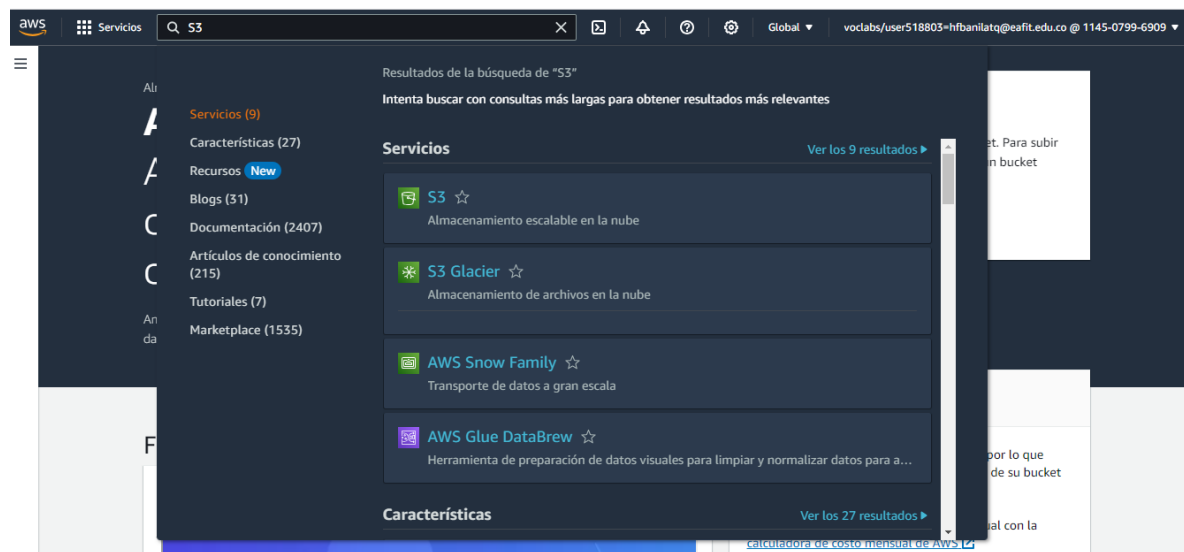


Laboratorio 3.0: Instalación de un cluster EMR en AWS

Para desarrollar este laboratorio necesitamos acceso a una cuenta de AWS, debemos iniciar sesión y una vez allí necesitaremos crear:

1. Un bucket de s3 para almacenar la información del cluster, logs y demás.
2. Un cluster EMR
3. Configurar el grupo de seguridad para habilitar los puertos necesarios de las aplicaciones a internet
4. Configurar el nodo maestro del cluster para solucionar un error de puertos.

Para crear un bucket de s3 debemos poner en el buscador "S3" y seleccionar el servicio de S3



Una vez seleccionado el servicio de S3 debemos dar click en crear bucket



El paso anterior nos enviará a la página de creación del bucket donde debes darle un nombre, Seleccionar la región y la propiedad del bucket como sigue

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

Configuración general

Nombre del bucket

El nombre del bucket debe ser único dentro del espacio de nombres global y seguir las reglas de nomenclatura del bucket. [Consulte las reglas para la asignación de nombres de buckets](#)

Región de AWS

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

Propiedad de objetos [Información](#)

Controle la propiedad de los objetos escritos en este bucket desde otras cuentas de AWS y el uso de listas de control de acceso (ACL). La propiedad de los objetos determina quién puede especificar el acceso a los objetos.

☒ ACL deshabilitadas (recomendado) ☐ ACL habilitadas

Propiedad de objetos [Información](#)

Controle la propiedad de los objetos escritos en este bucket desde otras cuentas de AWS y el uso de listas de control de acceso (ACL). La propiedad de los objetos determina quién puede especificar el acceso a los objetos.

☒ ACL deshabilitadas (recomendado)
Todos los objetos de este bucket son propiedad de esta cuenta. El acceso a este bucket y sus objetos se especifica solo mediante políticas.

☐ ACL habilitadas
Los objetos de este bucket pueden ser propiedad de otras cuentas de AWS. El acceso a este bucket y sus objetos se puede especificar mediante ACL.

Propiedad del objeto

Aplicada al propietario del bucket

Configuración de bloqueo de acceso público para este bucket

Se concede acceso público a los buckets y objetos a través de listas de control de acceso (ACL), políticas de bucket, políticas de puntos de acceso o todas las anteriores. A fin de garantizar que se bloquee el acceso público a todos sus buckets y objetos, active Bloquear todo el acceso público. Esta configuración se aplica exclusivamente a este bucket y a sus puntos de acceso. AWS recomienda activar Bloquear todo el acceso público, pero, antes de aplicar cualquiera de estos ajustes, asegúrese de que las aplicaciones funcionarán correctamente sin acceso público. Si necesita cierto nivel de acceso público a los buckets u objetos, puede personalizar la configuración individual a continuación para adaptarla a sus casos de uso de almacenamiento específicos. [Más información](#)

☒ **Bloquear todo el acceso público**
Activar esta configuración equivale a activar las cuatro opciones que aparecen a continuación. Cada uno de los siguientes ajustes son independientes entre sí.

☒ **Bloquear el acceso público a buckets y objetos concedido a través de nuevas listas de control de acceso (ACL)**
S3 bloqueará los permisos de acceso público aplicados a objetos o buckets agregados recientemente, y evitará la creación de

Una vez configurada esta información darle clic en crear bucket

Cifrado predeterminado [Información](#)

El cifrado del lado del servidor se aplica automáticamente a los nuevos objetos almacenados en este bucket.

Tipo de cifrado [Información](#)

- ☒ Cifrado del servidor con claves administradas de Amazon S3 (SSE-S3)
- ☐ Cifrado del servidor con claves de AWS Key Management Service (SSE-KMS)
- ☐ Cifrado de doble capa del servidor con claves de AWS Key Management Service (DSSE-KMS)
Proteja sus objetos con dos capas de cifrado independientes. Para obtener más información sobre los precios, consulte DSSE-KMS pricing (Precios de DSSE-KMS) en la pestaña Storage (Almacenamiento) de la [página de precios de Amazon S3](#).

Clave de bucket

El uso de una clave de bucket de S3 para SSE-KMS reduce los costos de cifrado al reducir las llamadas a AWS KMS. Las claves de bucket de S3 no son compatibles con DSSE-KMS. [Más información](#)

☐ Desactivar

☒ Habilitar

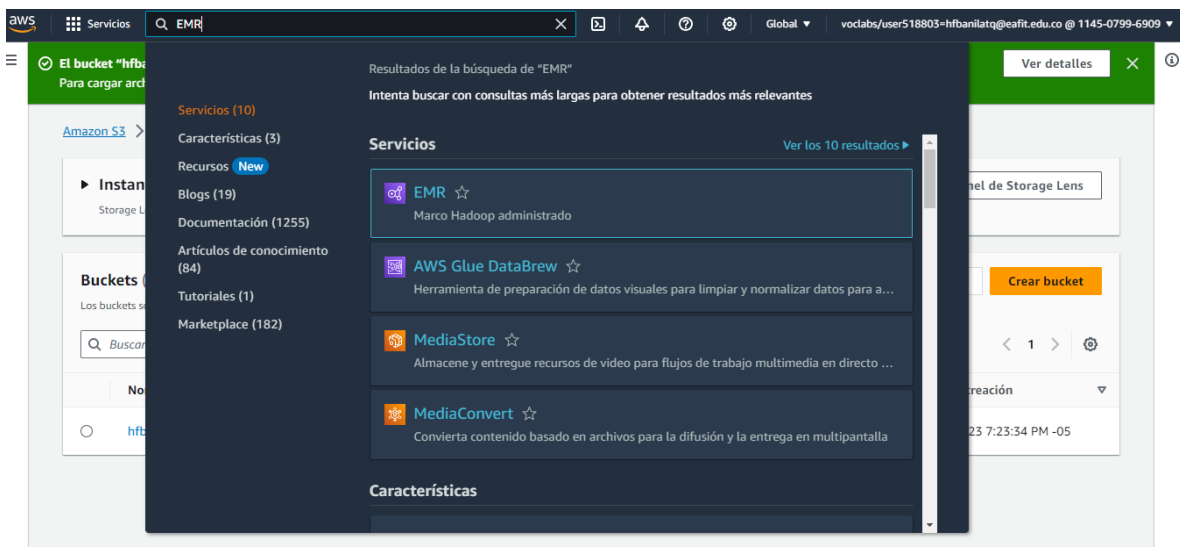
► Configuración avanzada

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

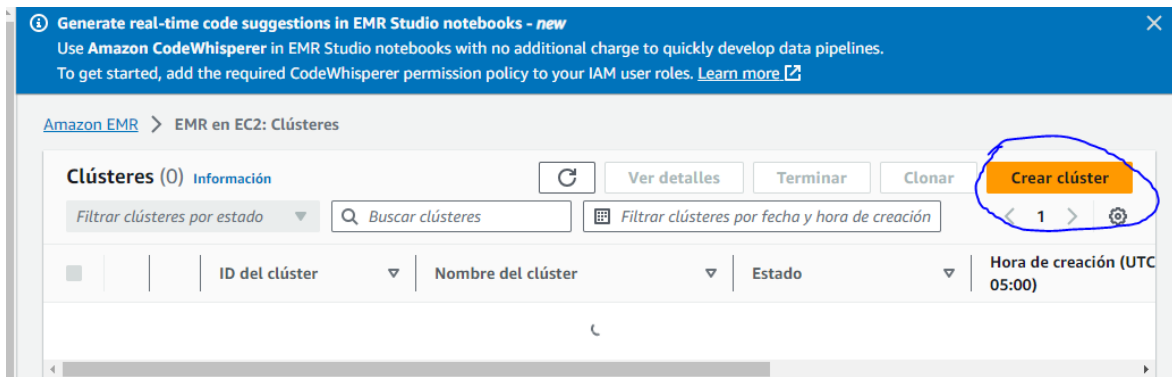
Cancelar **Crear bucket**

Esto debería crear el bucket de S3 y nos permitirá continuar al siguiente paso

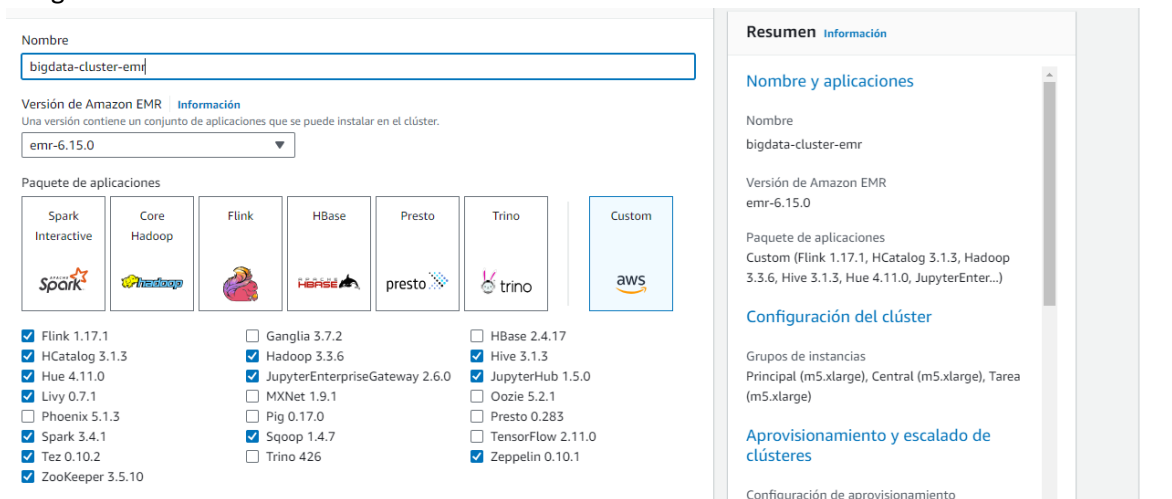
Crear un cluster de EMR: Para realizar este paso necesitaremos ir al panel de control del servicio EMR para ello buscaremos EMR y seleccionaremos el servicio



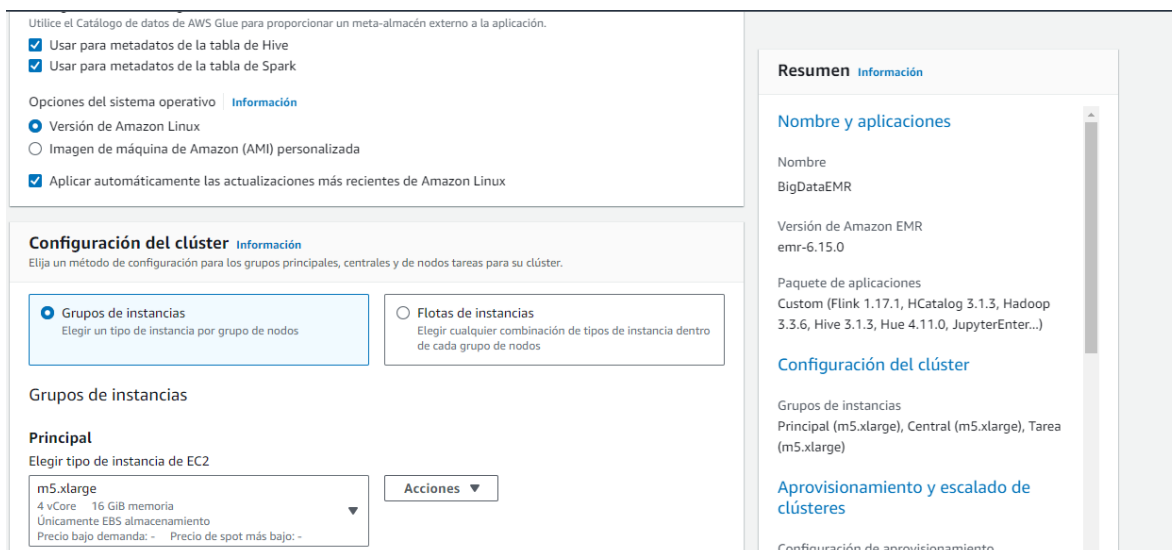
En este panel veremos un botón que nos enviará a la página de creación del cluster damos clic allí



En esta página asignamos un nombre al clúster y seleccionamos las aplicaciones que se ven en la imagen



La siguiente información debes configurarla igual, el tamaño mínimo sugerido para las maquinas de EC2 son m5.xlarge



Central

Elegir tipo de instancia de EC2

m5.xlarge

4 vCore 16 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

m5.xlarge

4 vCore 16 GiB memoria

Únicamente EBS almacenamiento

Precio bajo demanda: - Precio de spot más bajo: -

Acciones ▼

► Configuración de nodo - opcional

Agregar grupo de instancias de tareas

Puede agregar hasta 47 grupos más de instancias de tareas.

Resumen Información

Nombre y aplicaciones

Nombre

BigDataEMR

Versión de Amazon EMR

emr-6.15.0

Paquete de aplicaciones

Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnter...)

Configuración del clúster

Grupos de instancias

Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge)

Aprovisionamiento y escalado de clústeres

Configuración de aprovisionamiento

En mi caso seleccione un escalado administrado por AWS pero puedes dejarlo por defecto

Aprovisionamiento y escalado de clústeres Información

Establezca las configuraciones de escalado y aprovisionamiento para los grupos de nodos principales y los nodos de tarea del 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 escalado

Tamaño mínimo del clúster

1

Instancias

Tamaño máximo del clúster

20

Instancias

Cantidad máxima de nodos principales en el clúster

Limite la cantidad de nodos principales en su clúster.

20

Instancias

Número máximo de instancias bajo demanda en el clúster

Si desea aprovisionar el nodo principal para utilizar los precios bajo demanda y otros nodos del clúster para utilizar los precios de spot, establezca este valor en 1. Si desea aprovisionar todo el clúster para utilizar los precios bajo demanda, utilice el mismo valor que el tamaño máximo del clúster.

20

Instancias

Resumen Información

Nombre y aplicaciones

Nombre

BigDataEMR

Versión de Amazon EMR

emr-6.15.0

Paquete de aplicaciones

Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnter...)

Configuración del clúster

Grupos de instancias

Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge)

Aprovisionamiento y escalado de clústeres

Configuración de escalado

Se selecciona la VPC y la subred, si no tienes VPC actualmente puedes crear una nueva al dar clic en Crear PVC, de igual forma para la subnet

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	m5.xlarge	1	<input type="checkbox"/>
Tarea - 1	m5.xlarge	1	<input type="checkbox"/>

Redes Información

Virtual Private Cloud (VPC) Información

vpc-0ed6bb1f7a91ba16 [Examinar](#) [Crear VPC](#)

Subred Información

subnet-0148ac41181da0a0e [Examinar](#) [Crear subred](#)

► Grupos de seguridad de EC2 (firewall)

▼ Pasos: *opcional* (0) Información [Eliminar](#) [Editar](#) [Agregar](#)

Utilice comandos y scripts para indicar a su clúster dónde encontrar y cómo procesar los datos. Los pasos se ejecutan de forma consecutiva a menos que habilite la opción Simultaneidad.

Resumen Información

Nombre y aplicaciones

Nombre
BigDataEMR

Versión de Amazon EMR
emr-6.15.0

Paquete de aplicaciones
Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnter...)

Configuración del clúster

Grupos de instancias
Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge)

Aprovisionamiento y escalado de clústeres

Configuración de escalado

Puedes asignar un grupo de seguridad que ya tengas, como es mi caso o en su defecto usar el que crea automáticamente AWS

▼ Grupos de seguridad de EC2 (firewall)

Nodo principal

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

Crear ElasticMapReduce-Primary ▼

Grupos de seguridad adicionales - *opcional*
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales ▼

taller3-tec-security-group X
sg-015151991aa02a64b

Nodos principales y de tareas

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

Crear ElasticMapReduce-Core ▼

Grupos de seguridad adicionales - *opcional*
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales ▼

taller3-tec-security-group X
sg-015151991aa02a64b

▼ Pasos: *opcional* (0) Información [Eliminar](#) [Editar](#) [Agregar](#)

Utilice comandos y scripts para indicar a su clúster dónde encontrar y cómo procesar los datos. Los pasos se ejecutan de forma consecutiva a menos que habilite la opción Simultaneidad.

Filtrar pasos por estado ▼ < 1 > ⚙

Resumen Información

Nombre y aplicaciones

Nombre
BigDataEMR

Versión de Amazon EMR
emr-6.15.0

Paquete de aplicaciones
Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnter...)

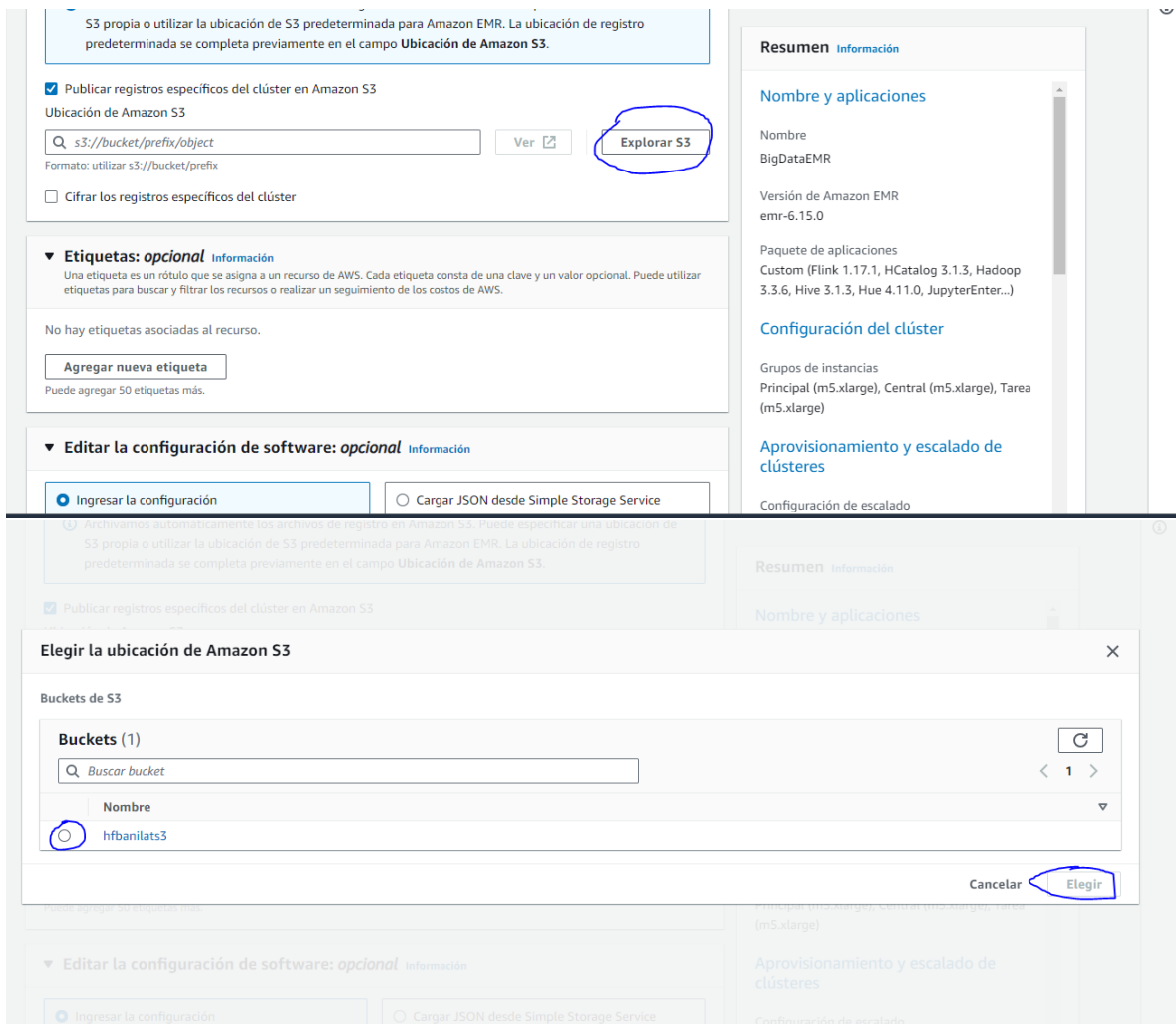
Configuración del clúster

Grupos de instancias
Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge)

Aprovisionamiento y escalado de clústeres

Configuración de escalado

Ahora vamos a relacionar el bucket de S3 al cluster, para ello damos clic en examinar y seleccionamos el bucket de S3 que creamos



Ahora debemos agregar configuración del software, para ello agregamos lo siguiente en el campo de texto:

```
[
{
  "Classification": "jupyter-s3-conf",
  "Properties": {
    "s3.persistence.bucket": "hfbanilats3",
    "s3.persistence.enabled": "true"
  }
}
]
```

En mi caso la propiedad "s3.persistence.bucket" es el nombre del bucket de s3 que creé, para tu caso debes poner el que hayas creado

▼ Editar la configuración de software: *opcional* [Información](#)

☒ Ingresar la configuración

☐ Cargar JSON desde Simple Storage Service (Amazon S3)

```
1 ▼ [  
2 ▼ {  
3   "Classification": "jupyter-s3-conf",  
4 ▼   "Properties": {  
5     "s3.persistence.enabled": "true",  
6     "s3.persistence.bucket": "hfbartilats3"  
7   }  
8 }  
9 ]
```

En los pares de claves debes seleccionar un par de claves si tienes, en caso de que no debes dar clic en crear par de claves, asignar un nombre y luego seleccionar ese par de claves

Configuración de seguridad

Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.

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

No ha especificado una clave EC2. Si está fuera de una VPN y desea habilitar SSH o utilizar el asistente Hue SQL con este clúster, debe escribir una clave EC2.

Par de claves

Un par de claves, compuesto por una clave privada y una clave pública, es un conjunto de credenciales de seguridad que se utilizan para demostrar su identidad cuando se conecta a una instancia.

Nombre

bigdata-key-pair

El nombre puede incluir hasta 255 caracteres ASCII. No puede incluir espacios al principio ni al final.

Tipo de par de claves [Información](#)

☒ RSA

☐ ED25519

Formato de archivo de clave privada

☒ .pem
Para usar con OpenSSH

☐ .ppk
Para usar con PuTTY

Etiquetas: *opcional*

No hay etiquetas asociadas a este recurso.

Agregar nueva etiqueta

Puede agregar hasta 50 etiquetas más.

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

Pares de claves (3)

Buscar pares de claves

ID	Name	Fingerprint
<input checked="" type="radio"/> key-0346b16e5422ee881	bigdata-key-pair	22:7e:27:6b:55:a5:06:47:47:a3:26:0c:be:7b:3c:73:94:e9:69:d9
<input type="radio"/> key-06fdd675d5509fe07	taller3-tec-key-pair	67:7c:c2:77:e7:06:4d:3b:50:88:71:3d:f0:3e:cf:f8:79:65:5a:50
<input type="radio"/> key-0f37ae482e0cc1c1a	vockey	84:6c:68:a3:aa:2a:23:aa:0f:77:51:64:f7:50:e7:db:73:ae:9f:d6

Cancelar **Elegir**

Debe seleccionar:

Service role: EMR_DefaultRole

Instance profile: EMR_EC2_DefaultRole

Custom automatic scaling role: LabRole

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
 Elegir rol de IAM

Rol de servicio

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

Rol de escalamiento automático personalizado - opcional

Cuando se activa una regla de escalamiento automático personalizada, Amazon EMR asume esta función para agregar y finalizar instancias de EC2. [Más información](#)

Rol de escalamiento automático personalizado
 EMR_AutoScaling_DefaultRole

Nombre y aplicaciones

Nombre
 BigDataEMR

Versión de Amazon EMR
 emr-6.15.0

Paquete de aplicaciones
 Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnter...)

Configuración del clúster

Grupos de instancias
 Principal (m5.xlarge), Central (m5.xlarge), Tarea (m5.xlarge)

Aprovisionamiento y escalado de clústeres

Configuración de escalado
 Tamaño mínimo del clúster: 1 instancia

Y ya simplemente dar clic en crear cluster

Una vez creado el cluster debes esperar que esté en estado “Esperando”, puede dardar hasta 20 min en completarse

<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	j-2XETAT8QJ7DXP	bigdata-cluster-emr	✓ Esperando	22 de noviembre de 2023
					Listo para ejecutar pasos	19:23

Ahora vamos a abrir los puertos necesarios en el grupo de seguridad, para ello debes primero en el panel de control de EMR en el menú izquierdo seleccionar “Bloquear acceso publico” en este lado debes desactivarlo y luego dirigirte al panel de EC2.

En el panel de EC2 busca la instancia en ejecución cuyo grupo de seguridad tenga la palabra master

Instancias (1/2) Información 🔄 Conectar Estado de la instancia ▼ Acciones ▼ Lanzar instancias ▼

🔍 Buscar Instance por atributo o etiqueta (case-sensitive)

Estado de la instancia = running ✕ Quitar los filtros < 1 > ⚙️

	Dirección IP...	IP elástica	Direcciones I...	Monitoreo	Nombre del grupo d...	Nombre de...
amazonaws.com	18.232.66.159	–	–	disabled	ElasticMapReduce-master	bigdata-key-pair
1.amazonaws.com	34.201.130.172	–	–	disabled	ElasticMapReduce-slave	bigdata-key-pair

Instancia: i-003db18a74ad478b6 ⚙️ ✕

Detalles Seguridad Redes Almacenamiento Comprobaciones de estado Monitoreo Etiquetas

▼ Resumen de instancia Información

ID de la instancia i-003db18a74ad478b6	Dirección IPv4 pública 18.232.66.159 dirección abierta	Direcciones IPv4 privadas 10.0.3.62
Dirección IPv6 –	Estado de la instancia En ejecución	DNS de IPv4 pública ec2-18-232-66-159.compute-1.amazonaws.com dirección abierta

Dale clic en el y dirígete a la parte de seguridad:

Seguridad

Dirección IP asignada automáticamente 18.232.66.159 [IP pública]	ID de VPC vpc-0edb6bb1f7a91ba16 (taller3-tec-vpc)	Hallazgo de AWS Compute Optimizer Suscribirse a AWS Compute Optimizer para recibir recomendaciones. Más información
Rol de IAM EMR_EC2_DefaultRole	ID de subred subnet-0148ac41181da0a0e (taller3-tec-subnet-public1-us-east-1a)	Nombre del grupo de Auto Scaling –
IMDSv2 Optional ⚠️ EC2 recommends setting IMDSv2 to required Más información		

Detalles **Seguridad** Redes Almacenamiento Comprobaciones de estado Monitoreo Etiquetas

▼ Detalles de seguridad

Rol de IAM EMR_EC2_DefaultRole	ID del propietario 114507996909	Hora de lanzamiento Tue Nov 21 2023 19:42:04 GMT-0500 (hora estándar de Colombia)
Grupos de seguridad sg-0b692d40dc27a080 (ElasticMapReduce-master)		

Selecciona el grupo de seguridad y dale clic, allí iremos al grupo de seguridad y debemos crear las siguientes reglas de entrada:

Regla	Protocolo	Porta	Origen	Destino	Acción
-	TCP personalizado	TCP	22	MI IP	Eliminar
-	TCP personalizado	TCP	8088	Any...	Eliminar
-	TCP personalizado	TCP	9443	Any...	Eliminar
-	TCP personalizado	TCP	8998	Any...	Eliminar
-	TCP personalizado	TCP	9870	Any...	Eliminar
-	TCP personalizado	TCP	18080	Any...	Eliminar
sgr-054b6cad5dba2e106	Todos los TCP	TCP	0 - 65535	Per...	Eliminar
-	TCP personalizado	TCP	14000	Any...	Eliminar
-	TCP personalizado	TCP	9878	Any...	Eliminar
-	TCP personalizado	TCP	18080	Any...	Eliminar
-	TCP personalizado	TCP	8888	Any...	Eliminar
-	TCP personalizado	TCP	8080	Any...	Eliminar
-	TCP personalizado	TCP	8890	Any...	Eliminar


Agregar regla

Rules with source of 0.0.0.0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

Cancelar Previsualizar los cambios Guardar reglas

Con esto las aplicaciones podrán comunicarse entre sí.

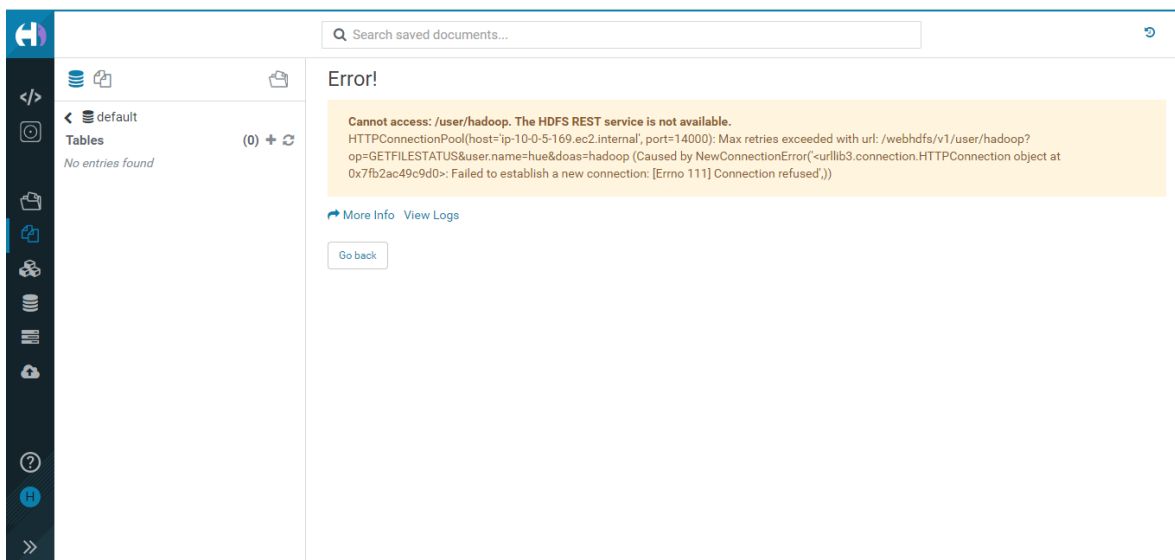
Ahora nos dirigimos al cluster nuevamente en en la parte de aplicaciones damos clic en Tonalidad o hue

IU de la aplicación en el nodo principal	
Estas requieren que el túnel de SSH esté habilitado.	
	Habilitar una conexión SSH
Aplicación	URL de la IU 
Administrador de recursos	http://ec2-54-81-140-223.compute-1.amazonaws.com:8088/
JupyterHub	https://ec2-54-81-140-223.compute-1.amazonaws.com:9443/
Livy	http://ec2-54-81-140-223.compute-1.amazonaws.com:8998/
Nodo del nombre de HDFS	http://ec2-54-81-140-223.compute-1.amazonaws.com:9870/
Servidor de historial de Spark	http://ec2-54-81-140-223.compute-1.amazonaws.com:18080/
Tonalidad	http://ec2-54-81-140-223.compute-1.amazonaws.com:8888/
UI de Tez	http://ec2-54-81-140-223.compute-1.amazonaws.com:8080/tez-ui
Zeppelin	http://ec2-54-81-140-223.compute-1.amazonaws.com:8890/

Asignamos el usuario hadoop y la contraseña de nuestra preferencia



Ahora vamos a ver si nos genera un error en los archivos, para ello en el menú damos clic en Files, si nos sale el error como viene:



Ahora como ultimo paso debemos configurar el nodo maestro para que se elimine el error que nos para ello accedemos con el par de claves

Información del clúster	Aplicaciones	Administración de clústeres	Estado y hora
<p>ID del clúster j-2XETAT8QJ7DXP</p> <p>Configuración del clúster Grupos de instancias</p> <p>Capacidad 1 Primary (Principal) 1 Principal 0 Tarea</p>	<p>Versión de Amazon EMR emr-6.15.0</p> <p>Aplicaciones instaladas Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Livy 0.7.1, Spark 3.4.1, Sqoop 1.4.7, Tez 0.10.2, Zeppelin 0.10.1, ZooKeeper 3.5.10</p>	<p>Destino del registro en Amazon S3 hfbaniats3</p> <p>IU de aplicación persistente Servidor de historial de Spark Servidor de línea de tiempo de YARN UI de Tez</p> <p>DNS público del nodo principal ec2-54-81-140-223.compute-1.amazonaws.com</p> <p>Conectarse al nodo principal mediante SSH</p> <p>Conectarse al nodo principal mediante SSM</p>	<p>Estado Esperando</p> <p>Hora de creación 22 de noviembre de 2023 19:23 (UTC-05:00)</p> <p>Tiempo transcurrido 32 minutos, 31 segundos</p>

```
hfbaniatq@DESKTOP-DITK689 MINGW64 ~/Documents/universidad/telematica
$ chmod 400 bigdata-key-pair.pem

hfbaniatq@DESKTOP-DITK689 MINGW64 ~/Documents/universidad/telematica
$ ssh -i bigdata-key-pair.pem hadoop@ec2-54-81-140-223.compute-1.amazonaws.com
```

editamos el archivo hue.ini “sudo nano /etc/hue/conf/hue.ini” buscar la línea que contenga: ‘webhdfs-url’ y cambiar el puerto de 14000 a 9870 (en nano puede utilizar control-w para buscar la palabra)

```
# HA support by using HttpFs

[[[default]]]
# Enter the filesystem uri
fs_defaultfs = hdfs://ip-10-0-5-169.ec2.internal:8020

# NameNode logical name.
## logical_name=

# Use WebHdfs/HttpFs as the communication mechanism.
# Domain should be the NameNode or HttpFs host.
# Default port is 14000 for HttpFs.
webhdfs_url = http://ip-10-0-5-169.ec2.internal:9870/webhdfs/v1

# Change this if your HDFS cluster is Kerberos-secured
security_enabled = false

# In secure mode (HTTPS), if SSL certificates from YARN Rest APIs
# have to be verified against certificate authority

^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text   ^J Justify    ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Uncut Text ^T To Spell   ^_ Go To Line
```

Por ultimo reiniciamos el servicio

“sudo systemctl restart hue.service”

Ahora vamos a ingresar a Jupyter-hub para ello volvemos a cluster en aplicaciones y damos clic en Jupyter-hub

Aplicacion	URL de la IU
Administrador de recursos	http://ec2-54-81-140-223.compute-1.amazonaws.com:8088/
JupyterHub	https://ec2-54-81-140-223.compute-1.amazonaws.com:9443/
Livy	http://ec2-54-81-140-223.compute-1.amazonaws.com:8998/
Nodo del nombre de HDFS	http://ec2-54-81-140-223.compute-1.amazonaws.com:9870/
Servidor de historial de Spark	http://ec2-54-81-140-223.compute-1.amazonaws.com:18080/
Tonalidad	http://ec2-54-81-140-223.compute-1.amazonaws.com:8888/
UI de Tez	http://ec2-54-81-140-223.compute-1.amazonaws.com:8080/tez-ui
Zeppelin	http://ec2-54-81-140-223.compute-1.amazonaws.com:8890/

UI de aplicaciones en las redes principales y de tareas

Y una vez allí usaremos las credenciales por defecto:

Username: jovyan

Password: jupyter



Sign in

Username:

Password:

Sign in

Your server is starting up.

You will be redirected automatically when it's ready for you.

Server ready at /user/jovyan/

Event log

Ahora vamos a crear un notebook, para ello damos clic en new, y seleccionamos PySpark

