Prácticas con HDFS

En esta parte vamos a utilizar los comandos HDFS.

Comandos HDFS

El comando hdfs dfs es una interfaz de línea de comandos que permite interactuar con el sistema de archivos distribuido de Hadoop (HDFS). Proporciona una serie de subcomandos que permiten realizar operaciones de gestión de archivos y directorios dentro de HDFS, de forma similar a los comandos tradicionales de Unix/Linux.

¿Cómo Funciona?

Interfaz de Usuario: El comando “hdfs dfs” actúa como una puerta de enlace entre el usuario y HDFS, permitiendo ejecutar comandos desde la terminal para gestionar datos en el sistema de archivos distribuido.

Comunicación con HDFS

NameNode: Cuando ejecutas un comando hdfs dfs, este se comunica con el NameNode, que es el maestro que gestiona el espacio de nombres del sistema de archivos y los metadatos de HDFS.

DataNodes: Para operaciones que implican transferencia de datos (como lectura o escritura), el comando también interactúa con los DataNodes, que son los nodos que almacenan realmente los bloques de datos.

Sintaxis General

```
hdfs dfs –comando [opciones] [ruta(s)]
```

- **-**Especifica la operación que deseas realizar (por ej. -ls, -put, -get).
- **[opciones]:** Opcionalmente, puedes incluir opciones adicionales para modificar el comportamiento del comando.
- **[ruta(s)]:** Indica la ruta del archivo o directorio en HDFS o en el sistema local, según corresponda.

Lista de comandos HDFS

Apache Hadoop 2.4.1 - File System Shell Guide

hdfs dfs -ls

Lista los contenidos de un directorio en HDFS.

```
hdfs dfs –ls /ruta/del/directorio
```

hdfs dfs -mkdir

Crea un nuevo directorio en HDFS.

```
hdfs dfs –mkdir /ruta/del/nuevo/directorio
```

hdfs dfs -put

Copia un archivo o directorio desde el sistema de archivos local a HDFS.

```
hdfs dfs –put /ruta/local/archivo /ruta/hdfs/destino
```

```
hdfs dfs –put /home/hadoop/archivo.txt /user/hadoop/datos/
```

hdfs dfs -get

Descarga un archivo o directorio desde HDFS al sistema de archivos local.

```
hdfs dfs –get /ruta/hdfs/archivo /ruta/local/destino
```

hdfs dfs -rm

Elimina archivos o directorios en HDFS.

```
hdfs dfs –rm /ruta/hdfs/archivo
```

Nota: Para eliminar directorios y su contenido recursivamente, utiliza -rm -r.

hdfs dfs -rmdir

Elimina un directorio vacío en HDFS.

```
hdfs dfs –rmdir /ruta/hdfs/directorio
```

hdfs dfs -cat

Muestra el contenido de un archivo en HDFS.

```
hdfs dfs –cat /ruta/hdfs/archivo
```

hdfs dfs -moveFromLocal

Mueve un archivo desde el sistema local a HDFS y elimina el archivo local.

```
hdfs dfs –moveFromLocal /ruta/local/archivo /ruta/hdfs/destino
```

hdfs dfs -du

Muestra el uso de espacio en disco de archivos y directorios en HDFS.

```
hdfs dfs –du /ruta/hdfs
```

hdfs dfs -df

Muestra la capacidad, espacio libre y utilizado en HDFS. El modificador -h muestra la información en un formato legible para humanos.

```
hdfs dfs –df –h
```

hdfs dfs -chmod

Cambia los permisos de archivos y directorios en HDFS.

```
hdfs dfs –chmod [permisos] /ruta/hdfs/archivo
```

Los permisos en chmod se representan mediante un número de tres dígitos, donde cada dígito corresponde a los permisos del **usuario** (dueño), **grupo** y **otros**, respectivamente.

Estos números se derivan de la representación binaria de los permisos y se basan en la suma de los valores asignados a cada tipo de permiso:

- **Lectura (r)**: valor **4** (en binario: 100)
- **Escritura (w)**: valor **2** (en binario: 010)
- **Ejecución (x)**: valor **1** (en binario: 001)

Cuando ejecutas *chmod 755 archivo*, usuario propietario tiene todos los permisos sobre el archivo (lectura, escritura y ejecución). El grupo y otros usuarios pueden leer y ejecutar el archivo, pero no modificarlo.

hdfs dfs -chown

Cambia el propietario de archivos y directorios en HDFS.

```
hdfs dfs –chown [usuario][:grupo] /ruta/hdfs/archivo
```

hdfs dfs -chgrp

Cambia el grupo de archivos y directorios en HDFS.

```
hdfs dfs –chgrp [grupo] /ruta/hdfs/archivo
```

hdfs dfs -appendToFile

Añade contenido a un archivo existente en HDFS.

```
hdfs dfs –appendToFile /ruta/local/archivo_local /ruta/hdfs/archivo_hdfs
```

hdfs dfs -setrep

Cambia el factor de replicación de archivos y directorios en HDFS.

```
hdfs dfs –setrep [factor] /ruta/hdfs/archivo
```

hdfs dfs -stat

Muestra información sobre un archivo en HDFS.

```
hdfs dfs –stat [formato] /ruta/hdfs/archivo
```

Los especificadores de formato te permiten controlar qué información se muestra y cómo se muestra. Se representan mediante el símbolo % seguido de una letra que indica el tipo de información.

- %a: Muestra el último acceso al archivo en formato de fecha.
- %A: Muestra el último acceso al archivo en formato de fecha numérico (timestamp).
- %b: Muestra el tamaño del archivo en bytes.
- %F: Muestra el tipo de archivo.
- %g: Muestra el ID del grupo propietario del archivo.
- %n: Muestra el nombre del archivo.
- %o: Muestra la fecha de modificación en formato numérico (timestamp).
- %r: Muestra el factor de replicación del archivo.
- %u: Muestra el ID de usuario propietario del archivo.
- %y: Muestra la fecha de modificación en formato de fecha.
- %Y: Muestra la fecha de modificación en formato numérico (timestamp).

hdfs dfs -tail

Muestra las últimas líneas de un archivo en HDFS.

```
hdfs dfs –tail /ruta/hdfs/archivo
```

hdfs dfsadmin -report

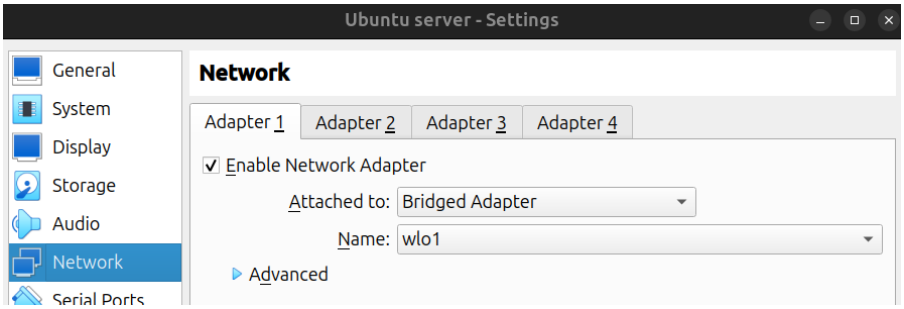
Proporciona un informe detallado del estado del clúster HDFS: información sobre el almacenamiento total, el espacio utilizado, el número de DataNodes, entre otros detalles.

```
hdfs dfsadmin –report
```

Arrancar Hadoop

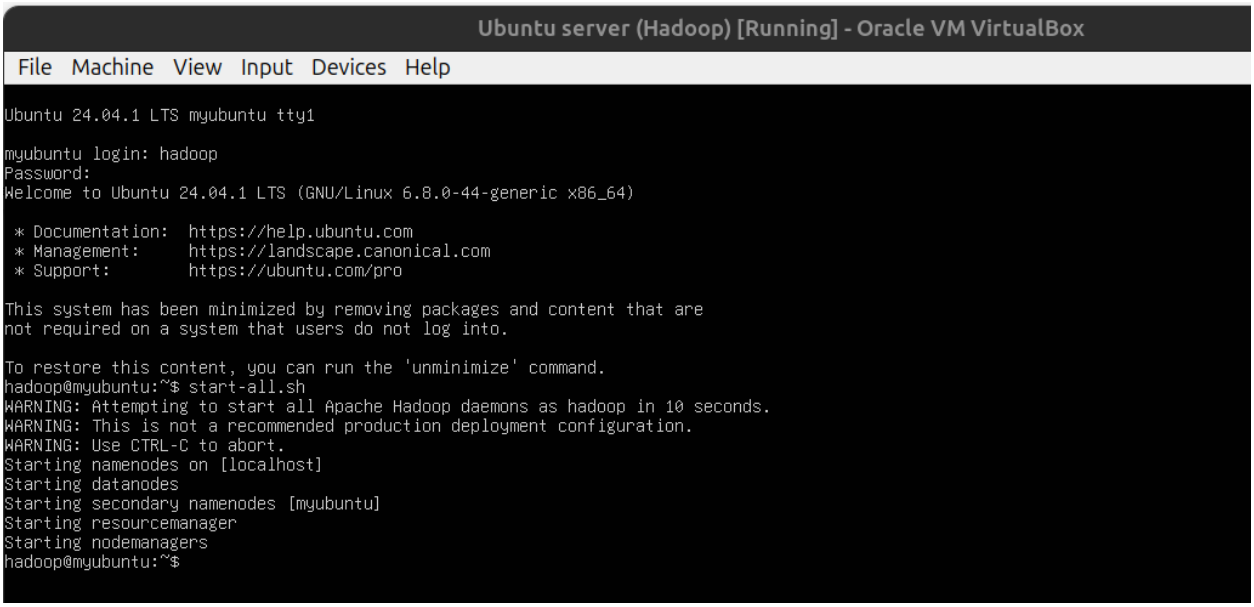
Abrimos VirtualBox.

Nos aseguramos que la configuración de Red de nuestra máquina virtual está en modo “Bridged Adapter”



Arrancamos la máquina virtual y hacemos login con el usuario hadoop (contraseña hadoop).

A continuación ejecutamos el script *start-all.sh* que arrancará los servicios asociados a hadoop



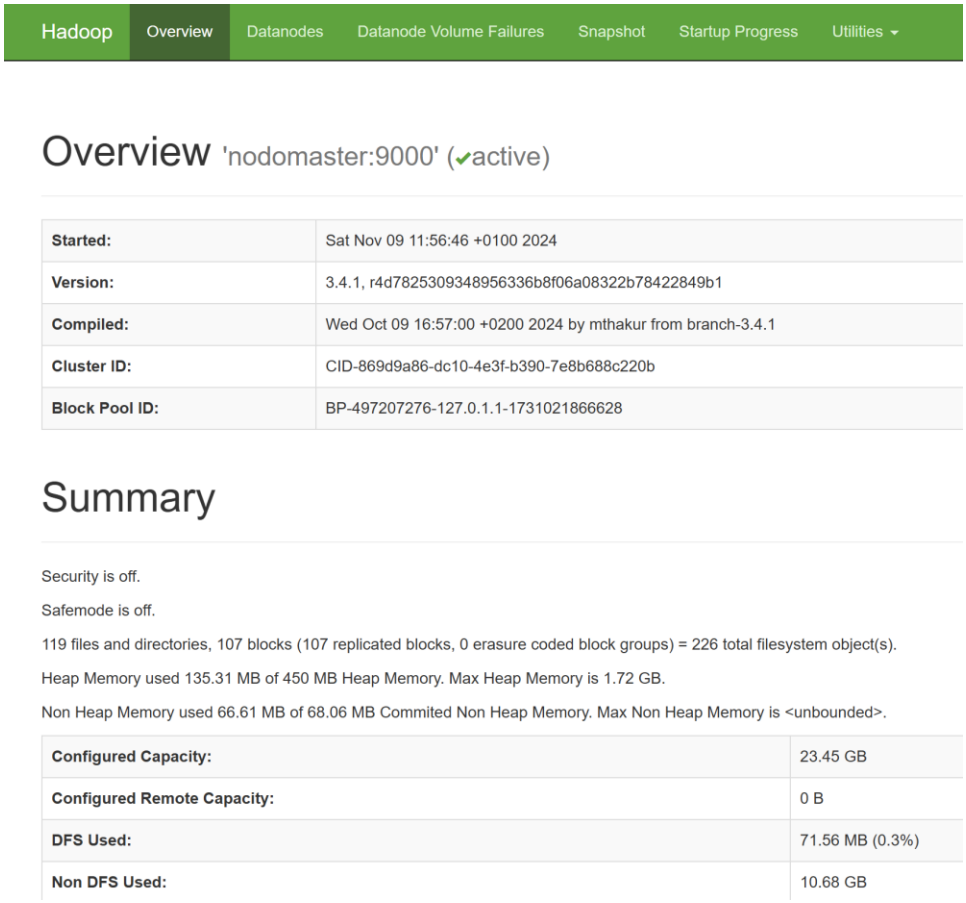
Una vez hecho login ejecutamos `ip a` en cada nodo para ver las IPs asignadas (enp0s3), y revisamos si es necesario cambiarlas en los ficheros hosts de cada nodo:

```
vi /etc/hosts
        192.168.0.45 nodomaster
        192.168.0.45 nodoworker1
```

Acceso NameNode

Comprobamos desde nuestro navegador que podemos acceder en esa IP al puerto 9870

El puerto 9870 en Hadoop es el puerto predeterminado donde se ejecuta la interfaz web del NameNode de HDFS. Cuando accedes a `http://<nombre_del_servidor>:9870/`, Hadoop muestra esta interfaz web que proporciona información detallada y herramientas de administración para el sistema de archivos distribuido de Hadoop (HDFS).

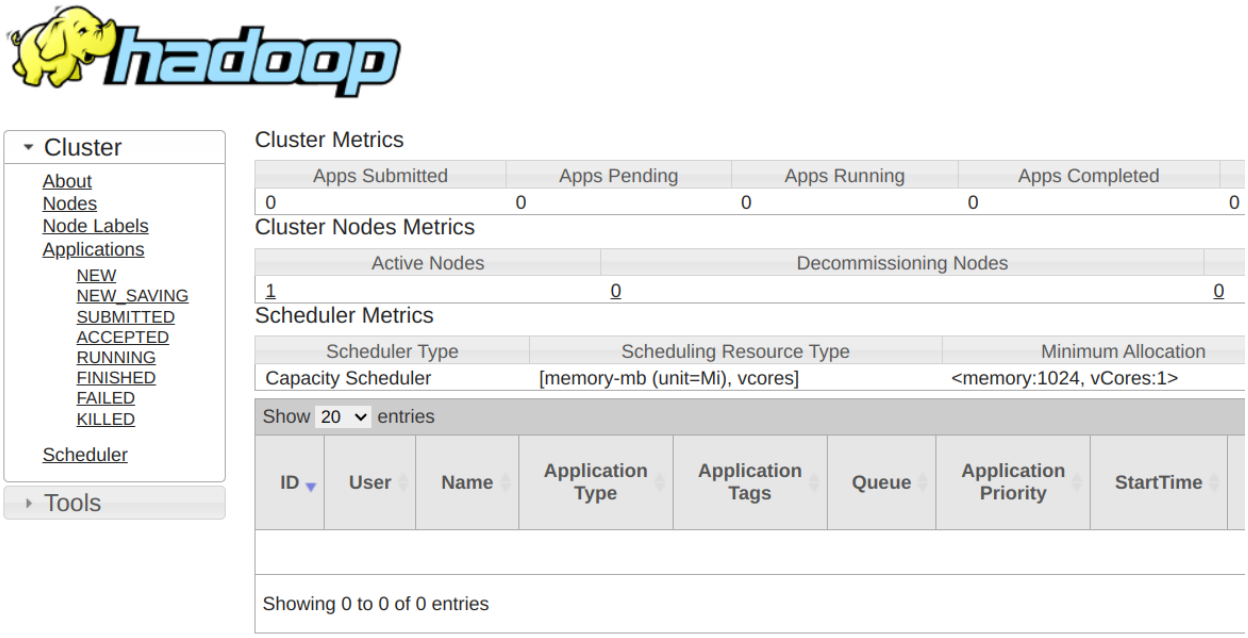


La interfaz web del NameNode en el puerto 9870 ofrece una variedad de información y funcionalidades para monitorizar y administrar HDFS:

- Muestra el estado operativo del NameNode (activo o en espera).
- Información sobre la capacidad total del sistema, espacio utilizado y espacio libre.
- Número de archivos y directorios almacenados en HDFS.
- Información de DataNodes: Lista de DataNodes vivos y muertos.
- Detalles sobre el espacio utilizado en cada DataNode. Estado de salud de los DataNodes y alertas si existen problemas.
- Explorador del Sistema de Archivos: Permite explorar el sistema de archivos HDFS de manera similar a un explorador de archivos.

Acceso Yarn ResourceManager

El puerto 8088 en Hadoop es el puerto predeterminado donde se ejecuta la interfaz web del ResourceManager de YARN. Al acceder a `http://<nombre_del_servidor>:8088/`, Hadoop muestra esta interfaz web que proporciona información detallada y herramientas de administración para YARN (Yet Another Resource Negotiator), el sistema de gestión de recursos y planificación de tareas de Hadoop.



Ejecución comandos HDFS

Desde la propia máquina virtual podemos ejecutar comandos HDFS que se conectan con nuestro cluster Hadoop.

Completa los siguientes pasos:

1- Sobre la máquina virtual, en el directorio home:
   a. Crea un directorio llamado “libros”
   b. Descarga varios libros del proyecto Gutenberg
2- Crea un directorio “entrada” y otro “salida” en el cluster hadoop
3- Copia los libros al directorio “entrada”
4- Comprueba que se han copiado, tanto por línea de comando como en la página web de monitorización de hdfs (`http://nodomaster:9870/`)

- Nota: No existe el comando para cambiar de directorio (cd). Además Hadoop no crea estructura de directorios, son metadatos almacenados en namenode.

5- Sobre el cluster, muestra el contenido de las 40 primeras líneas de cada fichero
6- Crea un snapshot del directorio “entrada” con el nombre “entrada_v1”

```
hdfs dfsadmin –allowSnapshot /entrada/
hdfs dfs –createSnapshot /entrada entrada_v1
```

7- Comprueba que se ha generado el snapshot:

```
hdfs dfs –ls /entrada/.snapshot
http://nodomaster:9870/dfshealth.html#tab-snapshot
```

8- Borra el fichero quijote.txt en el cluster y comprueba el contenido del directorio.
9- Restaura el fichero del snapshot:

```
hdfs dfs –cp /entrada/.snapshot/snapshot_v1/quijote.txt /entrada/quijote.txt
```

10- Borra el snapshot y deshabilita los snapshots

```
hdfs dfs –deleteSnapshot /entrada snapshot_v1
hdfs dfsadmin –disallowSnapshot /entrada
```

11- Ejecuta la aplicación mapreduce sobre los directorios de entrada y salida

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-
examples-*.jar wordcount /entrada /salida
```

12- Verifica los ficheros generados en el directorio “salida”, comprueba su contenido.
13- Copia a la máquina virtual los ficheros de salida y comprueba su contenido.
14- Navega por la estructura de archivos /data/hdfs
    a. namenode: ficheros fsimage con datos de la estructura de directorios
    b. datanode: bloques de datos

```
for i in {1..100}; do
  wget -O "libro_$i.txt" "https://www.gutenberg.org/cache/epub/$i/pg$i.txt"
done
```