PROYECTO 1 – PARTE 2: MANUAL DE REPRODUCIBILIDAD

ALEJANDRO ARANGO GIRALDO

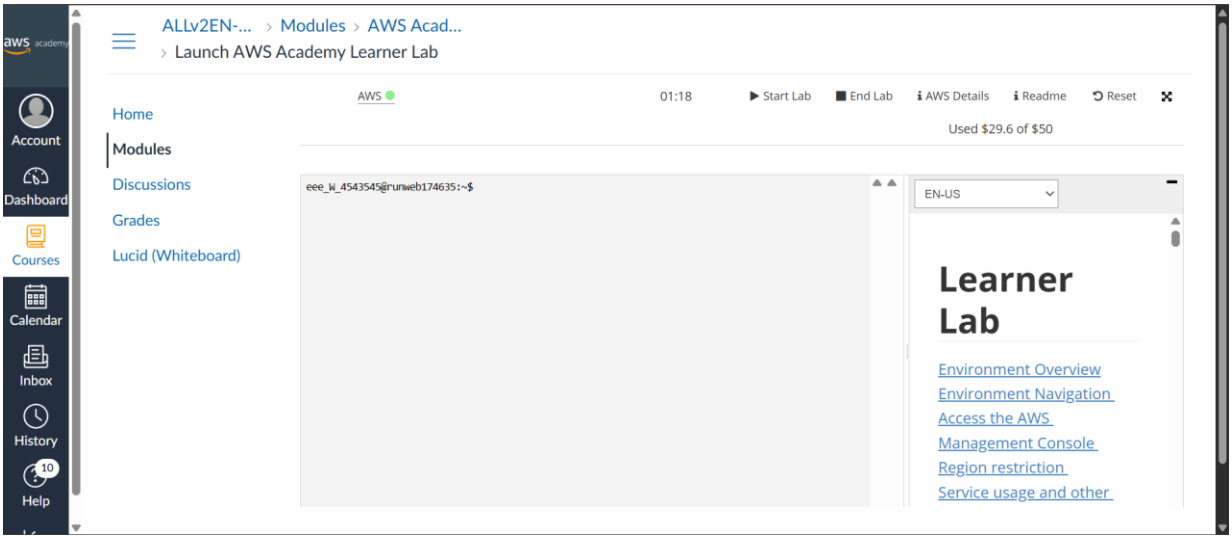
Línea de énfasis en Ciencias de los Datos

ST1800 – Almacenamiento y recuperación de la información

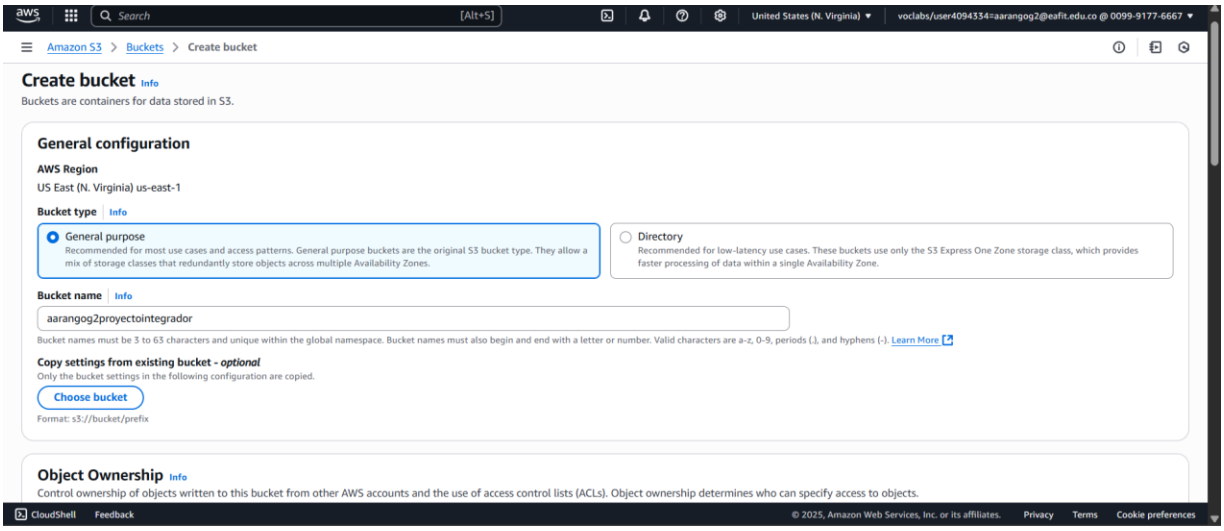
UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
CARRERA DE INGENIERÍA MECÁNICA
MEDELLÍN
2025 - 1

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

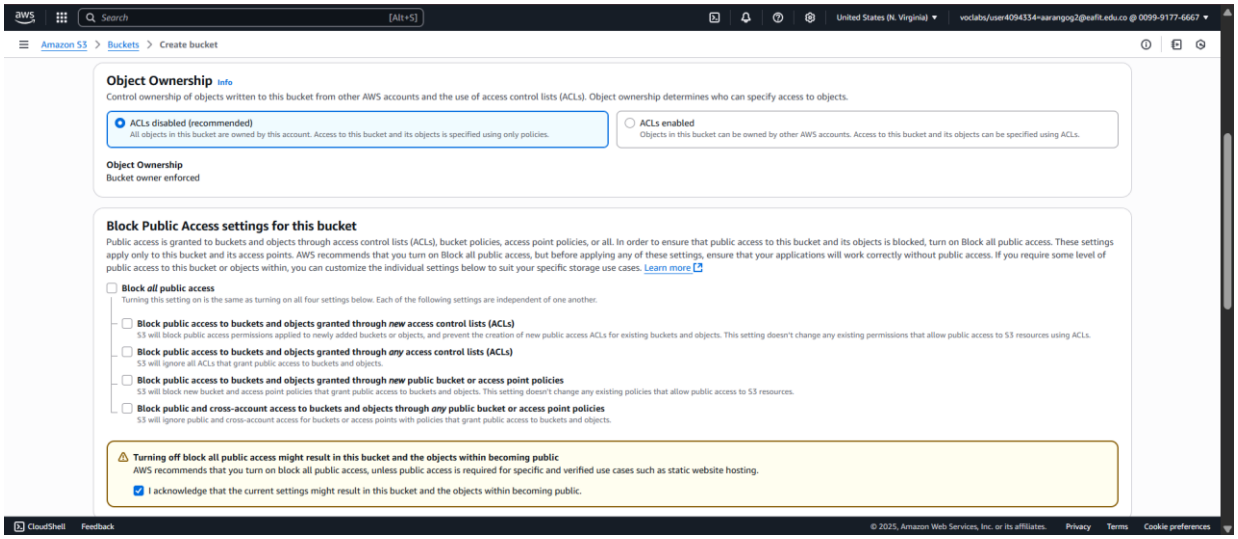
1. Activar una sesión de AWS Academy Learner Lab, o, si se cuenta con otras credenciales, activar la sesión.

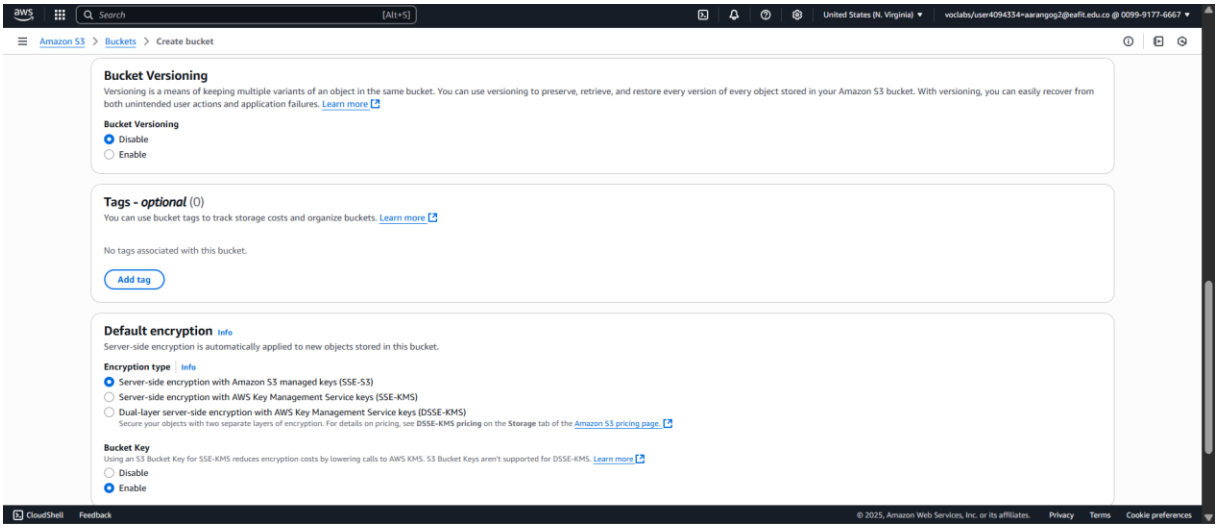


2. Entrar a la interfaz web de AWS y buscar el servicio de AWS S3. Oprimir en la opción “Create bucket” y mantener la siguiente configuración:

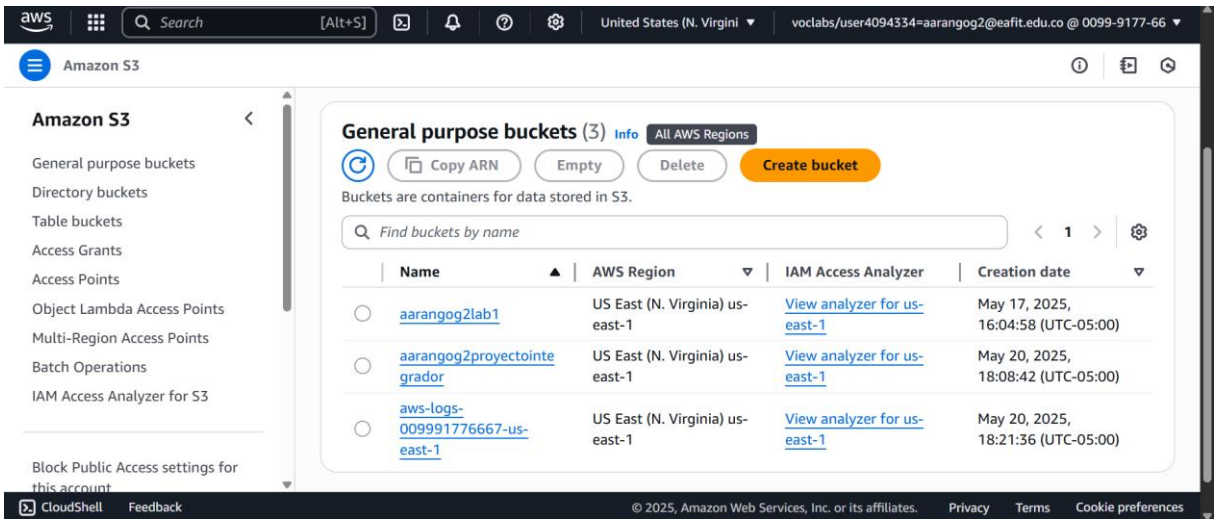


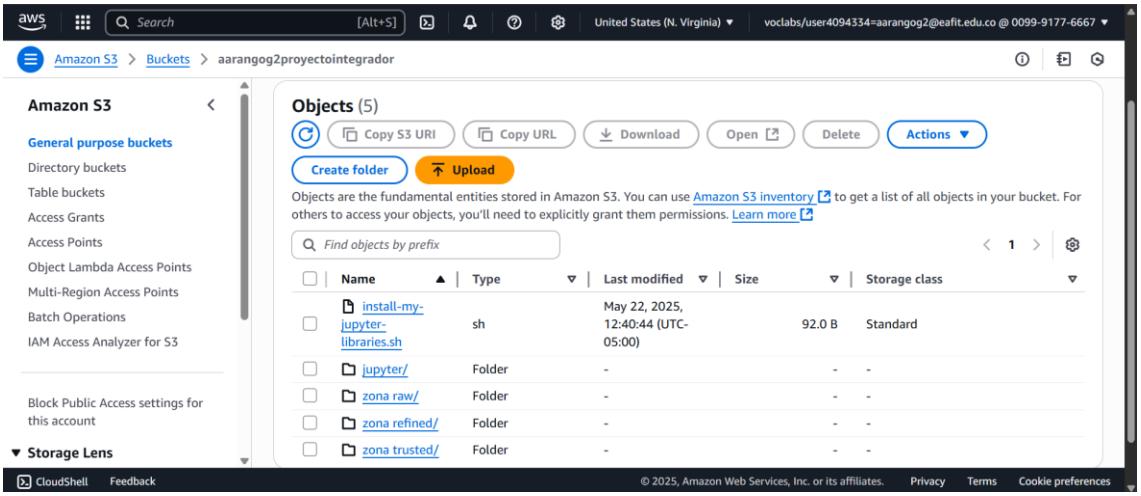
Si se cambia el nombre del bucket, se deben cambiar las rutas definidas en los códigos para leer y guardar los archivos allí almacenados.



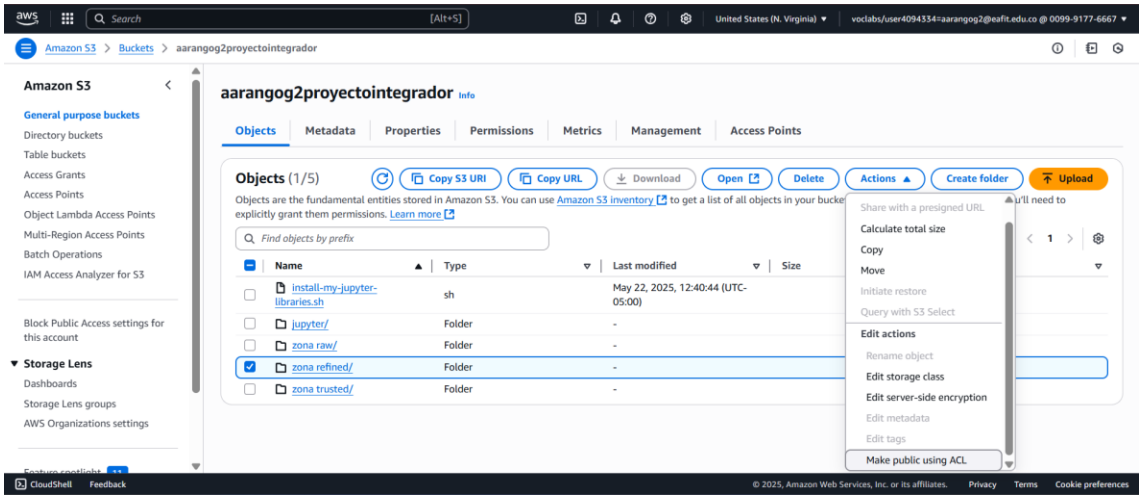


3. Una vez creado el bucket “aarangog2proyectointegrador”, se crean las carpetas: zona raw, zona trusted y zona refined.

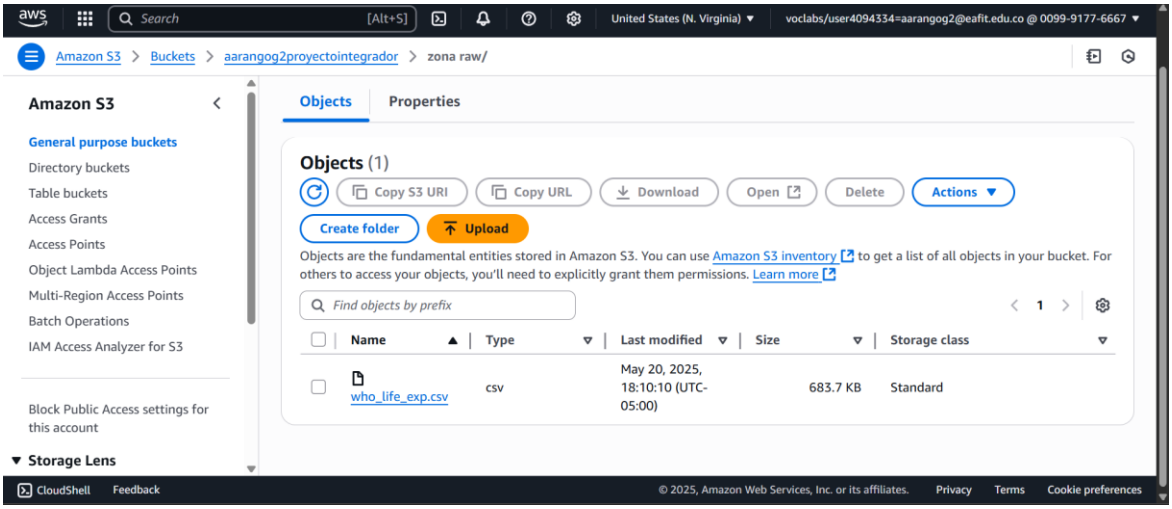




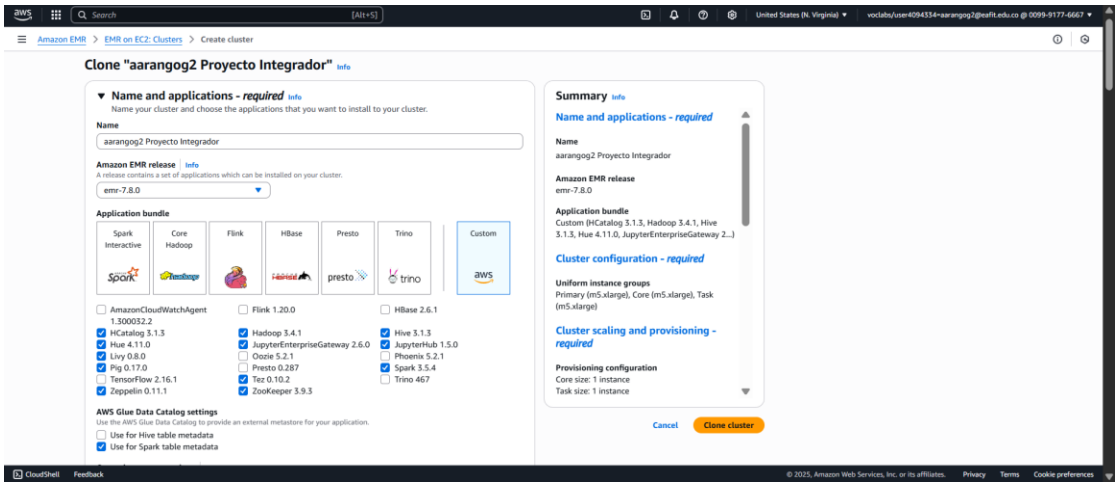
Las zonas “trusted” y “refined” deben volverse públicas para su futura consulta. Esto se realiza por medio de la acción “Make public using ACL”.

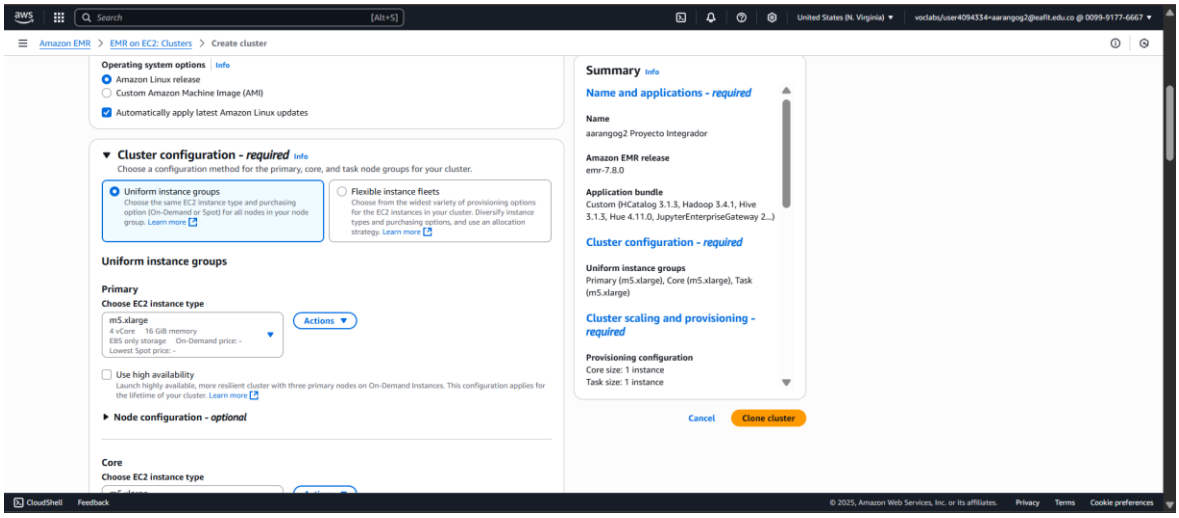


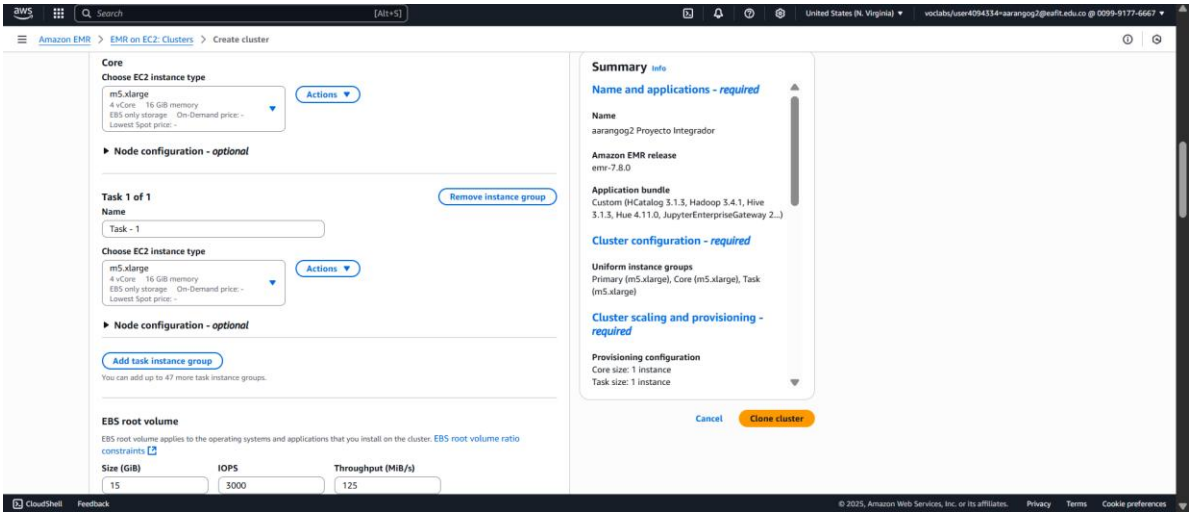
4. En la carpeta “zona raw”, se deben cargar manualmente los datos crudos. En nuestro caso, sería el archivo “who_life_exp.csv”.

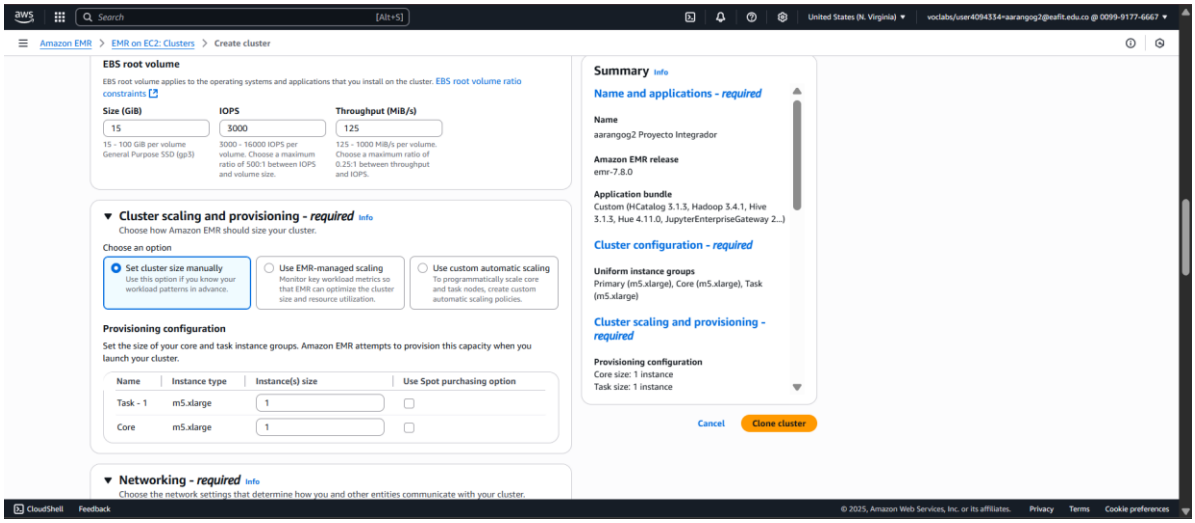


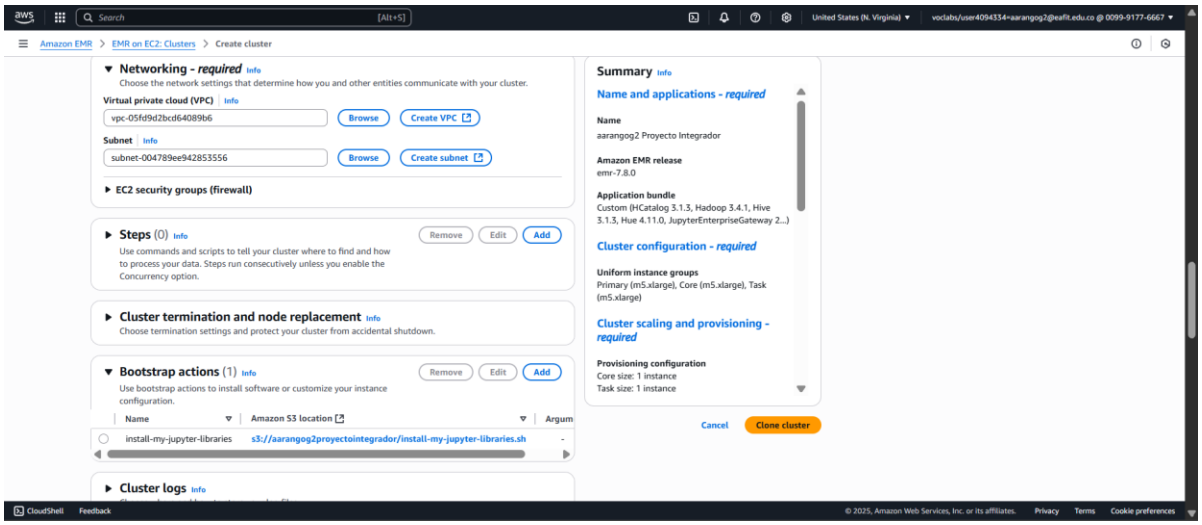
5. Buscar en la interfaz web el servicio EMR y crear un cluster con la siguiente configuración:



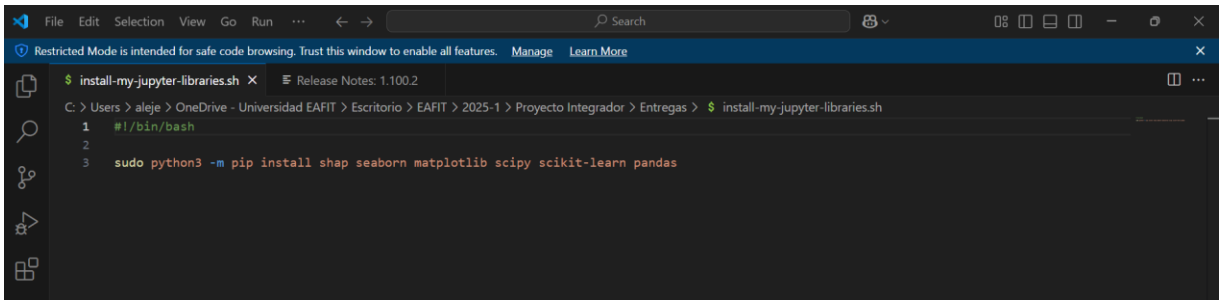


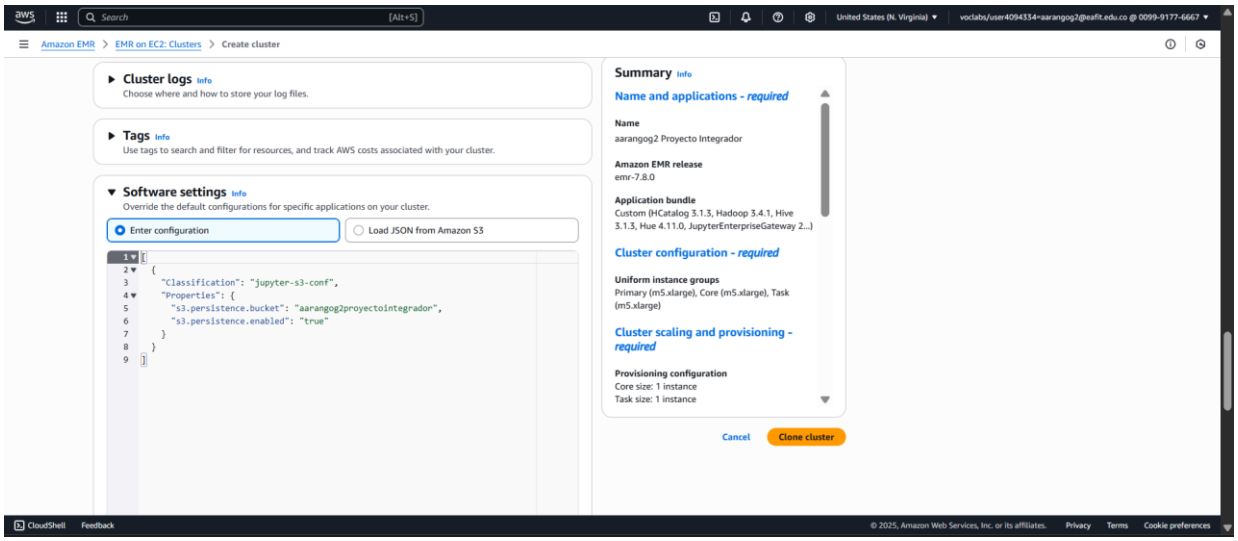


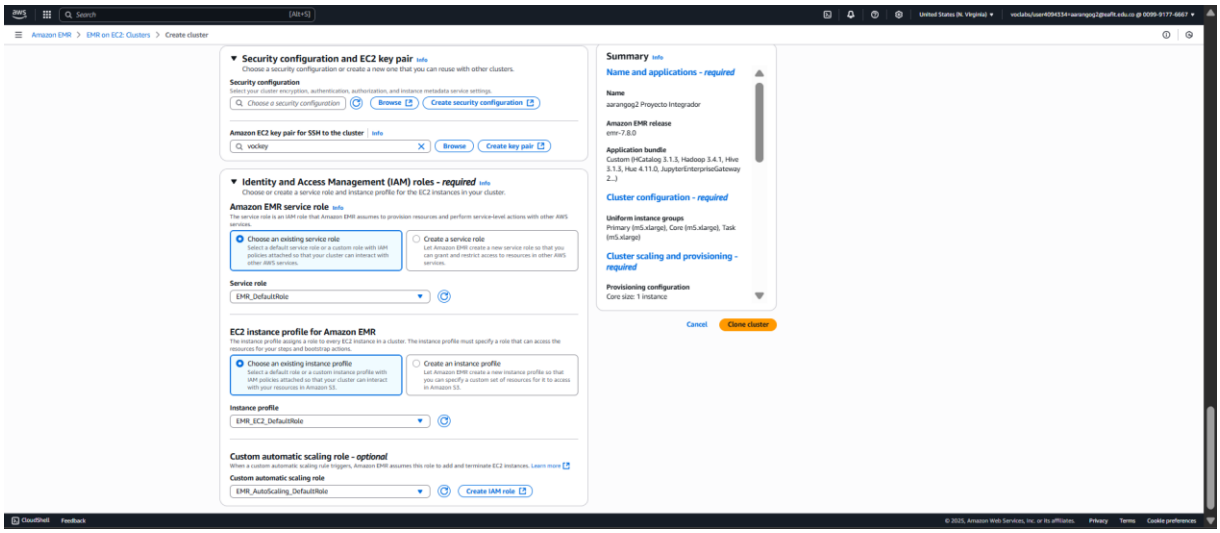




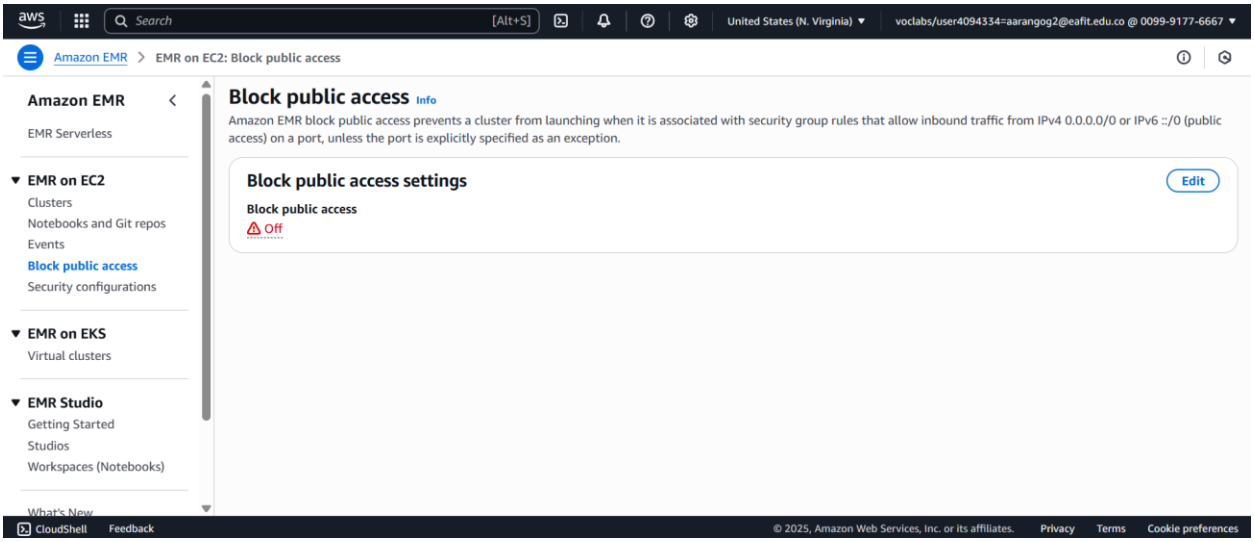
Adicionar “Bootstrap actions” es opcional en este caso de uso, puesto que las librerías utilizadas en las diferentes fases son SparkML y SparkSQL, las cuales forman parte de Apache Spark. Sin embargo, de necesitarse otras librería, se puede crear un código como el siguiente, guardarlo con la extensión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





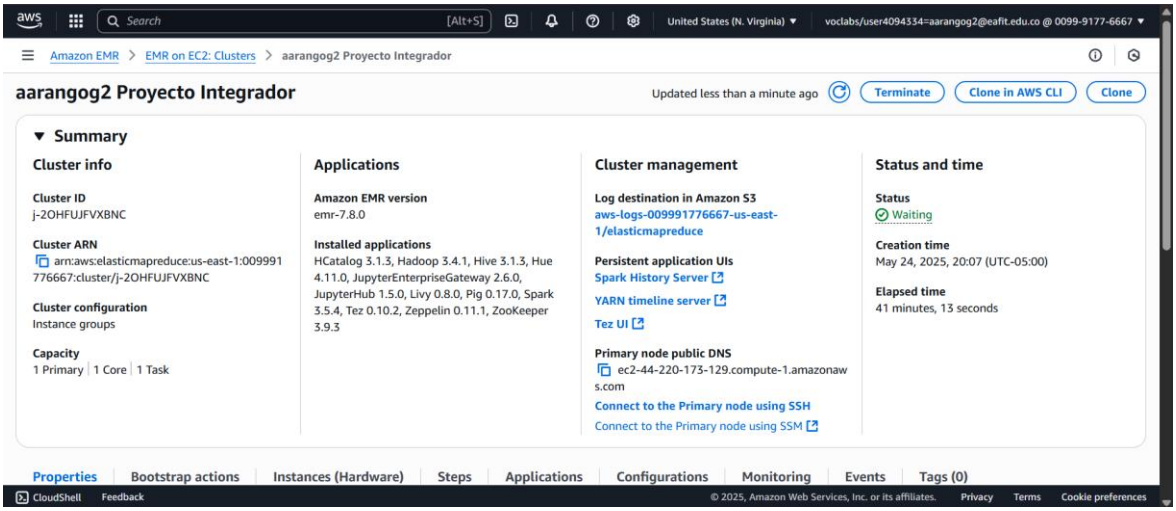


6. Entrar a la opción “Block public access” del menú de la izquierda y abrir todos los puertos TCP para acceso al clúster de la siguiente manera:

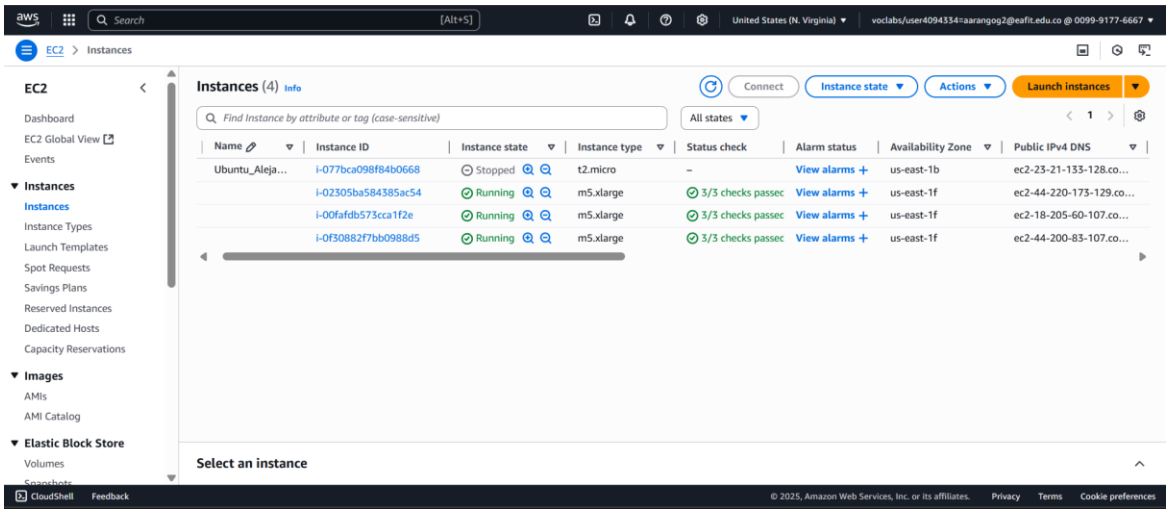


7. Abrir los puertos de las aplicaciones de hadoop/Spark en el Security Group del nodo MASTER del clúster como se muestra a continuación:

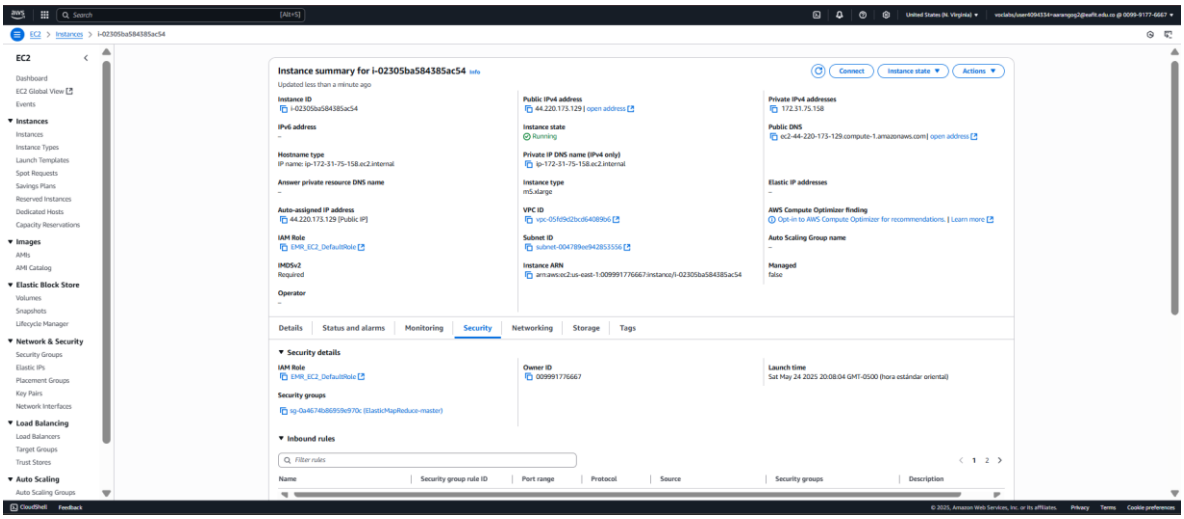
7.1 Identificar el nodo primario del cluster recién creado, el cual se muestra en "Primary node public DNS".



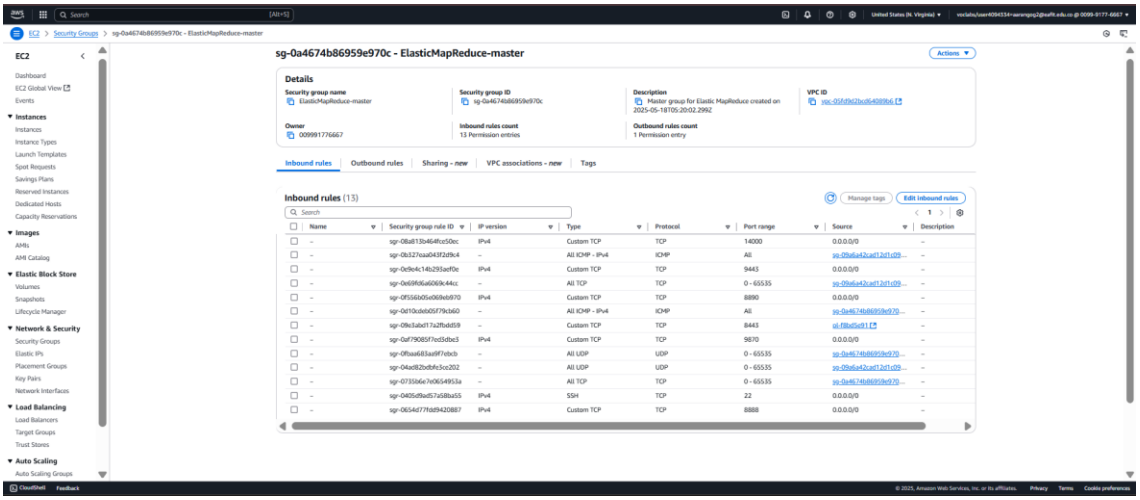
7.2 Buscar en la interfaz web el servicio E2C, donde se encontrarán tres máquinas. Abrir la que tenga el valor de la columna "Public IPv4 DNS" igual al valor del "Primary node public DNS" del cluster.



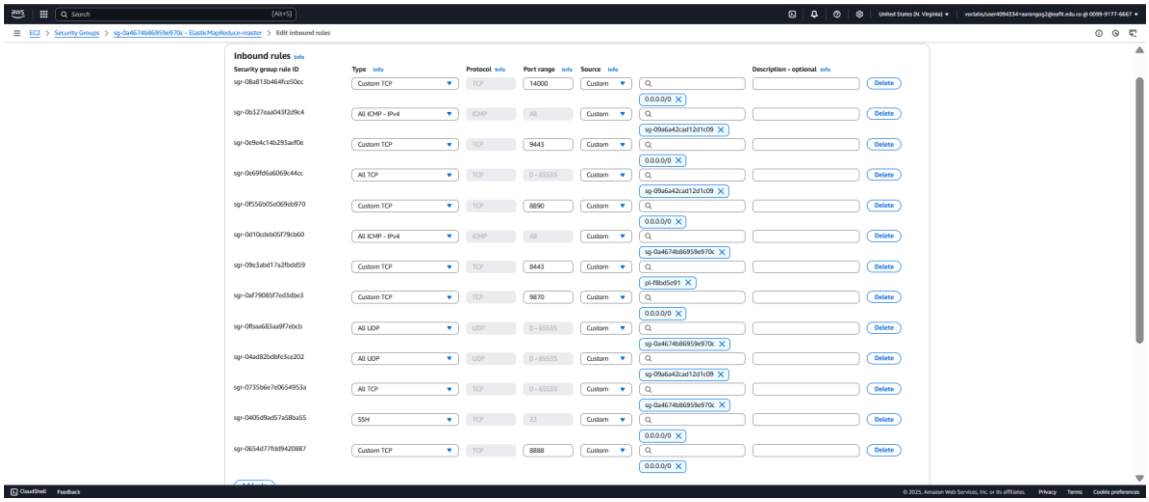
7.3 Entrar a la pestaña de seguridad de la Instancia EC2 del nodo master y abrir la opción “Security groups”.



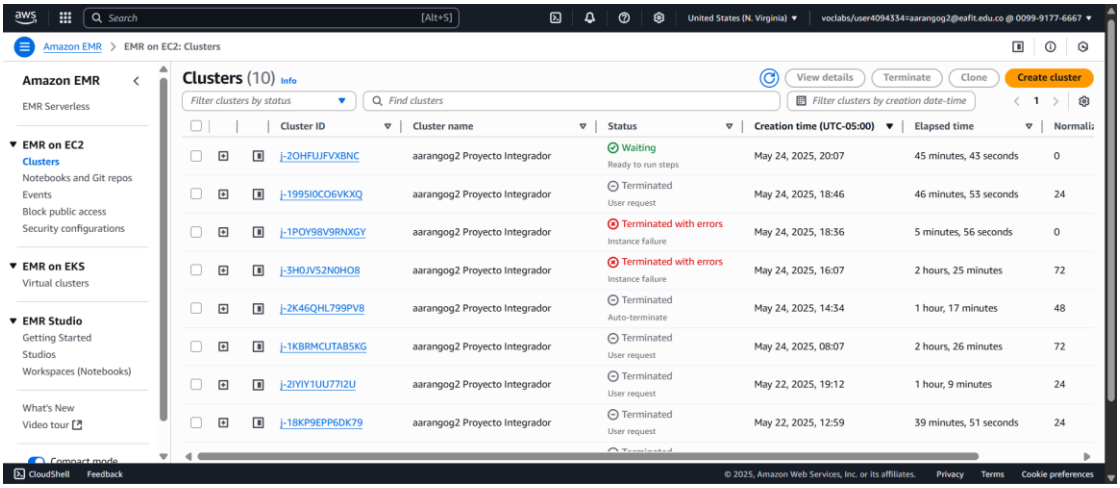
7.4 Entrar a la opción “Edit inbound rules”.

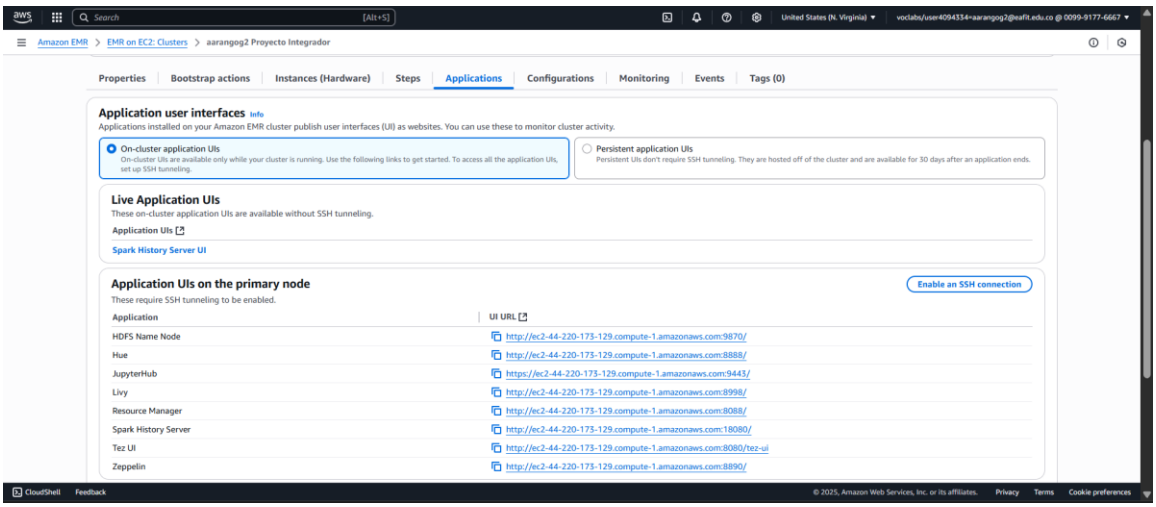


7.5 Habilitar los nodos: 22, 14000, 9870, 8888, 9443, y 8890.



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

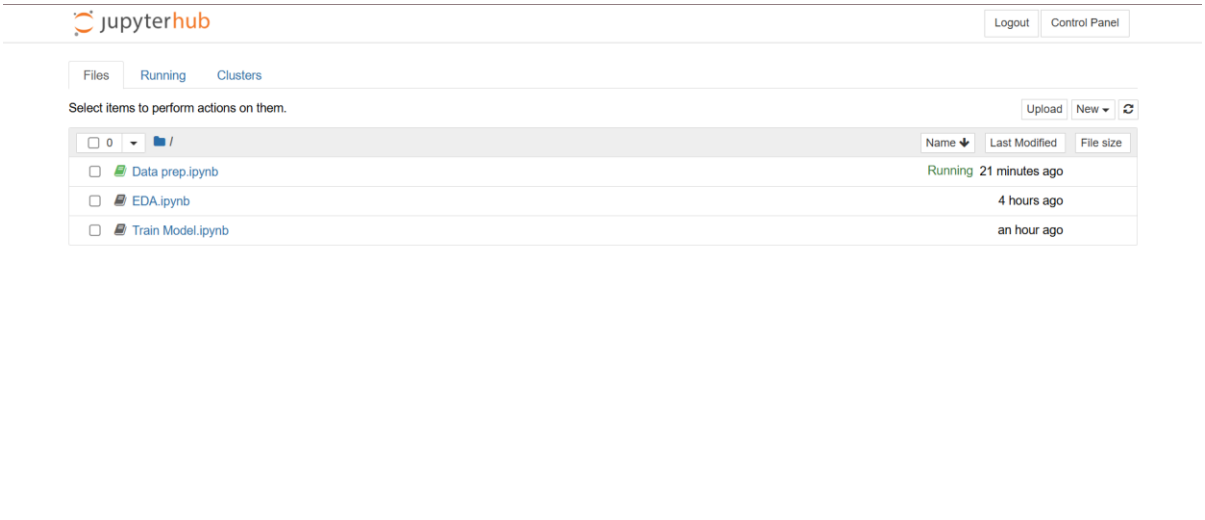




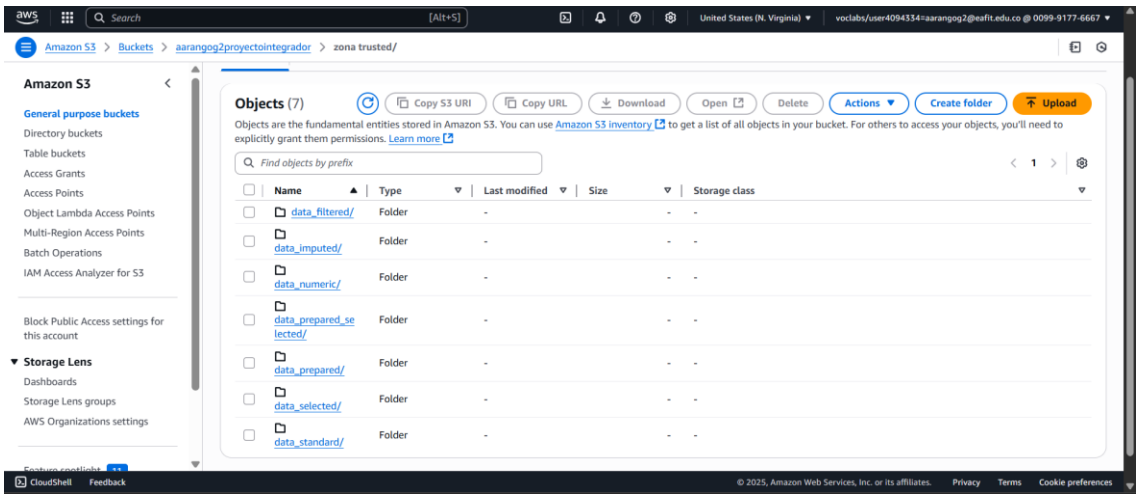
9. Ingresar con las credenciales:

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

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

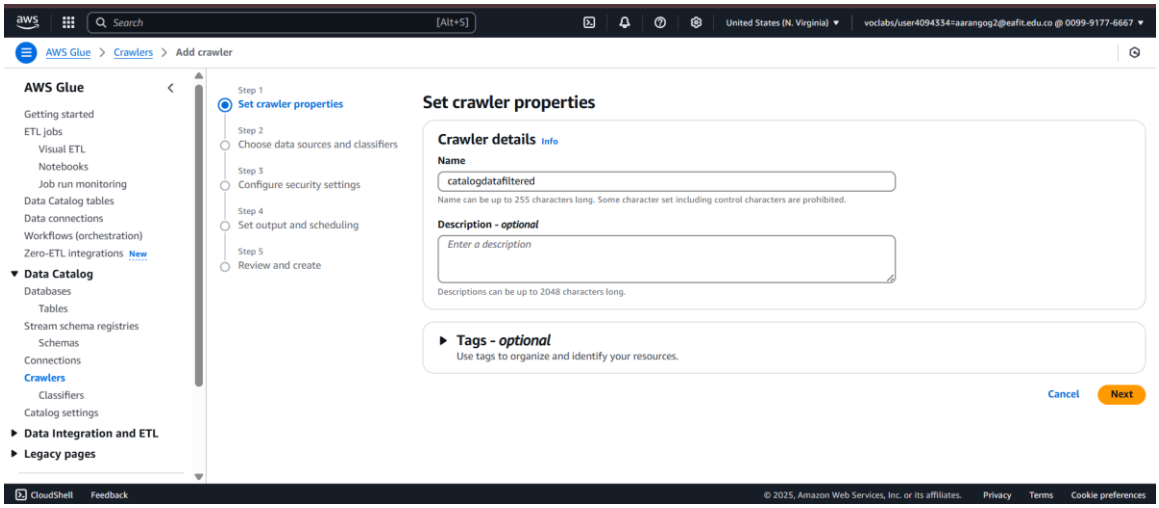


Tras ejecutar este notebook, se generarán los siguientes archivos con formato parquet en la zona trusted del S3:

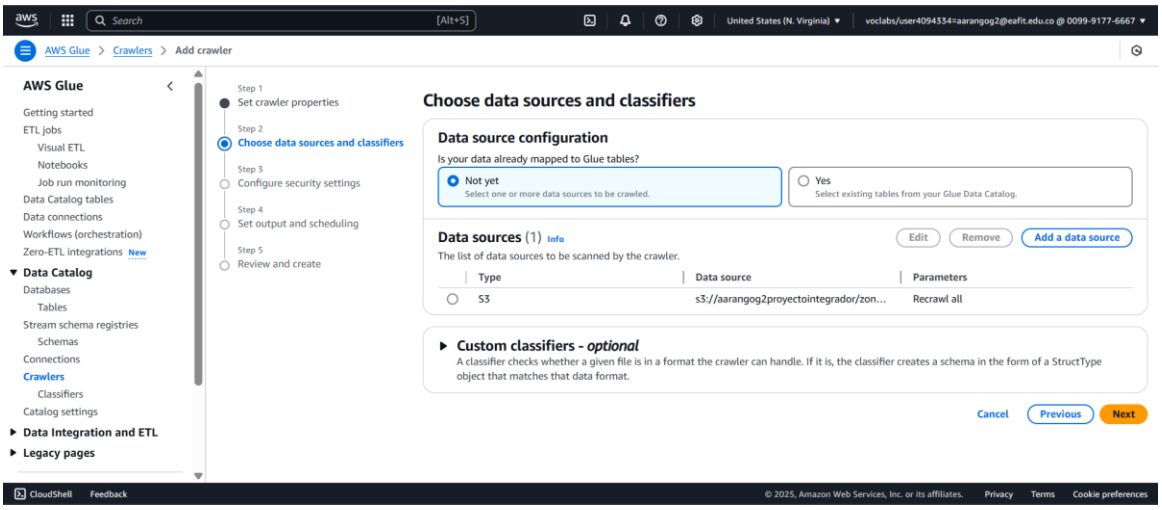


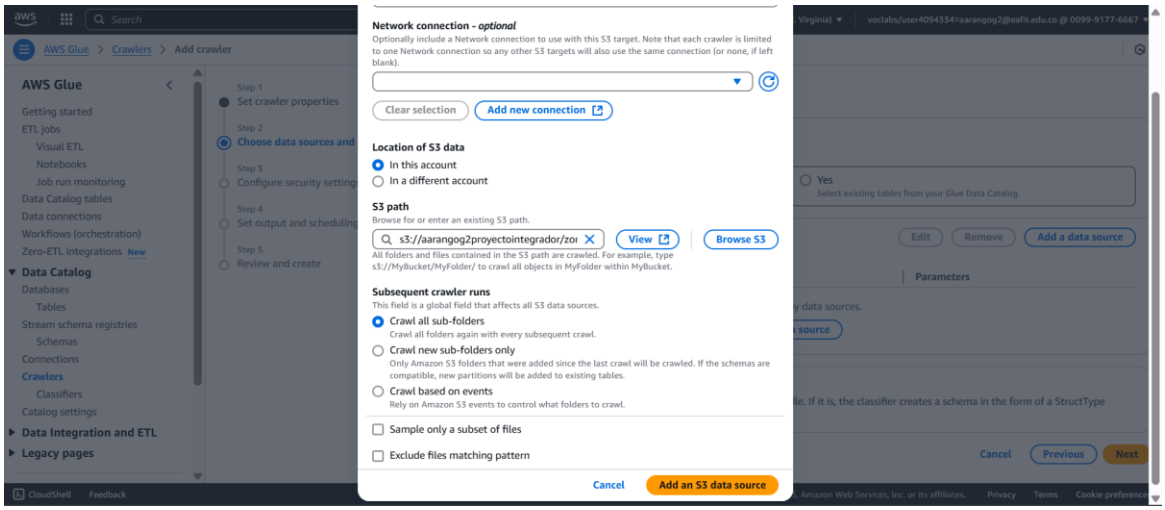
10. Catalogar los resultados del notebook “Data prep.ipynb” con el servicio de AWS Glue, creando un crawler por cada uno de los resultados. Para la creación de un crawler, se debe utilizar la siguiente configuración, y replicarla para cada uno de ellos:

10.1 Seleccionar el nombre del crawler.

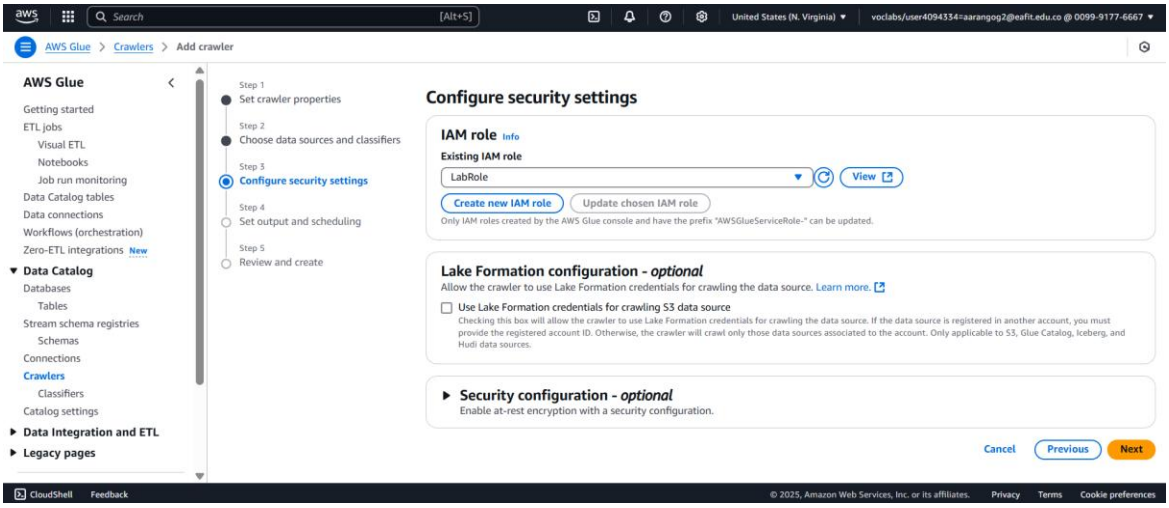


10.2 Seleccionar la ruta al archivo almacenado en S3 que se quiere catalogar.

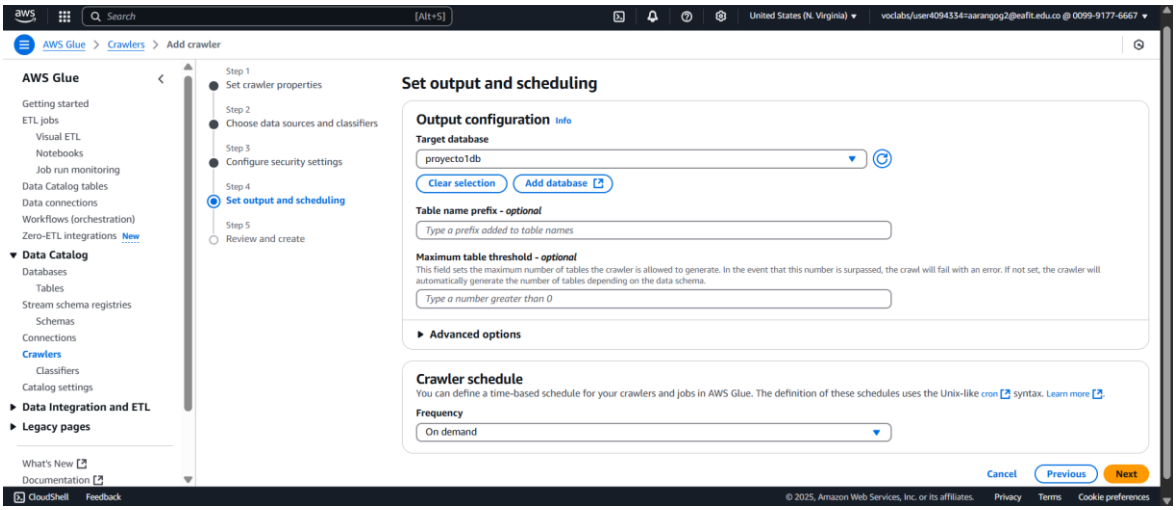




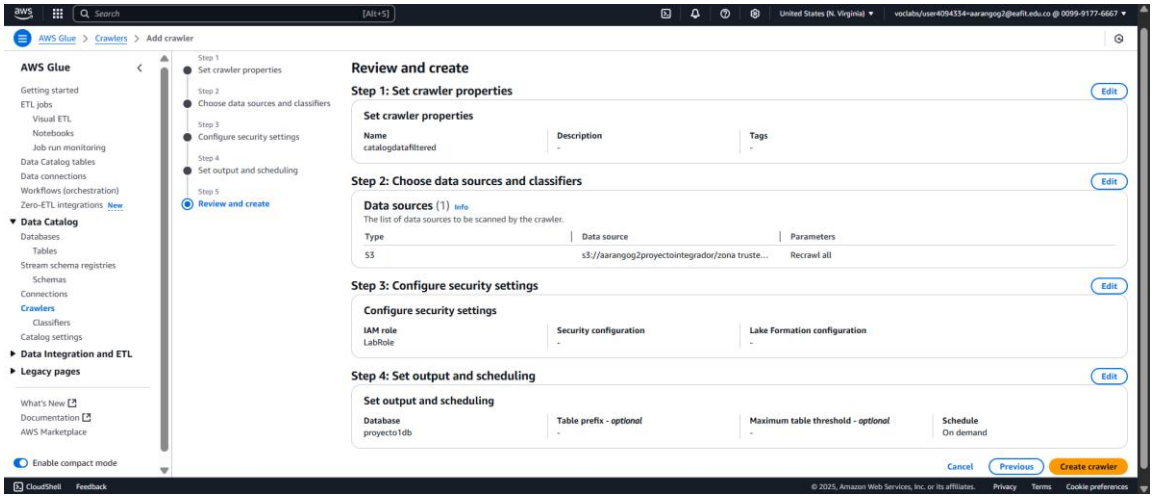
10.3 Seleccionar el rol “LabRole” para el “IAM role”.



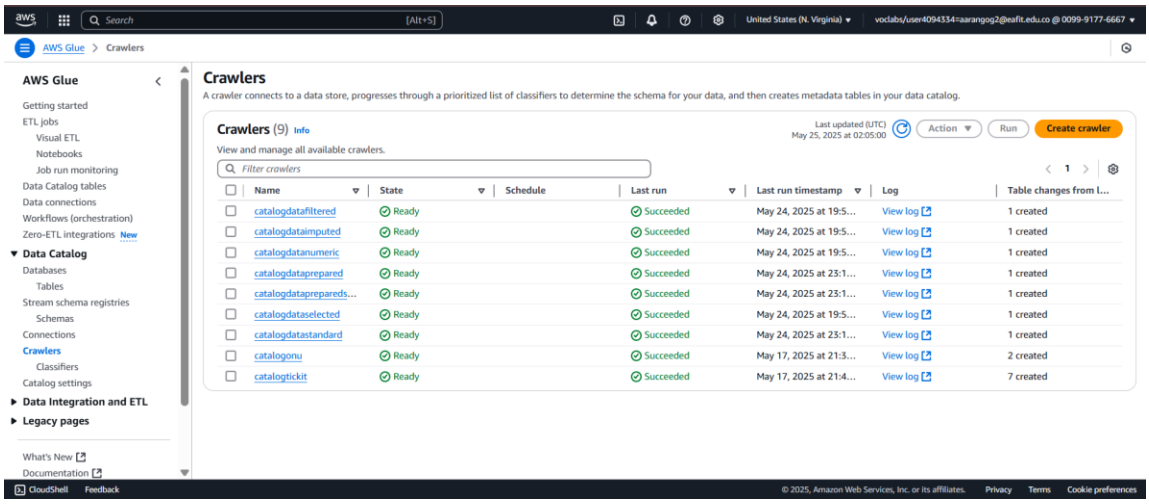
10.4 Crear una nueva base de datos llamada “proyecto1db” y seleccionarla para el almacenamiento de los esquemas.

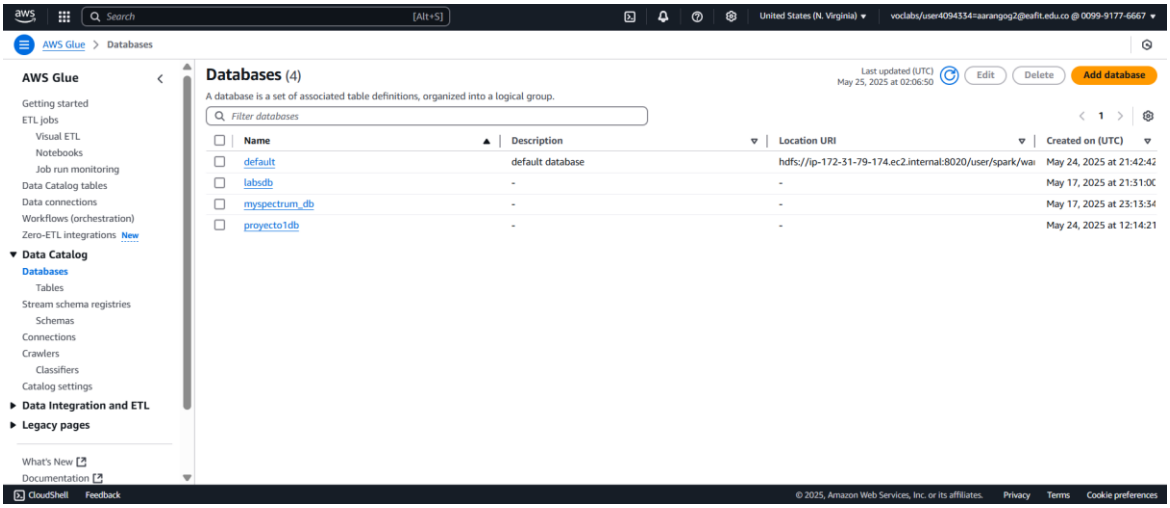


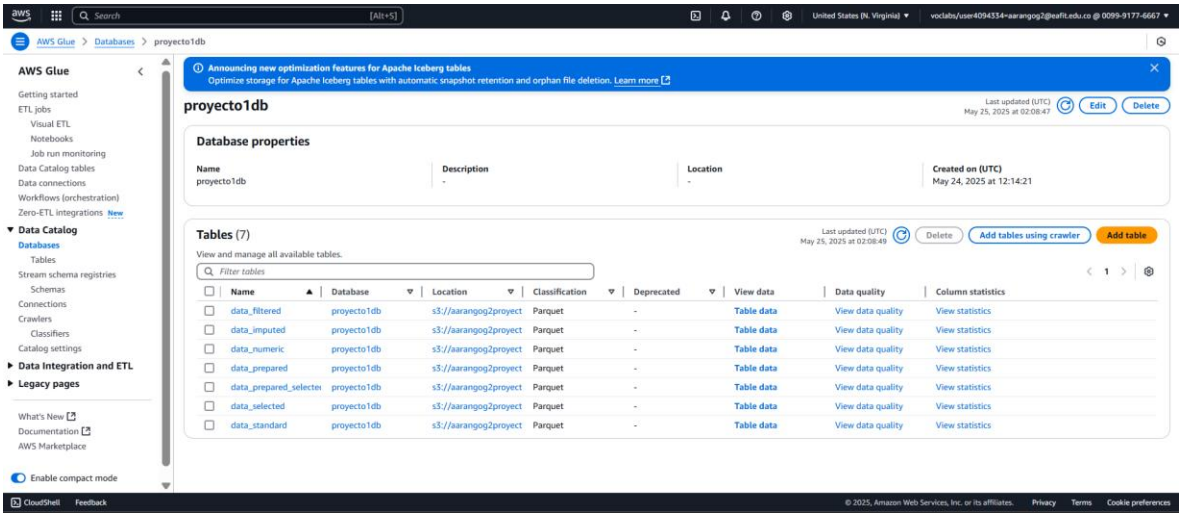
10.5 Crear y correr el crawler.



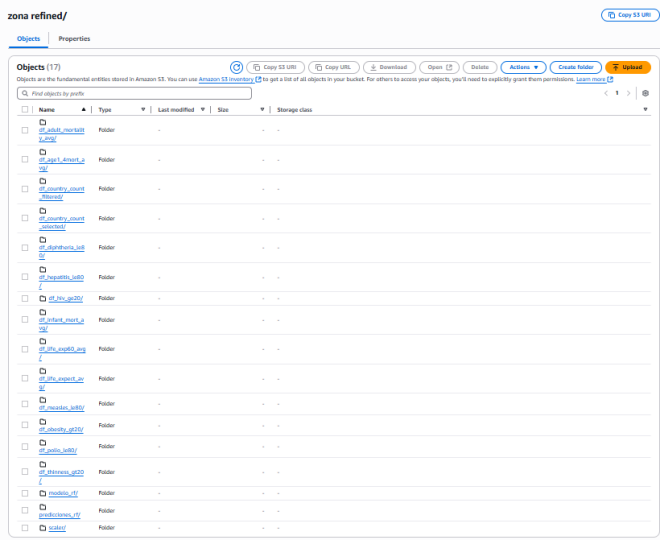
11. Correr todos los crawlers para obtener las siguientes tablas en la base de datos “proyecto1db”:







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.

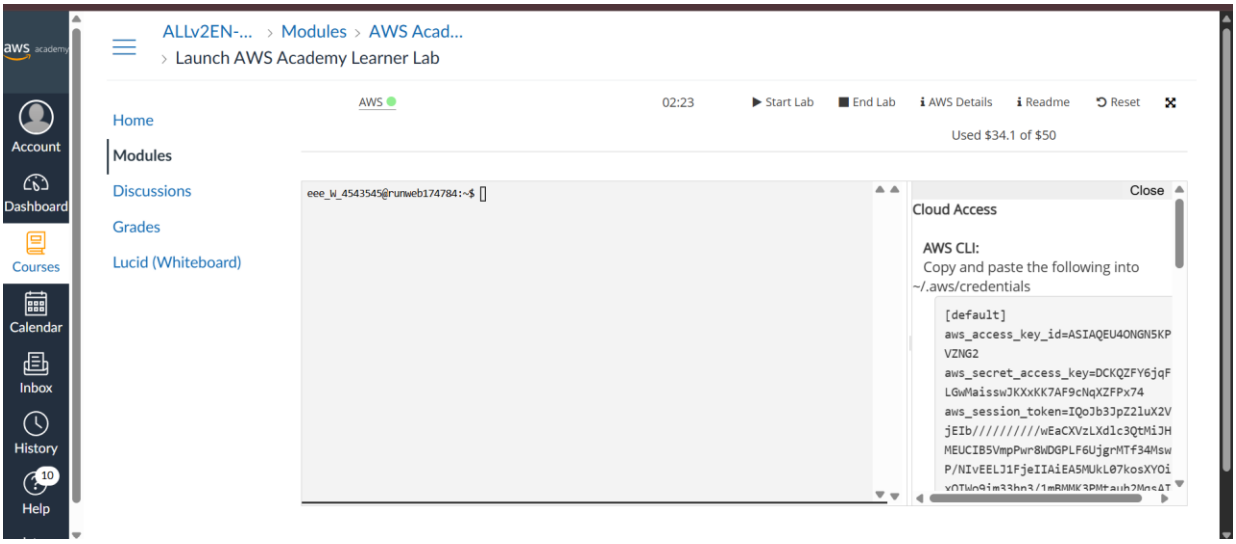
Para la ejecución exitosa del notebook, se deben modificar los siguientes parámetros:

```
# Credenciales
aws_access_key_id = "ASIAQEU4ONGN5KPVZNG2"
aws_secret_access_key = "DCKQZFY6jqFLGwMaisswJKXxKK7AF9cNqXZFPx74"
aws_session_token = "IQoJb3JpZ2luX2VjEIb//////////wEaCXVzLXdlc3QtMiJHMEUCIB5VmpPwr8WDGPLF6UjgrMTf34MswP/NIvEELJ1FjeIIAiEA5MUkL07kosXYOixOIWo9im33hn3/1mBMMK3PMtauh2MqsAI
s3_path = "s3://aarangog2proyectointegrador/zona refined/"
```

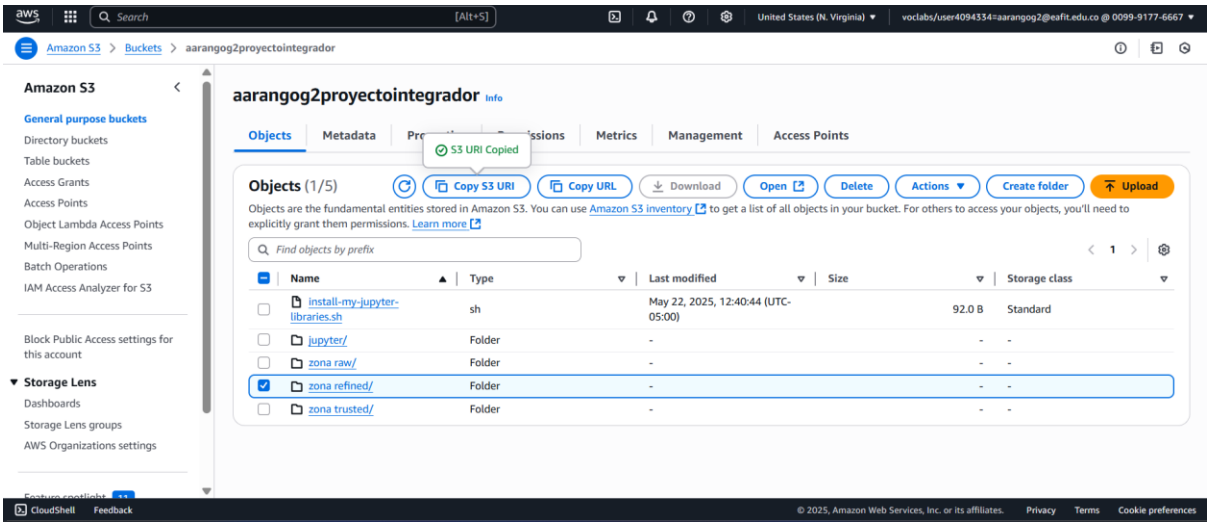
Los parámetros:

- 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”.



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.