PRINCIPLES OF BIG DATA MANAGEMENT PROJECT

Veeresha M Thotigar, Sai Sampath Kumar Raigiri and Sai Srinivas Vidiyala

Department of Computer Science Electrical Engineering, University of Missouri-Kansas City, Missouri

Introduction

In this project we design, develop and execute a application that would visualize interesting analytic queries executed on tweets on topic related to #YOGA , #EXERCISE, #GYM , #FITNESS and #DIET.

Objective

The main objective is to analyze, store and visualize the downloaded twitter big data on topic HEALTH.

We are interested in extracting the significance of health from this large data set

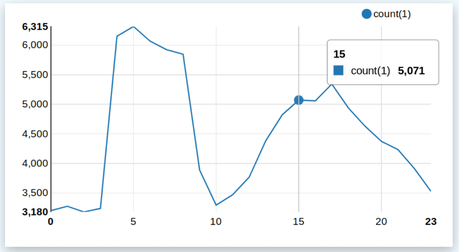
Technologies Used

- Hadoop (for distributed storage)
- Spark (query engine)
- Scala (programming language)
- Python (for collecting Twitter data)
- Zeppelin (visualization platform)
- Helium (for better visualization)

Methods

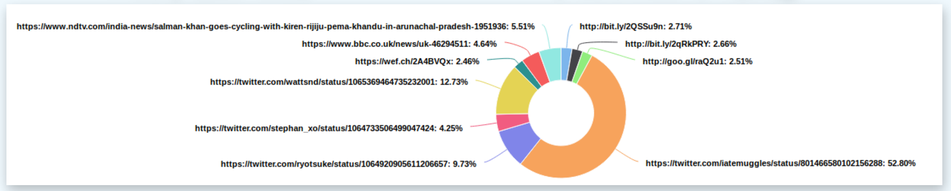
We have written a python code to download the tweets in JSON. After which stored it in the HDFS storage using “copyfromlocal” command. Later using “Zeppelin” package we visualized the big data using Spark query engine. Zeppelin is totally based on dependent on spark libraries. By starting the HDFS and Zeppelin we can load the JSON file through HDFS path in the zeppelin. By providing the analytical queries we have visualized the large data.

Results



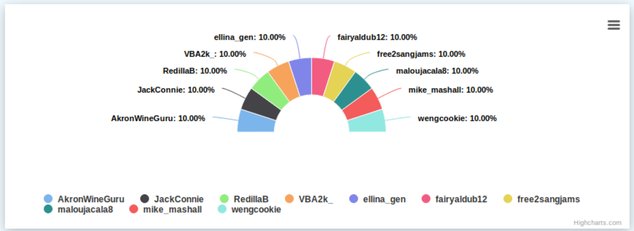
1. Hourly Based Tweets

2. Word Cloud for Hashtags



3. URL Based Tweets

4. Tweets Based on Places

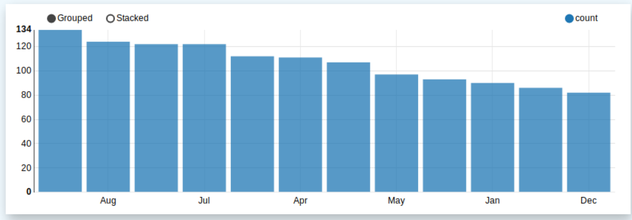


5. User Mentioned Tweets

The reports draw from the large data gives significant information like in fig-1 the people are most interested in the fitness ranging in day time from morning from 4am to 10am and also in the evening from 5pm to 8pm.

Like this we draw many reports that would help us how health can be maintained from various activities and food plan that we need to follow.

Reports also says that what kind of activities are followed most to maintain the good health over region of the place.



6. Number of Tweets by Users

Performance

When compared to traditional data base ware houses like MangoDB. The average query retrieval of traditional query engine 25 to 30 secs where as Spark query engine takes 5 to 8 seconds in a single node cluster and we were expecting it would reduce to milliseconds when we move to multiple nodes.

Conclusion

We get insights by analyzing and visualizing the Twitter data on the topic "Health" like what particular time hours the people are talking about health, what people are talking about health etc,.

We also got hands on how to save data in SparkQL data frames, run jobs and visualize them.

Under the Guidance

DR. PRAVEEN RAO (PH.D)

Associate Professor

Department of Computer Science Electrical Engineering

University of Missouri-Kansas CIty