UNIVERSIDAD MAYOR DE SAN SIMÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA

DIRECCIÓN DE POSGRADO

DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES

SEGUNDA VERSIÓN

LABORATORIO HDFS

NOMBRE : HEBERT JUAN DE DIOS DELGADILLO FERNANDEZ
DOCENTE : DANNY LUIS HUANCA SEVILLA

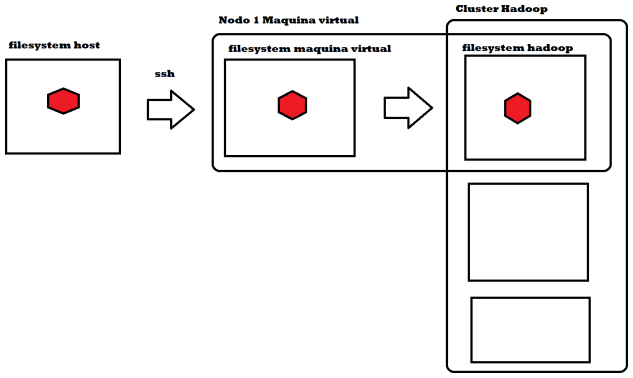
Cochabamba – Bolivia
2023

Laboratorio HDFS

Caso empresa Telco.

La empresa llego a un límite de capacidad en sus sistemas de datawarehouse y cree que es máseconómico realizar un backup de la información en su datalake.

Para tal efecto le piden realizar la operación descrita en el siguiente esquema:



El objetivo consiste en llevar los dos archivos churn_Operacion_1.csv y churn_Operacion_2.csval HDFS.

Para el efecto realizamos:

1. Primero ubicamos a la maquina virtual en la red

```
ens33: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.35  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 ::20c:29ff:fe6e:2e93  prefixlen 64  scopeid 0x0<global>
        inet6 fe80::20c:29ff:fe6e:2e93  prefixlen 64  scopeid 0x20<link>
        ether 00:0c:29:6e:2e:93  txqueuelen 1000  (Ethernet)
        RX packets 12396  bytes 18372036 (18.3 MB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 2841  bytes 209211 (209.2 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

   Nos conectaremos por medio de MobaXterm por el protocolo SSH a la IP 192.168.0.35.

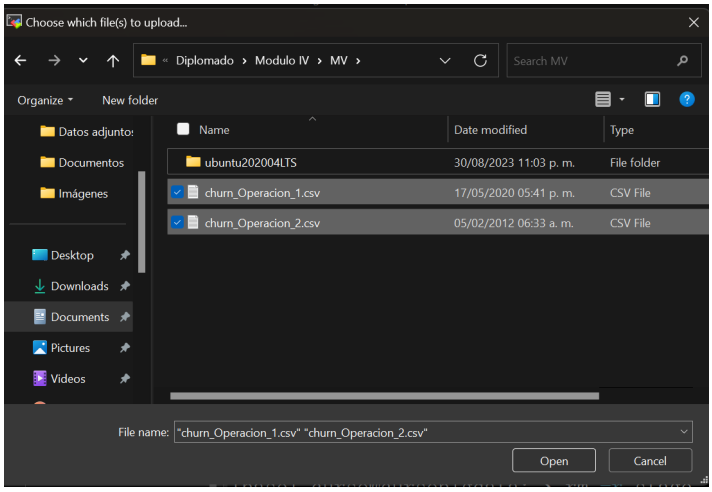
2. Creamos un directorio en el filesystem de la máquina virtual (en el home de curso) denominado **stage.**

   mkdir stage

```
(base) curso@cursobigdata:~$ mkdir stage
(base) curso@cursobigdata:~$ ls
Downloads          iniciarConfluent    iniciarLaboratorio
'Untitled Folder'  iniciarHadoop       pararLaboratorio
anaconda3          iniciarHive         spark-streaming-kafka-0-8-assembly_2.11-2.4.7.jar
datos              iniciarJupyter      stage
derby.log          iniciarKSQLserver
hiveserver2log     iniciarLabA
```

3. Llevar los archivos del caso telco desde el filesystem local de su computadora alfilesystem de la máquina virtual utilizando un cliente ssh (MobaXterm) mediante el protocolo sftp a la carpeta **stage**.

Primero seleccionamos los archivos a subir a la máquina virtual:



4. Verificar que los archivos se encuentren ahí, haciendo un ls al directorio stage.

```
(base) curso@cursobigdata:~$ cd stage
(base) curso@cursobigdata:~/stage$ ls
churn_Operacion_1.csv  churn_Operacion_2.csv
```

5. Iniciamos Hadoop

```
(base) curso@cursobigdata:~/stage$ iniciarHadoop
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-curso-namenode-cursobigdata.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-curso-datanode-cursobigdata.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-curso-secondarynamenode-c
ursobigdata.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-curso-resourcemanager-cursobigdata.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-curso-nodemanager-cursobigdata.
out
```

6. Verificamos los servicios de Hadoop

```
(base) curso@cursobigdata:~/stage$ jps
7361 NodeManager
7195 ResourceManager
6828 DataNode
7053 SecondaryNameNode
7678 Jps
```

7. Verificamos los directorios que tienen en el HDFS, mediante el comando HDFS

```
(base) curso@cursobigdata:~$ hdfs dfs -mkdir /stage
(base) curso@cursobigdata:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - curso supergroup          0 2023-08-31 19:00 /stage
```

8. Crear una carpeta en la raíz del HDFS que se denomine: STAGE/CASO_TELCO

```
(base) curso@cursobigdata:~$ hdfs dfs -mkdir /stage/CASO_TELCO
(base) curso@cursobigdata:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - curso supergroup          0 2023-08-31 19:02 /stage
(base) curso@cursobigdata:~$ hdfs dfs -ls /stage
Found 1 items
drwxr-xr-x   - curso supergroup          0 2023-08-31 19:02 /stage/CASO_TELCO
```

9. Cree una carpeta en la raíz del HDFS que se denomine: STAGE/CASO_TELCO2 en un solo paso con la opción –p de mkdir

```
(base) curso@cursobigdata:~$ hdfs dfs -mkdir -p /stage/CASO_TELCO2
(base) curso@cursobigdata:~$ hdfs dfs -ls /stage
Found 2 items
drwxr-xr-x   - curso supergroup          0 2023-08-31 19:02 /stage/CASO_TELCO
drwxr-xr-x   - curso supergroup          0 2023-08-31 19:03 /stage/CASO_TELCO2
```

10. Subir los archivos desde el directorio local stage al directorio del HDFS en el cluster /STAGE/CASO_TELCO

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -put * /stage/CASO_TELCO
(base) curso@cursobigdata:~/stage$ hdfs dfs -ls /stage/CASO_TELCO
Found 2 items
-rw-r--r--   1 curso supergroup     977501 2023-08-31 21:39 /stage/CASO_TELCO/churn_Operacion_1.csv
-rw-r--r--   1 curso supergroup     380191 2023-08-31 21:39 /stage/CASO_TELCO/churn_Operacion_2.csv
```

¿Qué es lo que sucedió?

Se copiaron los archivos que estaban en la maquina virtual, los que estaban en la carpeta stage a la que apuntábamos con la cli de Linux y se subieron todos los archivos que habían en esa carpeta a la carpeta CASO_TELCO.

11. Copiar los archivos en el HDFS desde el directorio CASO_TELCO al directorioCASO_TELCO2 que creamos antes y también verificamos que se copiaron los 2 archivos de la carpeta CASO_TELCO.

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -cp /stage/CASO_TELCO/* /stage/CASO_TELCO2
cp: `/stage/CASO_TELCO2/churn_Operacion_1.csv': File exists
cp: `/stage/CASO_TELCO2/churn_Operacion_2.csv': File exists
(base) curso@cursobigdata:~/stage$ hdfs dfs -ls /stage/CASO_TELCO2
Found 2 items
-rw-r--r--   1 curso supergroup     977501 2023-08-31 21:43 /stage/CASO_TELCO2/churn_Operacion_1.csv
-rw-r--r--   1 curso supergroup     380191 2023-08-31 21:43 /stage/CASO_TELCO2/churn_Operacion_2.csv
```

12. Borrar la carpeta CASO_TELCO2 con todos los archivos incluidos

```
(base) curso@cursobigdata:~/stage$ hdfs dfs -rm -r /stage/CASO_TELCO2
Deleted /stage/CASO_TELCO2
(base) curso@cursobigdata:~/stage$ hdfs dfs -ls /stage
Found 1 items
drwxr-xr-x   - curso supergroup          0 2023-08-31 21:39 /stage/CASO_TELCO
```

13. Pregunta reflexión. ¿Alguna vez interactuaron con algún filesystem en cluster?

    No, nunca interactúe con un sistema de archivos que maneje varias computadoras a la vez, sin embargo, es bastante similar a Linux, que es un sistema operativo que ya es bastante conocido debido a que en mi carrera se manejaba este sistema operativo. Además, es bastante interesante ver como funciona un clúster, la capacidad que puede almacenar y que también la escalabilidad es bastante sencilla, solo debe añadirse una computadora o servidor extra y conectarse al servidor hdfs, es muy bueno saber que podemos manejar grandes cantidades de datos con este tipo de tecnologías.

14. ¿Cómo creen que funciona Drive?

    Que por detrás existe un servidor gigante modular, no lo imagino como un cluster, sino como un servidor con hardware modular, es decir que basta con añadir mas tarjetas a ese sistema para tener mas memoria tanto de almacenamiento como de procesamiento y que un solo sistema operativo

lo administra a todo en general como 1 sola computadora. Sin embargo, viendo el tema actual, tomaría demasiado tiempo buscar en 1 sistema de archivos unificado como el anteriormente mencionado, por lo que un sistema de archivos distribuido es el sistema más probable para Google Drive.

Conclusiones

Aprendimos a usar el sistema de archivos HDFS sobre una maquina Linux, desde comandos para ver los archivos que existen, comandos para subir archivos a HDFS desde la máquina virtual, para realizar operaciones dentro de HDFS como copiar, mover y eliminar, también vimos como crear carpetas o directorios. Por último mencionar que es muy interesante ver que podemos obtener una cantidad de espacio enorme juntando varias maquinas como si fuera una computadora gigantesca y además tener los datos almacenados de manera confiable ya que en caso de perderse una computadora, los archivos están replicados en varios lugares, lo que hace a Hadoop una opción interesante de almacenamiento de BigData.