

Creación de un Cluster EMR en AWS

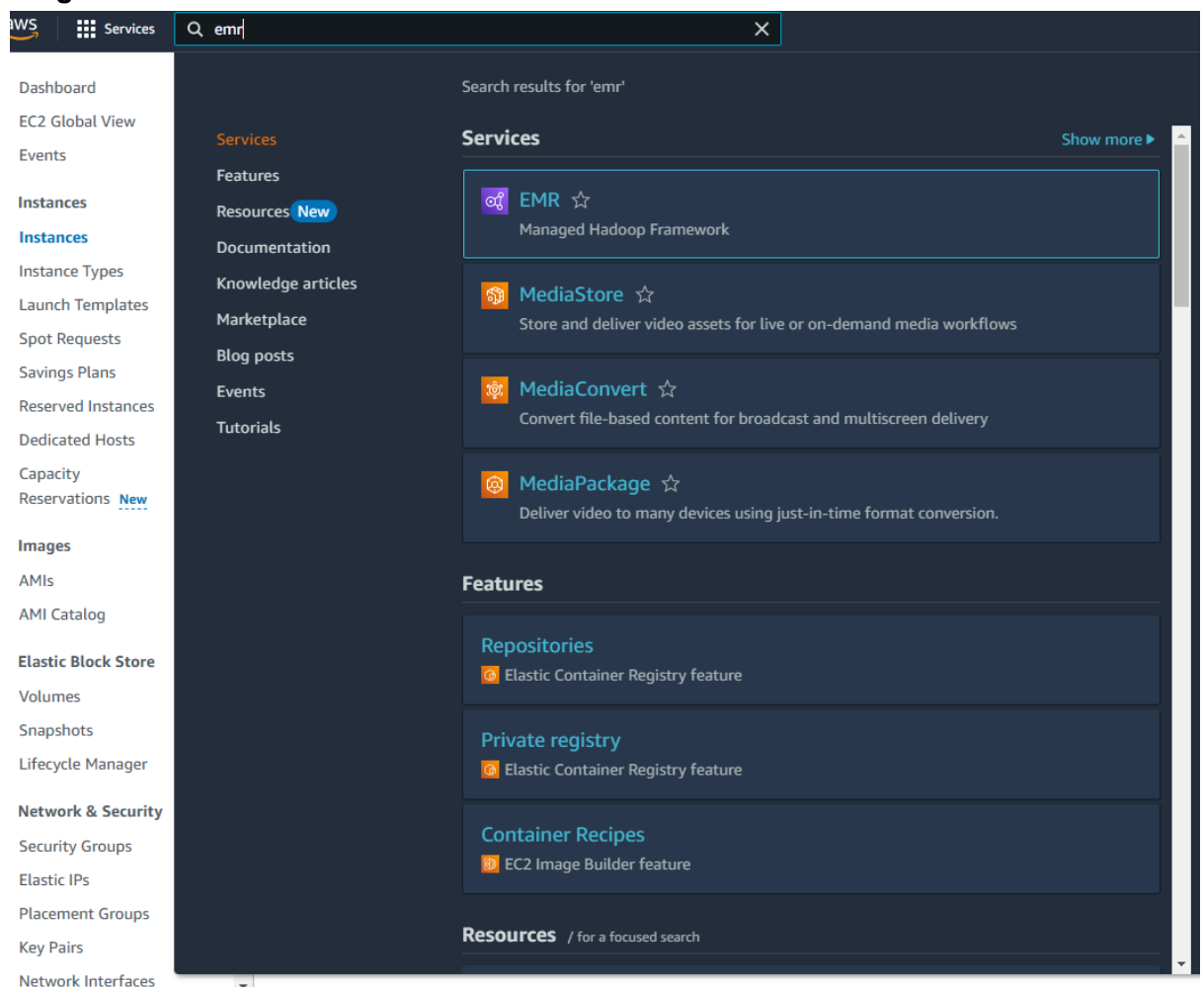
Introducción

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:**

Create cluster [Info](#)

▼ **Name and applications - required** [Info](#)
Name your cluster and choose the applications that you want to install to your cluster.

Name

Amazon EMR release [Info](#)
A release contains a set of applications which can be installed on your cluster.

▼ **Cluster configuration - required** [Info](#)
Choose a configuration method for the primary, core, and task node groups for your cluster.

Summary [Info](#)

Cluster configuration - required

Uniform instance groups
Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

Cluster scaling and provisioning - required

Provisioning configuration
Core size: 1 instance
Task size: 1 instance

Networking - required

VPC
[vpc-0c4a6520d...](#)

Subnet

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:**

Name

Amazon EMR release [Info](#)
A release contains a set of applications which can be installed on your cluster.

Application bundle

Spark Interactive
Core Hadoop
Flink
HBase
Presto
Trino
Custom

☐ AmazonCloudWatchAgent 1.300032.2
☒ HCatalog 3.1.3
☒ Hue 4.11.0
☒ Livy 0.8.0
☐ Pig 0.17.0
☒ Sqoop 1.4.7
☐ Trino 442

☐ Flink 1.18.1
☒ Hadoop 3.3.6
☒ JupyterEnterpriseGateway 2.6.0
☐ Oozie 5.2.1
☐ Presto 0.285
☐ TensorFlow 2.16.1
☒ Zeppelin 0.11.1

☐ HBase 2.4.17
☒ Hive 3.1.3
☒ JupyterHub 1.5.0
☐ Phoenix 5.1.3
☒ Spark 3.5.1
☐ Tez 0.10.2
☒ ZooKeeper 3.9.1

AWS Glue Data Catalog settings
Use the AWS Glue Data Catalog to provide an external metastore for your application.
☒ Use for Hive table metadata
☒ Use for Spark table metadata

Operating system options [Info](#)
☒ Amazon Linux release
☐ Custom Amazon Machine Image (AMI)
☒ Automatically apply latest Amazon Linux updates

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:**

▼ Cluster configuration - required Info
Choose a configuration method for the primary, core, and task node groups for your cluster.

☒ **Uniform instance groups**
Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. [Learn more](#)

☐ **Flexible instance fleets**
Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. [Learn more](#)

Uniform instance groups
Primary
Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

Actions ▼

☐ **Use high availability**
Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. [Learn more](#)

► **Node configuration - optional**

Core
Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

Actions ▼

► **Node configuration - optional**

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:**

▼
Identity and Access Management (IAM) roles - required
[Info](#)

Choose or create a service role and instance profile for the EC2 instances in your cluster.

Amazon EMR service role
[Info](#)

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

☒
Choose an existing service role

Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

☐
Create a service role

Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role

EMR_DefaultRole

▼

↺

EC2 instance profile for Amazon EMR

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

☒
Choose an existing instance profile

Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

☐
Create an instance profile

Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

EMR_EC2_DefaultRole

▼

↺

Custom automatic scaling role - optional

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. [Learn more](#)

Custom automatic scaling role

EMR_AutoScaling_DefaultRole

▼

↺

Create IAM role

Paso 6: Revisión y Lanzamiento del Cluster

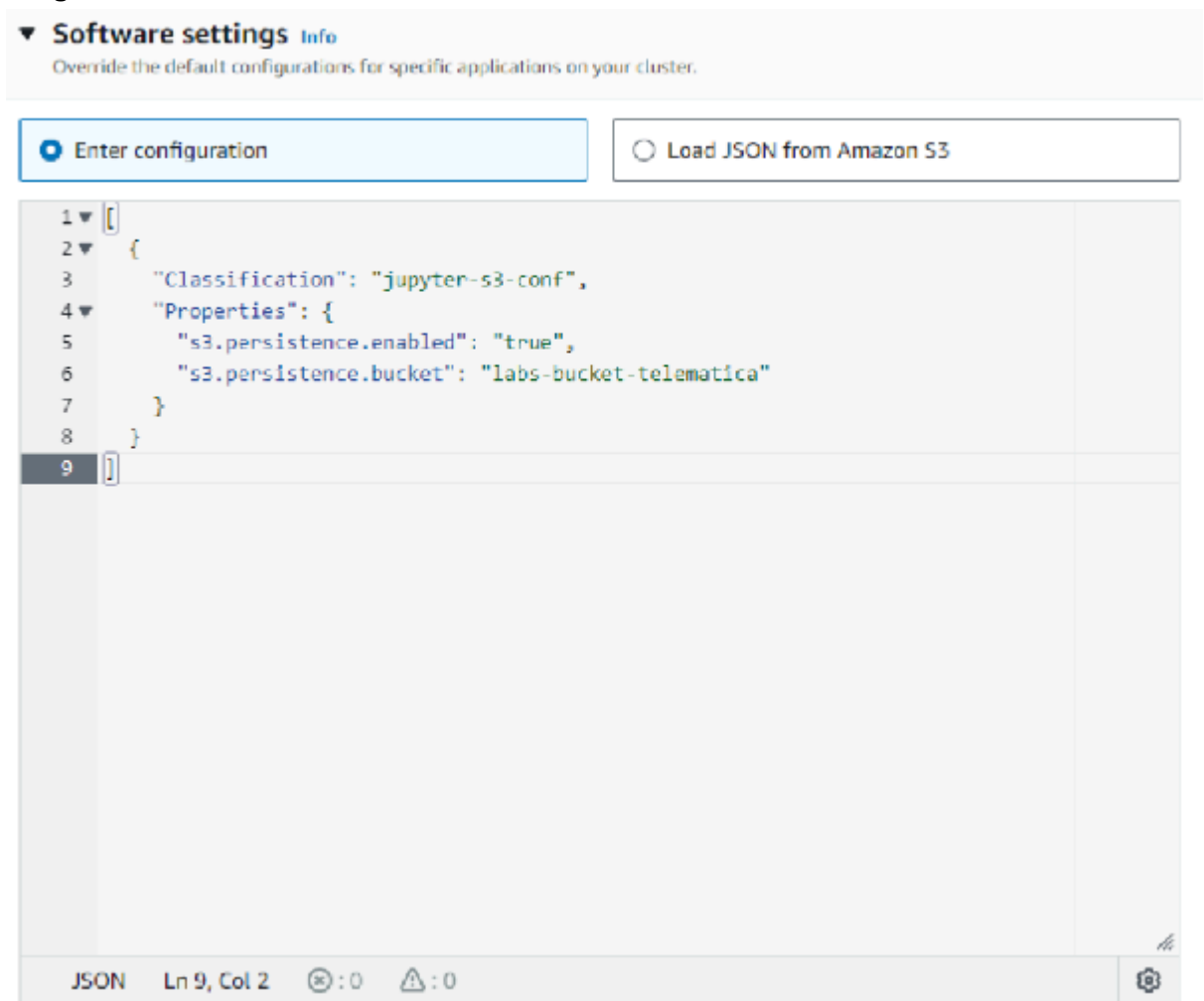
1. **Descripción:** Revisamos todas las configuraciones y, si todo está correcto, iniciamos la creación del cluster.
2. **Estado de Creación:** Una vez lanzado, la consola de EMR muestra el progreso de la creación del cluster.
3. **Imagen de referencia:** (agrega la imagen que muestra el resumen final 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.persistance.enabled": Activamos la persistencia en S3 estableciendo este valor en "true".
 - "s3.persistance.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.