

# Laboratorio N° 03: Introducción a big data con Amazon EMR

Derian Francisco Herrera Amezcuita (2017059489)

*Tacna, Perú*

---

## 1. Título

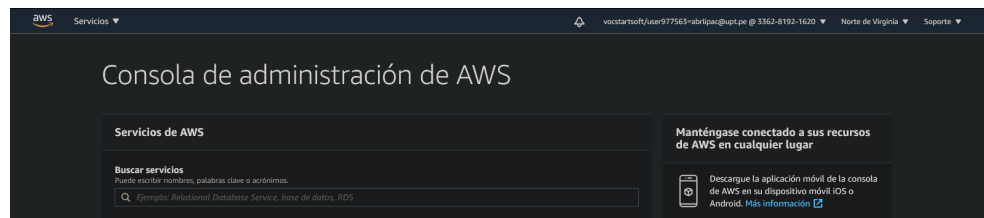
INFORME DE LABORATORIO N° 03  
Introducción a big data con Amazon EMR

## 2. Autores

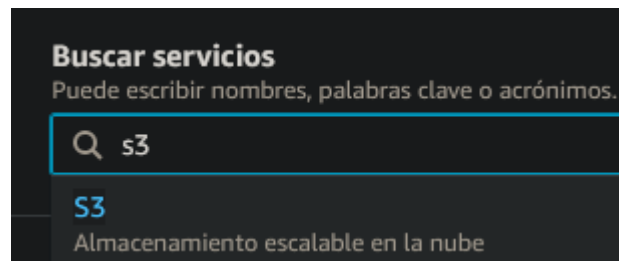
- Derian Francisco Herrera Amezcuita

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

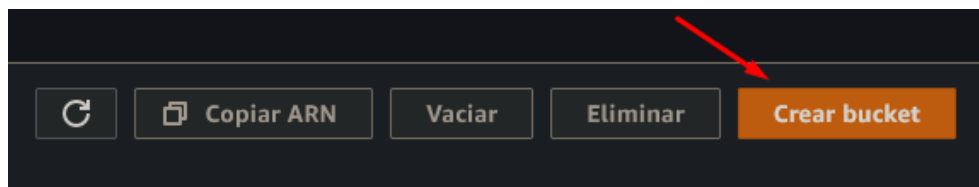
Iniciamos sesión en AWS Educate y entramos en la consola de administración de AWS.



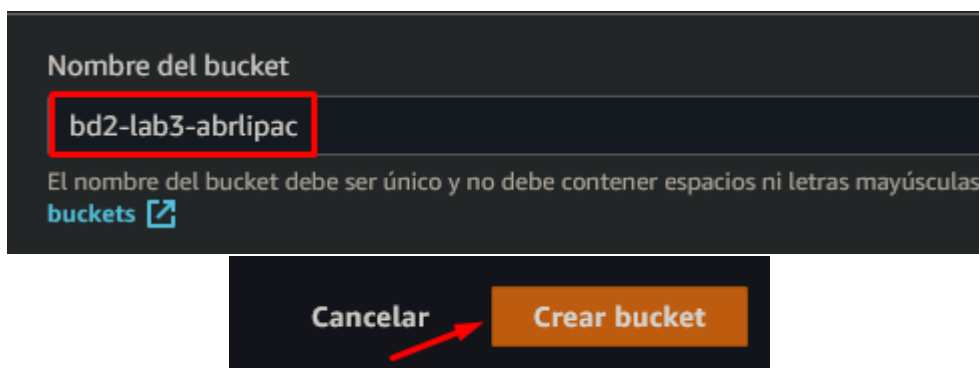
Buscamos el servicio de Amazon S3.



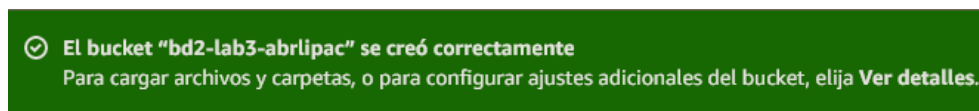
Una vez en la administración de S3, hacemos clic en Crear bucket.



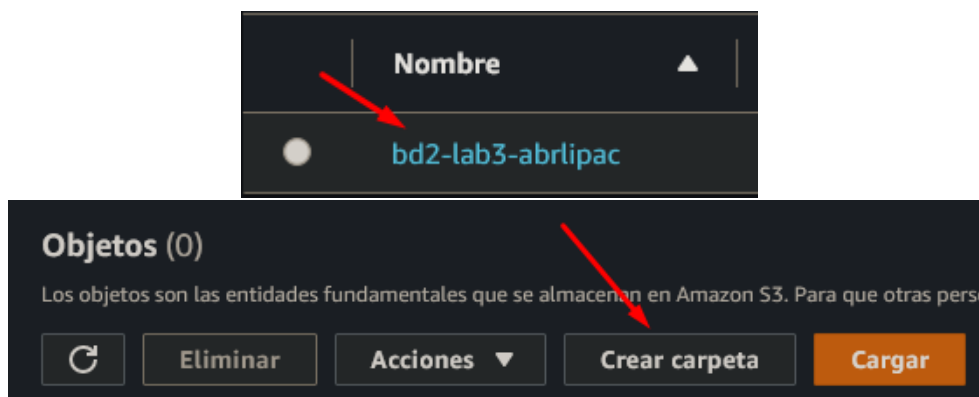
Especificamos un nombre y hacemos clic en Crear bucket.



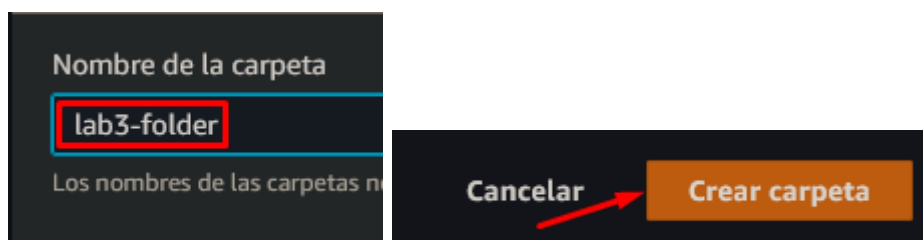
Aparecerá un mensaje que indica que se realizó la operación.



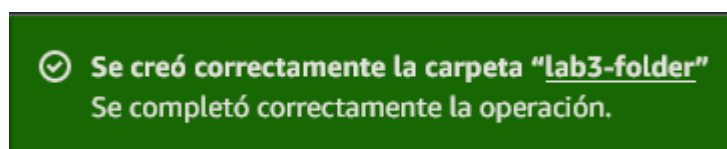
Hacemos clic en el bucket que hemos creado y luego hacemos clic en Crear carpeta.



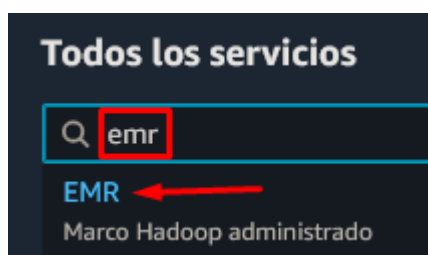
Especificamos el nombre de la carpeta y hacemos clic en Crear carpeta.



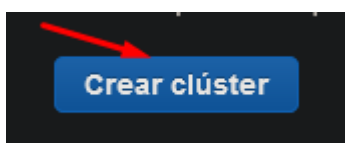
Veremos un mensaje que indica la operación exitosa.



Buscamos y abrimos el servicio de EMR.



Hacemos clic en crear cluster.



Especificamos el nombre único del cluster, el par de claves de EC2 (que hemos creado previamente) y hacemos clic en crear cluster.

Nombre del clúster: DB2 Lab3 ERM Cluster

Par de claves EC2: abrlipac\_key

Permisos: abrlipac\_key

Cancelar | Crear clúster

Veremos que el estado del Cluster es Comenzando.

Clúster: DB2 Lab3 ERM Cluster **Comenzando**

Abrimos el cluster que hemos creado.

[DB2 Lab3 ERM Cluster](#)

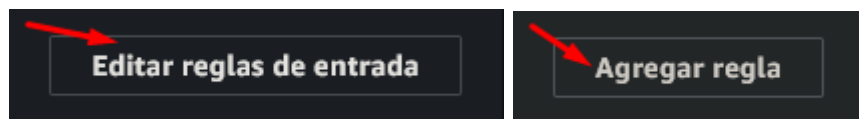
Accedemos al grupo de seguridad (correspondiente a master) y seleccionamos el grupo de seguridad ElasticMapReduce-master.

Grupos de seguridad para <b>sg-00252c8c2e0c949e7</b> (ElasticMapReduce-principal: master)	
ID del grupo de segu...	Nombre del grupo de s...
<b>sg-00252c8c2e0c949e7</b>	ElasticMapReduce-master

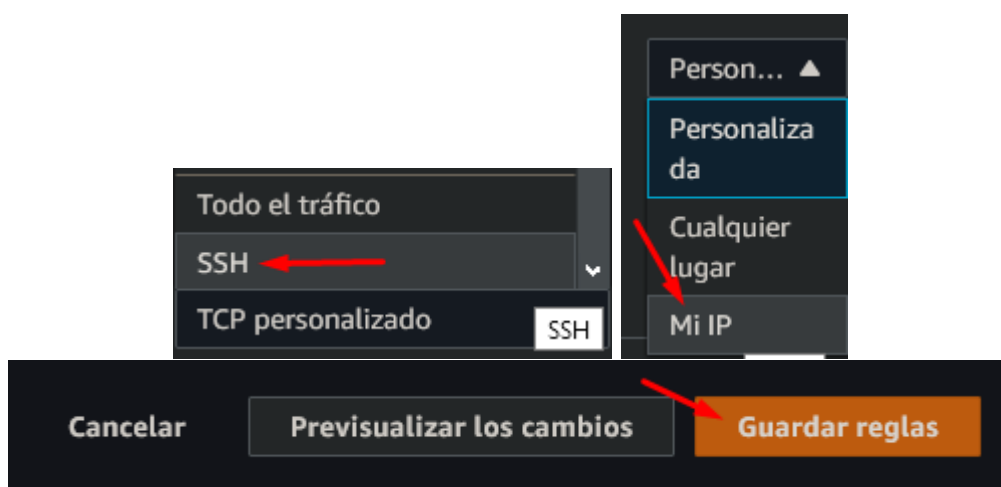
Nos aseguramos que estamos en la sección de Reglas de entrada.

**Reglas de entrada** | Reglas de salida

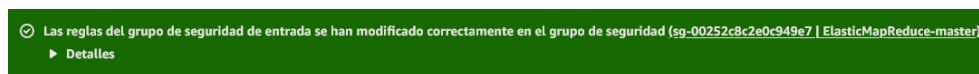
Hacemos clic en Editar reglas de entrada y luego en Agregar regla.



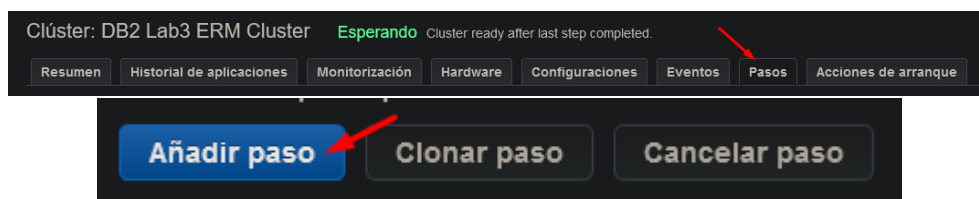
Especificamos el tipo de conexión como SSH y elegimos como origen My IP (es decir la IP actual de la conexión). Luego hacemos clic en Guardar reglas.



Veremos un mensaje que confirma la operación.



Nos ubicamos en el cluster y hacemos clic en la sección Pasos y hacemos clic en Añadir paso.



Especificamos el tipo de paso como Programa de Hive y el nombre del paso.

**Tipo de paso** Programa de Hive

**Nombre** JAR personalizado

**S3 del script\*** Programa de retransmisión

**S3 de entrada** Programa de Hive

**Nombre** First Hive Program

Agregamos las siguientes URIs:

- Ubicación S3 del script: s3://us-east-1.elasticmapreduce.samples / cloudfront/code/Hive\_CloudFront.q
- Ubicación S3 de entrada: s3://us-east-1.elasticmapreduce.samples
- Ubicación S3 de salida: s3://bd2-lab3-abrliapac/lab3-folder/

**Ubicación S3 del script\*** s3://us-east-1.elasticmapreduce.samples/cloudfront/co  
s3://<nombre del bucket>/<ruta al archivo>

**Ubicación S3 de entrada** s3://us-east-1.elasticmapreduce.samples  
s3://<nombre del bucket>/<carpeta>/

**Ubicación S3 de salida** s3://bd2-lab3-abrliapac/lab3-folder/  
s3://<nombre del bucket>/<carpeta>/

Nos aseguramos que la Acción sobre el error sea Continuar.

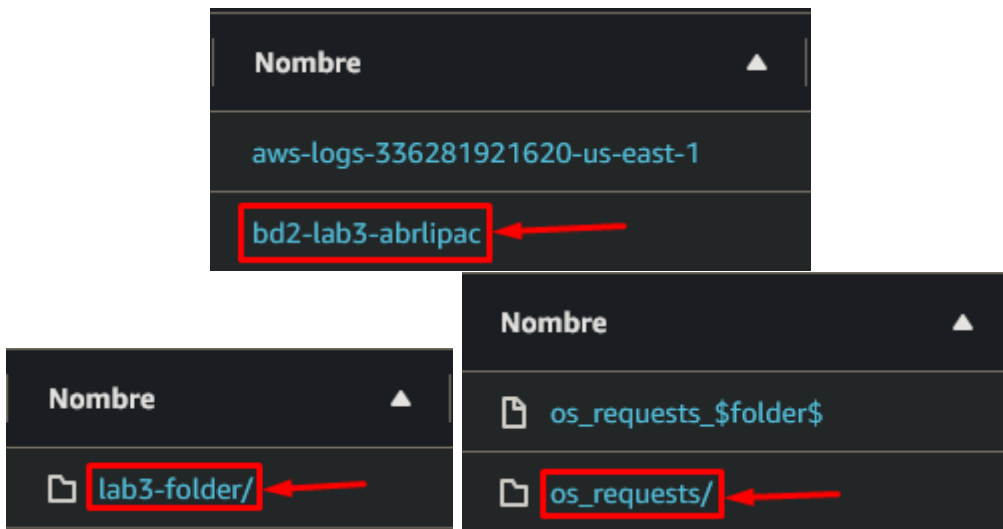
**Acción sobre el error** Continuar

Hacemos clic en Añadir y veremos que el programa tendrá un estado Pendiente.

Cancelar Añadir

Nombre	Estado
Programa de Hive	Pendiente

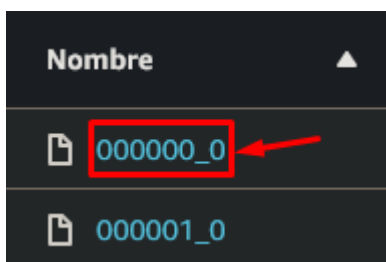
Abrimos Amazon S3 y elegimos el Bucket que hemos creado. Entramos a la carpeta, podremos ver y hacer clic en la carpeta `os_request/`.



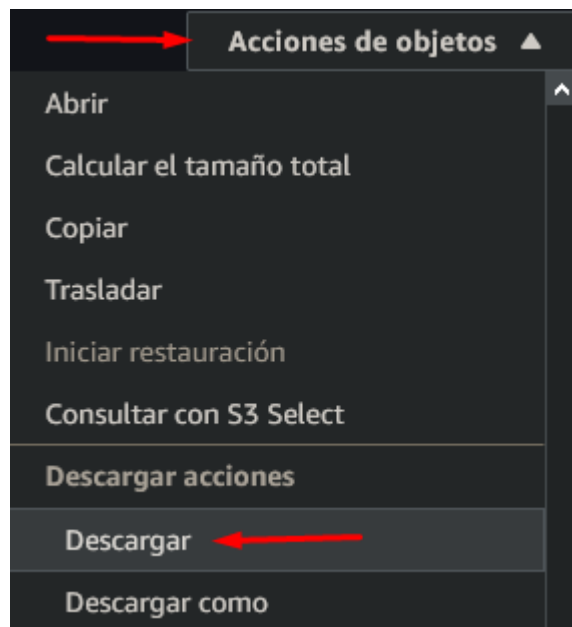
Veremos dos ficheros.

- 000000\_0
- 000001\_0

Hacemos clic en el primero.



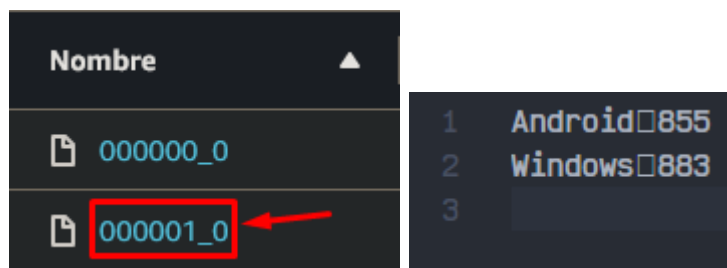
Nos dirigimos a Acciones de objetos y luego hacemos clic en Descargar.



Se descargará el fichero (sin extensión) y si la abrimos veremos el siguiente contenido.

```
1 Linux 813
2 MacOS 852
3 OSX 799
4 iOS 794
5 |
```

Realizamos el mismo proceso con el otro fichero y si lo abrimos , veremos el siguiente contenido.

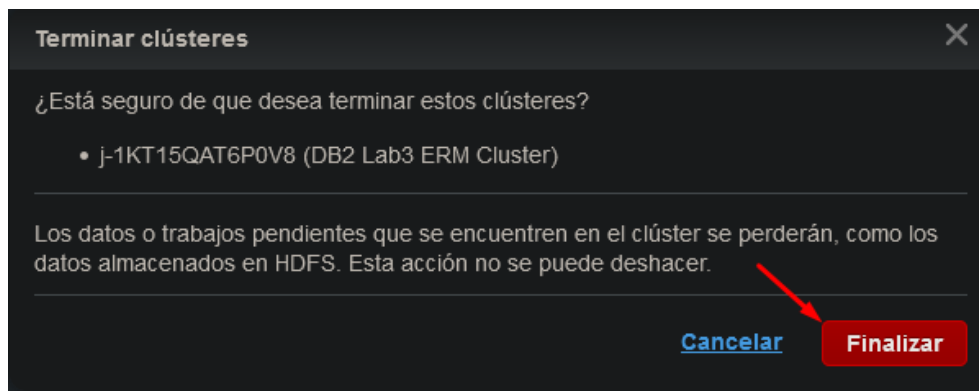




Para terminar con el cluster, seleccionamos el cluster que hemos creado y luego hacemos clic en Finalizar.



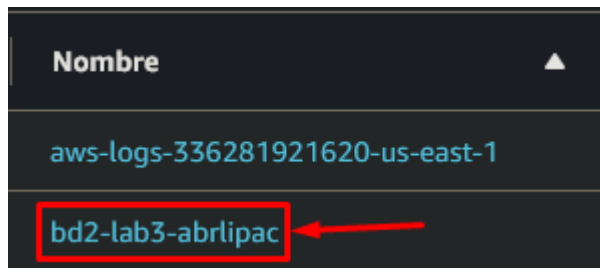
Veremos un cuadro de confirmación y hacemos clic en Finalizar.



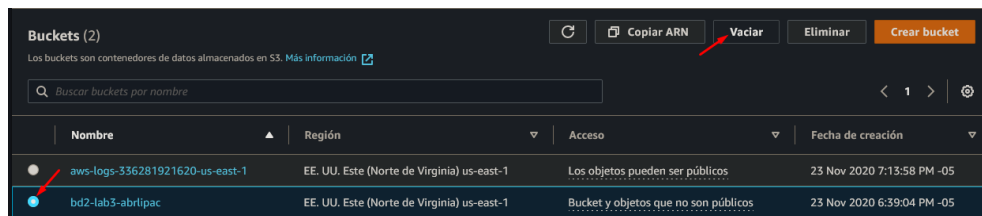
El cluster tendrá un estado Terminando.

ID	Estado
j-1KT15QAT6P0V8	Terminando Solicitud del usuario

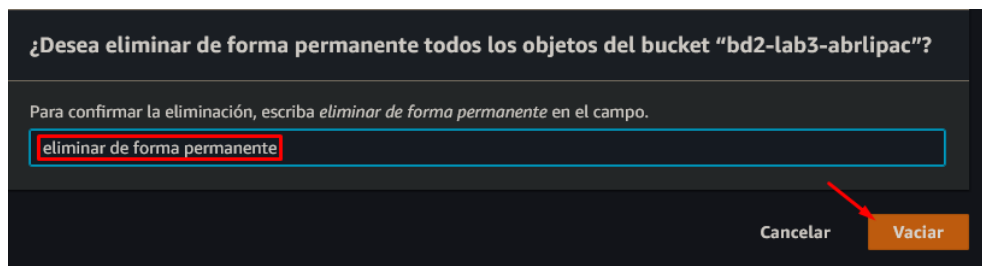
Para eliminar el Bucket que contiene la carpeta con los log que hemos generado, abrimos Amazon S3 y hacemos clic en el bucket creado.



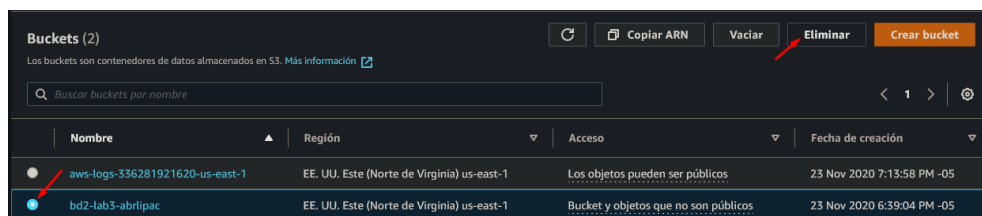
Seleccionamos el bucket y hacemos clic en Vaciar.



Escribimos el texto de confirmación requerido y hacemos clic en Vaciar.



Luego de vaciar todos los elementos del Bucket, lo seleccionamos nuevamente y hacemos clic en Eliminar.



Escribimos el texto de confirmación requerido y hacemos clic en Eliminar.

**Eliminar bucket bd2-lab3-abrlipac**

Escriba el nombre del bucket para confirmar que desea eliminarlo.

Cancelar Eliminar bucket

Veremos un mensaje que confirma la operación de eliminación.

