

GUIA PARA LA CREACIÓN DE UN CLÚSTER DE AMAZON EMR

Prerrequisitos

Antes de comenzar a configurar el clúster de Amazon EMR, usted debe completar los requisitos que a continuación se indican:

1. Registrarse en AWS

Si no dispone de una cuenta de AWS, utilice el siguiente procedimiento para crearla.

Para inscribirse en AWS

- Abra <https://aws.amazon.com/> y elija **Create an AWS Account**.
- Siga las instrucciones en línea.

2. Crear un bucket de Amazon S3

Debe especificar un bucket y una carpeta de Amazon S3 para almacenar los datos entrada o de salida de los programa que ejecute en el cluster. En esta guía, se utiliza la ubicación predeterminada para los registros de logs, sin embargo también puede especificar una ubicación personalizada si así lo desea.

De acuerdo a los requisitos de Hadoop, los nombres del bucket y de las carpetas que utilice con Amazon EMR tienen las siguientes restricciones:

- Deben contener solo letras en minúsculas, números, puntos (.) y guiones (-).
- No pueden terminar en números.

Si ya tiene acceso a una carpeta que cumpla estos requisitos, puede utilizarla para este guía.

La carpeta de salida debería estar vacía.

Otra consideración importante 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, elígelo de la lista y a continuación, elija **Create folder (Crear carpeta)**, sustituya **New folder (Carpeta nueva)** por un nombre que cumpla los requisitos antes nombrados y luego elija **Save (Guardar)**.

El nombre del bucket y de la carpeta utilizado más adelante en esta guía es *s3://mybucket/MyEMRFoLder*.

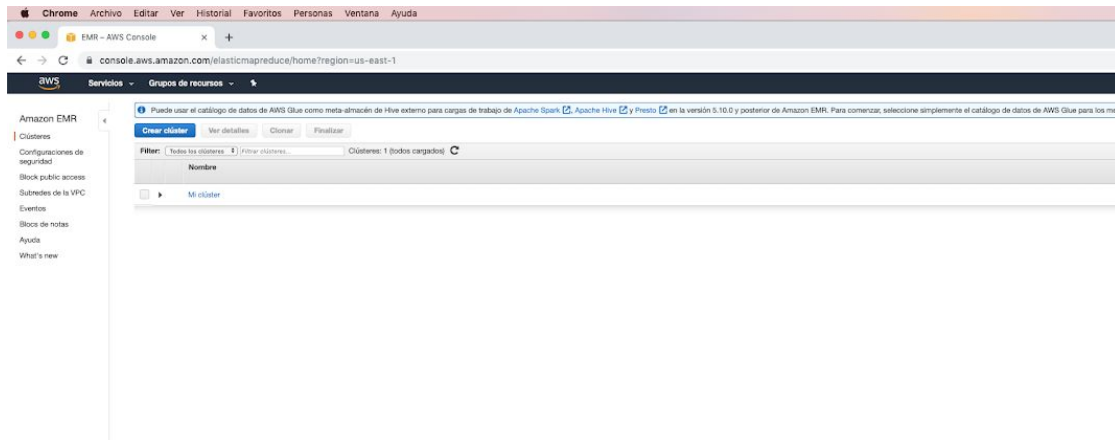
3. Crear un par de claves de Amazon EC2

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 la guía [GUÍA PARA LA CREACIÓN DE UN PAR DE CLAVES EN AWS EC2](#).

Procedimiento para crear el cluster EMR

En este procedimiento creará un clúster de EMR mediante las **Advanced Options (Opciones avanzadas)** de la consola de Amazon EMR cambiando algunas de las opciones para ajustarlas a nuestra necesidad. Para obtener más información sobre estas opciones, consulte [Resumen de las opciones rápidas](#) después del procedimiento. También puede seleccionar **Go to Quick options (Ir a las opciones rápidas)** para explorar las opciones de configuración rápidas de un clúster. Antes de crear el clúster, asegúrese de cumplir los prerequisites que se exponen al inicio de esta guía:

1. Inicie sesión en la Consola de administración de AWS y abra la consola de Amazon EMR en <https://console.aws.amazon.com/elasticmapreduce/>.



2. Elija **Create cluster (Crear clúster)**.
3. En la página **Create Cluster - Advanced Options (Crear clúster: opciones avanzadas)**:
 - En el paso 1. Software and steps
 - i. Elija la versión de emr-5.26.0 de EMR.
 - ii. Seleccione la siguiente lista de componentes:
 1. Hadoop 2.8.5
 - 2.
 3. Hive 2.3.5
 4. Hue 4.4.0
 5. Spark 2.4.3
 6. Sqoop 1.4.7
 7. Oozie 5.1.0

Chrome Archivo Editar Ver Historial Favoritos Personas Ventana Ayuda

EMR - AWS Console

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster:

aws Servicios Grupos de recursos

Crear clúster: Opciones avanzadas [Ir a las opciones rápidas](#)

Step 1: Software y pasos

Step 2: Hardware

Step 3: Configuración general del clúster

Step 4: Seguridad

Configuración de software

Versión:

| | | |
|--|---|---|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 0.9.6 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.0 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input type="checkbox"/> Presto 0.220 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input checked="" type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input checked="" type="checkbox"/> Oozie 5.1.0 |
| <input checked="" type="checkbox"/> Spark 2.4.3 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.13.1 |

Compatibilidad con instancias maestras múltiples

☐ Habilitar la compatibilidad con instancias maestras múltiples

Configuración del catálogo de datos de AWS Glue

☐ Usar para metadatos de la tabla de Hive

☐ Usar para metadatos de la tabla de Spark

Editar configuración de software

☒ Escribir la configuración ☐ Cargar JSON desde S3

`classification { config-file-name, properties { myKey1:myValue1, myKey2:myValue2 }`

Añadir pasos (opcional)

Tipo de paso:

☐ Terminar automáticamente el clúster después de que se complete el último paso

- En el paso 2, se le pide elegir las configuraciones de red, así como el tipo de instancias, la cantidad, el método de escalamiento y la forma de cobro. En este caso acepte los valores predeterminados.

Chrome Archivo Editar Ver Historial Favoritos Personas Ventana Ayuda

EMR - AWS Console

console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#create-cluster:

Crear clúster: Opciones avanzadas [Ir a las opciones rápidas](#)

Step 1: Software y pasos
Step 2: Hardware
 Step 3: Configuración general del clúster
 Step 4: Seguridad

Configuración de hardware

Si necesita más de 20 instancias EC2, consulte este tema.

Configuración del grupo de instancias

☒ **Grupos de instancias uniformes**
 Especifique un tipo de instancia única y la opción de compra para cada tipo de nodo.

☐ **Flotas de instancias**
 Especifique la capacidad de destino y cómo Amazon EMR la cumple para cada tipo de nodo. Combine los tipos de instancias y las opciones de compra. [Más información](#)

Red: [Creación de una VPC](#)

Subred EC2:

Tamaño del volumen de EBS del dispositivo raíz: GiB

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#)

| Tipo de nodo | Tipo de instancia | Recuento de instancias | Opción de compra | Auto Scaling |
|----------------------------|--|------------------------|---|----------------------------|
| Maestro Maestro - 1 | m5.xlarge 4 vCore, memoria de 16 GiB, almacenamiento solo EBS Almacenamiento de EBS: 64 GiB Agregar opciones de configuración | 1 instancias | <input checked="" type="radio"/> Bajo demanda <input type="radio"/> Spot <input type="text" value="Use on-demand as max price"/> | No disponible para maestro |
| Principal Principal - 2 | m5.xlarge 4 vCore, memoria de 16 GiB, almacenamiento solo EBS Almacenamiento de EBS: 64 GiB Agregar opciones de configuración | 2 instancias | <input checked="" type="radio"/> Bajo demanda <input type="radio"/> Spot <input type="text" value="Use on-demand as max price"/> | No habilitado |
| Tarea Tarea - 3 | m5.xlarge 4 vCore, memoria de 16 GiB, almacenamiento solo EBS Almacenamiento de EBS: 64 GiB Agregar opciones de configuración | 3 instancias | <input checked="" type="radio"/> Bajo demanda <input type="radio"/> Spot <input type="text" value="Use on-demand as max price"/> | No habilitado |

[+ Añadir grupo de instancias de tareas](#)

Cancel Previous **Next**

- En el paso 3, se le pide que introduzca un **Cluster name (Nombre del clúster)** que le ayude a identificar el clúster, por ejemplo, *Mi primer clúster de EMR*, así como elegir la carpeta destino de los registros de log del cluster y otras configuraciones generales, acepte los valores predeterminados.

- En el paso 4, **Security and access (Seguridad y acceso)**, elija el **EC2 key pair (Par de claves EC2)** que ha creado en el paso 3 de Prerrequisitos, al inicio de este documento.

4. Elija **Create cluster**.

Aparecerá la página de estado del clúster con el **Summary (Resumen)** del clúster. Puede utilizar esta página para monitorizar el progreso de creación del clúster y ver información detallada sobre el estado del clúster. A medida que finalizan las tareas de creación del clúster, los elementos de la página de estado se actualizan.