**Assignment 4: Kafka + Spark Streaming + Spark SQL**

In this assignment you are required to use Kafka and Spark Streaming to consume and process streaming data from Twitter.

**Details**:

1. Before doing the assignment you should learn how to use Kafka and Spark Streaming. It is highly recommended that you read the Kafka Guide [Apache Kafka Quickstart Guide](https://kafka.apache.org/quickstart) and follow steps 1-5 + 7. Then read the Spark Guide [Structured Streaming + Kafka Integration](https://spark.apache.org/docs/2.4.3/structured-streaming-kafka-integration.html), first part (Reading Data from Kafka, using Python).

**Assignment**

1. You should use Kafka to consume data from Twitter. You will need to have a Twitter account and API key.
2. Write a **push\_kafka.py** python file to create a "Corona" topic and consume all tweets that mention "Corona" during one minute. There should probably be 10s to 100s of such messages.
3. Write a **kafka\_to\_spark.py** python file to read the messages from Kafka into a Spark context.
4. For parts 2+3 we provided you with **push\_kafka.py** and **kafka\_to\_spark.py** template files that you can get from Moodle. **You will need to complete the missing parts**.
5. Add Spark code to **kafka\_to\_spark.py** to print the following every 10 seconds:
   1. First 5 tweets
   2. Tweets that contain the word "USA"
6. When you finish the assignment submit your **two** python files (push\_kafka.py & kafka\_to\_spark.py) to Moodle.
7. **When done delete all AWS resources**! Otherwise, you will keep paying for them.

**Tips**

1. Create an ec2 instance with type **t2.large** (image: ubuntu)
2. You need to run simultany 4 processes on your virtual machine (one for zookeeper, one for kafka, one for push data from twitter to kafka and one for run spark streaming on your data in kafka).  
   You can do that with mobaXterm by opening 4 tabs which are connected to your virtual machine.
3. Connect to your ec2 instance and run the following commands *one time* (after creating your virtual machine):

*sudo apt update*

*sudo apt install -y python python3-pip default-jdk*

*pip3 install pykafka tweepy pyspark*

*wget https://archive.apache.org/dist/kafka/2.3.1/kafka\_2.12-2.3.1.tgz*

*tar xzf kafka\_2.12-2.3.1.tgz*

*mv kafka\_2.12-2.3.1 kafka*

*wget https://downloads.apache.org/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz*

*tar -xzf spark-2.4.8-bin-hadoop2.7.tgz*

*mv spark-2.4.8-bin-hadoop2.7 spark*

1. Open a **new connection** to run zookeeper server:

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

1. Open a **new connection** to run kafka server:

./kafka/bin/kafka-server-start.sh ./kafka/config/server.properties

1. Open a **new connection** to run python file **push\_kafka.py** which push data to your kafka topic:

*python3* **push\_kafka.py (get this file from moodle and modify it)**

You can see the output of your code in your kafka topic by command: *./kafka/bin/kafka-console-consumer.sh --topic twitter --from-beginning --bootstrap-server localhost:9092*

1. Open a **new connection** to run a spark job which read data from your kafka topic:

*spark/bin/spark-submit --packages org.apache.spark:spark-streaming-kafka-0-8\_2.11:2.4.8* ***kafka\_to\_spark.py* (get this file from moodle and modify it)**