



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

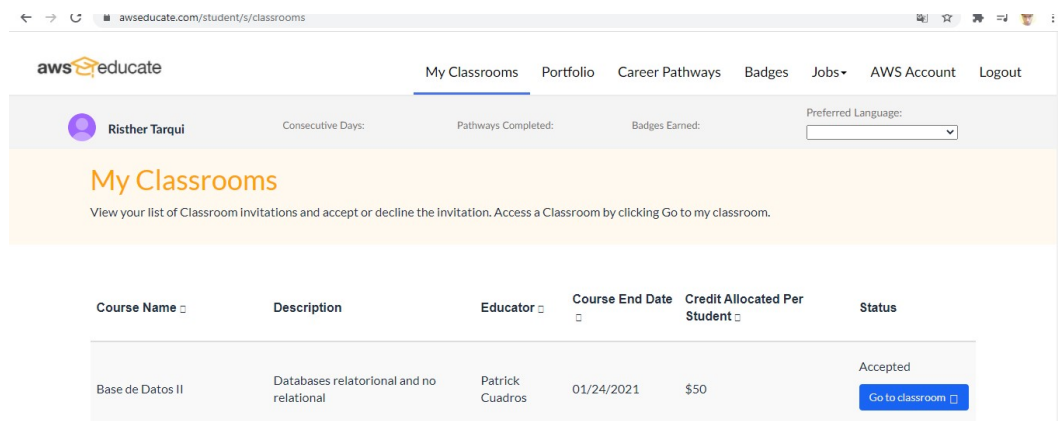
1. OBJETIVO

- 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).

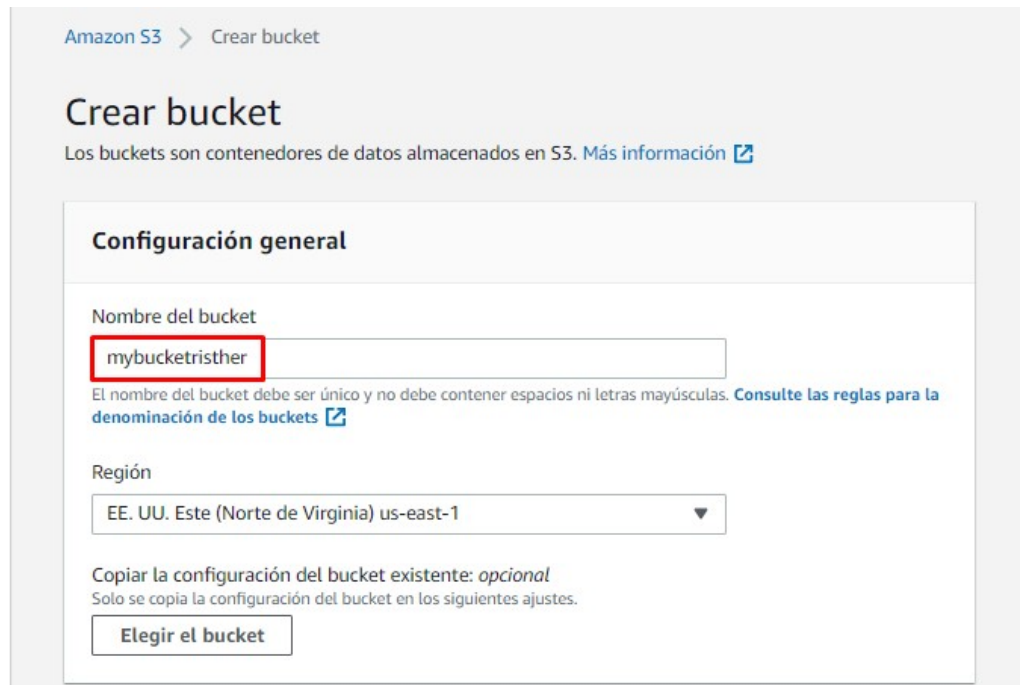
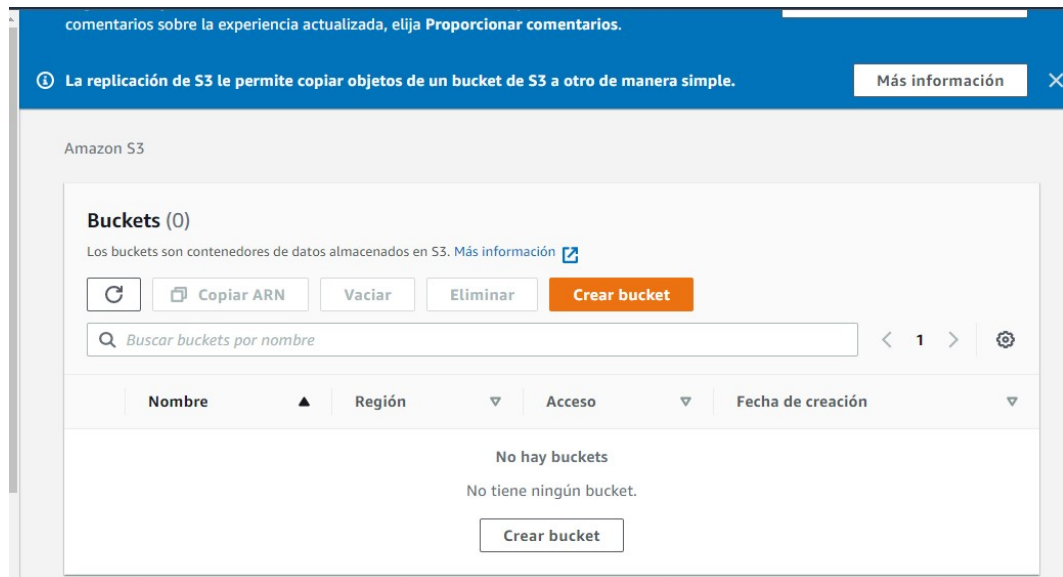
2. DESARROLLO

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



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`.



Cifrado predeterminado

Cifre automáticamente los nuevos objetos almacenados en este bucket. [Más información](#)

Cifrado del lado del servidor

☒ Deshabilitar

☐ Habilitar

► Configuración avanzada

Después de crear el bucket, puede cargar archivos y carpetas en el bucket y configurar ajustes adicionales del bucket.

Cancelar

Crear bucket

La replicación de S3 le permite copiar objetos de un bucket de S3 a otro de manera simple. [Más información](#)

Amazon S3

Buckets (1)

Los buckets son contenedores de datos almacenados en S3. [Más información](#)

Buscar buckets por nombre

	Nombre	Región	Acceso	Fecha de creación
<input type="radio"/>	mybucketristher	EE. UU. Este (Norte de Virginia) us-east-1	Bucket y objetos que no son públicos	29 Nov 2020 8:58:10 PM -03

Arrastre y suelte los archivos y las carpetas que desee cargar aquí, o elija Cargar.


Objetos (0)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Para que otras personas obtengan acceso a los objetos, tendrá que cono [información](#)

Eliminar Acciones **Crear carpeta** Cargar

Buscar objetos por prefijo

	Nombre	Tipo	Última modificación	Tamaño	Clase d
No hay objetos					

**Su política de bucket podría bloquear la creación de carpetas**
Si su política de bucket impide cargar objetos sin etiquetas, metadatos o beneficiarios específicos de la lista de control de acceso (ACL), no podrá crear una carpeta con esta configuración. En su lugar, puede utilizar la configuración de carga para cargar una carpeta vacía y especificar la configuración adecuada.


Carpeta

Nombre de la carpeta

MyHiveQueryResults /

Los nombres de las carpetas no pueden contener "/"". [Consulte las reglas de nomenclatura](#)

Cifrado del lado del servidor

 La siguiente configuración se aplica únicamente al nuevo objeto de carpeta y no a los objetos que contiene.

Cifrado del lado del servidor

☒ Deshabilitar

☐ Habilitar

Cancelar

Crear carpeta

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.

← → ↻ console.aws.amazon.com/ec2/v2/home?region=us-east-1#Home:

aws Servicios ▾

☐ New EC2 Experience Más información

Panel de EC2 New

Eventos New

Etiquetas

Límites

▼ Instancias

Instancias New

Tipos de instancia

Plantillas de lanzamiento

Solicitudes de spot

Savings Plans

Instancias reservadas

Hosts dedicados New

Instancias programadas

Reservas de capacidad

▼ Imágenes

AMI

▼ Elastic Block Store

Volúmenes

Lanzar la instancia

Para comenzar, lance una instancia de Amazon EC2, que es un servidor virtual en la nube.

Lanzar la instancia ▾

Nota: Sus instancias se lanzarán en la región EE.UU. Este (Norte de Virginia)

Eventos programados

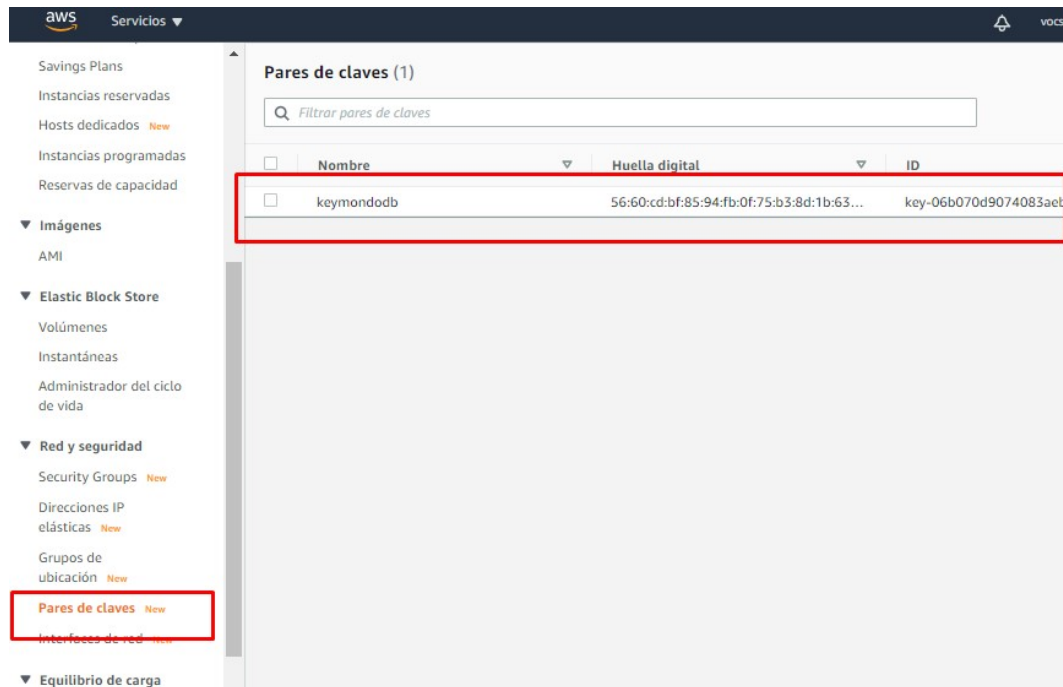
EE.UU. Este (Norte de Virginia)

No hay eventos programados

Migrar una máquina

Use CloudEndure Migration para simplificar, agilizar y automatizar las migraciones a gran escala desde infraestructuras físicas, virtuales y basadas en la nube a AWS.

[Introducción a CloudEndure Migration](#) ↗

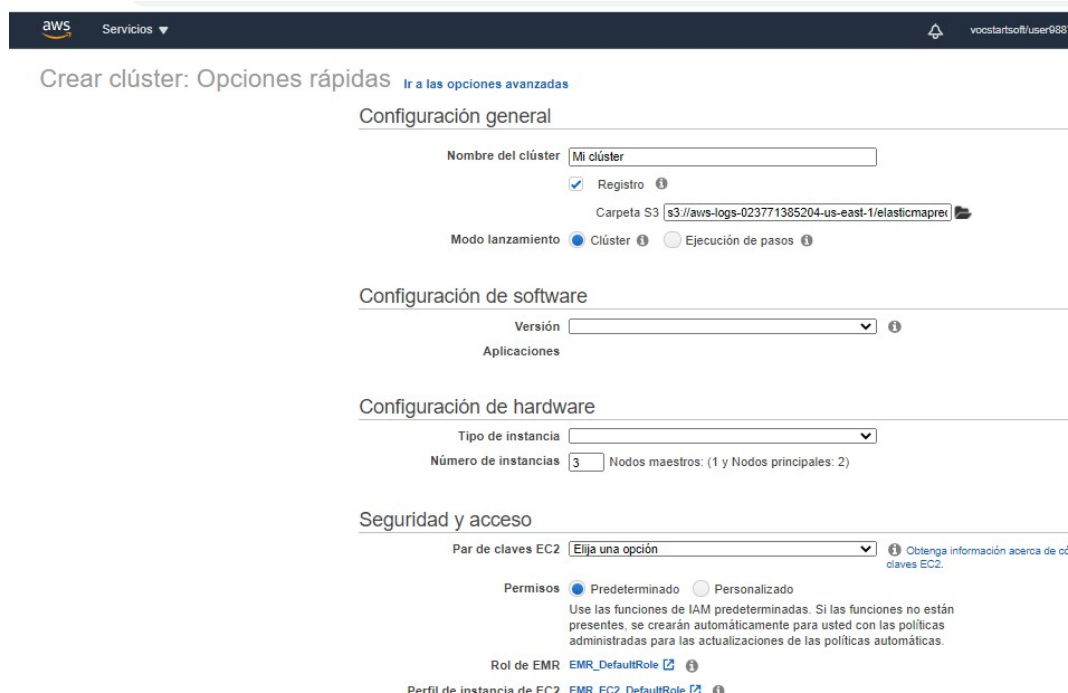


2.2. Paso 2: Lanzar el clúster de Amazon EMR de ejemplo

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.

Configuración de hardware

Tipo de instancia: El tipo de instancia seleccionado añade un volumen de EBS GP2 de 64 GiB predeterminado por instancia. [Más información](#)

Número de instancias: Nodos maestros: (1 y Nodos principales: 2)

Cluster scaling: ☐ scale cluster nodes based on workload

Seguridad y acceso

Par de claves EC2: [Obtenga información acerca de cómo crear un par de claves EC2.](#)

Permisos: ☒ Predeterminado ☐ Personalizado
Use las funciones de IAM predeterminadas. Si las funciones no están presentes, se crearán automáticamente para usted con las políticas administradas para las actualizaciones de las políticas automáticas.

Rol de EMR: [EMR_DefaultRole](#)

Perfil de instancia de EC2: [EMR_EC2_DefaultRole](#)

[Cancelar](#) [Crear clúster](#)

d. Elija Create cluster.

Amazon EMR

Clúster: ClousterRisther **Comenzando**

[Clonar](#) [Finalizar](#) [Exportación de la CLI de AWS](#)

[Resumen](#) [Historial de aplicaciones](#) [Monitorización](#) [Hardware](#) [Configuraciones](#) [Eventos](#) [Pasos](#) [Acciones de arranque](#)

Resumen

ID: j-2YQE8DZXD9GXN
Fecha de creación: 2020-12-03 12:26 (UTC-3)
Tiempo transcurrido: 0 segundos
Terminar automáticamente: Cluster waits
Protección contra la Desactivación: [Cambiar](#)
Etiquetas: -- [Ver todo / Editar](#)
DNS público principal: --

Detalles de las configuraciones

Etiqueta de la versión: emr-5.32.0
Distribución Hadoop: Amazon 2.10.1
Aplicaciones: Hive 2.3.7, Hue 4.8.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
URI de registro: s3://aws-logs-023771385204-us-east-1/elasticmapreduce/
Vista coherente de EMRFS: Deshabilitados
ID de AMI personalizada: --

Redes y hardware

Zona de disponibilidad: --
ID de subred: subnet-8417aeb5
Maestro: [Aprovisionamiento](#) 1 m5.xlarge
Principal: [Aprovisionamiento](#) 2 m5.xlarge
Tarea: --
Cluster scaling: Not enabled

Seguridad y acceso

Nombre de la clave: keymonddb
Perfil de instancia EC2: EMR_EC2_DefaultRole
Función de EMR: EMR_DefaultRole
[Mejorar para todos los Todo](#) [Cambiar](#)

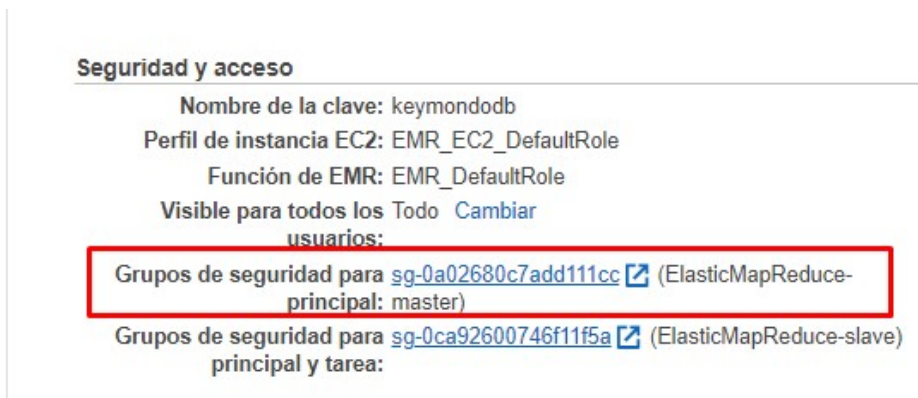
2.3. Paso 3: Permitir las conexiones SSH con el clúster desde el cliente

- 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.

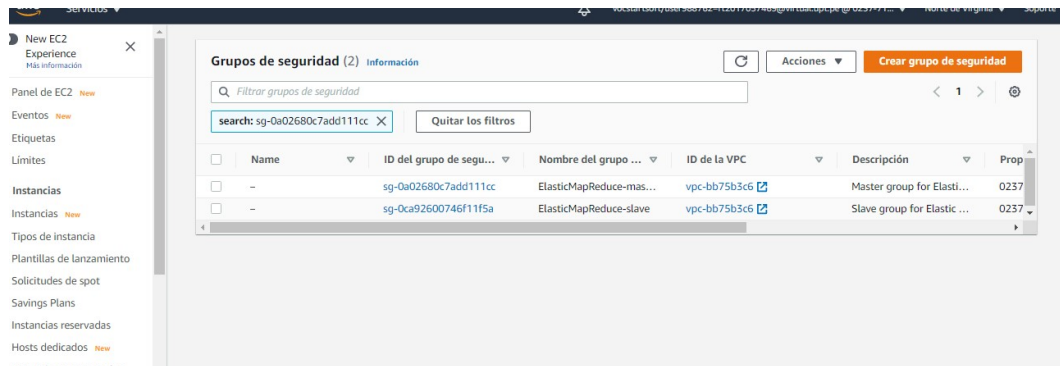
a) Abra la consola de Amazon EMR en <https://console.aws.amazon.com/elasticmapreduce/>



- b) Seleccione Clusters (Clústeres).
- c) Elija el Name (Nombre) del clúster.
- 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).

Reglas de entrada				
Reglas de entrada				
Tipo	Protocolo	Intervalo de puertos	Origen	Descripción: opcional
Todos los TCP	TCP	0 - 65535	sg-0a02680c7add111cc (ElasticMapReduce-master)	-
Todos los TCP	TCP	0 - 65535	sg-0ca92600746f11f5a (ElasticMapReduce-slave)	-
TCP personalizado	TCP	8443	207.171.167.25/32	-
TCP personalizado	TCP	8443	54.240.217.8/29	-
TCP personalizado	TCP	8443	72.21.196.64/29	-
TCP personalizado	TCP	8443	72.21.198.64/29	-
TCP personalizado	TCP	8443	54.240.217.16/29	-
TCP personalizado	TCP	8443	54.239.98.0/24	-
TCP personalizado	TCP	8443	207.171.167.101/32	-
TCP personalizado	TCP	8443	207.171.167.26/32	-
TCP personalizado	TCP	8443	72.21.217.0/24	-

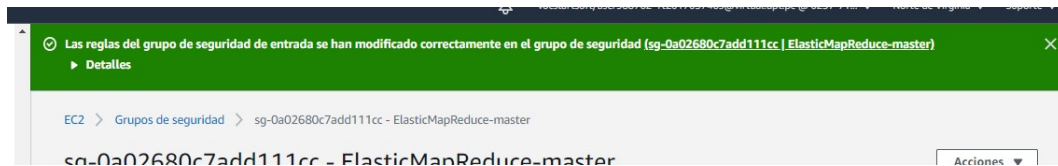
g) Compruebe si hay una regla de entrada que permite el acceso público con la siguiente configuración. Si existe, elija Delete (Eliminar) para eliminarla. • Type (Tipo) SSH • Port (Puerto) 22 • Source (Fuente) Personalizada 0.0.0.0/0

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

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



- 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 region por el identificador de la región. En Output S3 lo`

Añadir paso

Tipo de paso: Programa de Hive

Nombre: Programa de Hive

Ubicación S3 del script: s3://region.elasticmapreduce.samples/cloudfront/codi

Ubicación S3 de entrada: s3://region.elasticmapreduce.samples

Ubicación S3 de salida: s3://mybuckettristher/

Argumentos:

Acción sobre el error: Continuar

Cancelar Añadir

6. Elija Add (Añadir). El paso aparece en la consola con el estado Pending (Pendiente).

After last step completes: Cluster waits

Añadir paso Clonar paso Cancelar paso

View Jobs in the Application Histo

Filter: Todos los pasos Filtrar pasos... Pasos: 3 (todos cargados)

ID	Nombre	Estado	Hora de inicio (UTC-3)	Tiempo transcurrido	Archivos de registro P2
s-1S8E33YTR2JGZ	Programa de Hive	Completado	2020-12-03 13:06 (UTC-3)	52 segundos	Ver logs

7. 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/>.

After last step completes: Cluster waits

Añadir paso Clonar paso Cancelar paso

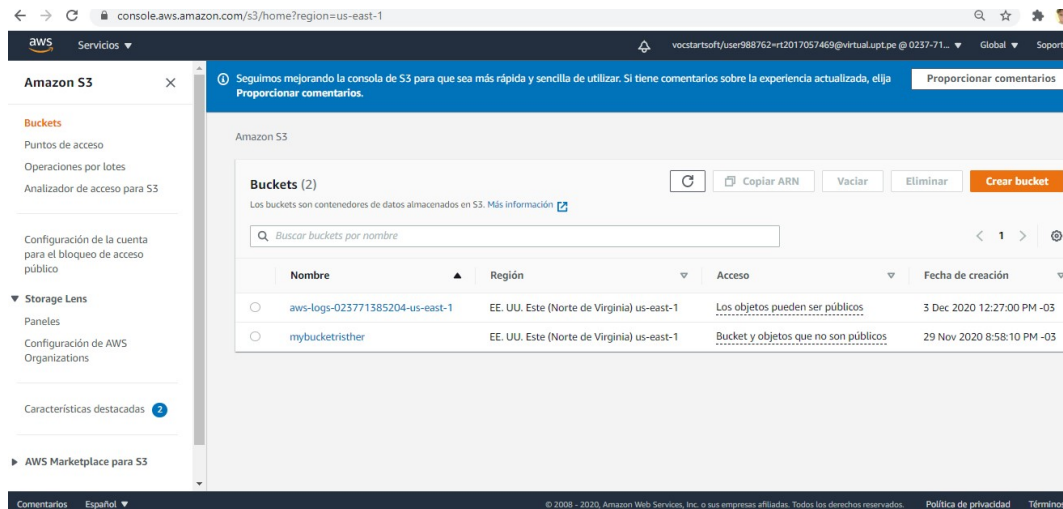
View Jobs in the Application Histo

Filter: Todos los pasos Filtrar pasos... Pasos: 3 (todos cargados)

ID	Nombre	Estado	Hora de inicio (UTC-3)	Tiempo transcurrido	Archivos de registro P2
s-1S8E33YTR2JGZ	Programa de Hive	Completado	2020-12-03 13:06 (UTC-3)	52 segundos	Ver logs

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 esa carpeta. Debería haber un único archivo denominado `00000000_end` en esa carpeta. Setra



*Sin título: Bloc de notas

Archivo Edición Formato Ver Ayuda

Android| 855

Linux| 813

MacOS| 852

OSX| 799

Windows| 883

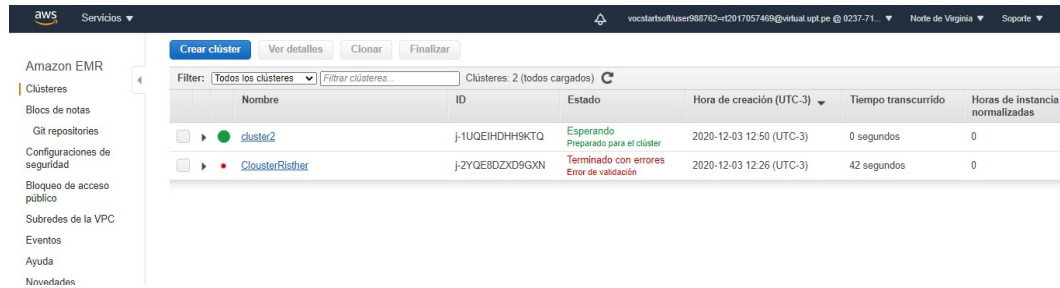
ios| 794|

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

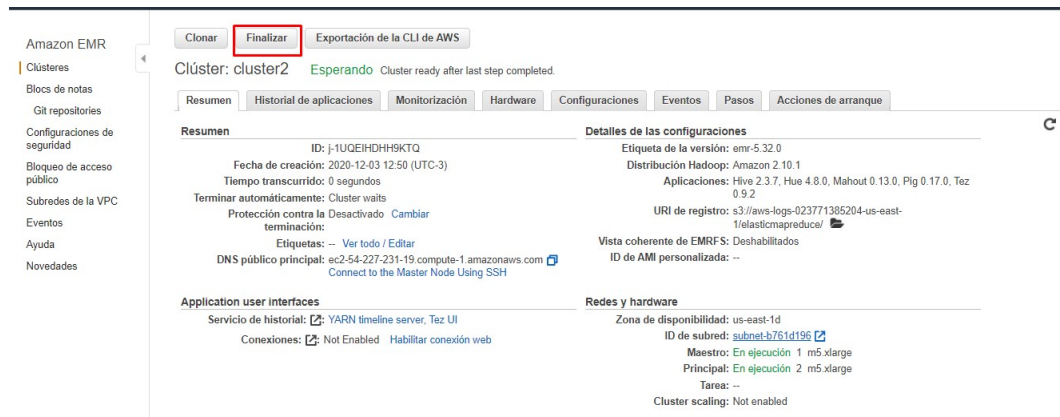
Es conveniente que termine el clúster y elimine el bucket de Amazon S3 para evitar cargos adicionales. Al terminar el clúster, terminan las instancias Amazon EC2 asociadas y se detiene la acumulación de cargos de Amazon EMR. Amazon EMR conserva la información de metadatos sobre los clústeres completados 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.

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



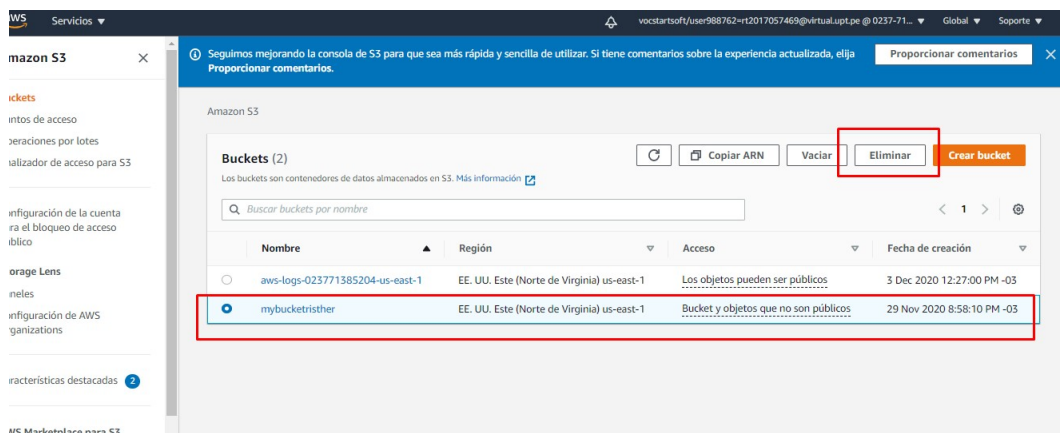
2. Elija Clusters (Clústeres), elija el clúster y, a continuación, Terminate (Terminar). Los clústeres suelen crearse con la protección de terminación activada, lo que ayuda a evitar que se cierren de forma accidental. Si ha seguido el tutorial al pie de la letra, la protección de terminación debería 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).



3. 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.



3. CONCLUSIONES

- Se creo un bucket y un par de keys para la creacion de cluster, dentro del cluster editamos la reglas de entrada para conexion mediante ssh.Finalmente creamos un paso con el cluster en la carpeta de bucket que se creo al inicio.