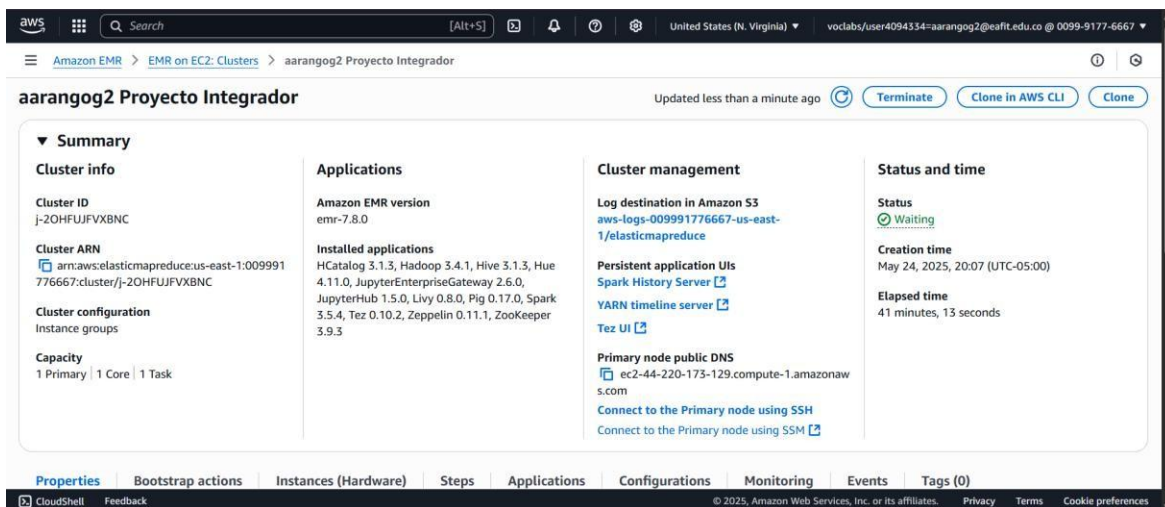



6. Entrar a la opción “Block public access” del menú de la izquierda y abrir todos los puertos TCP para acceso al clúster de la siguiente manera:

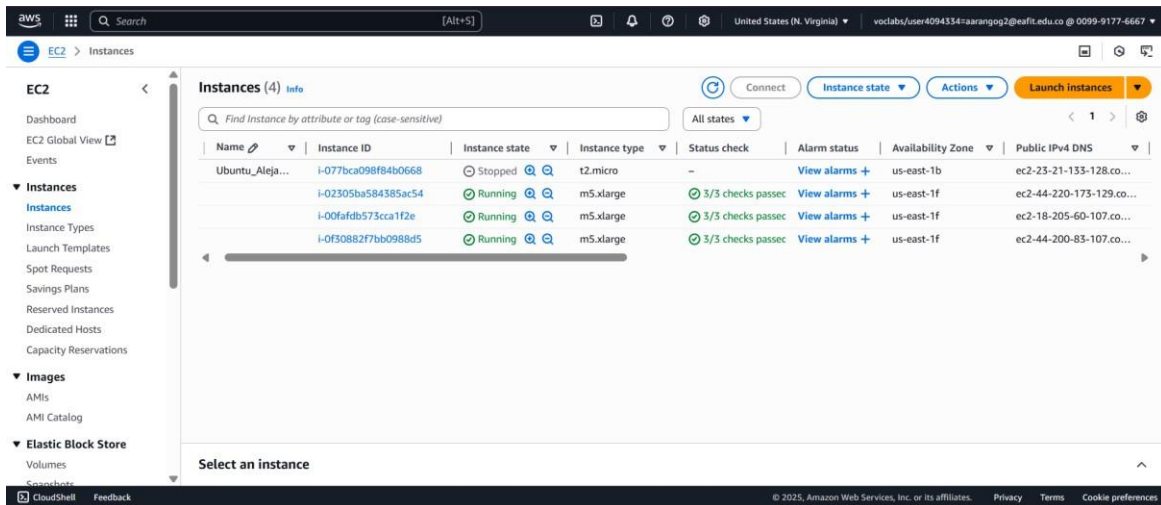


7. Abrir los puertos de las aplicaciones de hadoop/Spark en el Security Group del nodo MASTER del clúster como se muestra a continuación:

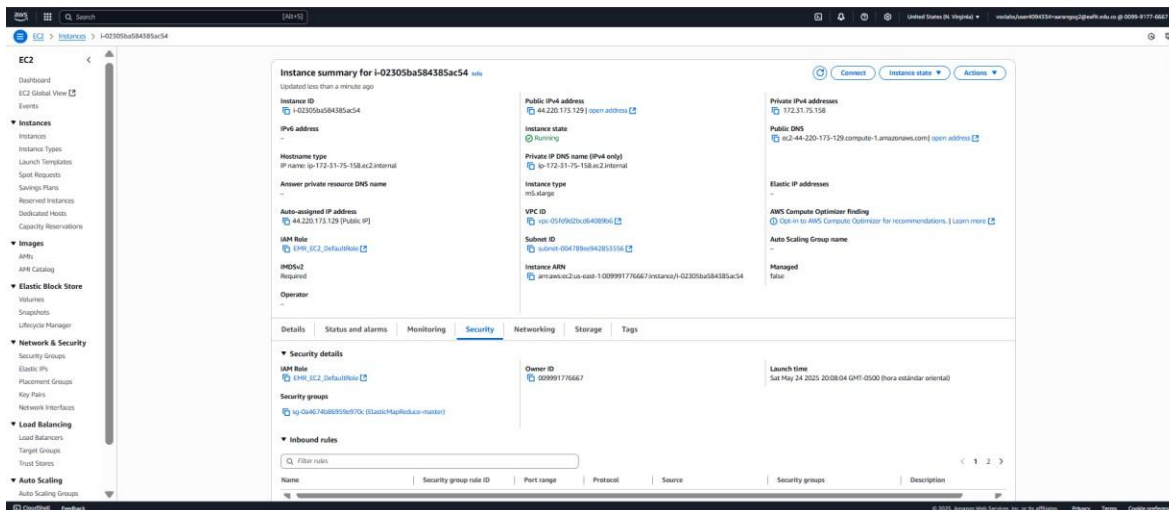
7.1 Identificar el nodo primario del cluster recién creado, el cual se muestra en “Primary node public DNS”.



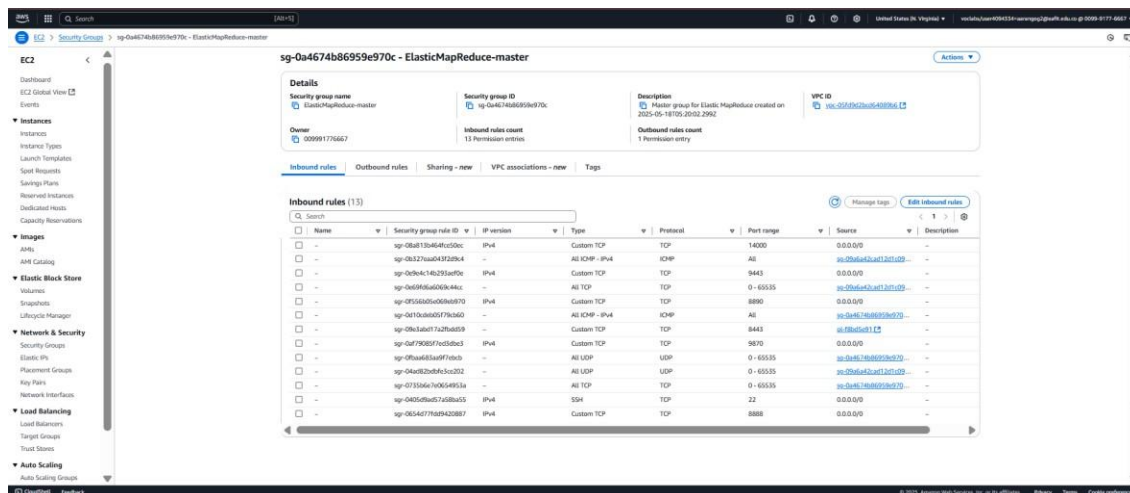
7.2 Buscar en la interfaz web el servicio E2C, donde se encontrarán tres máquinas. Abrir la que tenga el valor de la columna “Public IPv4 DNS” igual al valor del “Primary node public DNS” del cluster.



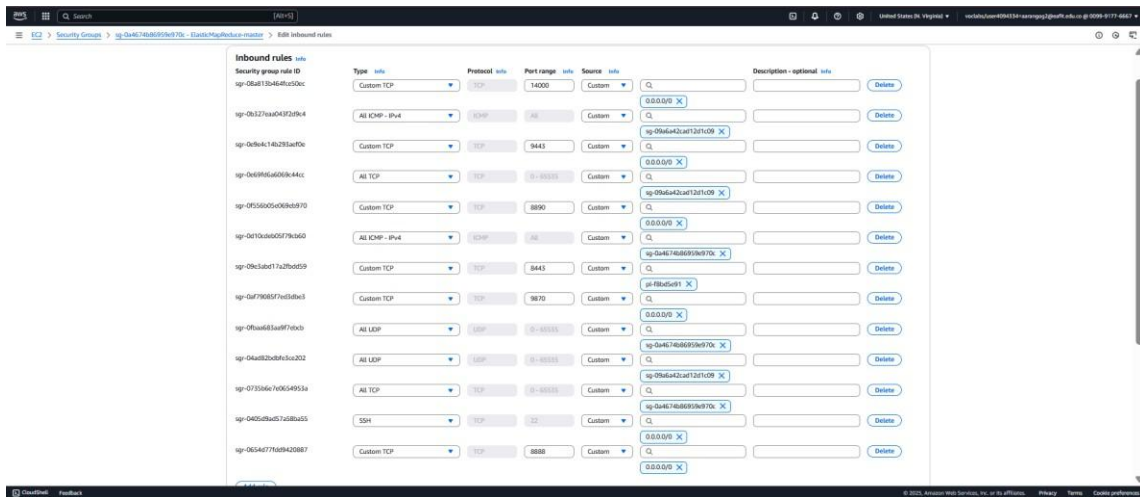
7.3 Entrar a la pestaña de seguridad de la Instancia EC2 del nodo master y abrir la opción “Security groups”.



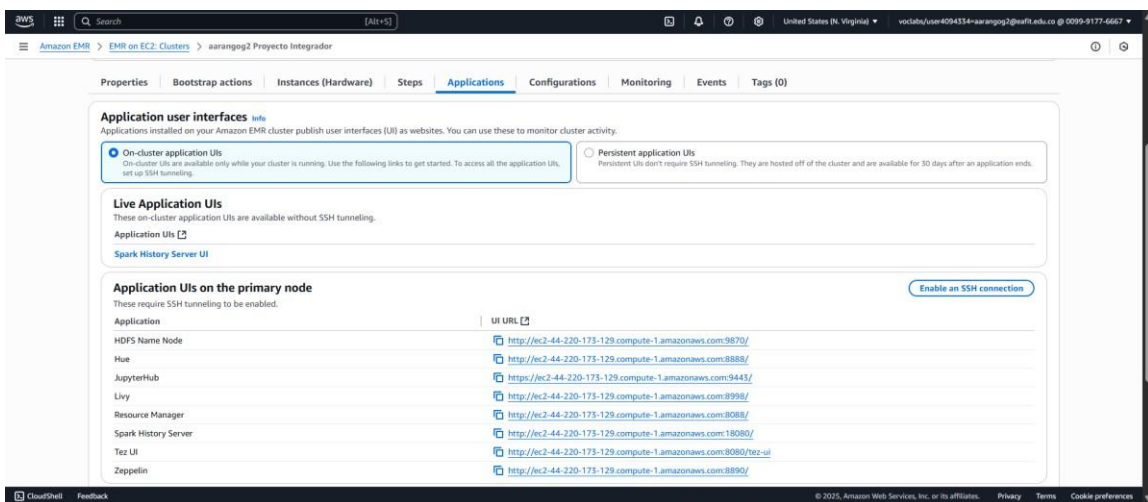
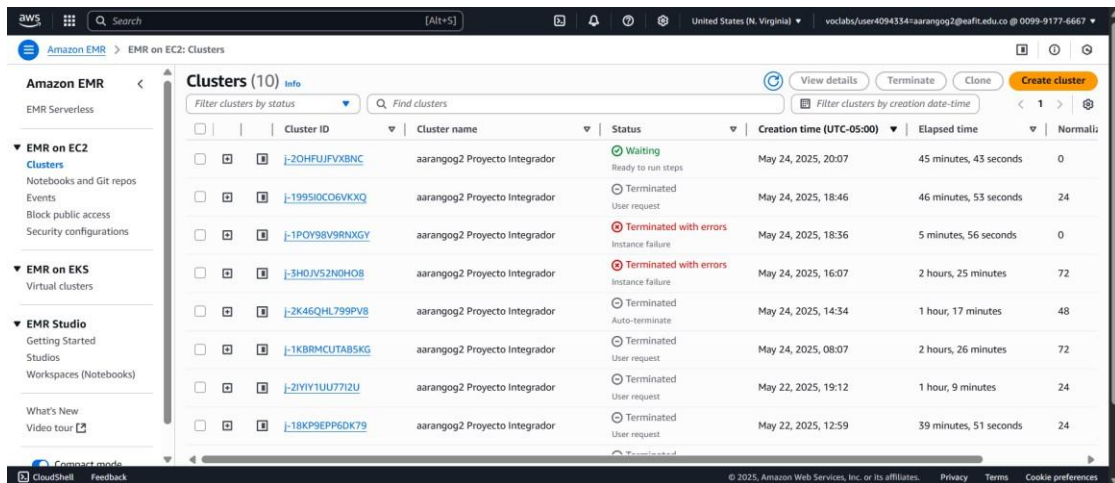
7.4 Entrar a la opción “Edit inbound rules”.



7.5 Habilitar los nodos: 22, 14000, 9870, 8888, 9443, y 8890.



8. Con el cluster en estado “Waiting”, seleccionarlo y en el menú “Applications”, oprimir el UI que lleva al servicio de Jupyterhub.



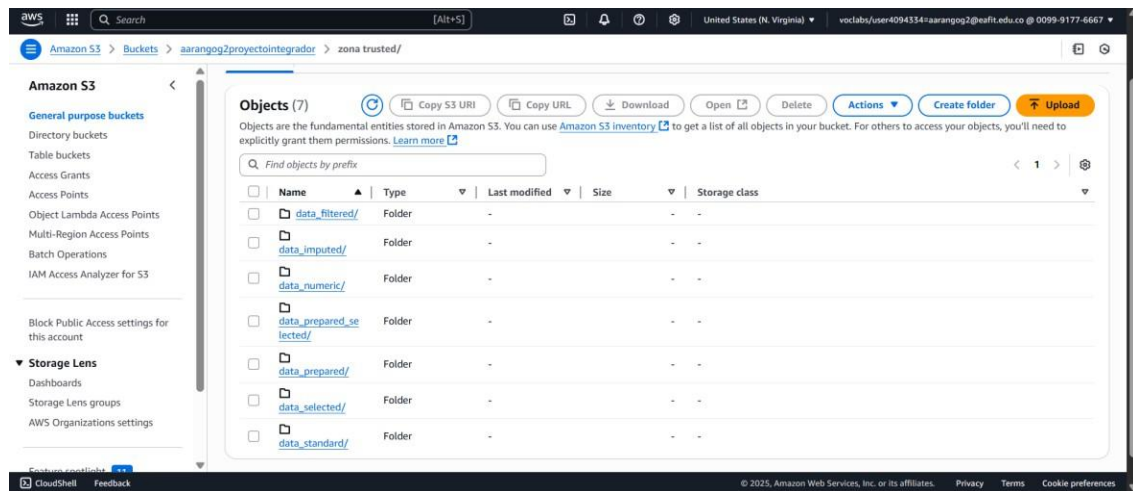
9. Ingresar con las credenciales:

- **Username:** jovyan
- **Password:** jupyter

Cargar y ejecutar el notebook: “Data prep.ipynb”.

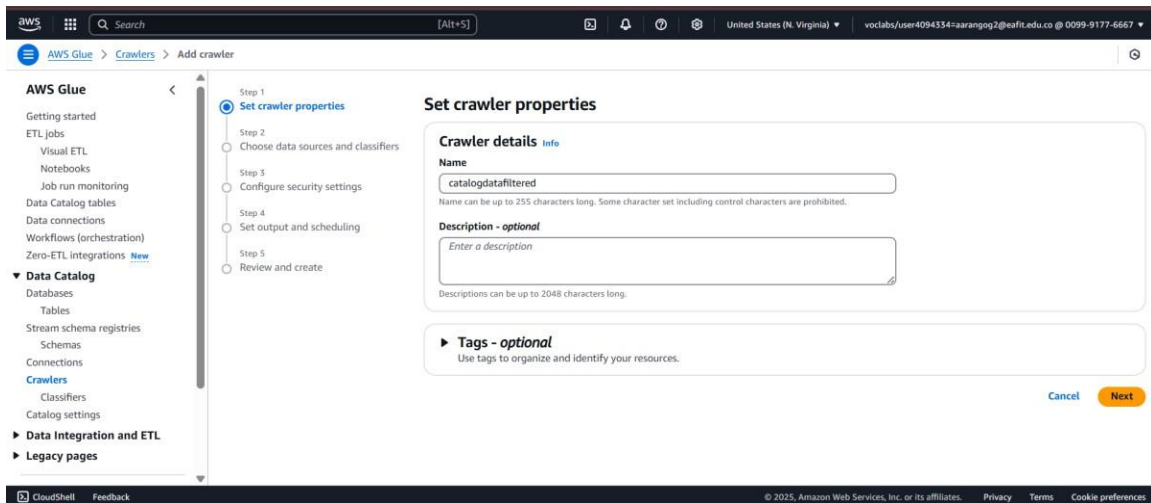


Tras ejecutar este notebook, se generarán los siguientes archivos con formato parquet en la zona trusted del S3:

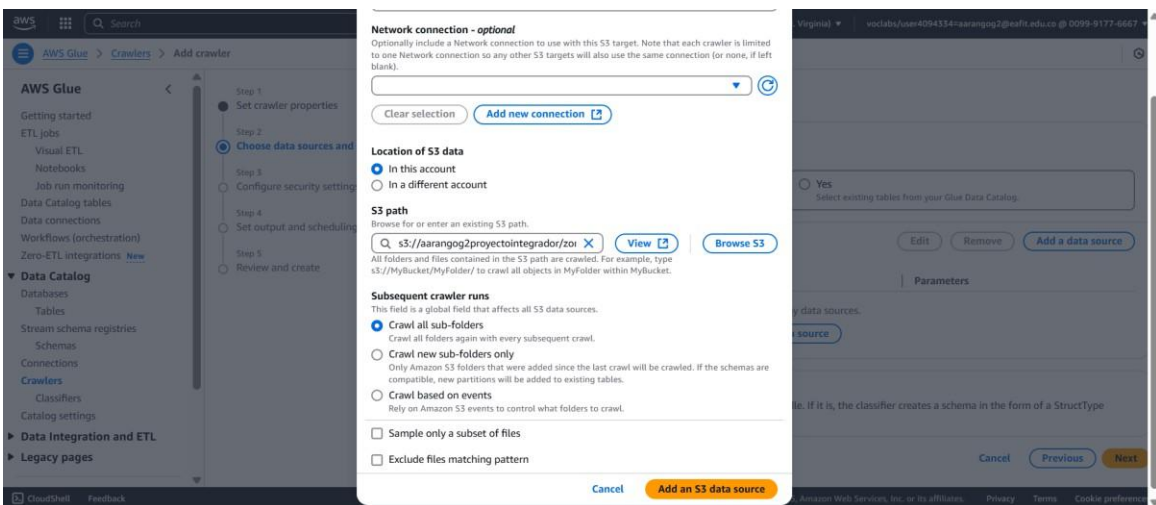
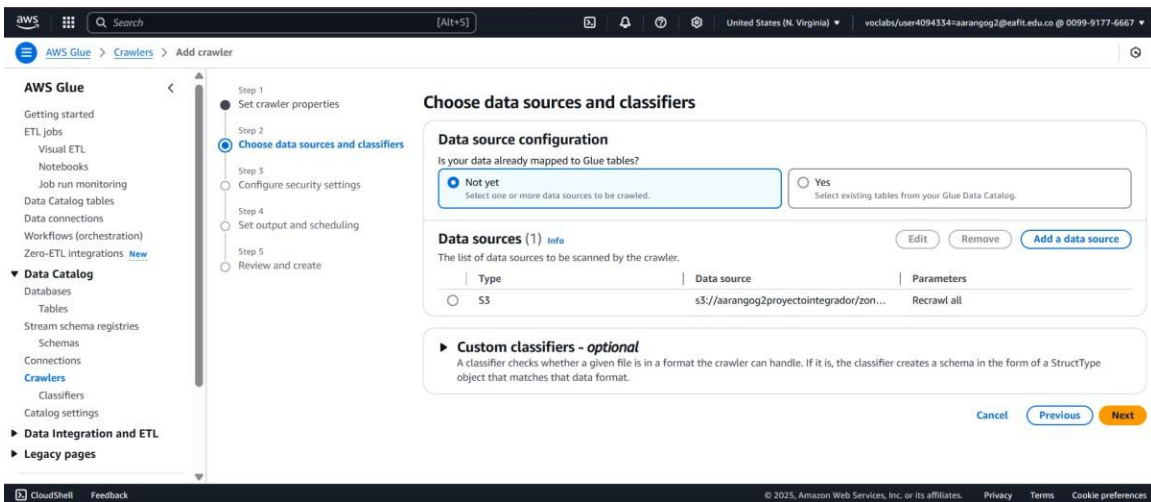


10. Catalogar los resultados del notebook “Data prep.ipynb” con el servicio de AWS Glue, creando un crawler por cada uno de los resultados. Para la creación de un crawler, se debe utilizar la siguiente configuración, y replicarla para cada uno de ellos:

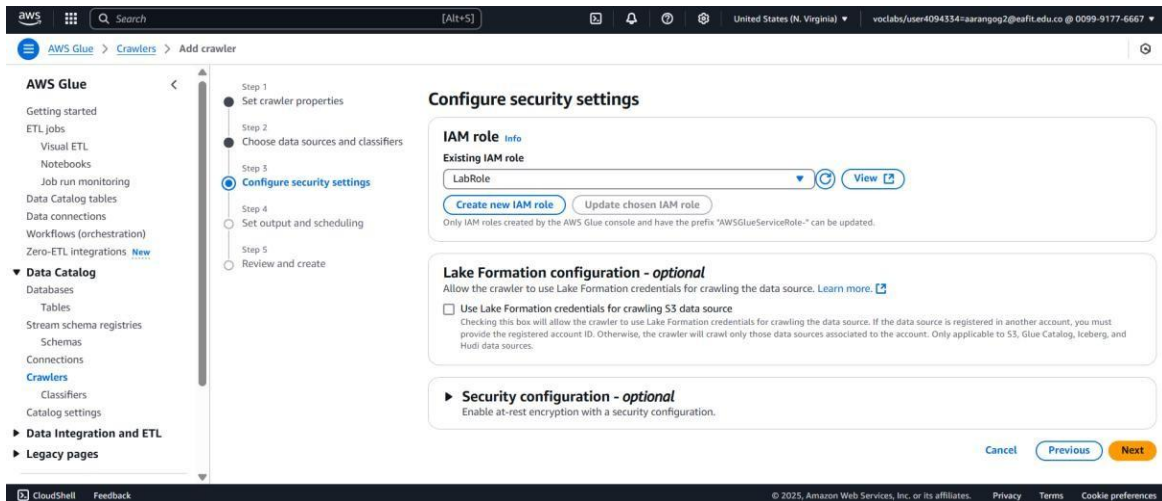
10.1 Seleccionar el nombre del crawler.



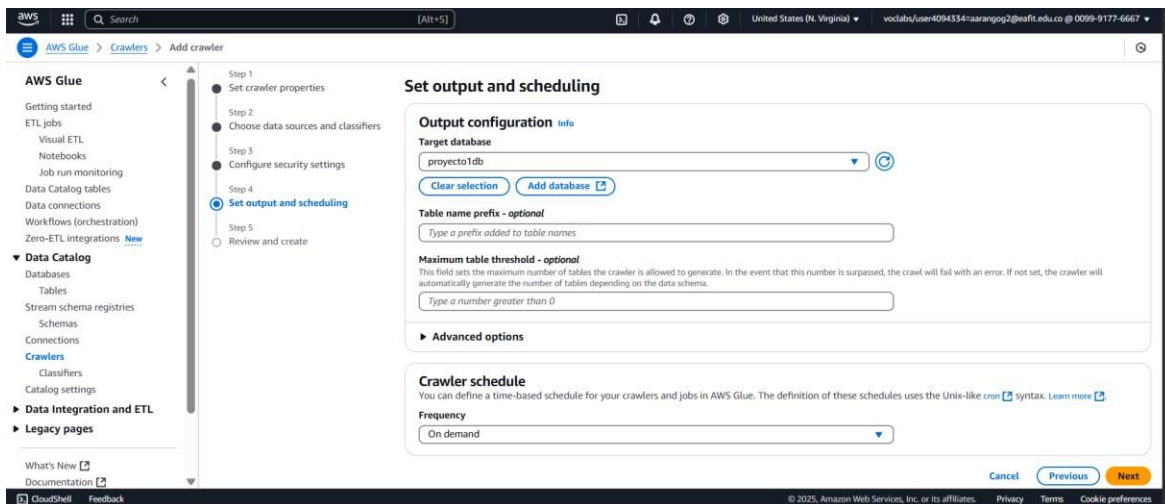
10.2 Seleccionar la ruta al archivo almacenado en S3 que se quiere catalogar.



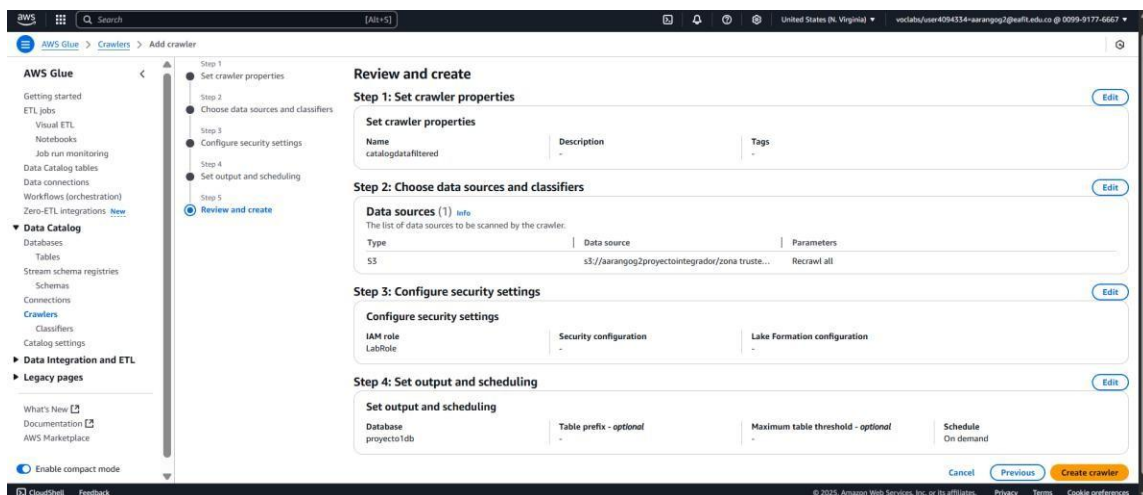
10.3 Seleccionar el rol "LabRole" para el "IAM role".



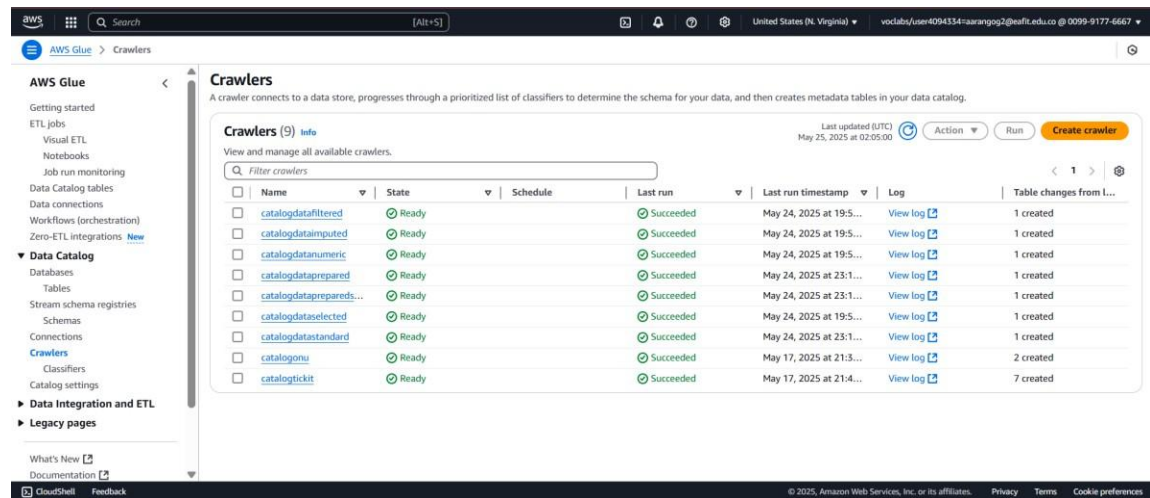
10.4 Crear una nueva base de datos llamada “proyecto1db” y seleccionarla para el almacenamiento de los esquemas.



10.5 Crear y correr el crawler.

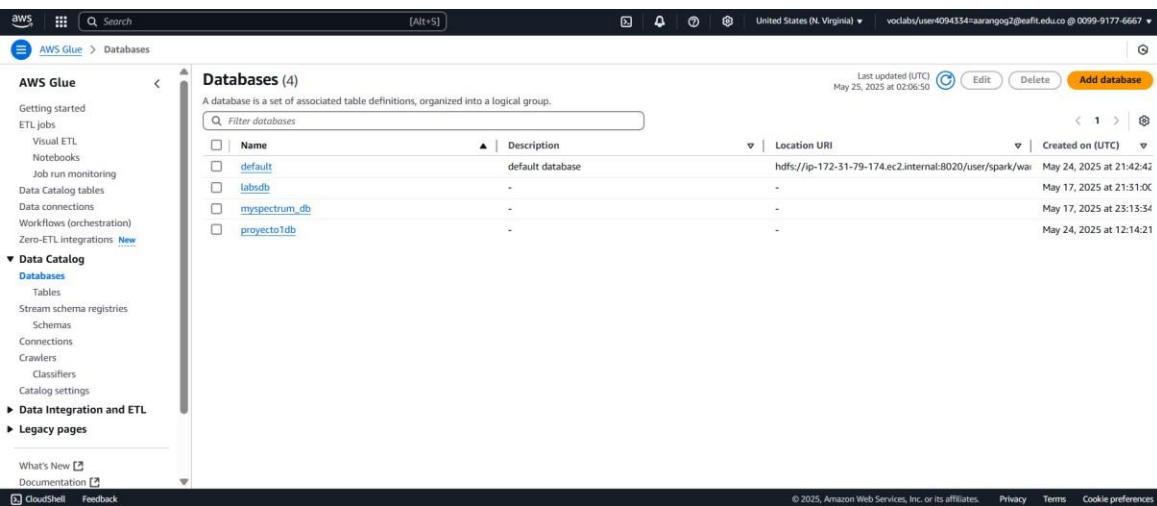


11. Correr todos los crawlers para obtener las siguientes tablas en la base de datos “proyecto1db”:



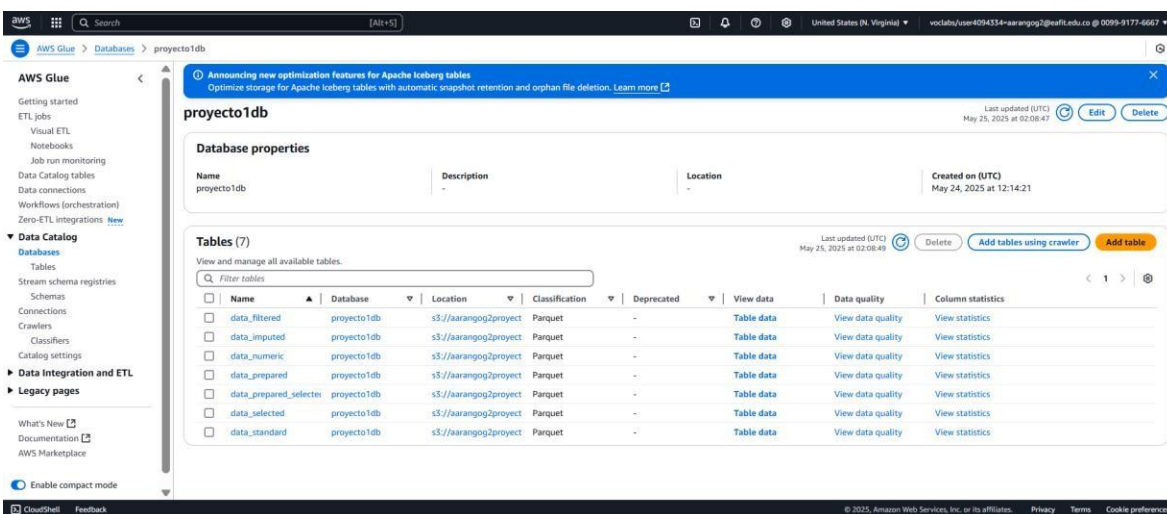
The screenshot shows the AWS Glue console with the 'Crawlers' page selected. The page title is 'Crawlers (9)' and it includes a description: 'A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.' The page has a 'Filter crawlers' search bar and a table listing 9 crawlers. All crawlers are in a 'Ready' state and have successfully completed their last run. The 'Table changes from last run' column shows the number of tables created by each crawler.

Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run
catalogdatafiltered	Ready		Succeeded	May 24, 2025 at 19:5...	View log	1 created
catalogdataimputed	Ready		Succeeded	May 24, 2025 at 19:5...	View log	1 created
catalogdatanumeric	Ready		Succeeded	May 24, 2025 at 19:5...	View log	1 created
catalogdataprepared	Ready		Succeeded	May 24, 2025 at 23:1...	View log	1 created
catalogdataprepareds...	Ready		Succeeded	May 24, 2025 at 23:1...	View log	1 created
catalogdataselected	Ready		Succeeded	May 24, 2025 at 19:5...	View log	1 created
catalogdatastandard	Ready		Succeeded	May 24, 2025 at 23:1...	View log	1 created
cataloggonu	Ready		Succeeded	May 17, 2025 at 21:3...	View log	2 created
catalogtickit	Ready		Succeeded	May 17, 2025 at 21:4...	View log	7 created



The screenshot shows the AWS Glue console with the 'Databases' page selected. The page title is 'Databases (4)' and it includes a description: 'A database is a set of associated table definitions, organized into a logical group.' The page has a 'Filter databases' search bar and a table listing 4 databases. The 'projecto1db' database is highlighted.

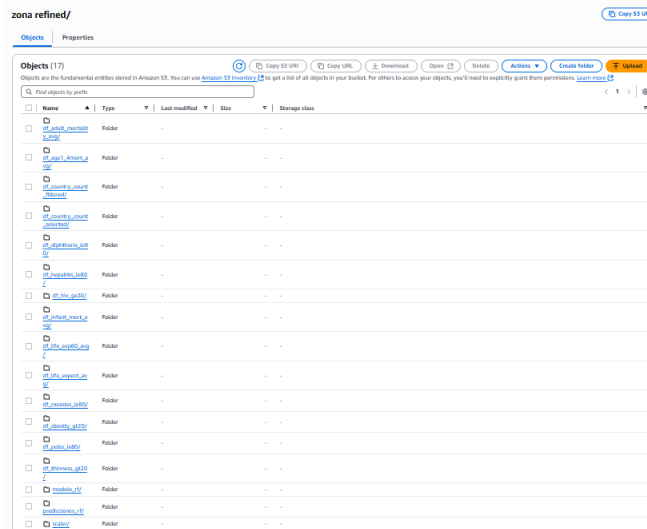
Name	Description	Location URI	Created on (UTC)
default	default database	hdfs://ip-172-31-79-174.ec2.internal:8020/user/spark/wai	May 24, 2025 at 21:42:42
labsdb	-	-	May 17, 2025 at 21:31:06
myspectrum_db	-	-	May 17, 2025 at 23:13:34
projecto1db	-	-	May 24, 2025 at 12:14:21



The screenshot shows the AWS Glue console with the 'projecto1db' database page selected. The page title is 'projecto1db' and it includes a description: 'Database properties'. The page has a 'Filter tables' search bar and a table listing 7 tables. The 'data_standard' table is highlighted.

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
data_filtered	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_imputed	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_numeric	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_prepared	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_prepared_selects	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_selected	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_standard	projecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics

12. Ejecutar los notebooks “EDA.ipynb”, y “Train Model.ipynb” en el EMR para obtener los siguientes resultados en la zona refined. Todos, a excepción del archivo “scaler” que es creado en “Data prep.ipynb”, son generados en el EDA y el entrenamiento del modelo.



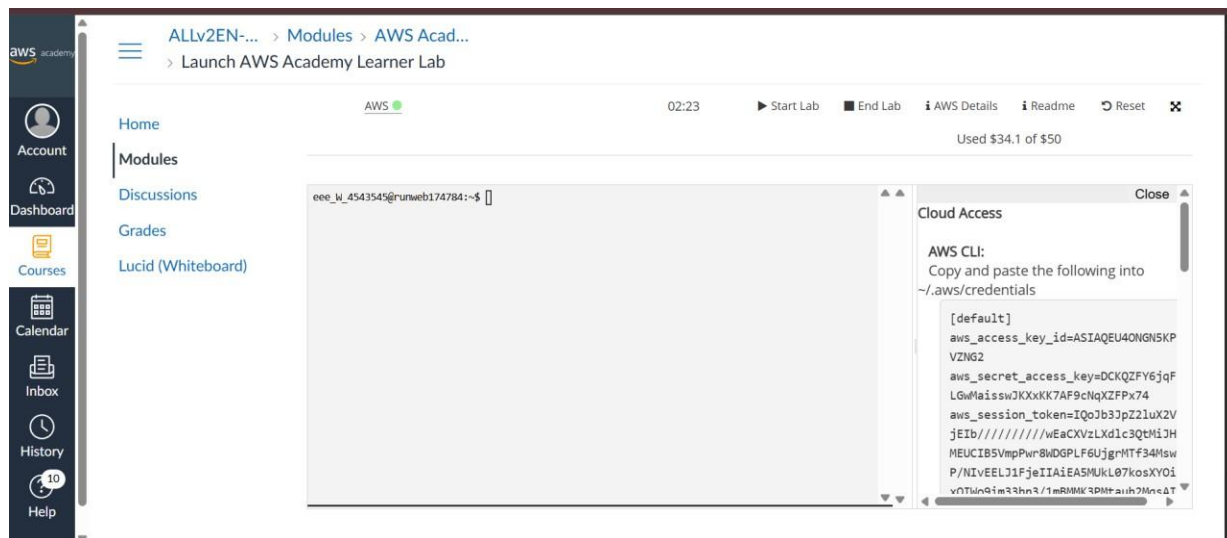
Name	Type	Last modified	Size	Storage class
01_eda	Folder	-	-	-
01_eda_model	Folder	-	-	-
01_eda_model_2	Folder	-	-	-
01_eda_model_3	Folder	-	-	-
01_eda_model_4	Folder	-	-	-
01_eda_model_5	Folder	-	-	-
01_eda_model_6	Folder	-	-	-
01_eda_model_7	Folder	-	-	-
01_eda_model_8	Folder	-	-	-
01_eda_model_9	Folder	-	-	-
01_eda_model_10	Folder	-	-	-
01_eda_model_11	Folder	-	-	-
01_eda_model_12	Folder	-	-	-
01_eda_model_13	Folder	-	-	-
01_eda_model_14	Folder	-	-	-
01_eda_model_15	Folder	-	-	-
01_eda_model_16	Folder	-	-	-
01_eda_model_17	Folder	-	-	-

13. Ejecutar en Google Colab el notebook “Visualizaciones.ipynb” para visualizar los resultados del EDA y el desempeño del modelo.

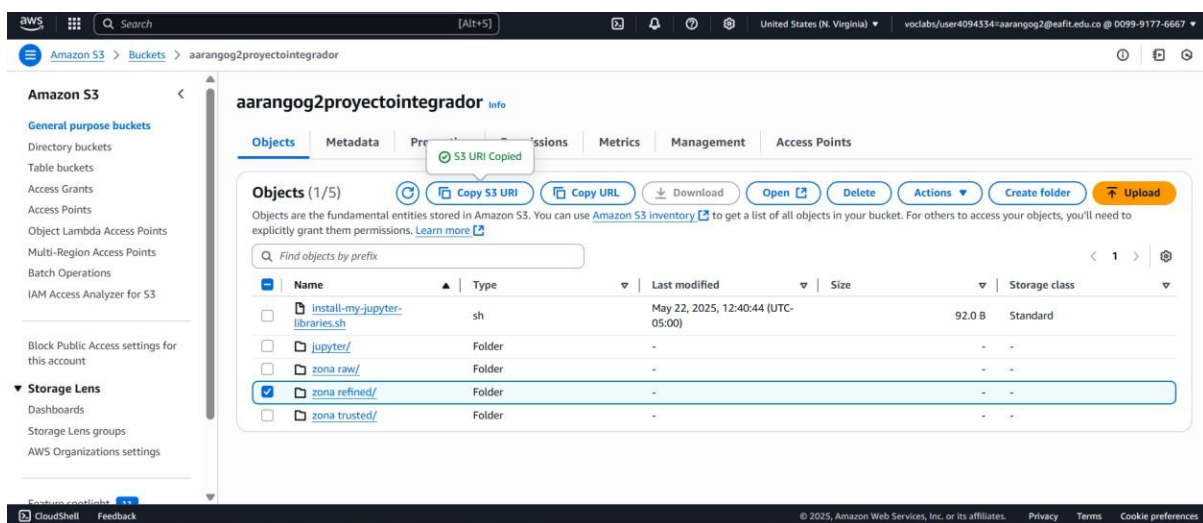
Para la ejecución exitosa del notebook, se deben modificar los siguientes parámetros:

- aws_access_key_id
- aws_secret_access_key
- aws_session_token

Se encuentran en la sección “AWS Details” de la terminal de la sesión de AWS creada en el “AWS Academy Learner Lab”.



El parámetro “s3_path” se puede obtener copiando la URI de la zona refined tras seleccionar la carpeta “zona refined/” y oprimiendo la opción “Copy S3 URI”.



Asumiendo los archivos se guardaron con los mismos nombres definidos en los notebooks “EDA.ipynb” y “Train Model.ipynb”, solo se debe correr el notebook para obtener las visualizaciones para “data_filtered”, “data_selected” y el desempeño del modelo.

