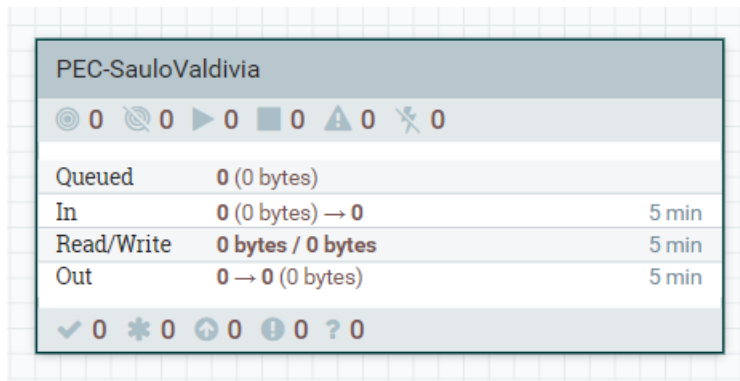
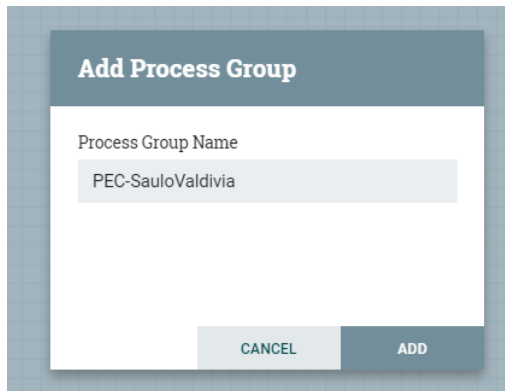
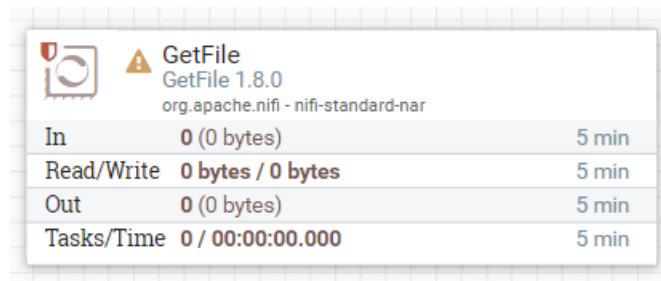


1. Ejercicio NiFi

- a. Creación de un grupo (Process Group) dentro de NiFi con el nombre PEC-<NombreAlumno>



- b. Crear un flujo en NiFi que coja solo los ficheros de la carpeta tweets (del zip proporcionado) y **no** de sus subcarpetas y los deje en el directorio tweets/ejercicio
- i. Creación de un proceso GetFile para recuperar los tweets



ii. Configuración de proceso GetFile

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field


+

Property	Value
Input Directory	<div>?</div> /home/ubuntu/bigdata/practica/FicherosPractica/tweets
File Filter	<div>?</div> [^\\].*.json
Path Filter	<div>?</div> No value set
Batch Size	<div>?</div> 10
Keep Source File	<div>?</div> false
Recurse Subdirectories	<div>?</div> false
Polling Interval	<div>?</div> 0 sec
Ignore Hidden Files	<div>?</div> true
Minimum File Age	<div>?</div> 0 sec
Maximum File Age	<div>?</div> No value set
Minimum File Size	<div>?</div> 0 B
Maximum File Size	<div>?</div> No value set

CANCEL

APPLY

iii. Creación de proceso PutFile

	<div><div>PutFile</div><div>PutFile 1.8.0</div><div>org.apache.nifi - nifi-standard-nar</div></div>	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

iv. Configuración de proceso PutFile

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

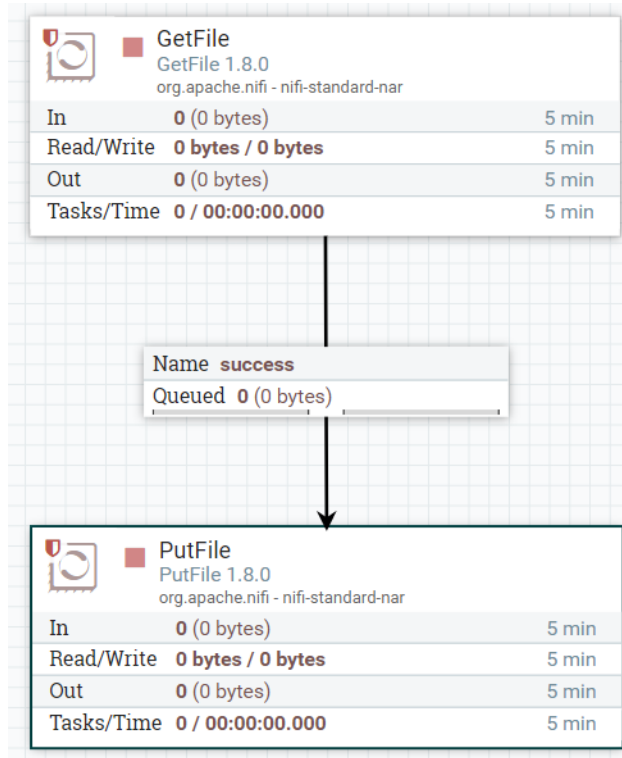
+

Property	Value
Directory	<div>?</div> /home/ubuntu/bigdata/practica/FicherosPractica/tweets/e...
Conflict Resolution Strategy	<div>?</div> replace
Create Missing Directories	<div>?</div> true
Maximum File Count	<div>?</div> No value set
Last Modified Time	<div>?</div> No value set
Permissions	<div>?</div> No value set
Owner	<div>?</div> No value set
Group	<div>?</div> No value set

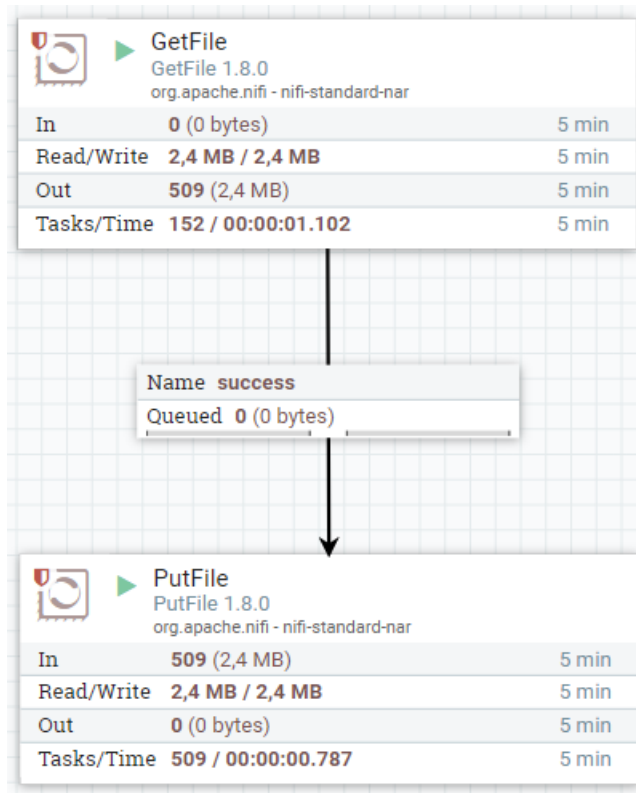
CANCEL

APPLY

v. Flujo completo



vi. Resultados





vii. Resultados en Filezilla

Remote site: /home/ubuntu/bigdata/practica/FicherosPractica/tweets/ejercicio					
<ul style="list-style-type: none"> namenode practica <ul style="list-style-type: none"> FicherosPractica <ul style="list-style-type: none"> tweets <ul style="list-style-type: none"> ejercicio ejercicio1 english spanish 					
Filename ^	Filesize	Filetype	Last modifi...	Permissi...	Owner/Gr...
..					
00a0e5c9-749f-4...	5,520	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
00bf1029-cbbd-...	9,295	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
01b1a4bc-3595-...	3,614	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
025e66c5-4824-...	9,173	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
029c7146-c227-...	3,055	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
02d3dd82-bae8-...	160	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
03ad5a01-ccdb-...	2,560	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
040c19ea-0fac-4...	7,270	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
04162878-0df1-...	7,589	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
053d82ea-6411-...	5,287	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
05e12814-ca9d-...	5,351	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
05ff063a-6d8c-4...	5,520	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
0625317d-bb83-...	5,438	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
065c6fa7-b42f-4...	2,333	JSON File	4/25/2021 ...	-rw-rw-r--	ubuntu u...
509 files. Total size: 2,512,261 bytes					

- c. Crear un nuevo flujo que coja los tweets de 5 en 5 de la carpeta tweets/english cada 10 segundos y los deje de nuevo en el directorio tweets/ejercicio

- i. Creación de proceso GetFile

	 GetFile GetFile 1.8.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

- ii. Configuración de proceso GetFile

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Name

GetFile

☒ Enabled

Automatically Terminate Relationships ?
☐ success
All files are routed to success

Id

0924759c-0179-1000-f244-e9df17e51316

Type

GetFile 1.8.0

Bundle

org.apache.nifi - nifi-standard-nar

Penalty Duration ?

30 sec

Yield Duration ?

10 sec

Bulletin Level ?

WARN

▼

CANCEL

APPLY

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field





Property		Value
Input Directory		/home/ubuntu/bigdata/practica/FicherosPractica/tweets/e...
File Filter		[^\\].*.json
Path Filter		No value set
Batch Size		5
Keep Source File		false
Recurse Subdirectories		true
Polling Interval		0 sec
Ignore Hidden Files		true
Minimum File Age		0 sec
Maximum File Age		No value set
Minimum File Size		0 B
Maximum File Size		No value set

CANCEL

APPLY

iii. Creación de proceso PutFile

	 PutFile PutFile 1.8.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

iv. Configuración de proceso PutFile

Configure Processor

SETTINGS

SCHEDULING

PROPERTIES

COMMENTS

Required field

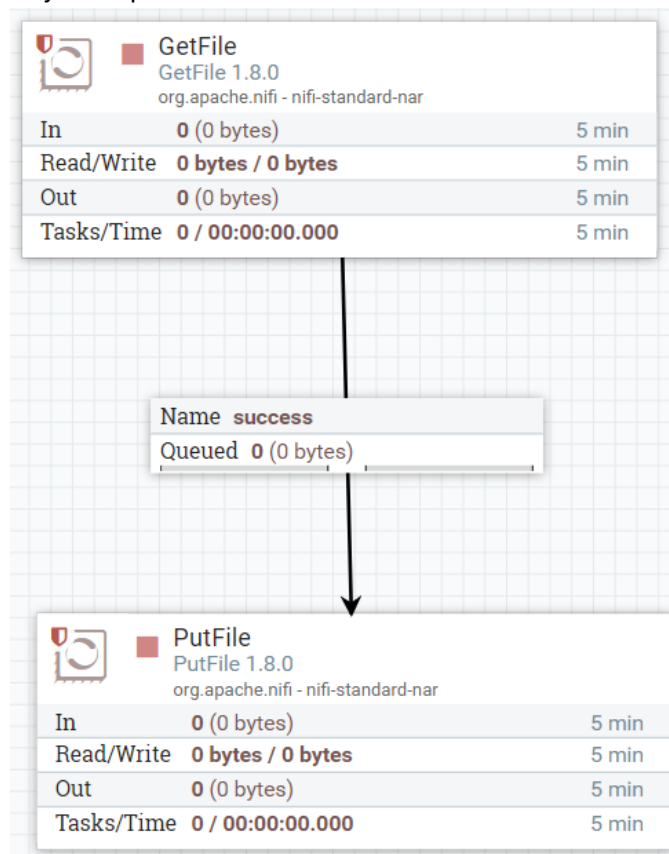
+

Property		Value
Directory	?	/home/ubuntu/bigdata/practica/FicherosPractica/tweets/e...
Conflict Resolution Strategy	?	replace
Create Missing Directories	?	true
Maximum File Count	?	No value set
Last Modified Time	?	No value set
Permissions	?	No value set
Owner	?	No value set
Group	?	No value set

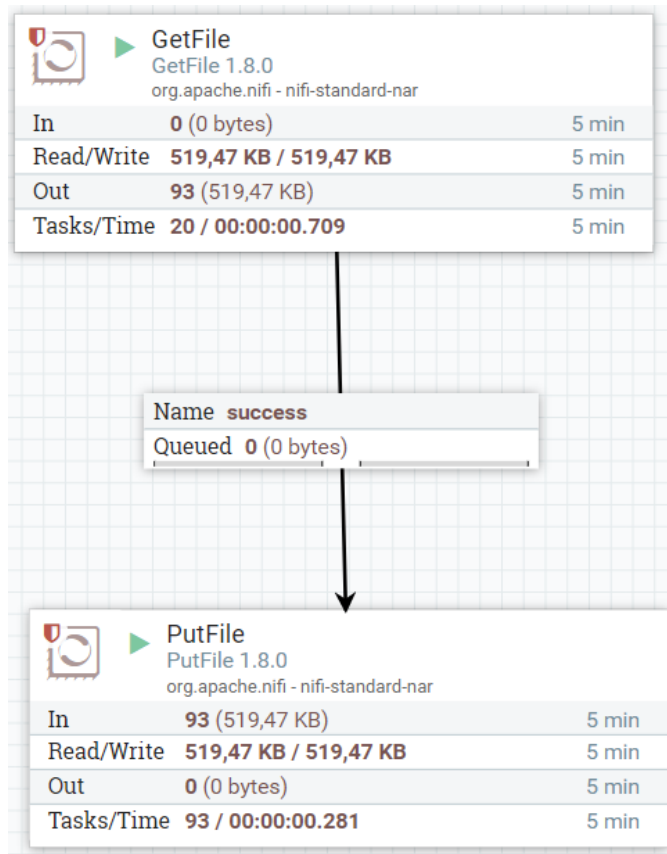
CANCEL

APPLY

v. Flujo Completo



vi. Resultados



vii. Resultados Filezilla

Remote site: /home/ubuntu/bigdata/practica/FicherosPractica/tweets/ejercicio1

Filename	Filesize	Filetype	Last modifi...	Permissi...	Owner/Gr...
..					
0015ead6-76d5-...	17,856	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
029796c9-73da-...	2,059	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
07c655e9-0f3e-4...	7,002	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
0809a32a-b575-...	2,091	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
0ac579ae-e6eb-...	7,087	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
0c55db63-5ee1-...	3,302	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
0cd4e309-1b47-...	2,105	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
0db07a8b-4596-...	7,455	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...
24170180-2d95-...	3,354	JSON File	4/30/2021 ...	-rw-rw-r--	ubuntu u...

93 files. Total size: 531,942 bytes

2. Ejercicio SparkSQL

- a. Cargar los datos del fichero dataset_coches.csv en un dataframe y mostrar las primeras 5 líneas del dataframe.

```
In [1]: ruta = "file:///home/ubuntu/bigdata/examples/spark/sparkSQL/data"
```

```
datosCoches = spark.read.format('csv') \
    .option('header','true') \
    .option('inferSchema','true') \
    .option('delimiter',';') \
    .load(ruta + "/dataset_coches.csv")
```

```
In [2]: datosCoches.show(5)
```

Car	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model	Origin
Chevrolet Chevell...	18.0	8	307.0	130.0	3504	12.0	70	US
Buick Skylark 320	15.0	8	350.0	165.0	3693	11.5	70	US
Plymouth Satellite	18.0	8	318.0	150.0	3436	11.0	70	US
AMC Rebel SST	16.0	8	304.0	150.0	3433	12.0	70	US
Ford Torino	17.0	8	302.0	140.0	3449	10.5	70	US

only showing top 5 rows

- b. Usando tanto SQL como el API de dataframes de Spark obtener las columnas “Car” y “Cylinders” de todos los que sean de Europa

i. SparkSQL Dataframe API

```
In [6]: from pyspark.sql.functions import count, desc, avg
```

```
europaCars = datosCoches.where("Origin == 'Europe'")
selectEuropaCars = europaCars.select(europaCars['Car'].alias('Car'), europaCars['Cylinders'].alias('Cylinders'))
selectEuropaCars.show()
```

Car	Cylinders
Citroen DS-21 Pallas	4
Volkswagen 1131 D...	4
Peugeot 504	4
Audi 100 LS	4
Saab 99e	4
BMW 2002	4
Volkswagen Super ...	4
Opel 1900	4
Peugeot 304	4
Fiat 124B	4
Volkswagen Model 111	4
Volkswagen Type 3	4
Volvo 145e (sw)	4
Volkswagen 411 (sw)	4
Peugeot 504 (sw)	4
Renault 12 (sw)	4
Volkswagen Super ...	4
Fiat 124 Sport Coupe	4
Fiat 128	4
Opel Manta	4

only showing top 20 rows

ii. SQL

```
In [13]: ▶ datosCoches.createOrReplaceTempView("carsSQL")
```

```
In [14]: ▶ rset = spark.sql("SELECT Car, Cylinders FROM carsSQL where Origin == 'Europe'")
rset.show()
```

Car	Cylinders
Citroen DS-21 Pallas	4
Volkswagen 1131 D...	4
Peugeot 504	4
Audi 100 LS	4
Saab 99e	4
BMW 2002	4
Volkswagen Super ...	4
Opel 1900	4
Peugeot 304	4
Fiat 124B	4
Volkswagen Model 111	4
Volkswagen Type 3	4
Volvo 145e (sw)	4
Volkswagen 411 (sw)	4
Peugeot 504 (sw)	4
Renault 12 (sw)	4
Volkswagen Super ...	4
Fiat 124 Sport Coupe	4
Fiat 128	4
Opel Manta	4

only showing top 20 rows

- c. Usando tanto SQL como el API de dataframes de Spark obtener el número de vehículos por “Origin”

i. SparkSQL Dataframe API

```
In [16]: ▶ datos2 = datosCoches.groupBy(datosCoches['Origin']) \
        .agg(count(datosCoches['Origin']).alias('Total'))
datos2.show(5)
```

Origin	Total
Europe	73
US	254
Japan	79

ii. SQL

```
In [17]: rset = spark.sql("SELECT Origin, count(*) as Total FROM carsSQL group by Origin")
rset.show()
```

```
+-----+-----+
|Origin|Total|
+-----+-----+
|Europe|   73|
|   US|  254|
| Japan|   79|
+-----+-----+
```

3. Ejercicio Kafka

a. Creación de topics

- Crear un nuevo topic denominado pec-topic1-<NombreAlumno> con dos particiones y con factor de replicación 1

```
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic pec-topic1-Saulo --create --partitions 2 --replication-factor 1
Created topic "pec-topic1-Saulo".
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --list
__confluent.support.metrics
__consumer_offsets
pec-topic1-Saulo
```

```
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic pec-topic1-Saulo --describe
Topic:pec-topic1-Saulo PartitionCount:2 ReplicationFactor:1 Configs:
Topic: pec-topic1-Saulo Partition: 0 Leader: 1 Replicas: 1 Isr: 1
Topic: pec-topic1-Saulo Partition: 1 Leader: 2 Replicas: 2 Isr: 2
```

- ¿Sería posible crearlo en el entorno que tienes con un factor de replicación de 2? ¿Por qué?

Nuestra arquitectura contiene 3 Brokers/Nodos y podría soportar un factor de replicación de 2.

```
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic pec-topic1-Saulo2 --create --partitions 2 --replication-factor 2
Created topic "pec-topic1-Saulo2".
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-topics.sh --zookeeper 127.0.0.1:2181 --topic pec-topic1-Saulo2 --describe
Topic:pec-topic1-Saulo2 PartitionCount:2 ReplicationFactor:2 Configs:
Topic: pec-topic1-Saulo2 Partition: 0 Leader: 2 Replicas: 2,3 Isr: 2,3
Topic: pec-topic1-Saulo2 Partition: 1 Leader: 3 Replicas: 3,1 Isr: 3,1
```

- Creación de productores

 - Crear un productor que empiece a pasar información desde la línea de comandos al topic creado anteriormente y simular el envío de información

```
(base) ubuntu@master-1:~/Docker/kafka-docker-compose$ kafka-console-producer.sh --broker-list 127.0.0.1:9092 --topic pec-topic1-Saulo
>This is a test
>tes1
>test2
>bye
```

c. Creación de consumidores

- Crear un consumidor que sea capaz de leer *toda* la información que contenga el topic creado

```
(base) ubuntu@master-1:~$ kafka-console-consumer.sh --bootstrap-server 127.0.0.1:9092 --topic pec-topic1-Saulo
This is a test
tes1
test2
bye
^CProcessed a total of 4 messages
```

4. Ejercicio Spark Streaming

- a. Saber en tiempo real cuantas veces pasa un vehículo por cada punto del sistema desde el día que se pone en marcha el mecanismo.
 - i. Iniciar el Producer

```
(base) ubuntu@master-1:~/bigdata/examples/spark/sparkStreaming$ python vehiculos_producer.py --port 9997 --interval 5
Listening at ('localhost', 9997)
```

- ii. Script para conectarse al stream y procesar las lineas

```
In [4]: ▶ def update_func(new_val, last_sum):
        return sum(new_val) + (last_sum or 0)
```

Es necesario tener HDFS arrancado, guarda información (sólo si no ponemos ruta local)

```
In [5]: ▶ # To maintain status info we need to use checkpoints. This way, if our
        # application fails it will be able to resume operations correctly
        # later on, using the updated information
        checkpointDir = "file:///home/ubuntu/bigdata/apps/spark/checkpoint"
        #checkpointDir = "/home/ubuntu/bigdata/apps/spark/checkpoint"
        ssc.checkpoint(checkpointDir) # create dir for checkpoint archives

        lines = ssc.socketTextStream("localhost", 9997)
        counts = lines.map(lambda word: (word, 1))\
            .updateStateByKey(update_func)

        counts.pprint()

        #ssc.awaitTermination()
```

- iii. Resultados

```
In [6]: ▶ ssc.start()
```

```
-----
Time: 2021-04-29 18:45:20
Matriculas
-----
Puntos de Control (1234AAA A7-KM-50, 2)
                  (1234AAA A7-KM-01, 3)
                  (1234EEE A7-KM-50, 1)
                  (1234AAA A7-KM-30, 2)
                  (1234UUU A7-KM-15, 1)
                  (1234UUU A7-KM-30, 1)
                  (1234EEE A7-KM-30, 1)
-----
Time: 2021-04-29 18:45:30
-----
('1234AAA A7-KM-50', 2)
('1234AAA A7-KM-01', 3)
('1234EEE A7-KM-50', 1)
('1234AAA A7-KM-30', 2)
('1234UUU A7-KM-15', 2)
```

- b. Saber en tiempo real cuántas veces pasa un vehículo por un determinado punto, pero solo teniendo en cuenta los últimos 7 días de información.

- i. Script para conectarse y procesar las líneas.

```
In [ ]: ▶ sc.stop()
sc = SparkContext(appName="NetWordCountStateWindow")
ssc = StreamingContext(sc, 100)
```

Es necesario tener HDFS arrancado, guarda información (sólo si no ponemos ruta local)

```
In [ ]: ▶ # To maintain status info we need to use checkpoints. This way, if our
# application fails it will be able to resume operations correctly
# later on, using the updated information
checkpointDir = "file:///home/ubuntu/bigdata/apps/spark/checkpoint"
ssc.checkpoint(checkpointDir) # create dir for checkpoint archives
#ssc.checkpoint("/tmp/checkpoint") # Necesario HDFS

lines = ssc.socketTextStream("localhost", 9997)
counts = lines.map(lambda word: (word, 1))\
    .reduceByKeyAndWindow(lambda x,y: x+y, lambda x, y: x-y, 604800, 600)

counts.pprint()
```

- ii. Resultados

Los resultados son similares al apartado anterior. Sin embargo, esperaríamos que transcurrido el tiempo de la ventana temporal, estos datos se actualicen.

```
In [6]: ▶ ssc.start()
```

```
-----
Time: 2021-04-29 18:45:20
Matriculas
-----
Puntos de Control (1234AAA A7-KM-50, 2)
(1234AAA A7-KM-01, 3)
(1234EEE A7-KM-50, 1)
(1234AAA A7-KM-30, 2)
(1234UUU A7-KM-15, 1)
(1234UUU A7-KM-30, 1)
(1234EEE A7-KM-30, 1)

-----
Time: 2021-04-29 18:45:30
-----
('1234AAA A7-KM-50', 2)
('1234AAA A7-KM-01', 3)
('1234EEE A7-KM-50', 1)
('1234AAA A7-KM-30', 2)
('1234UUU A7-KM-15', 2)
```