Lab Assignment 3

Authors: Andrei Capastru, Luca Vavassori

I. Code explanation

The file `KafkaSpark.scala` consists of an object `KafkaSpark` that includes the only method main.

In the main, as first step we create a cluster and link it to Cassandra. Then we create a `keyspace` and in that `keyspace` we create a table. The second step is to create a `Streaming Context` though Spark. We also set the checkpoint folder because it is a best practice and it was required to avoid an error at compile time. After that, we create a `Map` with the configuration for the `Kafka Stream` and a `Set` of topics the consumer will subscribe to. As final setting step, we create a `Kafka DirectStream` defining the `Key` and `Value` classes as `String` and the `KeyDecoder` and `ValueDecoder` classes ad `StringDecoder`.

To process the stream of data, we map the received pairs of key and value to a tuple that consists of the letter and the related double. The function `mappingFunc` is the one used to perform a stateful map transformation. The state consists of a tuple made with a `Double` that represents the average and an `Int` that represents the number of times the letter related to that average appeared. Into the function we use the `state` to compute the new average and update the new state consequently. The output of the function is a tuple that consists of the letter and the new average. The output is then saved dynamically into the Cassandra table we created previously.

II. How to run the code

The preliminary steps, after having installed Cassandra and Kafka, are:

Start Zookeeper

```
>zkServer start
```

Start Kafka

```
>kafka-server-start /path/to/server.properties
```

Create a topic ‘avg’

```
>kafka-topics --create --zookeeper localhost:2181 --replication-
factor 1 --partitions 1 --topic avg
```

Start Cassandra

```
>cassandra -f
```

Run the generator

```
>cd /path/to/generator/
>sbt run
```

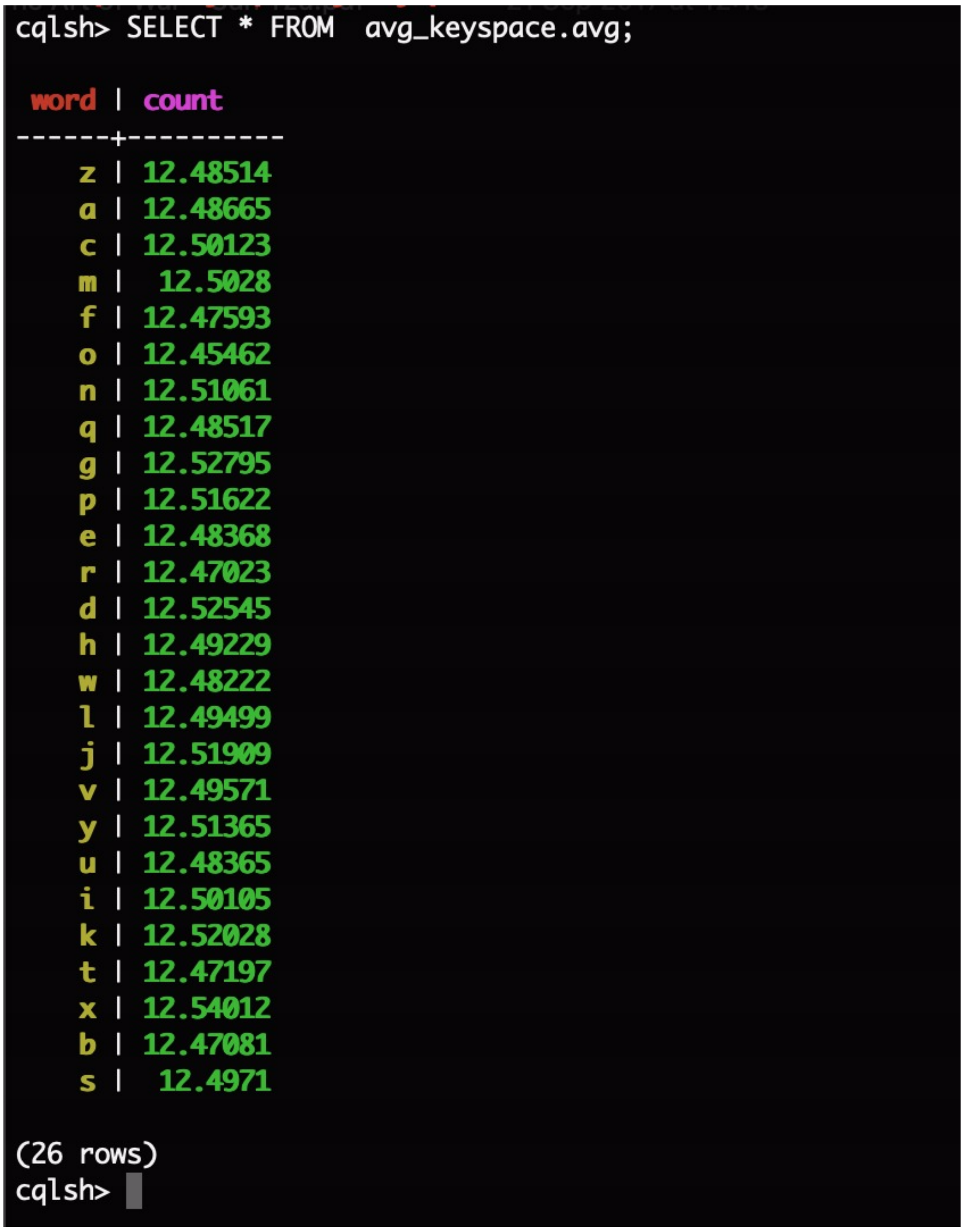
Run the project

```
>cd /path/to/sparkstreaming/
>sbt run
```

To check the results, run:

```
>cqlsh
>SELECT * FROM avg_keyspace.avg
```

III. Results



```
cqlsh> SELECT * FROM  avg_keyspace.avg;

 word | count
------+----------
    z | 12.48514
    a | 12.48665
    c | 12.50123
    m |  12.5028
    f | 12.47593
    o | 12.45462
    n | 12.51061
    q | 12.48517
    g | 12.52795
    p | 12.51622
    e | 12.48368
    r | 12.47023
    d | 12.52545
    h | 12.49229
    w | 12.48222
    l | 12.49499
    j | 12.51909
    v | 12.49571
    y | 12.51365
    u | 12.48365
    i | 12.50105
    k | 12.52028
    t | 12.47197
    x | 12.54012
    b | 12.47081
    s |  12.4971

(26 rows)
cqlsh>
```