V Big Data and Machine Learning Bootcamp

Data Processing

PRACTICA

1. Hemos de crear una estructura de Kafka-Streaming PRODUCTOR - CONSUMIDOR de dos formas.

   1.1 Usando los métodos de Kafka propios para ello.

   En este caso, comprobaremos que si enviamos un fichero .json dado (personal.json) por el productor, el contenido de este fichero nos aparece en el consumidor.

   1.2 Creando el código en Scala para montar un productor y un consumidor.

   En este caso, ademas de comprobar el envío - recepción de este fichero, deberá contar el número de palabras del fichero y deberá filtrar (que no aparezcan) dos palabras a elegir. El resultado del .json filtrado se mostrará por consola.

2. Instalar Zepellin y configurarlo con Spark.

   2.1 Mostrar los pasos necesarios para la instalación.

   2.2 Calcular el número de registros que tenemos en el fichero amigos.csv

SOLUCION

Para la resolución de la práctica, he instalado en local en mi ordenador MAC todas las herramientas necesarias, por lo que los comandos no coinciden exactamente con los vistos en clase, pero son los necesarios para la realización de la práctica.

Las herramientas instaladas son:

- Zookeeper 3.4.14
- Kafka 2.4.0
- Spark 2.4.4
- Scala
- IntelliJ IDEA 2019.3.2

Punto 1.1

- Nos desplazamos al directorio donde se encuentran los ejecutables de Kafka

  ```
  $ cd /usr/local/Cellar/kafka/2.4.0/bin
  ```

- Lanzamos el servidor de zookeeper

  ```
  $ sudo zookeeper-server-start /usr/local/etc/kafka/zookeeper.properties
  ```

- A continuación, lanzamos el servidor de kafka

  ```
  $ sudo kafka-server-start /usr/local/etc/kafka/server.properties
  ```

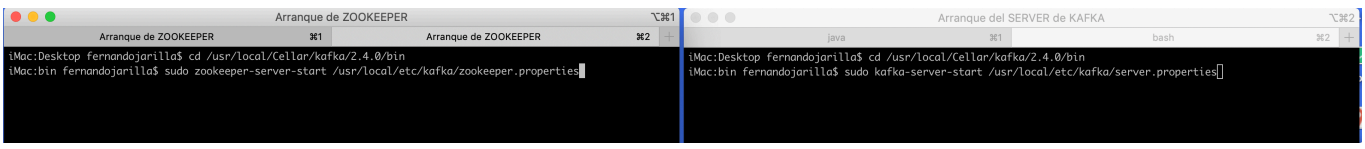

Figura 1. Comandos antes de ejecutarse*

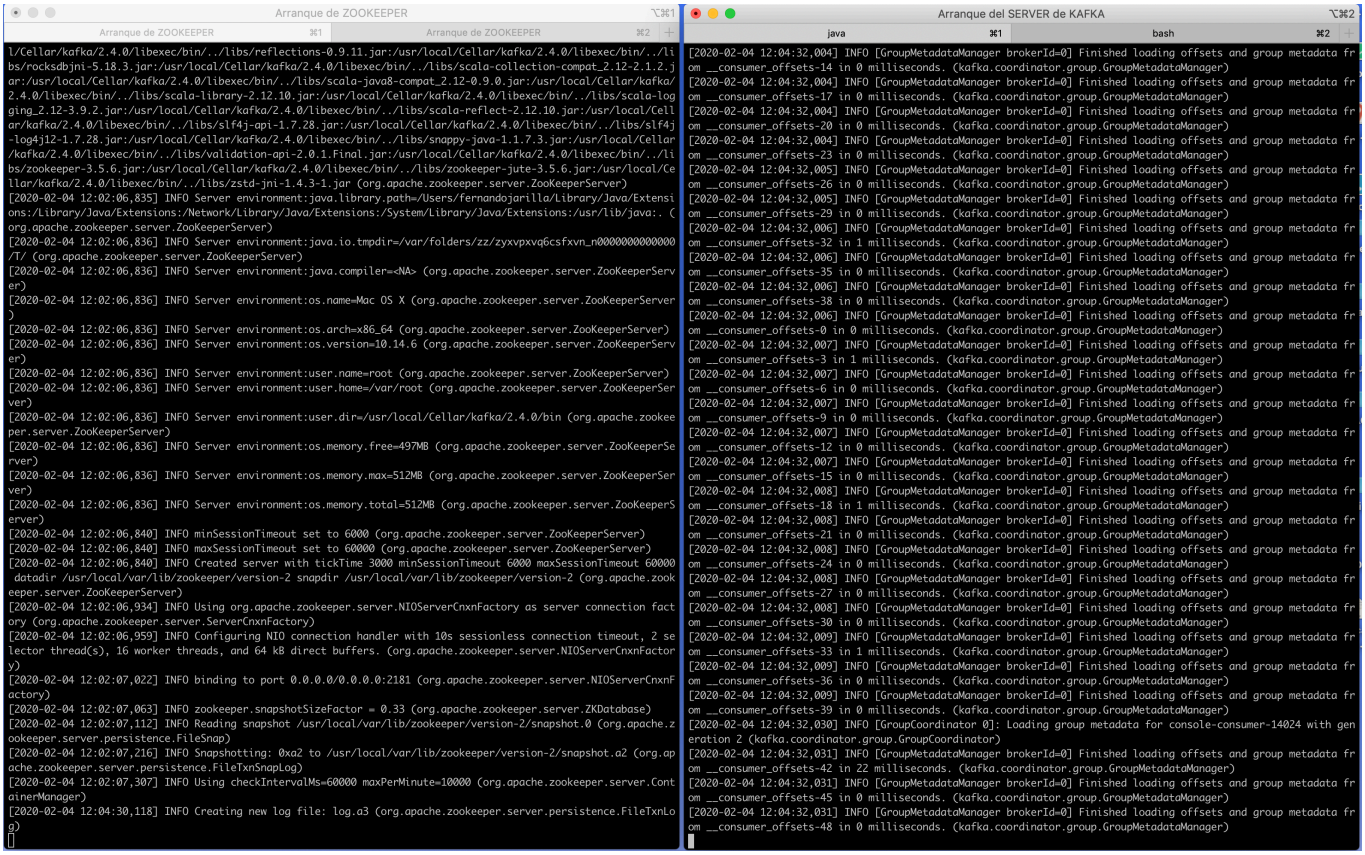

Figura 2. Comandos de arranque se servers ejecutados

- Una vez lanzados los servers, crearé un tópico que llamaré **practica**

```
$ sudo kafka-topics --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic practica
```

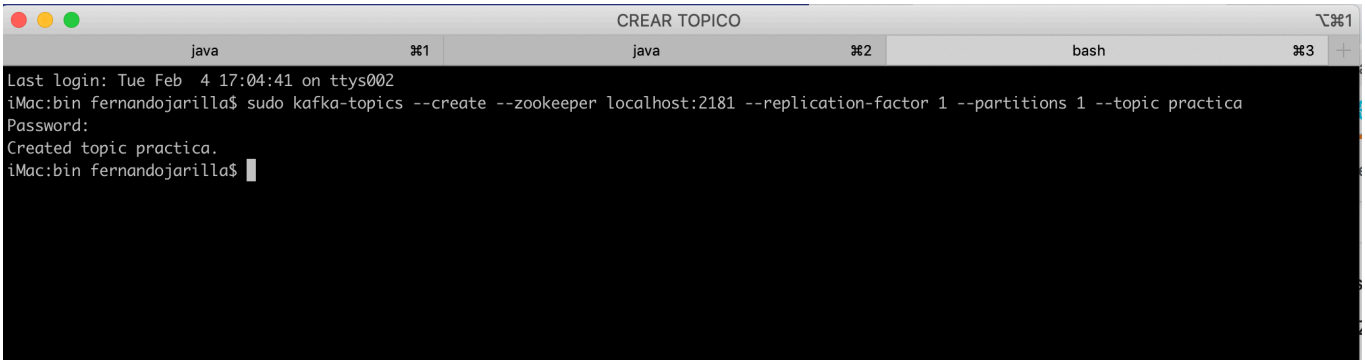

Figura 3. Crear tópico

- A continuación, crearé el **PRODUCTOR**

```
$ sudo kafka-console-producer --broker-list localhost:9092 --topic practica
```

- Y finalmente el **CONSUMIDOR**

```
$ sudo kafka-console-consumer --bootstrap-server localhost:9092 --topic practica --from-beginning
```

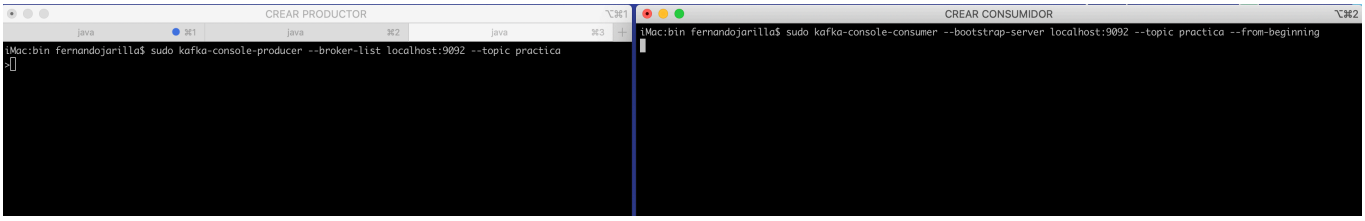

Figura 4. Lanzados productor y consumidor

- Por fin, lo que escriba en la ventana del productor, aparecerá en la ventana del consumidor. La información se envia tras pulsar *Return* en el productor

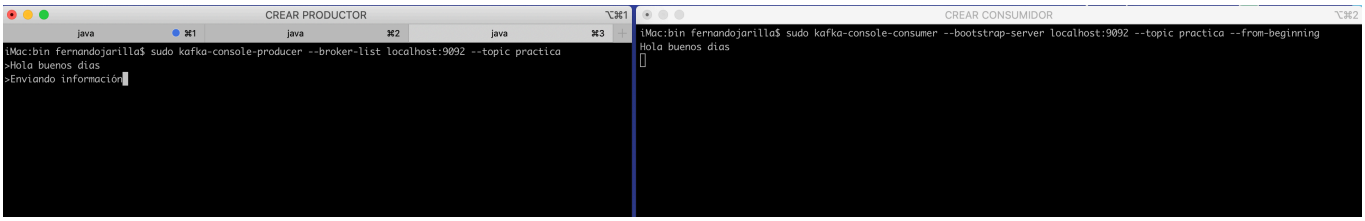

Figura 5. Primeros datos enviados y segunda linea antes de enviarse

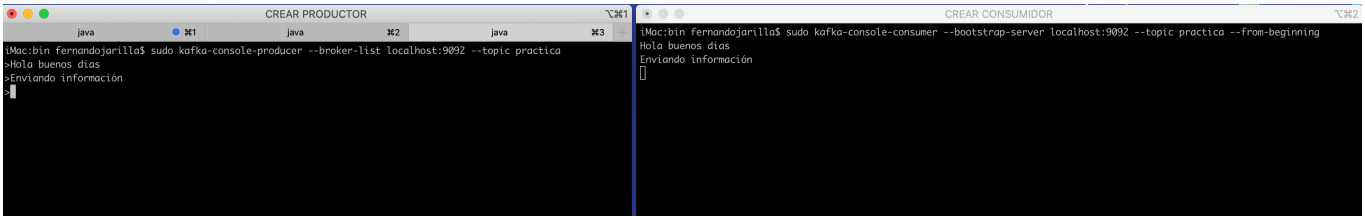

Figura 6. Segunda línea enviada

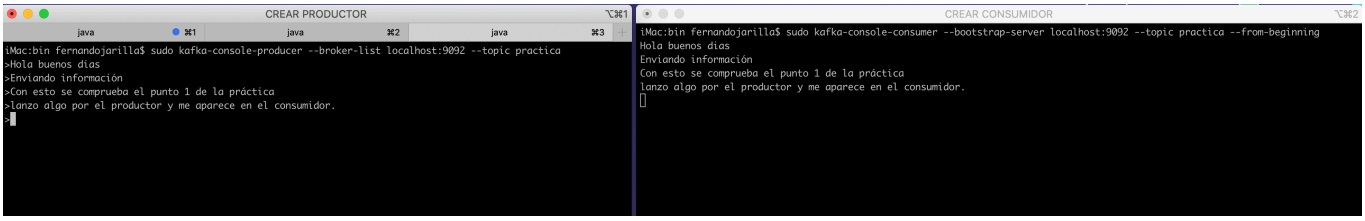

Figura 7. Envío de algo mas de información

- Finalmente, para enviar el fichero *.json* salgo del productor (con *Ctrl+C*) y lanzo el siguiente comando.

```
$ cat /users/fernandojarilla/Desktop/personal.json | sudo kafka-console-producer --broker-list localhost:9092 --topic practica
```

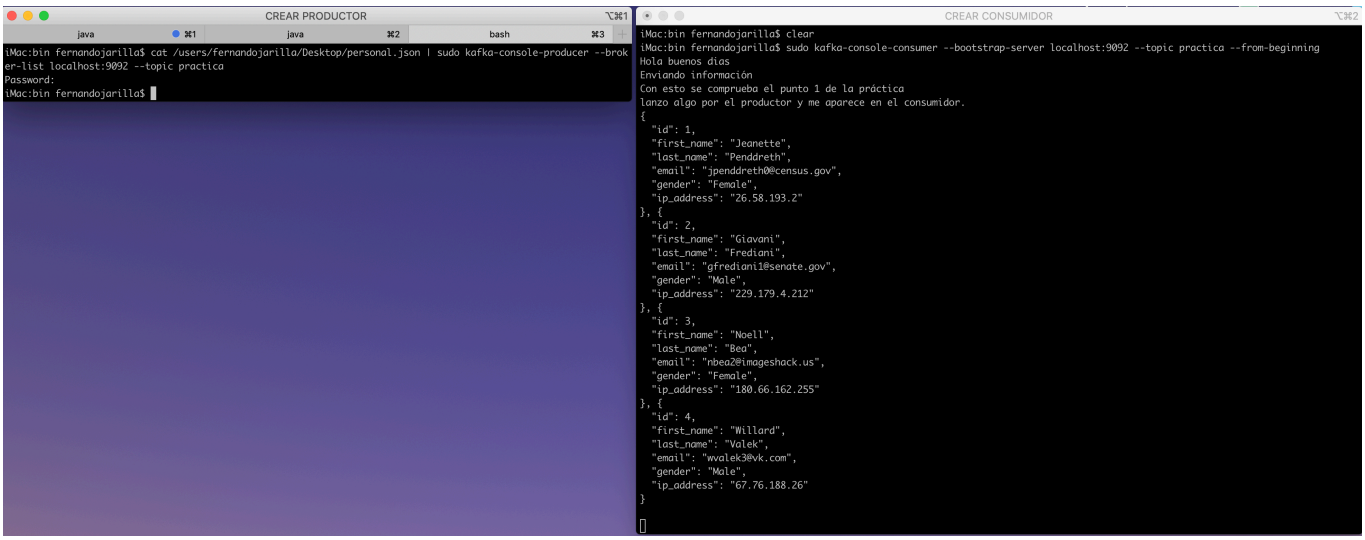

Figura 8. Envio y visualizacíon de la recepción del fichero personal.json

Con esto queda terminada la primera parte de la práctica.