# **Report on Twitter Question 1**

AVERAGE NO. OF TWEETS – **26** for the 17th hour

This report analyzes twitter dataset question 1 where we need to find the hour of the day PrezOno tweets the most on average. We use SPARK to analyze the trends of a particular twitter user. It is run in Hadoop clusters through YARN and it processes data in HDFS.

The data is fetched through hdfs and its patterns were carefully observed. It was analyzed that the total number of tweets by PrezOno was 341 for the given year.

The first part of this implementation was analyzing the collection of tweets using **sc.textFile**. The second part was to filter the tweets such that only tweets by PrezOno are stored using **[“user”][“screen\_name”]**. Next, the tweets were mapped against each hour of the day using **‘created\_at’** and split function. They were then reduced by key pair for quick results using **reduceByKey(lambda a, b: a+b)**. The output was stored in a text file using **saveAsTextFile()**.

The code was run in Hadoop cluster using spark-submit. The output file obtained helped me to plot the graph of the expected number of tweets for each hour of the day, for those PrezOno tweeted.