



Universidad Internacional de La Rioja

Facultad de Ingeniería y Tecnología

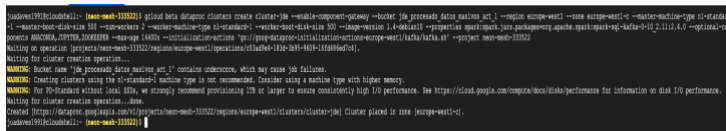
Máster Universitario en Análisis y Visualización
de Datos Masivos / Visual Analytics & Big Data

Actividad HDFS, Spark SQL y MLlib

Actividad de estudio presentado por:	Juan David Escobar Escobar
Tipo de trabajo:	Actividad
Modalidad:	Individual
Profesor/a:	Dr. Pablo J. Villacorta
Fecha:	Enero 2022

Comando creación de clúster de Spark mediante Google Shell:

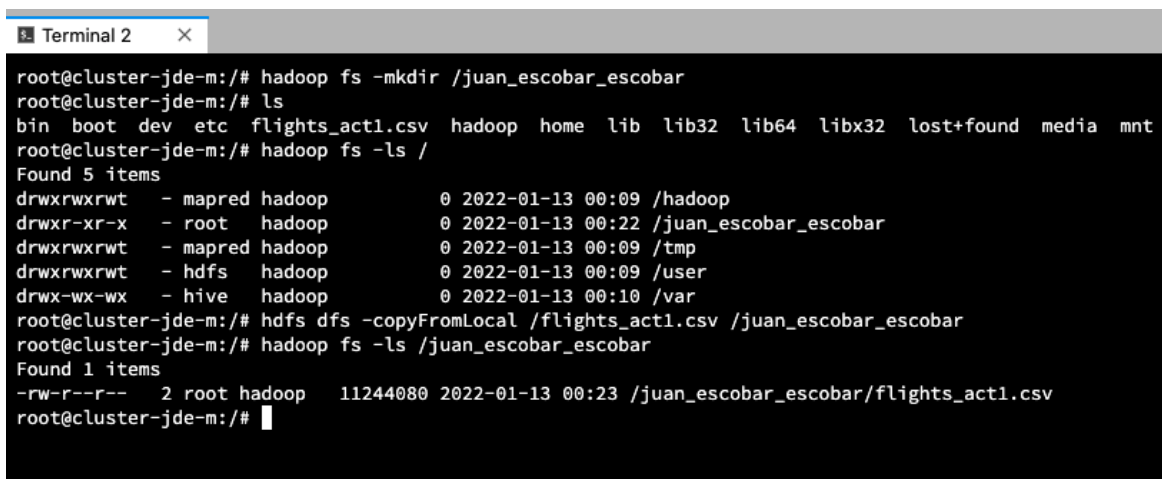
```
gcloud beta dataproc clusters create cluster-jde --enable-component-gateway --bucket
jde_procesado_datos_masivos_act_1 --region europe-west1 --zone europe-west1-c --
master-machine-type n1-standard-1 --master-boot-disk-size 500 --num-workers 2 --
worker-machine-type n1-standard-1 --worker-boot-disk-size 500 --image-version 1.4-
debian10 --properties spark:spark.jars.packages=org.apache.spark:spark-sql-kafka-0-
10_2.11:2.4.0 --optional-components ANACONDA,JUPYTER,ZOOKEEPER --max-age
14400s --initialization-actions 'gs://goog-dataproc-initialization-actions-europe-
west1/kafka/kafka.sh' --project neon-mesh-333522
```



```
cloudshell@cloudshell: ~$ gcloud beta dataproc clusters create cluster-jde --enable-component-gateway --bucket jde_procesado_datos_masivos_act_1 --region europe-west1 --zone europe-west1-c --master-machine-type n1-standard-1 --master-boot-disk-size 500 --num-workers 2 --worker-machine-type n1-standard-1 --worker-boot-disk-size 500 --image-version 1.4-debian10 --properties spark:spark.jars.packages=org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.0 --optional-components ANACONDA,JUPYTER,ZOOKEEPER --max-age 14400s --initialization-actions 'gs://goog-dataproc-initialization-actions-europe-west1/kafka/kafka.sh' --project neon-mesh-333522
Waiting for operation [projects/neon-mesh-333522/regions/europe-west1/operations/cluster-jde-2022-01-13-15480456] to complete...
WARNING: Recent some 'gs://goog-dataproc-initialization-actions-europe-west1/kafka/kafka.sh' contains undesired, which may cause job failures.
WARNING: Creating clusters using the n1-standard-1 machine type is not recommended. Consider using a machine type with higher memory.
WARNING: For n1-standard-1 machine type, we strongly recommend provisioning 128 or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/instance-templates for information on disk I/O performance.
Waiting for cluster creation operation... done
Created https://dataproc.googleapi.com/v1/projects/neon-mesh-333522/regions/europe-west1/clusters/cluster-jde Cluster placed in zone /europe-west1-c.
cloudshell@cloudshell: ~$
```

Comandos HDFS

- `hadoop fs -mkdir /juan_escobar_escobar`
- `hadoop fs -ls /`
- `hdfs dfs -copyFromLocal /flights_act1.csv /juan_escobar_escobar`
- `hadoop fs -ls /juan_escobar_escobar`



```
Terminal 2
root@cluster-jde-m:/# hadoop fs -mkdir /juan_escobar_escobar
root@cluster-jde-m:/# ls
bin boot dev etc flights_act1.csv hadoop home lib lib32 lib64 libx32 lost+found media mnt
root@cluster-jde-m:/# hadoop fs -ls /
Found 5 items
drwxrwxrwt - mapred hadoop 0 2022-01-13 00:09 /hadoop
drwxr-xr-x - root hadoop 0 2022-01-13 00:22 /juan_escobar_escobar
drwxrwxrwt - mapred hadoop 0 2022-01-13 00:09 /tmp
drwxrwxrwt - hdfs hadoop 0 2022-01-13 00:09 /user
drwx-wx-wx - hive hadoop 0 2022-01-13 00:10 /var
root@cluster-jde-m:/# hdfs dfs -copyFromLocal /flights_act1.csv /juan_escobar_escobar
root@cluster-jde-m:/# hadoop fs -ls /juan_escobar_escobar
Found 1 items
-rw-r--r-- 2 root hadoop 11244080 2022-01-13 00:23 /juan_escobar_escobar/flights_act1.csv
root@cluster-jde-m:/#
```

Se hace referencia a los recursos utilizados para el desarrollo de la actividad en el siguiente repositorio de [GitHub](#)

Herramientas:

- Jupyter Nootbooks Versión (v2.4.8),
- Python (Python 3.6.13 :: Anaconda, Inc)

Librerías y Módulos:

- Spark v2.4.8
- `pyspark.sql.functions`
- `pyspark.ml.feature`

- `pyspark.ml.classification`
- `pyspark.ml`.