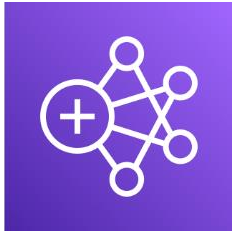


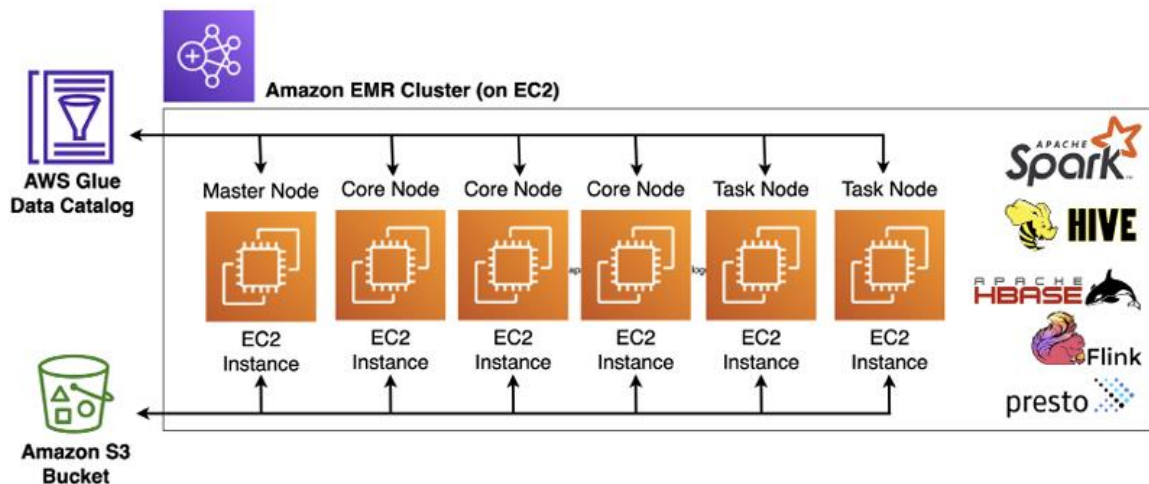
Elastic Map Reduce



[Amazon EMR](#) es una plataforma administrada para ejecutar los servicios del ecosistema Hadoop como *Apache Spark*, *Apache Hive*, *Apache Hudi*, *Apache HBase*, *Presto*, *Pig* y otras. *Amazon EMR* se encarga de las complejidades de implementar estas herramientas y administrar los recursos informáticos en clúster subyacentes.

Así pues, EMR permite crear clústeres para realizar analíticas sobre datos y cargas de BI, así como transformar y mover grandes volúmenes de datos, tanto cargando como almacenando datos en servicios de AWS como *S3* y *DynamoDB*.

Al igual que muchos otros servicios de AWS, *Amazon EMR* se puede ejecutar en modo aprovisionado (indicando los tipos de instancias a emplear) o en modo *serverless*.



Clúster EMR

Para ello, utiliza una distribución propia de AWS que permite seleccionar los componentes que van a lanzarse en el clúster (Hive, Spark, Presto, etc...)

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 3.3.3 | <input type="checkbox"/> Zeppelin 0.10.1 | <input type="checkbox"/> Livy 0.7.1 |
| <input type="checkbox"/> JupyterHub 1.4.1 | <input type="checkbox"/> Tez 0.10.2 | <input type="checkbox"/> Flink 1.15.2 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 2.4.13 | <input type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 3.1.3 | <input type="checkbox"/> Presto 0.276 | <input type="checkbox"/> ZooKeeper 3.5.10 |
| <input type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input type="checkbox"/> MXNet 1.9.1 | <input checked="" type="checkbox"/> Sqoop 1.4.7 |
| <input checked="" type="checkbox"/> Hue 4.10.0 | <input type="checkbox"/> Phoenix 5.1.2 | <input type="checkbox"/> Trino 398 |
| <input type="checkbox"/> Oozie 5.2.1 | <input type="checkbox"/> Spark 3.3.0 | <input type="checkbox"/> HCatalog 3.1.3 |
| <input type="checkbox"/> TensorFlow 2.10.0 | | |

Servicios Hadoop en EMR

Ofrece elasticidad sobre el clúster, pudiendo modificar dinámicamente el dimensionamiento del mismo según necesidades, tanto escalándolo hacia arriba como hacia abajo.

Respecto al hardware, se ejecuta sobre máquinas EC2 (IaaS), las cuales configuraremos según necesidades. **Utiliza HDFS y S3 para el almacenamiento, de manera que podemos guardar los datos de entrada y los de salida en S3**, mientras que los resultados intermedios los almacenamos en HDFS.

Cada clúster de EMR requiere de un nodo principal (*master*) y al menos otro nodo *core* (un nodo *worker* que incluye almacenamiento local), y opcionalmente un número de nodos tarea (nodos *worker* que no incluyen almacenamiento local).

Los clústeres de EMR se componen de:

- un nodo **principal central**, encargado de gestionar el clúster y ejecutar los servicios de coordinación de datos.
- varios nodos **principales**, los cuales ejecutan las tareas y almacenan los datos en el clúster HDFS.
- nodos **tareas**, los cuales son opcionales, y no almacenan datos, y podemos añadir a un clúster para incrementar la capacidad de procesamiento (y eliminarlos una vez no los necesitemos para reducir costes).

A nivel de servicios, podemos definir su arquitectura en cuatro capas:

- Almacenamiento: mediante *HDFS*, *EMR FS* o el sistema de archivos local (almacenamiento de las instancias EC2).
- Gestor de recursos del clúster: *YARN*, *Tez*.
- *Frameworks* de procesamiento de datos: *Hadoop MapReduce* y *Apache Spark*
- Aplicaciones: *Apache Spark*, *Apache Hive*, etc...

Lanzando EMR



EMR

Antes de lanzar EMR, necesitamos crear un repositorio de información donde guardar los datos de entrada y salida, así como los logs que EMR genere. Por tanto, previamente hemos de crear un *bucket* S3 para ello.

Podemos lanzar un clúster de EMR de tres formas, mediante la consola, el CLI o con un API. Vamos a centrar en el uso de la consola.

Accederemos al servicio de EMR, y creamos un clúster al cual hemos de ponerle un nombre, seleccionamos la última versión de EMR (a día de hoy es la 7.8) y seleccionamos algunos de los paquetes ya definidos, o seleccionamos las aplicaciones que deseemos (en este ejemplo seleccionaremos el conjunto predefinido etiquetado como *Core Hadoop*):

▼ Nombre y aplicaciones - **required** Información

Name your cluster and choose the applications that you want to install to your cluster.

Nombre

EMR_1

Versión de Amazon EMR | Información

Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.

emr-7.0.0

Paquete de aplicaciones

Spark Interactive Core Hadoop Flink HBase Presto Trino Custom

AmazonCloudWatchAgent 1.300031.1

HCatalog 3.1.3

Hue 4.11.0

Livy 0.7.1

Phoenix 5.1.3

Spark 3.5.0

Tez 0.10.2

ZooKeeper 3.5.10

Flink 1.18.0

Hadoop 3.3.6

JupyterEnterpriseGateway 2.6.0

MXNet 1.9.1

Pig 0.17.0

Sqoop 1.4.7

Trino 426

HBase 2.4.17

Hive 3.1.3

JupyterHub 1.5.0

Oozie 5.2.1

Presto 0.283

TensorFlow 2.11.0

Zeppelin 0.10.1

Configuración del Catálogo de datos de AWS Glue

Utilice el Catálogo de datos de AWS Glue para proporcionar un meta-almacén externo a la aplicación.

Usar para metadatos de la tabla de Hive

Opciones del sistema operativo | Información

Versión de Amazon Linux

Imagen de máquina de Amazon (AMI) personalizada

Aplicar automáticamente las actualizaciones más recientes de Amazon Linux

Creando un clúster en EMR

A continuación, vamos a seleccionar los tipos de instancias de nuestro clúster. Para ahorrar costes, únicamente elegiremos un nodo principal y otro central (los nodos tareas como son opcionales, para este ejemplo, vamos a descartarlos). En cuanto al tipo de instancia, escogeremos las más económicas que nos permite *AWS Academy*, actualmente las instancias **m4.large (2 CPU y 8 GiB RAM)**. Finalmente, escogemos el tamaño del almacenamiento del volumen EBS para las instancias EC2.

Configuración del clúster Información

Elija un método de configuración para los grupos principales, centrales y de nodos tareas para su clúster.

☒ **Grupos de instancias uniformes**
Elija el mismo tipo de instancia de EC2 y la misma opción de compra (bajo demanda o de spot) para todos los nodos de su grupo de nodos. [Más información](#)

☐ **Flotas de instancias flexibles**
Elija entre la más amplia variedad de opciones de aprovisionamiento para las instancias de EC2 de su clúster. Diversifique los tipos de instancias y las opciones de compra, y utilice una estrategia de asignación. [Más información](#)

Grupos de instancias uniformes

Principal

Elegir tipo de instancia de EC2

m4.large
2 vCore 8 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: - Precio de spot más bajo: -

Acciones ▼

☐ **Utilice la alta disponibilidad**
Lance un clúster más resiliente y de alta disponibilidad con tres nodos principales en instancias bajo demanda. Esta configuración se aplica durante toda la vida útil del clúster. [Más información](#)

► Configuración de nodo - *opcional*

Central

Elegir tipo de instancia de EC2

m4.large
2 vCore 8 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: - Precio de spot más bajo: -

Acciones ▼

Eliminar grupo de instancias

► Configuración de nodo - *opcional*

Agregar grupo de instancias de tareas

Puede agregar hasta 48 grupos más de instancias de tareas.

Volumen raíz de EBS

El volumen raíz de EBS se aplica a los sistemas operativos y las aplicaciones que instale en el clúster. [Restricciones de relación de volumen raíz de EBS](#)

Tamaño (GiB)	IOPS	Rendimiento (MiB/s)
<div>15</div> <div>15- 100 GiB por volumen SSD de uso general (gp3)</div>	<div>3000</div> <div>3000-16000 IOPS por volumen. Elija una relación máxima de 500:1 entre IOPS y el tamaño del volumen.</div>	<div>125</div> <div>125-1000 MiB/s por volumen. Elija una relación máxima de 0.25:1 entre el rendimiento y las IOPS.</div>

Elegiendo instancias y almacenamiento

Tras la configuración de las instancias, podemos definir el tamaño del clúster. Lo normal es que los clúster de EMR se compongan de un gran número de instancias (si no, más que Big Data, sería *Small Data* y probablemente habrá mejores soluciones que EMR). En nuestro caso, para probar el servicio, con

una única instancia nos es suficiente. Otras opciones es que el redimensionado del clúster los gestione AWS mediante métricas o programar nosotros el escalado mediante políticas personalizadas:

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 aprovisionamiento

Establezca el tamaño del principal grupo 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	m4.large	<input type="text" value="1"/>	<input type="checkbox"/>

Eligiendo el tamaño del clúster

De los pasos siguientes, podemos dejar los valores por defecto y bajamos hasta los ajustes de seguridad, donde elegimos el par de claves (si ya tuviéramos unas creadas, podrían ser esas) y luego seleccionamos los roles IAM **EMR_DefaultRole** en el apartado de *Rol de servicio* y **EMR_EC2_DefaultRole** en el de *Perfil de Instancia* respectivamente:

Configuración de seguridad y par de claves de EC2: opcional [Información](#)

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](#)

Roles de Identity and Access Management (IAM) [Información](#)

Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

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

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

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

Configurando la seguridad del clúster

En el lado derecho podremos ver un resumen de las opciones seleccionadas, y tras darle a crear, a los 10 minutos aproximadamente, nuestro clúster estará listo para trabajar con él.

No se olvide de asignar al clúster las claves de acceso, ya que sino no nos podríamos conectar a él.

▼ **Configuración de seguridad y par de claves de EC2** [Información](#)
Elija una configuración de seguridad o cree una nueva que pueda reutilizar con otros clústeres.

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](#)

Modos de lanzamiento

Normalmente, cuando utilizamos un clúster para procesar analíticas, interactuar con aplicaciones de *big data* o procesamiento de *datasets* de forma periódica, el clúster está siempre corriendo, a no ser que lo detengamos nosotros de forma explícita. Pero si queremos que sólo exista durante la ejecución de uno o más trabajos, el cual se le conoce como clúster *transient* o de ejecución por pasos, al terminar de ejecutar los pasos indicados, el clúster se detendrá.

Preparando al clúster

En un clúster EMR, el nodo maestro es una instancia EC2 que coordina al resto de instancias EC2 que corren los nodos principales y de tareas. Este nodo expone un DNS público el cual podemos utilizar para conectarnos.

Por defecto, EMR bloquea el arranque de los clústeres que permitan las conexiones del exterior (esto nos puede pasar si al crear el clúster le asignamos un grupo de seguridad que ya teníamos creado y que permitía el tráfico entrante).

Desbloqueando el acceso público

Para permitir el acceso desde el exterior, podemos arrancar un clúster con el tráfico cerrado, y una vez ya ha arrancado desactivar la opción de *Bloquear el acceso público* de EMR:

Amazon EMR ✕

EMR sin servidor

▼ **EMR en EC2**

- Clústeres
- Blocs de notas y repositorios de Git
- Eventos
- Bloquear el acceso público**
- Configuraciones de seguridad

▼ **EMR en EKS**

- Clústeres virtuales

Bloquear el acceso público [Información](#)

El bloqueo del acceso público de Amazon EMR impide el lanzamiento de un clúster cuando está asociado a reglas del grupo de seguridad que permiten el tráfico entrante desde IPv4 0.0.0.0/0 o IPv6 ::/0 (acceso público) en un puerto, a menos que el puerto se especifique explícitamente como una excepción.

Configuración del bloqueo del acceso público

Bloquear el acceso público
🟢 **Activado**

Excepciones de rango de puertos
Un clúster se puede lanzar con reglas del grupo de seguridad que permiten el tráfico entrante desde todas las direcciones IP públicas en estos puertos. El puerto 22 se agrega como una excepción de manera predeterminada por SSH.

22

Configurando el bloqueo del acceso público

Si editamos las opciones, podemos desbloquear el acceso público:

Amazon EMR > EMR en EC2: Bloquear el acceso público > Editar configuración

Editar configuración

Configuración del bloqueo del acceso público

Bloquear el acceso público
Este cambio solo afecta a los nuevos clústeres de EMR. Los clústeres de EMR existentes no se ven afectados.

☐ Activar *(recomendado)*
Bloquear el acceso público a todos los puertos, excepto los que se agregan como excepciones.

☒ Desactivar
Permitir el acceso público en función de las reglas del grupo de seguridad.

Cancelar Guardar

Desactivando el bloqueo del acceso público

Editando el grupo de seguridad

Por defecto, EMR crea un grupo de seguridad para el nodo maestro el cual determina el acceso. De inicio, ese grupo de seguridad no permite las conexiones SSH. Por ello, antes de poder conectarnos al clúster, también necesitamos modificar el grupo de seguridad del nodo principal para permitir todo el tráfico TCP (el tráfico SSH ya está abierto por defecto).

Para ello, en la pantalla de información del clúster que acabamos de crear, en la pestaña de *Propiedades*, en *Red y Seguridad*, editaremos el grupo de seguridad del nodo principal:

Red y seguridad Información

Red

Virtual Private Cloud (VPC)
vpc-0ed0e7eadd11cee6a

Subredes y zonas de disponibilidad (AZ)
subnet-043ccf3c665c3ae87 us-east-1a

▼ Grupos de seguridad de EC2 (firewall)

Nodo principal
Grupos de seguridad administrados de EMR
sg-0922c3aad1a8e5fdc

Grupos de seguridad adicionales
-

Nodos principales y de tareas
Grupos de seguridad administrados de EMR
sg-057ba74d1525a9b6d

Grupos de seguridad adicionales
-

Configuración de seguridad

Configuración de seguridad
Ninguna

Par de claves de EC2
vockey

Permisos

Rol de servicio para Amazon EMR
EMR_DefaultRole

Perfil de instancia EC2
EMR_EC2_DefaultRole

Rol de escalamiento automático personalizado
No configurado

Seleccionando el grupo de seguridad

Tras *Editar reglas de entrada*, permitimos todo el tráfico entrante (cuidado que esta es una práctica de seguridad muy mala, sólo lo hacemos así por comodidad):

Todos los TCP TCP 0 - 65535 Anywh... 0.0.0.0/0 X

Agregar regla

Rules with source of 0.0.0.0/0 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

Permitiendo todas las conexiones de entrada

Conectándonos al clúster

- El *namenode* de HDFS
- *Hue* (mal traducido en la interfaz como *Tonalidad*)
- El interfaz de *Tez*
- O la de *YARN*

Aplicaciones accesibles

IU de la aplicación en el nodo principal	
Estas requieren que el túnel de SSH esté habilitado.	
Aplicación	URL de la IU 
Administrador de recursos	http://ec2-44-200-116-159.compute-1.amazonaws.com:8088/
Nodo del nombre de HDFS	http://ec2-44-200-116-159.compute-1.amazonaws.com:9870/
Tonalidad	http://ec2-44-200-116-159.compute-1.amazonaws.com:8888/
UI de Tez	http://ec2-44-200-116-159.compute-1.amazonaws.com:8080/tez-ui

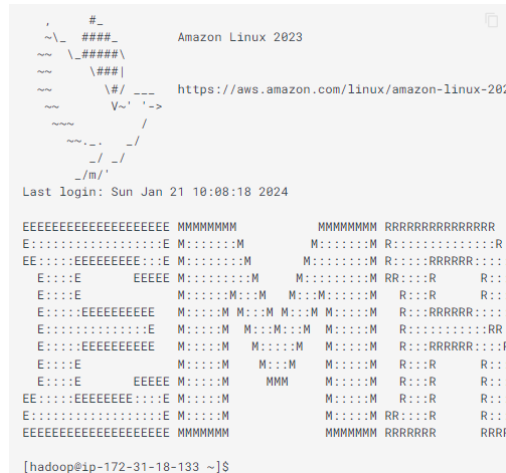
Una vez arrancado, podríamos tener un par de propiedades que modificar, ya que el acceso web a HDFS podría estar bloqueado, y la ruta de HDFS en *Hue* estar mal configurada.

Mediante SSH

Así pues, vamos a conectarnos vía SSH para activar el navegador web de HDFS. Con nuestras claves (nuestro archivo ***.pem**) descargadas, nos conectamos al clúster. El comando sería algo similar a:

```
ssh -i misclaves.pem hadoop@ec2-18-212-83-167.compute-1.amazonaws.com
```

Al conectarnos, tanto EC2 como EMR nos darán la bienvenida:



Visualizando HDFS

Si no pudiéramos visualizar el sistema de archivos de HDFS, primero le cambiamos los permisos al archivo `hdfs-site.xml` para poder editarlo sin problemas:

```
sudo chmod 777 /usr/lib/hadoop/etc/hadoop/hdfs-site.xml
```

```
nano /usr/lib/hadoop/etc/hadoop/hdfs-site.xml
```

Editamos la propiedad `dfs.webhdfs.enabled` y la ponemos a `true` (podemos buscarla mediante CTRL + W, editarla, y luego guardar y salir mediante CTRL + X):

<property>

<name>dfs.webhdfs.enabled</name>

<value>>false</value>

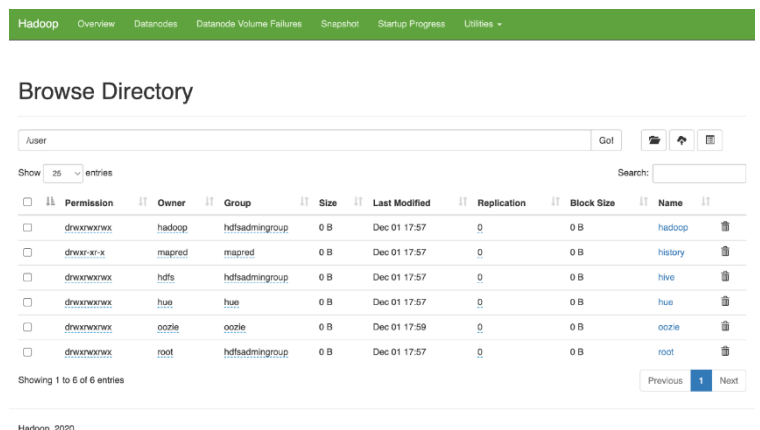
</property>

Sólo nos queda reiniciar el servicio de HDFS:

```
sudo systemctl restart hadoop-hdfs-namenode
```

Posteriormente agrega al grupo de seguridad del nodo principal una nueva regla que nos permita acceso a ese puerto (9870)

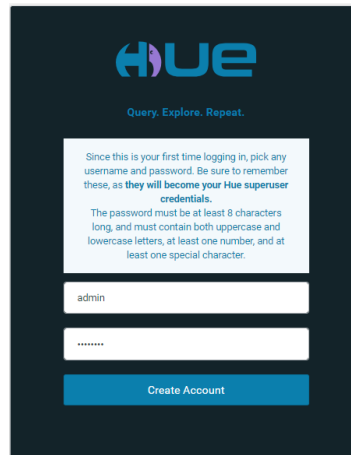
Y ahora desde el interfaz de HDFS ya podemos navegar por las carpetas y ver el contenido, por ejemplo, accediendo a `http://ec2-IP-Pública.compute-1.amazonaws.com:9870:`



Hue y HDFS

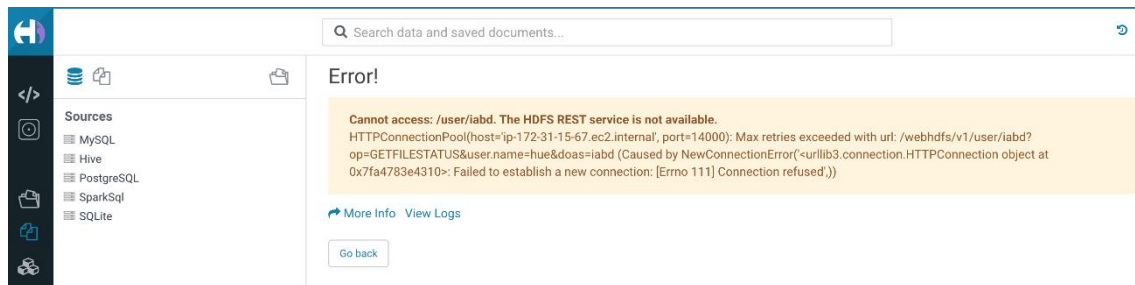
Cuando arranca Hue, la primera vez nos pide crear un usuario y contraseña.

Interfaz de HDFS



Login de acceso a Hue

Si intentamos visualizar los archivos y nos diera un error, de igual modo, pudiera ser que *Hue* no tiene bien configurado el acceso a HDFS:



Error al mostrar los archivos en Hue

Para ello, igual que hemos realizado antes, cambiamos los permisos y editamos el fichero de configuración:

```
sudo chmod 777 /usr/lib/hue/desktop/conf/hue.ini
```

```
nano /usr/lib/hue/desktop/conf/hue.ini
```

Y configuramos bien el puerto de acceso (debe ser 9870), dentro del grupo `[[hdfs_clusters]]` de `[hadoop]` editamos la propiedad `webhdfs_url` (sólo debes cambiar el puerto a **9870**):

```
webhdfs_url = http://ip-172-31-18-133.ec2.internal:9870/webhdfs/v1
```

Y reiniciamos el servicio:

```
sudo systemctl restart hue
```

Sólo nos queda crear una carpeta para nuestro usuario de HUE en HDFS:

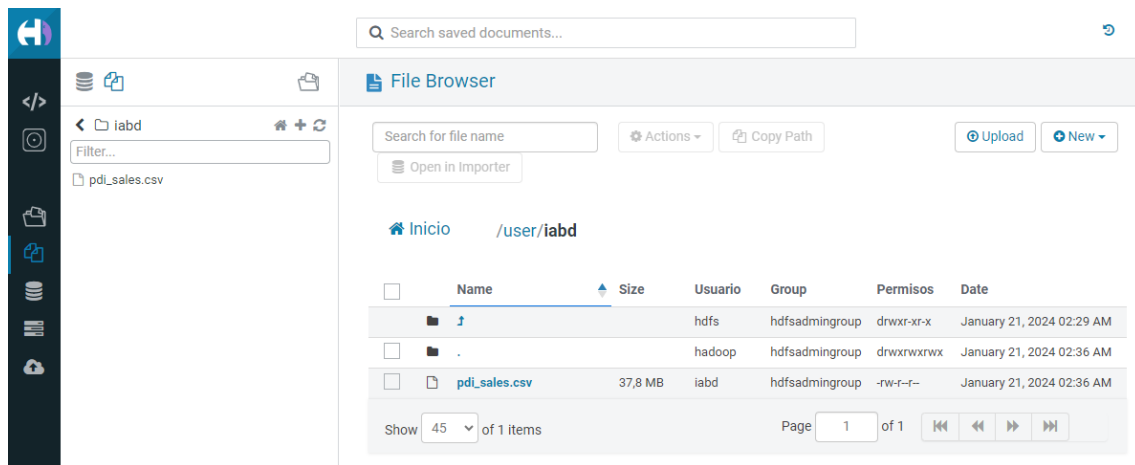
```
hdfs dfs -mkdir /user/miusuario
```

```
hdfs dfs -chmod 777 /user/miusuario
```

Por ejemplo, podemos subir un archivo cualquiera, ya sea mediante el interfaz gráfico o comandos y copiarlo mediante el comando `hdfs` adecuado:

```
hdfs dfs -put archive.csv /user/miusuario
```

Y a continuación visualizarlo desde *Hue*:



Visualizando HDFS desde Hue

Escalando

Podemos ajustar el número de instancias EC2 con las que trabaja nuestro clúster de EMR, ya sea manualmente o de forma automática en respuesta a la demanda que reciba.

Para ello, en la pestaña de *Instancias (hardware)*, tenemos la opción de *Editar opción de escalado de clústeres*:

Amazon EMR > EMR en EC2: Clústeres > ClusterEMRs8a > Editar escalado de clúster

Editar escalado de clúster Información

☐ 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.

Tamaño mínimo del clúster

1 Instancias

Tamaño máximo del clúster

2 Instancias

Cantidad máxima de nodos principales en el clúster

Limite la cantidad de nodos principales en su clúster.

2 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.

2 Instancias

Cancelar

Guardar cambios

Escalando un clúster EMR

Para ello, podemos activar el escalado gestionado mediante EMR o crear una política de escalado a medida. Independiente del modo, hemos de considerar que siempre hemos de tener de uno a tres nodos maestros, y que una vez creado el clúster, este número no lo podemos cambiar. Lo que sí que podemos es añadir y eliminar nodos principales o de tareas.

[Amazon EMR](#) > [EMR en EC2: Clústeres](#) > [ClusterEMRs8a](#) >
Agregar grupo de instancias de tareas para el clúster j-F8W0GH1V0EMM

Agregar grupo de instancias de tareas para el clúster j-F8W0GH1V0EMM

Grupo de instancias de tareas [Información](#)
Agregue un grupo de instancias de tareas para aumentar las instancias de Amazon EC2 en respuesta al aumento de la carga de trabajo.

Nombre

Elegir tipo de instancia de EC2

m4.large
2 vCore 8 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: - Precio de spot más bajo: -

Acciones ▼

► Configuración de nodo - opcional

Tamaño de grupo de instancias
 Instancias

☐ Usar la opción de compra de spot

[Cancelar](#) [Agregar grupo de instancias de tareas](#)

Añadiendo un nodo de tipo tarea

Conviene destacar que no podemos reconfigurar y redimensionar el clúster al mismo tiempo, de manera que hasta que no acabe la reconfiguración de un grupo de instancias no se puede iniciar el redimensionado.

Costes

Es muy importante ser conscientes de los [costes](#) que lleva utilizar EMR. *Grosso modo*, EMR supone un 25% de sobrecoste a las instancias EC2, es decir, pagaremos el coste del alquiler de las máquinas EC2 más un sobre un incremento del 25%.

Por ejemplo, para 20 nodos con 122 Gb RAM y 16 CPU, pagaríamos unos 32 €/h. En cambio, si sólo usamos las instancias que nos permite *AWS Academy* para practicar, pagaremos 0,13\$/hora por instancia de m4.large (2 CPU y 8GB RAM).