

# 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

Welcome to AWS Skill Builder. Explore [what's new](#) and [tell us how we're doing](#)

### Analyze Big Data with Hadoop (Español de España)

AWS Builder Lab | ★ 4.0 (3) | 1h | Español (España) [+10 more](#)

You're viewing this training as part of a learning plan: [Data Analytics Learning Plan \(includes Labs\) \(Español de España\)](#)

Details

Outline

Not started

●

Análisis de Big Data con Hadoop (Español de España)

View only

●

Feedback

●

Achievements

Análisis de Big Data con Hadoop (Español de España)

Start ↗

View next

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

Iniciar laboratorio

### Analyze Big Data with Hadoop

SPL-166 - Version 1.0.27

© 2025 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at [AWS Training and Certification](#).

#### Lab overview

In this lab, you deploy a fully functional Hadoop cluster, ready to analyze log data in just a few minutes. You start by launching an Amazon EMR cluster and then use a HiveQL script to process sample log data stored in an Amazon Simple Storage Service (Amazon S3) bucket. *HiveQL* is a SQL-like scripting language for data warehousing and analysis. You can then use a similar setup to analyze your own log files.

[Learn more](#): Refer to *Tutorial: Getting started with Amazon EMR* in the **Additional resources** section for more information.

#### Amazon EMR Overview

De 26 de Agosto de 2025

Presentar

#### Objectives

Validación del laboratorio >

Este laboratorio incluye pruebas de conocimientos diseñadas para evaluar tu comprensión del material tratado. Inicia el laboratorio, completa las tareas y, a continuación, inicia las pruebas de conocimientos y responde las preguntas que se incluyen en las evaluaciones.


#### Prueba de conocimientos

Iniciar ↗

Para poder completar este laboratorio, debes obtener un aprobado en la prueba de conocimientos.

### Task 1: Create an Amazon S3 bucket

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


- At the top of the AWS Management Console, in the search bar, search for and choose  **S3**.
- 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.

- At the top of the AWS Management Console, in the search bar, search for and choose  **EMR**.
- 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**.
- Choose **Create cluster**.
- 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”

End Lab

Abrir consola

El laboratorio está preparado.  
Abre la consola para empezar. Mantén la región predeterminada. Tu laboratorio estará activo hasta el 14 ene a las 8:58  
Consejo: abre la consola en una ventana nueva para verla junto a estas instrucciones.


Validación del laboratorio

Resultado

## 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.


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

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

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

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

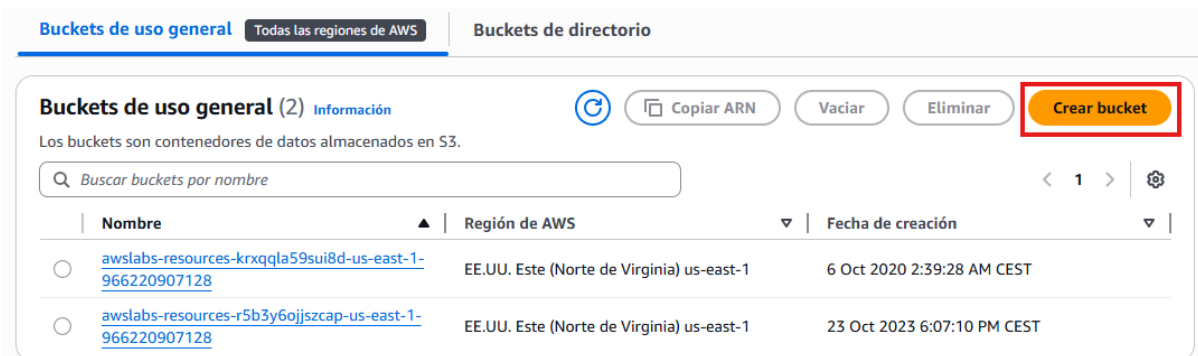
- Choose **Create bucket**.

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

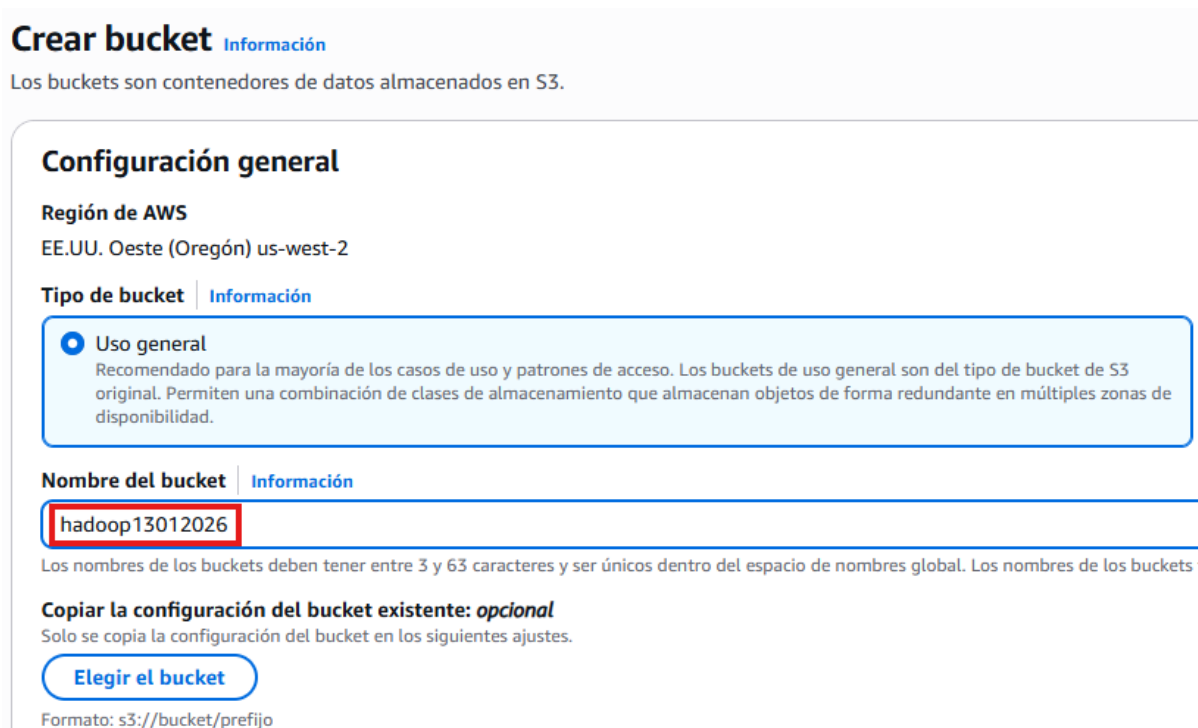
El primero paso es buscar “S3” en la búsqueda de AWS



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

El bucket "hadoop13012026" se creó correctamente  
Para cargar archivos y carpetas, o para configurar ajustes adicionales del bucket, elija [Ver detalles](#).

Buckets de uso general Todas las regiones de AWS Buckets de directorio

**Buckets de uso general (3)** Información   Copiar ARN  Vaciar  Eliminar  Crear bucket

Los buckets son contenedores de datos almacenados en S3.

Buscar buckets por nombre

	Nombre	Región de AWS	Fecha de creación
<input type="radio"/>	<a href="#">awslabs-resources-krxqqla59sui8d-us-east-1-966220907128</a>	EE.UU. Este (Norte de Virginia) us-east-1	6 Oct 2020 2:39:28 AM CEST
<input type="radio"/>	<a href="#">awslabs-resources-r5b3y6ojjszcap-us-east-1-966220907128</a>	EE.UU. Este (Norte de Virginia) us-east-1	23 Oct 2023 6:07:10 PM CEST
<input type="radio"/>	<a href="#">hadoop13012026</a>	EE.UU. Oeste (Oregón) us-west-2	13 Jan 2026 9:08:03 AM CET

## Tarea 2

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

aws    Preguntar a Amazon Q 

 Amazon

**Servicios** [Ver los 11 resultados](#)

**Servicios (11)**

- Características (3)
- Publicaciones de blog (20)
- Documentación (1978)
- Artículos de conocimiento (93)
- Tutoriales (1)
- Marketplace (992)

 **EMR**   
Marco Hadoop administrado


 **Amazon SageMaker**   
El centro de datos, análisis e inteligencia artificial (IA)

 **AWS Glue DataBrew**   
Herramienta de preparación de datos visuales para limpiar y normalizar datos para a...

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

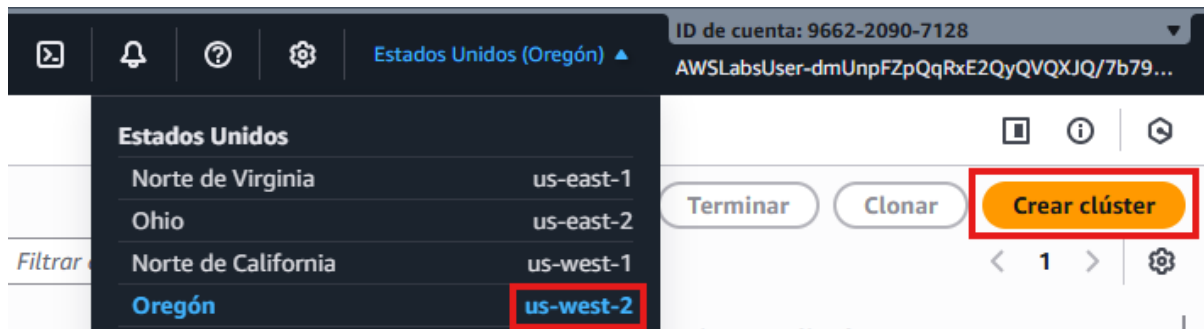
## Recursos

CommandHostSessionManagementUrl



Region





Ahora tenemos que configurar la siguiente sección, “Nombre y aplicaciones”

▼ **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

☐ Flink 1.14.2 ☐ Ganglia 3.7.2 ☐ HBase 1.4.13

☐ HCatalog 2.3.9 ☒ Hadoop 2.10.1 ☒ Hive 2.3.9

☒ Hue 4.10.0 ☐ JupyterEnterpriseGateway 2.6.0 ☐ JupyterHub 1.4.1

☐ Livy 0.7.1 ☐ MXNet 1.8.0 ☐ Mahout 0.13.0

☐ Oozie 5.2.1 ☐ Phoenix 4.14.3 ☒ Pig 0.17.0

☐ Presto 0.267 ☐ Spark 2.4.8 ☐ Sqoop 1.4.7

☐ TensorFlow 2.4.1 ☐ Tez 0.9.2 ☐ Zeppelin 0.10.0

☐ ZooKeeper 3.4.14

Ahora tenemos que configurar la sección de “Configuración del clúster”

## ▼ 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*

### Central

#### 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 ▼

## ► Configuración de nodo - *opcional*

### Tarea 1 de 1

#### Nombre

Tarea - 1

#### 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 ▼

## ► Configuración de nodo - *opcional*

Ahora configuramos la sección de “Redes”

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

Elegir VPC

VPC (2)

Filtrar VPC

< 1 >

Nombre	ID de VPC	Condición	CIDR IPv4	CIDR IPv6
<input checked="" type="radio"/> Lab VPC	<a href="#">vpc-0c43b4d9ccae09611</a>	✓ Disponible	10.1.0.0/16	
<input type="radio"/> -	<a href="#">vpc-0a3f7acb3923059f9</a>	✓ Disponible	172.31.0.0/16	

Cancelar

Elegir

Ahora le damos al desplegable y seleccionamos esta opción

LabStack-7b790a58-2210-4960-8830-1da0f6...  
sg-04e78bdfd058e096a

Elegir grupos

Q |

Crear ElasticMapReduce-Primary

default  
sg-0af83bd918220602d

GuardDutyManagedSecurityGroup-vpc-0c43b4d9ccae09611  
sg-0325f82b3ec3ea9f9

LabStack-7b790a58-2210-4960-8830-1da0f6d43754-  
dmUnpFZpQqRxE2QyQVQXJQ-0-CommandHostSG-c7NLGbY2HsYb  
sg-05cf206e80cb0f5c4

LabStack-7b790a58-2210-4960-8830-1da0f6d43754-  
dmUnpFZpQqRxE2QyQVQXJQ-0-EmrSecurityGroup-  
phzCdxAzGTxU  
sg-04e78bdfd058e096a

✓

Nos tendría que quedar tal que así



▼ **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-0c43b4d9ccae09611

[Examinar](#)

[Crear VPC](#) [↗](#)

**Subred** [Información](#)

subnet-012b90529429fc98a

[Examinar](#)

[Crear subred](#) [↗](#)

▼ **Grupos de seguridad de EC2 (firewall)**

**Nodo principal**

**Grupos de seguridad administrados de EMR**

EMR actualizará automáticamente el grupo seleccionado.

LabStack-7b790a58-2210-4960-8830-1da0f6...  
sg-04e78bdfd058e096a ▼

**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.

LabStack-7b790a58-2210-4960-8830-1da0f6...  
sg-04e78bdfd058e096a ▼

**Grupos de seguridad adicionales - *opcional***

Seleccione hasta 4 grupos de seguridad adicionales.

[Elegir grupos de seguridad adicionales](#) ▼

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

### ▼ Terminación del clúster y reemplazo de nodos [Información](#)

Elija la configuración de terminación y proteja su clúster contra un apagado accidental.

#### Opción de terminación

- ☐ Terminar manualmente el clúster
- ☐ Terminar automáticamente el clúster después de que finalice el último paso
- ☒ Terminar el clúster después del tiempo de inactividad (recomendado)

#### Tiempo de inactividad

Ingrese el tiempo hasta que el clúster termine.

0 días ▼

01:00:00

Elija una hora mayor a 1 minuto (00:01:00) y menor a 7 días. La hora está en formato hh:mm:ss (24 horas).

#### ☒ Use la protección contra la terminación

Protege al clúster para evitar una terminación accidental. Si está activada, deberá primero desactivar la protección para terminar el clúster. Recomendamos activar la protección frente a terminaciones para los clústeres de larga duración.

**i** Para garantizar que el reemplazo de nodos en mal estado no afecte a sus flujos de trabajo actuales en las versiones 7.0.0 y anteriores de EMR, la desactivamos cuando activa la protección de terminación. Puede cambiar esta configuración al crear un clúster o yendo a la configuración del clúster.

#### Reemplazo de nodos en mal estado - *novedad* | [Información](#)

- ☐ Activar  
Amazon EMR detiene correctamente los procesos en los nodos en mal estado para minimizar la pérdida de datos y las interrupciones del trabajo. Reemplaza rápidamente los nodos en mal estado por nuevas instancias de EC2 para que sus trabajos funcionen sin problemas.
- ☒ Desactivar  
Amazon EMR agrega los nodos en mal estado a una lista de denegación mientras los mantiene en el clúster, lo que le permite tener acceso continuo para solucionar problemas.

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

## Elegir la ubicación de Amazon S3

### Buckets de S3

#### Buckets (1/3)

🔍 *Buscar bucket*

| Nombre

☐ awslabs-resources-krxqqla59sui8d-us-east-1-966220907128

☐ awslabs-resources-r5b3y6ojjszcap-us-east-1-966220907128

☒ **hadoop13012026**

Y nos quedaría de esta manera;

**▼ Registros de clúster** [Información](#)  
Elija dónde y cómo almacenar los archivos de registro.

*i* Archivamos automáticamente los archivos de registro en Amazon S3. Puede especificar una ubicación de S3 propia o utilizar la ubicación de S3 predeterminada para Amazon EMR. La ubicación de registro predeterminada se completa previamente en el campo **Ubicación de Amazon S3**.

☒ Publicar registros específicos del clúster en Amazon S3

**Ubicación de Amazon S3**

[Ver ↗](#) [Explorar S3](#)

Formato: utilizar s3://bucket/prefix

☐ Cifrar los registros específicos del clúster

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”

**Elegir una clave de Amazon EC2 para el protocolo SSH al clúster** ✕

**Pares de claves (1)** 🔄

ID	Name	Fingerprint
<input checked="" type="radio"/> key-010694b6f51c9d032	EMRKey-lab	fc:98:74:f8:11:7d:f0:99:20:d3:0f:bb:c 4:76:15:88:9a:40:79:c4

[Cancelar](#) [Elegir](#)

Nos quedará de la siguiente manera;

**▼ 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.

[🔄](#) [Examinar ↗](#) [Crear configuración de seguridad ↗](#)

---

**Par de claves de Amazon EC2 para el protocolo SSH al clúster** [Información](#)

[✕](#) [Examinar](#) [Crear par de claves ↗](#)

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

## ▼ Roles de Identity and Access Management (IAM) - obligatorio [Información](#)

Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

### Rol de servicio de Amazon EMR [Información](#)

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

- ☒ **Elegir un rol de servicio existente**  
Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

- ☐ **Crear un rol de servicio**  
Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

#### Rol de servicio

EMR\_DefaultRole



### Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

- ☒ **Elegir un perfil de instancia existente**  
Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

- ☐ **Crear un perfil de instancia**  
Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

#### Perfil de instancia

EMR\_EC2\_DefaultRole



Finalmente, le daremos al botón de “Crear Cluster”

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

Cancelar

Crear clúster

El clúster "My cluster" se ha creado correctamente.

Se ha actualizado hace 3 minutos

Terminar

Clonar en AWS CLI

Clonar

My cluster

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

▼ Resumen

Información del clúster

ID del clúster  
j-3U7W2PFNGZBCS

ARN del clúster  
arn:aws:elasticmapreduce:us-west-2:966220907128:cluster/j-3U7W2PFNGZBCS

Configuración del clúster

Grupos de instancias

Capacidad  
1 Primary (Principal) | 1 Principal | 1 Tarea

Aplicaciones

Versión de Amazon EMR  
emr-5.36.1

Aplicaciones instaladas  
Hadoop 2.10.1, Hive 2.3.9, Hue 4.10.0, Pig 0.17.0

Administración de clústeres

Destino del registro en Amazon S3  
hadoop13012026

IU de aplicación persistente  
Servidor de línea de tiempo de YARN  
UI de Tez

DNS público del nodo principal  
ec2-44-255-123-74.us-west-2.compute.amazonaws.com  
Conectarse al nodo principal mediante SSH  
Conectarse al nodo principal mediante SSM

Estado y hora

Estado  
Esperando

Hora de creación  
13 de enero de 2026 9:41 (UTC+01:00)

Tiempo transcurrido  
12 minutos, 35 segundos

## Tarea 3

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

Propiedades | Acciones de arranque | Instancias (hardware) | **Pasos** | Aplicaciones | Configuraciones | Monitorización | Eventos | Etiquetas (0)

**Pasos (0)** [Información](#) [Actualizar tabla](#) [Cancelar pasos](#) [Clonar paso](#) [Agregar paso](#)

Cada paso es una unidad de trabajo que contiene instrucciones para manipular los datos para su procesamiento por software instalado en el clúster.

Pasos simultáneos: 1 [🔗](#)

Filtrar pasos por estado [▼](#)

ID de paso	Estado	Nombre	Archivos de registro <a href="#">🔗</a>	Hora de creación (UTC+01:00)	Hora de inicio (UTC+01:00)	Tiempo transcurrido
No hay coincidencias						
No se encuentra ninguna coincidencia						

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

### 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 Hive**  
Agrega un paso que envía un script de Hive para las interacciones de almacenamiento de datos.

☐ Script de shell  
Soluciona los problemas que se presentan con el clúster.

☐ 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 Pig  
Agrega un paso que envía un script de Pig para analizar conjuntos de datos de gran tamaño.

**Nombre**

**Ubicación del script de Hive**  
La ubicación Amazon S3 del script de Hive.  
 [Ver 🔗](#) [Explorar S3](#)

**Entrada de la ubicación de Amazon S3 - opcional**  
La ubicación Amazon S3 de los archivos de entrada de Hive.  
 [Ver 🔗](#) [Explorar S3](#)

**Salida de la ubicación de Amazon S3 - opcional**  
La ubicación Amazon S3 de los archivos de salida de Hive.  
 [Ver 🔗](#) [Explorar S3](#)

**Argumentos - opcional** [Información](#)  
Especifique los argumentos opcionales para su script.

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

**Pasos (1)** [Información](#) [Actualizar tabla](#) [Cancelar pasos](#) [Clonar paso](#) [Agregar paso](#)

Cada paso es una unidad de trabajo que contiene instrucciones para manipular los datos para su procesamiento por software instalado en el clúster.

Pasos simultáneos: 1 [🔗](#)

Filtrar pasos por estado [▼](#)

ID de paso	Estado	Nombre	Archivos de registro <a href="#">🔗</a>	Hora de creación (UTC+01:00)	Hora de inicio (UTC+01:00)	Tiempo transcurrido
<input type="checkbox"/> <a href="#">🔗</a> s-06449872E25XMPFF543Y	Completed	Process logs	No se han creado registros aún <a href="#">🔗</a>	13 de enero de 2026, 10:00	13 de enero de 2026, 10:00	1 minuto, 34 segundos

## Tarea 4






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

aws  Preguntar a Amazon Q [✕](#)

**Servicios** [Ver los 9 resultados](#)

- Servicios (9)**
  - Características (40)
  - Publicaciones de blog (35)
  - Documentación (3029)
  - Artículos de conocimiento (280)
  - Tutoriales (7)
  - Marketplace (3131)
- S3**  
Almacenamiento escalable en la nube
- S3 Glacier**  
Almacenamiento de archivos en la nube
- AWS Snow Family**  
Transporte de datos a gran escala




## Una vez dentro, hacemos clic en nuestro bucket

**Buckets de uso general (3)** [Información](#)   Copiar ARN  Vaciar  Eliminar  Crear bucket





Los buckets son contenedores de datos almacenados en S3.

	Nombre	Región de AWS	Fecha de creación
<input type="radio"/>	<a href="#">awslabs-resources-krxqqla59sui8d-us-east-1-966220907128</a>	EE.UU. Este (Norte de Virginia) us-east-1	6 Oct 2020 2:39:28 AM CEST
<input type="radio"/>	<a href="#">awslabs-resources-r5b3y6ojjszcap-us-east-1-966220907128</a>	EE.UU. Este (Norte de Virginia) us-east-1	23 Oct 2023 6:07:10 PM CEST
<input type="radio"/>	<b>hadoop13012026</b>	EE.UU. Oeste (Oregón) us-west-2	13 Jan 2026 9:08:03 AM CET



## Entramos en la siguiente carpeta

<input type="checkbox"/>	Nombre	Tipo
<input type="checkbox"/>	 <a href="#">j-3U7W2PFNGZBC3/</a>	Carpeta
<input type="checkbox"/>	 <a href="#">os_requests_\$folder\$</a>	-
<input type="checkbox"/>	 <b><a href="#">os_requests/</a></b>	Carpeta


## Descargamos los archivos

**Objetos (1/2)**   Copiar URI de S3  Copiar URL  **Descargar**

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de t que concederles permisos de forma explícita. [Más información](#)


<input type="checkbox"/>	Nombre	Tipo	Última modificación
<input checked="" type="checkbox"/>	 <b>000000_0</b>	-	13 Jan 2026 10:02:07 AM CET
<input type="checkbox"/>	 <b>000001_0</b>	-	13 Jan 2026 10:02:07 AM CET

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

 **000000\_0**

Archivo Editar Ver

Linux@813  
MacOS@852  
OSX@799  
iOS@794

 **000001\_0**

Archivo Editar Ver

Android@855  
Windows@883

## Tarea 5

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





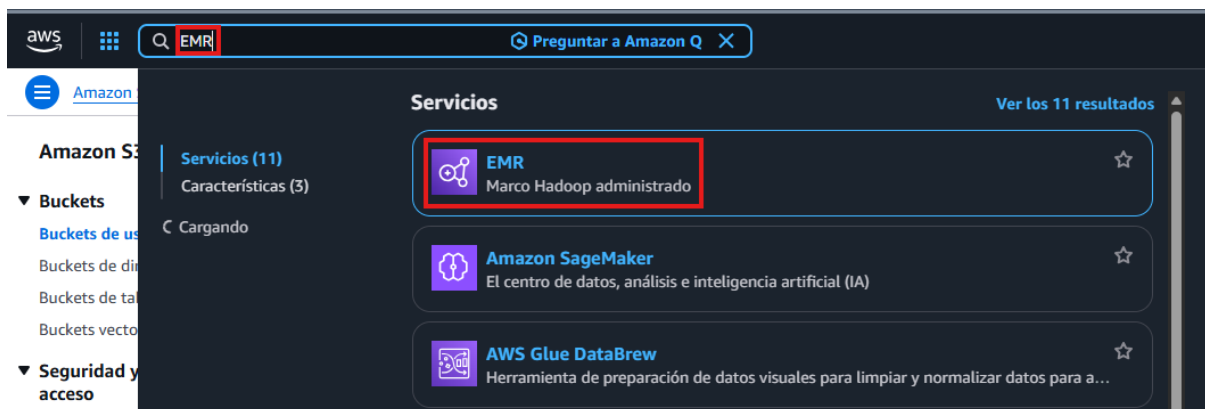
```
hive> SELECT os, COUNT(*) count FROM cloudfront_logs WHERE dateobject BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY os;
Query ID = hadoop_20260113091916_cb509209-t5db-4525-acd9-21182dc0e9a8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768294025176_0002)
```

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

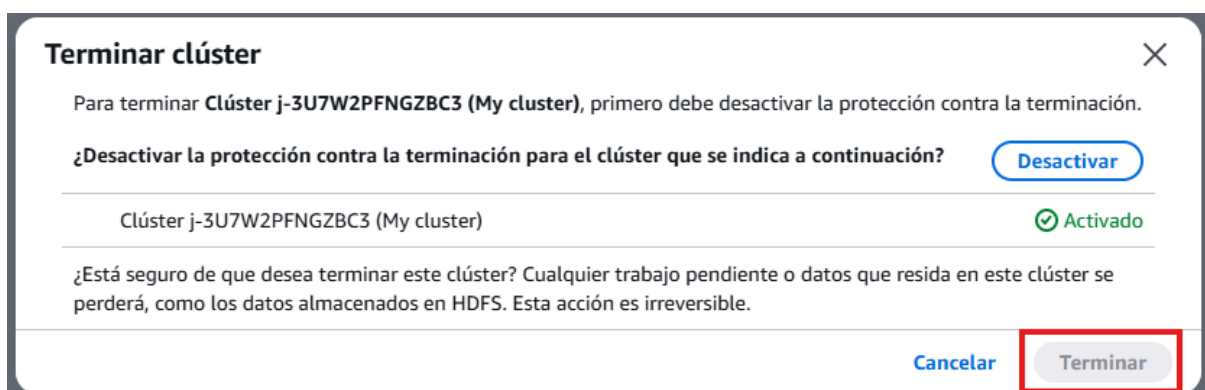
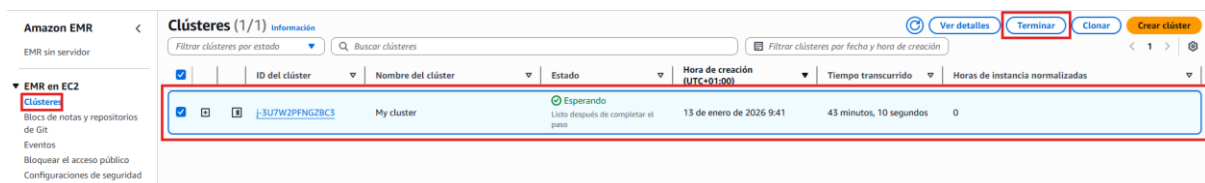
```
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 29.90 s
OK
Linux 813
MacOS 852
OSX 799
iOS 794
Android 855
Windows 883
Time taken: 36.434 seconds, Fetched: 6 row(s)
```

## 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 debido a la protección contra la terminación del cluster, simplemente la desactivamos y le damos a terminar



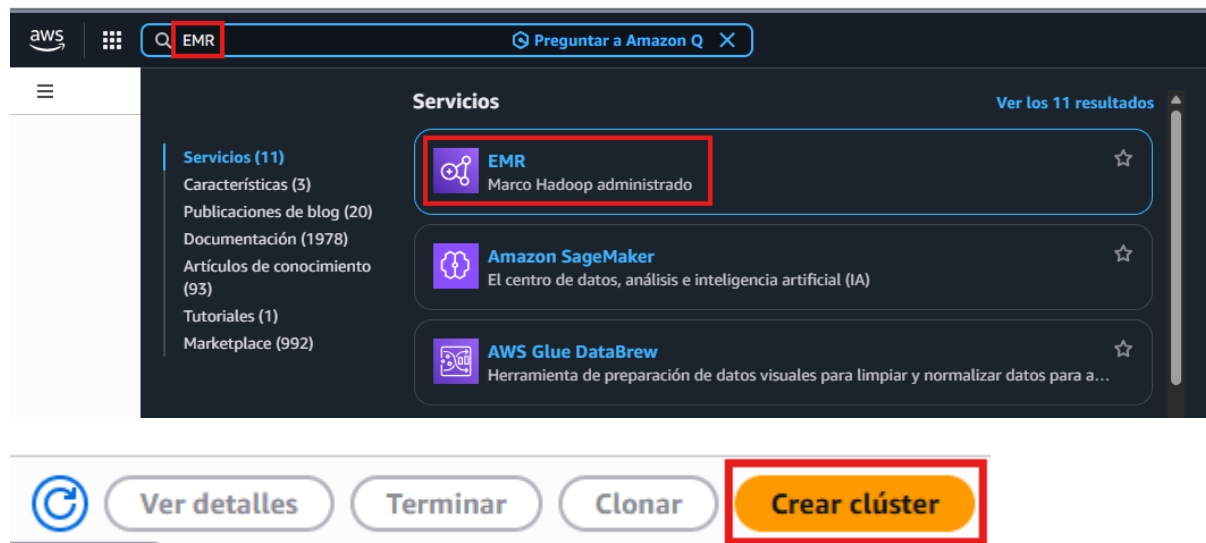
# Apartado B

## 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
---	---	--	--	---	---	---

- |   |   |  |
|---|---|--|
| <input type="checkbox"/> AmazonCloudWatchAgent 1.300032.2 | <input type="checkbox"/> Flink 1.20.0                   | <input type="checkbox"/> HBase 2.6.2           |
| <input type="checkbox"/> HCatalog 3.1.3                   | <input checked="" type="checkbox"/> Hadoop 3.4.1        | <input checked="" type="checkbox"/> Hive 3.1.3 |
| <input type="checkbox"/> Hue 4.11.0                       | <input type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input type="checkbox"/> JupyterHub 1.5.0      |
| <input type="checkbox"/> Livy 0.8.0                       | <input type="checkbox"/> Oozie 5.2.1                    | <input type="checkbox"/> Phoenix 5.2.1         |
| <input type="checkbox"/> Pig 0.17.0                       | <input type="checkbox"/> Presto 0.287                   | <input type="checkbox"/> Spark 3.5.6           |
| <input type="checkbox"/> TensorFlow 2.19.0                | <input type="checkbox"/> Tez 0.10.2                     | <input type="checkbox"/> Trino 476             |
| <input type="checkbox"/> Zeppelin 0.11.1                  | <input type="checkbox"/> ZooKeeper 3.9.3                |  |

### ▼ 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*

## Central

### 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 ▼

### ► Configuración de nodo - *opcional*

## Tarea 1 de 1

Eliminar grupo de instancias

### Nombre

Tarea - 1

### 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 ▼

### ► 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	<input type="checkbox"/>
Tarea - 1	m4.large	1	<input type="checkbox"/>

### ▼ 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 seguridad



Examinar [↗](#)

Crear configuración de seguridad [↗](#)

#### Par de claves de Amazon EC2 para el protocolo SSH al clúster [Información](#)

 emr



Examinar

Crear par de claves [↗](#)

## ▼ Roles de Identity and Access Management (IAM) - obligatorio [Información](#)

Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

### Rol de servicio de Amazon EMR [Información](#)

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

☒ Elegir un rol de servicio existente  
Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

☐ Crear un rol de servicio  
Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

#### Rol de servicio

EMR\_DefaultRole



### Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

☒ Elegir un perfil de instancia existente  
Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

☐ Crear un perfil de instancia  
Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

#### Perfil de instancia

EMR\_EC2\_DefaultRole



## Finalmente le damos al botón de “Crear Cluster”

El clúster "Ngram cluster" se ha creado correctamente.

### Ngram cluster

Se ha actualizado hace menos de un minuto

Terminar

Clonar en AWS CLI

Clonar

#### ▼ Resumen

##### Información del clúster

ID del clúster  
j-18IIBEJ54T217

##### ARN del clúster

arn:aws:elasticmapreduce:us-east-1:716757242964:cluster/j-18IIBEJ54T217

Configuración del clúster  
Grupos de instancias

##### Capacidad

1 Primary (Principal) | 1 Principal | 1 Tarea

##### Aplicaciones

Versión de Amazon EMR  
emr-7.12.0

Aplicaciones instaladas  
Hadoop 3.4.1, Hive 3.1.3

##### Administración de clústeres

Destino del registro en Amazon S3  
Registro no configurado

Destino del registro en Amazon CloudWatch  
[/aws/emr/j-18IIBEJ54T217](#)

UI de aplicación persistente  
[Servidor de línea de tiempo de YARN](#)  
[UI de Tez](#)

DNS público del nodo principal  
[ec2-52-91-135-33.compute-1.amazonaws.com](#)  
[Conectarse al nodo principal mediante SSH](#)  
[Conectarse al nodo principal mediante SSM](#)

##### Estado y hora

Estado  
Esperando

Hora de creación  
14 de enero de 2026 11:15 (UTC+01:00)

Tiempo transcurrido  
9 minutos, 3 segundos

## Tarea 2

### Ahora nos vamos a conectar mediante “ssh”





```
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-8146-7a27ede5b765  
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	0
Reducer 2 .....	container	SUCCEEDED	4	4	0	0	0	0	0
Reducer 3 .....	container	SUCCEEDED	1	1	0	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.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\_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	0
Reducer 5 .....	container	SUCCEEDED	2	2	0	0	0	0	0
Map 1 .....	container	SUCCEEDED	11	11	0	0	0	0	0
Reducer 2 .....	container	SUCCEEDED	4	4	0	0	0	0	0
Reducer 3 .....	container	SUCCEEDED	1	1	0	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-73dalbdc57f2

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	0
Map 1 .....	container	SUCCEEDED	10	10	0	0	0	0	0
Reducer 2 .....	container	SUCCEEDED	1	1	0	0	0	0	0
Reducer 3 .....	container	SUCCEEDED	1	1	0	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\_fe766bct-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	0
Reducer 2 .....	container	SUCCEEDED	1	1	0	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	0
Reducer 2 .....	container	SUCCEEDED	2	2	0	0	0	0	0
Reducer 3 .....	container	SUCCEEDED	6	6	0	0	0	0	0
Reducer 4 .....	container	SUCCEEDED	1	1	0	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 ooooooooooooooooooooooooooooo
28 trimethoprim sulfamethoxazole
29 methylenedioxymethamphetamine
30 dipalmitoylphosphatidylcholine
31 dichlorodiphenyltrichloroethane
32 ooooooooooooooooooooooooooooo
33 ooooooooooooooooooooooooooooo
34 ooooooooooooooooooooooooooooo
35 ooooooooooooooooooooooooooooo
36 ooooooooooooooooooooooooooooo
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	0
Reducer 2 .....	container	SUCCEEDED	1	1	0	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:

```
hive> SELECT MIN(year), MAX(year) FROM ngrams;
Query ID = hadoop_20260114110934_75c9f3d0-9281-45b2-9342-33e9750d4307
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	0
Reducer 2 .....	container	SUCCEEDED	1	1	0	0	0	0	0

VERTICES: 02/02 [=====>>] 100% ELAPSED TIME: 223.21 s

```
OK
1520      2008
Time taken: 223.619 seconds, Fetched: 1 row(s)
```

### 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:

```
hive> SELECT COUNT(*) FROM normalized;
OK
20803439
Time taken: 0.205 seconds, Fetched: 1 row(s)
```