



UNIVERSIDAD PRIVADA DE TACNA
FACULTAD DE INGENIERIA
Escuela Profesional de Ingeniería de Sistemas

INFORME DE LABORATORIO N°03
“Introducción a big data con Amazon EMR”

CURSO:

Base de Datos II

DOCENTE:

Ing. Patrick Jose Cuadros Quiroga

ALUMNO:

Risther Jaime Tarqui Montalico

(2017057469)

Tacna - Perú

2020

Introducción a big data con Amazon EMR

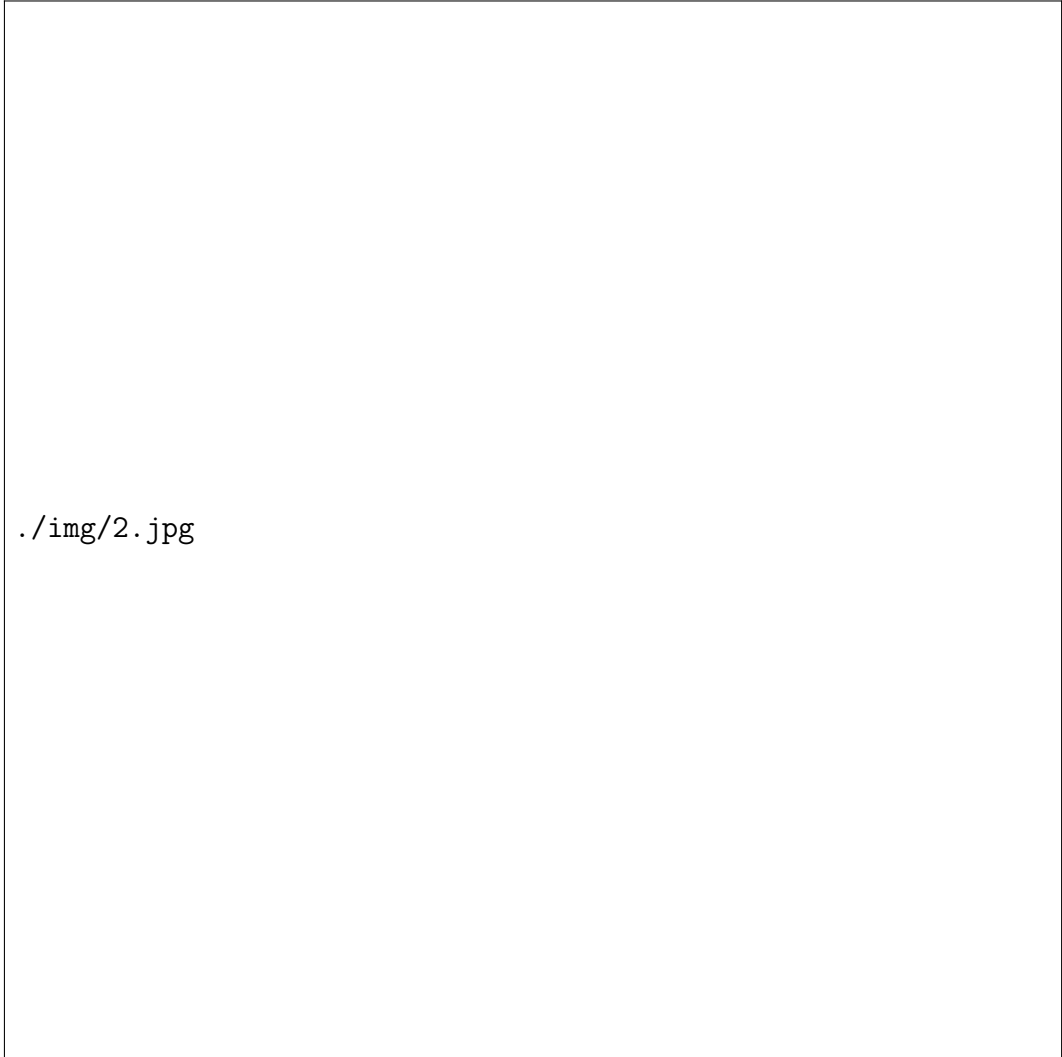
1. OBJETIVO

- Este laboratorio tiene como objetivo guiar a través del proceso de creación de un clúster de Amazon EMR de ejemplo con las opciones de Creación rápida en la Consola de administración de AWS. Después de crear el clúster, enviará un script de Hive como un paso para procesar datos de ejemplo almacenados en Amazon Simple Storage Service (Amazon S3).

2. DESARROLLO

2.1. Paso 1: Configurar los requisitos previos para el clúster de ejemplo

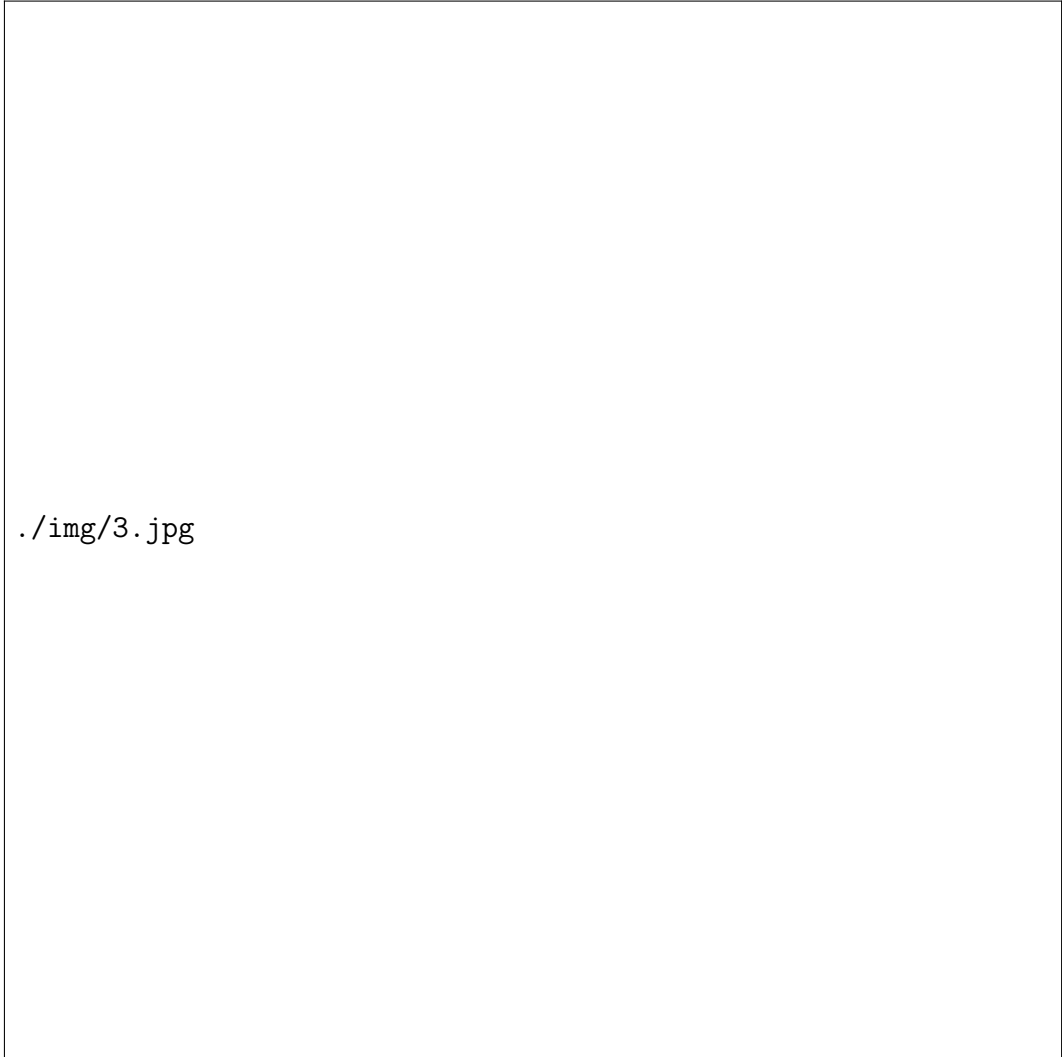
1. Inicie Sesión en AWS Educate, dirigirse a la Consola de Administración



./img/2.jpg

2. Crear un bucket de Amazon S3 En este laboratorio, debe especificar un bucket y una carpeta de Amazon S3 para almacenar los datos de salida de una consulta de Hive. En este laboratorio,

se utiliza la ubicación predeterminada para los registros, pero también puede especificar una ubicación personalizada si lo desea. Debido a los requisitos de Hadoop, los nombres del bucket y de la carpeta que utilice con Amazon EMR tienen las siguientes limitaciones: • Deben incluir únicamente letras, números, puntos (.) y guiones (-). • No pueden terminar en números. Si ya tiene acceso a una carpeta que cumpla estos requisitos, puede utilizarla para este tutorial. La carpeta de salida debería estar vacía. Otro requisito que no hay que olvidar es que los nombres de los buckets deben ser únicos en todas las cuentas de AWS. Después de crear el bucket, elíjalo en la lista y, a continuación, elija Create folder (Crear carpeta), sustituya New folder (Carpeta nueva) por un nombre que cumpla los requisitos y, por último, elija Save (Guardar). El nombre del bucket y de la carpeta utilizado más adelante en el tutorial es s3://mybucket/ MyHiveQueryResults.



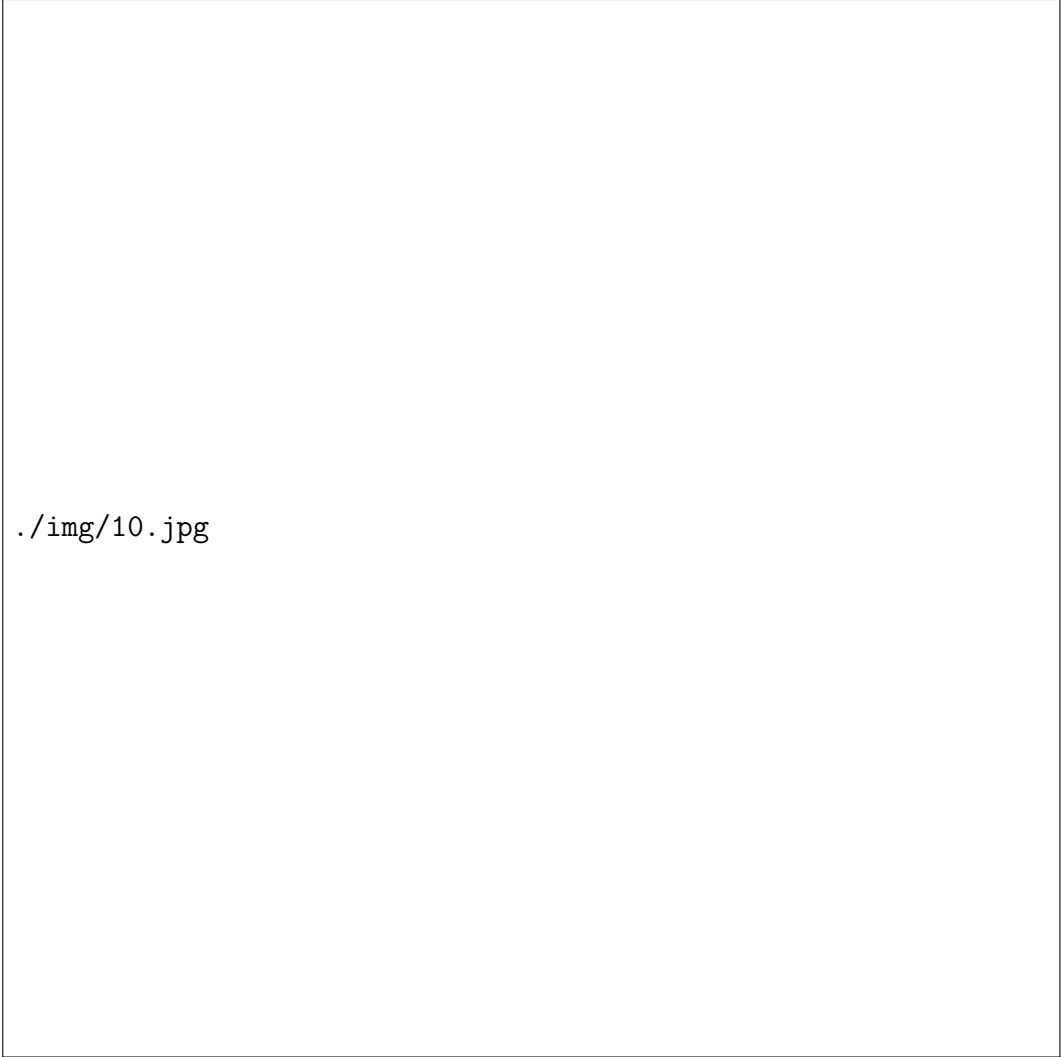
`./img/3.jpg`

3. Crear un par de claves de Amazon EC2 Debe disponer de un par de claves de Amazon Elastic Compute Cloud (Amazon EC2) para conectarse a los nodos del clúster a través de un canal seguro mediante el protocolo Secure Shell (SSH). Puede omitir este paso si ya dispone del par de claves que desea utilizar. Si no dispone de un par de claves, siga uno de los procedimientos que se indican a continuación en función de su sistema operativo.

./img/4.jpg

2.2. Paso 2: Lanzar el clúster de Amazon EMR de ejemplo

4. En este paso, lanzará el clúster de ejemplo mediante las Quick Options (Opciones rápidas) de la consola de Amazon EMR dejando la mayoría de las opciones con sus valores predeterminados. También puede seleccionar Go to advanced options (Ir a las opciones avanzadas) para explorar las opciones de configuración adicionales disponibles para un clúster.
 - a) Lanzar el clúster de ejemplo Para lanzar el clúster de Amazon EMR de ejemplo
 - a. Inicie sesión en la Consola de administración de AWS y abra la consola de Amazon EMR ([https:// console.aws.amazon.com/elasticmapreduce/](https://console.aws.amazon.com/elasticmapreduce/)).
 - b. Elija Create cluster (Crear clúster).
 - c. En la página Create Cluster - Quick Options (Crear clúster: opciones rápidas), acepte los valores predeterminados, excepto para los campos siguientes: • Introduzca un Cluster name (Nombre del clúster) que le ayude a identificar el clúster, por ejemplo, Mi primer clúster de EMR. • En Security and access (Seguridad y acceso), elija el EC2 key pair (Par de claves EC2) que ha creado en Crear un par de claves de Amazon EC2.
 - d. Elija Create cluster.

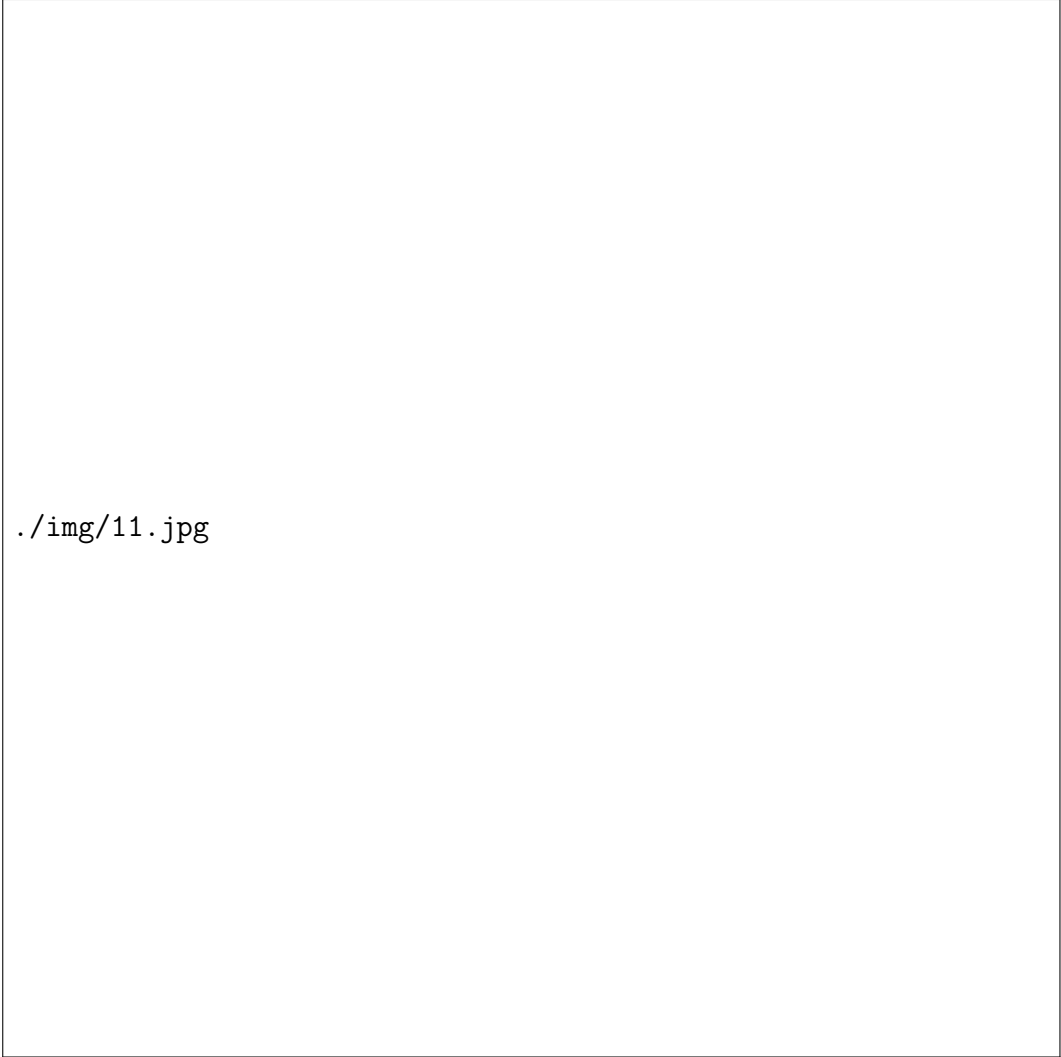


./img/10.jpg

2.3. Paso 3: Permitir las conexiones SSH con el clúster desde el cliente

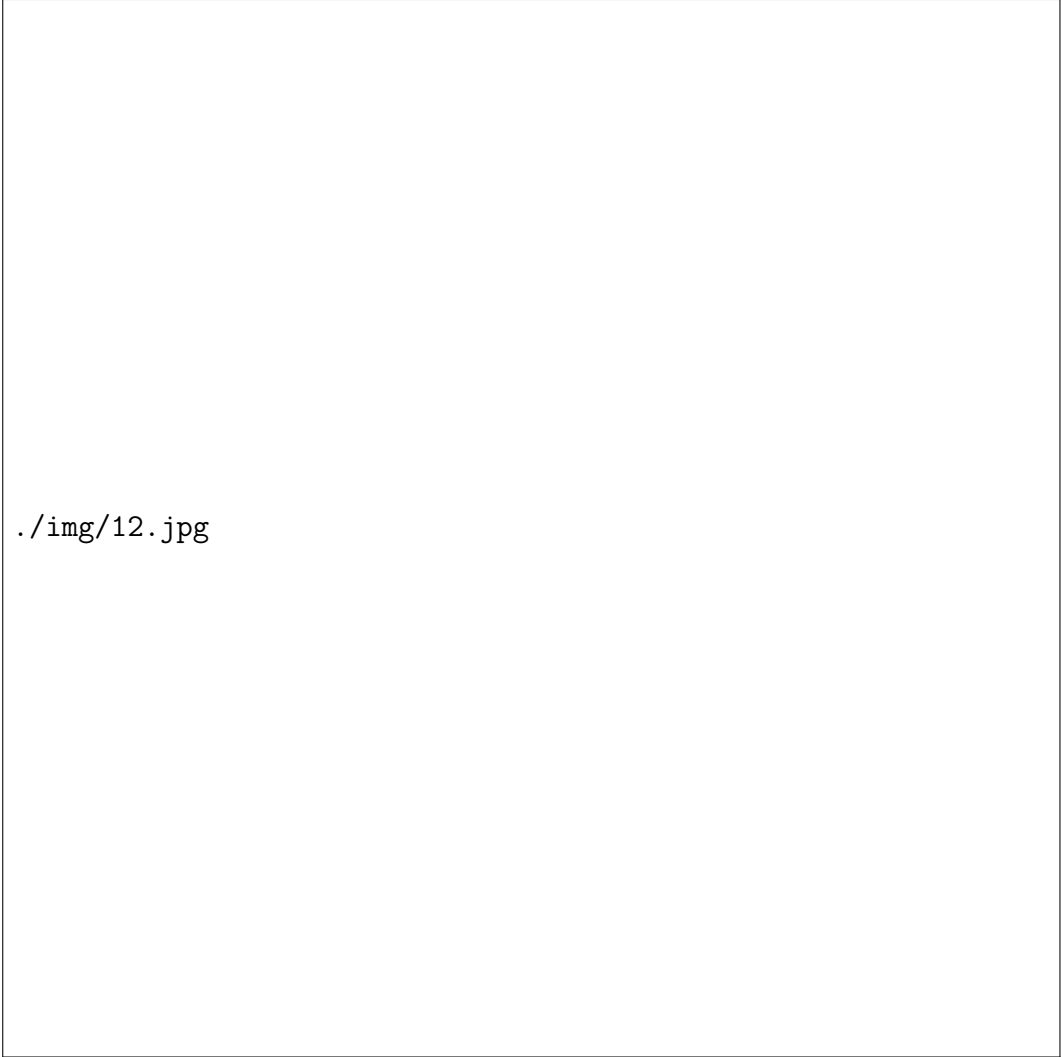
5. Los grupos de seguridad funcionan como firewalls virtuales para controlar el tráfico entrante y saliente del clúster. Al crear el primer clúster, Amazon EMR crea el grupo de seguridad administrado por Amazon EMR por defecto asociado a la instancia principal, ElasticMapReduce-master y el grupo de seguridad asociado a los nodos principal y de tareas, ElasticMapReduce-slave. Para restringir el acceso mediante SSH para el grupo de seguridad ElasticMapReduce-master Se debe haber iniciado sesión primero en AWS como usuario raíz o como principal de IAM con permiso para administrar grupos de seguridad para la VPC en la que se encuentra el clúster. Para más información, consulte Cambio de los permisos de un usuario de IAM y el Ejemplo de política que permite administrar grupos de seguridad de EC2 en la Guía del usuario de IAM.

a) Abra la consola de Amazon EMR en <https://console.aws.amazon.com/elasticmapreduce/>



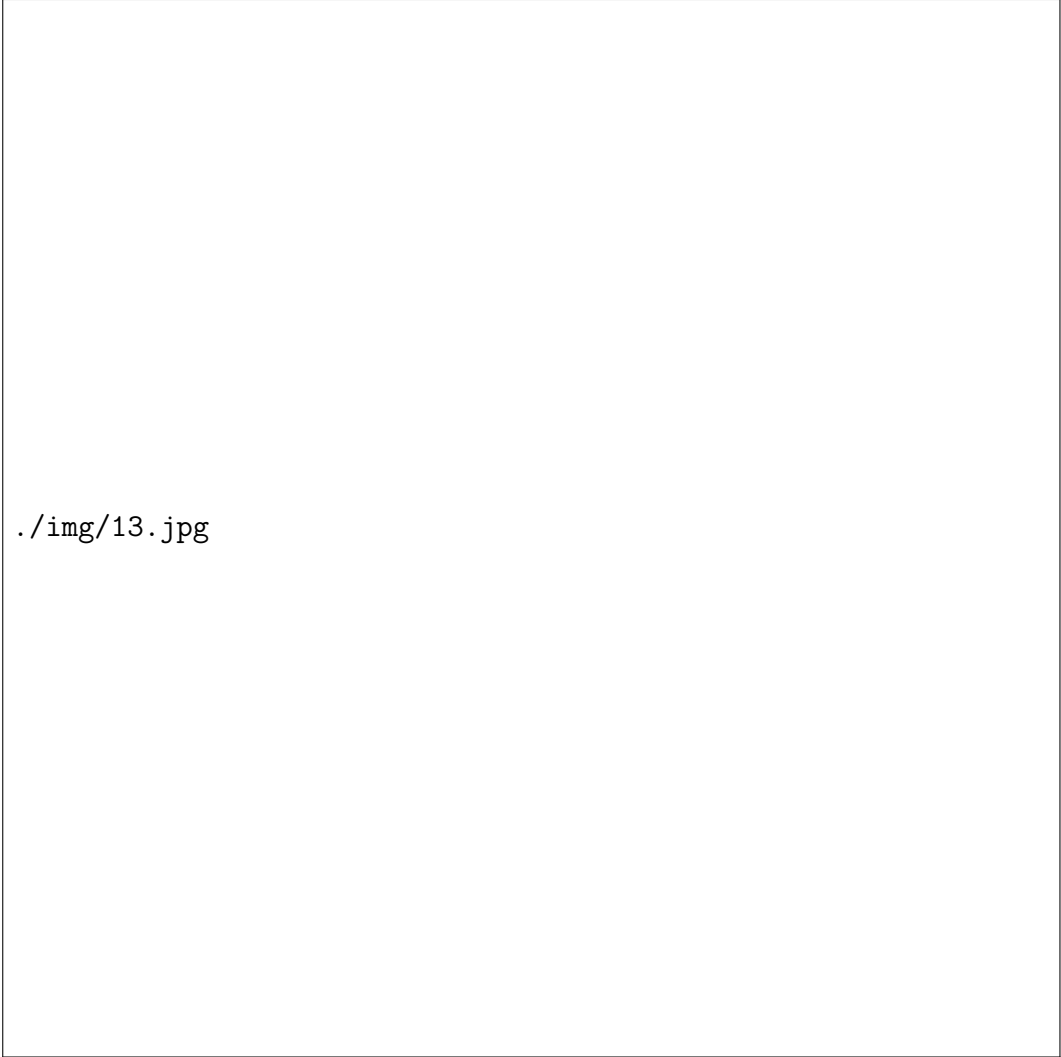
`./img/11.jpg`

b) Seleccione Clusters (Clústeres).



`./img/12.jpg`

c) Elija el Name (Nombre) del clúster.



./img/13.jpg

- d)* En Security and access (Seguridad y acceso), elija el enlace Security groups for Master (Grupos de seguridad para principal).
- e)* Elija ElasticMapReduce-master en la lista.
- f)* Elija Inbound (Entrada), Edit (Editar).
- g)* Compruebe si hay una regla de entrada que permite el acceso público con la siguiente configuración. Si existe, elija Delete (Eliminar) para eliminarla. • Type (Tipo) SSH • Port (Puerto) 22 • Source (Fuente) Personalizada 0.0.0.0/0
- h)* Desplácese a la parte inferior de la lista y elija Add Rule (Añadir regla).
- i)* En Type (Tipo), seleccione SSH. Esto introduce automáticamente TCP para Protocol (Protocolo) y 22 para Port Range (Rango de puertos).
- j)* Como origen, seleccione My IP (Mi IP). Esto añade automáticamente la dirección IP del equipo cliente como la dirección de origen. También puede añadir un rango de direcciones IP de clientes de confianza Custom (Personalizadas) y elegir Add rule (Añadir regla) para crear reglas adicionales para otros clientes. Muchos entornos de red asignan dinámicamente direcciones IP, por lo que es posible que necesite editar periódicamente

las reglas de grupos de seguridad para actualizar las direcciones IP de los clientes de confianza.

k) Elija Save (Guardar).

2.4. Paso 4: Procesar los datos ejecutando el script de Hive como paso

- a) Abra la consola de Amazon EMR en <https://console.aws.amazon.com/elasticmapreduce/>
- b) En Cluster List (Lista de clústeres), seleccione el nombre del clúster. Asegúrese de que el clúster está en el estado Waiting (Esperando).
- c) Elija Steps (Pasos) y, a continuación Add step (Añadir paso).

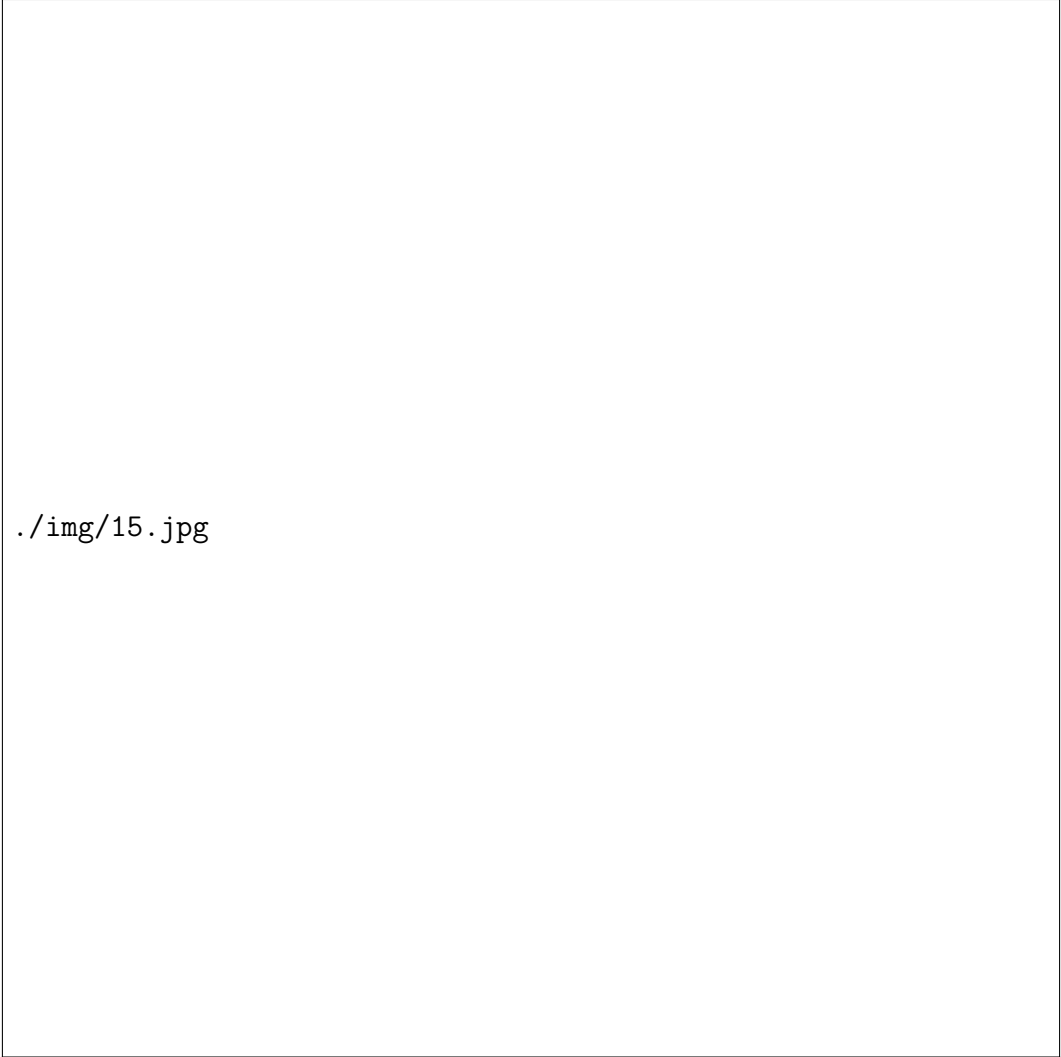
d) Configure el paso de acuerdo con las directrices siguientes: • En Step type (Tipo de paso), elija Hive program (Programa de Hive). • En Name (Nombre), puede dejar el valor predeterminado o escribir un nombre nuevo. Si tiene muchos pasos en un clúster, el nombre le ayuda a realizar un seguimiento de ellos. • En Script S3 location (Ubicación en S3 del script), escriba `s3://region.elasticmapreduce.samples/cloudfront/code/HiveCloudFront.q.Sustituya` `s3://uswest2.elasticmapreduce.samples/cloudfront/code/HiveCloudFront.q` si está trabajando en la región `us-west-2`. `s3://region.elasticmapreduce.samples` Sustituya la región por el identificador de la región. En Output S3 location (Ubicación en S3 de la salida), escriba `s3://region.elasticmapreduce.samples/cloudfront/output/HiveCloudFront.q` si está trabajando en la región `us-west-2`. Sustituya la región por el identificador de la región.

ℳ) El estado del paso cambia de Pending (Pendiente) a Running (En ejecución) y a Completed (Completado) a medida que se ejecuta. Para actualizar el estado, elija el icono de actualización situado a la derecha de Filter (Filtro). El script tarda aproximadamente un minuto en ejecutarse.

1. Abra la consola de Amazon S3 en <https://console.aws.amazon.com/s3/>.

2. Elija el Bucket name (Nombre del bucket) y, a continuación, elija la carpeta que ha configurado anteriormente. Por ejemplo, mybucket y luego MyHiveQueryResults.

3. La consulta escribe los resultados en una carpeta ubicada en la carpeta de salida denominada `os_requests`. Elija la carpeta. Debería haber un único archivo denominado `000000000end` en dicha carpeta. Se trata de un archivo de texto que contiene los resultados de la consulta.

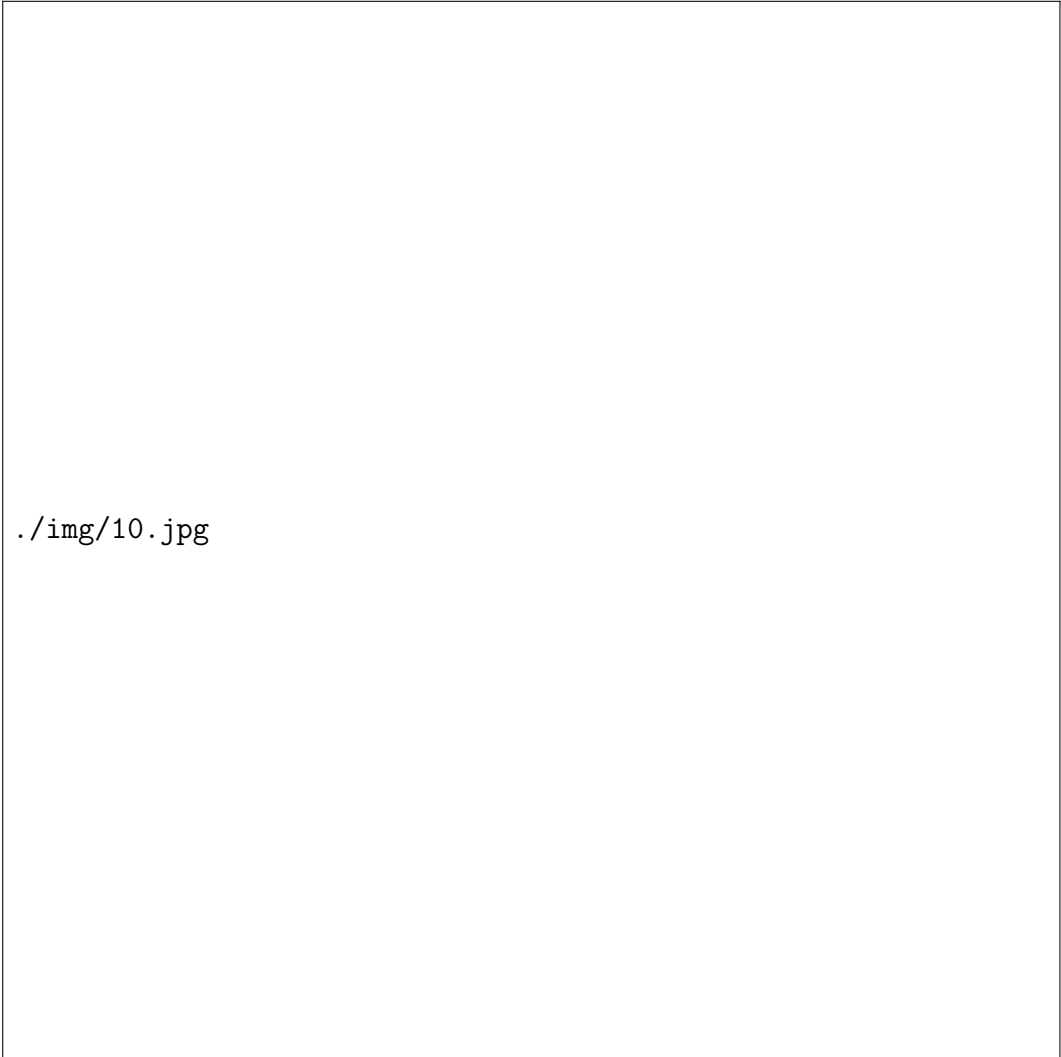


`./img/15.jpg`

2.5. Paso 5: Terminar el clúster y eliminar el bucket

6. Es conveniente que termine el clúster y elimine el bucket de Amazon S3 para evitar cargos adicionales. Al terminar el clúster, terminan las instancias Amazon EC2 asociadas y se detiene la acumulación de cargos de Amazon EMR. Amazon EMR conserva la información de metadatos sobre los clústeres completados para su referencia, gratuitamente, durante dos meses. La consola no proporciona una forma de eliminar clústeres terminados, por lo que no se pueden ver en la consola. Los clústeres terminados se eliminan del clúster al eliminar los metadatos.
 - a) . Abra la consola de Amazon EMR en <https://console.aws.amazon.com/elasticmapreduce/>.
 - b) Elija Clusters (Clústeres), elija el clúster y, a continuación, Terminate (Terminar). Los clústeres suelen crearse con la protección de terminación activada, lo que ayuda a evitar que se cierren de forma accidental. Si ha seguido el tutorial al pie de la letra, la protección de terminación debería estar desactivada. Si la protección de terminación está activada, se le pedirá que cambie esta opción como medida de precaución antes de terminar el clúster. Elija Change (Cambiar), Off (Desactivada).

- c) Para eliminar el bucket de salida
1. Abra la consola de Amazon S3 en <https://console.aws.amazon.com/s3/>
 2. Elija el bucket en la lista, de forma que toda la fila del bucket esté seleccionada.
 3. Elija eliminar el bucket, escriba el nombre de este y, a continuación, haga clic en Confirm (Confirmar). Para obtener más información sobre la eliminación de carpetas y buckets, vaya a ¿Cómo elimino un bucket de S3? en la Guía de introducción a Amazon Simple Storage Service.



`./img/10.jpg`

3. CONCLUSIONES

- Se realizó con éxito la creación de tablas y agregación de datos
Se analizó realizando consultas de datos y finalmente se eliminó la tabla