1. ***Cree una nueva VPC con las subredes que considere necesarias***

1.1Creación y configuración de VPC

Se realiza la configuración de la VPC con dos subredes una pública y una privada con el fin de garantizar la accesibilidad que permite la red publica y la seguridad de la red privada.

Una captura de pantalla de una red social

Descripción generada automáticamente

Imagen1. Configuración de VPC

1. ***Cree un nuevo bucket S3 con una carpeta para almacenar datos de entrada.***
   1. Creación y configuración inicial de Bucket S3

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Imagen2. Creación de bucket s3

* 1. Carga de la información en el bucket creado.

Previo a cargar la información en el bucket s3 se crean tres carpetas *“/input, /output y /scripts”*

Texto

Descripción generada automáticamente

Imagen3. Creación de carpetas internas bucket s3

* 1. Carga de información en el bucket s3 /input

Procedemos a cargar la base de datos Saber11.csv en la carpeta /*input*

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen4. Carga de CSV en bucket s3

1. ***Cree un clúster EMR donde se ejecutarán las consultas***
   1. Creación clúster EMR

Creamos el EMR teniendo en cuenta la correcta asignación de los roles, claves y VPC, adicional se crea con jupyterhub, aunque en este ejercicio no se usará dicha extensión.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Imagen5. Creación clúster EMR

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen6. Configuración seguridad EMR

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen7. Configuración roles EMR

Asignamos la VPC creada en el paso 1.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen8. Asignación de VPC a clúster EMR

Seleccionamos la subred pública.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen8. Asignación subred a clúster EMR

1. ***Cree un NoteBook EMR que lea los datos de S3 y resuelva las actividades planteadas***
   1. *Creación y configuración de Studio-EMR*

En este caso y dado que no se especifica que tipo de notebook se debe trabajar, se creará el notebook en Jupyter, razón por la cual se debe crear y configurar un Studio, teniendo en cuenta asignarle la VPC creada en el punto 1, asignándole ambas subred tanto la pública como la privada, adicionalmente se debe configurar para trabajar con el rol LabRole.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Imagen9. Configuración inicial Studio

* 1. *Configuración wokspace*

Creado el workspace, se procede a asignarle un clúster EMR (el cual fue creado en el punto anterior), para ello es necesario detener el workspace y proceder con la asignación, teniendo en cuenta la configuración de los grupos de seguridad.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Imagen10. Asignación de clúster EMR a Workspace

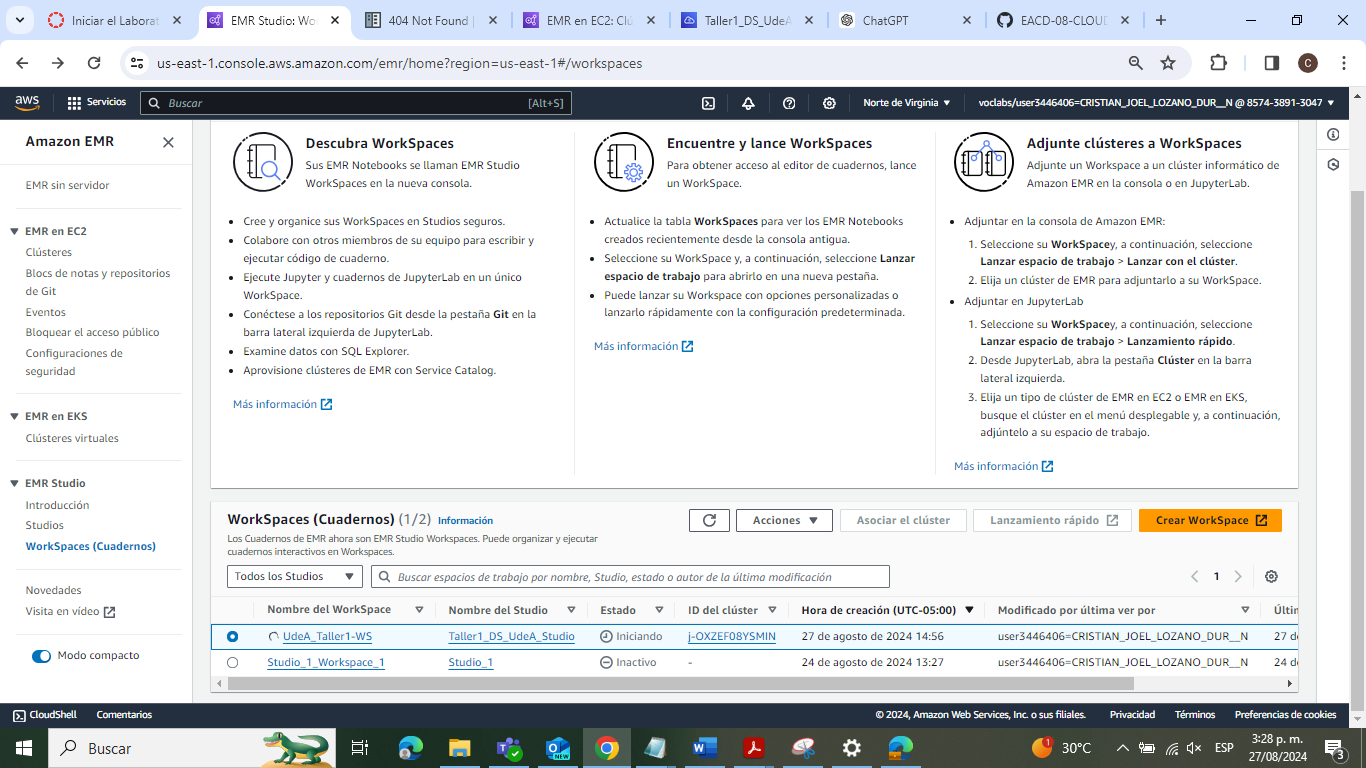


Imagen11. Inicio de workspace

* 1. *Notebook y descarga*

Creado en workspace se procede a conectarlo al kernel PySpark y se programa para que trabaje con el archivo cargado en el bucket s3 ya configurado el cual se encuentra en *(“https://s3-udea-ds.s3.amazonaws.com/input/Saber11.csv”)*

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen11. Espacio de trabajo en jupyter con código de ML

1. ***Descargue el NoteBook en formato ipynb***

Se procede a descargar el archivo en formato .ipynb

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Imagen12. Descarga de código en .ipynb

1. ***Revise las métricas EMR y de las instancias EC2 creadas para el EMR e identifique cuales métricas requieren de seguimiento***
2. ***Cree alarmas de CloudWatch que notifiquen cuando considere que el comportamiento de las métricas debe ser advertido y adicione acciones si lo considera oportuno*.**

Para desarrollar los últimos dos numerales (6 y 7) es necesario tener acceso al panel de administración de facturación y costos; sin embargo, el tipo de licencia con el que se está trabajando no lo permite, razón por la cual no se pudieron desarrollar estas configuraciones.

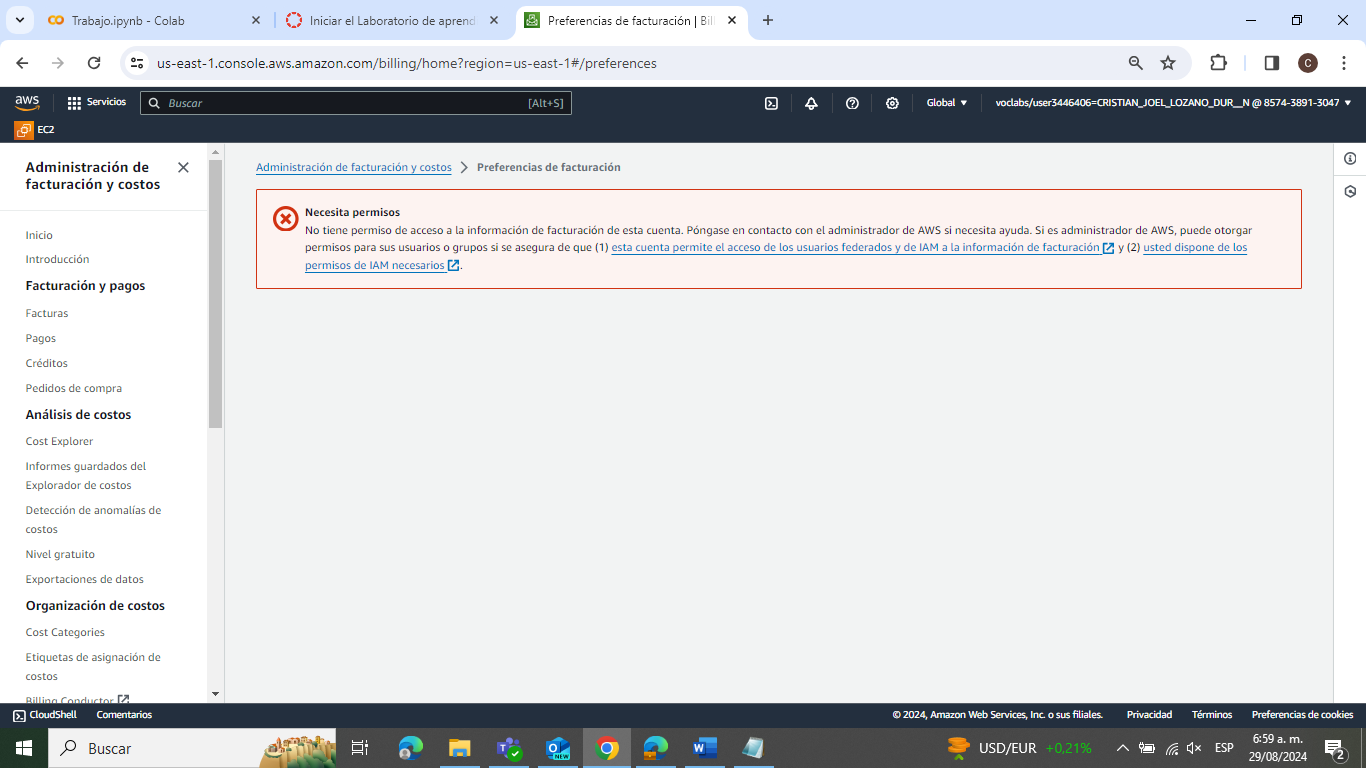


Imagen13. Error al ingresar a panel de administración de facturación y costos.

***Nota:*** *algunos archivos y ajustes cambian de nombre durante a lo largo del informe debido a que, durante el desarrollo de la actividad fue necesario iniciar varias veces el laboratorio, razón por la cual se reinició la actividad diversas veces para garantizar su correcto desarrollo.*