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 a 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 guia.

La carpeta de salida debería estar vacía.

Otra consideración importe 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 guia es `s3://mybucket/MyEMRFolder`.

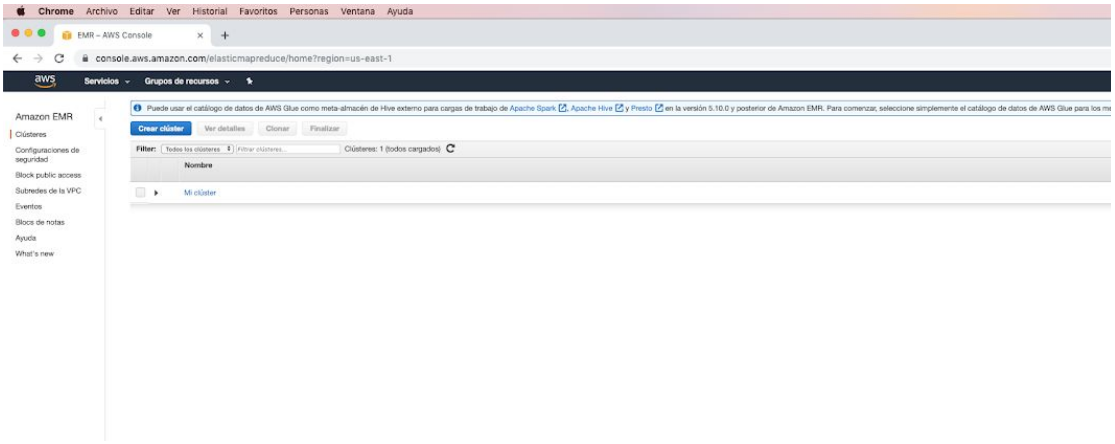
3. Crear un par de claves de Amazon EC2

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 guia GUIA PARA LA CREACIÓN DE UN PAR DE CLAVES EN AWS EC2.

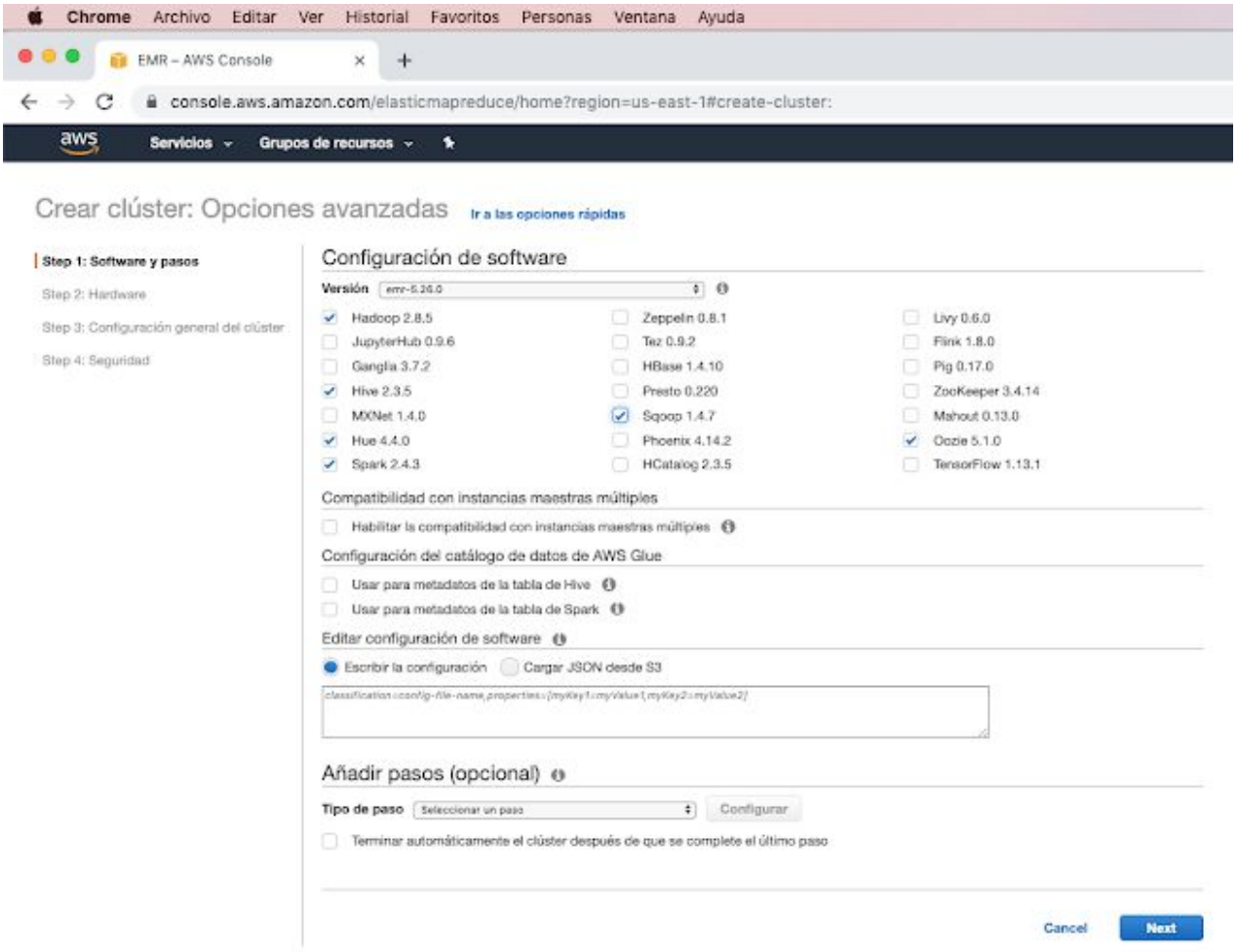
Procedimiento para crear el cluster EMR

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 prerrequisitos que se exponen al inicio de esta guía:

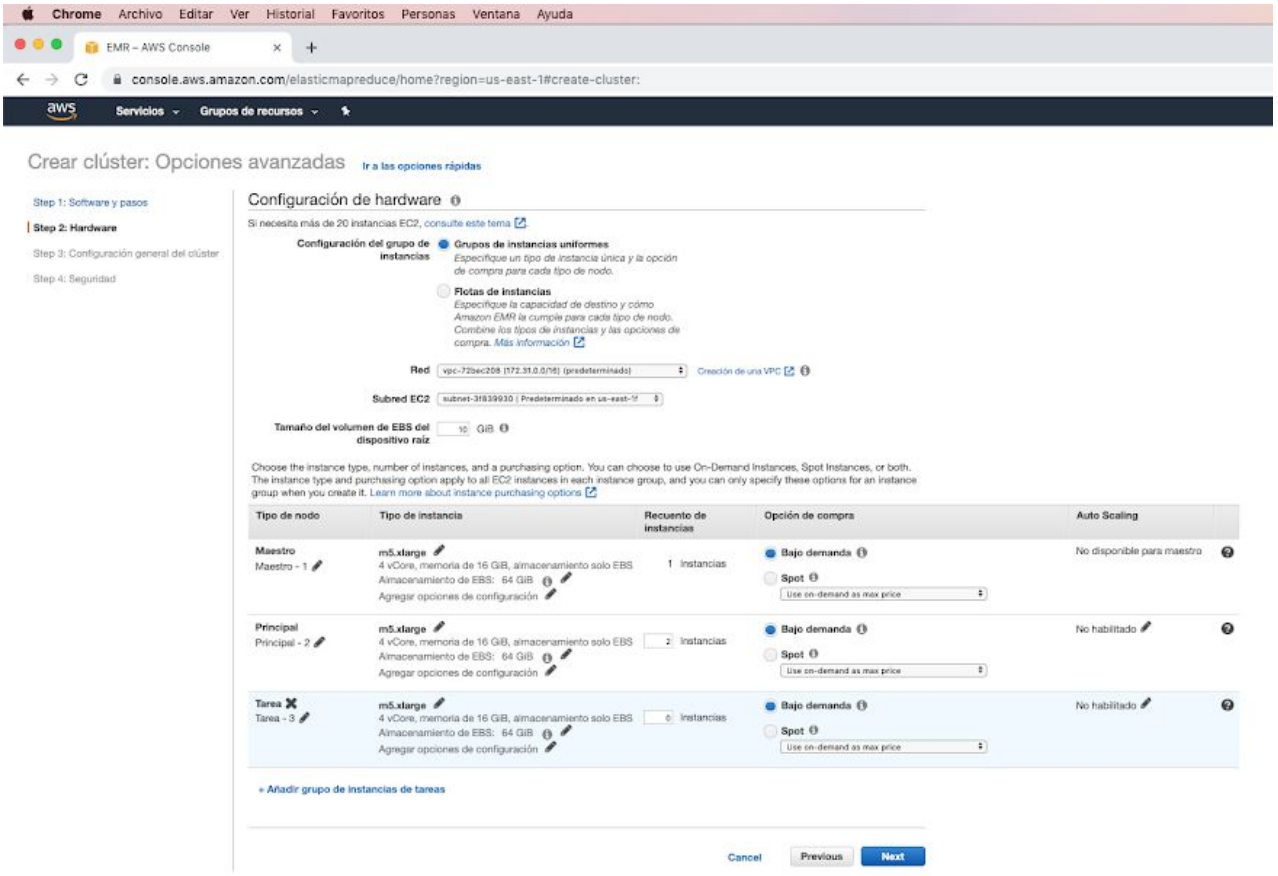
1. Inicie sesión en la Consola de administración de AWS y abra la consola de Amazon EMR en https://console.aws.amazon.com/elasticmapreduce/.



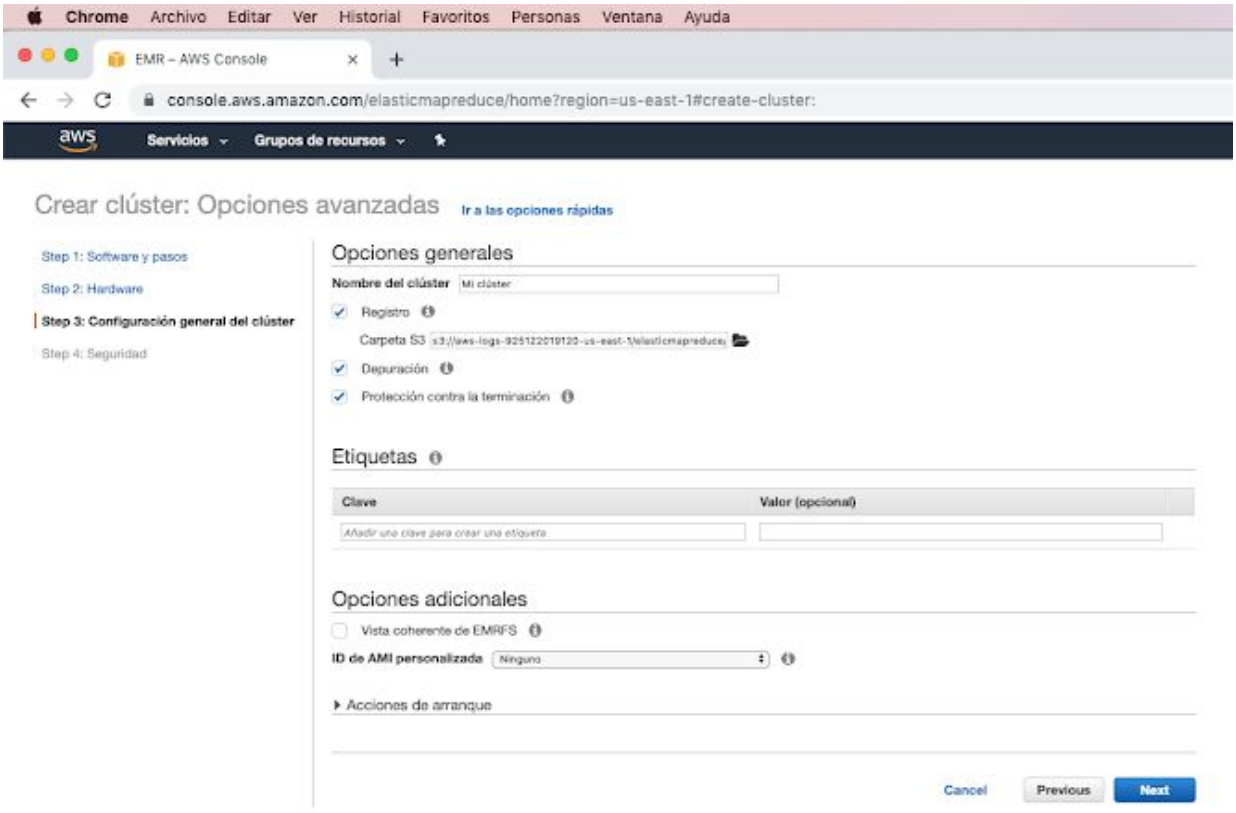
2. Elija **Create cluster (Crear clúster)**.
3. En la página **Create Cluster - Advanced Options (Crear clúster: opciones avanzadas)**:
    - En el paso 1. Software and steps
        i. Elija la versión de emr-5.26.0 de EMR.
        ii. Selecciones la siguiente lista de componentes:
            1. Hadoop 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



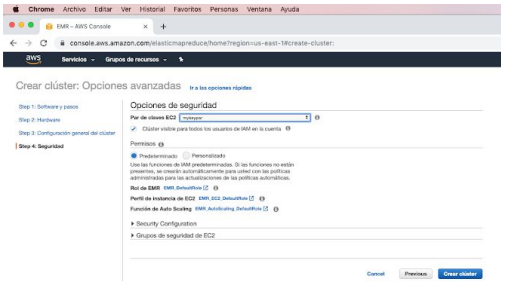
- En el paso 2, se le pide elegir las configuraciones de red, así como el tipo de instancias, la cantidad, el método de escalamiento y la forma de cobro. En este caso acepte los valores predeterminados.



- En el paso 3, se el 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.