Actividad 1: HDFS, Spark SQL y MLlib

Parte 1

1) **Creamos el directorio raíz de HDFS en una carpeta llamada como el alumno**

```
root@ipmd-act-cluster-m:/# hdfs dfs -mkdir /Araceli_Ruiz_Vallecillo
root@ipmd-act-cluster-m:/# hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - root hadoop          0 2024-12-01 19:24 /Araceli_Ruiz_Vallecillo
drwxrwxrwt   - hdfs hadoop          0 2024-12-01 19:15 /tmp
drwxrwxrwt   - hdfs hadoop          0 2024-12-01 19:14 /user
drwxrwxrwt   - hdfs hadoop          0 2024-12-01 19:14 /var
```

2) **Subimos el fichero CSV a la carpeta Local Disk de JupyterLab**
   - Primero añadimos el CSV de forma manual al bucket de GCP.
   - Luego mediante el comando ***"gsutil cp"*** copiamos el CSV del bucket al entorno local de JupyterLab.

```
root@ipmd-act-cluster-m:/# gsutil cp gs://ipmd-act-bucket/flights.csv /flights.csv
Copying gs://ipmd-act-bucket/flights.csv...
/ [1 files][ 10.7 MiB/ 10.7 MiB]
Operation completed over 1 objects/10.7 MiB.
```

3) **Una vez subido el CSV al disco local de JupyterLab, subimos el mismo a la carpeta creada en HDFS**

```
root@ipmd-act-cluster-m:/# hdfs dfs -copyFromLocal /flights.csv /Araceli_Ruiz_Vallecillo
```

Comprobamos que se haya subido correctamente

```
root@ipmd-act-cluster-m:/# hdfs dfs -ls /Araceli_Ruiz_Vallecillo
Found 1 items
-rw-r--r--   2 root hadoop   11244080 2024-12-01 19:35 /Araceli_Ruiz_Vallecillo/flights.csv
```

4) **Ejecutamos el comando que nos da información sobre cómo está almacenado ese fichero en HDFS.**

```
root@ipmd-act-cluster-m:/# hdfs fsck /Araceli_Ruiz_Vallecillo/flights.csv -blocks -files
Connecting to namenode via http://ipmd-act-cluster-m:9870/fsck?ugi=root&blocks=1&files=1&path=%2FAraceli_Ruiz_Vallecillo%2Fflights.csv
FSCK started by root (auth:SIMPLE) from /10.132.0.3 for path /Araceli_Ruiz_Vallecillo/flights.csv at Sun Dec 01 19:44:10 UTC 2024

/Araceli_Ruiz_Vallecillo/flights.csv 11244080 bytes, replicated: replication=2, 1 block(s):  OK
0. BP-1176185793-10.132.0.3-1733080380698:blk_1073741825_1001 len=11244080 Live_repl=2


Status: HEALTHY
 Number of data-nodes:  2
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    11244080 B
 Total files:   1
 Total blocks (validated):      1 (avg. block size 11244080 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    2
 Average block replication:     2.0
 Missing blocks:                0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)

Erasure Coded Block Groups:
 Total size:    0 B
 Total files:   0
 Total block groups (validated):        0
 Minimally erasure-coded block groups:  0
 Over-erasure-coded block groups:       0
 Under-erasure-coded block groups:      0
 Unsatisfactory placement block groups: 0
 Average block group size:      0.0
 Missing block groups:          0
 Corrupt block groups:          0
 Missing internal blocks:       0
FSCK ended at Sun Dec 01 19:44:10 UTC 2024 in 13 milliseconds


The filesystem under path '/Araceli_Ruiz_Vallecillo/flights.csv' is HEALTHY
root@ipmd-act-cluster-m:/#
```