**Interim Project Report – Principles of Big Data Management**

**Teammates:**

1. Avinash Ganguri

2. Sri Sai Nikhil Kantipudi

3. Roshna Toke

**Introduction:**

Phase-1, Collected tweets in JSON format and separated the URL’s and Hashtags from the tweets and stored the data in HDFS and performed the Word Count in Spark and Hadoop, and secured the log file and output in local destination folder.

Phase-2, Plans to Write analytic queries to surf and understand the data and develop the visualizations as pie charts, bar graphs, and heat maps etc. Also, to work on tweets related to La Liga (2019 - 20) Football matches.

**Analytic Queries:**

1. What is the most popular team in La Liga?
2. Who is the most popular player among all the teams?
3. The positive and negative tweets about the popular player?
4. The tweets from different countries about La Liga.
5. The frequency of Retweets ?
6. Who is the popular fan of La Liga that tweets most than any other?
7. Tweets count for different types of franchises in La Liga
8. What is the most frequently occurring tweet?
9. What is the frequency of tweets posted on different days?
10. A query on tweets made through verified accounts.

**Visualization:**

Planned to visualize the queries in the form of bar graphs, Pie charts, line graphs, heat maps and etc. using the D3.JS and frontend using HTML and CSS.

**Distribution of Work:**

* All the team members will be working together to visualize the queries.
* All the 10 different queries will be shared among the team members and will be working accordingly.
* Documentation will be done by each member respectively.

**Software Testing:**

* By Unit Testing, the individual units or components are tested. Unit Testing of software is done during the development of application.