

PROYECTO 1 – PARTE 2: MANUAL DE REPRODUCIBILIDAD

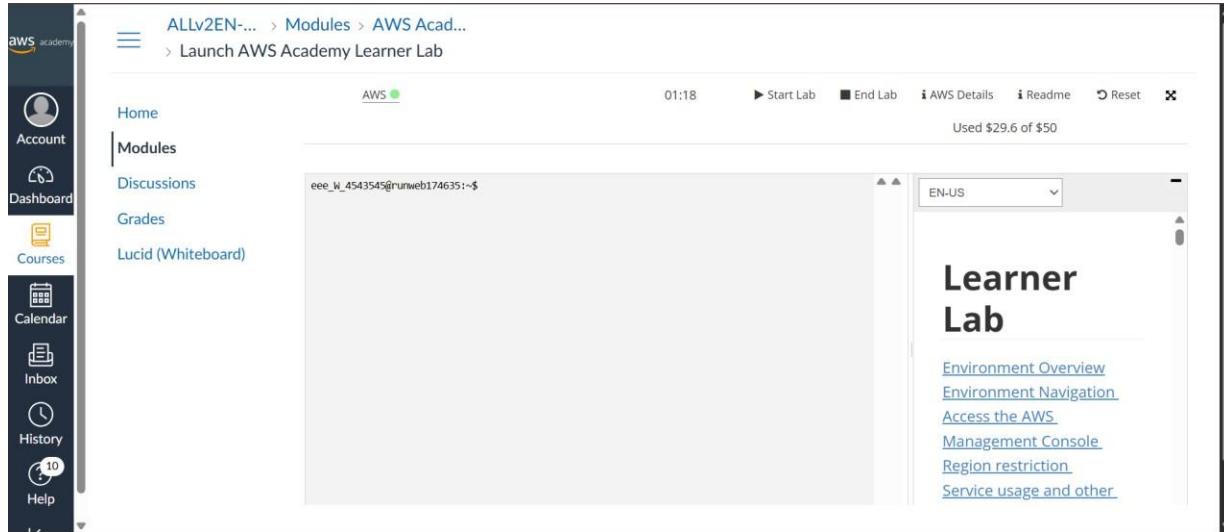
ALEJANDRO ARANGO GIRALDO

Línea de énfasis en Ciencias de los Datos
ST1800 – Almacenamiento y recuperación de la información

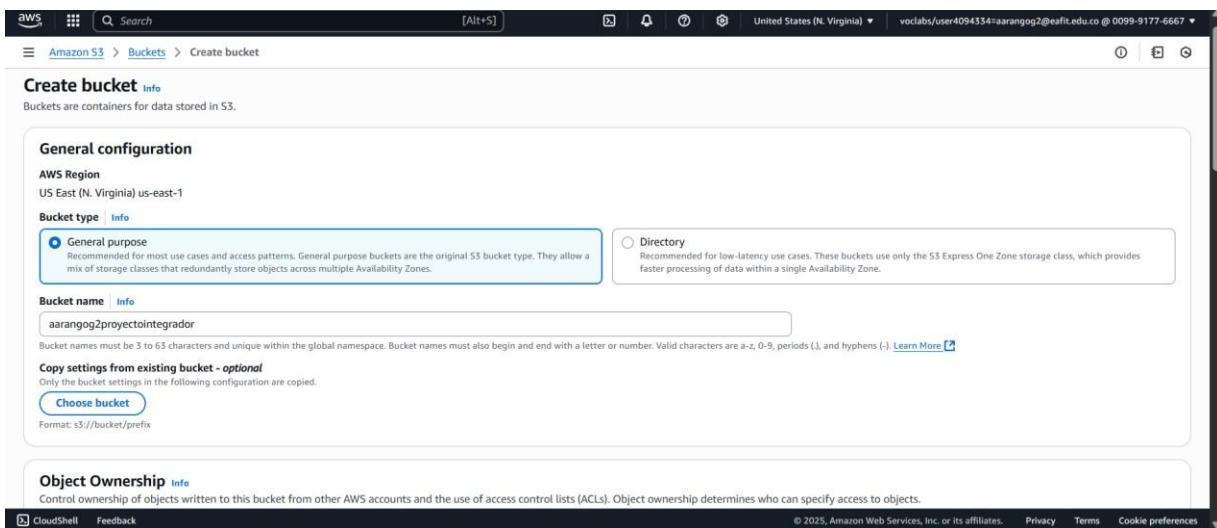
**UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
CARRERA DE INGENIERÍA MECÁNICA
MEDELLÍN
2025 - 1**

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.

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

Name	AWS Region	IAM Access Analyzer	Creation date
aarangog2lab1	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 17, 2025, 16:04:58 (UTC-05:00)
aarangog2proyectointegrador	US East (N. Virginia) us-east-1	View analyzer for us-east-1	May 20, 2025, 18:08:42 (UTC-05:00)
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)

Objects (5)

Name	Type	Last modified	Size	Storage class
install-my-jupyter-libraries.sh	sh	May 22, 2025, 12:40:44 (UTC-05:00)	92.0 B	Standard
jupyter/	Folder	-	-	-
zona raw/	Folder	-	-	-
zona refined/	Folder	-	-	-
zona trusted/	Folder	-	-	-

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

Objects (1/5)

Name	Type	Last modified	Size
install-my-jupyter-libraries.sh	sh	May 22, 2025, 12:40:44 (UTC-05:00)	92.0 B
jupyter/	Folder	-	-
zona raw/	Folder	-	-
zona refined/	Folder	-	-
zona trusted/	Folder	-	-

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

Objects (1)

Name	Type	Last modified	Size	Storage class
who_life_exp.csv	CSV	May 20, 2025, 18:10:10 (UTC-05:00)	683.7 KB	Standard

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

This screenshot shows the first step of the AWS EMR 'Create cluster' wizard, titled 'Clone "aarangog2 Proyecto Integrador"'. The 'Name and applications - required' section is expanded, showing a name input field containing 'aarangog2 Proyecto Integrador', an 'Amazon EMR release' dropdown set to 'emr-7.8.0', and an 'Application bundle' section with various services selected: AmazonCloudWatchAgent 1.3000.2.2, HCatalog 3.1.3, Hue 4.11.0, Livy 0.8.0, Pig 0.17.0, TensorFlow 2.16.1, Zepplin 0.11.1, Flink 1.20.0, Hadoop 3.4.1, Oozie 5.2.1, Presto 0.287, Tez 0.10.2, ZooKeeper 3.9.3, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Phoenix 5.2.1, Spark 3.5.4, and Trino 2.16.1. Below this, the 'AWS Glue Data Catalog settings' section is collapsed. The 'Summary' section on the right shows the cloned cluster details.

This screenshot shows the second step of the AWS EMR 'Create cluster' wizard, titled 'Cluster configuration - required'. It includes sections for 'Operating system options' (Amazon Linux release selected), 'Cluster configuration method' (Uniform instance groups selected), 'Uniform instance groups' (Primary group with m5.xlarge instance type selected), and 'Core' and 'Task' instance groups (both collapsed). The 'Summary' section on the right shows the cluster configuration.

This screenshot shows the third step of the AWS EMR 'Create cluster' wizard, titled 'Core'. It includes sections for 'Core' (instance type m5.xlarge selected), 'Task 1 of 1' (name 'Task - 1' and instance type m5.xlarge selected), 'Node configuration - optional' (collapsed), 'Add task instance group' (button), 'EBS root volume' (constraints: Size (GB) 15, IOPS 3000, Throughput (MiB/s) 125), and 'Cluster scaling and provisioning - required' (Provisioning configuration: Core size 1 instance, Task size 1 instance). The 'Summary' section on the right shows the cluster configuration.

The screenshot shows the 'Create cluster' step in the AWS EMR wizard. Under 'EBS root volume', the size is set to 15 GiB, IOPS to 3000, and Throughput to 125. A note states: '15 - 100 GiB per volume. General Purpose SSD (gp3)'. Below this, the 'Cluster scaling and provisioning - required' section is expanded, showing three options: 'Set cluster size manually' (selected), 'Use EMR-managed scaling', and 'Use custom automatic scaling'. The 'Provisioning configuration' section shows two instance groups: 'Task - 1' with 1 m5.xlarge instance and 'Core' with 1 m5.xlarge instance. The 'Networking - required' section is also expanded.

The screenshot continues the 'Create cluster' wizard. The 'Networking - required' section is expanded, showing 'Virtual private cloud (VPC)' configuration with a selected VPC ID and subnet. The 'EC2 security groups (firewall)' and 'Steps (0)' sections are also visible. The 'Cluster termination and node replacement' section is collapsed. The 'Bootstrap actions (1)' section is expanded, showing a single action named 'install-my-jupyter-libraries' that installs packages from an S3 location. The 'Cluster logs' section is partially visible at the bottom.

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, guardararlo con la extensión “sh”, y adicionarlo a la sección “Bootstrap actions” en la configuración del cluster.

A screenshot of a terminal window titled '\$ install-my-jupyter-libraries.sh'. The window shows the following code:

```

$ install-my-jupyter-libraries.sh
$ !#/bin/bash
$ 2
$ 3 sudo python3 -m pip install shap seaborn matplotlib scikit-learn pandas

```

The terminal is running in 'Restricted Mode' and displays a message about enabling features. The path shown is C: > Users > alejo > OneDrive - Universidad EAFIT > Escritorio > EAFIT > 2025-1 > Proyecto Integrador > Entregas > \$ install-my-jupyter-libraries.sh

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:

The screenshot shows the AWS Amazon EMR console. In the left sidebar, under 'EMR on EC2', the 'Block public access' option is selected. The main content area is titled 'Block public access' with a sub-section 'Block public access settings'. It shows that 'Block public access' is set to 'Off'. There is an 'Edit' button in the top right corner of this section.

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”.

The screenshot shows the AWS Amazon EMR Clusters summary page for a cluster named 'aarangog2 Proyecto Integrador'. Under the 'Summary' tab, the 'Status' section indicates 'Waiting'. The 'Primary node public DNS' field is listed as 'ec2-44-220-173-129.compute-1.amazonaws.com'. Below this, there are links to 'Connect to the Primary node using SSH' and 'Connect to the Primary node using SSM'.

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.

Instances (4) Info

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
Ubuntu_Aleja...	i-077bca098f84b0668	Stopped	t2.micro	-	View alarms +	us-east-1b	ec2-23-21-133-128.co...
	i-02305ba584385ac54	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1f	ec2-44-220-173-129.co...
	i-00fafdb575cca1f2e	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1f	ec2-18-205-60-107.co...
	i-0f30882f7bb0988d5	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1f	ec2-44-200-83-107.co...

Select an instance

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

Instance summary for i-02305ba584385ac54

Updated less than a minute ago

Details

Security

Inbound rules

Name	Security group rule ID	IP range	Protocol	Source	Description
sg-0a4674b86959e970c (ElasticMapReduce-master)	ng-0a4674b86959e970c	0.0.0.0/0	TCP	All	Allow ssh traffic

7.4 Entrar a la opción “Edit inbound rules”.

Security groups

sg-0a4674b86959e970c - ElasticMapReduce-master

Inbound rules (13)

Name	Security group rule ID	IP version	Type	Protocol	Port range	Source	Description
sg-08ab13b464c050c	ng-08ab13b464c050c	IPv4	Custom TCP	TCP	14000	0.0.0.0/0	Allow RDP traffic
sg-08237aa04332b4	ng-08237aa04332b4	-	All ICMP - IPv4	ICMP	All	0.0.0.0/0	Allow ICMP traffic
sg-0d6e14e295ef9e	ng-0d6e14e295ef9e	-	Custom TCP	TCP	9443	0.0.0.0/0	Allow port 9443
sg-0d6094a06944c	ng-0d6094a06944c	-	All TCP	TCP	0-65535	0.0.0.0/0	Allow all TCP traffic
sg-0f550b01a0694070	ng-0f550b01a0694070	-	Custom TCP	TCP	8880	0.0.0.0/0	Allow port 8880
sg-0d10a0a0579e640	ng-0d10a0a0579e640	-	All ICMP - IPv4	ICMP	All	0.0.0.0/0	Allow ICMP traffic
sg-0f81a0ef72a0d010	ng-0f81a0ef72a0d010	-	Custom TCP	TCP	8443	0.0.0.0/0	Allow port 8443
sg-0f7005079f0d8e3	ng-0f7005079f0d8e3	IPv4	Custom TCP	TCP	9870	0.0.0.0/0	Allow port 9870
sg-0faa6d4a0f0fb0	ng-0faa6d4a0f0fb0	-	All UDP	UDP	0-65535	0.0.0.0/0	Allow all UDP traffic
sg-04a02b2bfbc5c02	ng-04a02b2bfbc5c02	-	All UDP	UDP	0-65535	0.0.0.0/0	Allow all UDP traffic
sg-0315a629004983a	ng-0315a629004983a	-	All TCP	TCP	0-65535	0.0.0.0/0	Allow all TCP traffic
sg-04508d5ca7a89a55	ng-04508d5ca7a89a55	IPv4	SSH	TCP	22	0.0.0.0/0	Allow SSH traffic
sg-0644077bd420887	ng-0644077bd420887	IPv4	Custom TCP	TCP	8888	0.0.0.0/0	Allow port 8888

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”.

Name	Last Modified	File size
Data prep.ipynb	Running 21 minutes ago	
EDA.ipynb	4 hours ago	
Train Model.ipynb	an hour ago	

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

Name	Type	Last modified	Size	Storage class
data_filtered/	Folder	-	-	-
data_imputed/	Folder	-	-	-
data_numeric/	Folder	-	-	-
data_prepared_selected/	Folder	-	-	-
data_prepared/	Folder	-	-	-
data_selected/	Folder	-	-	-
data_standard/	Folder	-	-	-

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.

Crawler details

Name: catalogodatafiltered

Description - optional: Enter a description

Tags - optional: Use tags to organize and identify your resources.

Cancel **Next**

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

Data source configuration

Is your data already mapped to Glue tables?

Not yet Select one or more data sources to be crawled.

Yes Select existing tables from your Glue Data Catalog.

Data sources (1)

Type	Data source	Parameters
S3	s3://aarangog2/proyectointegrador/zon...	Recrawl all

Custom classifiers - optional

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Cancel **Previous** **Next**

Network connection - optional

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Location of S3 data

In this account

In a different account

S3 path

Browse for or enter an existing S3 path.

s3://aarangog2/proyectointegrador/zon

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

This field is a global field that affects all S3 data sources.

Crawl all sub-folders Crawl all folders again with every subsequent crawl.

Crawl new sub-folders only Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

Crawl based on events Rely on Amazon S3 events to control what folders to crawl.

Sample only a subset of files

Exclude files matching pattern

Add an S3 data source

Cancel **Previous** **Next**

10.3 Seleccionar el rol “LabRole” para el “IAM role”.

AWS Glue > Crawlers > Add crawler

Configure security settings

IAM role [Info](#)

- Existing IAM role: LabRole
- Create new IAM role
- Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - optional

Allow the crawler to use Lake Formation credentials for crawling the data source. [Learn more](#)

Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

Security configuration - optional

Enable at-rest encryption with a security configuration.

Cancel Previous Next

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

AWS Glue > Crawlers > Add crawler

Set output and scheduling

Output configuration [Info](#)

Target database: proyecto1db

Clear selection Add database

Table name prefix - optional

Type a prefix added to table names

Maximum table threshold - optional

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

Type a number greater than 0

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. [Learn more](#)

Frequency

On demand

Cancel Previous Next

10.5 Crear y correr el crawler.

AWS Glue > Crawlers > Add crawler

Review and create

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
catalogdatafilterred	-	-

Step 2: Choose data sources and classifiers

Data sources (1) [Info](#)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
S3	s3://aarangog2/proyectointegrador/zona truste...	Recrawl all

Step 3: Configure security settings

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
projecto1db	-	-	On demand

Create crawler

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

The screenshot shows three sequential views of the AWS Glue interface, illustrating the process of running crawlers to extract data into a database.

Step 1: Crawlers

The first view shows the "Crawlers" list. There are 9 crawlers listed, all in a "Ready" state. The last run was successful for all, with the most recent being on May 24, 2025, at 19:55 UTC. The crawler names include catalogdatafiltered, catalogdataimpuned, catalogdatanumeric, catalogdataprepared, catalogdataprepared, catalogdataselected, catalogdatastandard, catalogonu, and catalogtickit.

Name	State	Last run	Last run timestamp	Log	Table changes from last run
catalogdatafiltered	Ready	Succeeded	May 24, 2025 at 19:55...	View log	1 created
catalogdataimpuned	Ready	Succeeded	May 24, 2025 at 19:55...	View log	1 created
catalogdatanumeric	Ready	Succeeded	May 24, 2025 at 19:55...	View log	1 created
catalogdataprepared	Ready	Succeeded	May 24, 2025 at 23:10...	View log	1 created
catalogdataprepared	Ready	Succeeded	May 24, 2025 at 23:10...	View log	1 created
catalogdataselected	Ready	Succeeded	May 24, 2025 at 19:55...	View log	1 created
catalogdatastandard	Ready	Succeeded	May 24, 2025 at 23:10...	View log	1 created
catalogonu	Ready	Succeeded	May 17, 2025 at 21:30...	View log	2 created
catalogtickit	Ready	Succeeded	May 17, 2025 at 21:44...	View log	7 created

Step 2: Databases

The second view shows the "Databases" list. A single database named "proyecto1db" is listed. It was created on May 24, 2025, at 12:14:21 UTC. The location URI is hdfs://ip-172-31-79-174.ec2.internal:8020/user/spark/wa...

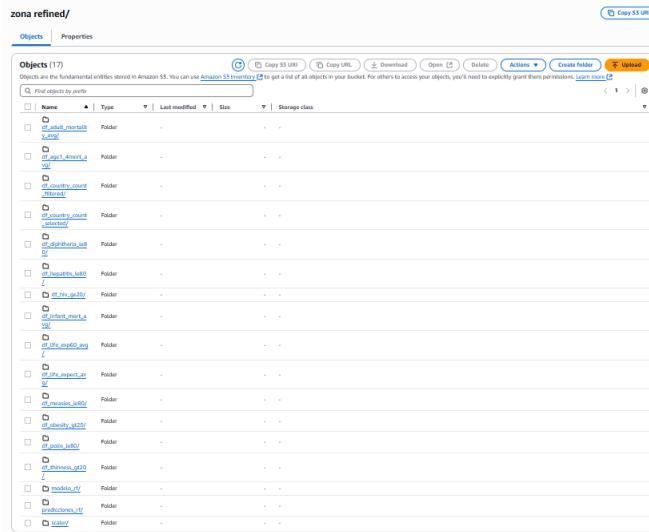
Name	Description	Location URI	Created on (UTC)
default	default database	hdfs://ip-172-31-79-174.ec2.internal:8020/user/spark/wa...	May 24, 2025 at 21:42:42
labsdb	-	-	May 17, 2025 at 21:31:00
myspectrum_db	-	-	May 17, 2025 at 23:13:34
proyecto1db	-	-	May 24, 2025 at 12:14:21

Step 3: Database Properties

The third view shows the "Database properties" for "proyecto1db". The database was created on May 24, 2025, at 12:14:21 UTC. The "Tables" section lists 7 tables: data_filtered, data_imputed, data_numeric, data_prepare, data_prepared_selecte, data_selected, and data_standard. All tables are located in the "proyecto1db" database and are Parquet files.

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
data_filtered	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_imputed	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_numeric	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_prepare	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_prepared_selecte	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_selected	proyecto1db	s3://aarangog2project	Parquet	-	Table data	View data quality	View statistics
data_standard	proyecto1db	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.



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”.

The screenshot shows the AWS Academy Learner Lab interface. On the left is a sidebar with navigation links: Account, Dashboard, Courses, Calendar, Inbox, History, and Help. The main area has a breadcrumb trail: ALLv2EN... > Modules > AWS Acad... > Launch AWS Academy Learner Lab. The top right shows the time (02:23), and buttons for Start Lab, End Lab, AWS Details, Readme, and Reset. A progress bar indicates 'Used \$34.1 of \$50'. On the right, a modal window titled 'Cloud Access' displays the AWS CLI configuration. It includes instructions to copy and paste the following into `~/.aws/credentials`. The configuration file content is as follows:

```
[default]
aws_access_key_id=ASIAQEU40NGN5KP
VZNG2
aws_secret_access_key=DCKQZFY6jqF
LGwMaisswJKXxKK7Af9cNgXZFPx74
aws_session_token=I0qj3JpZZluX2V
jEib//////////wEaCXvLXlcl3QtHiH
MEUCIB5VmpPwr8WDGPF6UjgrHTf34Hsw
P/NiVEELj1FjeIIaiEASMULkL07kosXYoi
vOTlw9inm3hnr3/1nBWW3PMauh7MncdT
```

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”.

The screenshot shows the Amazon S3 console. The left sidebar lists '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) and 'Storage Lens' (Dashboards, Storage Lens groups, AWS Organizations settings). The main area shows the 'aarangog2proyectointegrador' bucket. The 'Objects' tab is selected, showing 1/5 objects. A tooltip 'S3 URI Copied' appears over the 'Copy S3 URI' button for the 'zona refined/' folder. The table lists the objects:

Name	Type	Last modified	Size	Storage class
<input type="checkbox"/> install-my-jupyter-libraries.sh	sh	May 22, 2025, 12:40:44 (UTC-05:00)	92.0 B	Standard
<input type="checkbox"/> jupyter/	Folder	-	-	-
<input type="checkbox"/> zona raw/	Folder	-	-	-
<input checked="" type="checkbox"/> zona refined/	Folder	-	-	-
<input type="checkbox"/> zona trusted/	Folder	-	-	-

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.

