**UNIVERSIDAD EAFIT**

**ST0263 Tópicos Especiales en Telemática**

**2023-2**

Documentación y Evidencias Laboratorios Unidad 3 – Big Data

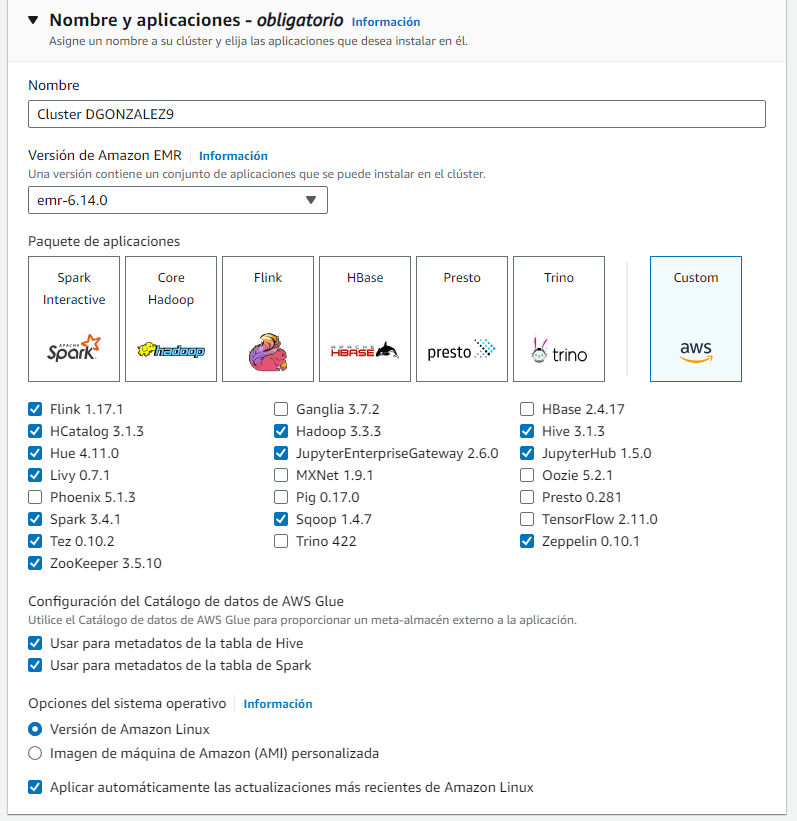
**Nombre :** David González Tamayo **- email :** dgonzalez2@eafit.edu.co

LABORATORIO 3-0: Crear un Clúster AWS EMR en Amazon para trabajar todos los laboratorios.

Para la creación del clúster se siguió la guía dada por el profesor Edwin Nelson Montoya Munera. Para más detalles dar click en el siguiente link:

* <https://github.com/st0263eafit/st0263-241/blob/main/bigdata/00-lab-aws-emr/Install-AWS-EMR.pdf>

Para la creación del clúster EMR se utilizó la misma configuración del tutorial dado por el profesor:



Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

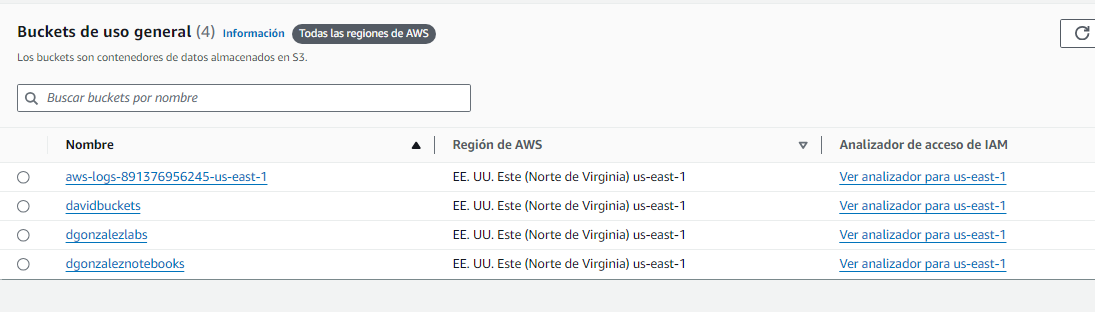
Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se utilizó la VPC y las subredes default que trae el AWS por defecto dichas en el tutorial.

Para guardar lo notebooks generados utilizaremos S3, entonces crearemos un bucket llamado de la siguiente manera:



Ahora, en la sección de software settings, se coloca la siguiente configuración, tenga en cuenta el nombre del bucket S3 previamente creado. Se cambia ‘MyJupyterBackups’ por el bucket creado anteriormente.

Texto, Carta

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Posterior se configura la clave privada llamada ‘dgtkey’para conectarnos al máster via SSH:

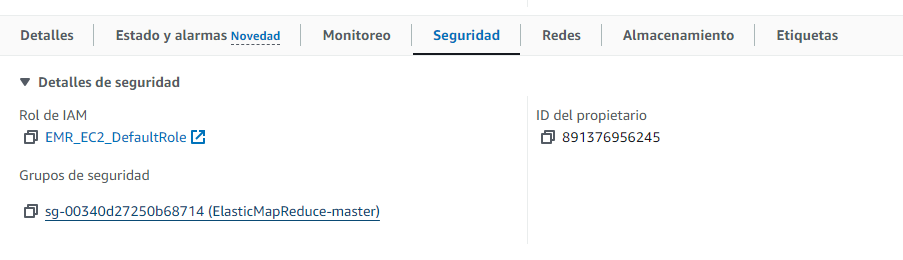
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

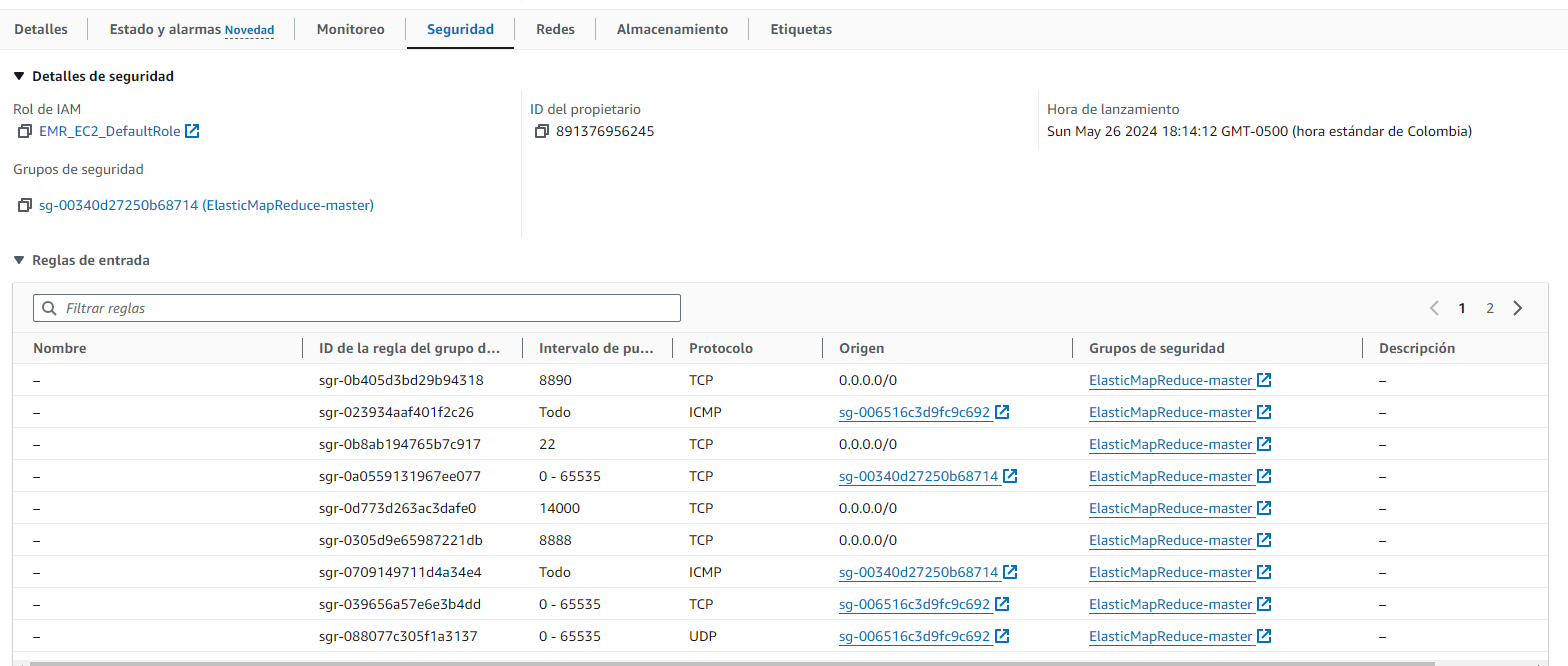
Y por último para crear este clúster, se deben colocar los roles necesarios para crear el EMR:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



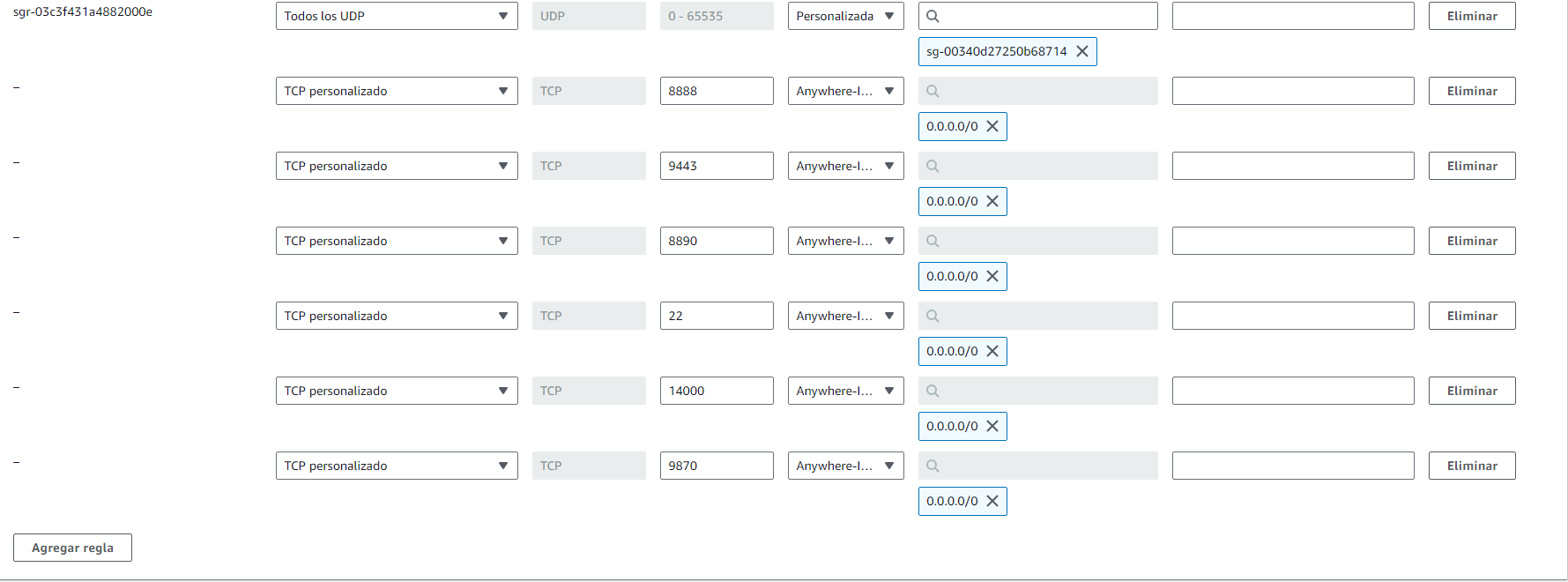
Ahora se presiona el botón de crear el clúster, este proceso se tardará alrededor de 20 minutos, mientras tanto puede ir configurando las reglas de seguridad. Necesitas abrir los siguientes puertos: **8888, 9443, 8890, 22, 14000, 9870**

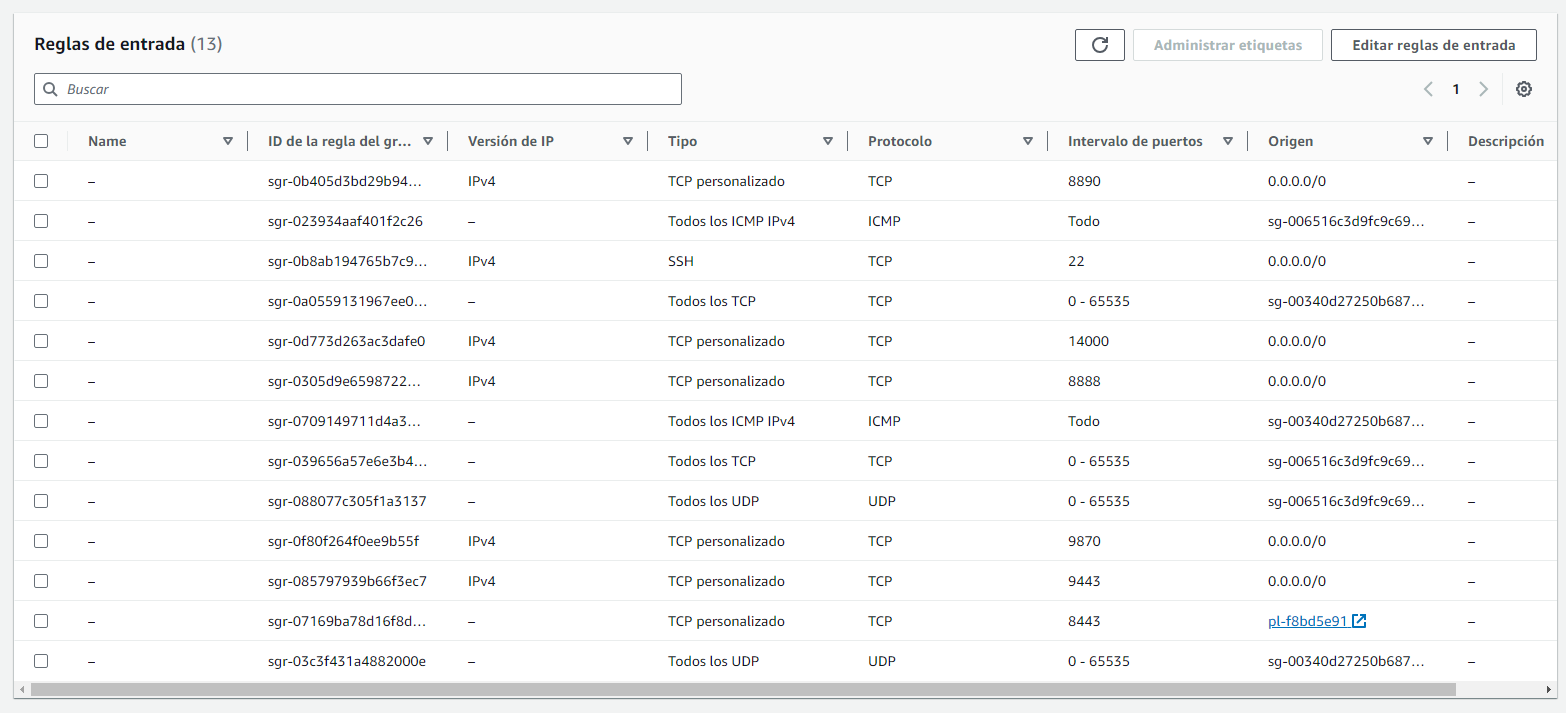


Se presiona ‘Editar reglas de entrada’ y se configura los puertos mencionados anteriormente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente





Patrón de fondo

Descripción generada automáticamente con confianza baja

Accedemos a las aplicaciones y entramos enlace de HUE con el puerto 8888:

<http://ec2-18-207-1-128.compute-1.amazonaws.com:8888/hue>

Le asignamos el nombre de usuario ‘hadoop’ y la contraseña que cumpla con lo requerido que sale en la siguiente foto:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Una vez iniciado sesión en HUE debe salir algo así:

Interfaz de usuario gráfica, Texto, Correo electrónico

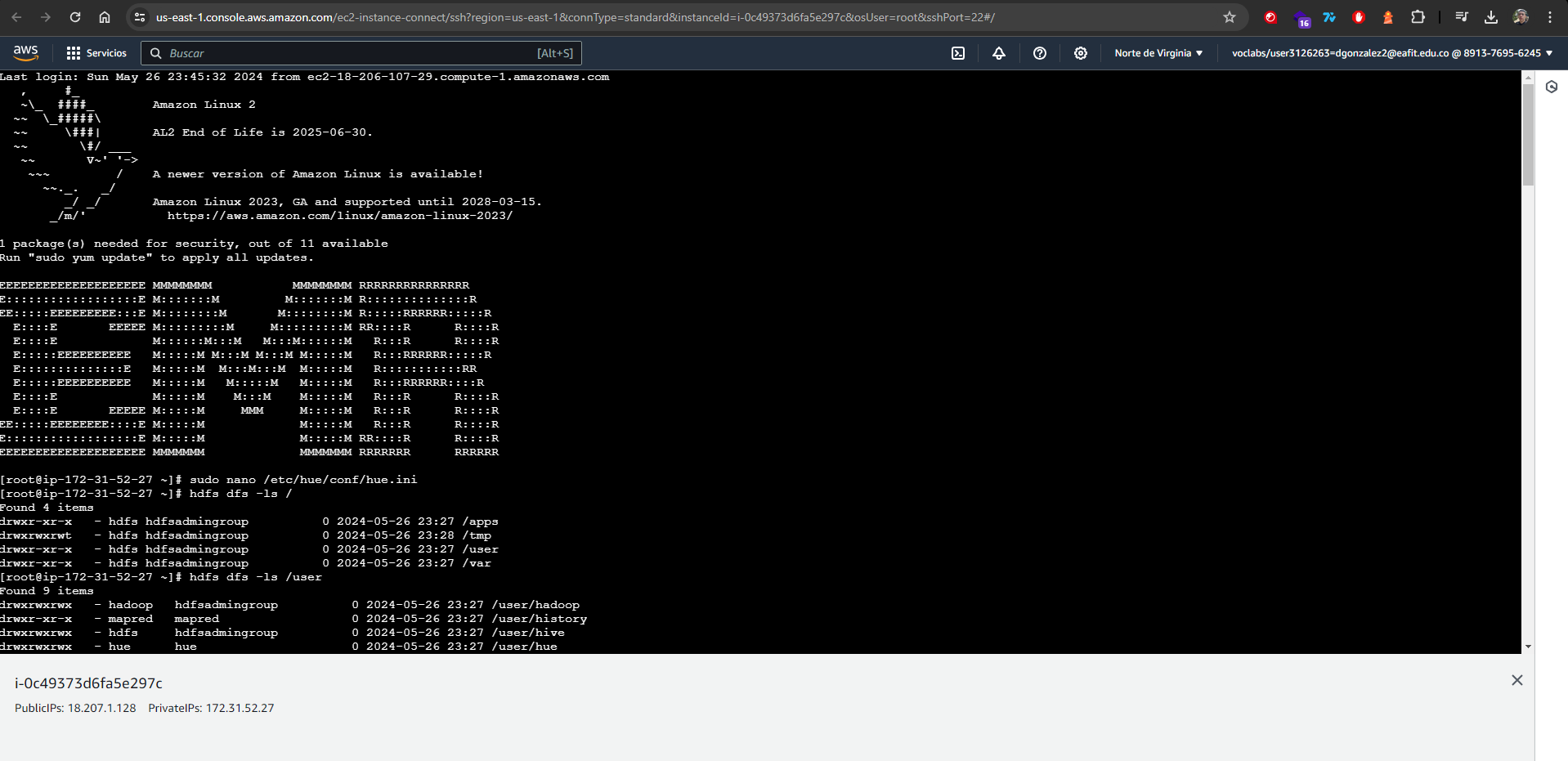
Descripción generada automáticamente

Ahora si nos fijamos bien nos vamos en el icono de files posteriormente a archivos en el menú. Nos indica un error mostrado en la imagen:

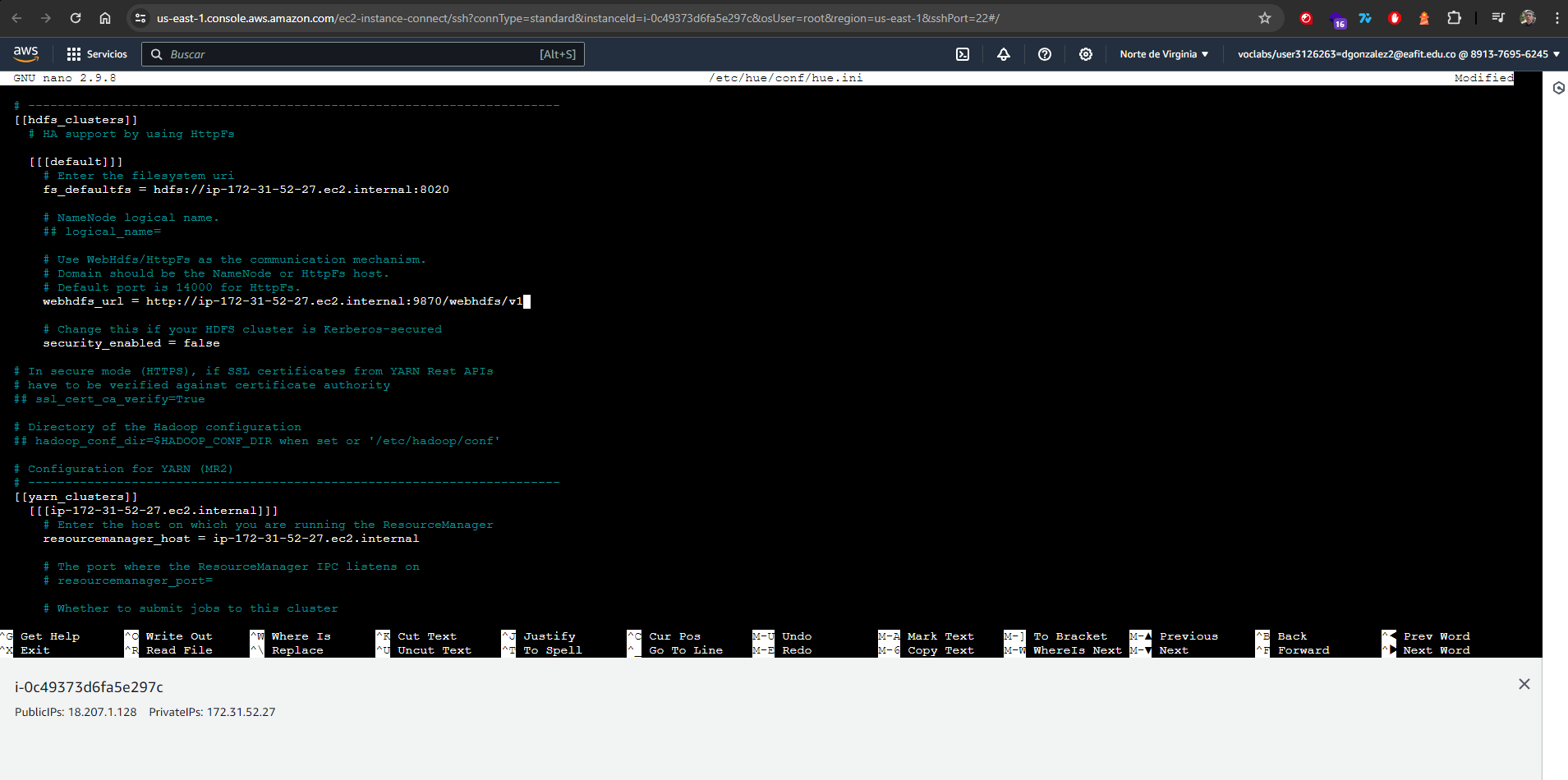
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Para solucionar este error debemos de entrar al nodo máster por medio de SSH. Para identificar el master nos fijamos en la ip pública que se pone en el navegador cuando abrimos hue. Nos conectamos y nos debería salir algo como esto:



Ahora localizaremos el archivo `hue.ini`. Para ello, usa el comando `sudo nano /etc/hue/conf/hue.ini`. Luego, busca la línea que contiene `webhdfs\_url` presionando `Control + W`. Una vez ahí, cambia el puerto 14000 por 9870.



Una vez cambiado el puerto guardamos y salimos para así aplicar el siguiente comando ‘sudo systemctl restart hue.service’

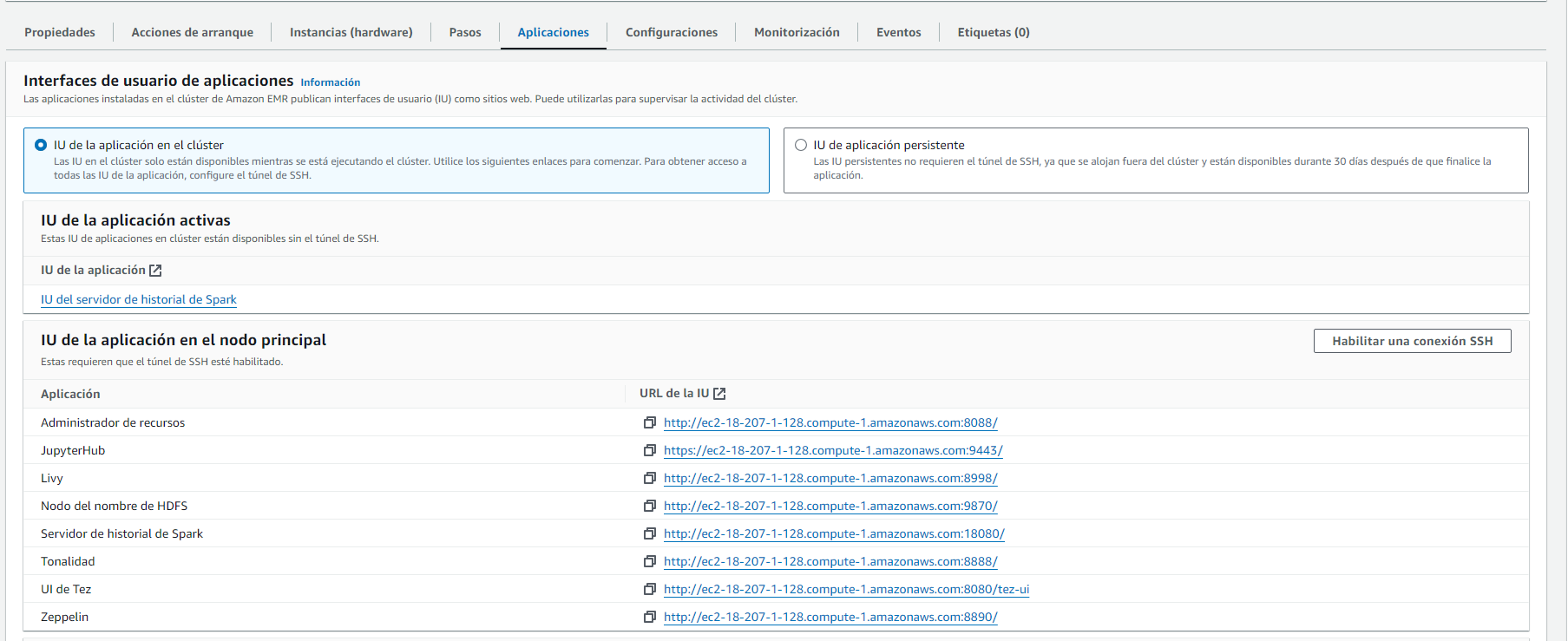
Volvemos a la página de HUE nos ubicamos en files nos debería aparecer lo siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Luego abrimos el enlace de ‘Jupyterhub’ que nos da en aplicaciones

<https://ec2-18-207-1-128.compute-1.amazonaws.com:9443/>



La información que este asignado es:

Username: jovyan

Password: jupyter

Interfaz de usuario gráfica

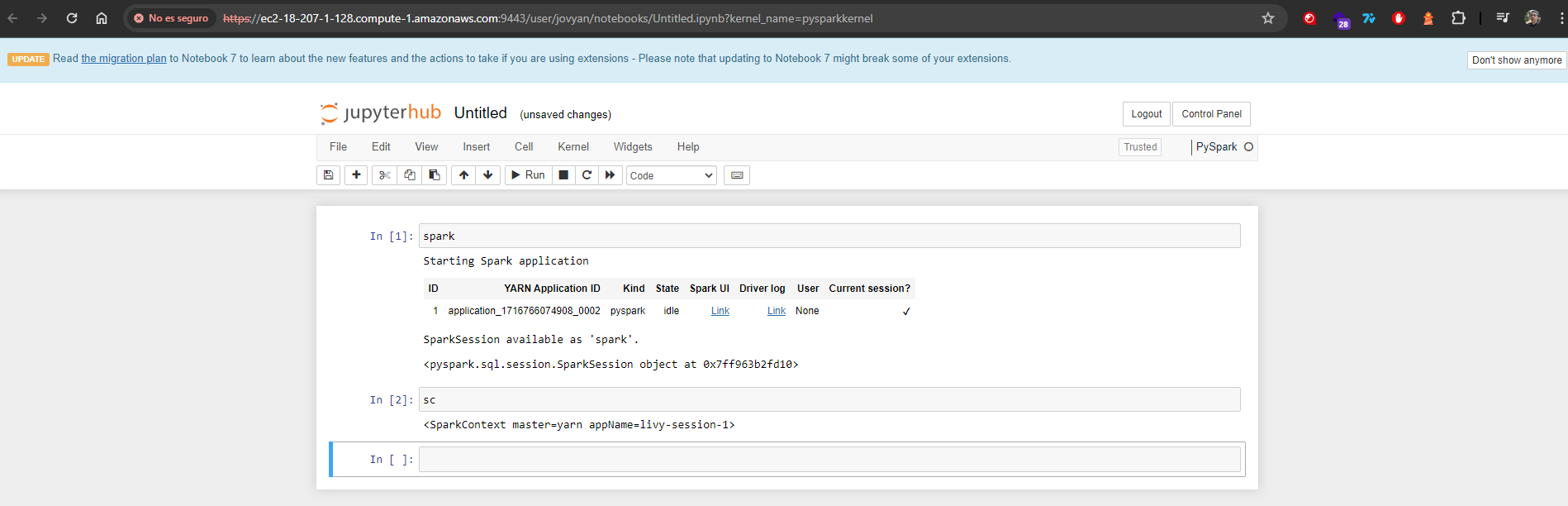
Descripción generada automáticamente

Una vez iniciado sesión nos debería aparecer así, luego seleccionamos en files la opción ‘New’ y oprimimos ‘PySpark’

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Posteriormente escribimos ‘spark’ y le damos ‘Run’ y así igualmente para ‘sc’.



Para verificar si está corriendo el programa nos regresamos al ‘Control Panel’ y nos indica que se guardó correctamente.

Texto

Descripción generada automáticamente con confianza media