**Project:**

Spark ETL on Twitter Data, Visualization.

**Goals and Objectives:**

In this data-driven world, handling data has become vital in the decision-making process in many industries such as Telecom, Banking, Financial and Health sector servicing industries. Managing the sheer volumes of data and getting insights from it would be the main factors. One of the amazing frameworks that can handle big data in real-time and perform different analysis, using Apache Spark.

**Objectives:**

Our Project's main idea is to do the ETL process using Spark Batch Processing and Integrating Spark with Web UI. The source of our system is Twitter data and we would be using Spark batch Processing to collect the data. Then we would be performing the transformations on the set of RDD’s and load the data into the Hive system which is similar to basic ETL process.

**Features:**

The project’s features includes collecting real-time tweets from twitter streaming API, performing ETL (Pre-Processing the data, Extracting necessary information and loading the data in to the Hive), and using TextBlob, predict the sentiment for each tweet. We are trying to feed the data into HDFS and implement Hive, SQL queries and we are using Sqoop to transfer data between SQL and HDFS.

**Significance:**

For sentiment analysis we are using a existing ML tool(TextBlob) to predict the sentiment of the tweets. Use spark to write queries on visualization using panda.

**Implementation:**

1. Collect tweets from Twitter api.

2. Import the collected data into hive from HDFS.

3. Export the data into RDBMS using sqoop.

4. Sentimental analysis on collected tweets.

5. Sentimental analysis on collected data.

6. Use spark sql to write queries and visualize(bar graphs/pie charts etc) the results using panda.

**References:**

* <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json>
* [https://spark.apache.org/docs/latest/](https://spark.apache.org/docs/latest/streaming-programming-guide.html)