**CLOUD COMPUTING (P4) – *Spark for Real Problems***

***Twitter data question 5:***

What twitter user tweeted the most? What is the top 5 longest tweeters over each’s average tweet length? Bottom 5?

***Required data to answer the questions:***

1. User data in twitter data set
   1. Only screen name is required to answer the question.
2. Twitter text – Only length of the text.

According to the question, we get to know that we have to find tweet length of each person. At the end we have to write 3 different queries to calculate the user who tweeted the most, top 5 longest tweeters and bottom 5 tweeters over each’s average tweet length. This can be done using Spark by creating a python file (twitter 5) and submitting it to Spark by following the below approach.

***Approach:***

Using Spark and writing a simple python code we can solve this question. Find below the steps followed in solving the question:

* Initially, print\_function, SparkContext, SQLContext are the main functions that have to be imported to the python file.
* Then a function is created a mapper which parses each input line using json library and extracts the Screen Name from user data and Twitter text.
* The dataset is then loaded into a RDD using the sc.textFile(sys.argv[1]) when the file name is sent from the command line parameters.
* Using “flatMap” the RDD is then transformed by using the function getInfo which was initially written to parse each input line.
* Using “sqlContext” and “createDataFrame”, a map is implemented to create a new transformed RDD as shown below

sqlContext.createDataFrame(texts.map(lambda (u,t): Row(text = t, length = len(t), user = u)))

* A temporary table is created using the function registerTempTable(“Name\_Of\_Table”).
* 3 SQL queries are written to find out the user who tweeted the most and the top/bottom users from the temporary table created in the above step.
* The program is then sent to the Hadoop cluster using the following command:

spark-submit --master yarn-client twitter5.py hdfs://hadoop2-0-0/data/twitter/part-03212

* At the end the results are printed using few for statements and then printed in the console.

***Results:***

marilyn9743 3419

Huntersweat 416

RoyalEliteKiva 350

blackxhole 320

KelleeMichele 272

pizzadellarry 253

Laila\_Lafrai 0

Fun\_Size20 0

Trevorsturgill5 0

2013Afi9 0

lm\_Lil\_Wanie 0

***Final result and observations:***

The first line is the output for twitter user who tweeted the most who is **marilyn9743** and the tweet length = **3419 characters**

The second set of output (next 5 lines) are the top 5 longest tweeters over each’s average tweet length. We can observe that the 1st longest tweeter is approximately double the 5th longest tweeter.

The third set of output (next 5 lines) are the bottom 5 tweeters over each’s average tweet length, we can observe that their tweet length is empty which means they just submitted a tweet with length of the tweet as 0.