Como resultado de la práctica has de entregar un archivo en formato *"pdf"* con breves explicaciones y capturas de pantalla de los pasos principales que has dado para realizar los ejercicios.

Realizaremos dos apartados del curso del enlace de abajo en AWS Skillbuilder:

https://skillbuilder.aws/learning-plan/J38YWQY59M/data-analytics-learning-plan-includes-labs-espaol-de-espaa/TDYZZ22A7S

Data Analytics Learning Plan (includes Labs) (Español de España)

CONTENIDO

APARTADO A

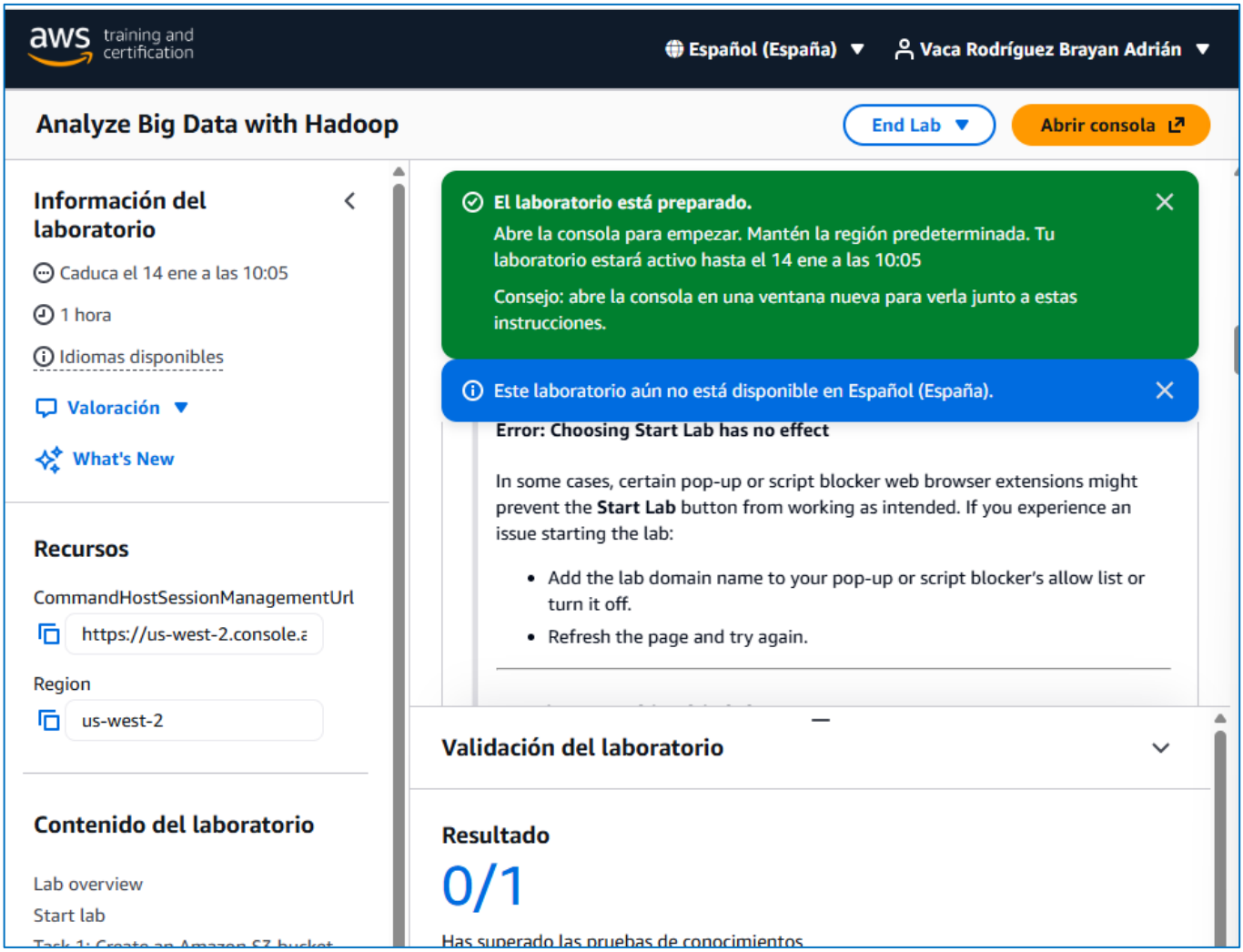
INTRODUCCIÓN

Realiza el laboratorio:

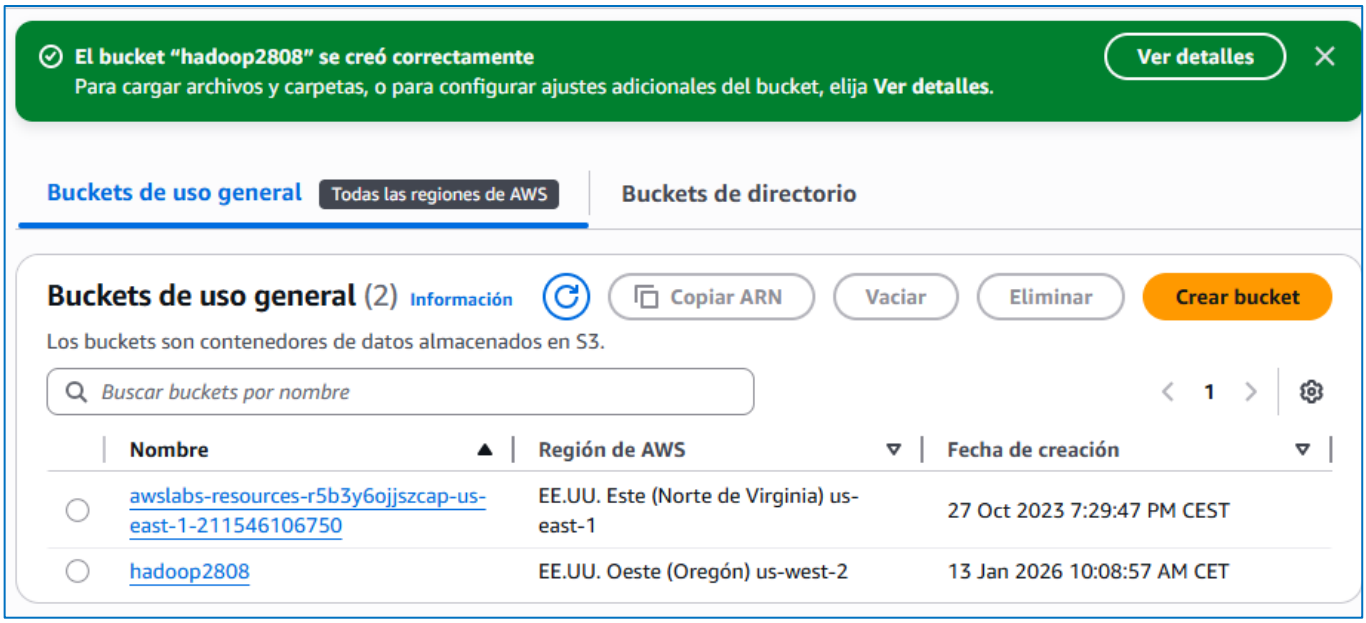
En él se analizan datos de Amazon CloudFront, servicio de AWS que genera registros de acceso que muestran todos los datos solicitados por los usuarios, con el siguiente formato (se explica el contenido de cada campo en el laboratorio):

```
#########################

#### SAMPLE LOG DATA ####

#########################

2017-07-05 20:05:47 SEA4 4261 10.0.0.15 eabcd12345678.cloudfront.net /test-image-2.jpeg
Mozilla/5.0%20(MacOS;%20U;%20Windows%20NT%205.1;%20enUS;%20rv:1.9.0.9)%20Gecko/2009040821%20Chrome/3.0.9
```

No necesitas usar AWS Academy, el curso incluye su propio laboratorio.



Cremaos el bucket



Crear clúster Información

Nombre y aplicaciones - *obligatorio* Información

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

Nombre

My cluster

Versión de Amazon EMR Información

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

emr-5.36.1

⚠ El soporte para esta versión de EMR finalizará May-01-2026, por lo que ya no podrá recibir soporte técnico. AWS recomienda encarecidamente que ponga en marcha sus cargas de trabajo en la versión más reciente de Amazon EMR para recibir actualizaciones y correcciones críticas para la seguridad. También puede usar el nuevo agente de actualización de Spark para actualizar las aplicaciones existentes en la versión 5.40 o superior a la última versión de EMR. Para obtener más información, consulte Política de soporte estándar de EMR y Actualizaciones de Spark

Paquete de aplicaciones

Spark	Core Hadoop	HBase	Presto	Custom
Apache Spark	hadoop	Apache HBase	presto	aws

- [ ] Flink 1.14.2
- [ ] HCatalog 2.3.9
- [x] Hue 4.10.0
- [ ] Livy 0.7.1
- [ ] Oozie 5.2.1
- [ ] Presto 0.267
- [ ] TensorFlow 2.4.1
- [ ] ZooKeeper 3.4.14
- [ ] Ganglia 3.7.2
- [x] Hadoop 2.10.1
- [ ] JupyterEnterpriseGateway 2.6.0
- [ ] MXNet 1.8.0
- [ ] Phoenix 4.14.3
- [ ] Spark 2.4.8
- [ ] Tez 0.9.2
- [ ] HBase 1.4.13
- [x] Hive 2.3.9
- [ ] JupyterHub 1.4.1
- [ ] Mahout 0.13.0
- [x] Pig 0.17.0
- [ ] Sqoop 1.4.7
- [ ] Zeppelin 0.10.0

Resumen Información

Nombre y aplicaciones - *obligatorio*

Nombre
My cluster

Versión de Amazon EMR
emr-5.36.1

Paquete de aplicaciones
Custom (Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Pig 0.17.0)

Configuración del clúster - *obligatorio*

Grupos de instancias uniformes
Principal (m4.large), Central (m4.large), Tarea (m4.large)

Aprovisionamiento y escalado de clústeres - *obligatorio*

Configuración de aprovisionamiento
Tamaño del núcleo: 1 instancia
Tamaño de la tarea: 1 instancia

Redes - *obligatorio*

Cancelar Crear clúster

Resumen Información

Redes - *obligatorio*

VPC
vpc-06b1370da...

Subred
subnet-057276...

Grupos de seguridad de nodos principales
sg-0657de1ede...

Grupos de seguridad de nodos básicos
sg-0657de1ede...

Terminación del clúster

Terminación del clúster
Terminar el clúster después del tiempo de inactividad
Tiempo de inactividad: 1 hora

Registros de clúster

Ubicación de Amazon S3
s3://hadoop28...

Cancelar | Crear clúster

Registros de clúster

Ubicación de Amazon S3
s3://hadoop28...

Configuración de seguridad y par de claves de EC2

Par de claves de Amazon EC2
EMRKey-lab

Roles de Identity and Access Management (IAM) - *obligatorio*

Rol de servicio
EMR_DefaultRole

Perfil de instancia
EMR_EC2_DefaultRole

Cancelar | Crear clúster

Agregar paso Información

Configuración de pasos

Tipo

- ○ **JAR personalizado**
  Agrega un paso que le permite escribir un script personalizado para procesar los datos utilizando el lenguaje de programación Java.
- ○ **Programa de transmisión**
  Agrega un paso que utiliza la entrada estándar para ejecutar scripts de asignación/reducción y enviar los resultados a la salida estándar.
- ◉ **Programa de Hive**
  Agrega un paso que envía un script de Hive para las interacciones de almacenamiento de datos.
- ○ **Programa de Pig**
  Agrega un paso que envía un script de Pig para analizar conjuntos de datos de gran tamaño.
- ○ **Script de shell**
  Solucione los problemas que se presentan con el clúster.

Nombre

Process logs

Ubicación del script de Hive

La ubicación Amazon S3 del script de Hive.

s3://us-west-2.elasticmapreduce.samples/cloudfront/code/Hive_CloudFront.q [Ver] [Explorar S3]

Entrada de la ubicación de Amazon S3 - *opcional*

La ubicación Amazon S3 de los archivos de entrada de Hive.

s3://us-west-2.elasticmapreduce.samples [Ver] [Explorar S3]

Salida de la ubicación de Amazon S3 - *opcional*

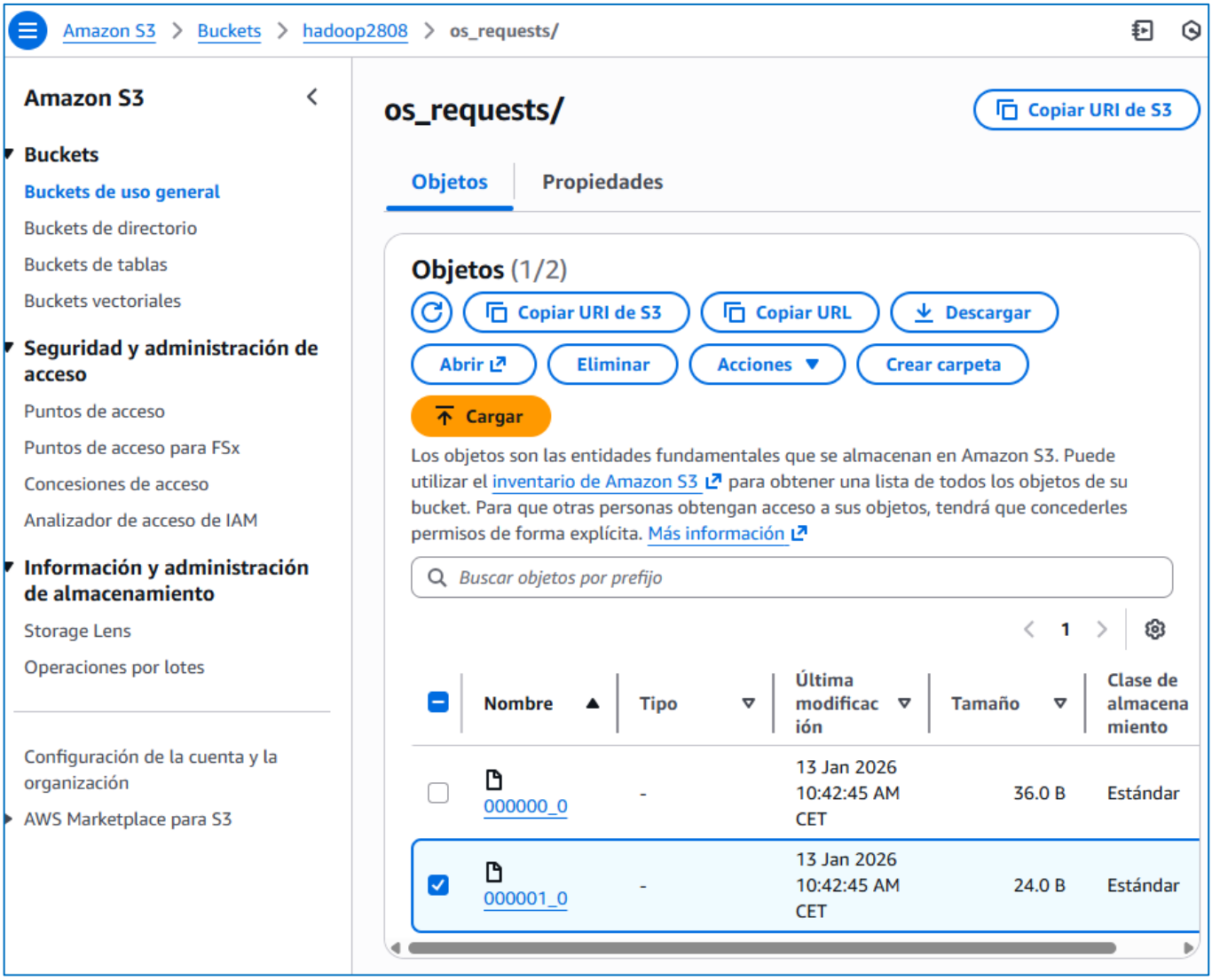
La ubicación Amazon S3 de los archivos de salida de Hive.

s3://hadoop2808 [Ver] [Explorar S3]

Argumentos - *opcional* Información

Especifique los argumentos opcionales para su script.

```
-hiveconf hive.support.sql11.reserved.keywords=false
```


Amazon S3 > Buckets > hadoop2808 > os_requests/
Amazon S3
Buckets
Buckets de uso general
Buckets de directorio
Buckets de tablas
Buckets vectoriales
Seguridad y administración de acceso
Puntos de acceso
Puntos de acceso para FSx
Concesiones de acceso
Analizador de acceso de IAM
Información y administración de almacenamiento
Storage Lens
Operaciones por lotes
Configuración de la cuenta y la organización
AWS Marketplace para S3
os_requests/
Copiar URI de S3
Objetos
Propiedades
Objetos (1/2)
Copiar URI de S3
Copiar URL
Descargar
Abrir
Eliminar
Acciones
Crear carpeta
Cargar
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3 para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. Más información
Buscar objetos por prefijo
1
Nombre
Tipo
Última modificación
Tamaño
Clase de almacenamiento
000000_0
-
13 Jan 2026 10:42:45 AM CET
36.0 B
Estándar
000001_0
-
13 Jan 2026 10:42:45 AM CET
24.0 B
Estándar

ID de sesión: 60b66c8d-09a7-44bf-9cb8-2e92c43c6b4c-rrsk5gkf5foat9or7iqdllu4

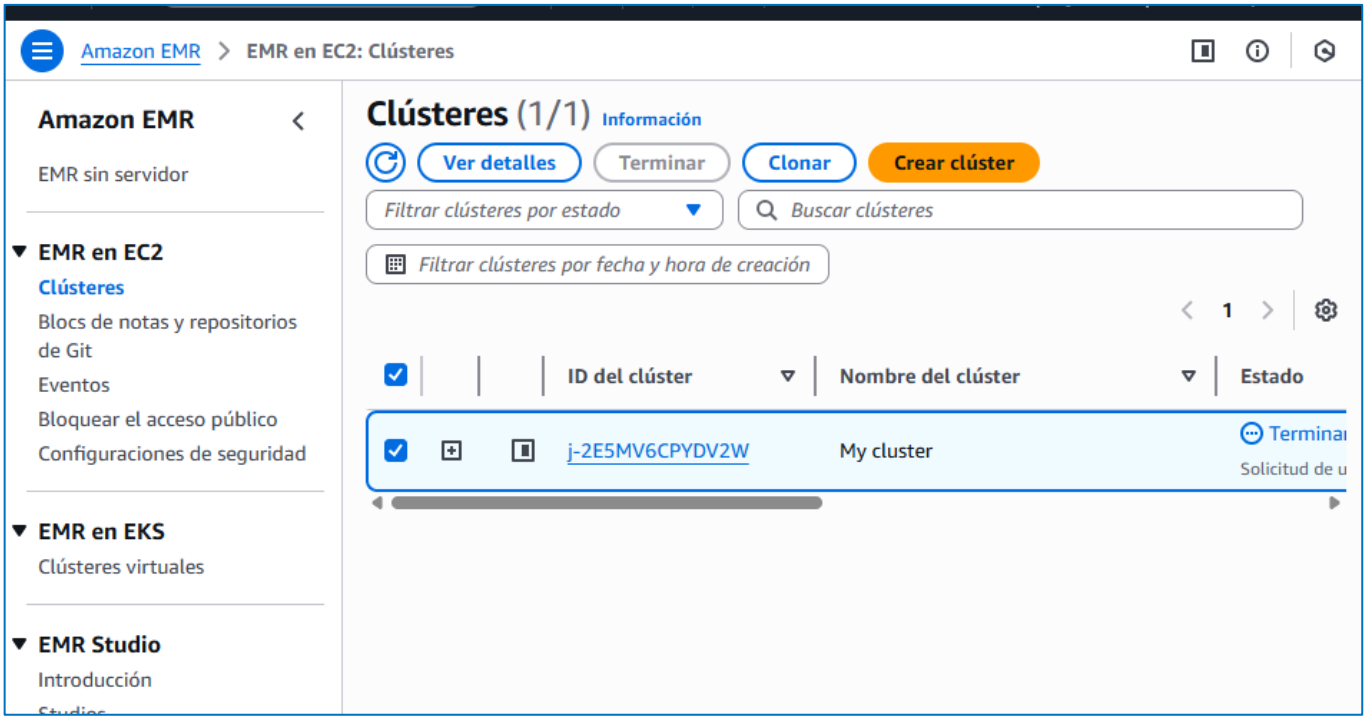
ID de instancia: i-0c15cc4e76751f2d1

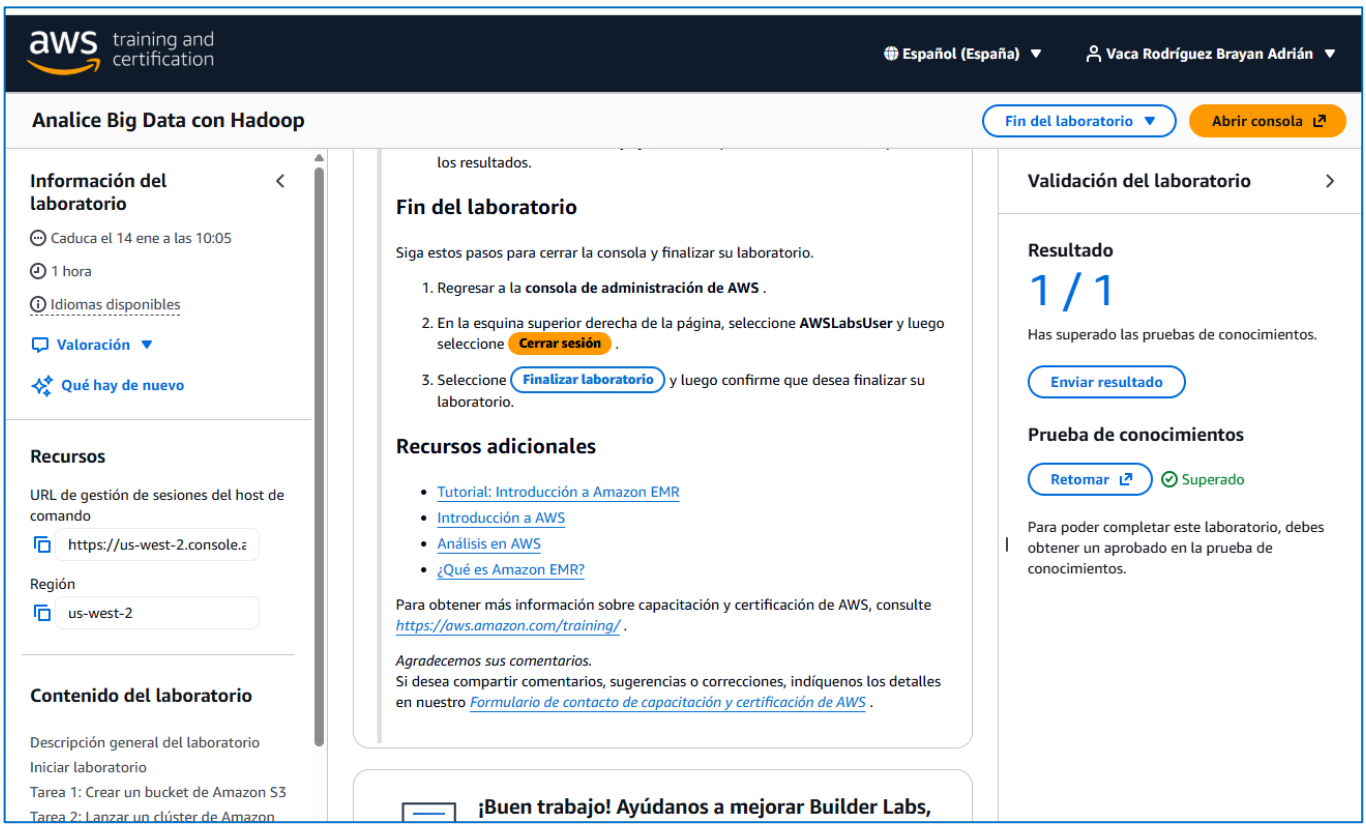
Terminar

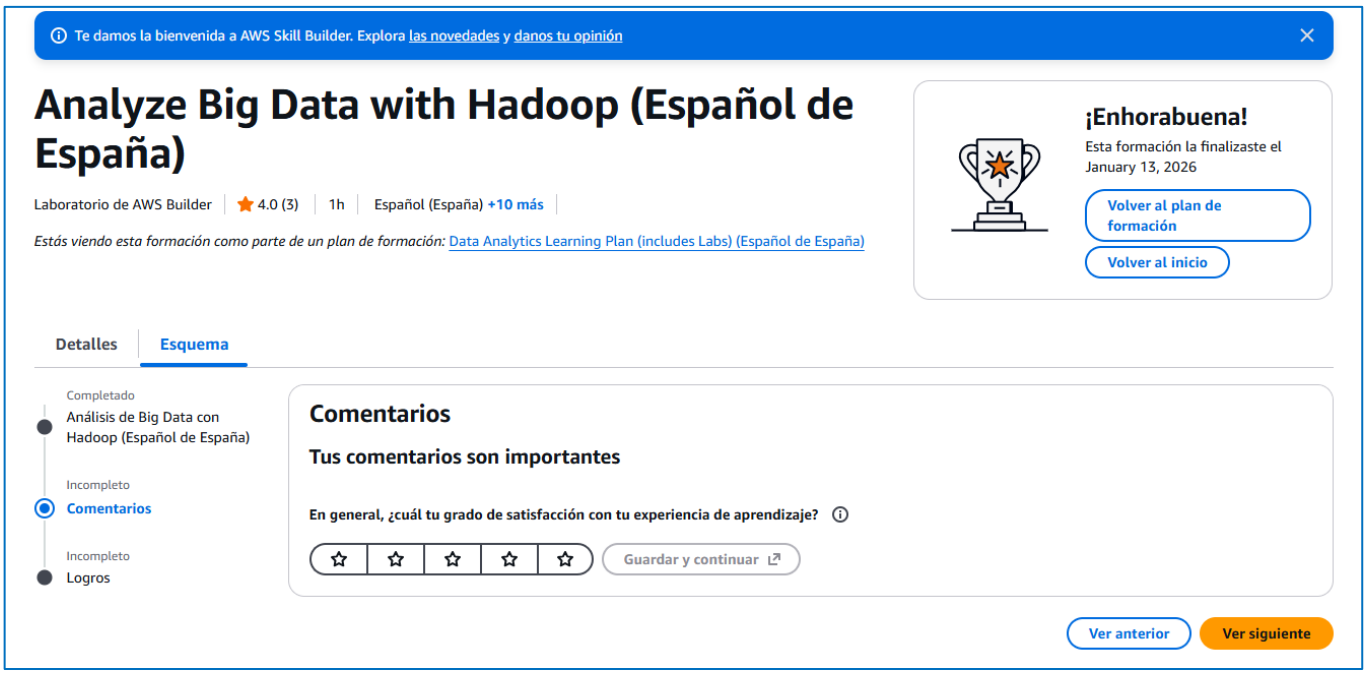
cy Shortcuts

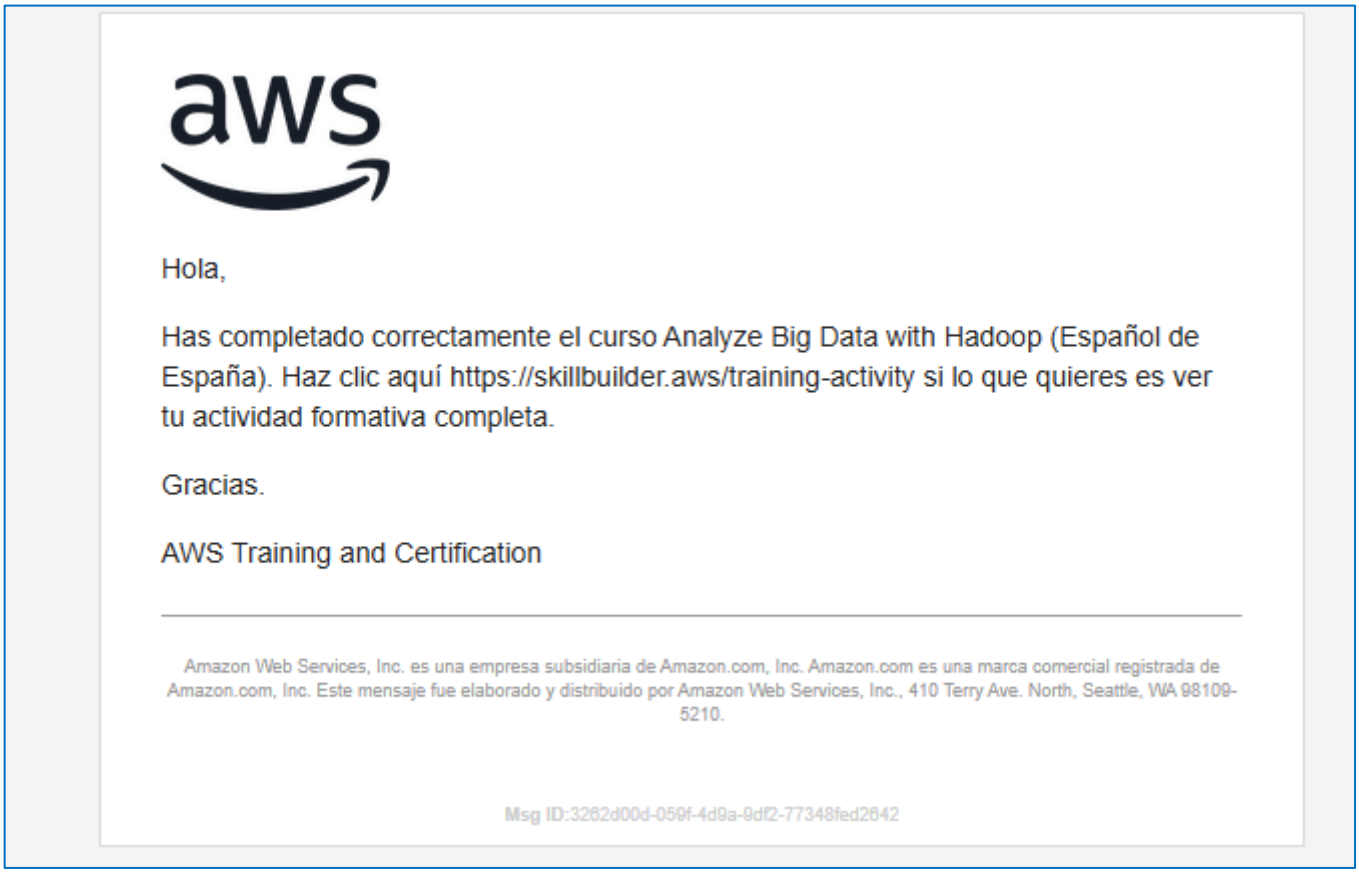
```
sh-4.2$ export ID=$(aws emr list-clusters | jq '.Clusters[0].Id' | tr -d '"')
sh-4.2$
sh-4.2$ export HOST=$(aws emr describe-cluster --cluster-id $ID | jq '.Cluster.MasterPublicDnsName' | tr
 -d '"')
sh-4.2$
sh-4.2$ ssh -i ~/EMRKey-lab.pem hadoop@$HOST
The authenticity of host 'ec2-52-13-122-212.us-west-2.compute.amazonaws.com (10.1.12.161)' can't be esta
blished.
ECDSA key fingerprint is SHA256:TBvH+Ort6MkfJ6MJ5+CIs6xln61aQAJVhkOify6c/VM.
ECDSA key fingerprint is MD5:21:81:c9:ea:e2:bc:3b:e7:f3:a3:10:ac:e0:cc:37:7f.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-52-13-122-212.us-west-2.compute.amazonaws.com,10.1.12.161' (ECDSA) to th
e list of known hosts.
Last login: Tue Jan 13 10:05:20 2026
       __|  __|_  )
       _|  (     /   Amazon Linux 2
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
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-10-1-12-161 ~]$ colmena
-bash: colmena: command not found
[hadoop@ip-10-1-12-161 ~]$ hive
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> SELECT
    >   os,
    >   COUNT(*) count
    > FROM cloudfront_logs
    > WHERE dateobject
    > BETWEEN '2014-07-05' AND '2014-08-05'
    > GROUP BY os;
Query ID = hadoop_20260113101954_d9afe82d-9ecc-4f93-9786-ea027334731b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768296653571_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: 30.46 s
--------------------------------------------------------------------------------
OK
Linux   813
MacOS   852
OSX     799
iOS     794
Android 855
Windows 883
Time taken: 38.838 seconds, Fetched: 6 row(s)
```


Amazon EMR > EMR en EC2: Clústeres
Amazon EMR
EMR sin servidor
EMR en EC2
Clústeres
Blocs de notas y repositorios de Git
Eventos
Bloquear el acceso público
Configuraciones de seguridad
EMR en EKS
Clústeres virtuales
EMR Studio
Introducción
Clústeres (1/1) Información
Ver detalles
Terminar
Clonar
Crear clúster
Filtrar clústeres por estado
Buscar clústeres
Filtrar clústeres por fecha y hora de creación
1
ID del clúster
Nombre del clúster
Estado
j-2E5MV6CPYDV2W
My cluster
Terminar
Solicitud de u


aws training and certification
Español (España)
Vaca Rodríguez Brayan Adrián
Analice Big Data con Hadoop
Fin del laboratorio
Abrir consola
Información del laboratorio
Caduca el 14 ene a las 10:05
1 hora
Idiomas disponibles
Valoración
Qué hay de nuevo
Recursos
URL de gestión de sesiones del host de comando
https://us-west-2.console.a
Región
us-west-2
Contenido del laboratorio
Descripción general del laboratorio
Iniciar laboratorio
Tarea 1: Crear un bucket de Amazon S3
Tarea 2: Lanzar un clúster de Amazon
los resultados.
Fin del laboratorio
Siga estos pasos para cerrar la consola y finalizar su laboratorio.
1. Regresar a la consola de administración de AWS .
2. En la esquina superior derecha de la página, seleccione AWSLabsUser y luego seleccione Cerrar sesión .
3. Seleccione Finalizar laboratorio y luego confirme que desea finalizar su laboratorio.
Recursos adicionales
Tutorial: Introducción a Amazon EMR
Introducción a AWS
Análisis en AWS
¿Qué es Amazon EMR?
Para obtener más información sobre capacitación y certificación de AWS, consulte https://aws.amazon.com/training/ .
Agradecemos sus comentarios.
Si desea compartir comentarios, sugerencias o correcciones, indíquenos los detalles en nuestro Formulario de contacto de capacitación y certificación de AWS .
¡Buen trabajo! Ayúdanos a mejorar Builder Labs,
Validación del laboratorio
Resultado
1 / 1
Has superado las pruebas de conocimientos.
Enviar resultado
Prueba de conocimientos
Retomar
Superado
Para poder completar este laboratorio, debes obtener un aprobado en la prueba de conocimientos.


Te damos la bienvenida a AWS Skill Builder. Explora las novedades y danos tu opinión
Analyze Big Data with Hadoop (Español de España)
Laboratorio de AWS Builder | 4.0 (3) | 1h | Español (España) +10 más
Estás viendo esta formación como parte de un plan de formación: Data Analytics Learning Plan (includes Labs) (Español de España)
¡Enhorabuena!
Esta formación la finalizaste el January 13, 2026
Volver al plan de formación
Volver al inicio
Detalles
Esquema
Completado
Análisis de Big Data con Hadoop (Español de España)
Incompleto
Comentarios
Incompleto
Logros
Comentarios
Tus comentarios son importantes
En general, ¿cuál tu grado de satisfacción con tu experiencia de aprendizaje?
Guardar y continuar
Ver anterior
Ver siguiente


aws
Hola,
Has completado correctamente el curso Analyze Big Data with Hadoop (Español de España). Haz clic aquí https://skillbuilder.aws/training-activity si lo que quieres es ver tu actividad formativa completa.
Gracias.
AWS Training and Certification
Amazon Web Services, Inc. es una empresa subsidiaria de Amazon.com, Inc. Amazon.com es una marca comercial registrada de Amazon.com, Inc. Este mensaje fue elaborado y distribuido por Amazon Web Services, Inc., 410 Terry Ave. North, Seattle, WA 98109-5210.
Msg ID:3262d00d-059f-4d9a-9df2-77348fed2642

CONTENIDO

APARTADO B

INTRODUCCIÓN

Realiza el laboratorio:

Big Data

Realiza el apartado:

Se trabajan con datos de Google Ngram:

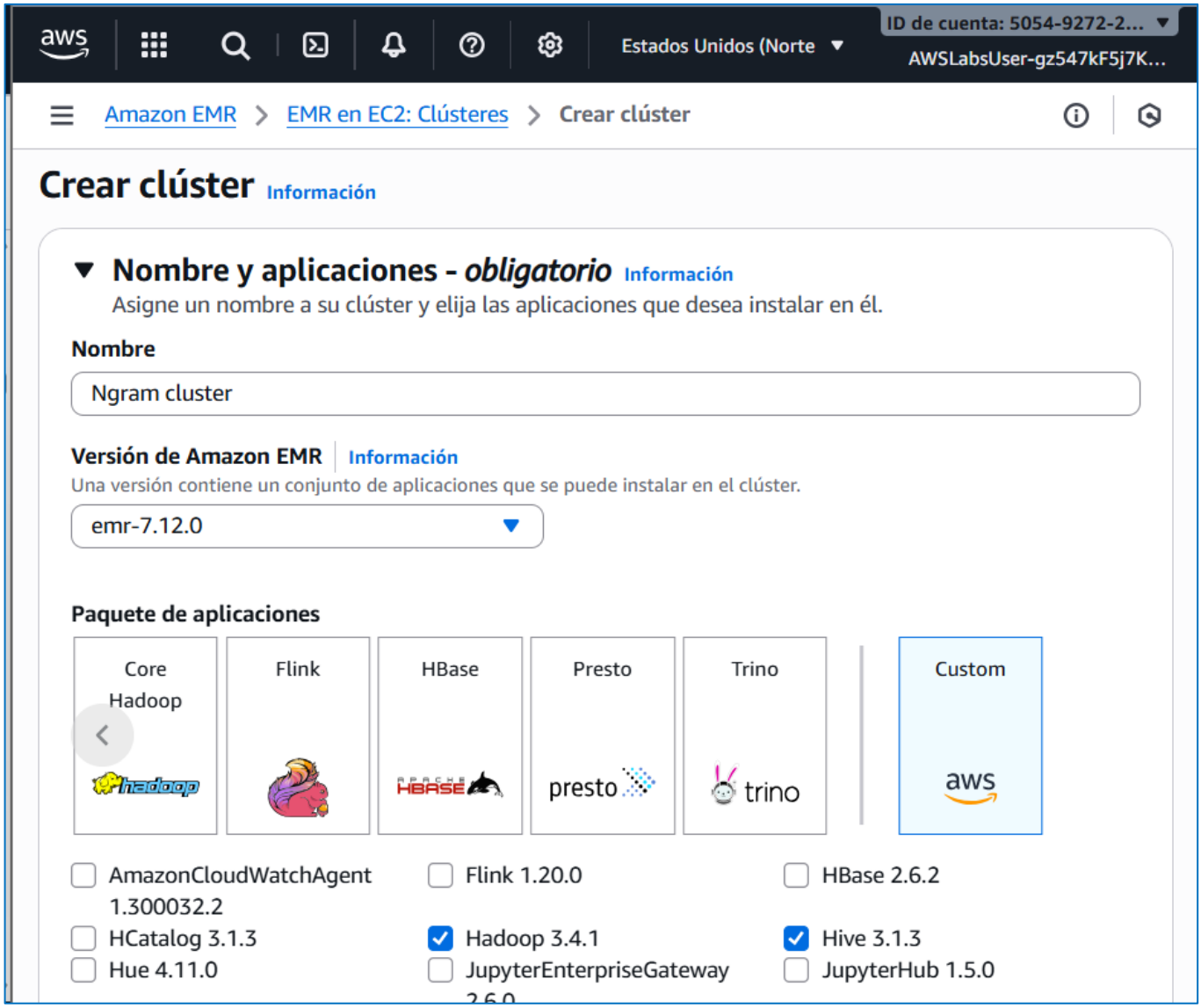
Google Ngram Viewer: Albert Einstein,Sherlock Holmes,Frankenstein

El *bucket* de Amazon S3 `s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/` contiene un conjunto de datos que es parte de un proyecto de análisis de n-gramas en libros en inglés. Este proyecto se utiliza principalmente para el análisis lingüístico y de texto a gran escala. A continuación, te doy una idea más clara de lo que puedes encontrar en ese *bucket*:

Este *bucket* en particular contiene archivos relacionados con **n-gramas de un solo término (1-gram)** de libros en inglés, recopilados en grandes cantidades. En particular, este conjunto de datos contiene un millón de palabras o términos (de ahí el nombre eng-1M).

https://lab.builder-labs.skillbuilder.aws/sa/lab/arn%3Aaws%3Alearningcontent%3Aus-east-1%3A470679935125%3Ablueprintversion%2FSPL-DD-300-ANGNGR-4%3A4.0.24-b2a8b332/es-ES

Realiza este laboratorio creando el clúster EMR desde AWS ACADEMY (en estes caso no uses el entorno que te propone el curso)


aws
Estados Unidos (Norte
ID de cuenta: 5054-9272-2...
AWSLabsUser-gz547kF5j7K...
Amazon EMR
EMR en EC2: Clústeres
Crear clúster
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
Core Hadoop
Flink
HBase
Presto
Trino
Custom
AmazonCloudWatchAgent 1.300032.2
Flink 1.20.0
HBase 2.6.2
HCatalog 3.1.3
Hadoop 3.4.1
Hive 3.1.3
Hue 4.11.0
JupyterEnterpriseGateway
JupyterHub 1.5.0

Grupos de instancias uniformes
Principal (m4.large), Central (m4.large), Tarea (m4.large)

Aprovisionamiento y escalado de clústeres - *obligatorio*

Configuración de aprovisionamiento
Tamaño del núcleo: 1 instancia
Tamaño de la tarea: 1 instancia

Redes - *obligatorio*

VPC
vpc-084ee45e9...

Subred
subnet-0058f9...

Grupos de seguridad de nodos principales
sg-07d010b53f...

Grupos de seguridad de nodos básicos
sg-0710a72ebb...

Cancelar | Crear clúster

Terminación del clúster
Terminar el clúster después del tiempo de inactividad
Tiempo de inactividad: 1 hora

Registros de clúster

Ubicación de Amazon S3
s3://aws-logs...

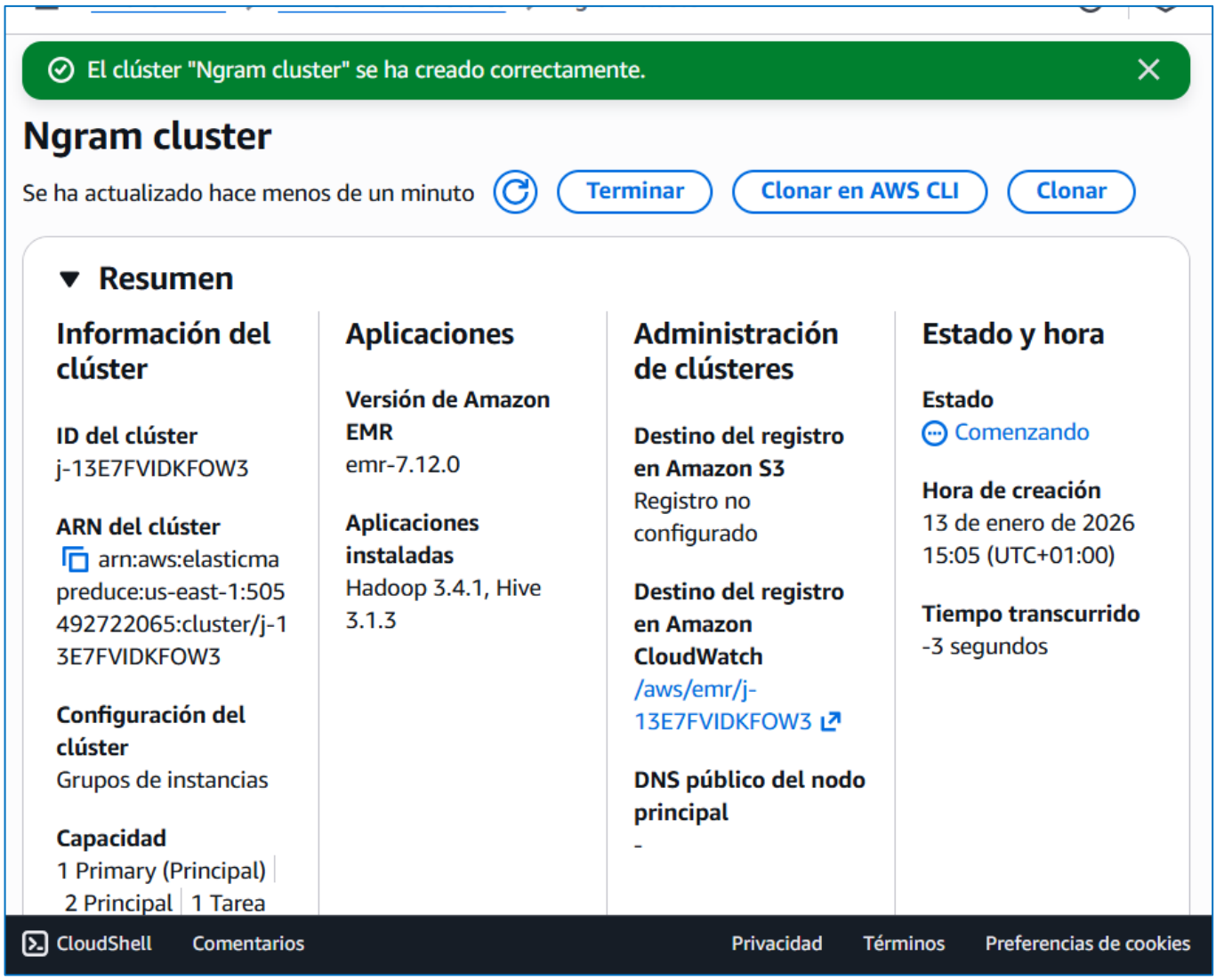
Configuración de seguridad y par de claves de EC2

Par de claves de Amazon EC2
MiClaveEMR

Roles de Identity and Access Management (IAM) - *obligatorio*

Rol de servicio
EMR_DefaultRole

Perfil de instancia
EMR_EC2_DefaultRole




```
PS E:\2025 AI IA\BIG DATA\GitHub\BigData2526\PR_07.4_EMR> ssh -i ~/MiClaveEMR.pem hadoop@ec2-18-206-83-142.compute-1.amazonaws.com
```

```
The authenticity of host '10.1.21.225 (10.1.21.225)' can't be established.
ECDSA key fingerprint is SHA256:CJ2qHyZNtgik09+bcQk20yfiTUzNSX5HoWRYBIMKHiE.
ECDSA key fingerprint is MD5:28:86:17:90:9c:37:c4:7a:ed:f5:65:b3:a6:e5:62:68.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '10.1.21.225' (ECDSA) to the list of known hosts.
   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'

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-10-1-21-225 ~]$
```

```
[hadoop@ip-10-1-21-225 ~]$ hive
Hive Session ID = b2eaaaca-ed1f-4992-b1a6-c02581f3c2c7

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.prope
rties Async: true
hive> CREATE EXTERNAL TABLE ngrams
    > (gram string, year int, occurrences bigint, pages bigint, books bigint)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
    > STORED AS SEQUENCEFILE
    > LOCATION 's3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/';
OK
Time taken: 6.148 seconds
hive>
```

```
hive> DESCRIBE ngrams;
OK
gram                    string
year                    int
occurrences             bigint
pages                   bigint
books                   bigint
Time taken: 0.199 seconds, Fetched: 5 row(s)
hive>
```

```
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: 1.815 seconds, Fetched: 10 row(s)
```

```
hive> INSERT OVERWRITE TABLE normalized
    > SELECT lower(gram), year, occurrences
    > FROM ngrams
    > WHERE year BETWEEN 1990 AND 2005
    > AND gram REGEXP "^[A-Za-z+\'-]{3,}$";
Query ID = hadoop_20260113141705_73f85b72-6300-4671-9995-a1a060f22b4d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0001
)

--------------------------------------------------------------------------------
---------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
 KILLED
--------------------------------------------------------------------------------
---------
Map 1            container  INITIALIZING     -1          0        0       -1       0
--------------------------------------------------------------------------------
---------         container        INITED      1          0        0        1       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
 KILLED  ------------------------------------------------------------------------
---------         container       RUNNING      1          0        1        0       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED
 KILLED  ------------------------------------------------------------------------
--------------------------------------------------------------------------------
--------- 00/02  [==========================>>---] 90%   ELAPSED TIME: 185.05 s
Map 1 .......... container     SUCCEEDED     21         21        0        0       0
      0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0
      0
--------------------------------------------------------------------------------
---------
VERTICES: 02/02  [============================>>] 100%  ELAPSED TIME: 185.65 s
--------------------------------------------------------------------------------
---------
Loading data to table default.normalized
OK
Time taken: 189.867 seconds
```

```
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: 0.119 seconds, Fetched: 20 row(s)
```

```
hive> SELECT
    >   gram,
    >   sum(occurrences) as total_occurrences
    > FROM normalized
    > GROUP BY gram
    > ORDER BY total_occurrences DESC
    > LIMIT 50;
Query ID = hadoop_20260113142635_de9e4683-b1f5-4c2c-bf18-d3dccbecbd7a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0001
)
```

```
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_20260113142757_48f40960-f5e2-47c2-abfc-1b7834f0cb16
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0001
)
```

```
hive> INSERT OVERWRITE TABLE ratios
    > SELECT
    >   a.gram,
    >   a.year,
    >   sum(a.occurrences) AS occurrences,
    >   sum(a.occurrences) / b.total 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_20260113143122_39edd4f0-d64f-445d-935f-b44f287825b1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0001
)
```

```
hive> -- Occurrences of 'internet' in books by year?
hive> SELECT
    >   year,
    >   occurrences
    > FROM ratios
    > WHERE gram = 'internet'
    > ORDER BY year;
Query ID = hadoop_20260113144014_2f11a04c-000b-4637-883f-4f84bd1dcceb
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_1768313387298_0002
)
```

```
hive> -- Most popular words of each length
hive> SELECT DISTINCT length, gram
    > FROM
    > (
    >   SELECT length(gram) AS length,
    >   gram,
    >   rank() OVER (partition by length(gram) order by occurrences desc) AS rank
    >   FROM ratios
    > ) x
    > WHERE rank = 1
    > ORDER BY length;
Query ID = hadoop_20260113144125_6393a2b6-0e54-48eb-ae3c-f78358686e2e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0002
)
```

```
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      oooooooooooooooooooooooooo
28      trimethoprimsulfamethoxazole
29      methylenedioxymethamphetamine
30      dipalmitoylphosphatidylcholine
31      dichlorodiphenyltrichloroethane
32      oooooooooooooooooooooooooooooo
33      ooooooooooooooooooooooooooooooo
34      oooooooooooooooooooooooooooooooo
35      ooooooooooooooooooooooooooooooooo
36      oooooooooooooooooooooooooooooooooo
Time taken: 42.576 seconds, Fetched: 34 row(s)
hive>
```

1.- ¿Qué contiene el *bucket* s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/?

El **visor de Ngramas de Google Books** es un motor de búsqueda en línea que grafica las frecuencias de cualquier conjunto de cadenas de búsqueda utilizando un recuento anual de *n* -gramas encontrados en fuentes impresas publicadas entre 1520 y 2008

¿Cuánto ocupa el archivo que contiene?
Tiene un tamaño aproximado de 2600 MB

2.- ¿Cuántos registros contiene la tabla **`ngrams`** que creaste en HIVE?

```
hive> SELECT COUNT(*) AS total_registros
    > FROM ngrams;
Query ID = hadoop_20260113145053_fe3f477b-86ca-4185-91c5-8ce85040e304
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768313387298_0003
)
```

```
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 189.66 s
--------------------------------------------------------------------------------
---------
OK
261823186
Time taken: 193.438 seconds, Fetched: 1 row(s)
hive>
```

Se compone de 261823186 registros

¿Desde qué año hasta qué año abarca la información que contiene?

```
hive> SELECT
    >   MIN(year) AS primer_anio,
    >   MAX(year) AS ultimo_anio
    > FROM ngrams;
---------
OK
1520    2008
Time taken: 185.117 seconds, Fetched: 1 row(s)
```

Va desde 1520 hasta 2008

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

```
hive> CREATE TABLE normalized AS
    > SELECT
    >   LOWER(gram) AS gram,
    >   year,
    >   match_count,
    >   page_count,
    >   book_count
    > FROM ngrams
    > WHERE gram REGEXP '^[A-Za-z+\'-]{3,}$';
```

Parte	Significado
^	Inicio de la palabra
[A-Za-z+\'-]	Letras mayúsculas/minúsculas, +, ', -
{3,}	Al menos 3 caracteres
$	Fin de la palabra

¿ Cuántos registros contiene la tabla **`normalized`**?

```
hive> SELECT COUNT(*) AS total_registros_normalized
    > FROM normalized;
OK
20803439
```

Tiene 20803439 registros

Hola,

Has completado correctamente el curso Exploring Google Ngrams with Amazon EMR and Hive (Español de España). Haz clic aquí https://skillbuilder.aws/training-activity si lo que quieres es ver tu actividad formativa completa.

Gracias.

AWS Training and Certification

Msg ID:08e6c18e-2c37-4e48-a24c-6acba61c3cac