**Big Data Programming**

**Team 4 Project Proposal**

**Title:**

Twitter Web series Data Visualization with Spark ETL

**Team Members:**

* Avinash Ganguri
* Akhil Teja Kanugolu
* Bhashitha Siddareddy
* Geetanjali Makineni

**Goals and Objectives:**

**Motivation:**

The motivation behind doing this analysis is we are living in a world where data handling and data using plays one most important role in making the decisions for most of the industries like Banking, Financial, Telecom and Health and IT sector serving ones. The main factors would be for getting the insights would be like managing its sheer volumes of data and its insights. Using the Apache Spark is one of the best amazing kind frameworks which will be handling big data and its real time performance of these analysis.

**Objectives:**

The main objective of our project is doing the ETL process using the Spark’s Batch Processing and then Spark Integration using Web UI. The main source of our data is twitter and then collecting the data with Spark Batch Process. We can perform our transactions on the set of RDD’s and later we load our data in our Hive which is similarly equal to the ETL basic process.

**Features:**

The main feature of the project is to collect the Real timed tweets from the twitter steaming API, also by performing the ETL which means we preprocess the data and extract the necessary data and then we load this data in our HIVE. Next, we use TextBlob for predicting the sentiment for each of the mentioned tweet. Later, we feed the data into our HDFS and then we implement SQL and Hive queries. Sqoop is used for transferring data between SQL and HDFS.

**Significance:**

Analyzing of sentimental analysis is done using one tool which is already existing named as ML tool which is TextBlob for prediction of sentiment on the tweets and Later we are using spark for writing the queries by visualizing with Panda.

**Implementation:**

1. We first collect the tweets from the Twitter api.

2. Later, we import the data that is collected into the hive from the HDFS.

3. Next we will export this data from there into RDBMS by usage of sqoop.

4. We later do the sentimental analysis on the tweets that are collected.

5. Now, we are using spark sql for writing required queries later by visualizing (examples like bar graphs and pie charts and some other type of graphs) the obtained results by using panda.

**WHO:**

Helps the advertiser’s for pushing the advertisement and series makers for making quick analysis.

**WHAT:**

Pulling the data helps them in predicting the TRP which in result helps the series makers and advertisers.

**WHEN:**

As a result of COVID-19 most of the people settled back in their homes lacking entertainment.

**WHERE:**

All over the world.

**WHY:**

Most of the people believe in IMDB rating rather than genre concept.

**Contribution:**

Increment-1:

* Downloading data batch from twitter (Avinash, Geetanjali)
* Implementation and Data preprocessing (Akhil, Bhashitha)

Increment-2:

* HDFS, Hive imports with HUE visualizations.
* Hive Queries
* Spark Sql Queries, Data Frames, and visualizations.

Final:

* Sentimental Analyzation (TextBlob)
* Solr Queries and Graph frames on the twitter data.
* Hosting on a Web Application.

**References:**

1. [developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json](https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json)
2. [www.info.com/**Twitter Data Visualization**](https://www.info.com/serp?q=twitter%20data%20visualization&segment=info.0376&s1aid=144007235&s1cid=357227086&s1agid=1264438724330452&s1kid=kwd-79027647541819:loc-190&utm_source=adcenter&msclkid=fa5513571bee1e925c7b4c24b6d1fd75)
3. [spark.apache.org/docs/latest/](https://spark.apache.org/docs/latest/streaming-programming-guide.html)