Universidad Internacional de La Rioja

Facultad de Ingeniería y Tecnología

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

Actividad: Spark Streaming y Kafka

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

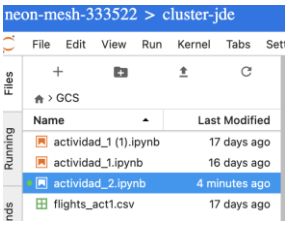
Comando creación de clúster de Spark mediante Google Shell:

```
gcloud projects list

gcloud config set project project neon-mesh-333522

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
```

Nombre	cluster-jde
UUID del clúster	efd912a7-04be-4d53-b08d-ebba3ecc883b
Tipo	Clúster de Dataproc
Estado	En ejecución

```
juadaves1991@cloudshell:~ (neon-mesh-333522)$ clear
juadaves1991@cloudshell:~ (neon-mesh-333522)$ gcloud beta dataproc clusters create cluster-jde --enable-component-gateway --bucket jde_procesado_datos_masivos_act_1 --reg
ion 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-d
isk-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,Z
OOKEEPER --max-age 14400s --initialization-actions 'gs://goog-dataproc-initialization-actions-europe-west1/kafka/kafka.sh' --project neon-mesh-333522
```

Upload Notebook en GCS

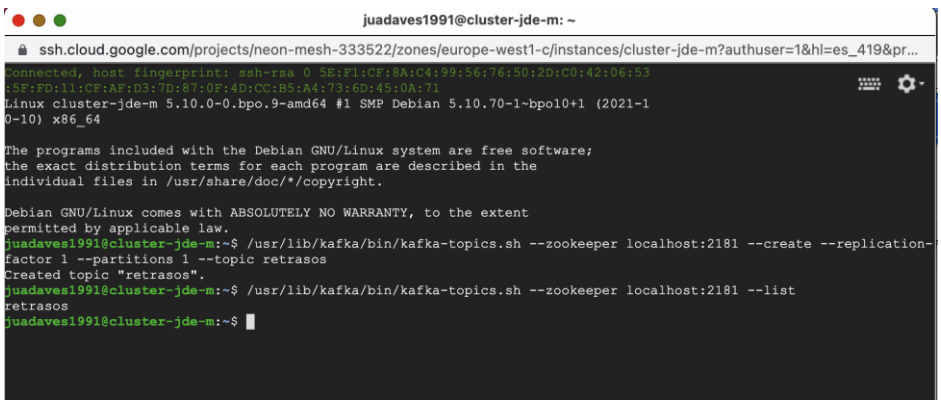


Creación de Topic “retrasos”

```
/usr/lib/kafka/bin/kafka-topics.sh --zookeeper localhost:2181 --create --
      replication-factor 1 --partitions 1 --topic retrasos
```

Creación de Topic “retrasos”

```
/usr/lib/kafka/bin/kafka-topics.sh --zookeeper localhost:2181 –list
```



Ejecución de Topic “retrasos”

```
/usr/lib/kafka/bin/kafka-console-producer.sh  --broker-list  cluster-jde-w-0:9092
--topic retrasos
```

En caso de error o no ejecución de Kafka validar:

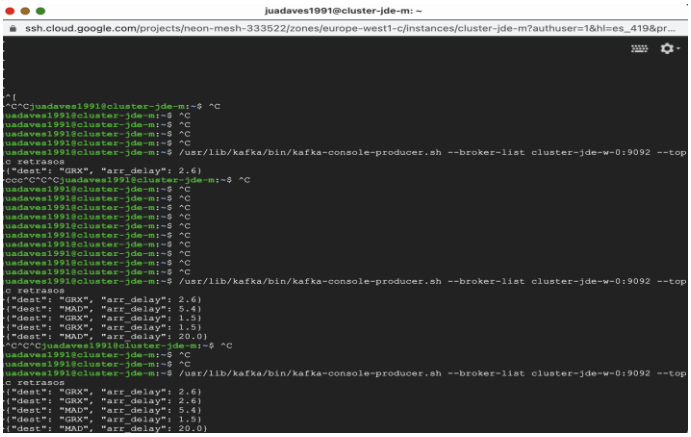
Kill process
```
sudo fuser -k 2181/tcp
```

run zookeeper
```
cd "/usr/lib/kafka/bin/"
./kafka-server-start.sh /usr/local/etc/kafka/server.properties

bin/zookeeper-server-start.sh config/zookeeper.properties
```

Run Kafka
```
bin/kafka-server-start.sh config/server.properties  '
```

Ejecución y resultado de mensajes sucesivamente

{"dest": "GRX", "arr_delay": 2.6} {"dest": "MAD", "arr_delay": 5.4} {"dest": "GRX", "arr_delay": 1.5} {"dest": "MAD", "arr_delay": 20.0}



Resultado 1

```
In [80]: agregadosDF.show()  # Ejecuta varias veces esta celda tras enviar el primer mensaje, hasta ver que el DataFrame no es vac
         retraso_medio_GRX_primer_mensaje = agregadosDF.where(F.col("dest") == "GRX")
         retraso_medio_GRX_primer_mensaje.show()
```

```
+----+-------------+
|dest|retraso_medio|
+----+-------------+
| GRX|          2.6|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| GRX|          2.6|
+----+-------------+
```

Resultado 2

```
In [82]: # Ejecuta varias veces esta celda tras enviar el segundo mensaje, hasta ver que el DataFrame ha cambiado
agregadosDF.show()
retraso_medio_GRX_primer_mensaje = agregadosDF.where(F.col("dest") == "GRX")
retraso_medio_MAD_segundo_mensaje = agregadosDF.where(F.col("dest") == "MAD")

retraso_medio_GRX_primer_mensaje.show()
retraso_medio_MAD_segundo_mensaje.show()
```

```
+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|          5.4|
| GRX|          2.6|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| GRX|          2.6|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|          5.4|
+----+-------------+
```

Resultado 3

```
In [85]: # Ejecuta varias veces esta celda tras enviar el tercer mensaje, hasta ver que el DataFrame ha cambiado
agregadosDF.show()
retraso_medio_GRX_tercer_mensaje = agregadosDF.where(F.col("dest") == "GRX")
retraso_medio_MAD_tercer_mensaje = agregadosDF.where(F.col("dest") == "MAD")

retraso_medio_GRX_tercer_mensaje.show()
retraso_medio_MAD_tercer_mensaje.show()
```

```
+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|          5.4|
| GRX|         2.05|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| GRX|         2.05|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|          5.4|
+----+-------------+
```

Resultado 4

```
In [87]: # Ejecuta varias veces esta celda tras enviar el cuarto mensaje, hasta ver que
agregadosDF.show()
retraso_medio_GRX_cuarto_mensaje = agregadosDF.where(F.col("dest") == "GRX")
retraso_medio_MAD_cuarto_mensaje = agregadosDF.where(F.col("dest") == "MAD")

retraso_medio_GRX_cuarto_mensaje.show()
retraso_medio_MAD_cuarto_mensaje.show()
```

```
+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|         12.7|
| GRX|         2.05|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| GRX|         2.05|
+----+-------------+

+----+-------------+
|dest|retraso_medio|
+----+-------------+
| MAD|         12.7|
+----+-------------+
```