PR_07.4 Dani Gayol Rodríguez

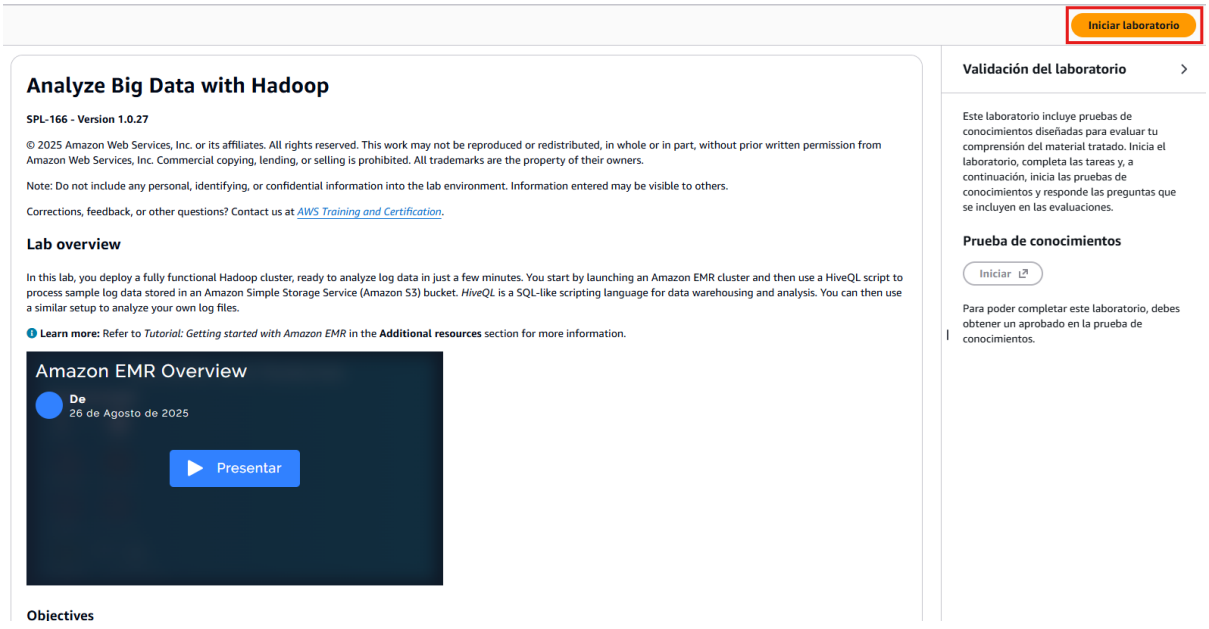
PR_07.4 Dani Gayol Rodríguez 1
Apartado A 1
Analyze Big Data with Hadoop 2
Tarea 1 3
Tarea 2 5
Tarea 3 13
Tarea 4 14
Tarea 5 15
Tarea 6 17
Apartado B 18
Exploring Google Ngrams with Amazon EMR and Hive 19
Tarea 1 19
Tarea 2 23
Tarea 3 24
1.) ¿Qué contiene el bucket s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/? ¿Cuánto ocupa el archivo que contiene? 29
2.) ¿Cuántos registros contiene la tabla ngrams que creaste en HIVE? ¿Desde qué año hasta qué año abarca la información que contiene? 30
3.) En la creación de la tabla normalized ¿qué significa la expresión REGEXP "^[A-Za-z+\'-]{3,}$"? ¿Cuántos registros contiene la tabla normalized? 31

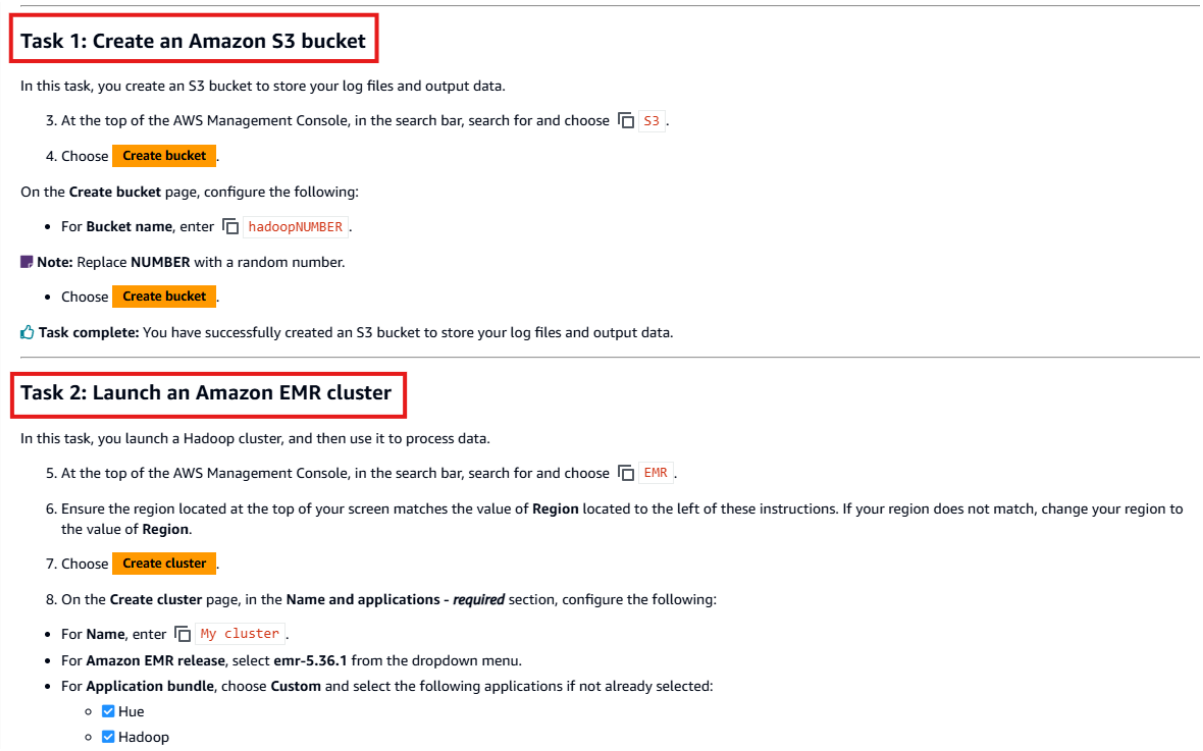
Apartado A

Analyze Big Data with Hadoop

Para comenzar con la práctica, tenemos que darle al botón de “start” en la pantalla principal

Una vez dentro, tenemos el botón de “iniciar el laboratorio” y si bajamos más abajo, nos aparecen las instrucciones a seguir para realizar el ejercicio





Task 1: Create an Amazon S3 bucket

In this task, you create an S3 bucket to store your log files and output data.

3. At the top of the AWS Management Console, in the search bar, search for and choose S3.
4. Choose **Create bucket**.

On the **Create bucket** page, configure the following:

- For **Bucket name**, enter hadoopNUMBER.

Note: Replace **NUMBER** with a random number.

- Choose **Create bucket**.

Task complete: You have successfully created an S3 bucket to store your log files and output data.

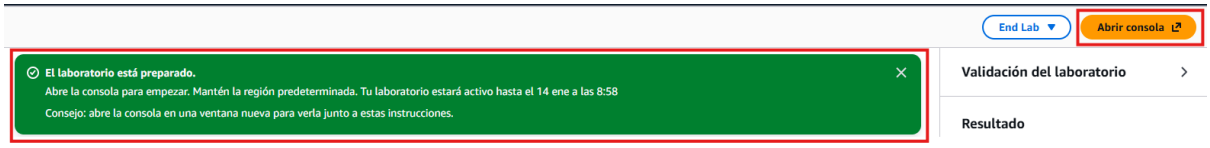
Task 2: Launch an Amazon EMR cluster

In this task, you launch a Hadoop cluster, and then use it to process data.

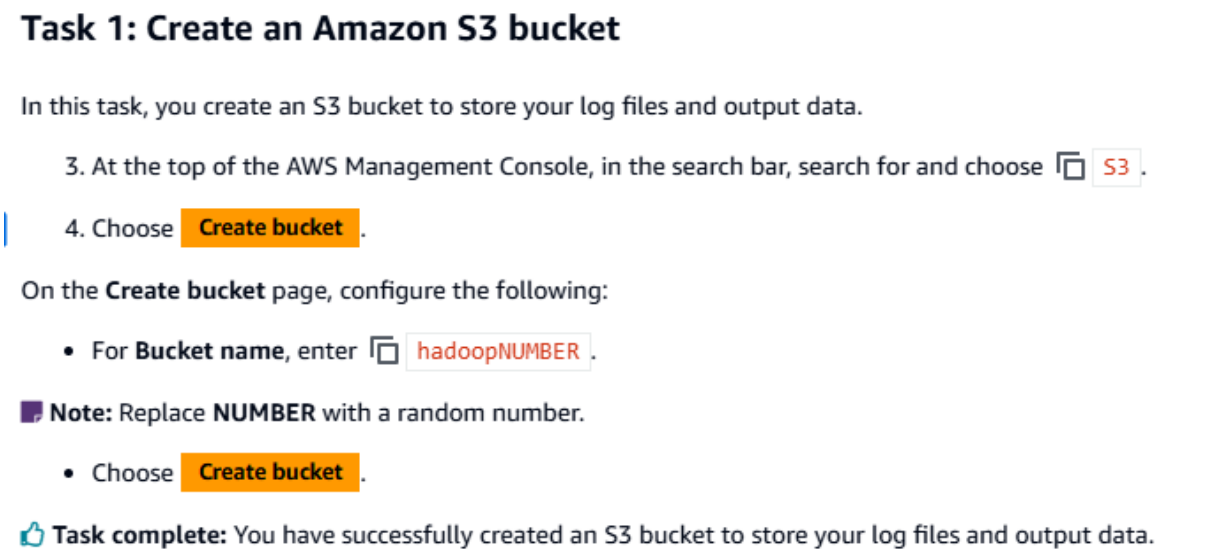
5. At the top of the AWS Management Console, in the search bar, search for and choose EMR.
6. Ensure the region located at the top of your screen matches the value of **Region** located to the left of these instructions. If your region does not match, change your region to the value of **Region**.
7. Choose **Create cluster**.
8. On the **Create cluster** page, in the **Name and applications - *required*** section, configure the following:

- For **Name**, enter My cluster.
- For **Amazon EMR release**, select **emr-5.36.1** from the dropdown menu.
- For **Application bundle**, choose **Custom** and select the following applications if not already selected:
  - Hue
  - Hadoop

Una vez pulsado el botón “iniciar laboratorio", nos aparecerá lo siguiente, y nos dará la opción de “abrir consola”



Tarea 1



Task 1: Create an Amazon S3 bucket

In this task, you create an S3 bucket to store your log files and output data.

3. At the top of the AWS Management Console, in the search bar, search for and choose S3.
4. Choose **Create bucket**.

On the **Create bucket** page, configure the following:

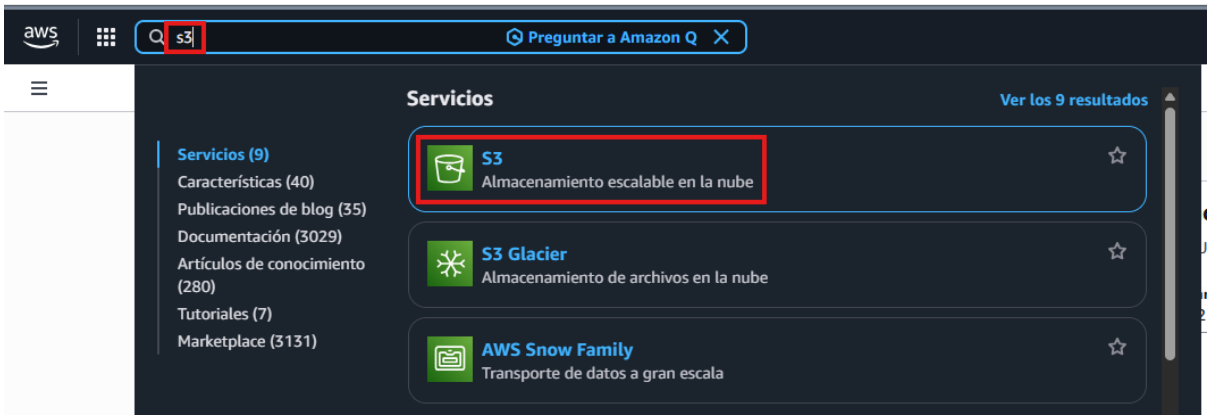
- For **Bucket name**, enter hadoopNUMBER.

Note: Replace **NUMBER** with a random number.

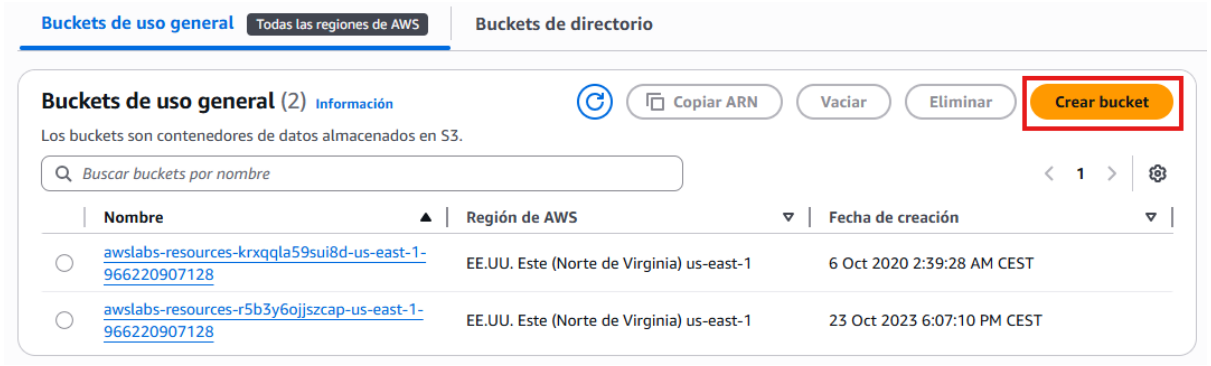
- Choose **Create bucket**.

Task complete: You have successfully created an S3 bucket to store your log files and output data.

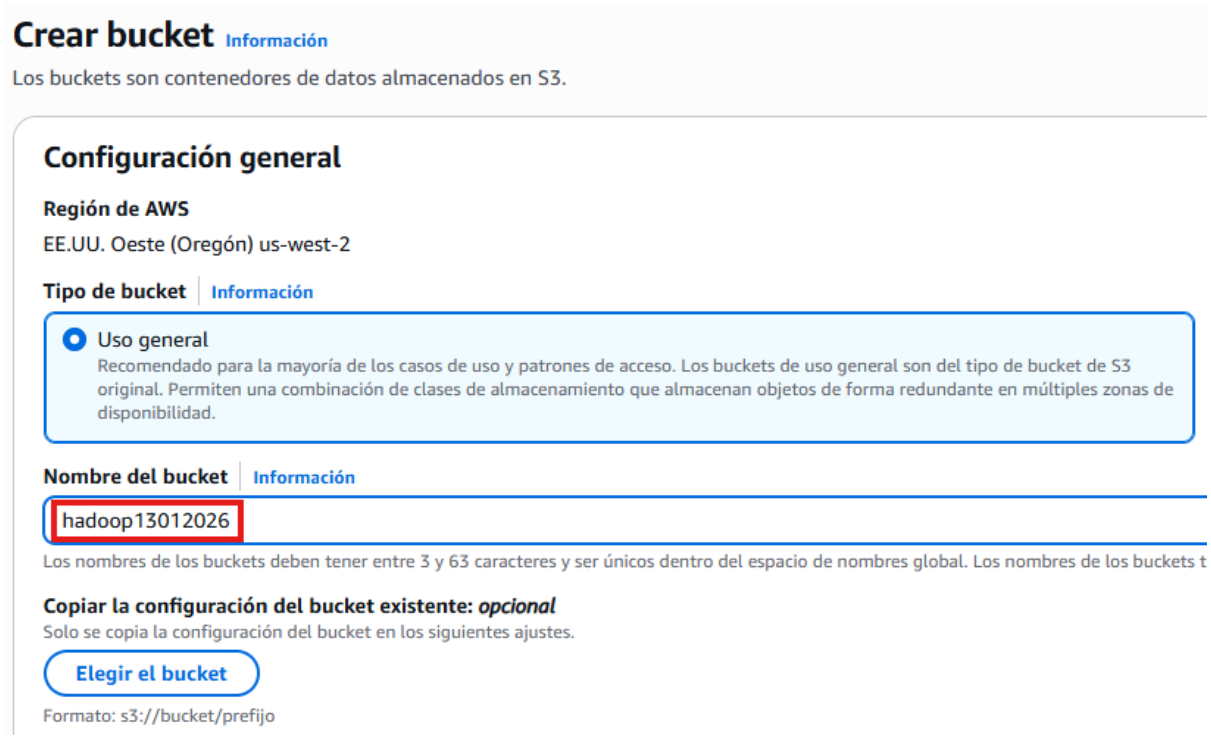
El primero paso es buscar “S3” en la búsqueda de AWS



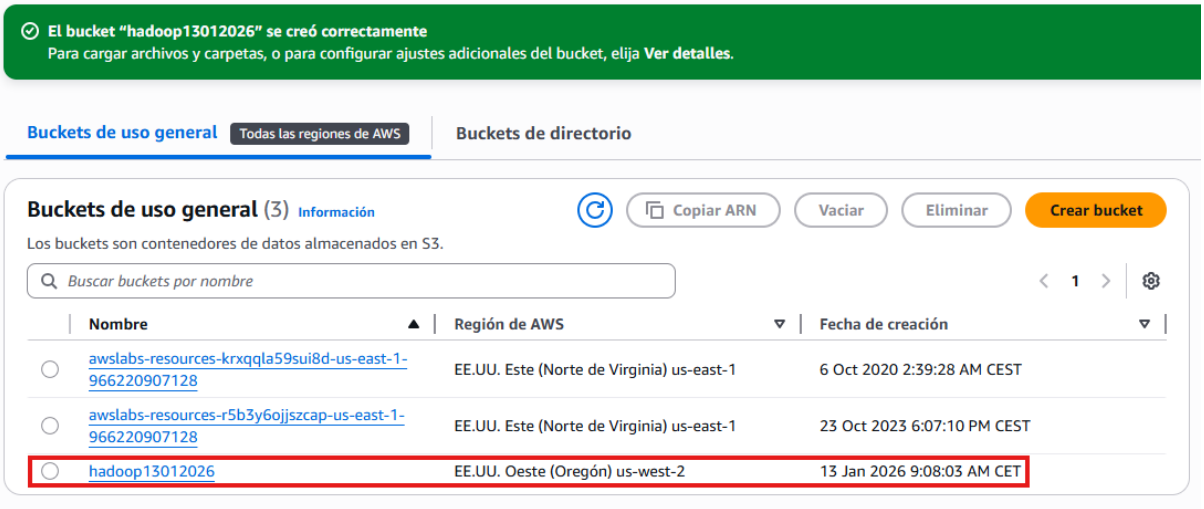
Una vez dentro, le tenemos que dar a "Create Bucket"



Le tenemos que dar un nombre, para ello ponemos "hadoopNUMBER" y el number le voy a poner el de la fecha de hoy, (hadoop13012026)

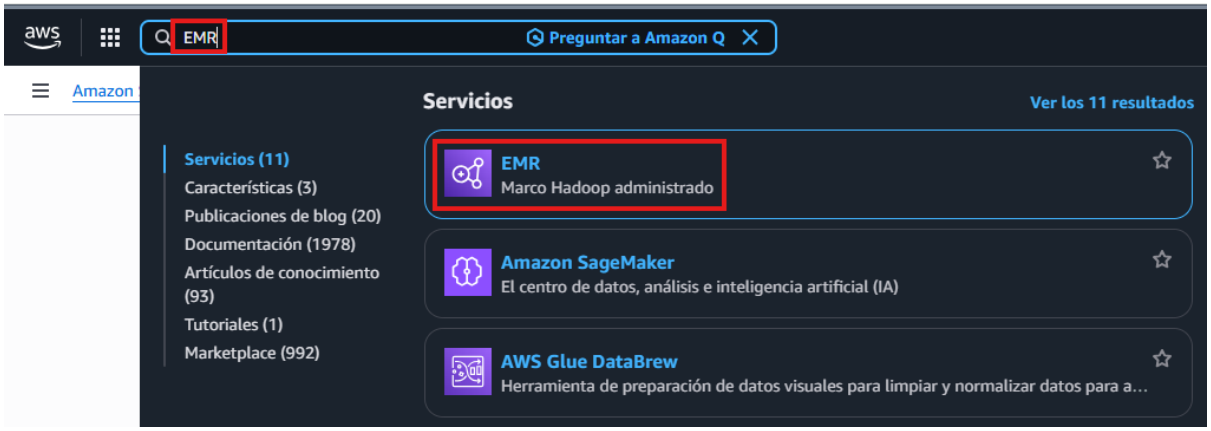


Finalmente, le daremos al botón de "Crear Bucket" abajo del todo y nos saldra lo siguiente al crearse

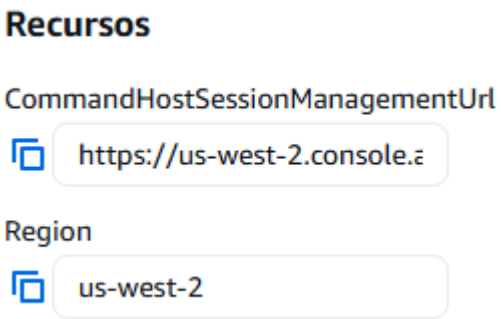


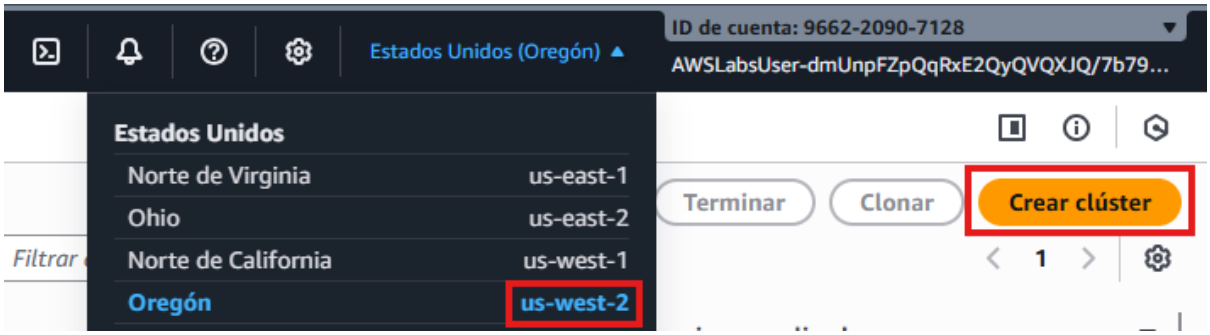
Tarea 2

Ahora tendremos que crear un EMR, por lo tanto, tenemos que buscar "EMR" en la búsqueda de AWS

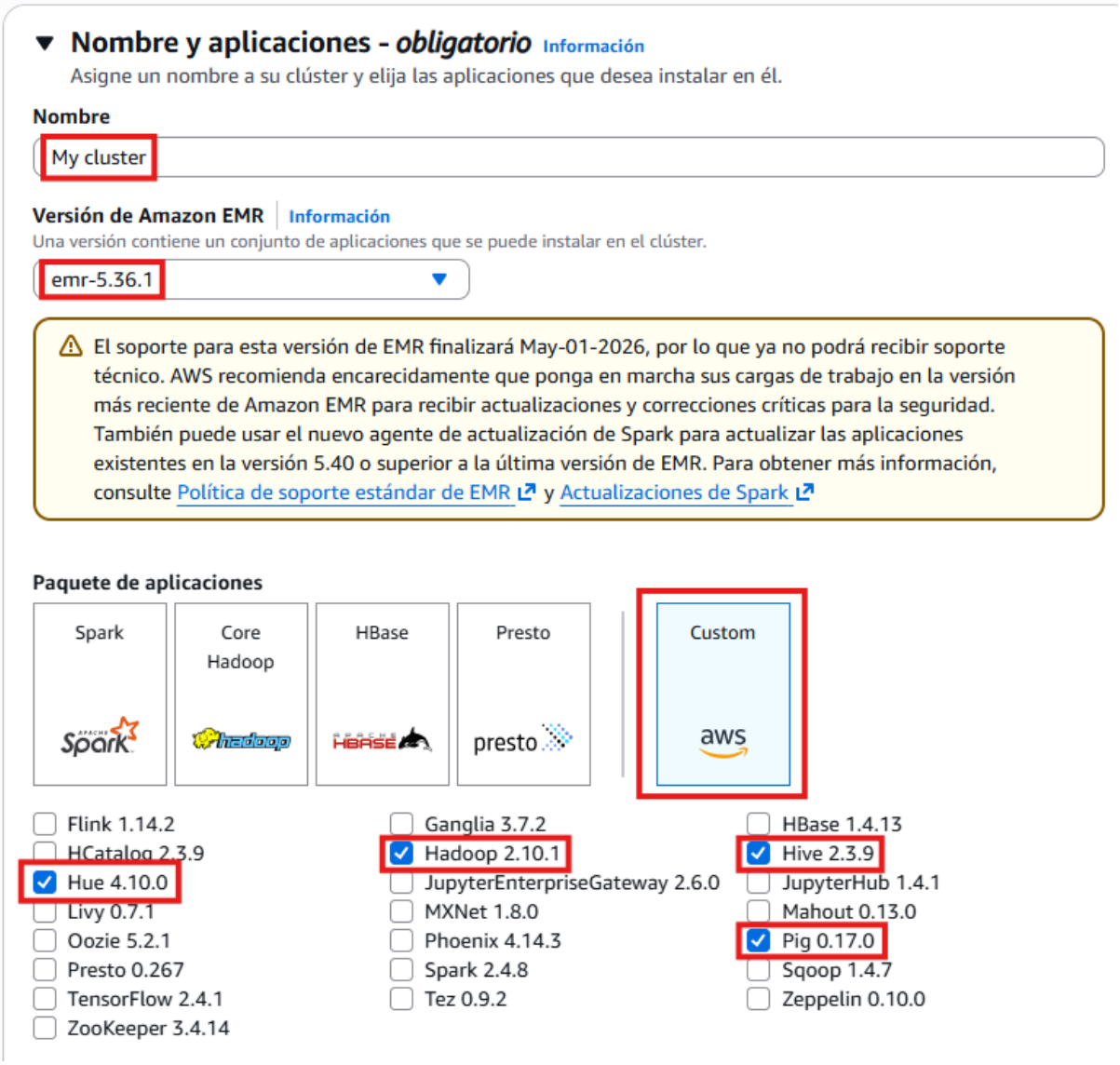


Antes de darle a "Crear Cluster", tenemos que verificar que la región está establecida correctamente

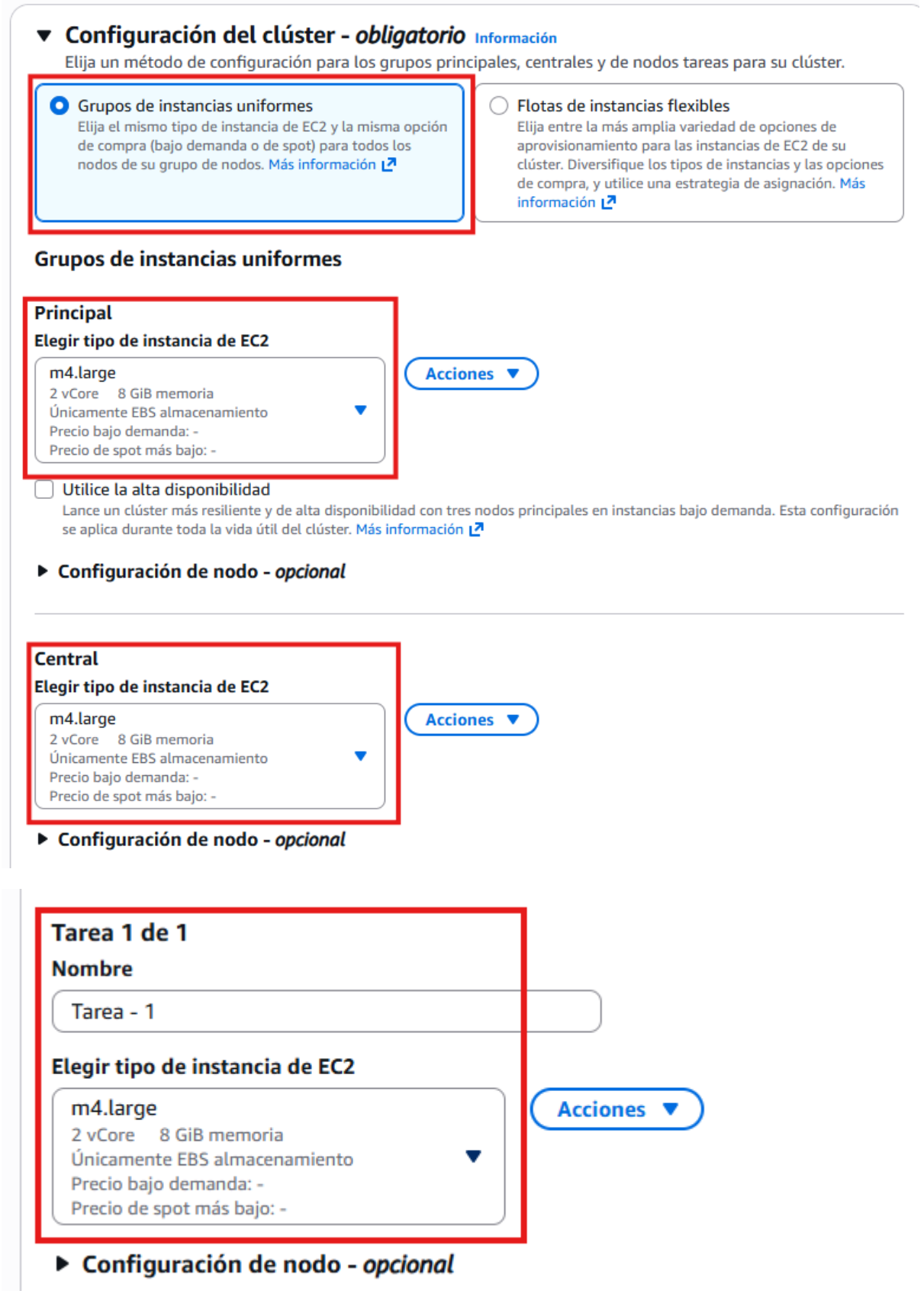




Ahora tenemos que configurar la siguiente sección, "Nombre y aplicaciones"



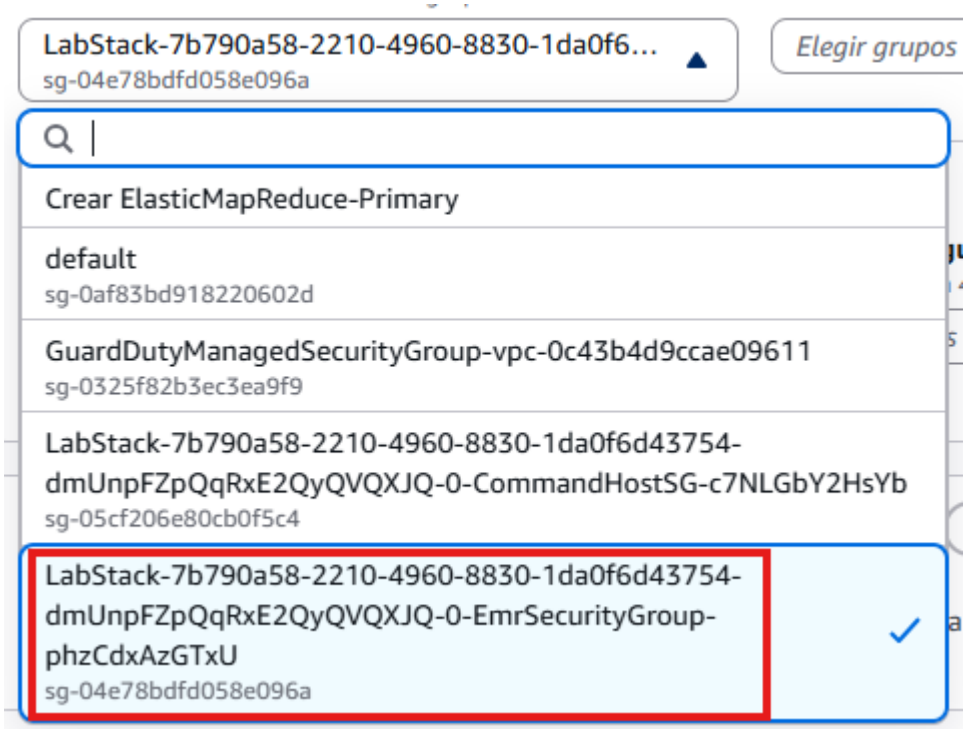
Ahora tenemos que configurar la sección de "Configuración del clúster"



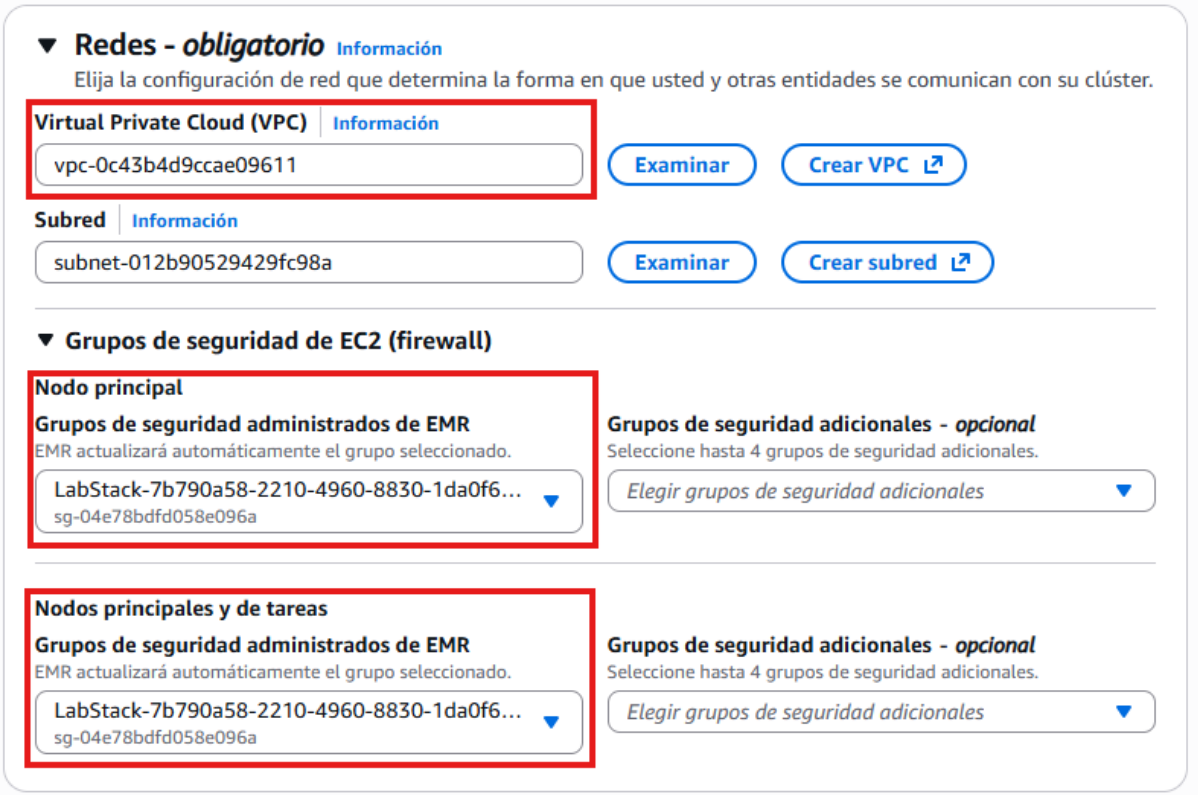
Ahora configuramos la sección de “Redes”

Tenemos que darle al botón de “examinar” al lado de “Virtual Private Cloud (VPC)”

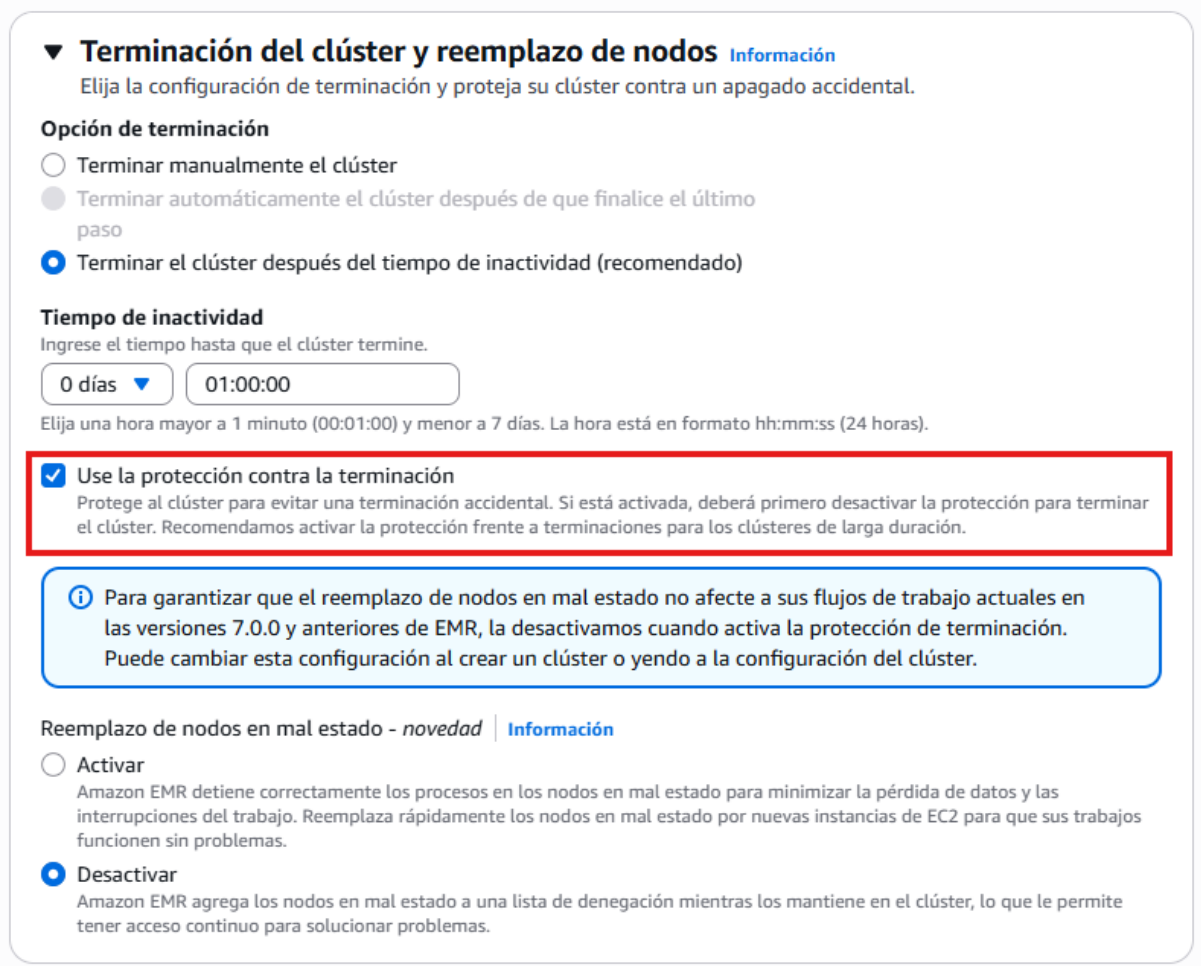
Ahora le damos al desplegable y seleccionamos esta opción



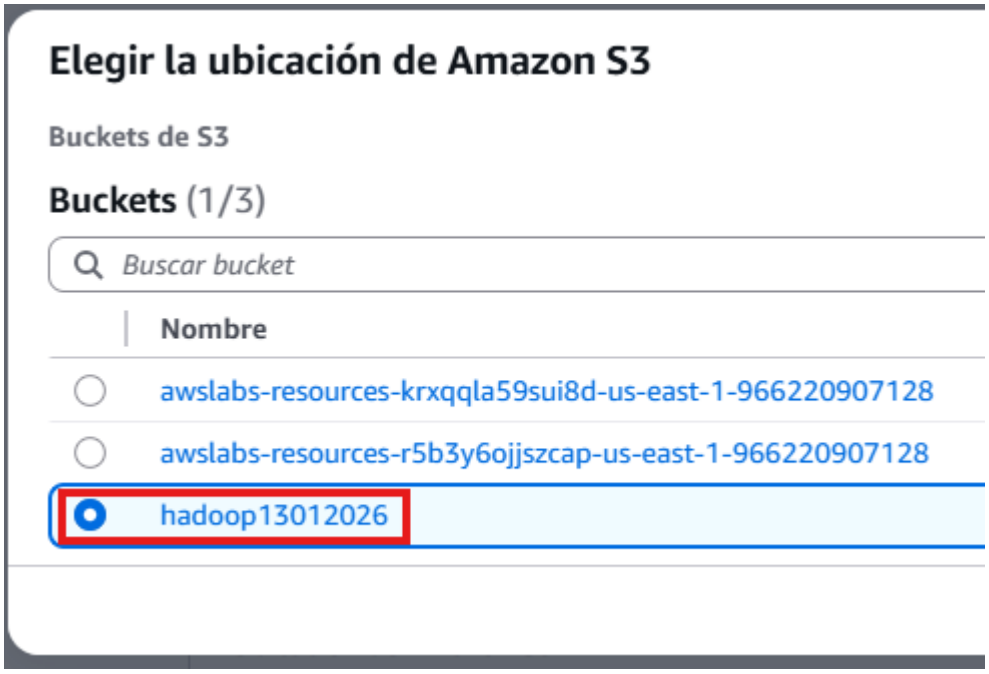
Nos tendría que quedar tal que así



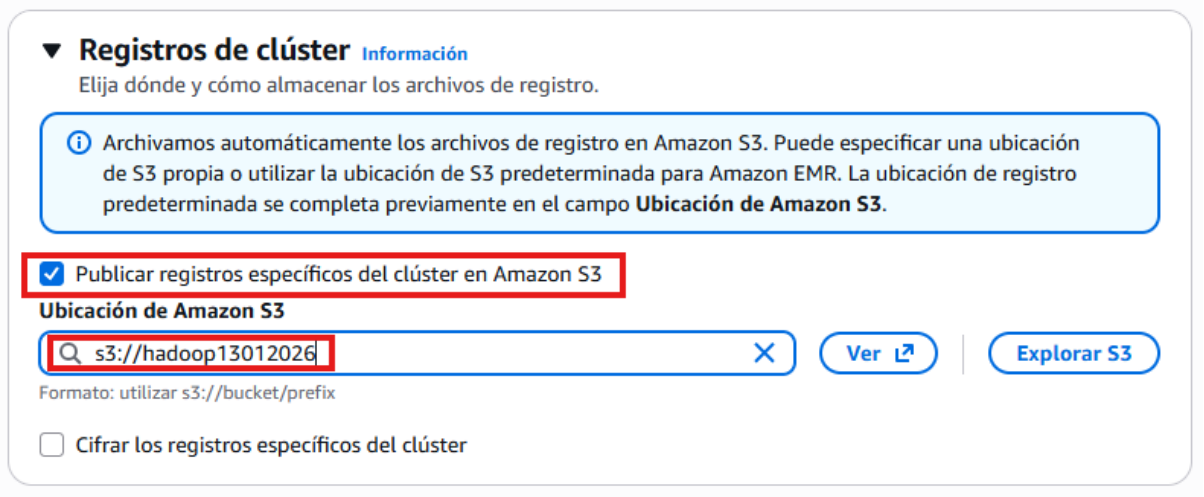
En la sección de “Terminación del clúster y reemplazo de nodos”, tenemos que marcar la siguiente opción



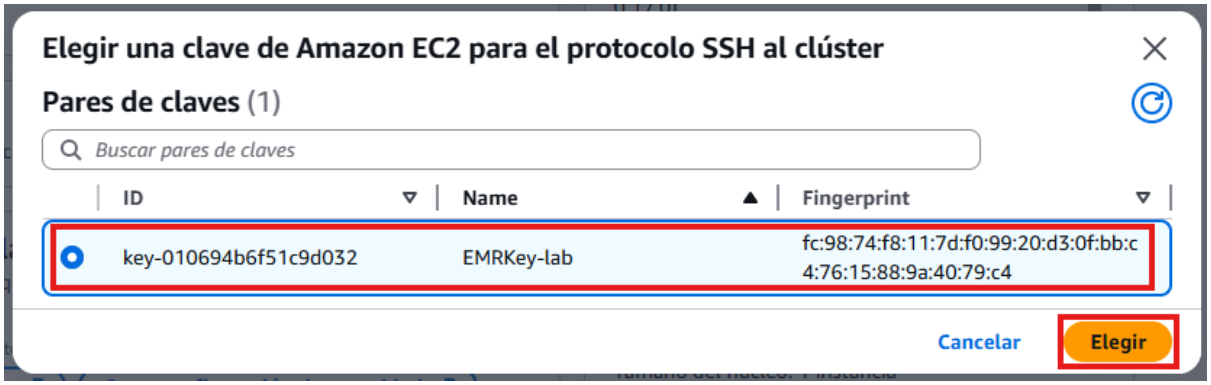
En la sección de “Registros de clúster” hacemos lo siguiente, le darmos al botón de “Explorar S3” y seleccionamos la opción siguiente



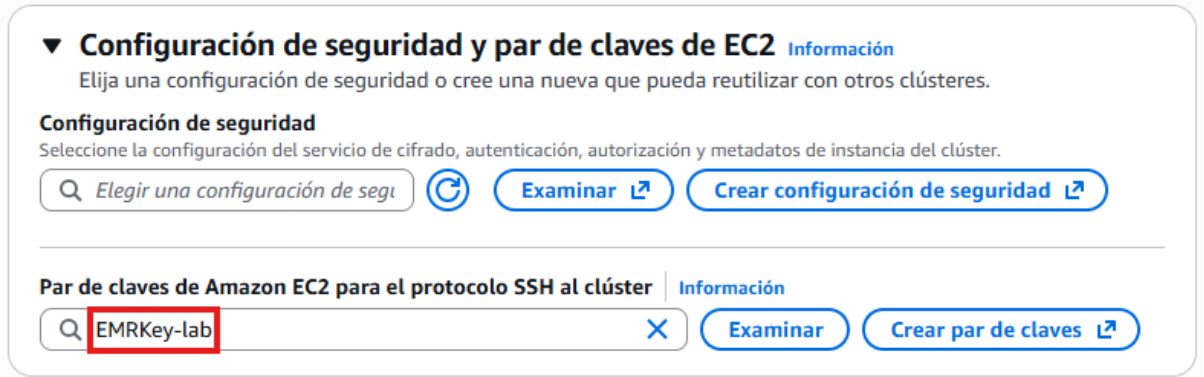
Y nos quedaría de esta manera;



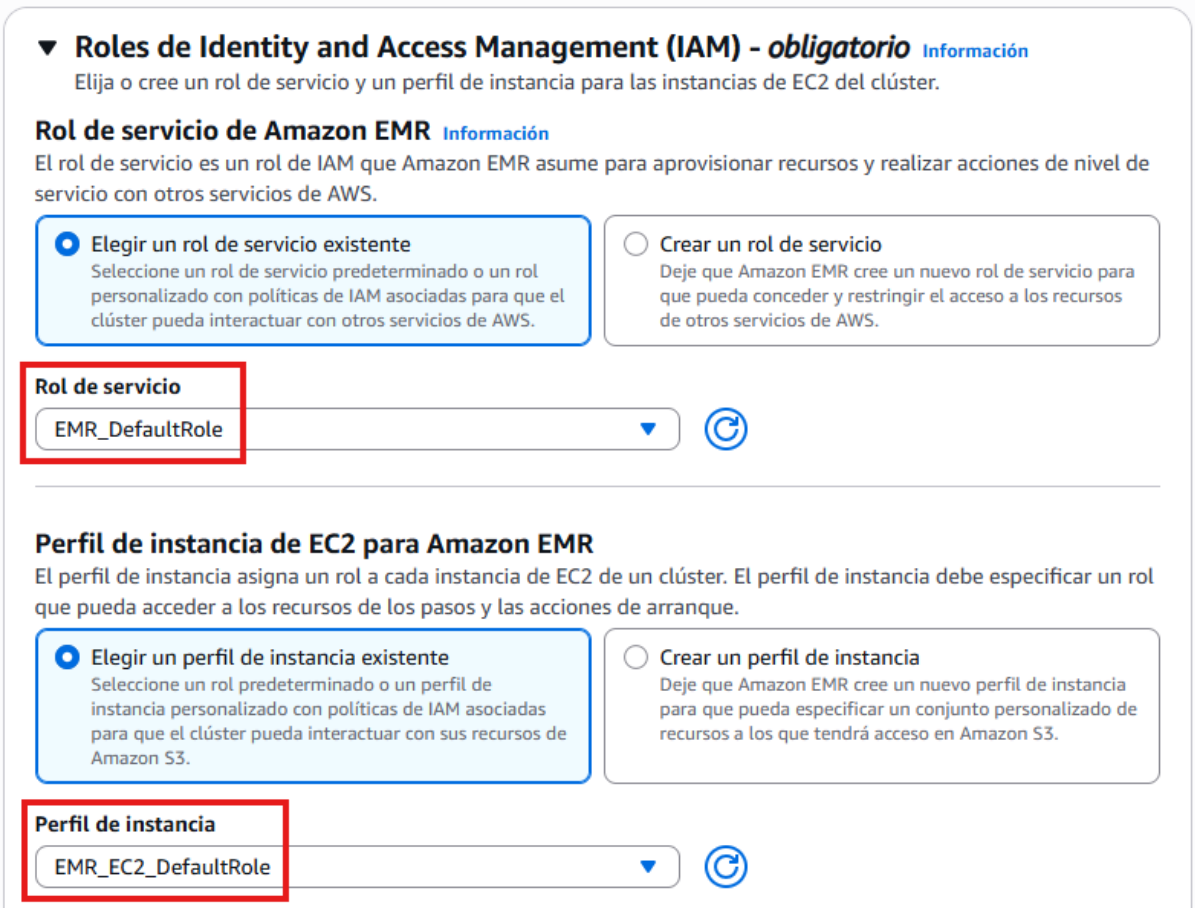
En la sección de "Configuración de seguridad y par de claves de EC2" le damos al botón de "examinar" en "Par de claves de Amazon EC2 para el protocolo SSH al clúster"



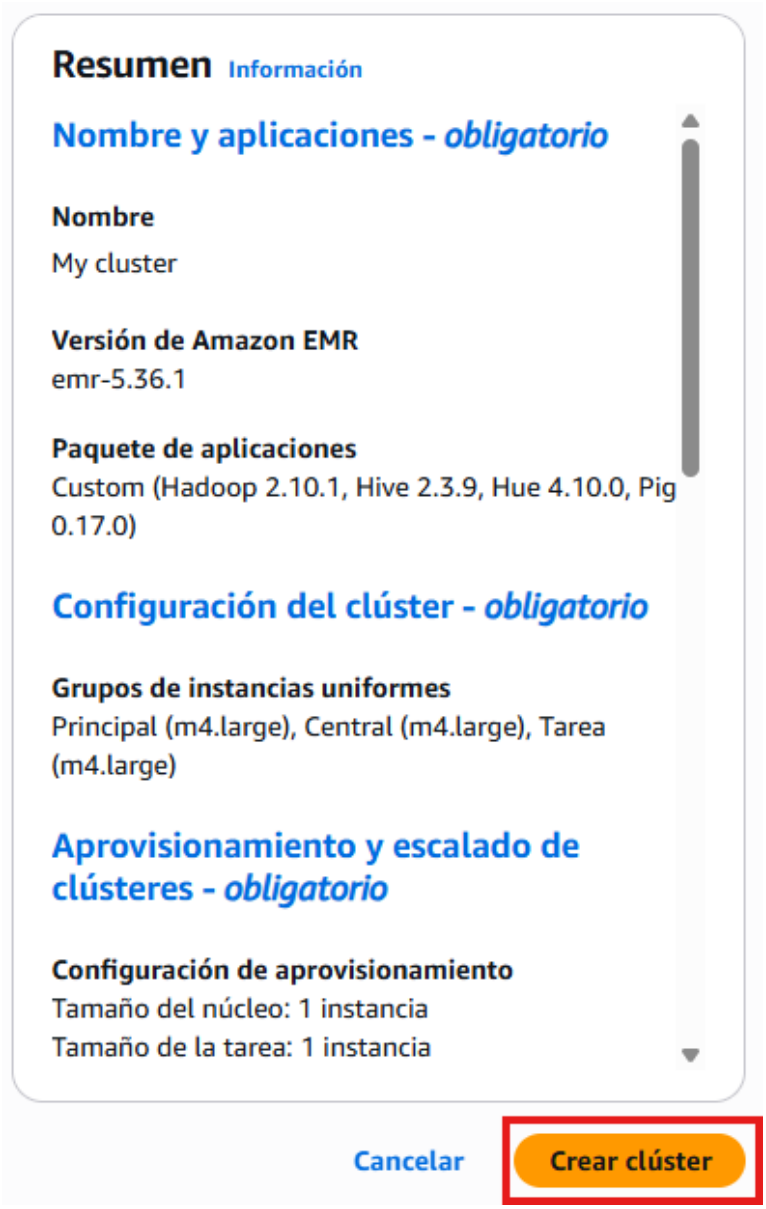
Nos quedará de la siguiente manera;

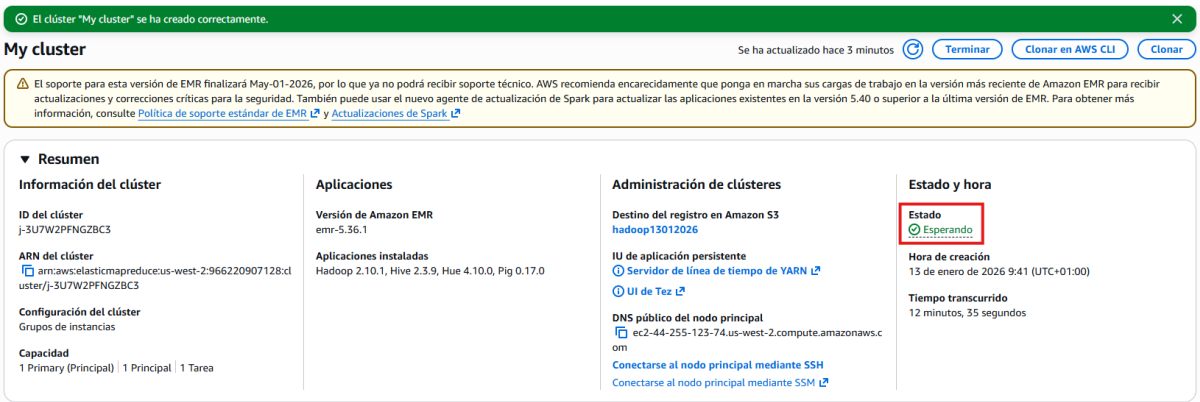


Ahora en la sección de "Roles de Identity and Access Management (IAM)" haremos esto;



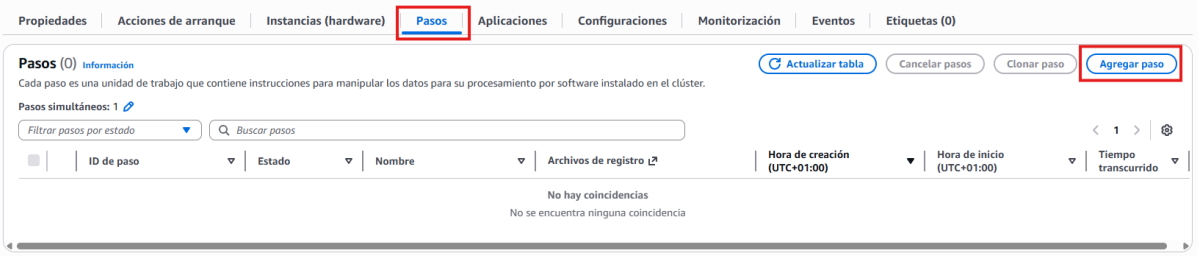
Finalmente, le daremos al botón de “Crear Cluster”



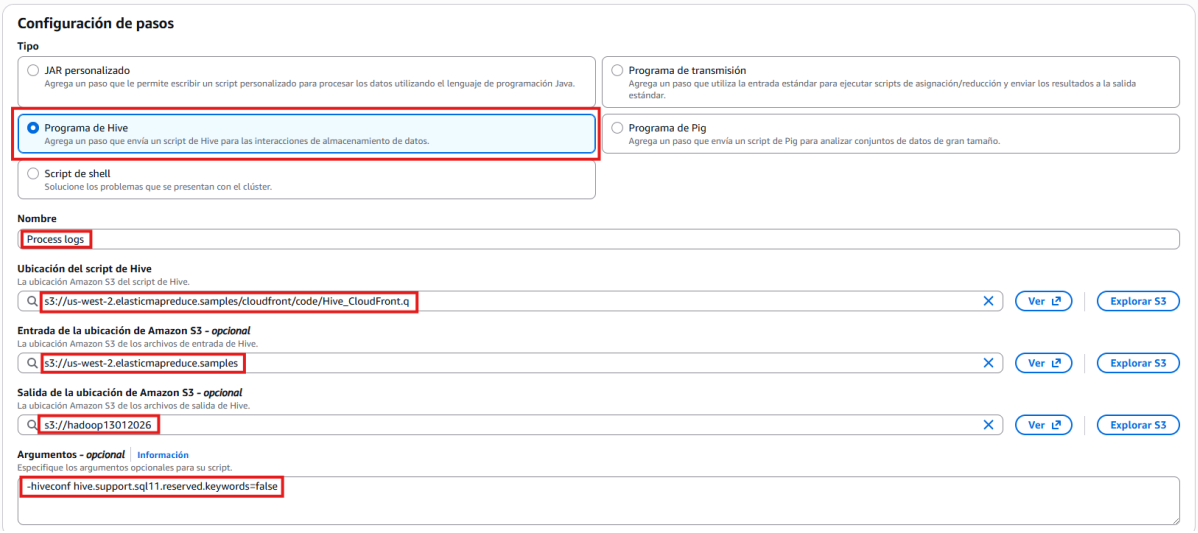


Tarea 3

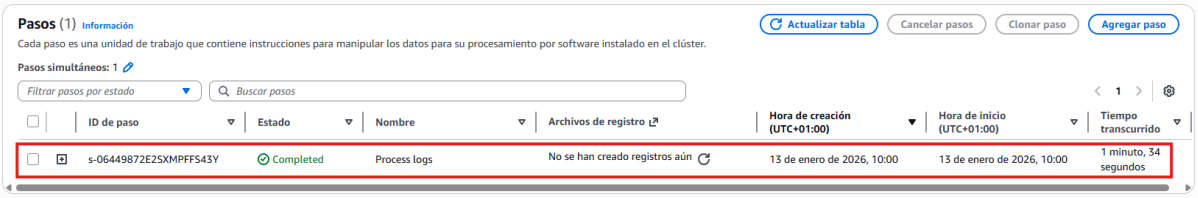
En el menú seleccionaremos “Pasos” y le daremos a “Agregar Paso”



Ahora tenemos que configurarlo de la siguiente manera, y ponerle nuestra región asignada, en mi caso es “us-west-2"

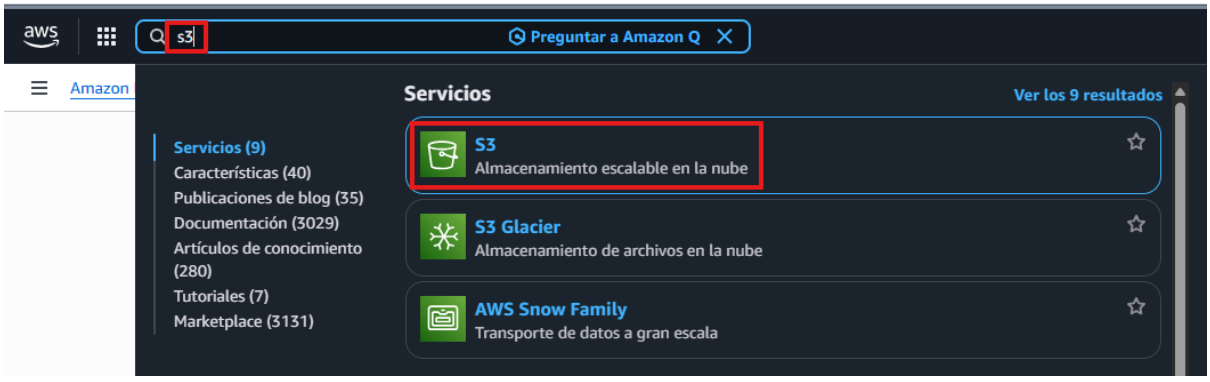


Finalmente, le damos a “Agregar Paso” y se nos creará

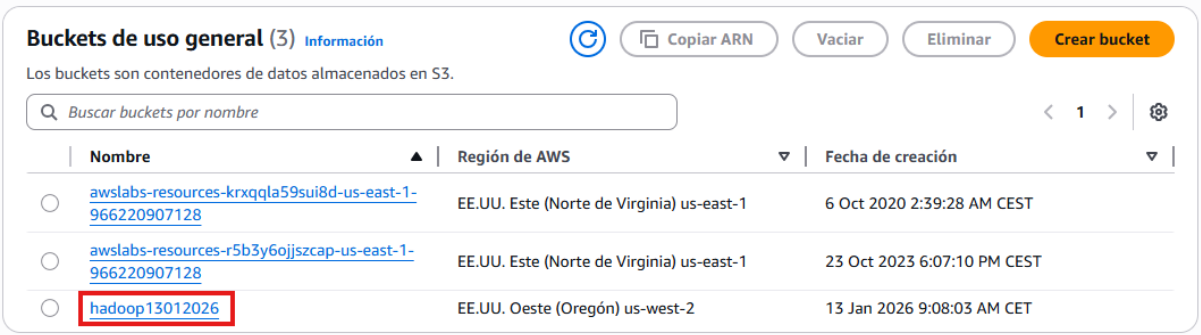


Tarea 4

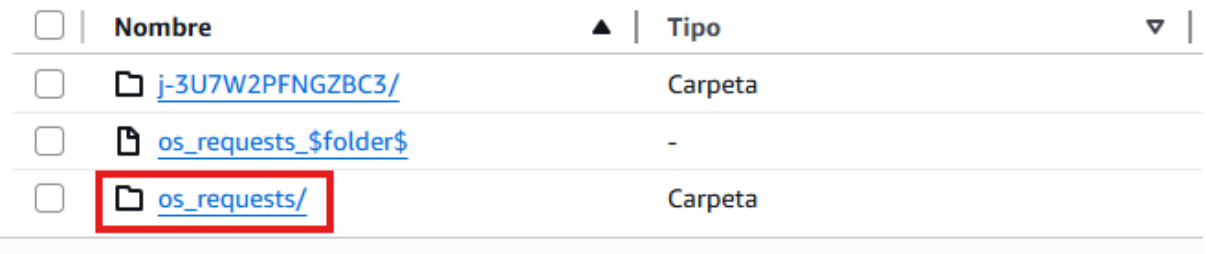
Ahora tenemos que volver al S3 de AWS, por lo tanto, lo volvemos a buscar en la lupa



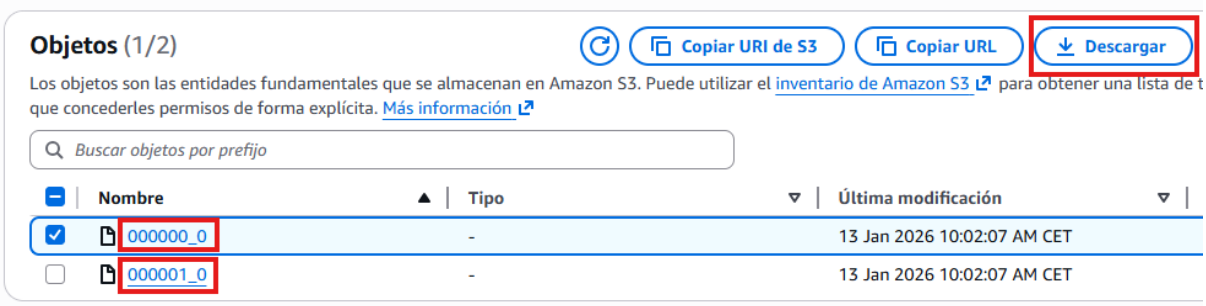
Una vez dentro, hacemos clic en nuestro bucket



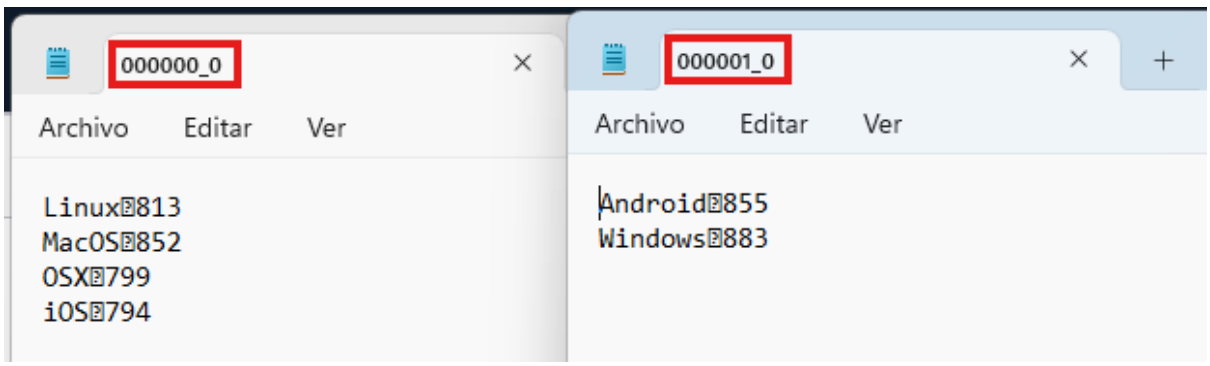
Entramos en la siguiente carpeta



Descargamos los archivos

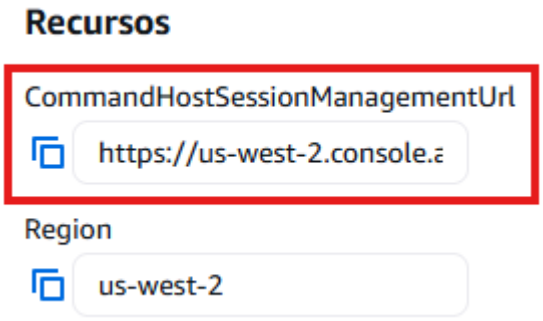


Finalmente, abrimos los archivos descargados para ver el contenido que tienen



Tarea 5

Copiamos este enlace que nos aparece en las instrucciones y lo abrimos en una nueva ventana



Ahora en la terminal ponemos los siguientes comandos;

```
sh-4.2$ export ID=$(aws emr list-clusters | jq '.Clusters[0].Id' | tr -d '"')
sh-4.2$ export HOST=$(aws emr describe-cluster --cluster-id $ID | jq '.Cluster.MasterPublicDnsName' | tr -d '"')
```

```
sh-4.2$ ssh -i ~/EMRKey-lab.pem hadoop@$HOST
The authenticity of host 'ec2-44-255-123-74.us-west-2.compute.amazonaws.com (10.1.12.67)' can't be established.
ECDSA key fingerprint is SHA256:B+oOOtPR3LY+zXpWHKtIcI6NoZLBA9+ZubT5yBU01vU.
ECDSA key fingerprint is MD5:b6:75:ee:17:e9:83:76:fe:bb:5d:bb:12:ae:dd:fa:c6.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-44-255-123-74.us-west-2.compute.amazonaws.com,10.1.12.67' (ECDSA) to the list of known hosts.
Last login: Tue Jan 13 09:02:04 2026
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2026-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

30 package(s) needed for security, out of 61 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-10-1-12-67 ~]$
```

Una vez conectados, escribimos “hive”

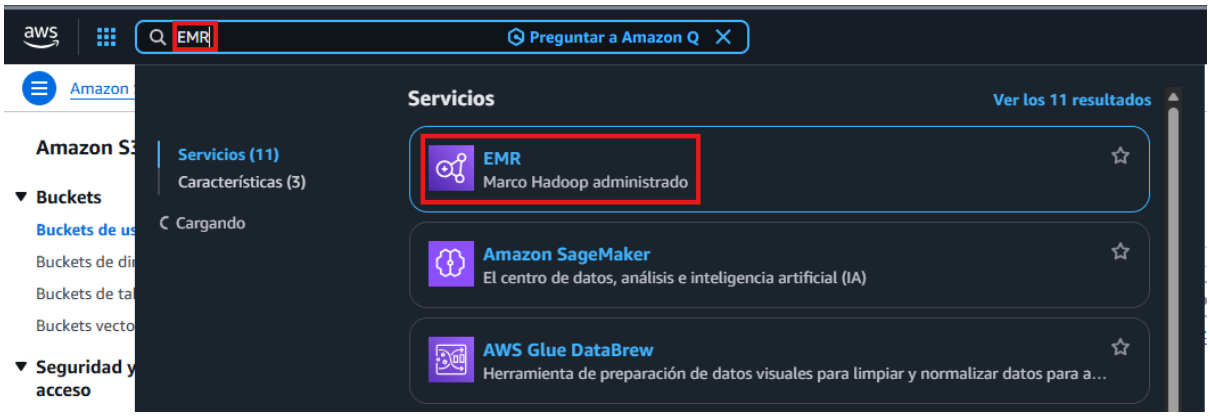
```
[hadoop@ip-10-1-12-67 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

Y ahora ponemos estos comandos

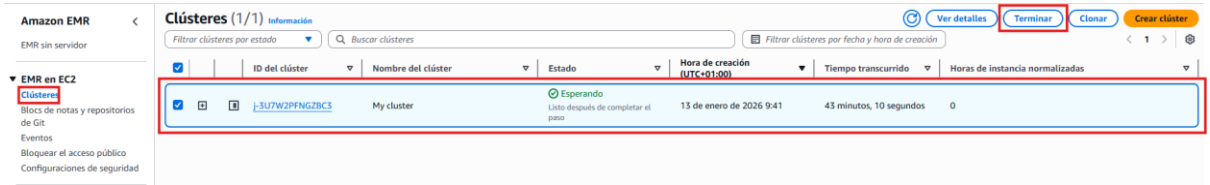
```
hive> SELECT os, COUNT(*) count FROM cloudfront_logs WHERE dateobject BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY os;
Query ID = hadoop_20260113091916_cb509209-f5db-4525-acd9-21182dc0e9a8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768294025176_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 29.90 s
--------------------------------------------------------------------------------
OK
Linux   813
MacOS   852
OSX     799
iOS     794
Android 855
Windows 883
Time taken: 36.434 seconds, Fetched: 6 row(s)
```

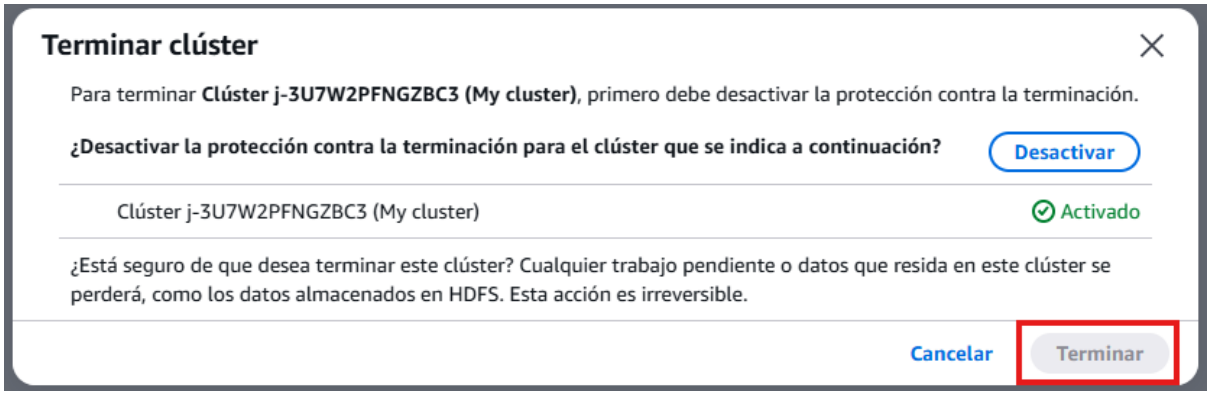
Tarea 6

Ahora volvemos a la búsqueda de AWS y ponemos “EMR”



Ahora iremos a nuestro Cluster y lo seleccionaremos y le daremos a “Terminar”





Si no nos deja terminarlo es devido a la protección contra la terminación del cluster, simplemente la desactivamos y le damos a terminar

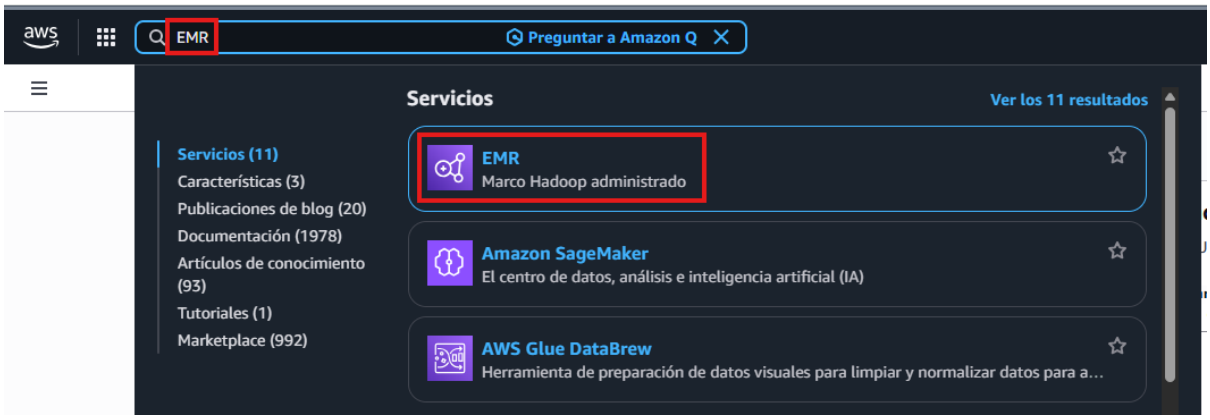
Apartado B

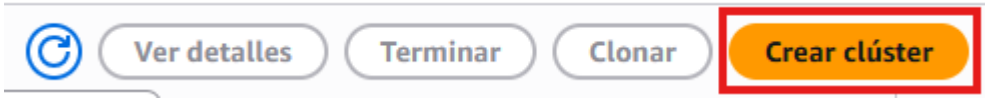
Exploring Google Ngrams with Amazon EMR and Hive

Este ejercicio vamos a realizarlo desde AWS ACADEMY y no desde entorno que nos propone el curso

Tarea 1

Primero, buscamos en la barra de búsqueda de AWS "EMR" y una vez dentro le damos al botón de "Crear Cluster"





Ahora configuramos el Cluster de la siguiente manera;

Crear clúster Información

Nombre y aplicaciones - *obligatorio* Información

Asigne un nombre a su clúster y elija las aplicaciones que desea instalar en él.

Nombre

Ngram cluster

Versión de Amazon EMR Información

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.

emr-7.12.0

Paquete de aplicaciones

Spark Interactive | Core Hadoop | Flink | HBase | Presto | Trino | Custom

- [ ] AmazonCloudWatchAgent 1.300032.2
- [ ] HCatalog 3.1.3
- [ ] Hue 4.11.0
- [ ] Livy 0.8.0
- [ ] Pig 0.17.0
- [ ] TensorFlow 2.19.0
- [ ] Zeppelin 0.11.1
- [ ] Flink 1.20.0
- [x] Hadoop 3.4.1
- [ ] JupyterEnterpriseGateway 2.6.0
- [ ] Oozie 5.2.1
- [ ] Presto 0.287
- [ ] Tez 0.10.2
- [ ] ZooKeeper 3.9.3
- [ ] HBase 2.6.2
- [x] Hive 3.1.3
- [ ] JupyterHub 1.5.0
- [ ] Phoenix 5.2.1
- [ ] Spark 3.5.6
- [ ] Trino 476

Configuración del clúster - *obligatorio* Información

Elija un método de configuración para los grupos principales, centrales y de nodos tareas para su clúster.

(•) Grupos de instancias uniformes
Elija el mismo tipo de instancia de EC2 y la misma opción de compra (bajo demanda o de spot) para todos los nodos de su grupo de nodos. Más información

( ) Flotas de instancias flexibles
Elija entre la más amplia variedad de opciones de aprovisionamiento para las instancias de EC2 de su clúster. Diversifique los tipos de instancias y las opciones de compra, y utilice una estrategia de asignación. Más información

Grupos de instancias uniformes

Principal

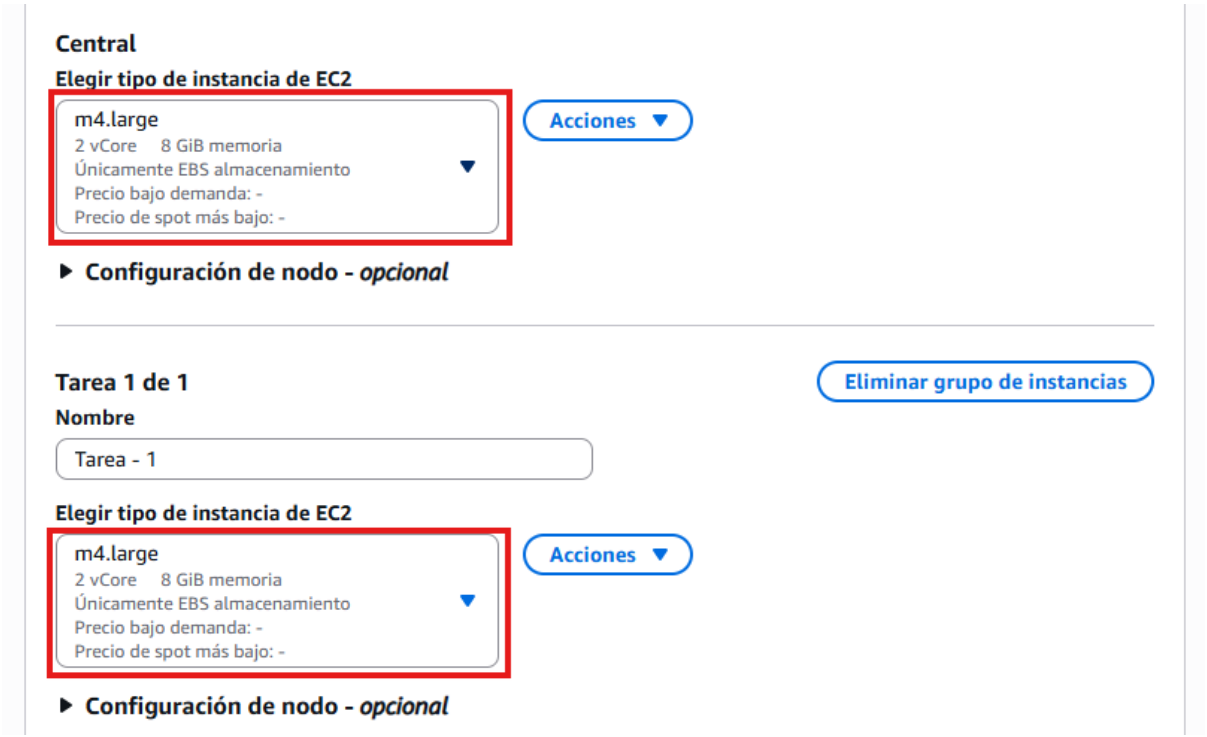
Elegir tipo de instancia de EC2

m4.large
2 vCore 8 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: -
Precio de spot más bajo: -

Acciones

- [ ] Utilice la alta disponibilidad
Lance un clúster más resiliente y de alta disponibilidad con tres nodos principales en instancias bajo demanda. Esta configuración se aplica durante toda la vida útil del clúster. Más información

▶ **Configuración de nodo - *opcional***



Aprovisionamiento y escalado de clústeres - *obligatorio* Información

Elija cómo Amazon EMR debe dimensionar su clúster.

Elija una opción

- **Establecer el tamaño del clúster manualmente** — Utilice esta opción si conoce los patrones de la carga de trabajo de antemano.
- **Utilizar escalado administrado por EMR** — Supervise las métricas clave de la carga de trabajo de modo que EMR pueda optimizar el tamaño del clúster y la utilización de los recursos.
- **Utilizar el escalamiento automático personalizado** — Para escalar mediante programación los nodos principales y los nodos de tarea, cree políticas de escalamiento automático personalizadas.

Configuración de aprovisionamiento

Establezca el tamaño del principal y tarea grupos de instancias. Amazon EMR intenta aprovisionar esta capacidad al lanzar el clúster.

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Central	m4.large	1	☐
Tarea - 1	m4.large	1	☐

Redes - *obligatorio* Información

Elija la configuración de red que determina la forma en que usted y otras entidades se comunican con su clúster.

Virtual Private Cloud (VPC) Información

vpc-0da527f677e65a9bf | Examinar | Crear VPC

Subred Información

subnet-07c8d2c469c627a2d | Examinar | Crear subred

Grupos de seguridad de EC2 (firewall)

Aviso de cambio
Hemos actualizado los nombres de algunos grupos de seguridad para utilizar un lenguaje más inclusivo. Por ejemplo, los grupos que incluían términos como "maestro" y "esclavo" ahora utilizan en su lugar los términos "principal" y "central".

Nodo principal

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

ElasticMapReduce-Primary
sg-0a91caae8b3aa783f

Grupos de seguridad adicionales - *opcional*
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales

Nodos principales y de tareas

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

ElasticMapReduce-Core
sg-02c2c569d408cc571

Grupos de seguridad adicionales - *opcional*
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales

Registros de clúster Información

Elija dónde y cómo almacenar los archivos de registro.

- [ ] Publicar registros específicos del clúster en Amazon S3
- [ ] Publicar registros específicos del clúster en Amazon CloudWatch

Configuración de seguridad y par de claves de EC2 Información

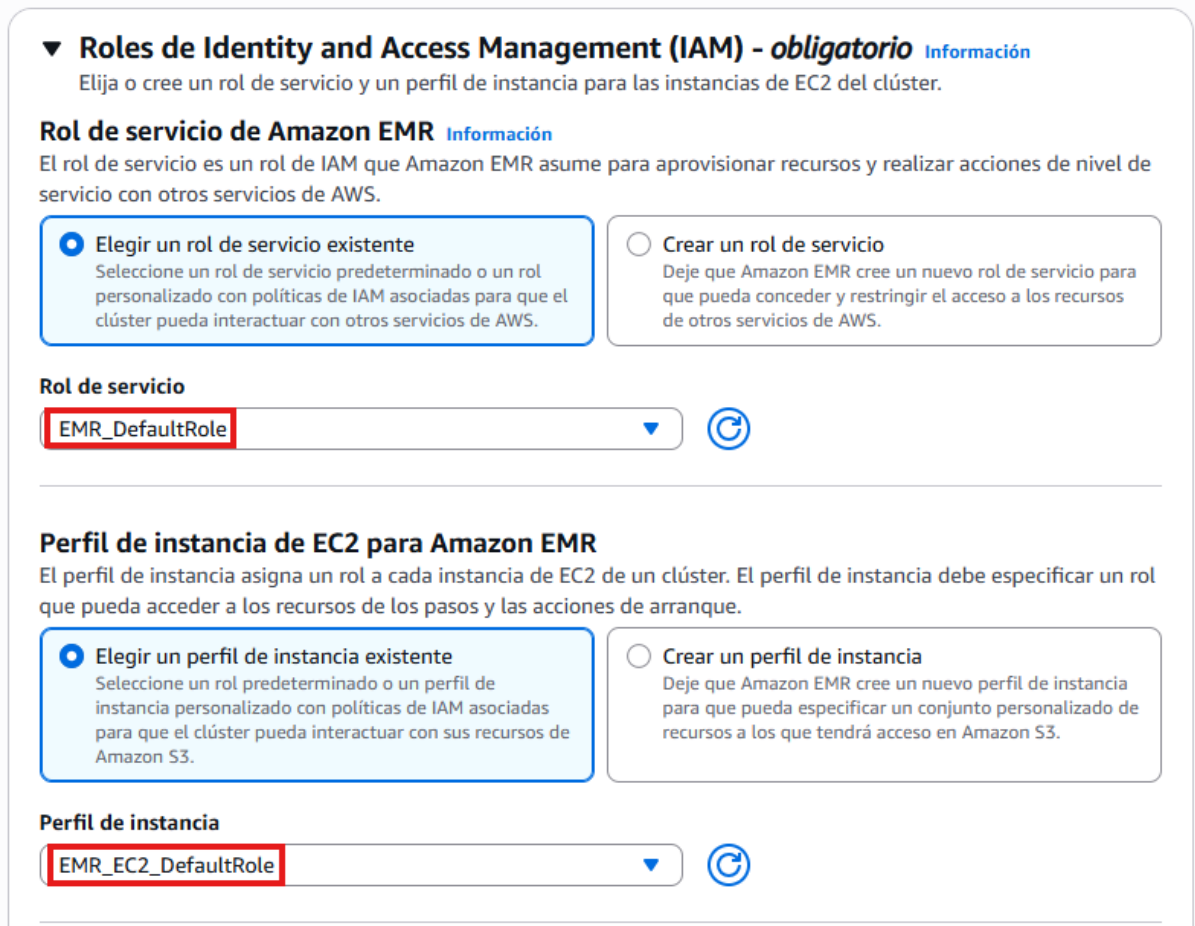
Elija una configuración de seguridad o cree una nueva que pueda reutilizar con otros clústeres.

Configuración de seguridad
Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.

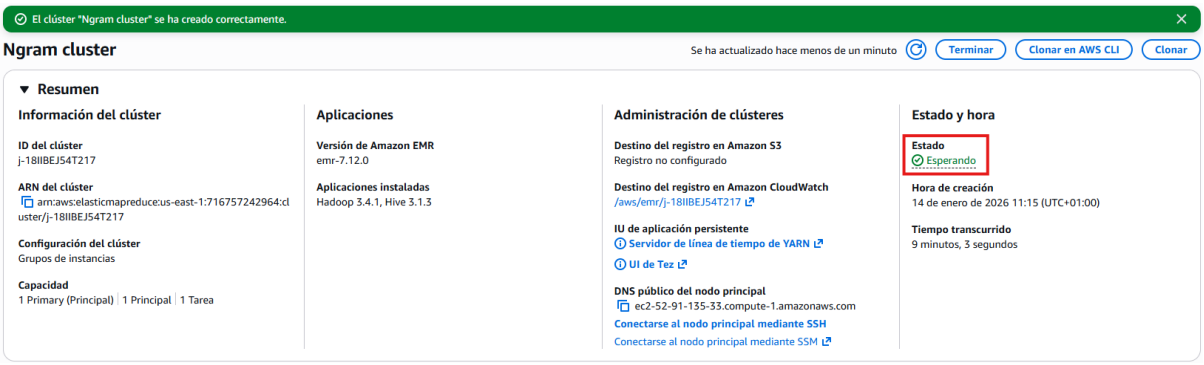
Elegir una configuración de segu | Examinar | Crear configuración de seguridad

Par de claves de Amazon EC2 para el protocolo SSH al clúster Información

emr | Examinar | Crear par de claves

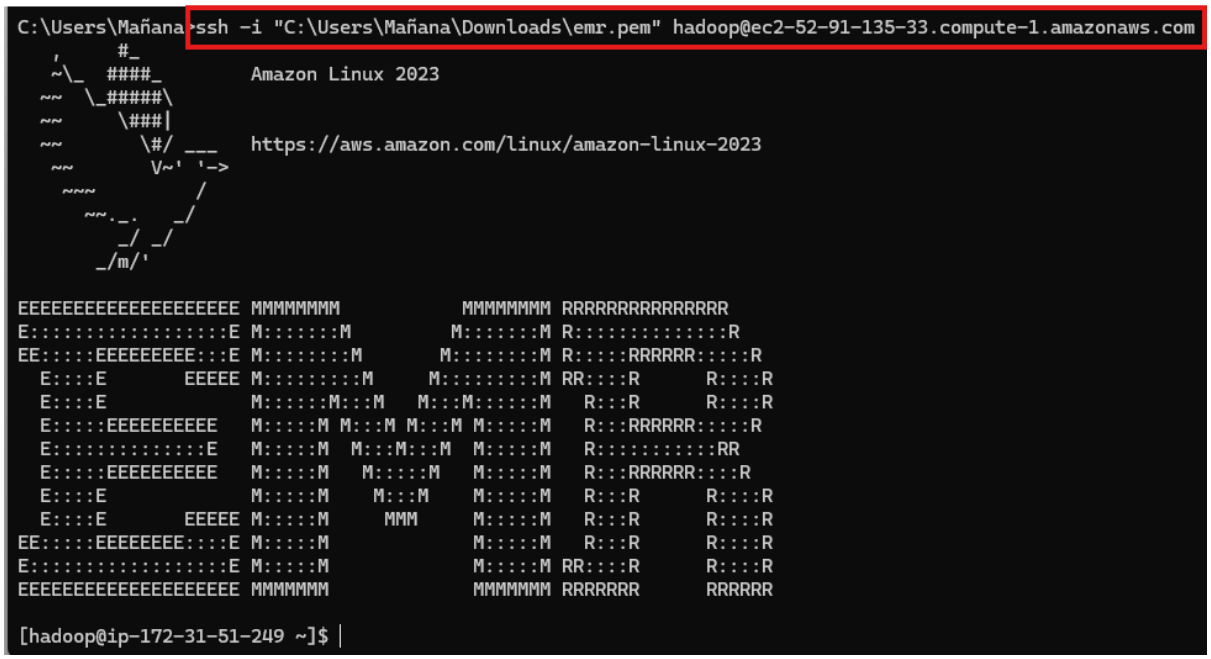


Finalmente le damos al botón de “Crear Cluster”



Tarea 2

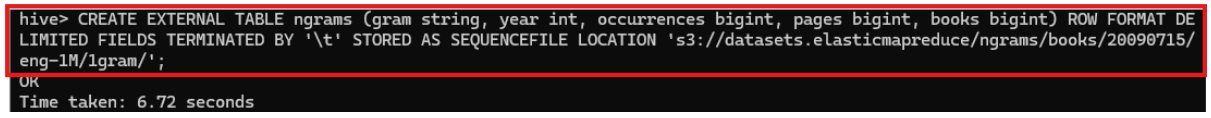
Ahora nos vamos a conectar mediante “ssh”



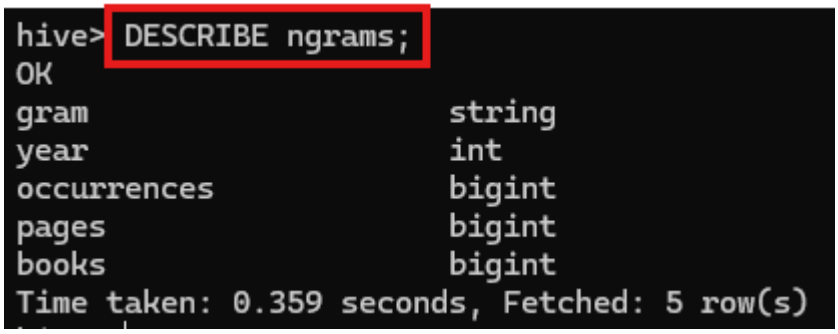
Una vez conectados, escribimos “hive”

Tarea 3

Escribimos el siguiente comando en la terminal;



Ahora ponemos lo siguiente;



Ahora vamos a mostrar los 10 primeros registros de la tabla:

```
hive> SELECT * FROM ngrams LIMIT 10;
OK
#       1574    1       1       1
#       1584    6       6       1
#       1614    1       1       1
#       1631    115     100     1
#       1632    3       3       1
#       1635    1       1       1
#       1640    1       1       1
#       1641    1       1       1
#       1642    5       5       1
#       1644    234     193     1
Time taken: 1.586 seconds, Fetched: 10 row(s)
```

Ahora vamos a almacenar los resultados de la normalización

```
hive> CREATE TABLE normalized (gram string, year int, occurrences bigint);
OK
Time taken: 0.581 seconds
```

```
hive> INSERT OVERWRITE TABLE normalized SELECT lower(gram), year, occurrences FROM ngrams WHERE year BETWEEN 1990 AND 20
05 AND gram REGEXP "^[A-Za-z+\'-]{3,}$";
Query ID = hadoop_20260114103640_17ad7c07-60cf-40ba-a180-738a840085e0
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 214.38 s
--------------------------------------------------------------------------------
Loading data to table default.normalized
OK
Time taken: 227.882 seconds
```

Ahora vamos a ver las 20 primeras filas de datos:

```
hive> SELECT * FROM normalized LIMIT 20;
OK
ingermany       1991    1
ingermany       1993    1
ingermany       1994    3
ingermany       1996    1
ingermany       2001    1
ingermany       2004    1
ingermany       2005    1
ingreece        1990    1
ingreece        2001    1
ingreece        2004    1
injuly  1990    7
injuly  1991    3
injuly  1992    6
injuly  1993    4
injuly  1994    1
injuly  1995    5
injuly  1996    4
injuly  1998    4
injuly  1999    3
injuly  2000    6
Time taken: 0.137 seconds, Fetched: 20 row(s)
```

Ahora vamos a ver las 50 palabras más usadas en todos los libros de todos los años:

```
hive> SELECT gram, sum(occurrences) as total_occurrences FROM normalized GROUP BY gram ORDER BY total_occurrences DESC L
IMIT 50;
Query ID = hadoop_20260114104325_bbee81fb-6f8e-4135-9cd6-b6e12126dd18
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 36.22 s
--------------------------------------------------------------------------------
OK
the     600731810
and     269591500
that    94084329
for     80649257
with    61620362
was     57843905
this    45202579
are     44749547
from    40039900
```

Ahora vamos a ver las 50 palabras más usadas de más de 10 caracteres:

```
hive> SELECT gram, sum(occurrences) as total_occurrences FROM normalized WHERE length(gram) > 10 GROUP BY gram ORDER BY
total_occurrences DESC LIMIT 50;
Query ID = hadoop_20260114104552_d294dc1b-b92b-478b-8146-7a27ede5bf65
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 31.33 s
--------------------------------------------------------------------------------
OK
development     4584319
information     4419750
international   2731441
relationship    2013252
significant     1762598
particularly    1709008
performance     1669887
```

Después de esto, vamos a crear la tabla “ratios”:

```
hive> CREATE TABLE ratios (gram string, year int, occurrences bigint, ratio double);
OK
Time taken: 0.071 seconds
```

Ahora vamos a rellenar la tabla “ratios” con los datos calculados a partir de la tabla “normalized”:

```
hive> INSERT OVERWRITE TABLE ratios SELECT a.gram, a.year, sum(a.occurrences) AS occurrences, sum(a.occurrences) / b.tot
al AS ratio FROM normalized a JOIN (SELECT year, sum(occurrences) AS total FROM normalized GROUP BY year) b ON (a.year =
 b.year) GROUP BY a.gram, a.year, b.total;
Query ID = hadoop_20260114104939_0496bbc5-7a48-401e-bd62-fb901d8ffe39
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 5 ...... container     SUCCEEDED      2          2        0        0       0       0
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      4          4        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 130.71 s
--------------------------------------------------------------------------------
Loading data to table default.ratios
OK
Time taken: 132.152 seconds
```

Ahora vamos a calcular la diferencia de proporciones año tras año:

```
hive> SELECT year, gram, occurrences, CONCAT(CAST(increase AS INT), 'x increase') as increase FROM ( SELECT y2.gram, y2.year, y2
.occurrences, y2.ratio / y1.ratio as increase, rank() OVER (PARTITION BY y2.year ORDER BY y2.ratio / y1.ratio DESC) AS rank FROM
 ratios y2 JOIN ratios y1 ON y1.gram = y2.gram and y2.year = y1.year + 1 WHERE y2.year BETWEEN 1991 and 2005 AND y1.occurrences
> 1000 AND y2.occurrences > 1000 ) grams WHERE rank = 1 ORDER BY year;
Query ID = hadoop_20260114105430_c53dad37-bf8c-4a1c-8aff-73da1bdc57f2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED     10         10        0        0       0       0
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 46.12 s
--------------------------------------------------------------------------------
OK
1991    amyloid 6405    5x increase
1992    comm    18841   8x increase
1993    abstr   7033    6x increase
1994    carole  8358    7x increase
1995    mansfield       4570    3x increase
1996    polymerization  14442   8x increase
1997    tho     19259   8x increase
1998    oswald  8774    6x increase
1999    sql     12516   6x increase
2000    dlb     12369   10x increase
2001    dcs     6031    5x increase
2002    proust  6231    5x increase
2003    olfactory       8538    6x increase
2004    eeg     8873    5x increase
2005    rectum  6981    6x increase
Time taken: 47.367 seconds, Fetched: 15 row(s)
```

Ahora vamos a ver cómo ha aumentado el uso de la palabra “internet”:

```
hive> SELECT year, occurrences FROM ratios WHERE gram = 'internet' ORDER BY year;
Query ID = hadoop_20260114105727_fe766bcf-4c57-4c59-8bfb-cc92ce7e35aa
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 25.50 s
--------------------------------------------------------------------------------
OK
1990    1201
1991    828
1992    1981
1993    5265
1994    8132
1995    14491
1996    21064
1997    26982
1998    30317
1999    40579
2000    50505
2001    55799
2002    55137
2003    55793
2004    40861
2005    39483
Time taken: 25.87 seconds, Fetched: 16 row(s)
```

Finalmente, escribimos el siguiente comando:

```
hive> SELECT DISTINCT length, gram FROM ( SELECT length(gram) AS length, gram, rank() OVER (partition by length(gram) order by o
ccurrences desc) AS rank FROM ratios ) x WHERE rank = 1 ORDER BY length;
Query ID = hadoop_20260114110003_2e8337b9-90b7-40ee-99c3-022a5b7ed687
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED     10         10        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ......  container     SUCCEEDED      6          6        0        0       0       0
Reducer 4 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 48.64 s
--------------------------------------------------------------------------------
OK
```

```
3       the
4       that
5       which
6       people
7       between
8       american
9       different
10      university
11      development
12      relationship
13      international
14      administration
15      characteristics
16      responsibilities
17      industrialization
18      telecommunications
19      hyperparathyroidism
20      institutionalization
21      psychopharmacological
22      electroencephalography
23      electroencephalographic
24      cholangiopancreatography
25      methylenetetrahydrofolate
26      abcdefghijklmnopqrstuvwxyz
27      ooooooooooooooooooooooooooo
28      trimethoprimsulfamethoxazole
29      methylenedioxymethamphetamine
30      dipalmitoylphosphatidylcholine
31      dichlorodiphenyltrichloroethane
32      oooooooooooooooooooooooooooooooo
33      ooooooooooooooooooooooooooooooooo
34      oooooooooooooooooooooooooooooooooo
35      ooooooooooooooooooooooooooooooooooo
36      oooooooooooooooooooooooooooooooooooo
Time taken: 49.271 seconds, Fetched: 34 row(s)
```

Ya acabamos todos los apartados que venian en la práctica, por lo tanto, ya podemos darle al botón de “Finalizar Laboratorio”

1.) ¿Qué contiene el bucket s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/? ¿Cuánto ocupa el archivo que contiene?

Ese Bucket, proporciona un recuento de las "palabras sueltas" que se encuentran en todos los libros. Estos datos tienen 261 millones de entradas y ocupan 2,6 GB de almacenamiento en disco.

La información la encontramos en este apartado de la guía:

Tarea 3: Analizar los datos

En esta tarea, analizarás los datos de Ngrams en una sesión interactiva de Hive.

3.1 Examinar los datos sin procesar

En esta sección, accederás a los datos sin procesar de Ngrams. Accederás a los datos sobre *1-grams*, que proporcionan un recuento de las *palabras sueltas* que se encuentran en todos los libros. Estos datos tienen 261 millones de entradas y ocupan 2,6 GB de almacenamiento en disco.

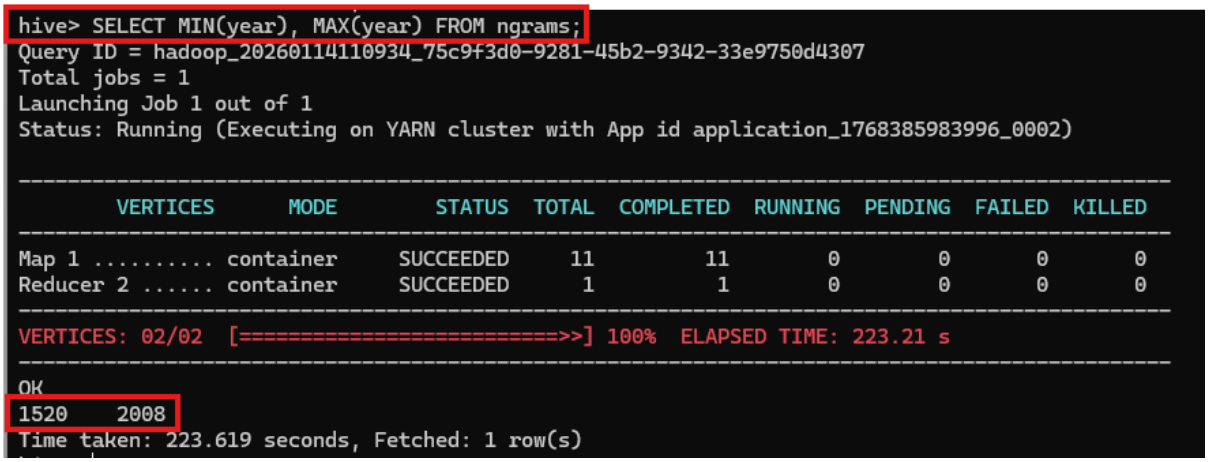
Los datos de Ngram son accesibles desde Amazon S3 y se puede acceder a ellos directamente desde Amazon EMR mediante la creación de una *tabla externa*. Esta definición indica a Amazon EMR qué formato tienen los datos y dónde se encuentran.

2.) ¿Cuántos registros contiene la tabla ngrams que creaste en HIVE? ¿Desde qué año hasta qué año abarca la información que contiene?

Para ver cuántos registros contiene la tabla "ngrams" usamos el siguiente comando:

```
hive> SELECT COUNT(*) FROM ngrams;
Query ID = hadoop_20260114110456_6b4111b1-588c-47b4-b37a-8858fef8e381
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768385983996_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11         11        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 221.02 s
----------------------------------------------------------------------------------------------
OK
261823186
Time taken: 221.431 seconds, Fetched: 1 row(s)
```

Y para ver desde que año hasta que año abarca utilizamos el siguiente comando:



3.) En la creación de la tabla normalized ¿qué significa la expresión REGEXP "^[A-Za-z+\'-]{3,}$"? ¿Cuántos registros contiene la tabla normalized?

Esa expresión significa que solo se pueden utilizar palabras, sin números, sin símbolos raros y que al menos tenga 3 letras

Para saber cuántos registros contiene la tabla “normalized” usamos el siguiente comando: