unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Máster Universitario en Análisis y Visualización de Datos Masivos

Actividad 1: Pipeline de procesamiento de datos con HDFS y Spark

Trabajo presentado por:	Henry Alejandro Gerena Ricardo
Asignatura:	Ingeniería para el procesamiento Masivo de Datos
Docente:	Pablo Villacorta Iglesias
Fecha:	07 diciembre 2025

- Creación de directorio dentro de HDFS, y comprobación:

```
neo@dataproc-unir-hagr27-m:~$ hdfs dfs -mkdir /alejandro_gerena
neo@dataproc-unir-hagr27-m:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - neo  hadoop          0 2025-11-29 13:35 /alejandro_gerena
drwxrwxrwt   - hdfs hadoop          0 2025-11-29 10:37 /tmp
drwxrwxrwt   - hdfs hadoop          0 2025-11-29 12:40 /user
drwxrwxrwt   - hdfs hadoop          0 2025-11-29 10:35 /var
```

- Copiar el fichero CSV de GCP a HDFS, y comprobación:

```
neo@dataproc-unir-hagr27-m:~$ hdfs dfs -cp gs://bucket-unir-hagr27/datos/flights.csv /alejandro_gerena/flights.csv
neo@dataproc-unir-hagr27-m:~$ hdfs dfs -ls -h /alejandro_gerena
Found 1 items
-rw-r--r--   2 neo hadoop      9.5 M 2025-11-29 13:36 /alejandro_gerena/flights.csv
```

- Ver metadatos del fichero en HDFS:

```
neo@dataproc-unir-hagr27-m:~$ hdfs dfs -ls -h /alejandro_gerena/flights.csv
hdfs fsck /alejandro_gerena/flights.csv -files -blocks -locations
-rw-r--r--   2 neo hadoop      9.5 M 2025-11-29 13:36 /alejandro_gerena/flights.csv
Connecting to namenode via http://dataproc-unir-hagr27-m.europe-west1-b.c.proyecto-master-unir.internal.:98
FSCK started by neo (auth:SIMPLE) from /10.132.0.25 for path /alejandro_gerena/flights.csv at Sat Nov 29 13

/alejandro_gerena/flights.csv 9947656 bytes, replicated: replication=2, 1 block(s):  OK
0. BP-493595081-10.132.0.25-1764412462015:blk_1073741830_1006 len=9947656 Live_repl=2  [DatanodeInfoWithSto
3:9866,DS-ee3b4739-ca8e-45cb-883c-7711916387fc,DISK]]


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

Replicated Blocks:
 Total size:    9947656 B
 Total files:   1
 Total blocks (validated):      1 (avg. block size 9947656 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 %)
 Blocks queued for replication: 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
 Blocks queued for replication: 0
FSCK ended at Sat Nov 29 13:39:48 UTC 2025 in 6 milliseconds


The filesystem under path '/alejandro_gerena/flights.csv' is HEALTHY
```