

Laboratorio N6.

Curso: ST0263 – Tópicos Especiales de Telemática

Título: Creación de Clúster EMR - Hadoop.

Objetivo: Desplegar un clúster de Hadoop utilizando el servicio de AWS EMR.

Duración: 20 mins

Entrega:

- La entrega se debe realizar por el buzón de interactiva virtual. SOLO por este medio.
- Se debe replicar la guía de manera completa para poder realizar la entrega.
- Para verificar la realización del laboratorio y verificar su funcionamiento, una vez termine y funcione, debe realizar un video con captura de su pantalla de no más de cinco (5) mins.
- En el video se debe explicar el funcionamiento de la arquitectura empleada y los detalles de la implementación.
- De esta forma el estudiante debe recorrer en su estación de trabajo y consola de AWS con el fin de mostrar la evidencia, así como el resultado de la implementación de cada uno de los pasos desarrollados en esta guía.
- El orden en que se debe presentar las funcionalidades es la siguiente:
 - Creación y explicación del bucket S3.
 - Creación y explicación paso a paso de un cluster EMR.
 - Explique como queda su cluster EMR con las opciones seleccionadas.
 - Explicar clara y detalladamente cada una de las opciones de aplicación seleccionadas para la creación del cluster (p.ej., apache flink, spark, hive, hue, etc)
 - Accesos al clúster EMR
 - Gestión del sistema de archivos HDFS

Contenido del Laboratorio

1	Introducción.	2
2	Background	2
3	Recursos	3
4	Desarrollo	3
4.1	Creación de Bucket S3.	3
4.2	Creación Key	4
4.3	Creando Clúster EMR	5
4.4	Accediendo al clúster EMR.	8
4.4.1	Modificando permisos Security Group Main Node.....	8
4.4.2	Estableciendo una session ssh con el Main Node.....	8
4.4.3	Accediendo el Clúster via Web Interface.	9
4.4.4	Gestionando el sistema de archivos HDFS desde la interfaz de consola.....	12

1 Introducción.

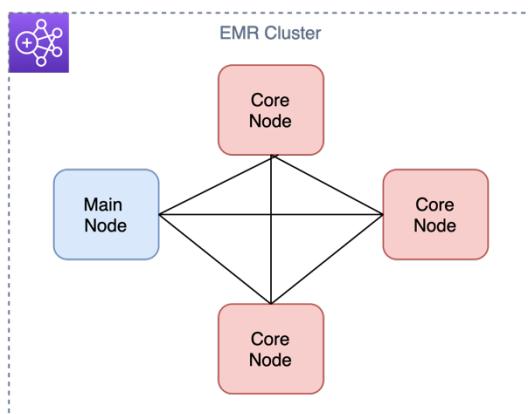
Amazon EMR es un servicio gestionando que permite la ejecución de frameworks para el procesamiento de grandes volúmenes de datos tales como Apache Hadoop, Spark, etc.

En este laboratorio vamos a desplegar un clúster con el fin de tener la infraestructura necesaria para el procesamiento de grandes volúmenes de datos.

2 Background

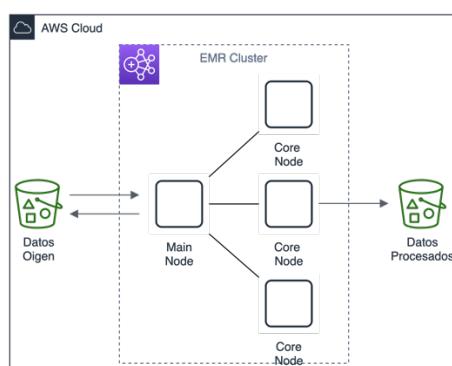
Amazon EMR es un servicio gestionando que permite la ejecución de frameworks para el procesamiento de grandes volúmenes de datos tales como Apache Hadoop, Spark, etc.

A nivel arquitectónico, el componente central es un elemento denominado clúster. En términos generales, un clúster es un conjunto de instancias EC2. Cada instancia es denominado nodo donde cada uno tiene un role (main node y core node). Sobre este clúster, se instalan y despliegan los diferentes frameworks o aplicaciones.



De esta forma, la instancia con el role de main node se encarga de la gestión del clúster, así como la coordinación de las tareas y distribución de los datos hacia los core node. Es importante resaltar que todo clúster tiene un main node. Por otro lado, los core node, son responsables de la ejecución de las tareas y el almacenamiento de los datos en el clúster HDFS.

Una vez desplegado el clúster, se realizará el procesamiento de los datos considerando el siguiente esquema:



En términos generales, para el procesamiento de los datos, éstos pueden estar almacenados como archivos en un tipo de sistema de archivos, en el caso de este laboratorio, puede ser S3 o HDFS.

Es importante tener en cuenta que los servicios/aplicaciones que se desplegarán en este laboratorio son los siguientes:

- **Hadoop User Experience (Hue):** Es un proyecto apache que permite la gestión vía interfaz web. De esta forma se facilita la creación, mantenimiento, así como ejecución de muchos trabajos en el ecosistema Hadoop.
- **Apache Hadoop:** Es un framework que permite el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras utilizando modelos de programación simples.
- **Apache Hive:** Es un sistema de almacenamiento de bodegas de datos tolerante a fallas y distribuido para el procesamiento de datos a gran escala. Todo el análisis de los datos que residen en el almacenamiento distribuido puede ser desarrollado a través de SQL. Hive es construido sobre Apache Hadoop.
- **Apache Spark:** Spark es un motor para el procesamiento de datos compatible con hadoop. De esta forma, puede ejecutarse en clústeres de Hadoop a través de YARN y puede procesar datos en HDFS, HBase, Cassandra, Hive y cualquier formato de entrada de Hadoop. Está diseñado para realizar procesamiento por lotes (similar a MapReduce) o bien sea por streaming, queries interactivos y aprendizaje de máquina.
- **Sqoop:** es una aplicación con interfaz de línea de comando para transferir datos entre bases de datos relacionales y Hadoop.

3 Recursos

Para el desarrollo de este laboratorio, se emplearán los siguientes servicios:

- Simple Storage Service (S3).
- Elastic Map Reduce (EMR).

4 Desarrollo

A continuación, se desarrollarán el conjunto de pasos para desplegar un clúster EMR. Para efectos de este laboratorio, se va a gestionar el clúster desde la interfaz de consola. Vale resaltar, que es posible interactuar con este servicio igualmente vía AWS CLI o API.

4.1 Creación de Bucket S3.

S3 es un servicio de almacenamiento de objetos en AWS. Es así como los datos (objetos) se almacenan en un bucket. Es importante comprender que cada bucket debe tener un nombre único a nivel global. Cada objeto que se pueda almacenar puede tener un tamaño que va desde los 0 bytes hasta los 5 TB.

Uno de los casos de uso es que S3 es empleado como origen de datos para su procesamiento y análisis. Para efectos de este laboratorio se va a crear un bucket S3 con el fin de que sirva para ser el origen/destino de los datos de entrada/salida.

Es importante mencionar que el nombre de este bucket debe cumplir con las siguientes características:

- Letras en minúsculas. Es posible utilizar símbolos como puntos(.) y guiones (-).
- Tenga en cuenta que el nombre no debe terminar en un número.

A continuación, se va a crear el bucket:

- En la consola, en la casilla de búsqueda digite S3 y seleccione este servicio.
- Click en Create bucket.
- BucketName: <username>-lab-emr
- Click en Create bucket.

Name	AWS Region	Access	Creation date
jcmontoy-lab-emr	US East (N. Virginia) us-east-1	Bucket and objects not public	May 22, 2023, 06:28:09 (UTC-05:00)

Ahora, seleccione el bucket creado y de click en este.

Name	Type	Last modified	Size	Storage class
No objects You don't have any objects in this bucket.				

- Click en **Create folder**
- Folder name: **data**
- Click en **Create folder**
- Click en **Create folder**
- Folder name: **logs**
- Click en **Create folder**

4.2 Creación Key

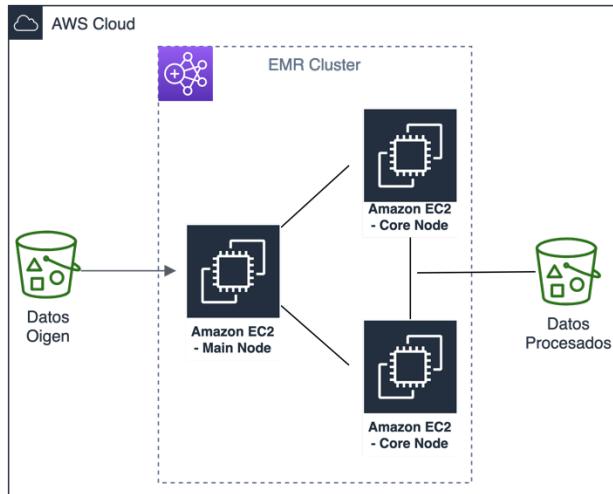
Para la creación del clúster se hace necesario crear una llave para poder tener acceso luego a éste. Para esto se deben ejecutar los siguientes pasos:

- En el menú de EC2, seleccione en la sección de Network and Security.
- Click en la opción Key Pairs
- Click en Create key pair
- Key pair:

- Name: **emr-key.pem**
- Key pair type: **Seleccione RSA.**
- Private key file format: **Seleccione .pem**
- Click en Create key pair

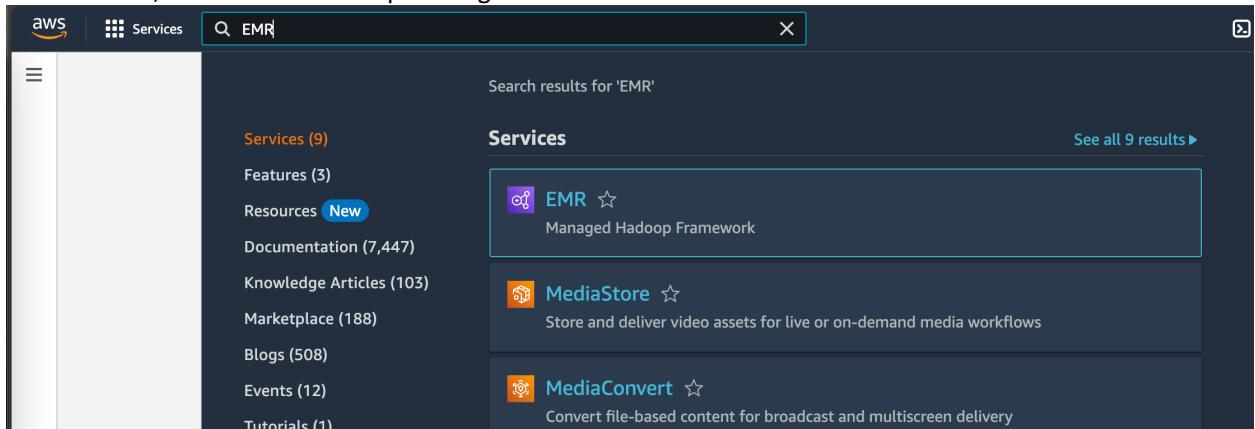
4.3 Creando Clúster EMR

En esta sección se va a crear el clúster EMR. El clúster EMR a implementar en este laboratorio estará compuesto de un (1) main node y dos (2) core tal como se ilustra en la figura a continuación:

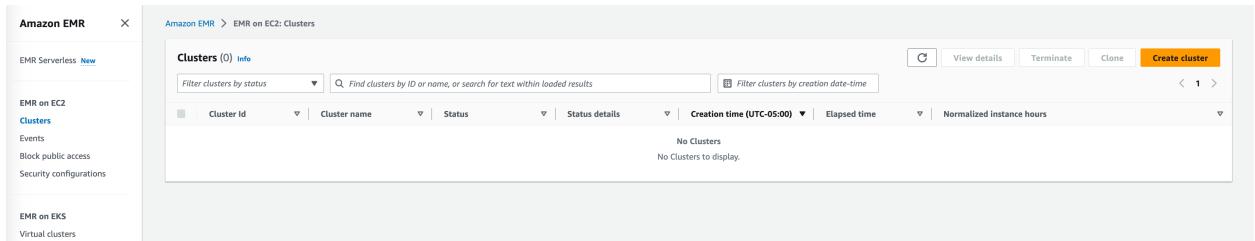


Para lograr esto, se va a ejecutar los siguientes pasos:

- En la consola, en la casilla de búsqueda digite EMR.



- Ahora, proceda a crear y configurar un cluster de Map/reduce



- Click en la opción, **Create cluster**.
- **Name and Applications:**
 - Name: **emr-MyClusterEMR**
 - Amazon EMR release: **emr-6.14.0**
 - Application Bundle: **Custom**
 - Click en Customize your application bundle para desplegar lista de opciones.
 - Applications included in bundle. Seleccione la siguiente lista de componentes para el clúster EMR:
 - **Flink 1.17.1**
 - **HCatalog 3.1.3**
 - **Hue 4.11.0**
 - **Livy 0.7.1**
 - **Spark 3.4.1**
 - **Tez 0.10.2**
 - **ZooKeeper 3.5.10**
 - **Hadoop 3.3.3**
 - **Sqoop 1.4.7**
 - **Hive 3.1.3**
 - **JupyterHub 1.5.0**
 - **Zeppelin 0.10.1**
 - **Oozie 5.1.0**

Name
emr-MyClusterEMR

Amazon EMR release | [Info](#)
A release contains a set of applications which can be installed on your cluster.
emr-6.14.0

Application bundle

Spark Interactive 	Core Hadoop 	Flink 	HBase 	Presto 	Trino 	Custom
-----------------------	-----------------	-----------	-----------	------------	-----------	------------

Flink 1.17.1 Ganglia 3.7.2 HBase 2.4.17
 HCatalog 3.1.3 Hadoop 3.3.3 Hive 3.1.3
 Hue 4.11.0 JupyterEnterpriseGateway 2.6.0 JupyterHub 1.5.0
 Livy 0.7.1 MXNet 1.9.1 Oozie 5.2.1
 Phoenix 5.1.3 Pig 0.17.0 Presto 0.281
 Spark 3.4.1 Sqoop 1.4.7 TensorFlow 2.11.0
 Tez 0.10.2 Trino 422 Zeppelin 0.10.1
 ZooKeeper 3.5.10

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

- En la opción: AWS Glue Data Catalog settings
 - Active la casilla: Use the AWS Glue Data Catalog to provide an external metastore for your application.
 - Active la casilla: Use for Hive table metadataUse for Spark table metadata

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

- **Cluster Configuration:**
 - Seleccione Uniform Instance groups
 - Uniform Instance groups.
 - Primary. Choose EC2 instance type: **Seleccione m4.large.**
 - Core. Choose EC2 instance type: **Seleccione m4.large**
 - Task. Choose EC2 instance type: **Seleccione m4.large**
 - Cluster Scaling and provisioning option.
 - Mantenga la opción seleccionada de **Set cluster size manually.**
 - Mantenga la opción **size para Core y Task-1 a los valores por defecto.**
- **Cluster termination**

- Desactive la casilla: **Use termination protection**
- **Cluster logs – optional:**
 - Active la casilla **Publish cluster-specific logs to Amazon S3.**
 - **Amazon S3 location.** Click en **Browse S3** y seleccione la carpeta creada anteriormente **logs** (**s3://jcmontoy-lab-emr/logs**).
- **Security configuration and EC2 key pair – optional:**
 - Amazon EC2 key pair for SSH to the cluster – *optional*: **Seleccione la llave creada emr-key.pem.**
- **Identity and Access Management (IAM) roles:**
 - **Amazon EMR service role:** Click en **Choose an existing service role**.
 - **Service Role:** **Seleccione EMR_DefaultRole.**
 - **EC2 instance profile for Amazon EMR:** Click en **Choose an existing instance profile**.
 - **Instance profile:** **Seleccione EMR_EC2_DefaultRole.**

Finalmente verifique que el clúster haya sido desplegado correctamente. Tenga presente que esta operación puede tardar alrededor de unos 10 mins aprox. Para esto debe estar en estado “Waiting”.

4.4 Accediendo al clúster EMR.

Con el fin de acceder al clúster se deben ejecutar los siguientes pasos:

4.4.1 Modificando permisos Security Group Main Node.

- En la consola, diríjase al servicio EC2. Escoja la opción de **Security Groups**.
- Seleccione el security group para el master/main node.

- Seleccione la pestaña de **Inbound Rules**.
- Click en **Edit inbound rules**.
- Click en **Add rule**.
 - Type : **Seleccione protocol SSH.**
 - Source: **Anywhere**
- Click en **Save rules**.

4.4.2 Estableciendo una session ssh con el Main Node.

En primer lugar, cambie los permisos para la llave (emr-key.pem) creada anteriormente. Para esto y ubicado en el directorio aplique el comando:

```
$ chmod 400 emr-key.pem
```

Ahora, se hace necesario localizar el nombre dns público para el main node. Para esto, ejecute los siguientes pasos:

- Seleccione el clúster creado.
- Click en View details.
- Seleccione la pestaña Instances.
- Identifique y seleccione el nodo primary (main node). De click en este nodo.
- Copie el valor de Public DNS name.

The screenshot shows the AWS EMR console interface. At the top, it displays the navigation path: Amazon EMR > EMR on EC2: Clusters > emr-MyClusterEMR > ig-2SLHSJOV5000D. Below this, the title "Instance group ig-2SLHSJOV5000D" is shown. Under "Summary", there's a table with columns: Node type & name (Primary & Primary), Creation date (May 22, 2023, 08:48 (UTC-05:00)), Elapsed time (39 minutes, 44 seconds), and Status (Running). In the "Instances (1) Info" section, a table lists one instance: ID (ci-00061071XQJH...), EC2 Instance ID (i-06eb5bd92cd957e49), Instance type (m5.xlarge), Status (Running), EBS volume per instance (2), Public DNS name (ec2-3-234-255-116.compute-1.amazonaws.com), Private DNS name (ip-172-31-0-85.ec2.ir), Public IP address (3.234.255.116), Private IP address (172.31.0.85), and Purchasing option (On-Demand).

Para conectarse al clúster, vamos a utilizar el usuario “hadoop”. Ahora, se debe ejecutar el siguiente comando (utilice el nombre del archivo .pem que usted creo y el nombre DNS público de su main node):

```
$ ssh -i emr-key.pem hadoop@ ec2-###-##-##-##.compute-1.amazonaws.com
```

Debe observar un mensaje similar al que se muestra en la figura a continuación.

```
Last login: Mon May 22 14:31:57 2023
[          ] /   Amazon Linux AMI
[          ] /   Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
49 package(s) needed for security, out of 75 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEE MMMMMMM MRRRRRRRRRRRRRR
E:::::::EE:::::E M:::::M M:::::M R:::::R
EE:::::EEEEE:::E M:::::M M:::::M R:::::RRRRR:::::R
E:::::E     EEEE M:::::M M:::::M RR:::::R      R:::::R
E:::::E     M::::M M:::::M M:::::M R:::::R      R:::::R
E:::::EEEEE::::E M:::::M M:::::M M:::::M R:::::RRRRR:::::R
E:::::E     M::::M M:::::M M:::::M M:::::M R:::::RRRRR:::::R
E:::::EEEEE:::::E M:::::M M:::::M M:::::M R:::::RRRRR:::::R
E:::::E     EEEE M:::::M M:::::M M:::::M R:::::R      R:::::R
E:::::E     M:::::M MMM M:::::M R:::::R      R:::::R
EE:::::EEEEE:::::E M:::::M M:::::M R:::::R      R:::::R
E:::::E     M:::::M M:::::M R:::::R      R:::::R
EEEEEEEEEEEEEEEEEE MMMMMMM MRRRRRR RRRRRR

[hadoop@ip-172-31-0-85 ~]$
```

4.4.3 Accediendo el Clúster via Web Interface.

Es posible gestionar el clúster EMR así como las aplicaciones/servicios desplegadas sobre éste (p.ej., Hue, Hadoop, etc) vía un servidor web que corre de manera local en el main node. Con el fin realizar una conexión de forma segura, debemos levantar un túnel vía ssh entre la estación cliente de trabajo y la instancia main

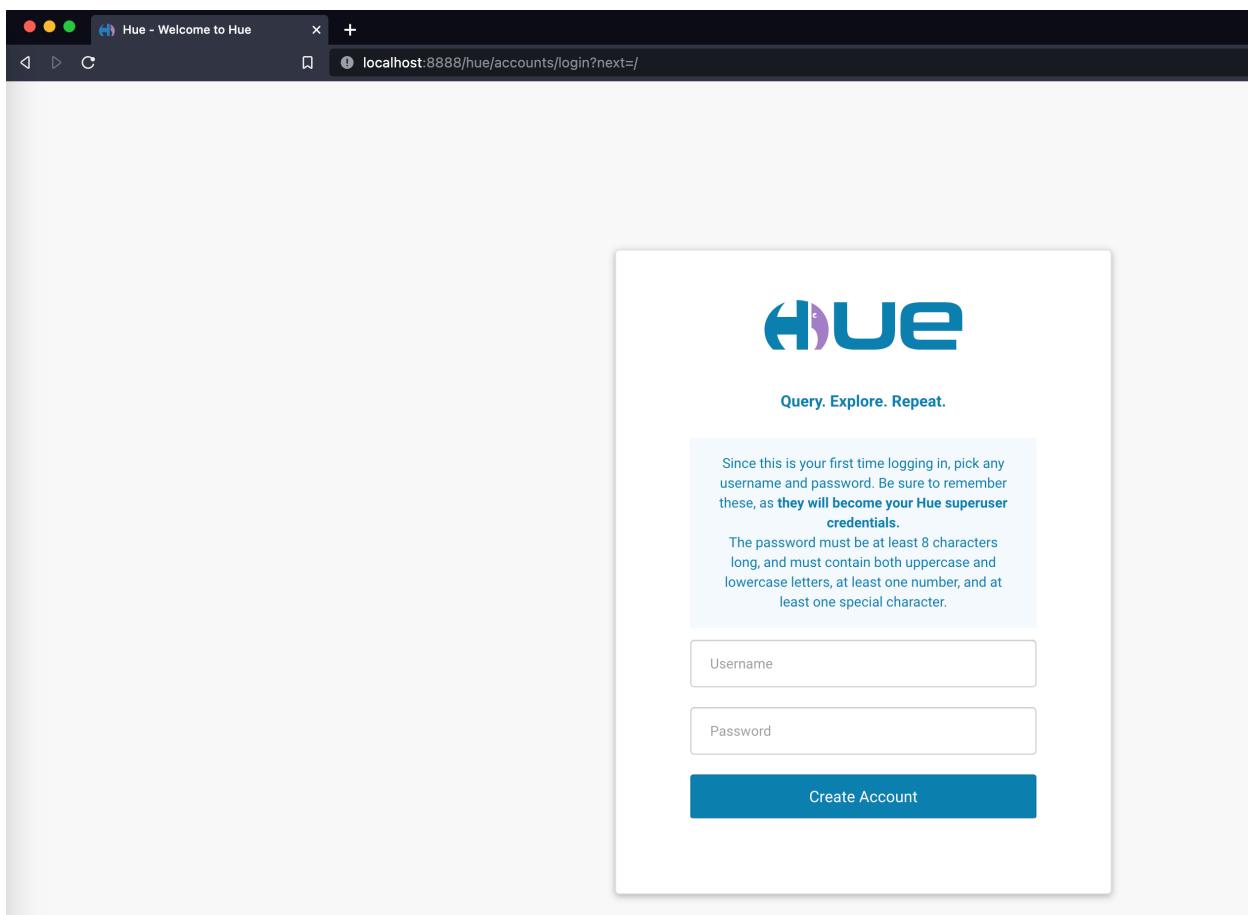
node del clúster. Todo esto empleado la característica de ssh de “local port forwarding”. Para lograr esto, ejecute los siguientes pasos:

- En un terminal de consola, digite el siguiente comando:

```
$ ssh -i emr-key.pem -N -L 8888:ec2-###-##-##-##.compute-1.amazonaws.com:8888 hadoop@ec2-###-##-##-##.compute-1.amazonaws.com
```

El flag – L significa que se está implementando el “local port forwarding”. Básicamente esto lo que permite es que se reciba el tráfico en un puerto local y se haga el forwarding hacia la estación remota (en este caso el main node) en el puerto en el cual se está ejecutando el servicio. En este caso, lo que se está accediendo es el servicio Hue el cual se ejecuta en el puerto 8888. Recuerde que ec2-###-##-##-##.compute-1.amazonaws.com es el nombre DNS público del main node.

Ahora para poder acceder y gestionar el servicio Hue vía web, abrá un browser y digite la URL <http://localhost:8888>



Cree una cuenta con el username “admin” y una contraseña que cumpla con las políticas y pueda recordar.

Ahora, de click en el ícono de menú, seleccione la opción Browers-S3 y puede observar el bucket S3 que se asoció en la creación del clúster EMR.

Name	Size	User	Group	Permissions	Date
.				drwxrwxrwx	
jcmonroy-lab-emr				drwxrwxrwx	

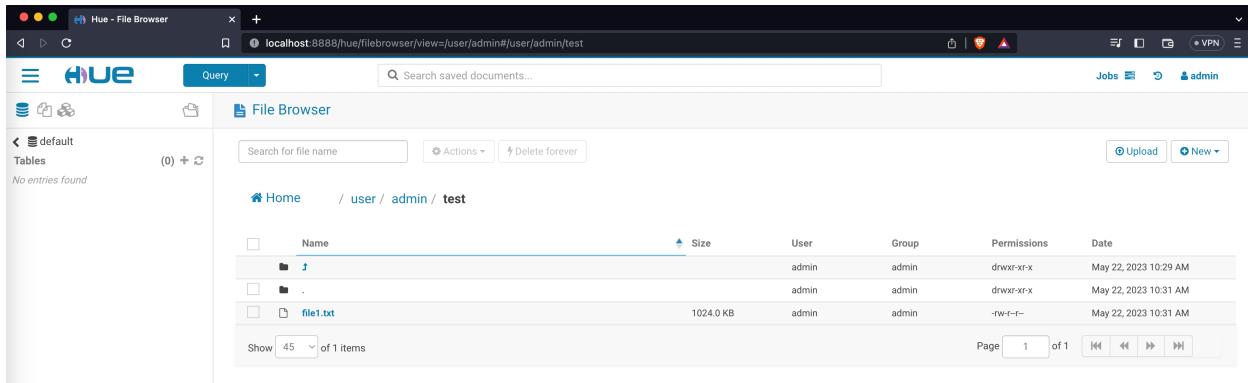
Recuerde que estamos desplegando un clúster EMR que está basado en Hadoop cuyo sistema de archivos es HDFS. A través de Hue podemos interactuar también con el sistema de archivos HDFS del cluster Hadoop. De click en el ícono de menú principal, seleccione la opción Browers-Files

Name	Size	User	Group	Permissions	Date
test		hdfs	hadoop	drwxr-xr-x	May 22, 2023 09:55 AM
.		admin	admin	drwxr-xr-x	May 22, 2023 09:55 AM

De click en la opción “New”, “Directory”. Cree un directorio llamado “test”.

Name	Size	User	Group	Permissions	Date
test		admin	admin	drwxr-xr-x	May 22, 2023 10:29 AM
..		admin	admin	drwxr-xr-x	May 22, 2023 09:55 AM
.		hdfs	hadoop	drwxr-xr-x	May 22, 2023 09:55 AM

Ahora, cree y suba un archivo de texto (text1.txt) en la carpeta que creo (test) al clúster EMR. Para esto seleccione la opción “Upload”, click en “Select Files” y seleccione el archivo de su sistema de archivos local.



4.4.4 Gestionando el sistema de archivos HDFS desde la interfaz de consola.

Para realizar esto, en una nueva terminal de consola, establezca una sesión ssh con el main node. Digite el siguiente comando:

```
$ ssh -i emr-key.pem hadoop@ ec2-###-##-##-##.compute-1.amazonaws.com
```

Ahora visualicemos el directorio así como archivo que creó desde la interfaz web, para esto digitemos el siguiente comando:

```
$ hdfs dfs -ls /user/admin
```

```
[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -ls /user/admin
Found 1 items
drwxr-xr-x - admin admin          0 2023-05-22 17:31 /user/admin/test
[hadoop@ip-172-31-11-95 ~]$
```

Como se puede observar, aparece en la consola el directorio "test". Ahora, vamos a listar el archivo, para esto digite el siguiente comando:

```
[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -ls /user/admin/test
Found 1 items
-rw-r--r-- 1 admin admin 1048575 2023-05-22 17:31 /user/admin/test/file1.txt
[hadoop@ip-172-31-11-95 ~]$
```

Como puede ver, el archivo file1.txt es visible.

Para efectos de esta guía, es equivalente el comando "hadoop fs" y "hdfs dfs". La diferencia es que "hdfs dfs" es solo para sistemas de archivos HDFS, pero "hadoop fs" soporta otros adicionales como Amazon S3.

A continuación, se va a crear un segundo directorio denominado "test2". Para esto utilice el siguiente comando:

```
$ hdfs dfs -mkdir /user/admin/test2
$ hdfs dfs -ls /user/admin
```

```
[[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -mkdir /user/admin/test2
[[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -ls /user/admin
Found 2 items
drwxr-xr-x  - admin  admin          0 2023-05-22 17:31 /user/admin/test
drwxr-xr-x  - hadoop admin          0 2023-05-22 18:14 /user/admin/test2
[hadoop@ip-172-31-11-95 ~]$
```

Se puede observar que el directorio "test2" ha sido creado. Por favor verifique desde la interfaz web.

Name	Size	User	Group	Permissions	Date
..		hdfs	hadoop	drwxr-xr-x	May 22, 2023 09:55 AM
.		admin	admin	drwxr-xr-x	May 22, 2023 11:14 AM
test		admin	admin	drwxr-xr-x	May 22, 2023 10:31 AM
test2		hadoop	admin	drwxr-xr-x	May 22, 2023 11:14 AM

Ahora vamos a subir un archivo file2.txt desde el sistema de ficheros local de la máquina main node al sistema de archivos distribuidos HDFS. Para esto, desde la terminal de consola y ubicado en el directorio /home/hadoop/ cree un archivo denominado file2.txt (diligencie cualquier información):

```
$ sudo nano file2.txt
$ hdfs dfs -copyFromLocal file2.txt /user/admin/test2/
```

Verifique que el archivo ya exista en el sistema de archivos HDFS vía consola y la interfaz web:

```
$ hdfs dfs -ls /user/admin/test2
```

```
[[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -copyFromLocal file2.txt /user/admin/test2/
[[hadoop@ip-172-31-11-95 ~]$ hdfs dfs -ls /user/admin/test2
Found 1 items
-rw-r--r--  1 hadoop  admin      53 2023-05-22 18:45 /user/admin/test2/file2.txt
[hadoop@ip-172-31-11-95 ~]$
```

Name	Size	User	Group	Permissions	Date
..		admin	admin	drwxr-xr-x	May 22, 2023 11:14 AM
.		hadoop	admin	drwxr-xr-x	May 22, 2023 11:45 AM
file2.txt	53 bytes	hadoop	admin	-rw-r--r--	May 22, 2023 11:45 AM