

# **PROYECTO 1 – PARTE 2: MANUAL DE REPRODUCIBILIDAD**

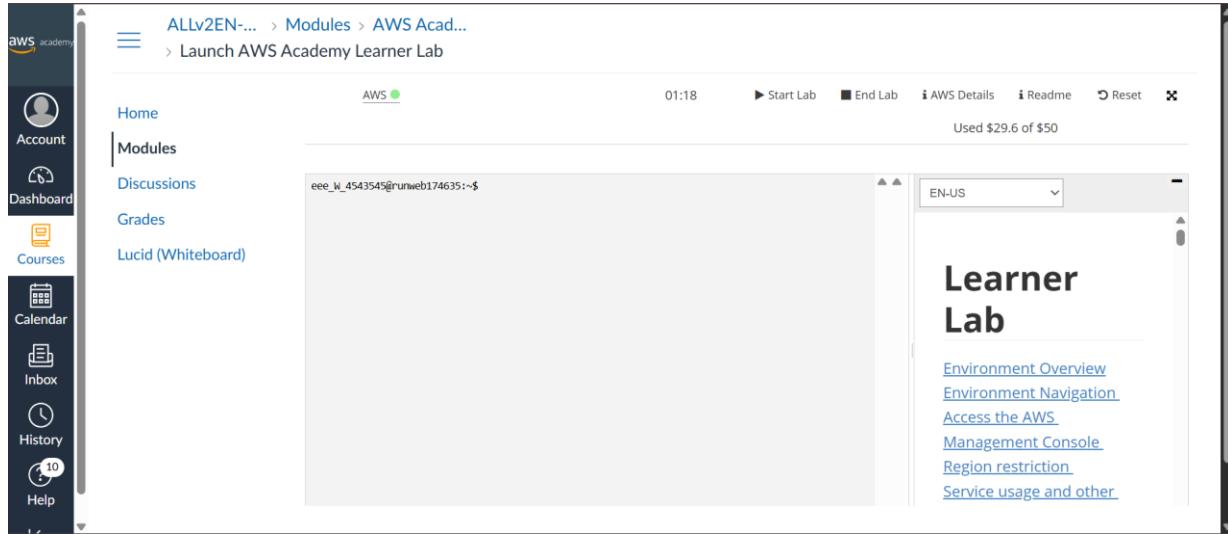
ALEJANDRO ARANGO GIRALDO

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

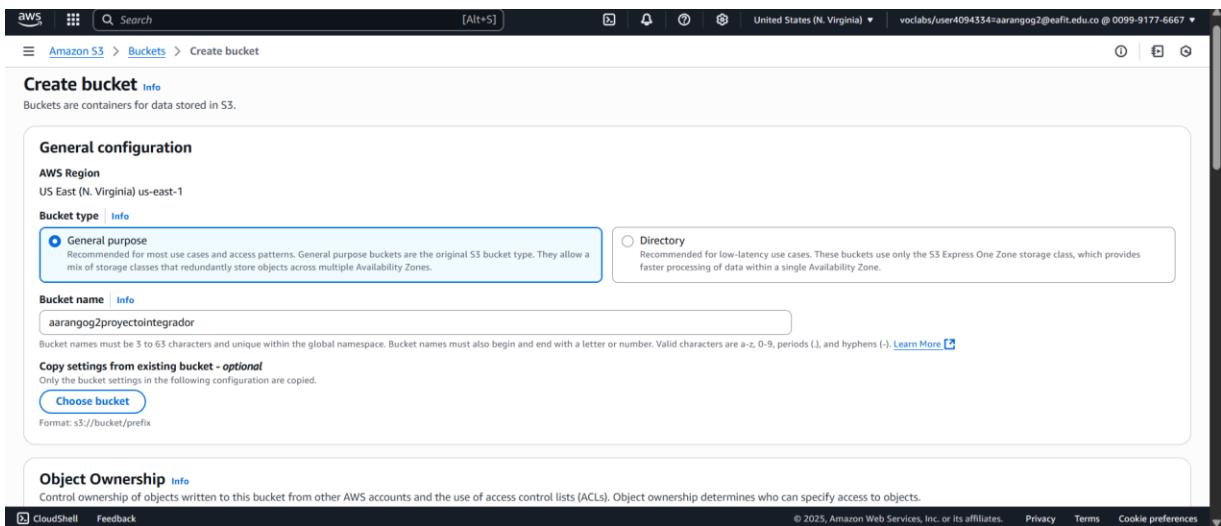
**UNIVERSIDAD EAFIT  
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA  
CARRERA DE INGENIERÍA MECÁNICA  
MEDELLÍN  
2025 - 1**

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.

Screenshot of the AWS S3 'Create bucket' page under 'Object Ownership' settings.

**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.

Screenshot of the AWS S3 'Create bucket' page under 'Bucket Versioning' settings.

**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](#)

**Disable**

**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 (DSS-E-KMS)  
Secure your objects with two separate layers of encryption. For details on pricing, see DSS-E-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 DSS-E-KMS. [Learn more](#)

**Disable**  
 **Enable**

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

Screenshot of the AWS S3 'General purpose buckets' list.

**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**

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

Buckets are containers for data stored in S3.

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

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

Storage Lens

CloudShell Feedback

Objects (5)

	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 type="checkbox"/>	zona refined/	Folder	-	-	-
<input type="checkbox"/>	zona trusted/	Folder	-	-	-

© 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”.

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

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

CloudShell Feedback

aarangog2proyectointegrador Info

Objects (1/5)

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

Actions ▾

- Share with a presigned URL
- Calculate total size
- Copy
- Move
- Initiate restore
- Query with S3 Select
- Edit actions
- Rename object
- Edit storage class
- Edit server-side encryption
- Edit metadata
- Edit tags
- Make public using ACL

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

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

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

Storage Lens

CloudShell Feedback

Objects (1)

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	who_life_exp.csv	CSV	May 20, 2025, 18:10:10 (UTC-05:00)	683.7 KB	Standard

Properties

Actions ▾

- Share with a presigned URL
- Calculate total size
- Copy
- Move
- Initiate restore
- Query with S3 Select
- Edit actions
- Rename object
- Edit storage class
- Edit server-side encryption
- Edit metadata
- Edit tags
- Make public using ACL

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

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

The screenshot shows three sequential steps for creating an Amazon EMR cluster, each showing a different configuration screen.

**Step 1: Clone "aarangog2 Proyecto Integrador"**

**Summary**  
**Name and applications - required**  
 Name: aarangog2 Proyecto 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 2.0)

**Cluster configuration - required**  
**Uniform instance groups**  
 Primary (m5.xlarge), Core (m5.xlarge), Task (m5.large)

**Cluster scaling and provisioning - required**  
**Provisioning configuration**  
 Core size: 1 instance  
 Task size: 1 instance

**AWS Glue Data Catalog settings**  
 Use the AWS Glue Data Catalog to provide an external metastore for your application.  
 Use for Hive table metadata  
 Use for Spark table metadata

**Clone cluster**

**Step 2: Create cluster**

**Operating system options**  
 Amazon Linux release  
 Custom Amazon Machine Image (AMI)  
 Automatically apply latest Amazon Linux updates

**Cluster configuration - required**  
 Uniform instance fleets  
 Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. [Learn more](#)

**Uniform instance groups**  
**Primary**  
 Choose EC2 instance type: m5.xlarge  
 4 vCore - 16 GB memory  
 EBS only storage - On-Demand price: Lowest Spot price:  
 Use high availability  
 Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. [Learn more](#)

**Core**  
 Choose EC2 instance type: m5.xlarge  
 4 vCore - 16 GB memory  
 EBS only storage - On-Demand price: Lowest Spot price:  
**Node configuration - optional**

**Node configuration - optional**

**Task 1 of 1**  
**Name**  
 Task - 1  
**Remove instance group**

**Choose EC2 instance type**  
 m5.xlarge  
 4 vCore - 16 GB memory  
 EBS only storage - On-Demand price: Lowest Spot price:  
**Node configuration - optional**

**Add task instance group**  
 You can add up to 47 more task instance groups.

**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 (GB) 15 | IOPS 3000 | Throughput (MiB/s) 125

**Summary**  
**Name and applications - required**  
 Name: aarangog2 Proyecto 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 2.0)

**Cluster configuration - required**  
**Uniform instance groups**  
 Primary (m5.xlarge), Core (m5.xlarge), Task (m5.large)

**Cluster scaling and provisioning - required**  
**Provisioning configuration**  
 Core size: 1 instance  
 Task size: 1 instance

**Clone cluster**

**Step 3: Create cluster**

**Core**  
**Choose EC2 instance type**  
 m5.xlarge  
 4 vCore - 16 GB memory  
 EBS only storage - On-Demand price: Lowest Spot price:  
**Node configuration - optional**

**Task 1 of 1**  
**Name**  
 Task - 1  
**Remove instance group**

**Choose EC2 instance type**  
 m5.xlarge  
 4 vCore - 16 GB memory  
 EBS only storage - On-Demand price: Lowest Spot price:  
**Node configuration - optional**

**Add task instance group**  
 You can add up to 47 more task instance groups.

**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 (GB) 15 | IOPS 3000 | Throughput (MiB/s) 125

**Summary**  
**Name and applications - required**  
 Name: aarangog2 Proyecto 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 2.0)

**Cluster configuration - required**  
**Uniform instance groups**  
 Primary (m5.xlarge), Core (m5.xlarge), Task (m5.large)

**Cluster scaling and provisioning - required**  
**Provisioning configuration**  
 Core size: 1 instance  
 Task size: 1 instance

**Clone cluster**

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.

```
$ install-my-jupyter-libraries.sh
$ install-my-jupyter-libraries.sh
$ sudo python3 -m pip install numpy scipy matplotlib seaborn scikit-learn pandas
```

**Cluster logs** info  
Choose where and how to store your log files.

**Tags** info  
Use tags to search and filter for resources, and track AWS costs associated with your cluster.

**Software settings** info  
Override the default configurations for specific applications on your cluster.

**Enter configuration**  Load JSON from Amazon S3

```

1  [
2    {
3      "Classification": "jupyter-s3-conf",
4      "Properties": {
5        "s3.persistence.bucket": "aarangog2/proyectointegrador",
6        "s3.persistence.enabled": "true"
7      }
8    }
9  ]

```

**Summary** info  
**Name and applications - required**

**Name**: aarangog2 Proyecto 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

**Clone cluster**

**Security configuration and EC2 key pair** info  
Choose a security configuration or create a new one that you can reuse with other clusters.

**Security configuration**: Select your cluster encryption, authentication, administration, and instance metadata service settings.  
 Choose a security configuration

**Amazon EC2 key pair for SSH to the cluster** info

**Identity and Access Management (IAM) roles - required** info  
Choose or create a service role and instance profile for the EC2 instances in your cluster.

**Amazon EMR service role** info  
The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.  
 Choose an existing service role Select a default service role or a custom role with IAM permissions that your Cluster can interact with other AWS services.  
 Create a service role Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

**Service role**: EMR\_DefaultRole

**EC2 Instance profile for Amazon EMR** info  
The instance profile manages EC2 instance roles in a cluster. The instance profile must specify a role that can access resources for your steps and bootstrap actions.  
 Choose an existing instance profile Select a default role or a custom instance profile with IAM permissions that your Cluster can interact with your resources in Amazon S3.  
 Create an instance profile Let Amazon EMR create a new instance profile so that you can specify a different set of resources for it to access in Amazon S3.

**Instance profile**: EMR\_EC2\_DefaultRole

**Custom automatic scaling role - optional** info  
Custom automatic scaling role When a custom automatic scaling role is specified, Amazon EMR assumes this role to add and terminate EC2 instances. Learn more  
 EMR\_AutoScaling\_DefaultRole

**Summary** info  
**Name and applications - required**

**Name**: aarangog2 Proyecto 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

**Clone cluster**

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 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 'Info' link. It states: 'Amazon EMR block public access prevents a cluster from launching when it is associated with security group rules that allow inbound traffic from IPv4 0.0.0.0/0 or IPv6 ::/0 (public access) on a port, unless the port is explicitly specified as an exception.' Below this is a 'Block public access settings' section with a 'Block public access' dropdown set to 'Off'. There is also an 'Edit' button.

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 EMR 'Clusters' summary page for a cluster named 'aarangog2 Proyecto Integrador'. The 'Summary' tab is selected. Under 'Cluster info', the 'Cluster ID' is listed as 'j-2OHFUJFVXBNC'. Under 'Cluster ARN', it shows 'arn:aws:elasticmapreduce:us-east-1:009991776667:cluster-j-2OHFUJFVXBNC'. Under 'Cluster configuration', there is one 'Instance groups'. Under 'Capacity', there is 1 Primary, 1 Core, and 1 Task. The 'Applications' section lists 'Amazon EMR version emr-7.8.0' and 'Installed applications' including HCatalog 3.1.3, Hadoop 3.4.1, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Livy 0.8.0, Pig 0.17.0, Spark 3.5.4, Tez 0.10.2, Zeppelin 0.11.1, ZooKeeper 3.9.3. The 'Cluster management' section includes links for 'Log destination in Amazon S3', 'Persistent application UIs', 'Spark History Server', 'YARN timeline server', and 'Tez UI'. The 'Status and time' section shows the 'Status' as 'Waiting', 'Creation time' as 'May 24, 2025, 20:07 (UTC-05:00)', and 'Elapsed time' as '41 minutes, 13 seconds'. The 'Primary node public DNS' is listed as 'ec2-44-220-173-129.compute-1.amazonaws.com'. Navigation tabs at the bottom include Properties, Bootstrap actions, Instances (Hardware), Steps, Applications, Configurations, Monitoring, Events, and Tags (0).

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.

The screenshot shows the AWS EC2 Instances page. On the left, there's a navigation sidebar with options like Dashboard, EC2 Global View, Events, Instances, Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Capacity Reservations, Images, AMIs, AMI Catalog, Elastic Block Store, Volumes, Snapshots, and CloudShell. The main area is titled "Instances (4) Info" and lists four instances:

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
Ubuntu_Alejandra	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-00fafdb575cca1fe	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1f	ec2-18-205-60-107.co...
	i-0f30882f7bb098bd5	Running	m5.xlarge	3/3 checks passed	View alarms +	us-east-1f	ec2-44-200-83-107.co...

At the bottom, there are buttons for Connect, Instance state, Actions, and Launch instances.

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

The screenshot shows the Instance summary for instance i-02305ba584385ac54. The left sidebar includes options for Dashboard, EC2 Global View, Events, Instances, Instance Types, Launch Templates, Spot Requests, Savings Plans, Reserved Instances, Dedicated Hosts, Capacity Reservations, Images, AMIs, AMI Catalog, Elastic Block Store, Volumes, Snapshots, Lifecycle Manager, Network & Security, Security Groups, EBS IOPS, Placement Groups, Key Pairs, Network Interfaces, Load Balancing, Auto Scaling, and CloudShell. The main area displays the instance summary with tabs for Details, Status and alarms, Monitoring, Security (selected), Networking, Storage, and Tags. Under the Security tab, it shows the IAM Role (EMR\_EC2\_DefaultRole), Security groups (sg-0a4674b86959e970c (ElasticMapReduce-master)), and Inbound rules.

## 7.4 Entrar a la opción “Edit inbound rules”.

The screenshot shows the AWS EC2 Security Groups page for the sg-0a4674b86959e970c group. The left sidebar is identical to the previous screenshot. The main area shows the security group details (Security group name: ElasticMapReduce-master, Security group ID: sg-0a4674b86959e970c, Owner: 009991776667, Description: New security group for Elastic MapReduce created on 2023-05-18T09:20:02Z, VPC ID: vpc-05f479e942853556) and the Inbound rules (13). The Inbound rules table has columns for Name, Security group rule ID, IP version, Type, Protocol, Port range, Source, and Description. The rules include various entries for TCP, UDP, and Custom TCP protocols on ports 14000, 9445, 8888, 8889, 9870, 9871, 0-65535, and 22.

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

The screenshot shows the 'Inbound rules' section of an AWS Security Group. There are 14 rules listed:

- sg-08d811ea64f4fc0c0c: Custom TCP, Port range 14000, Source 0.0.0.0/0
- sg-08327ea043f2b04: All ICMP - IPv4, Port range All, Source 0.0.0.0/0
- sg-0c9e614b293ae0e: Custom TCP, Port range 9443, Source 0.0.0.0/0
- sg-0e6946a60d69e44cc: All TCP, Port range 0-65535, Source 0.0.0.0/0
- sg-0f556025e09eb970: Custom TCP, Port range 8990, Source 0.0.0.0/0
- sg-0110a0ab0279b60: All ICMP - IPv4, Port range All, Source 0.0.0.0/0
- sg-083a11712fb6d59: Custom TCP, Port range 8443, Source 0.0.0.0/0
- sg-0f7908bf7cd3b53: Custom TCP, Port range 9870, Source 0.0.0.0/0
- sg-0baa683aa97ebcb: All UDP, Port range 0-65535, Source 0.0.0.0/0
- sg-04a822bd83c202: All UDP, Port range 0-65535, Source 0.0.0.0/0
- sg-0755a6e7e0544953a: All TCP, Port range 0-65535, Source 0.0.0.0/0
- sg-0405effbd571dbbd5: SSH, Port range 22, Source 0.0.0.0/0
- sg-0654d77bd9429887: Custom TCP, Port range 8888, Source 0.0.0.0/0

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

Cluster ID	Cluster name	Status	Creation time (UTC-05:00)	Elapsed time
j-20HFUJFVXBNC	ararangog2 Proyecto Integrador	Waiting	May 24, 2025, 20:07	45 minutes, 45 seconds
j-199510C06VKXQ	ararangog2 Proyecto Integrador	Terminated	May 24, 2025, 18:46	46 minutes, 53 seconds
j-1POY98V9RNXGY	ararangog2 Proyecto Integrador	Terminated with errors	May 24, 2025, 18:36	5 minutes, 56 seconds
j-3HOJV52N0HOB	ararangog2 Proyecto Integrador	Terminated with errors	May 24, 2025, 16:07	2 hours, 25 minutes
j-2K46QHL799PV8	ararangog2 Proyecto Integrador	Terminated	May 24, 2025, 14:34	1 hour, 17 minutes
j-1KBRMCUTAB5KG	ararangog2 Proyecto Integrador	Terminated	May 24, 2025, 08:07	2 hours, 26 minutes
j-2IYY1UU77IU	ararangog2 Proyecto Integrador	Terminated	May 22, 2025, 19:12	1 hour, 9 minutes
j-18KP9EPP6DK79	ararangog2 Proyecto Integrador	Terminated	May 22, 2025, 12:59	39 minutes, 51 seconds

**Application user interfaces**

On-cluster application UIs: Available only while your cluster is running. Use the following links to get started. To access all the application UIs, set up SSH tunneling.

Persistent application UIs: Persistent UIs don't require SSH tunneling. They are hosted off the cluster and are available for 30 days after an application ends.

**Live Application UIs**

These on-cluster application UIs are available without SSH tunneling.

**Application UIs**

**Application UIs on the primary node**

These require SSH tunneling to be enabled.

Application	UI URL
HDFS Name Node	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:9870/">http://ec2-44-220-175-129.compute-1.amazonaws.com:9870/</a>
Hue	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:8888/">http://ec2-44-220-175-129.compute-1.amazonaws.com:8888/</a>
JupyterHub	<a href="https://ec2-44-220-175-129.compute-1.amazonaws.com:9443/">https://ec2-44-220-175-129.compute-1.amazonaws.com:9443/</a>
Livy	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:8998/">http://ec2-44-220-175-129.compute-1.amazonaws.com:8998/</a>
Resource Manager	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:8088/">http://ec2-44-220-175-129.compute-1.amazonaws.com:8088/</a>
Spark History Server	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:18080/">http://ec2-44-220-175-129.compute-1.amazonaws.com:18080/</a>
Tez UI	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:9080/tez-ui/">http://ec2-44-220-175-129.compute-1.amazonaws.com:9080/tez-ui/</a>
Zeppelin	<a href="http://ec2-44-220-175-129.compute-1.amazonaws.com:8890/">http://ec2-44-220-175-129.compute-1.amazonaws.com:8890/</a>

**9.** Ingresar con las credenciales:

- **Username:** joyyan
- **Password:** jupyter

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

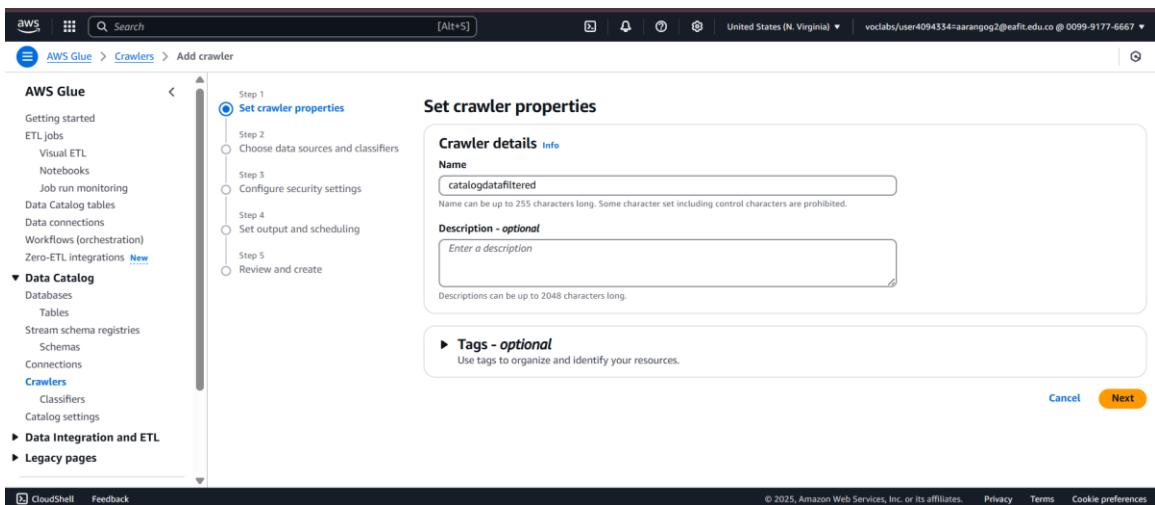
Name	Last Modified	File size
Data prep.ipynb	Running 21 minutes ago	
EDA.ipynb	4 hours ago	
Train Model.ipynb	an hour ago	

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.



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

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

**Configure security settings**

**IAM role** [Info](#)  
 Existing IAM role: LabRole  
  [View](#)

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

**Lake Formation configuration - optional**  
 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 Hull 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.

**Set output and scheduling**

**Output configuration** [Info](#)  
 Target database: proyecto1db  
  [View](#)

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.

**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.

**Review and create**

**Step 1: Set crawler properties**  
 Set crawler properties  
 Name: catalogdatafiltered

**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/proyecto1integrador/zona trusted...	Recrawl all

**Step 3: Configure security settings**  
 Configure security settings  
 IAM role: LabRole

**Step 4: Set output and scheduling**  
 Set output and scheduling  
 Database: proyecto1db  
 Table prefix - optional:   
 Maximum table threshold - optional:   
 Schedule: On demand

[Cancel](#) [Previous](#) [Create crawler](#)

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

The screenshot displays three sequential views of the AWS Glue interface, illustrating the process of running crawlers to create tables in the 'proyecto1db' database.

**Screenshot 1: Crawlers**

This screen shows the list of available crawlers. There are 9 crawlers listed, all in a 'Ready' state and have succeeded in their latest run. The table includes columns for Name, State, Last run, Last run timestamp, Log, and Table changes from last run.

Name	State	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
catalogdataprepared	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
catalogonu	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

**Screenshot 2: Databases**

This screen shows the list of databases. There are four databases listed: 'default', 'labsdb', 'myspectrum\_db', and 'proyecto1db'. The 'proyecto1db' database is selected. The table includes columns for Name, Description, Location URI, and Created on (UTC).

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 21:31:00
myspectrum_db	-	-	May 17, 2025 at 23:13:34
proyecto1db	-	-	May 24, 2025 at 12:14:21

**Screenshot 3: Database Properties for proyecto1db**

This screen shows the properties of the 'proyecto1db' database. It lists the database name, description, location, and creation date. Below this, it shows the list of tables in the database.

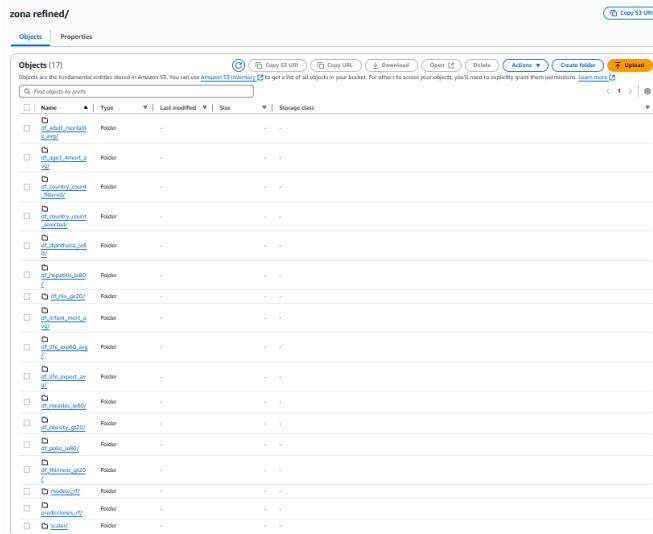
Name	Description	Location	Created on (UTC)
proyecto1db	-	-	May 24, 2025 at 12:14:21

**Tables (7)**

This table lists the seven tables created by the crawlers. The table includes columns for Name, Database, Location, Classification, Deprecated, View data, Data quality, and Column statistics.

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_selected	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.

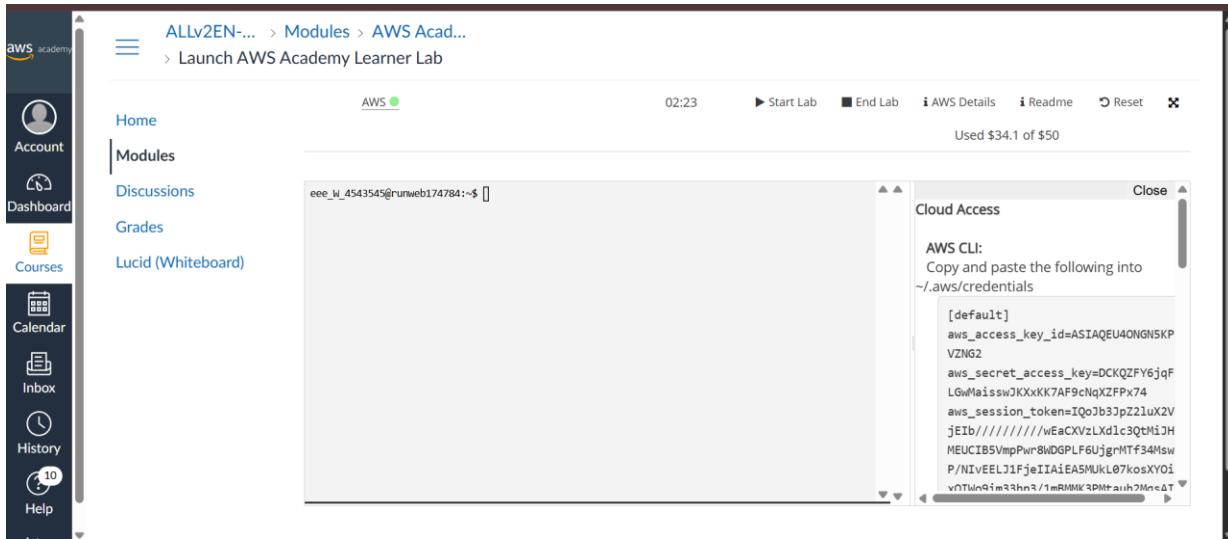
Para la ejecución exitosa del notebook, se deben modificar los siguientes parámetros:

```
# Credenciales
aws_access_key_id = "ASIAQUEU4ONGN5KPVZNG2"
aws_secret_access_key = "DCKQZFY6jqFLgwMaisswwJKXxKK7AF9cNqXZFPx74"
aws_session_token = "IQoJb3JpZ2luX2VjEib//////////wEAxVzLXd1c3QtMijHMEUCIB5VmpPwr8WDGPLE6UjgrMTf34MsP/NiVUELJ1FjeIIAiEA5MuKl07kosXY0ix0IWo9im33hn3/1mBMMK3PMtauh2MqsAI
s3_path = "s3://aarangog2/proyectointegrador/zona refined/"
```

Los 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”.

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	-	-	-
<b>zona refined/</b>	Folder	-	-	-
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.

