To follow our own lab, you should start Hadoop, MariaDB and Kafka in the virtual machine.

Here some instructions in case you want to run the whole lab:

* **Raw to STD – Dataframes.ipynb** was used to promote in Hadoop the tweets from the raw layer to the std layer and saving them with a parquet format. We download them in case you want to simulate everything, as the **tweets\_war.zip**. This python notebook is only to show how we made the promotion as we created our own schema.
* **tweets\_war.zip**, has all the tweets that we got from the twitter API after promoting in a parquet format.

IF YOU NEED THIS TWEETS\_WAR.ZIP, PLEASE CONTACT ME AS IT WAS NOT POSSIBLE TO UPLOAD IT INTO GITHUB

* **Final\_war.ipynb** is the notebook used in the batch pipeline. Here is where you should read all the tweets from **tweets\_war.zip**.

Within this file you have the exploration analysis, the graphing part, and the machine learning part.

**IMPORTANT: THE PROCESSING AND THE SERVING PART SHOULD BE RUN INDEPENDENTLY. If you want to check the processing run that part, but later restart kernel and run only the serving part.**

* **Processing and Serving in Streaming.ipynb** is the notebook to run the streaming part. Here you have the streaming processing part, which needs call first from the console the producer.py, and the serving part.
* **tweets\_producer.py** is the file you should call from the console before running the streaming notebooks to send messages to the tweets topic in kafka. The source of this streaming part is the **tweets\_war.txt**. The query is the following:

*“python3 tweets\_producer.py send -m “This is a message from the Python client” -t tweets”*

* **tweets\_consumer.py** is the file you could call from another console to check how the tweets are being storage in kafka.

*“python3 tweets\_consumer.py tweets IE 300”*

* **tweets\_war.txt** is the simulated database with all the tweets in a txt.file. Is called by the producer.
* Before running the serving part in the **Processing and Serving in Streaming.ipynb** notebook, is mandatory to run from the console the **tweets-db** file with SQL commands.

*“mariadb -u osbdet -p < tweets-db.sql”*

* Once you run the sql file and the SQL table is created in Mariadb, run the producer, and run **only** the serving part of the notebook.
* The notebook will use the csv file **Strategies\_Activity.csv** file, to join the batch excel with the streaming part. It is a excel created by ourselves during the LAB with some “marketing information”.
* Enter mariadb and change to the DBS\_tweets database.

“mariadb -u osdbdet -p”

use DBS\_tweets

* Here select the queries that you want over the tweets\_final table. Some queries we launched:

“Select \* from tweets\_final where Activity = ‘Adict’ and user\_followers\_account>10000”