

**UNIVERSIDAD PRIVADA DE TACNA**



**FACULTAD DE INGENIERIA**

**Escuela Profesional de Ingeniería de Sistema**

**Informe de laboratorio 01: Introducción a big  
data con Amazon EMR**

**Curso: Base de datos II**

**DOCENTE: Ing. Patrick Cuadros Quiroga**

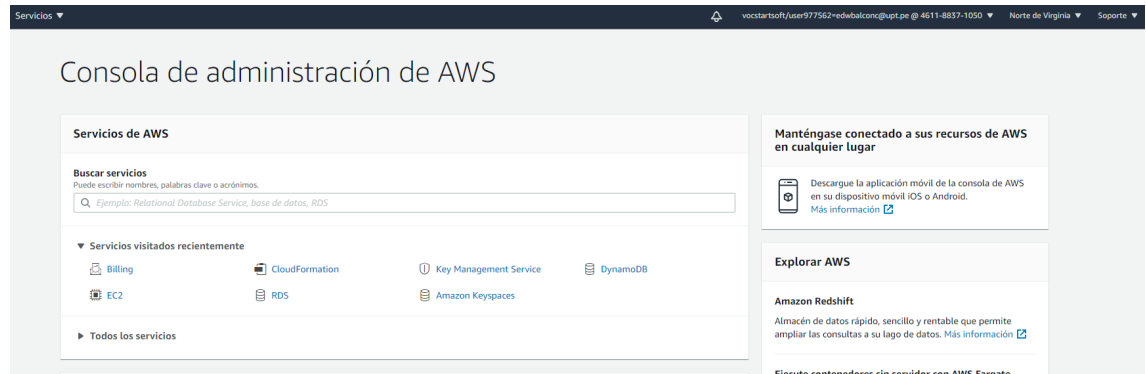
**Alumno: Balcon Coahila, Edwart Juan  
(2013046516)**

**Tacna – Perú**

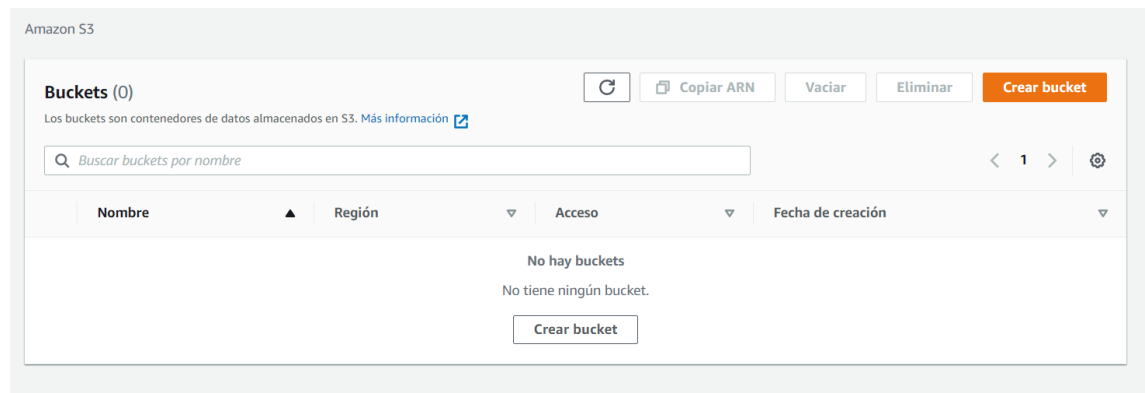
**2020**

# 1. Configurar los requisitos previos para el clúster de ejemplo

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



## 1.2. Crear un bucket de Amazon S3



## Crear bucket

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

### Configuración general

Nombre del bucket

myawsbucket

El nombre del bucket debe ser único y no debe contener espacios ni letras mayúsculas. [Consulte las reglas para la denominación de los buckets](#)

Región

EE. UU. Este (Norte de Virginia) us-east-1

Copiar la configuración del bucket existente: *opcional*

Solo se copia la configuración del bucket en los siguientes ajustes.

Elegir el bucket

Amazon S3

### Buckets (1)

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

Buscar buckets por nombre

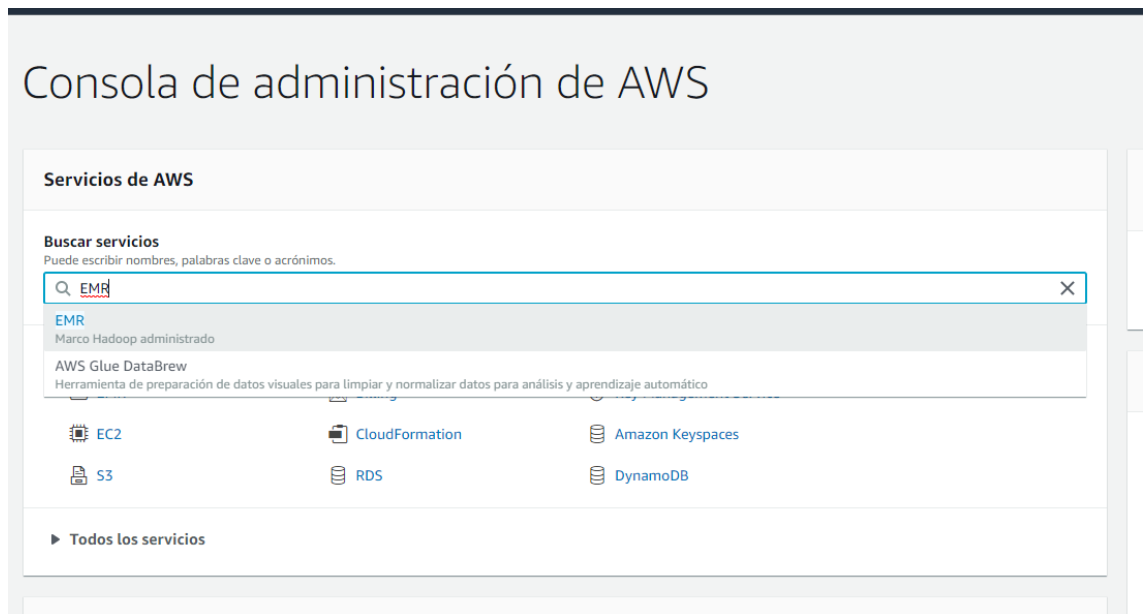
	Nombre ▲	Región ▼	Acceso ▼	Fecha de creación ▼
<input type="radio"/>	awsbucket-bdii	EE. UU. Este (Norte de Virginia) us-east-1	Bucket y objetos que no son públicos	6 Dec 2020 10:24:17 PM -05

### 1.3. Crear un par de claves de Amazon EC2

Pares de claves (1)			
<input type="text" value="Filtrar pares de claves"/>			
<input type="checkbox"/>	Nombre	Huella digital	ID
<input type="checkbox"/>	BDII	80:4b:1ca5:93:bb:45:bb:20:76:81:99:5...	key-05c45492e15e02d07

## 2. Lanzar el clúster de Amazon EMR de ejemplo

2.1. 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/)).



## 2.2. Elija Create cluster (Crear clúster)

Amazon EMR

Clústeres

Blocs de notas

Git repositories

Configuraciones de seguridad

Bloqueo de acceso público

Subredes de la VPC

Eventos

Ayuda

Novedades

### Bienvenido/a a Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) es un servicio web que permite a empresas, investigadores, analistas de datos y desarrolladores procesar de forma fácil y rentable grandes volúmenes de datos.

Parece que no dispone de ningún clúster. Crear uno ahora:

[Crear clúster](#)

### Cómo funciona Elastic MapReduce

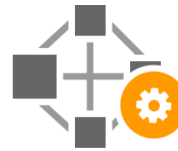
Cargar



Cargue sus datos y la aplicación de procesamiento en S3.

[Más información](#)

Crear



Configure y cree el clúster especificando las entradas de datos, los resultados, tamaño del clúster, configuración de seguridad, etc.

[Más información](#)

Monitorizar



Monitoree el estado y el progreso del clúster. Recupere el resultado en S3.

[Más información](#)


2.3. En la página Create Cluster - Quick Options (Crear clúster: opciones rápidas), acepte los valores predeterminados

---

### Configuración general

Nombre del clúster

☒ Registro ⓘ

Carpeta S3  

Modo lanzamiento ☒ Clúster ⓘ ☐ Ejecución de pasos ⓘ

---

### Configuración de software

Versión  ⓘ

Aplicaciones ☒ Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Phoenix 4.14.3, and ZooKeeper 3.4.14


☐ Presto: Presto 0.240.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore

☐ Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.8.2

☐ Usar el catálogo de datos de AWS Glue para metadatos de tabla ⓘ

---

### Configuración de hardware

Tipo de instancia  ⓘ El tipo de instancia seleccionado añade un volumen de EBS GP2 de 84 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](#) ⓘ

## 2.4. Elija Create cluster

ClonarFinalizarExportación de la CLI de AWS

Clúster: Mi primer clúster de EMRComenzando

ResumenHistorial de aplicacionesMonitorizaciónHardwareConfiguracionesEventosPasosAcciones de arranque

Resumen

ID: j-1TG9P684OPB75

Fecha de creación: 2020-12-06 22:35 (UTC-5)

Tiempo transcurrido: 0 segundos

Terminar automáticamente: Cluster waits

Protección contra la terminación: Desactivado [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-461188371050-us-east-1/elasticmapreduce/

Vista coherente de EMRFS: Deshabilitados

ID de AMI personalizada: --

Application user interfaces

Servicio de historial: : --

Conexiones: : --

Seguridad y acceso

Nombre de la clave: BDI

Perfil de instancia EC2: EMR\_EC2\_DefaultRole

Función de EMR: EMR\_DefaultRole

Redes y hardware

Zona de disponibilidad: --

ID de subred: [subnet-e077c7c1](#)

Maestro: [Aprovisionamiento](#) 1 m5.xlarge

Principal: [Aprovisionamiento](#) 2 m5.xlarge

Tarea: --

Cluster scaling: Not enabled

## 3. Permitir las conexiones SSH con el clúster desde el cliente

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

awsServicios

vocstartsoft/user977562-edwbalconc@upt.pe @ 4611-8837-1050

Norte de Virginia

Soporte

Crear clústerVer detallesClonarFinalizar

Filter: Todos los clústeres [Filtrar clústeres...](#) Clústeres: 1 (todos cargados)

	Nombre	ID	Estado	Hora de creación (UTC-5)	Tiempo transcurrido	Horas de instancia normalizadas		
			<a href="#">Mi primer clúster de EMR</a>	j-1TG9P684OPB75	Comenzando	2020-12-06 22:35 (UTC-5)	3 minutos	0

Amazon EMR

Clústeres

Bloqs de notas

Git repositories

Configuraciones de seguridad

Bloqueo de acceso público

Subredes de la VPC

Eventos

Ayuda

Novedades

### 3.2. Seleccione Clusters (Clústeres).

aws Servicios

vocstartsoft/user977562-edwbalconc@upl.pe @ 4611-8837-1050 Norte de Virginia Soporte

Crear clúster Ver detalles Clonar Finalizar

Amazon EMR

Clústeres

Blocs de notas

Git repositories

Configuraciones de seguridad

Bloqueo de acceso público

Subredes de la VPC

Eventos

Ayuda

Novedades

Filter: Todos los clústeres Filtrar clústeres Clústeres: 1 (todos cargados)

	Nombre	ID	Estado	Hora de creación (UTC-5)	Tiempo transcurrido	Horas de instancia normalizadas
	Mi primer clúster de EMR	j-1TG9P684OPB75	Comenzando	2020-12-06 22:35 (UTC-5)	3 minutos	0

### 3.3. Elija el Name (Nombre) del clúster.

aws Servicios

vocstartsoft/user977562-edwbalconc@upl.pe @ 4611-8837-1050 Norte de Virginia Soporte

Crear clúster Ver detalles Clonar Finalizar

Amazon EMR

Clústeres

Blocs de notas

Git repositories

Configuraciones de seguridad

Bloqueo de acceso público

Subredes de la VPC

Eventos

Ayuda

Novedades

Filter: Todos los clústeres Filtrar clústeres Clústeres: 1 (todos cargados)

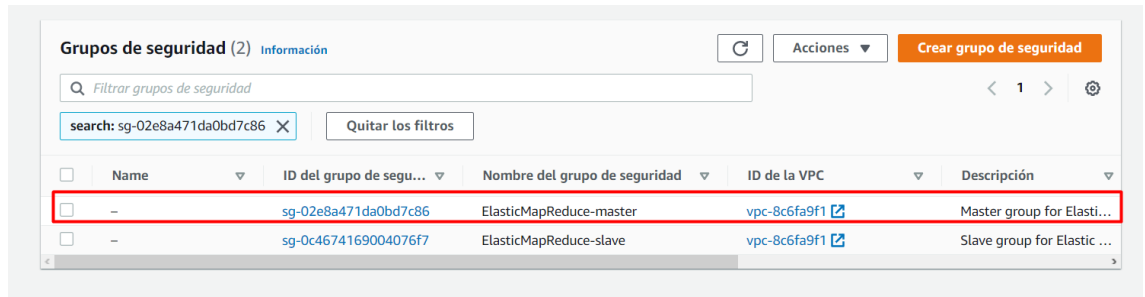
	Nombre	ID	Estado	Hora de creación (UTC-5)	Tiempo transcurrido	Horas de instancia normalizadas
	Mi primer clúster de EMR	j-1TG9P684OPB75	Esperando Preparado para el clúster	2020-12-06 22:35 (UTC-5)	37 minutos	0



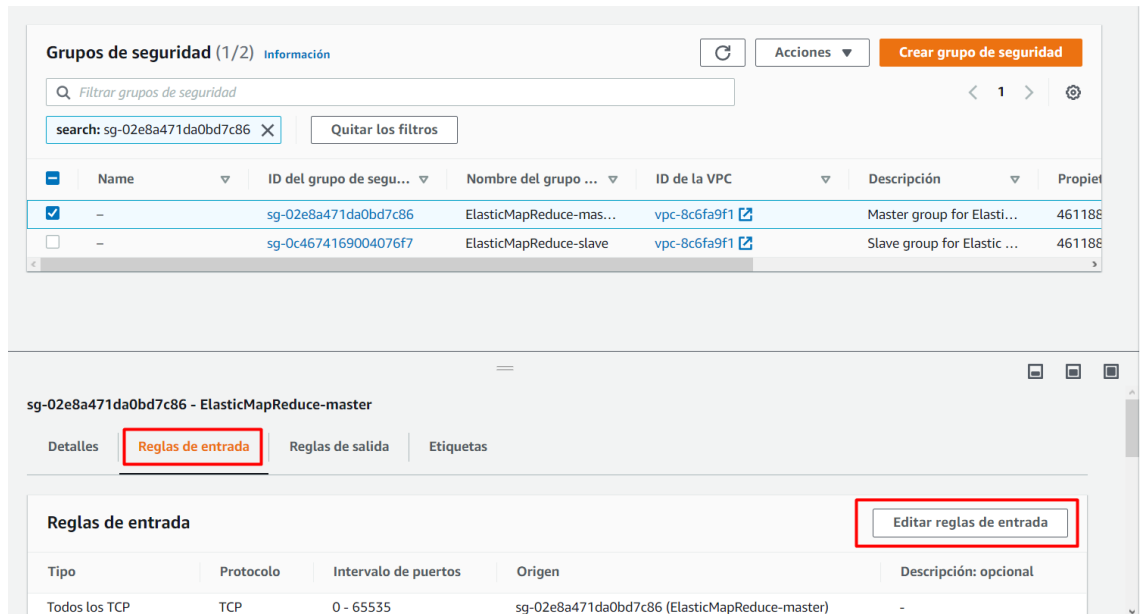
3.4. En Security and access (Seguridad y acceso), elija el enlace Security groups for Master (Grupos de seguridad para principal).

Resumen	Detalles de las
<p><b>ID:</b> j-1TG9P684OPB75</p> <p><b>Fecha de creación:</b> 2020-12-06 22:35 (UTC-5)</p> <p><b>Tiempo transcurrido:</b> 39 minutos</p> <p><b>Terminar automáticamente:</b> Cluster waits</p> <p><b>Protección contra la terminación:</b> Desactivado <a href="#">Cambiar</a></p> <p><b>Etiquetas:</b> -- <a href="#">Ver todo / Editar</a></p> <p><b>DNS público principal:</b> ec2-3-83-148-68.compute-1.amazonaws.com <a href="#">Connect to the Master Node Using SSH</a></p>	<p><b>Etiqueta</b></p> <p><b>Distrib</b></p> <p><b>U</b></p> <p><b>Vista cohere</b></p> <p><b>ID de AMI  </b></p>
Application user interfaces	Redes y hardw
<p><b>Servicio de historial:</b> <a href="#">YARN timeline server</a>, Tez UI</p> <p><b>Conexiones:</b> <a href="#">Not Enabled</a> <a href="#">Habilitar conexión web</a></p>	<p><b>Zona de c</b></p>
<b>Seguridad y acceso</b>	
<p><b>Nombre de la clave:</b> BDII</p> <p><b>Perfil de instancia EC2:</b> EMR_EC2_DefaultRole</p> <p><b>Función de EMR:</b> EMR_DefaultRole</p> <p><b>Visible para todos los usuarios:</b> Todo <a href="#">Cambiar</a></p> <p><b>Grupos de seguridad para principal:</b> <a href="#">sg-02e8a471da0bd7c86</a> <a href="#">(ElasticMapReduce-master)</a></p> <p><b>Grupos de seguridad para principal y tarea:</b> <a href="#">sg-0c4674169004076f7</a> <a href="#">(ElasticMapReduce-slave)</a></p>	

### 3.5. Elija ElasticMapReduce-master en la lista.



### 3.6. Elija Inbound (Entrada), Edit (Editar).



**3.7. Compruebe si hay una regla de entrada que permite el acceso público con la siguiente configuración. Si existe, elija Delete (Eliminar) para eliminarla**

Reglas de entrada <small>Información</small>					
Tipo <small>Información</small>	Protocolo <small>Información</small>	Intervalo de puertos <small>Información</small>	Origen <small>Información</small>	Descripción: opcional <small>Información</small>	
Todos los TCP	TCP	0 - 65535	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-02e8a471da0bd7c86"/>		
Todos los TCP	TCP	0 - 65535	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-0c4674169004076f7"/>		
TCP personalizado	TCP	8443	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="207.171.167.25/32"/>		
TCP personalizado	TCP	8443	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="54.240.217.8/29"/>		
TCP personalizado	TCP	8443	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="72.21.196.64/29"/>		
TCP personalizado	TCP	8443	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="72.21.198.64/29"/>		

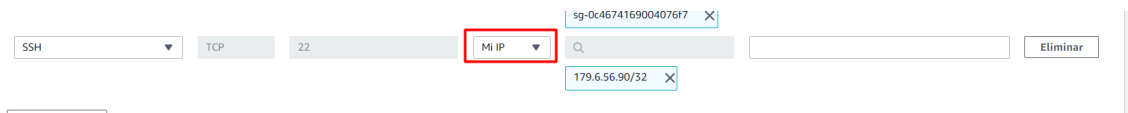
**3.8. Desplácese a la parte inferior de la lista y elija Add Rule (Añadir regla).**

TCP personalizado	TCP	8443	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="207.171.172.6/32"/>		
Todos los UDP	UDP	0 - 65535	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-02e8a471da0bd7c86"/>		
Todos los UDP	UDP	0 - 65535	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-0c4674169004076f7"/>		
Todos los ICMP IPv4	ICMP	Todo	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-02e8a471da0bd7c86"/>		
Todos los ICMP IPv4	ICMP	Todo	Persona... <input type="text" value="Q"/>		Eliminar
			<input type="text" value="sg-0c4674169004076f7"/>		

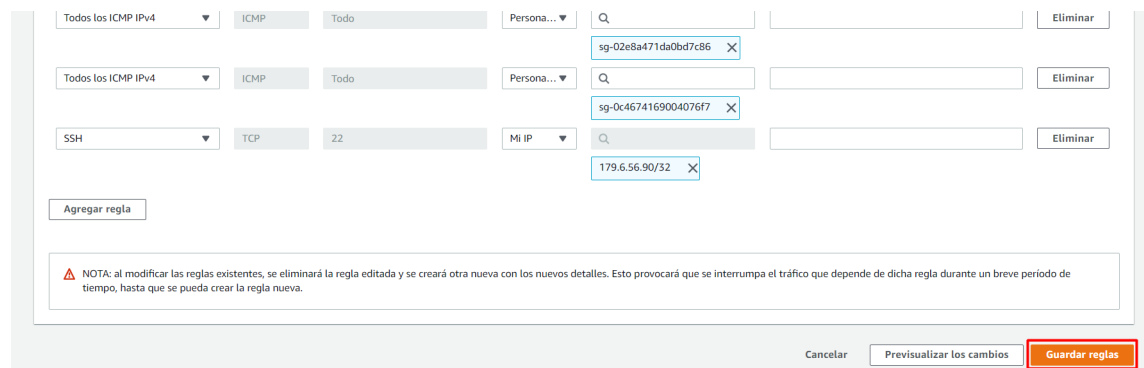
**3.9. En Type (Tipo), seleccione SSH. Esto introduce automáticamente TCP para Protocol (Protocolo) y 22 para Port Range (Rango de puertos).**



**3.10. 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.**



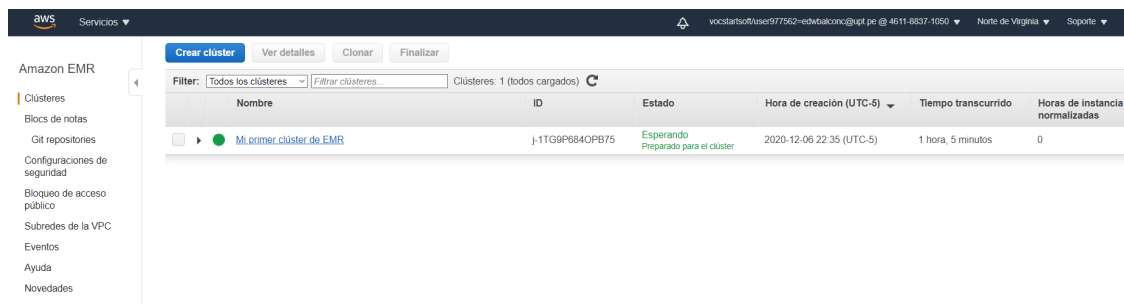
**3.11. Elija Save (Guardar).**



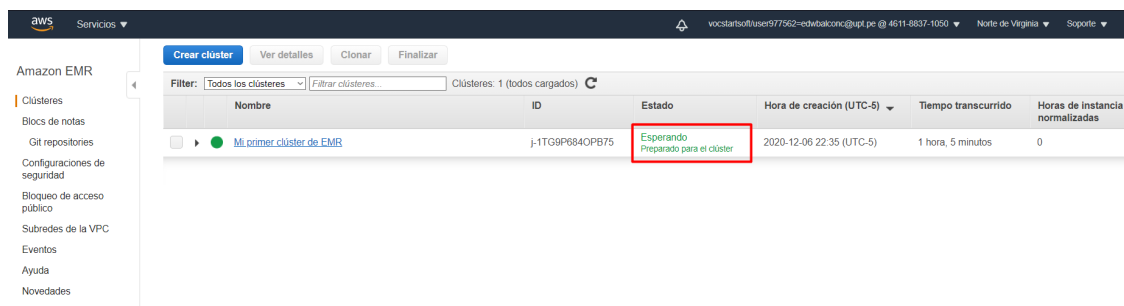
NOTA: al modificar las reglas existentes, se eliminará la regla editada y se creará otra nueva con los nuevos detalles. Esto provocará que se interrumpa el tráfico que depende de dicha regla durante un breve periodo de tiempo, hasta que se pueda crear la regla nueva.

## 4. Procesar los datos ejecutando el script de Hive como paso

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



4.2. En Cluster List (Lista de clústeres), seleccione el nombre del clúster. Asegúrese de que el clúster está en el estado Waiting (Esperando).



#### 4.3. Elija Steps (Pasos) y, a continuación Add step (Añadir paso).

Amazon EMR

Clúster: Mi primer clúster de EMR **Esperando** Cluster ready after last step completed.

Resumen

ID: j-1TG9P684OPB75

Fecha de creación: 2020-12-06 22:35 (UTC-5)

Tiempo transcurrido: 1 hora, 33 minutos

Terminar automáticamente: Cluster waits

Protección contra la Desactivación: [Cambiar](#)

Etiquetas: -- Ver todo / Editar

DNS público principal: ec2-3-83-148-68.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

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 0.8.2

URI de registro: s3://aws-logs-461188371050-us-east-1/elasticmapreduce/

Vista coherente de EMRFS: Deshabilitados

ID de AMI personalizada: --

Application user interfaces

Servicio de historial: ☒ YARN timeline server, Tez UI

Conexiones: ☒ Not Enabled [Habilitar conexión web](#)

Seguridad y acceso

Nombre de la clave: BDII

Perfil de instancia EC2: EMR\_EC2\_DefaultRole

Función de EMR: EMR\_DefaultRole

Visible para todos los usuarios: [Cambiar](#)

Redes y hardware

Zona de disponibilidad: us-east-1c

ID de subred: subnet-e077c7c1

Maestro: En ejecución 1 m5.xlarge

Principal: En ejecución 2 m5.xlarge

Tarea: --

Cluster scaling: Not enabled

#### 4.4. Configure el paso de acuerdo con las directrices siguientes:

**Añadir paso**

Tipo de paso: Programa de Hive

Nombre: Programa de Hive

Ubicación S3 del script\*: s3://uswest-2.elasticmapreduce.samples/cloudfront/c La ubicación S3 del script de Hive.  
s3://<nombre del bucket>/<ruta al archivo>

Ubicación S3 de entrada: s3://uswest-2.elasticmapreduce.samples La ubicación S3 de los archivos de entrada de Hive.  
s3://<nombre del bucket>/<carpeta>/

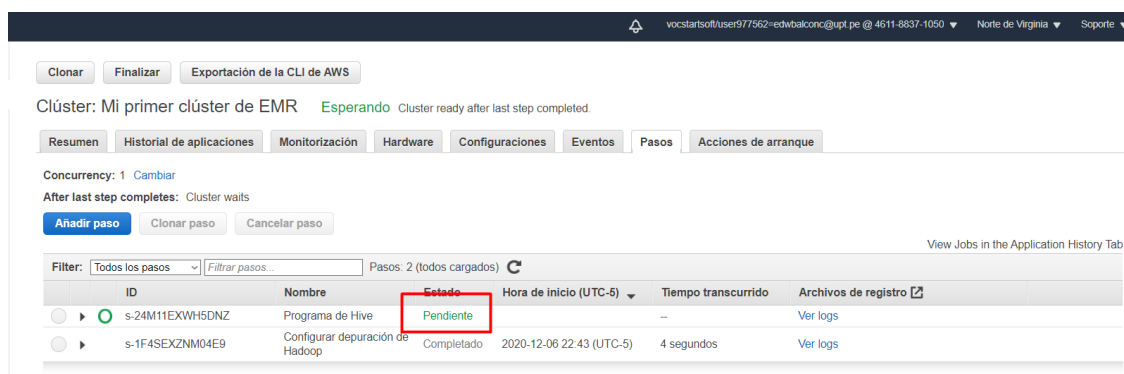
Ubicación S3 de salida: s3://awsbucket-bdii/ La ubicación S3 de los archivos de entrada de Hive.  
s3://<nombre del bucket>/<carpeta>/

Argumentos: Especifique los argumentos opcionales para su script.

Acción sobre el error: Continuar Qué hacer si se produce un error en el paso.

[Cancelar](#) [Añadir](#)

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



Clúster: Mi primer clúster de EMR **Esperando** Cluster ready after last step completed.

Resumen Historial de aplicaciones Monitorización Hardware Configuraciones Eventos **Pasos** Acciones de arranque

Concurrency: 1 [Cambiar](#)  
After last step completes: Cluster waits

[Añadir paso](#) [Clonar paso](#) [Cancelar paso](#) [View Jobs in the Application History Tab](#)

Filter:	ID	Nombre	Estado	Hora de inicio (UTC-5)	Tiempo transcurrido	Archivos de registro
Todos los pasos	s-24M11EXWH5DNZ	Programa de Hive	Pendiente	--	--	<a href="#">Ver logs</a>
	s-1F4SEXZNM04E9	Configurar depuración de Hadoop	Completado	2020-12-06 22:43 (UTC-5)	4 segundos	<a href="#">Ver logs</a>

#### 4.6. 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.



Clúster: Mi primer clúster de EMR **Esperando** Cluster ready after last step completed.

Resumen Historial de aplicaciones Monitorización Hardware Configuraciones Eventos **Pasos** Acciones de arranque

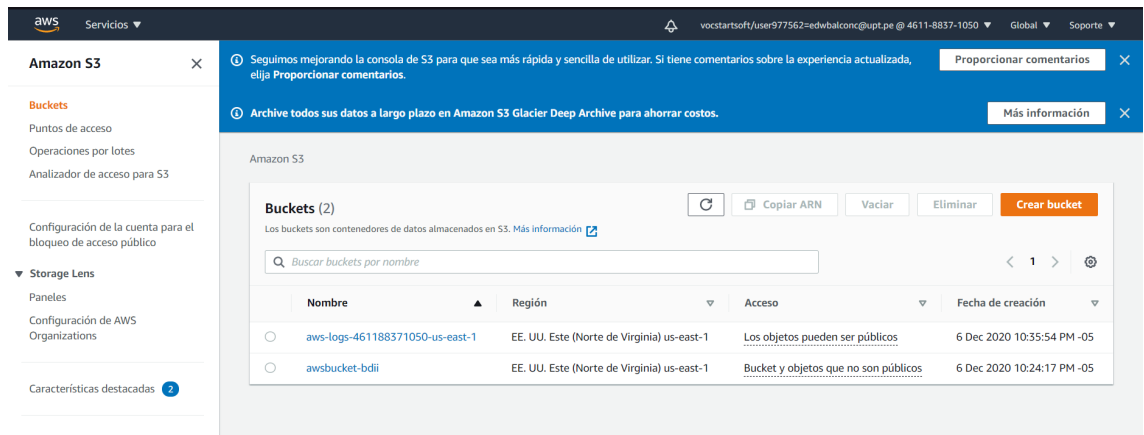
Concurrency: 1 [Cambiar](#)  
After last step completes: Cluster waits

[Añadir paso](#) [Clonar paso](#) [Cancelar paso](#) [View Jobs in the Application History Tab](#)

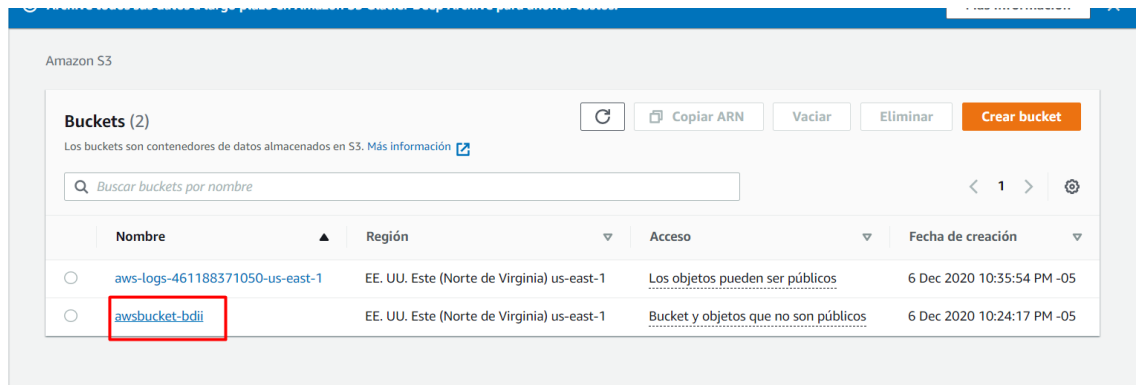
Filter:	ID	Nombre	Estado	Hora de inicio (UTC-5)	Tiempo transcurrido	Archivos de registro
Todos los pasos	s-34RJM965IRI3T	Programa de Hive	Completado	2020-12-07 00:21 (UTC-5)	48 segundos	<a href="#">Ver logs</a>
	s-1F4SEXZNM04E9	Configurar depuración de Hadoop	Completado	2020-12-06 22:43 (UTC-5)	4 segundos	<a href="#">Ver logs</a>

## 5. Ver los resultados Una vez que el paso se completa correctamente

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



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





5.3. 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 `000000_0`.

Objetos

Propiedades Permisos Métricas Administración Puntos de acceso

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

**Objetos (2)**

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

Eliminar Acciones ▼ Crear carpeta **Cargar**

Q Buscar objetos por prefijo

<input type="checkbox"/>	Nombre ▲	Tipo ▼	Última modificación ▼	Tamaño ▼	Clase de almacenamiento ▼
<input type="checkbox"/>	os_requests_\$folder\$	-	7 Dec 2020 12:21:43 AM -05	0 B	Estándar
<input type="checkbox"/>	os_requests/	Carpeta	-	-	-

5.4. Elija el archivo y, a continuación, elija **Download (Descargar)** para guardarlo localmente.

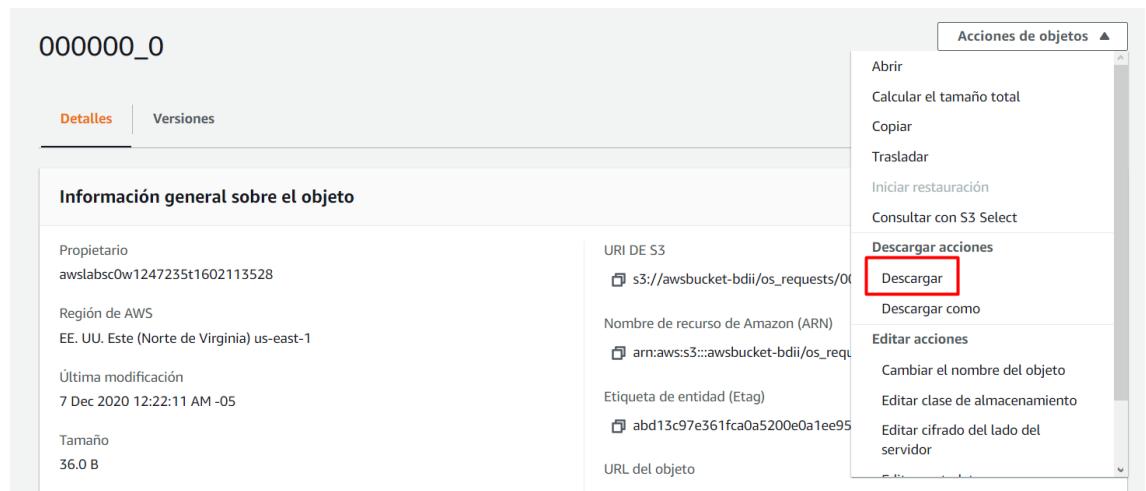
**Objetos (2)**

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

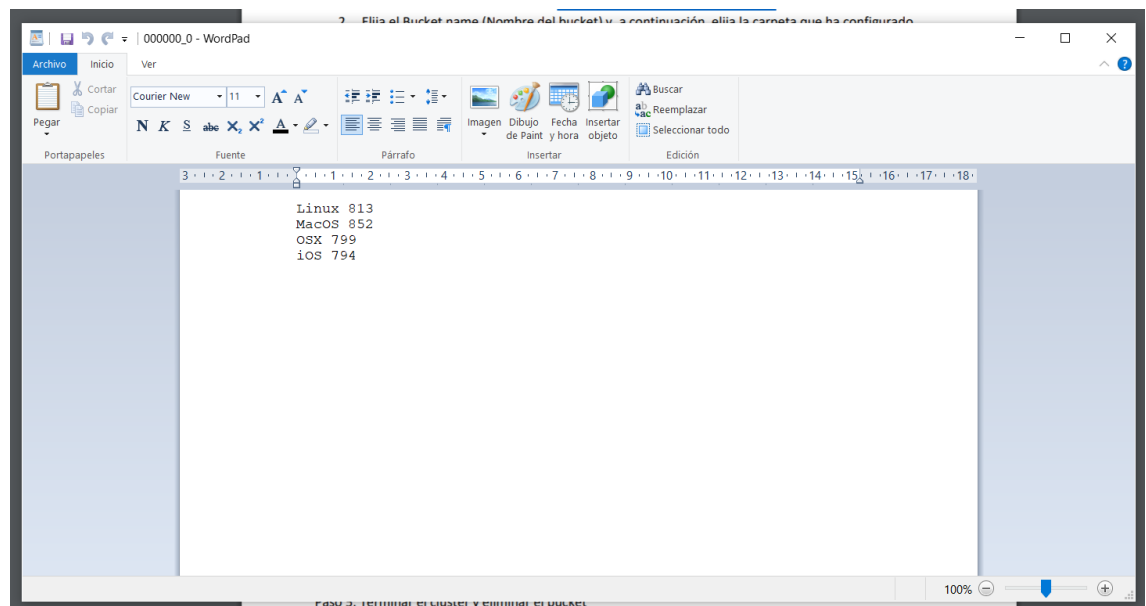
Eliminar Acciones ▼ Crear carpeta **Cargar**

Q Buscar objetos por prefijo

<input type="checkbox"/>	Nombre ▲	Tipo ▼	Última modificación ▼	Tamaño ▼	Clase de almacenamiento ▼
<input type="checkbox"/>	000000_0	-	7 Dec 2020 12:22:11 AM -05	36.0 B	Estándar
<input type="checkbox"/>	000001_0	-	7 Dec 2020 12:22:11 AM -05	24.0 B	Estándar

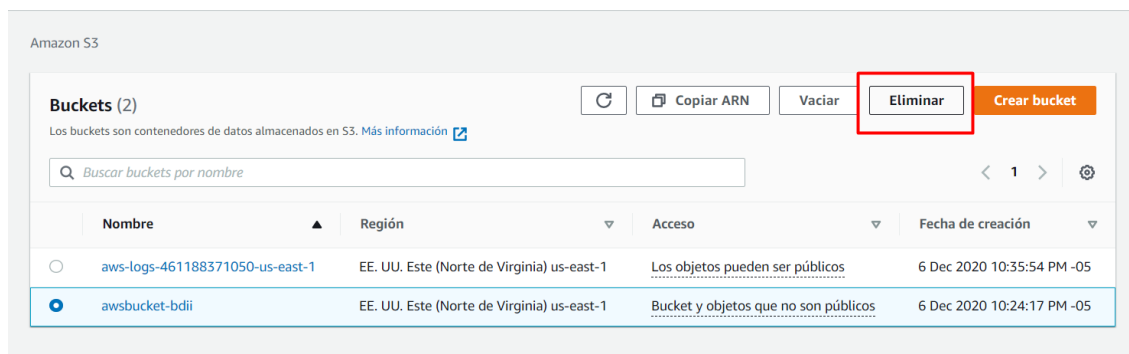


**5.5. Utilice el editor de texto que prefiera para abrir el archivo. El archivo de salida muestra el número de solicitudes de acceso ordenadas por sistema operativo. El siguiente ejemplo muestra la salida en WordPad:**



## 6. Terminar el clúster y eliminar el bucket

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



6.2. Elija Clusters (Clústeres), elija el clúster y, a continuación, Termine (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).

