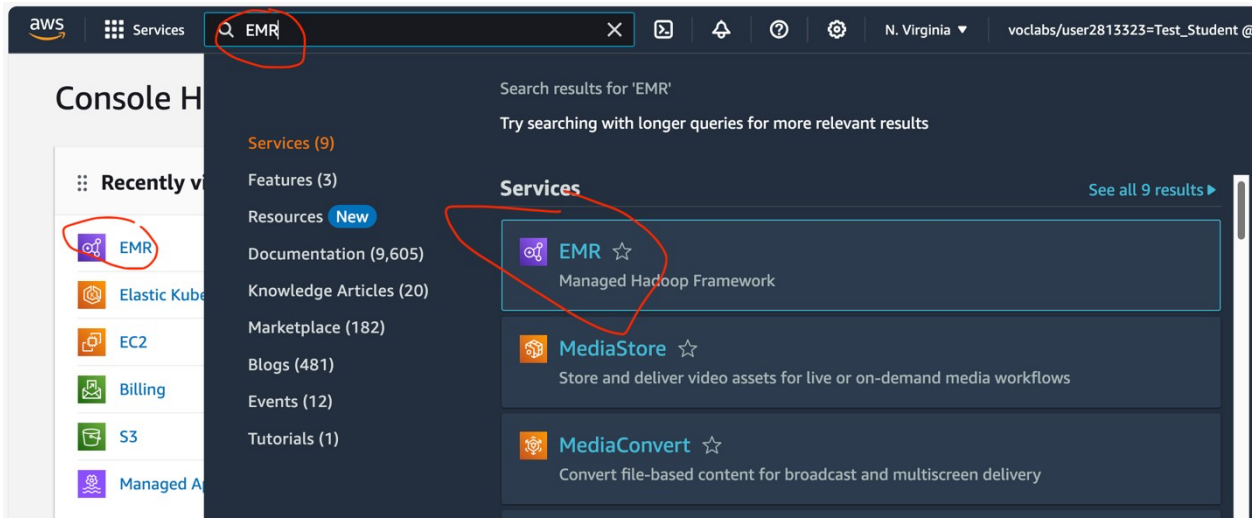
Universidad EAFIT

Laboratorio: Instalar un clúster EMR versión 7.9.0 Hadoop/Spark
Fecha: 22 mayo 2025

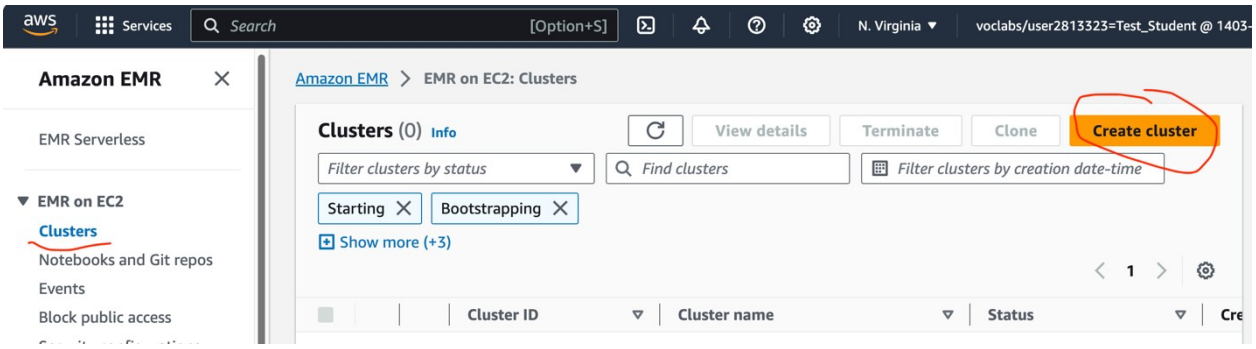
Parte 1: Crear un clúster AWS EMR versión 7.9.0

1 Instalar AWS EMR

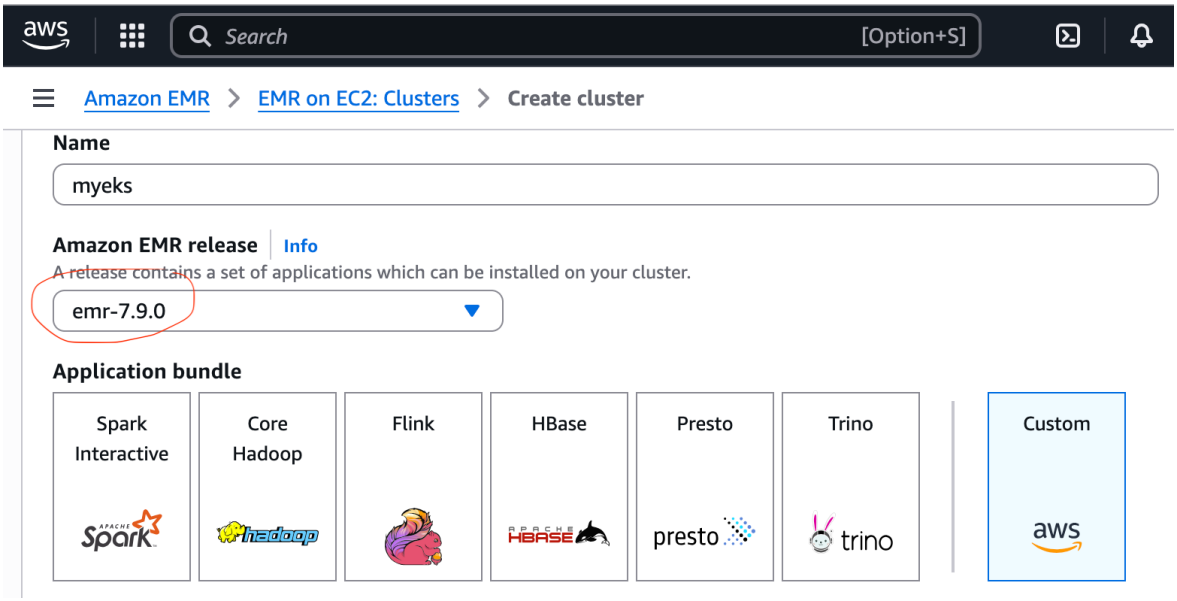
1. Buscar el servicio AWS EMR: Entrar a la consola web de AWS y buscar el servicio EMR:



2. crear clúster



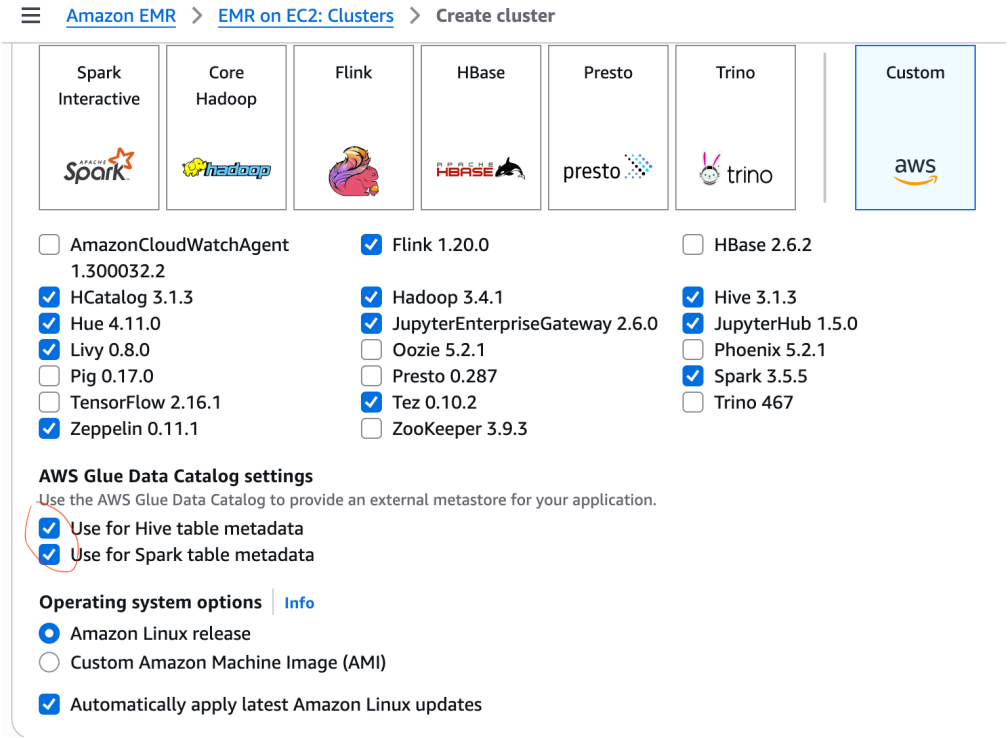
3. nombre, versión y Custom



4. seleccionando los paquetes adecuados para el curso y activando los catálogos Glue, Hive, Spark

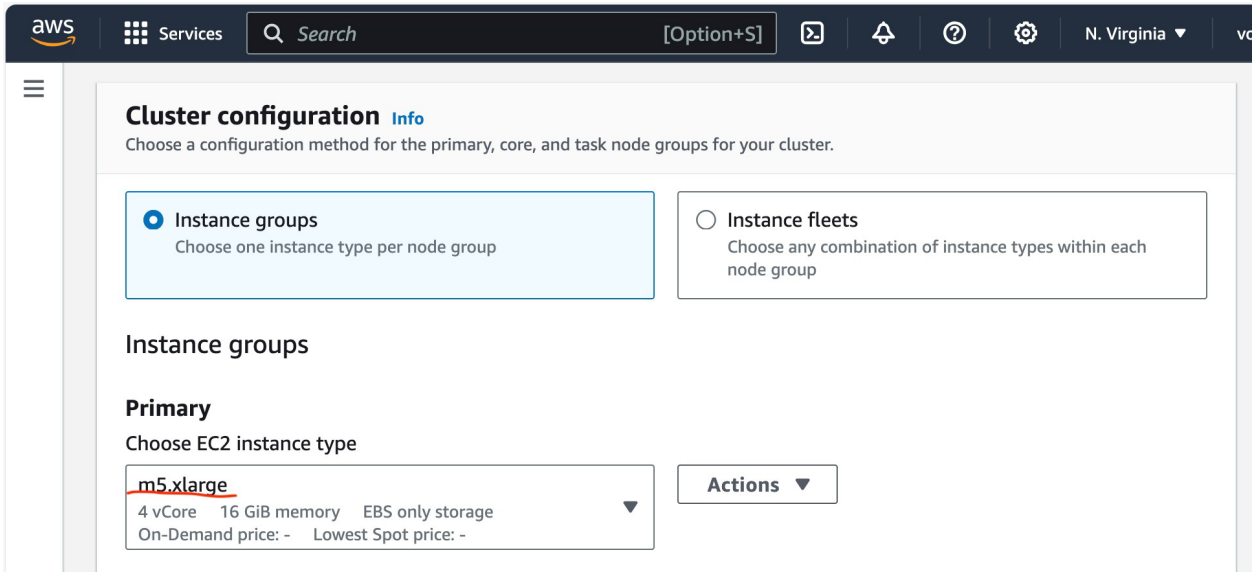
Nota: seleccionar los catalogos Hive y Spark permite ver las tablas AWS Glue en EMR, y las tablas Hive se podrán ver en Glue / Athena.

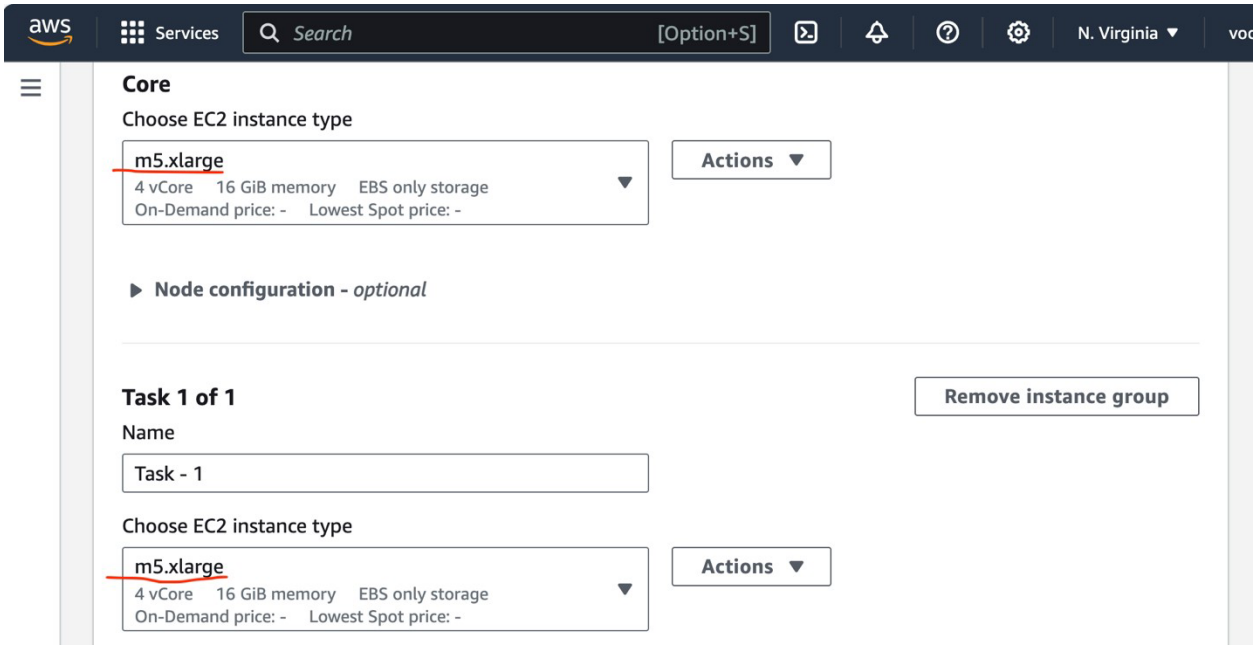
Seleccione los paquetes hadoop/Spark en azul.



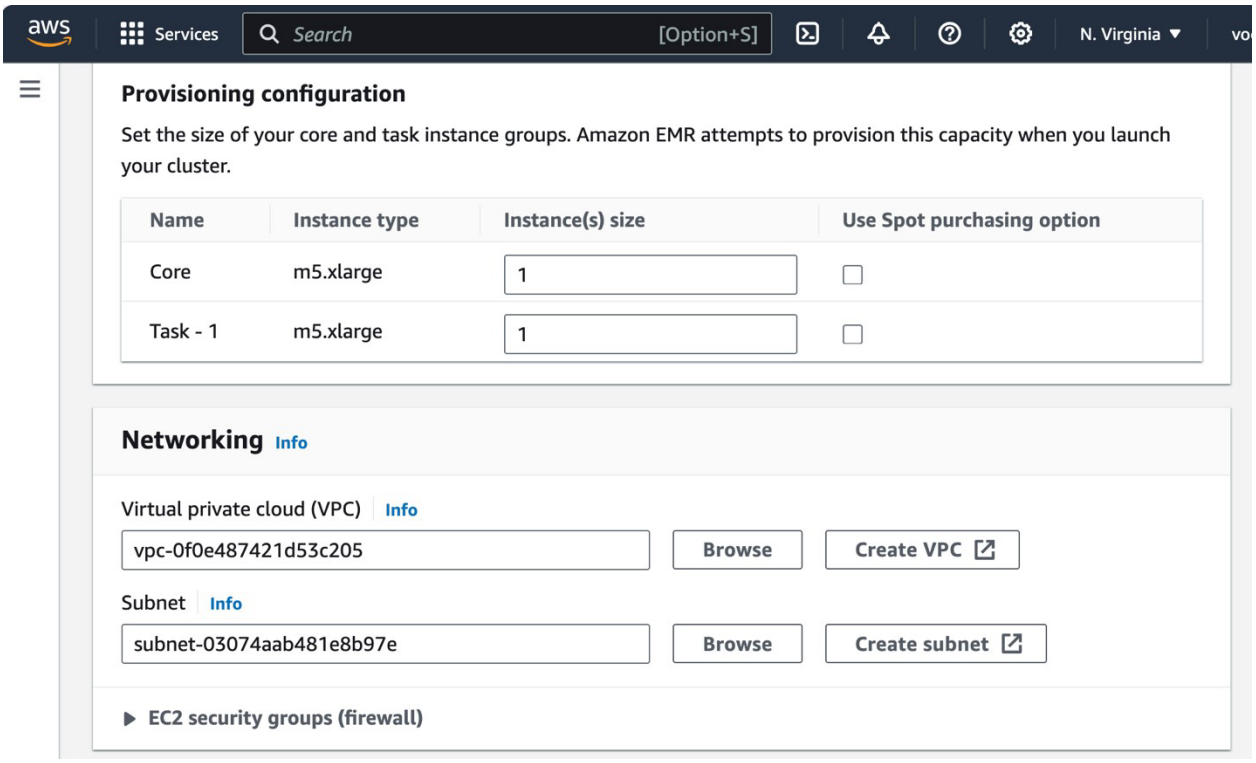
5. Máquinas EC2 del Clúster

Puede dejar las máquinas por defecto m5.xlarge, en algunos momentos puede fallar la creación del clúster porque no tiene suficientes recursos, puede cambiar estas máquinas a m4.xlarge, pero por defecto dejarlas como nos sugiere la creación del clúster EMR.

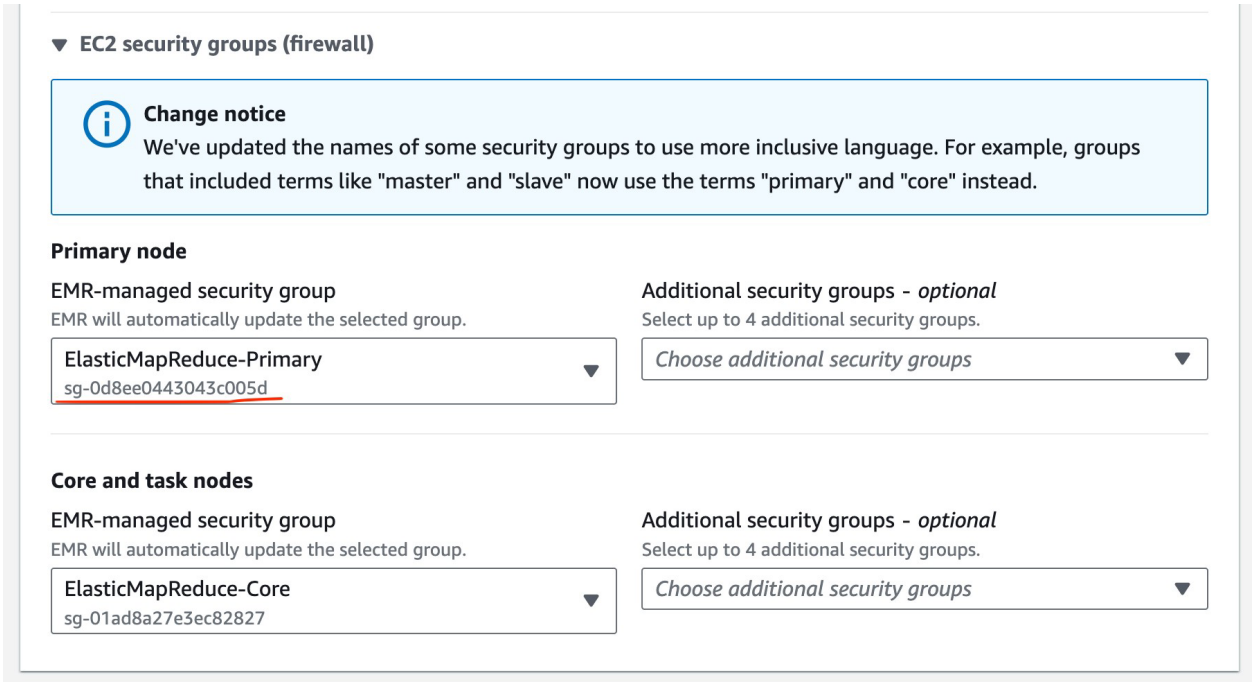




6. Dejar estas opciones por defecto

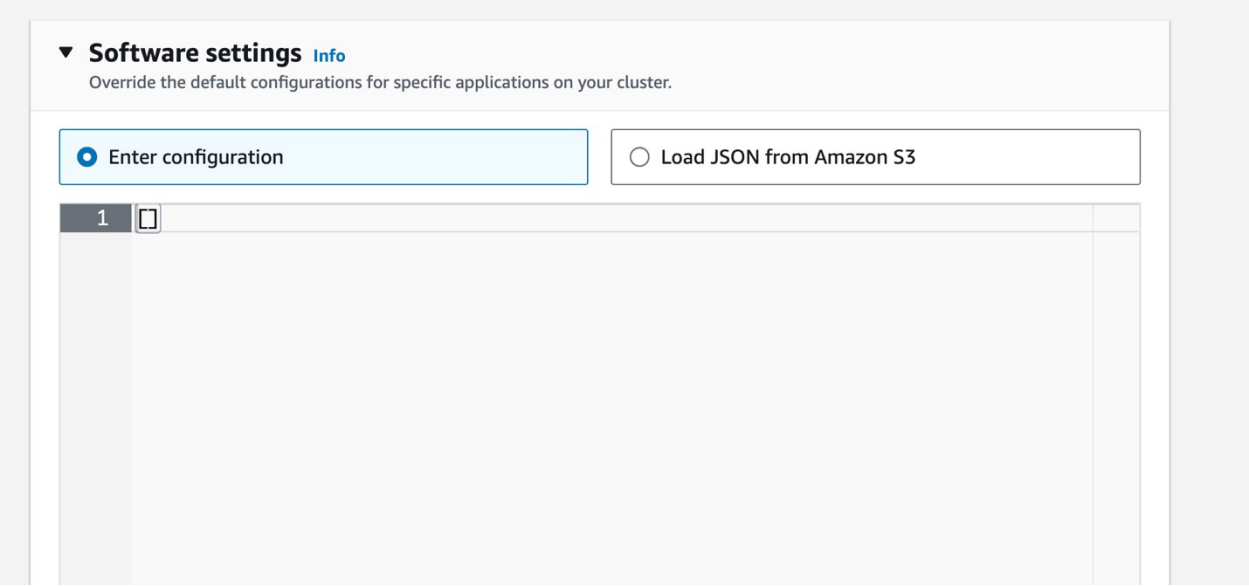


Tener en cuenta el EC2 security groups (firewall), para más adelante adicionar los diferentes puertos para las aplicaciones:



Dejar las siguientes opciones por defecto hasta: Software settings.

7. Configurar software settings



Acá va a configurar el bucket para guardar los notebooks jupyter y no se pierdan cuando se borre el clúster EMR.

Realizar una búsqueda sencilla Google: aws emr jupyterhub s3

Nos conduce al enlace: https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-jupyterhub-s3.html

Configurar con tu propio bucket (crear un bucket para esto)

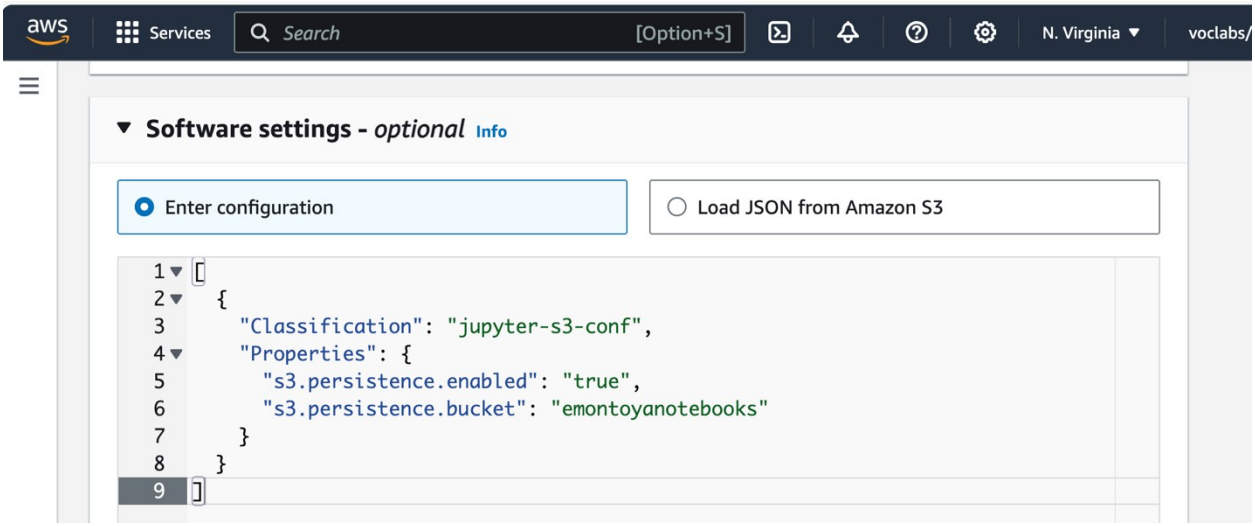
Antes:

```
[
  {
    "Classification": "jupyter-s3-conf",
"Properties": {
      "s3.persistence.enabled": "true",
      "s3.persistence.bucket": "MyJupyterBackups"
    }
  }
]
```

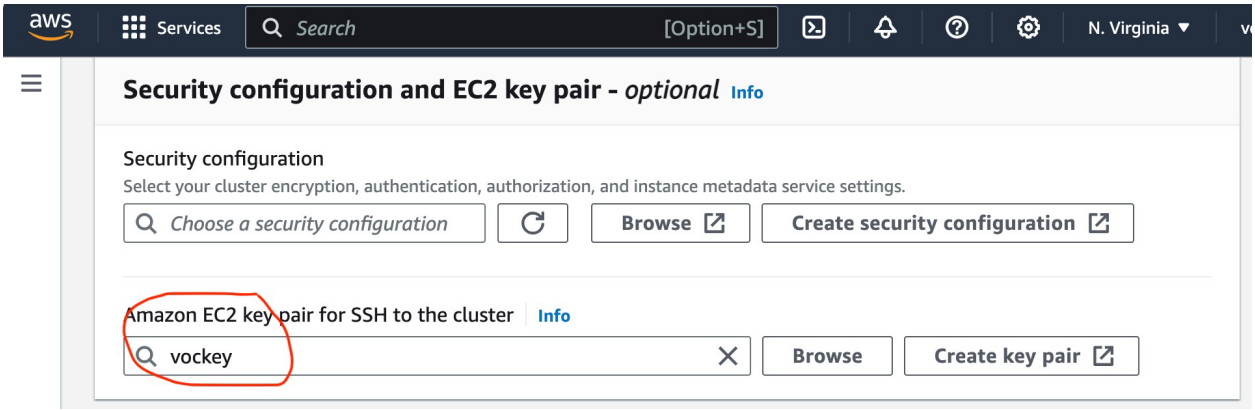
Con mi bucket:

```
[
  {
    "Classification": "jupyter-s3-conf",
    "Properties": {
      "s3.persistence.enabled": "true",
      "s3.persistence.bucket": "su-bucket"
    }
  }
]
```

Y pegue esta configuración en Software Settings así:



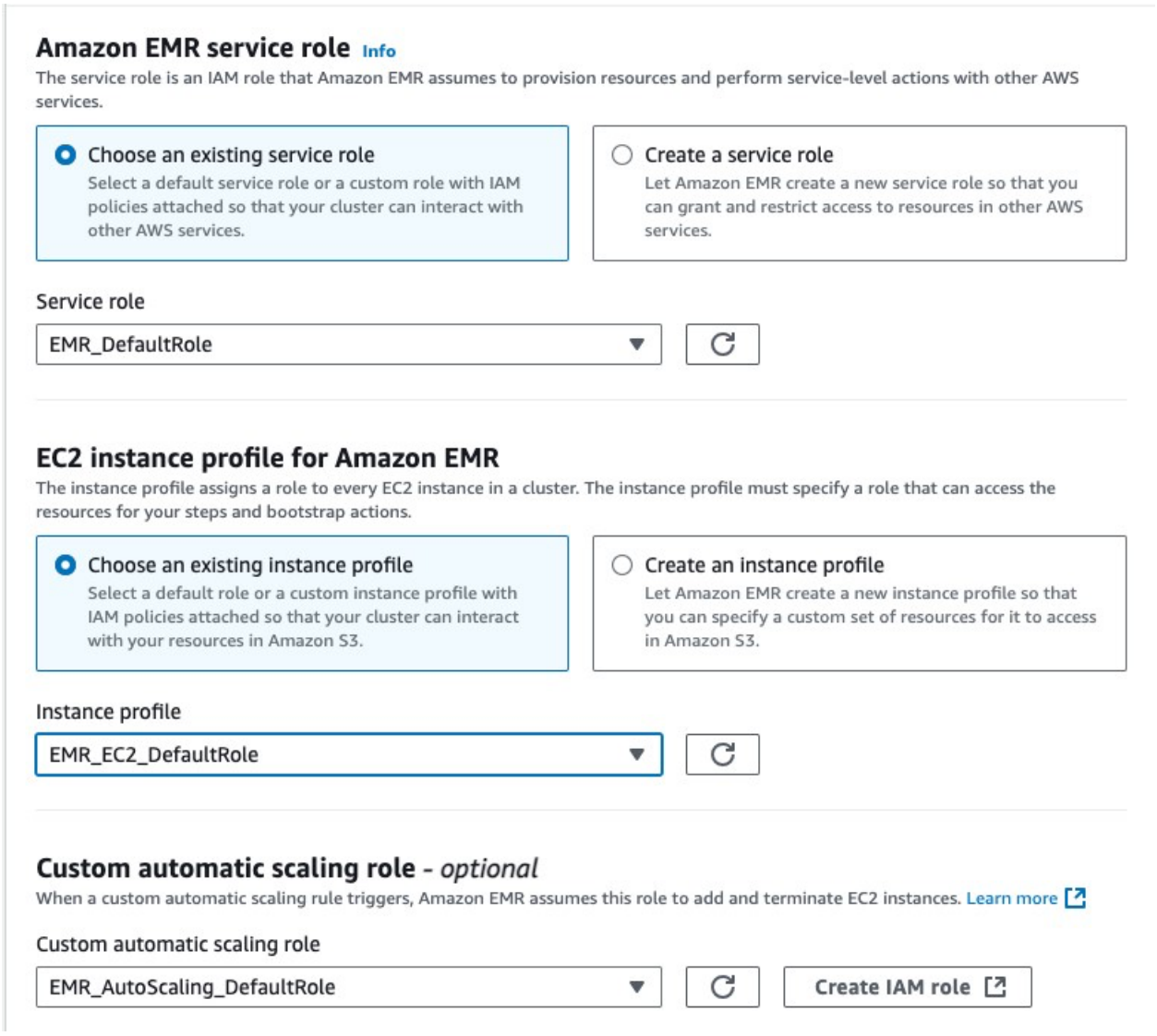
8. Security configuration and EC2 key pair



9. IAM roles

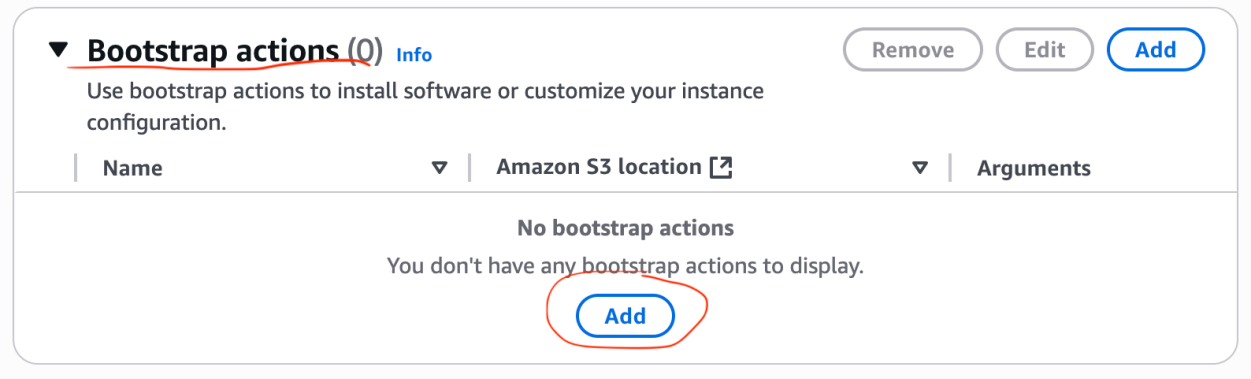
Debe seleccionar:

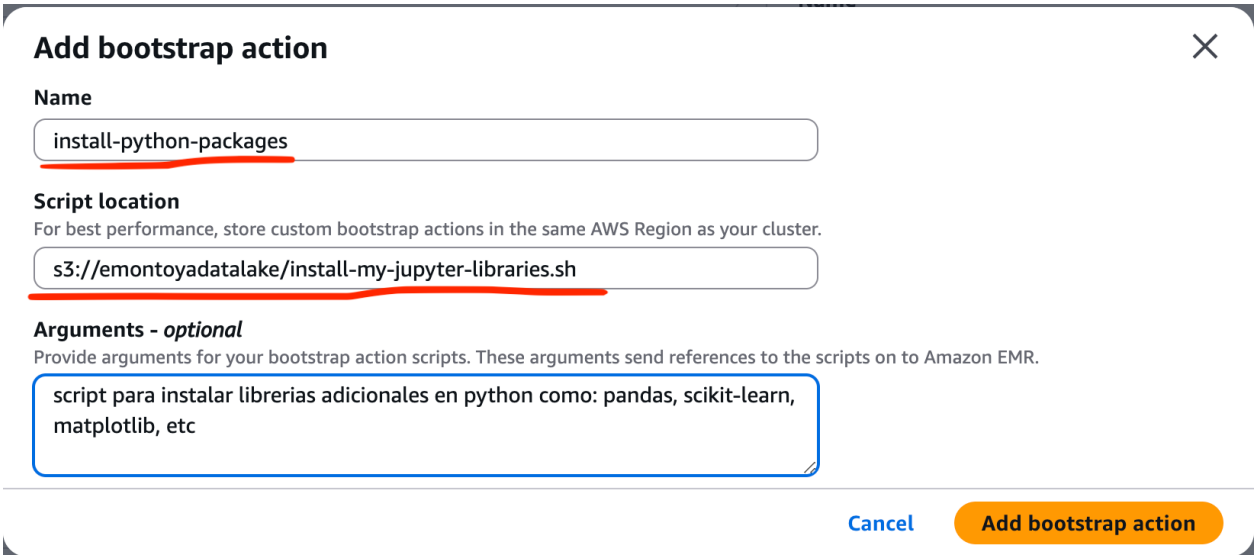
Service role: EMR_DefaultRole
Instance profile: EMR_EC2_DefaultRole
Custom automatic scaling role: EMR_AutoScaling_DefaultRole



10. Si desea instalar librerias adicionales en python para ejecutarse en los notebooks de jupyterhub, debe adicionar este script.sh en la sección de:

Bootstrap Actions(0)



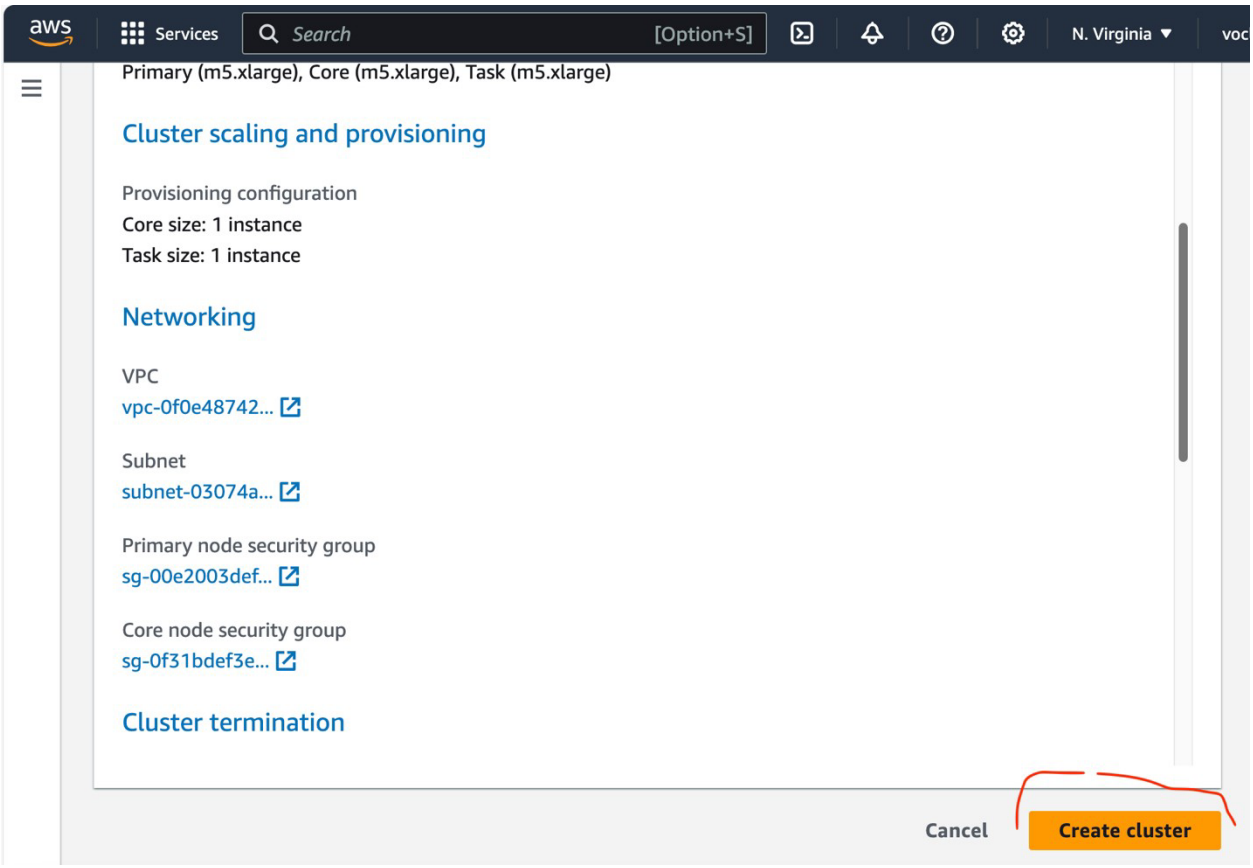


El contenido del archivo: `install-my-jupyter-libraries.sh`

```
#!/bin/bash
sudo python3 -m pip install boto3 nltk scipy scikit-learn pandas
```

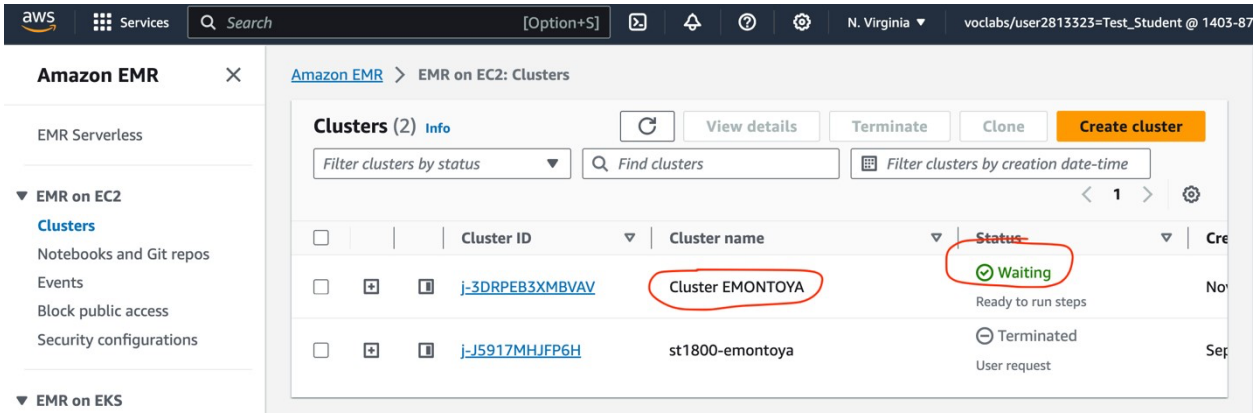
Asegurese antes de crear o clonar el cluster EMR, tener copiado el archivo: install-my-jupyter-libraries.sh en el bucket S3 requerido (en el github de la materia se encuentra un ejemplo de este archivo)

11. Finalmente, a crear el clúster



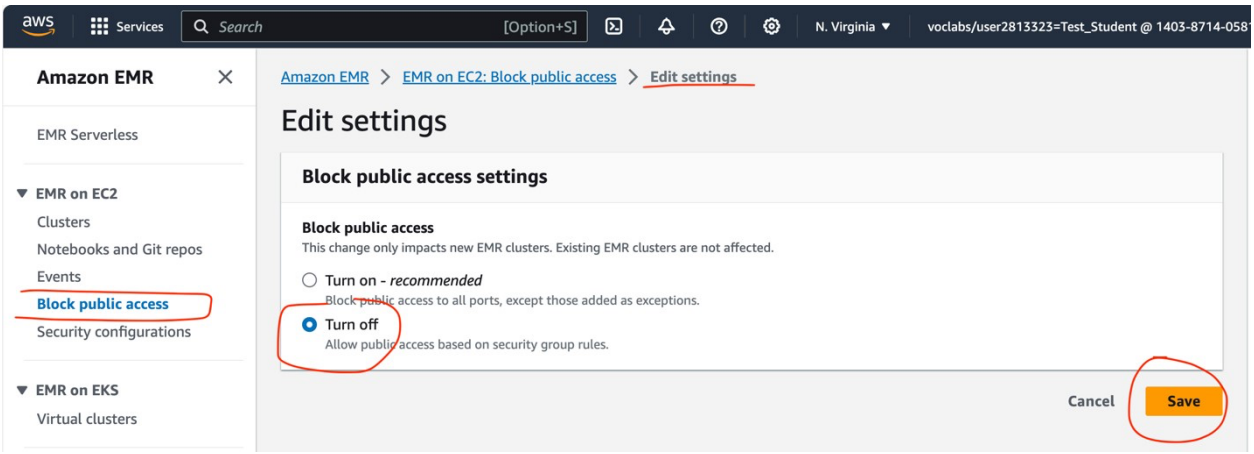
Este proceso demora aproximadamente 20 minutos, tenga paciencia.

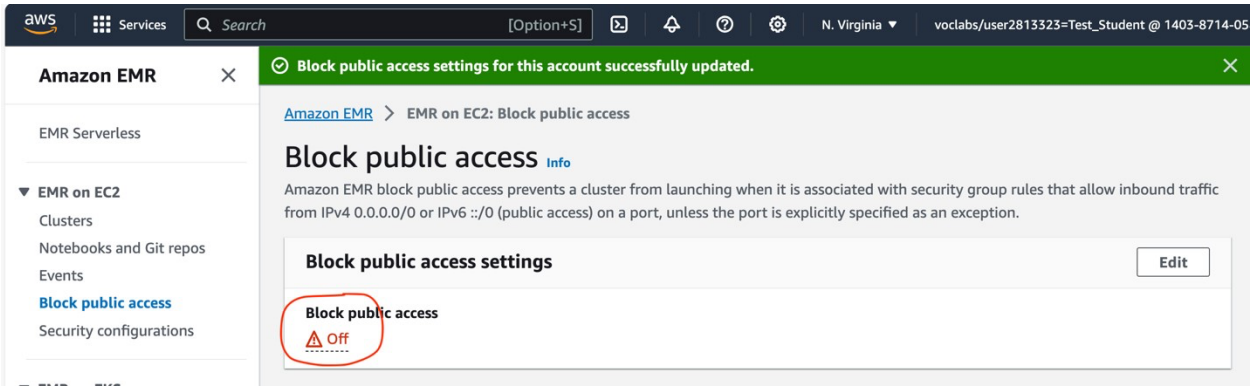
Debe salir con este mensaje de clúster exitosamente creado:



12. Debe abrir todos los puertos TCP para acceso al clúster así

(nota: esto solo se hace una vez, cada vez que crea, destruya o clone un clúster, ya quedan abiertos)



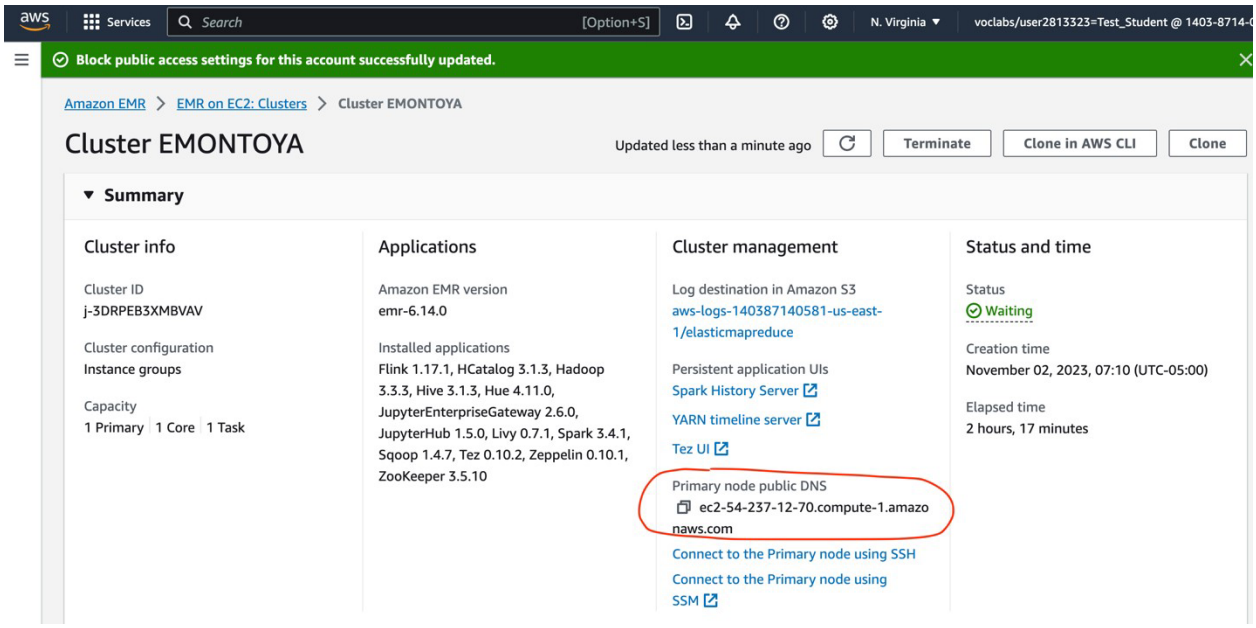


También debe abrir los puertos de las aplicaciones de hadoop/Spark en el Security Group del nodo MASTER del clúster.

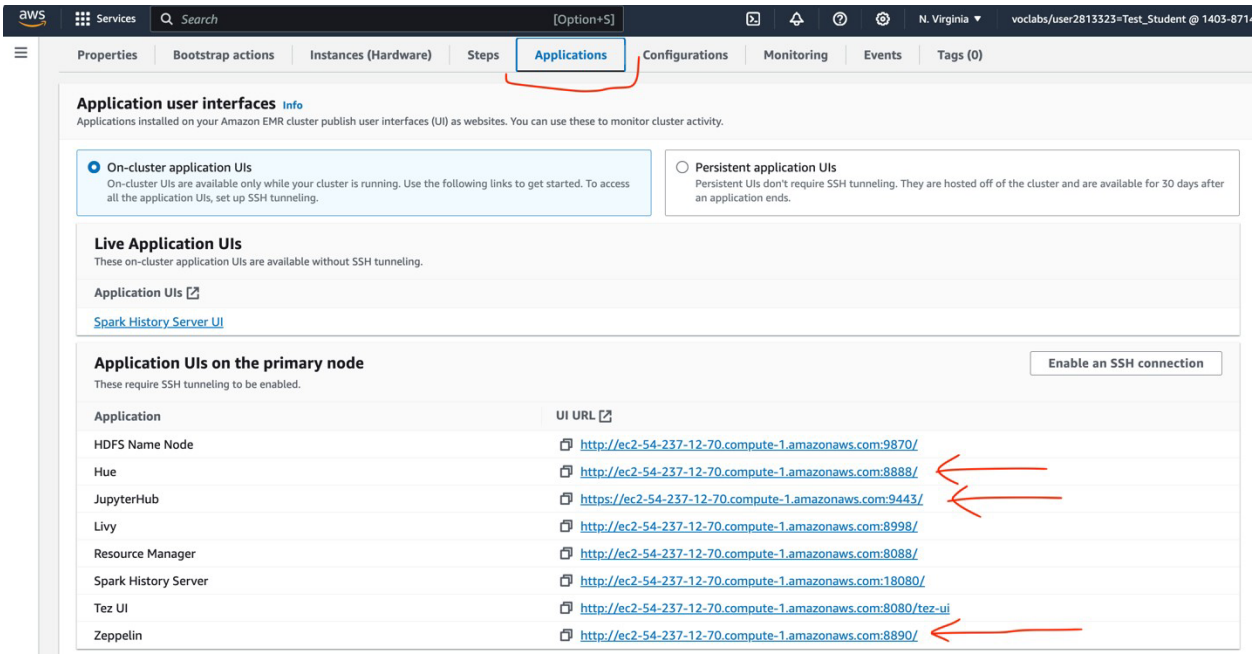
(nota: esto solo se hace una vez, cada vez que crea, destruya o clone un clúster, ya quedan abiertos)

Donde ubico el nodo master?

Dar Click en el clúster que acabas de crear, mirar la IP y nombre de la máquina EC2:



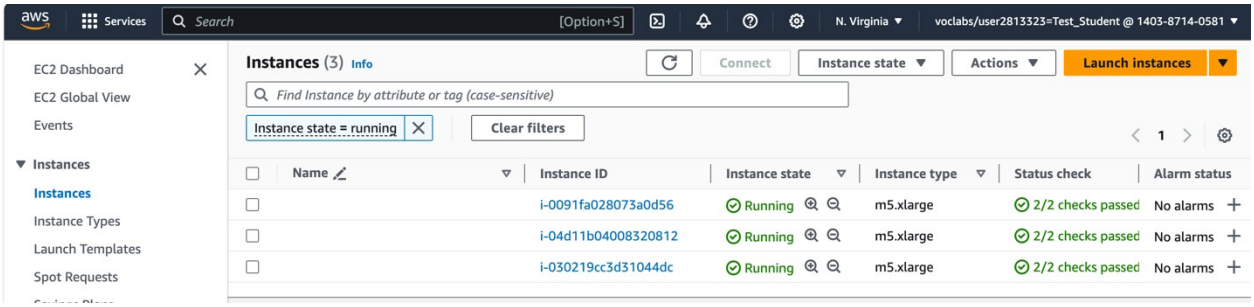
Luego entras al servicio EC2 de dicha máquina Master, y va a modificar el Security Group para agregar los siguientes puertos de las aplicaciones:

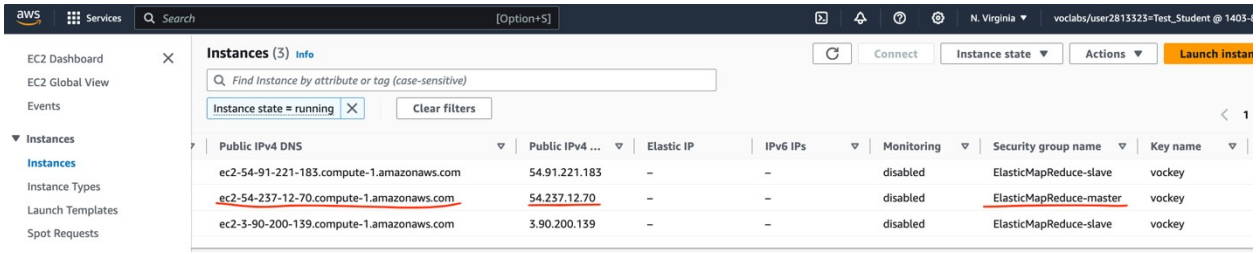


Además, abrir los puertos TCP:

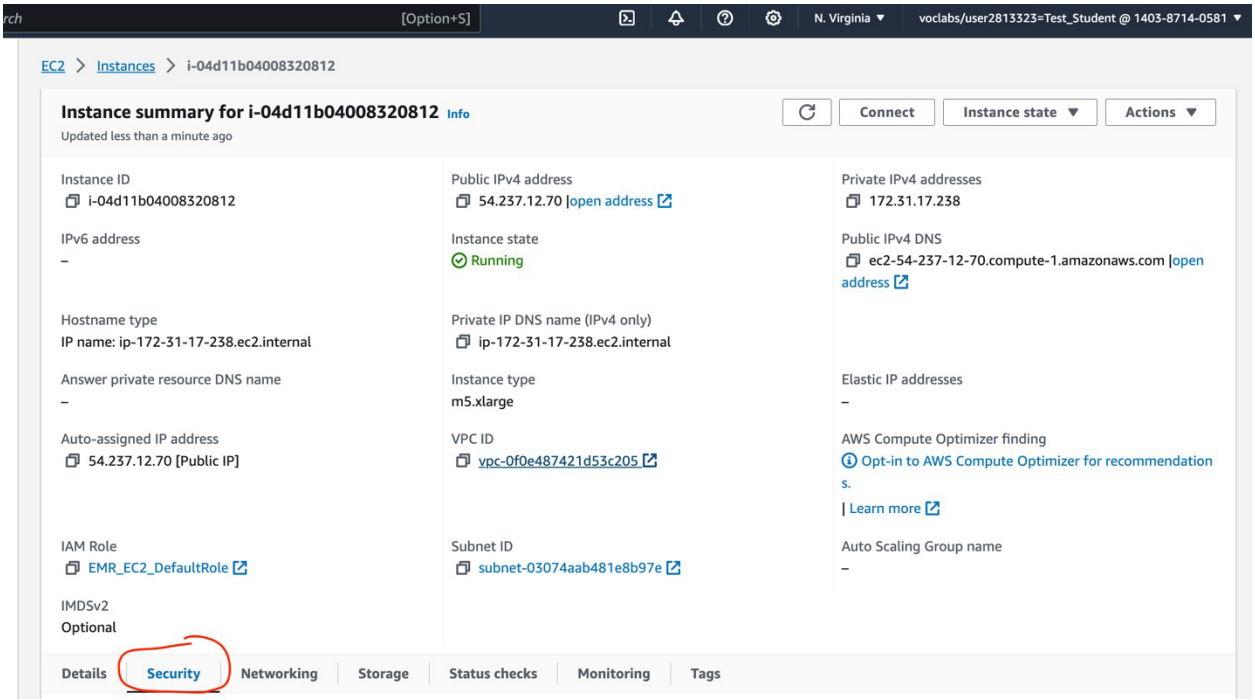
22
14000
9870

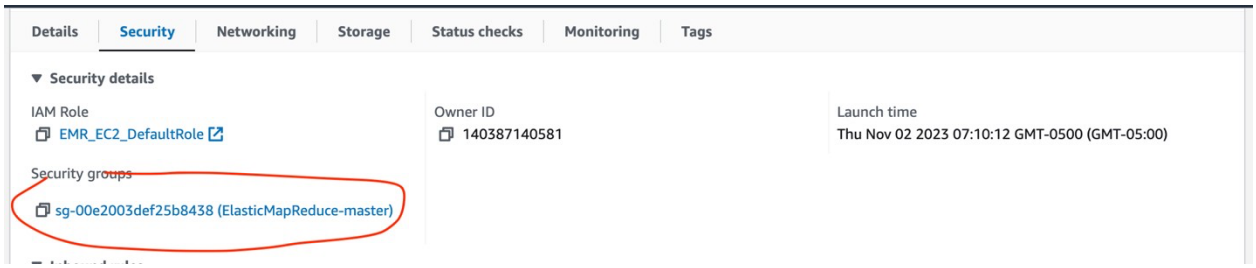
En AWS EC2, les debería mostrar 3 máquinas:

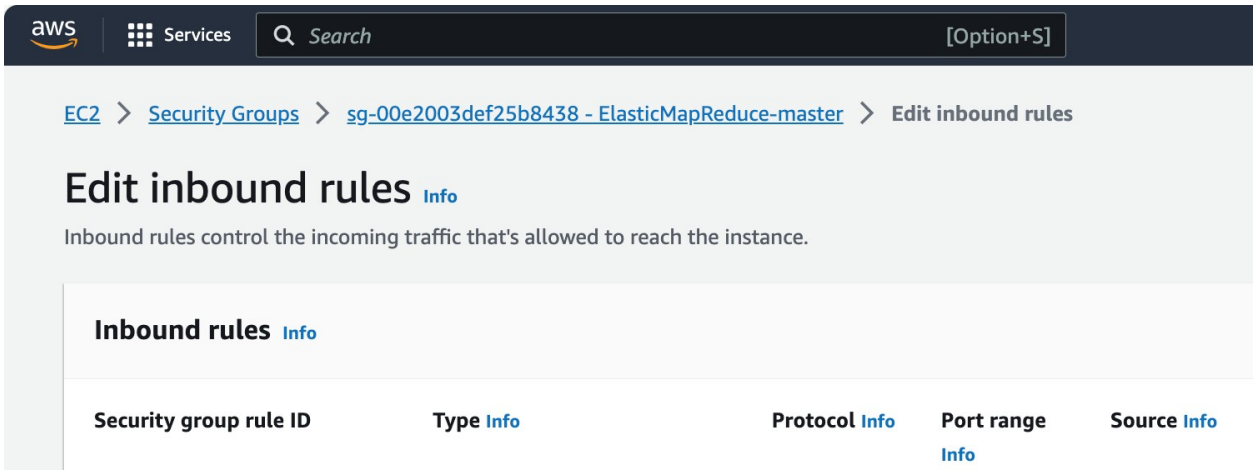




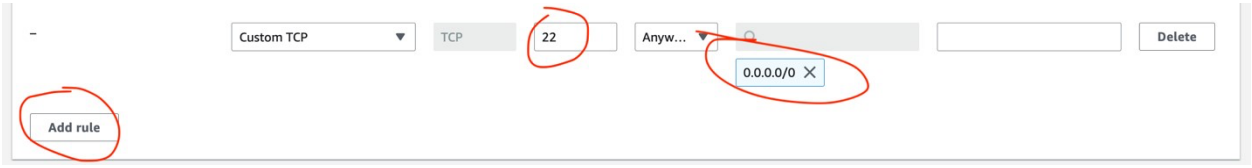
Entrar a la pestaña de seguridad de la Instancia EC2 del nodo master:







Uno a uno, va adicionando los puertos, aca se adiciono el puerto 22, haga lo mismo para los demás puertos:



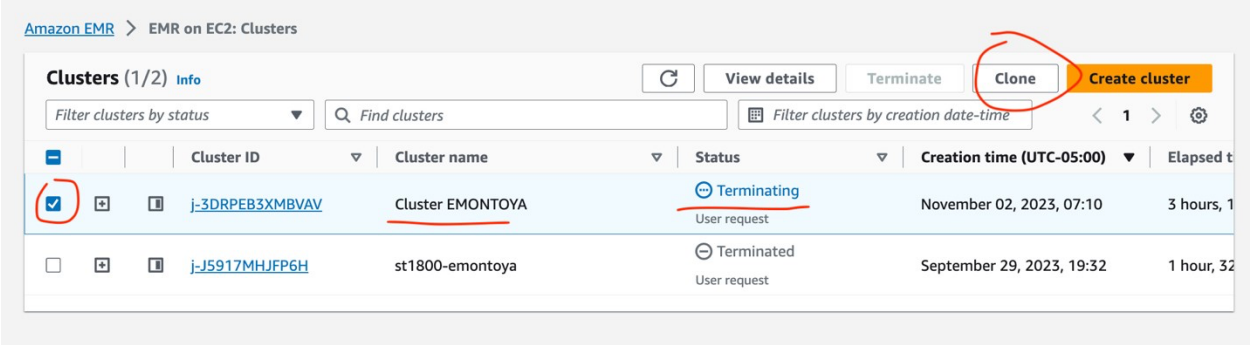
Parte 2: Borrar y recrear clúster

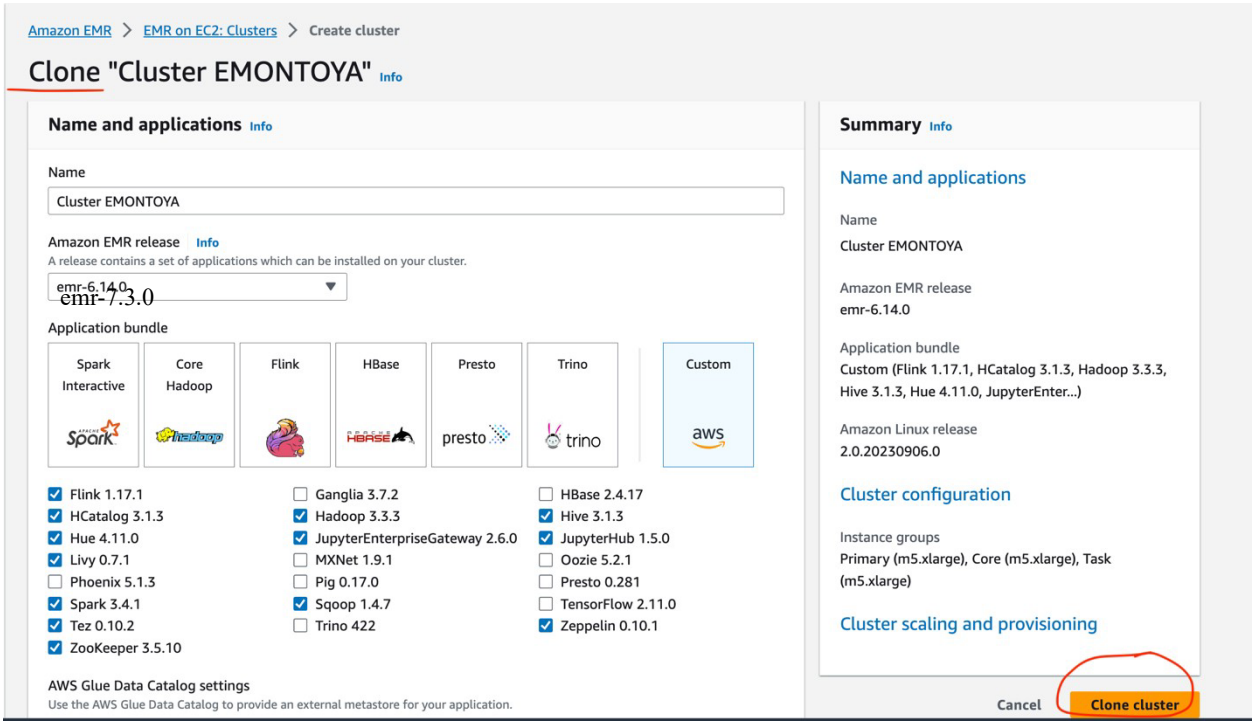
Los clúster EMR en amazon, son temporales.

Los clúster EMR no se pueden pausar

Cada que no requiera trabajar más con un clúster, DEBE BORRARLO:

Pero la próxima vez que lo requiera, puede Clonar y crear nuevamente un clúster, teniendo en cuenta la configuración de otro clúster previamente creado, esta es la opción que debe utilizar.



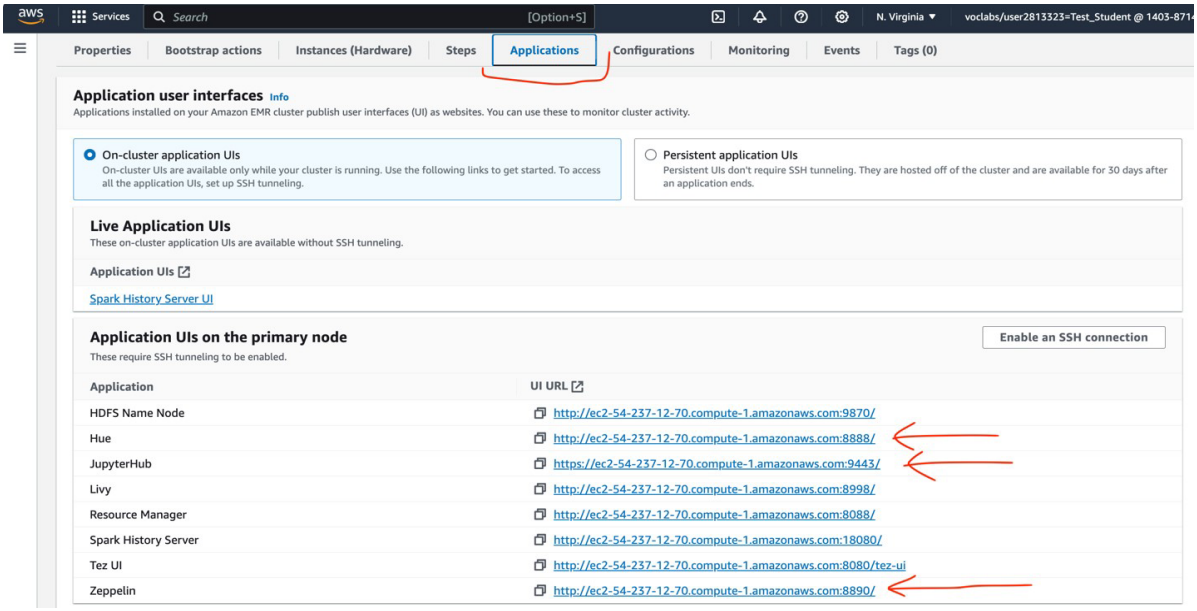


Cada vez que lo Clone, debe crear nuevamente el usuario hadoop / con su password de preferencia, así como realizar el arreglo del archivo hue.ini para cambiar el puerto 14000 a 9870 (esto lo entenderá más adelante)

Parte 3: Ingresar al clúster EMR por Hue

Utilice la aplicación hue, por el puerto 8888 desde un browser a la ip o nombre del nodo master.

Fijarse en la aplicación del clúster HUE:



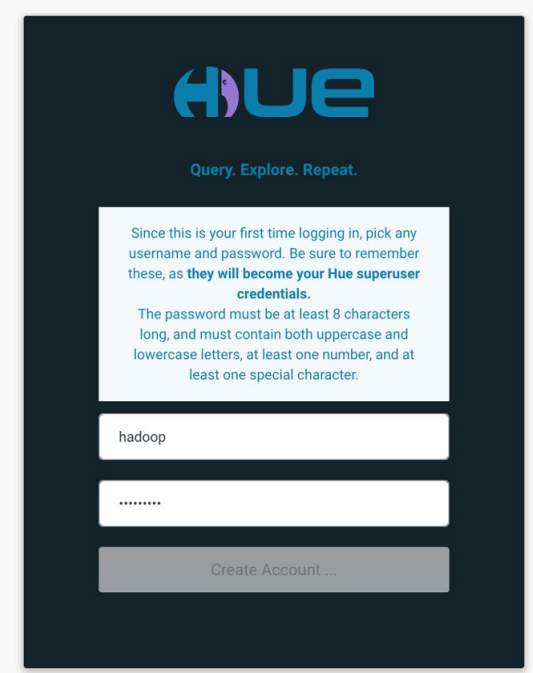
Y darle click a la URL de HUE, en este ejemplo:

http://ec2-54-237-12-70.compute-1.amazonaws.com:8888

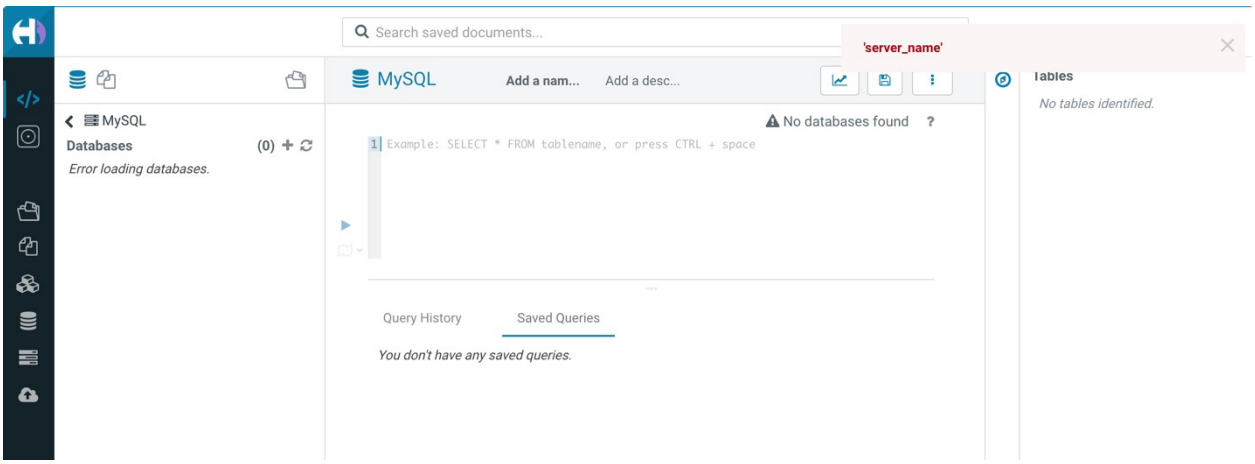
La primera vez, me pide crear un usuario y clave:

Username: hadoop
Passoword: <<el que quiera>>

Nota: el usuario tiene que ser ‘hadoop`

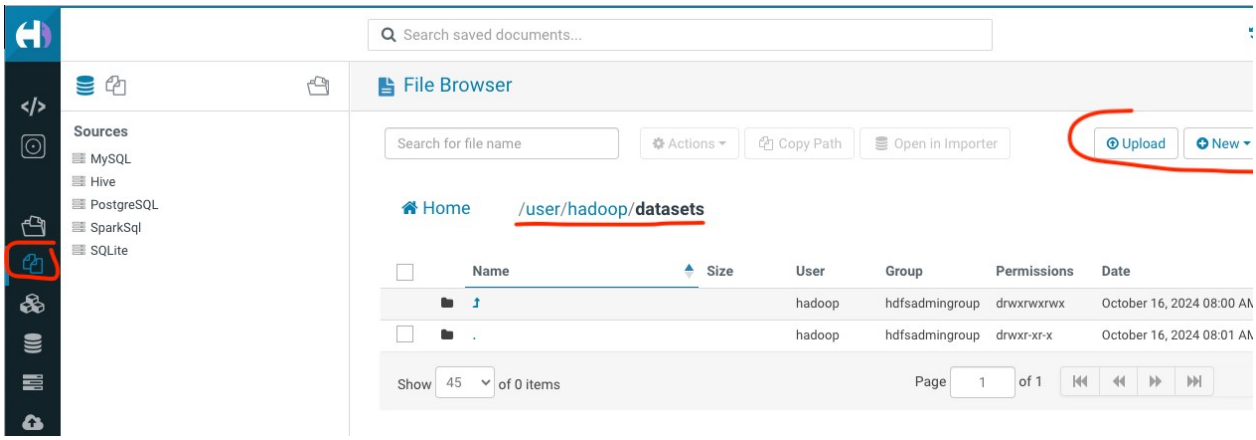


Deberá salir una interfaz así:



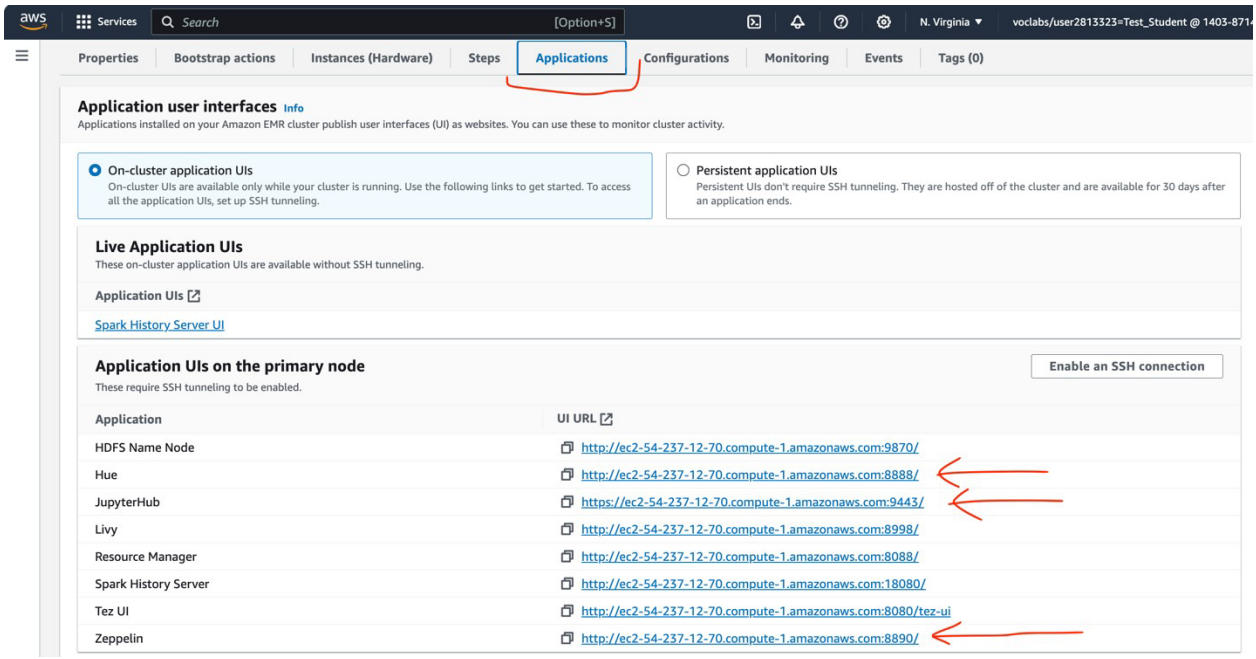
Podrá acceder los servicios Hive, Spark, S3, y HDFS.

Ya va a poder gestionar archivos sin problema por hue para HDFS

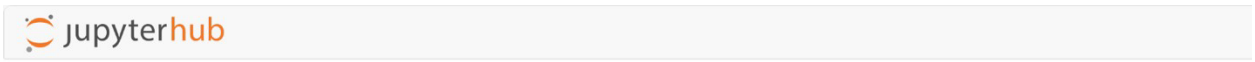


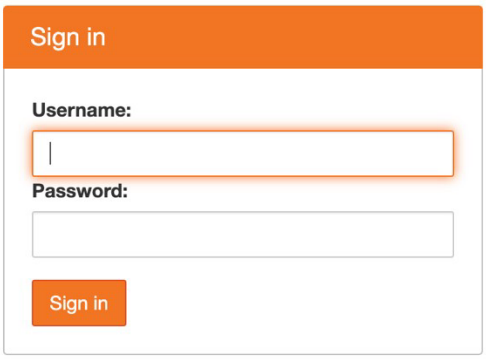
Parte 4: entrar a jupyter hub

Utilice la aplicación jupyterhub de:



Para este caso la URL es: https://ec2-54-237-12-70.compute-1.amazonaws.com:9443/





Utilice el usuario por defecto:

Username: jovyan
Password: jupyter

Tomado de:
https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-jupyterhub-user-access.html

Y listo, ya puede realizar notebooks pyspark, verifique que las 2 variables más importantes de contexto de spark esta activas en un notebook así: (primero debe crear un notebook pyspark)