Creación de un Cluster EMR en AWS

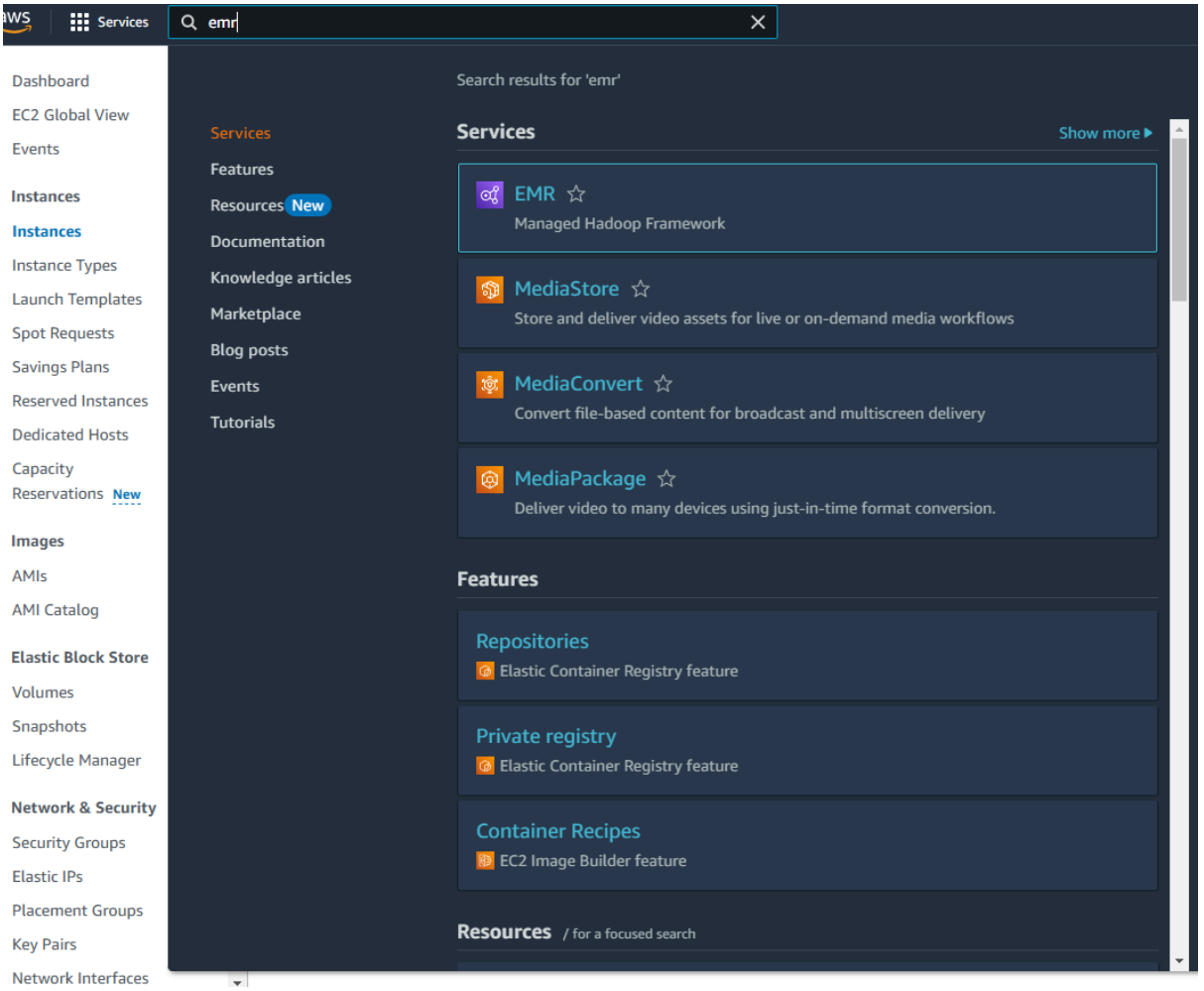
Introducción

En este documento, se detalla el proceso paso a paso para la creación de un cluster EMR en AWS, utilizando las capturas de pantalla del proceso como referencia visual.

Pasos para la Creación del Cluster EMR

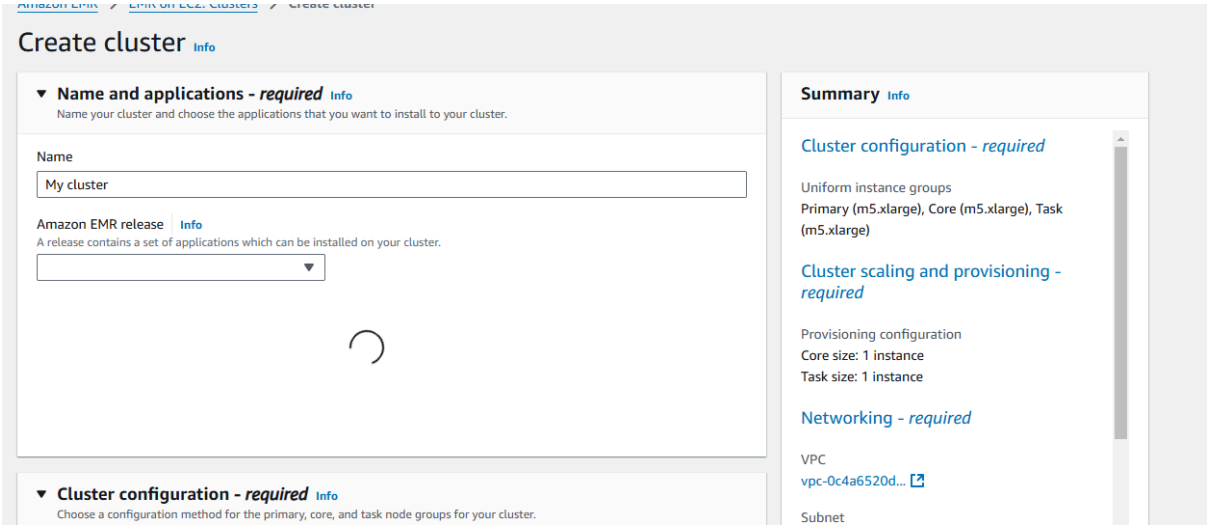
Paso 1: Acceder a AWS Management Console

1. **Descripción**: Accedemos a la consola de AWS y navegamos al servicio EMR (Elastic MapReduce).
2. **Imagen de referencia**:



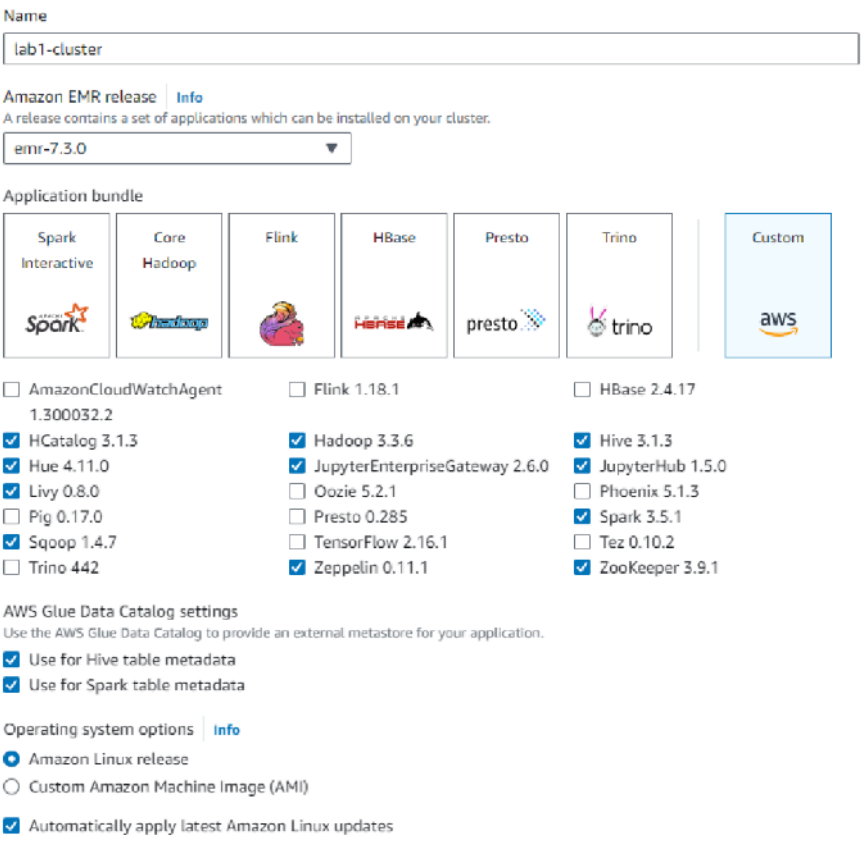
Paso 2: Configuración Inicial del Cluster EMR

1. **Descripción**: En el panel de EMR, seleccionamos la opción para crear un nuevo cluster.
2. **Configuraciones principales**:
    - **Nombre del Cluster**: Especificamos un nombre que identifique nuestro cluster.
    - **Región**: Seleccionamos la región en la que queremos desplegar el cluster.
3. **Imagen de referencia**:



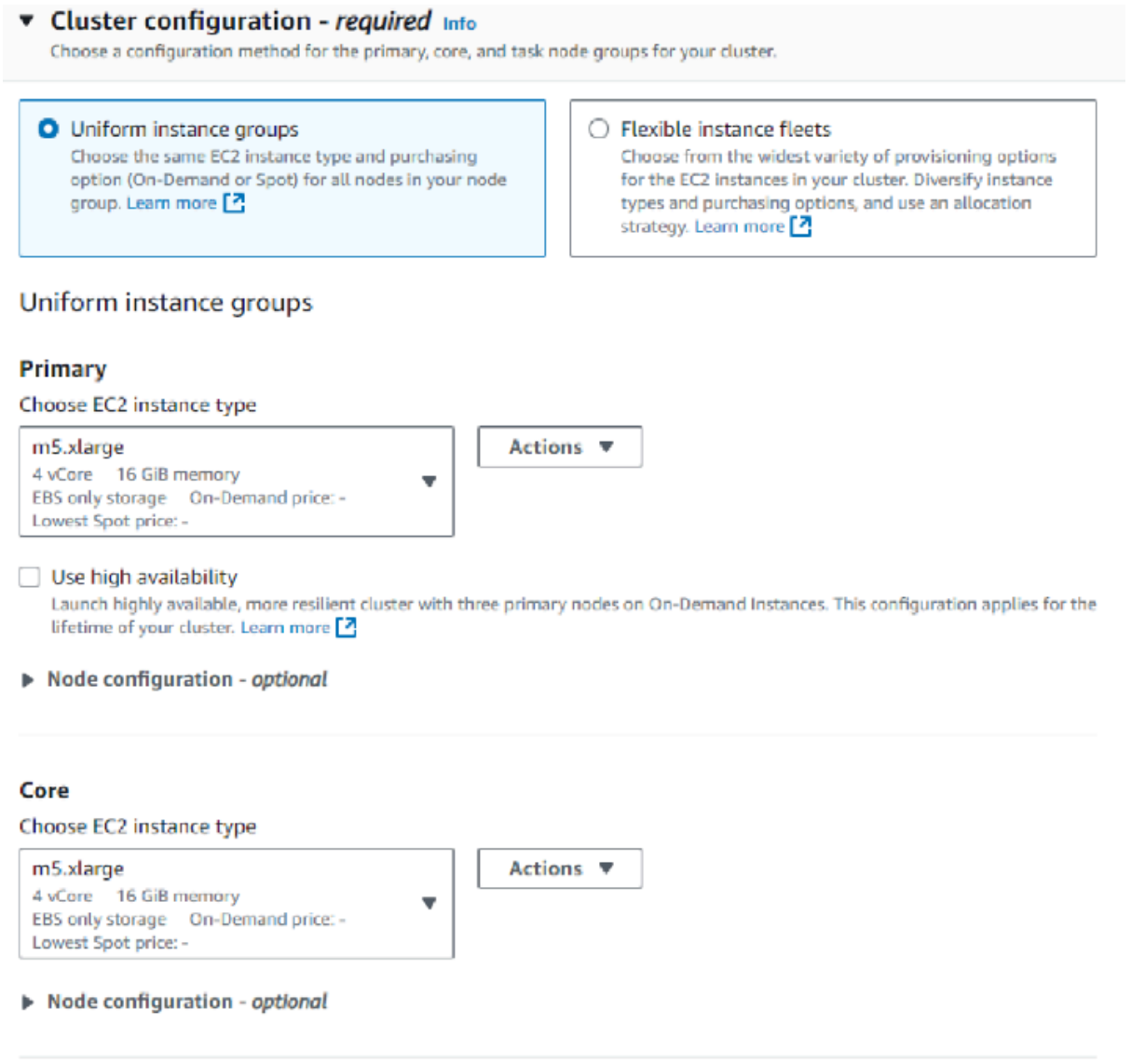
Paso 3: Selección del Software

1. **Descripción**: Elegimos el software que se instalará en el cluster, como Hadoop, Spark, o Hive.
2. **Versión**: Seleccionamos la versión adecuada de EMR y de las herramientas específicas necesarias.
3. **Imagen de referencia**:



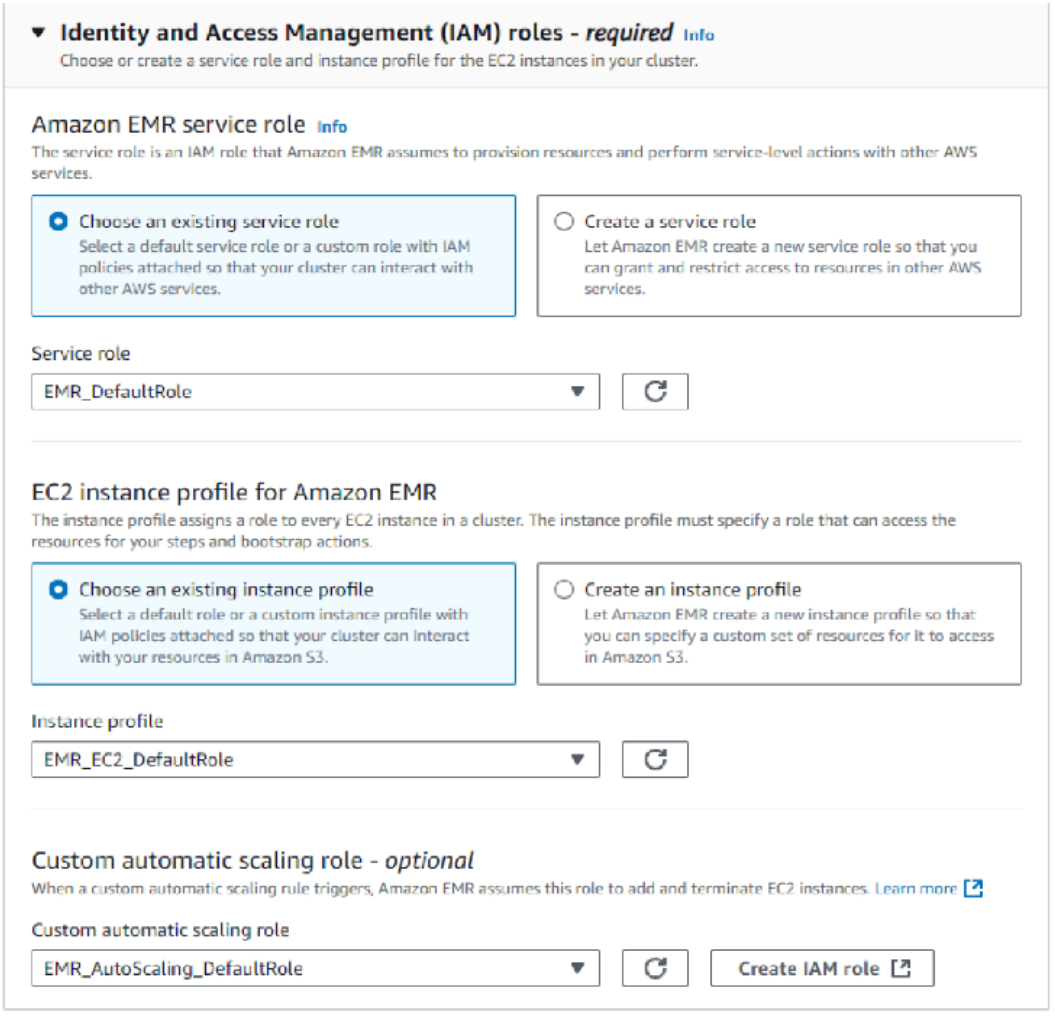
Paso 4: Configuración de Instancias

1. **Descripción**: En esta sección, configuramos las instancias para el master y los nodos de trabajo.
    - **Tipo de Instancia**: Seleccionamos el tipo de instancia (por ejemplo, `m5.xlarge`).
    - **Número de Instancias**: Especificamos cuántas instancias queremos para los nodos de trabajo.
2. **Imagen de referencia**:



Paso 5: Configuración de Permisos (IAM Roles)

1. **Descripción**: Asignamos los roles de IAM necesarios para que el cluster tenga los permisos de acceso adecuados.
    - **Rol de Servicio**: Seleccionamos un rol de servicio EMR predeterminado o configuramos uno personalizado.
2. **Imagen de referencia**:



Paso 6: Revisión y Lanzamiento del Cluster

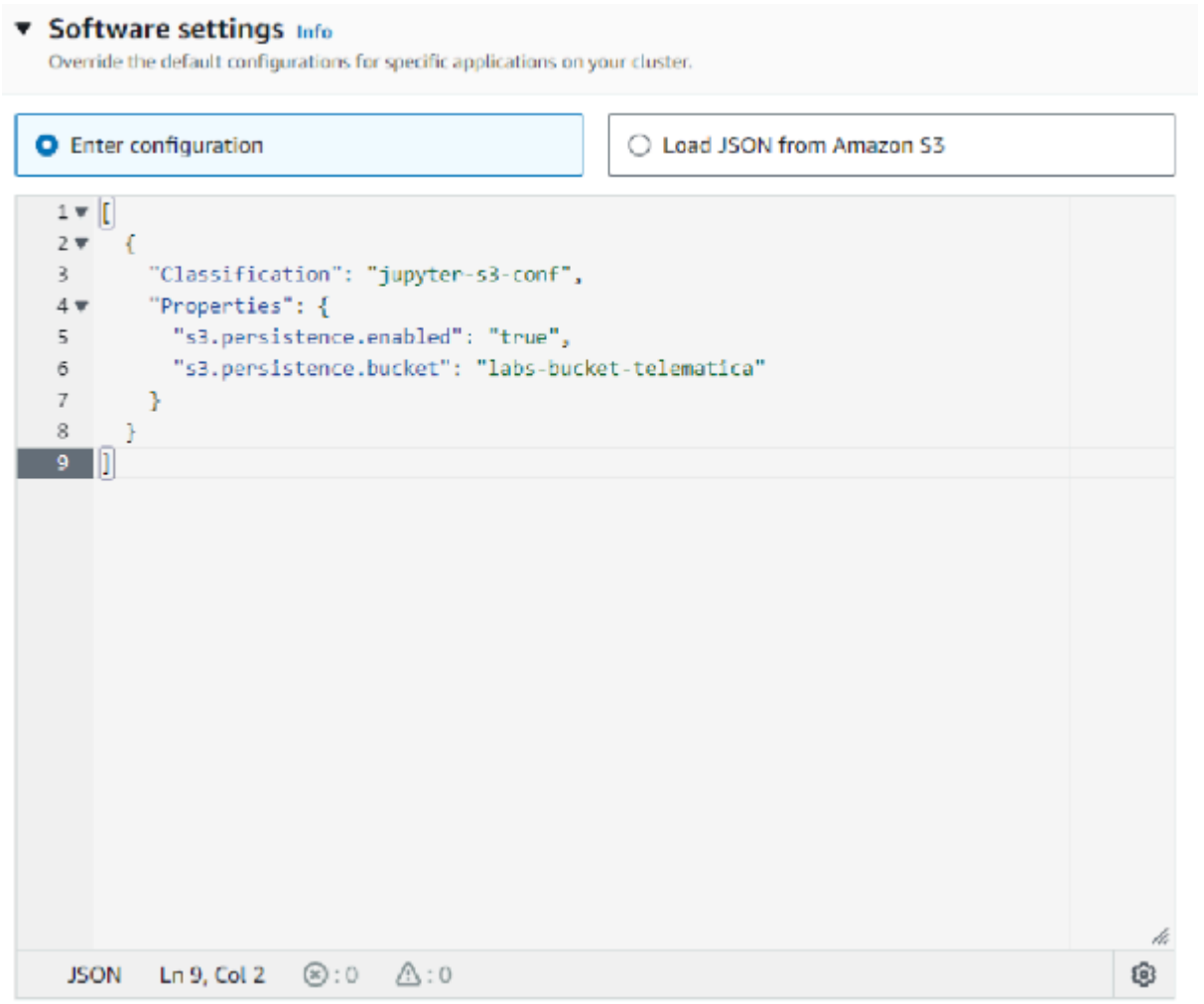
1. **Descripción**: Revisamos todas las configuraciones y, si todo está correcto, iniciamos la creación del cluster.
2. **Estado de Creación**: Una vez lanzado, la consola de EMR muestra el progreso de la creación del cluster.
3. **Imagen de referencia**: (agrega la imagen que muestra el resumen final antes de lanzar el cluster y/o el estado de creación).

Paso 7: Configuración del Software (Jupyter-S3-Conf)

1. **Descripción**: En esta sección, configuramos las opciones específicas del software para nuestro cluster EMR. En particular, estamos configurando la integración de Jupyter con Amazon S3 para almacenar de forma persistente los datos generados en los notebooks de Jupyter.
2. **Configuración JSON**: La configuración JSON utilizada especifica que los datos generados en Jupyter se guardarán en un bucket de S3 llamado `"labs-bucket-telematica"`.

- `"Classification"`: Este campo indica que estamos configurando `jupyter-s3-conf`, que permite definir opciones de almacenamiento en Amazon S3 para Jupyter.
- `"Properties"`: Dentro de este objeto, configuramos dos propiedades:
  - `"s3.persistence.enabled"`: Activamos la persistencia en S3 estableciendo este valor en `"true"`.
  - `"s3.persistence.bucket"`: Especificamos el nombre del bucket en S3 (`"labs-bucket-telematica"`) donde se guardarán los datos.

3. **Imagen de referencia**:



Esta configuración garantiza que cualquier dato o notebook generado en Jupyter dentro del cluster EMR se almacene automáticamente en el bucket S3 especificado, facilitando la recuperación y administración de los datos.

Paso 8: Verificación del Cluster Activo

1. **Descripción**: Una vez completada la creación, verificamos el estado del cluster en la consola de EMR.

Conclusión

Hemos creado exitosamente un cluster EMR en AWS siguiendo estos pasos. Esta configuración permite realizar análisis y procesamiento de grandes volúmenes de datos con las herramientas seleccionadas.