INTERNSHIP PROJECT REPORT

BIG DATA

TWITTER DATA ANALYSIS

Navya Dahiya
BIRLASOFT LTD

INDEX

I. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Mr. Akhil Kakkar for providing his invaluable guidance , comments and suggestions throughout the course of the internship. I would also like to thank the Project Manager- Mr. Manoj Bidichandani for helping me throughout the period.

II. INTRODUCTION

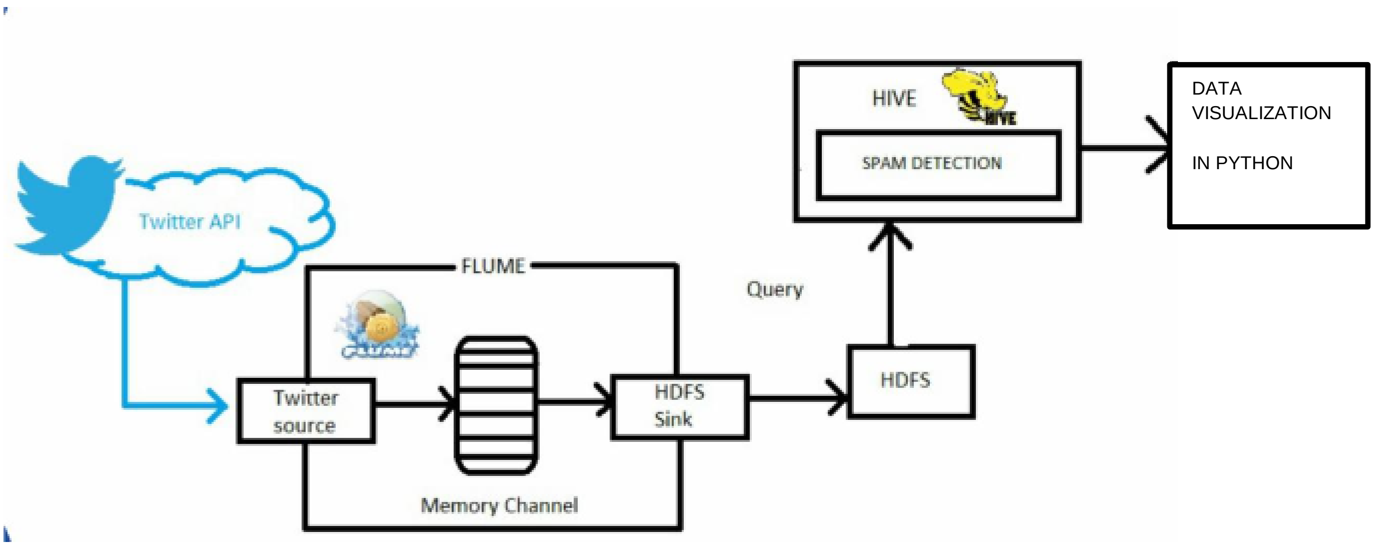
Sentiment essentially relates to feelings, attitudes, emotions and opinions. A person's opinion or feelings are for the most part subjective and not facts, which means to accurately analyze an individual's opinion or mood from a piece of text can be extremely difficult. With Sentiment Analysis at word level, from a text analytics point of view, we can get an understanding of the attitude of a writer with respect to a topic in a piece of text and its polarity: positive, negative or neutral.

In recent years, there has been a steady increase in interest from brands, companies and researchers in Sentiment Analysis and its application to business analytics. Many microblogging websites have evolved to become a source of varied kind of information as people post real time messages about their opinions on a variety of topics, discuss current issues,etc among which Twitter is one.

Twitter allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day .Twitter is a better approximation of public sentiment as opposed to conventional internet articles and web blogs.

Performing Sentiment Analysis on Twitter is trickier as the tweets are very short (only about 140 characters) and usually contain slangs, emoticons, hashtags ,etc. For the development purpose twitter provides streaming API which allows the developer an access to 1% of tweets tweeted at that time bases on the particular keyword. The object about which we want to perform sentiment analysis is submitted to the twitter API's which does further mining and provides the tweets related to only that object. Twitter data is generally unstructured i.e use of abbreviations is very high. Also it allows the use of emoticons which are direct indicators of the author's view on the subject. For doing twitter data analysis first data is collected using FLUME in local HDFS. Tweets are preprocesses for removing noise, meaningless symbols and spams.

III. WORKFLOW



IV. TECHNOLOGIES USED:

1) HADOOP

Apache Hadoop is an open source software framework for distributed storage and large scale distributed processing of data-sets on clusters.Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data.

Hadoop framework includes different modules like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase for different functionality .

Hadoop use HDFS (Hadoop Distributed File System) file system.HDFS uses a master/slave architecture where master consists of a single Name Node that manages the file system metadata and one or more slave Data Nodes that store the actual data. Benefit of using Hadoop is distributed storage, Distributed Processing, Security, Reliability, Speed, Efficiency, Availability, Scalability and lots more. This is the reason of using Hadoop for tweet processing.

2) FLUME

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS. For access to data on Twitter, a Twitter API account must be created.The Consumer Key, Consumer Secret, Owner Key and Owner Secret ID and the Access token would be . The flume.conf file needs to be configured in which the file configure sink is set as HDFS and the path is set to HDFS for storing the tweets.After running the configuration file, tweets start downloading in HDFS in specified path. After a couple of minutes the Tweets should appear in HDFS. The data downloaded in HDFS is in JSON format. That needs to be converted into readable format.

3) HIVE

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Apache Hive (HiveQL) with Hadoop Distributed file System is used for Analysis of data. Hive provides a SQL-like interface to process data stored in HDP. By default, Hive expects that input files use a delimited row format, but Twitter data is in a JSON format, so we can use the Hive SerDe interface to specify how to interpret what we've loaded. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process. JSON-SERDE jar file should be used for Hive to read and understand the unstructured json data.

4) PYTHON – DATA VISUALISATION

Have used Python libraries-Pandas, Numpy, Matplotlib and Codecs to visualise the analysed data. Codecs library has been used to decode the list of country names to utf-16 and encode it in utf-8 (to remove characters other than the alphabets in the python list)

V. DETAILED WORKFLOW

1) Cleaning of Twitter Data

Twitter data is the end product and hence does not need much cleaning. Cleaning here involves the following:-

a)Taking forward only those data that are useful in our analysis.
b)Removal of foreign language characters

2) Removal and detection of Spam

To get the best results it’s important to filter the content relevant to the use case, and to remove what is considered as spam.The following criteria has been used to remove the users who spamA user who is likely to follow lots of users, but be followed by very few users itself is a spam as the twitter profile is created just to post spam messages.In this case, the users’ followers’ ratio (number of users who follow the user, divided by the number of users they follow) is low.

- Users who post less tweets

If a Twitter profile has no description, again it could be from a bot or at least from a user who doesn't care about their public profile and has little concern for the quality of content they post.

- Large Numbers Of Hashtags

Poor quality content tends to include many hashtags. Many hashtags might be used by spam creators to hope that they can reach as many users listening to those tags as possible

- Short Content Length

Often users will write very short posts, such as '@friend ok' as a response to a question. This content has little value in analysis**.**

3). ANALYSIS:

a. Top 10 Hashtags
b. Top 5 locations with most tweets
c. Top 10 active users
d. Top sources used for tweeting
e. Top 10 most followed users
f. Top 10 users with maximum retweets
g. Country wise sentiment analysis
h. Overall sentiment analysis among users

4) DATA VISUALIZATION

The following was done to visualise data using Python :

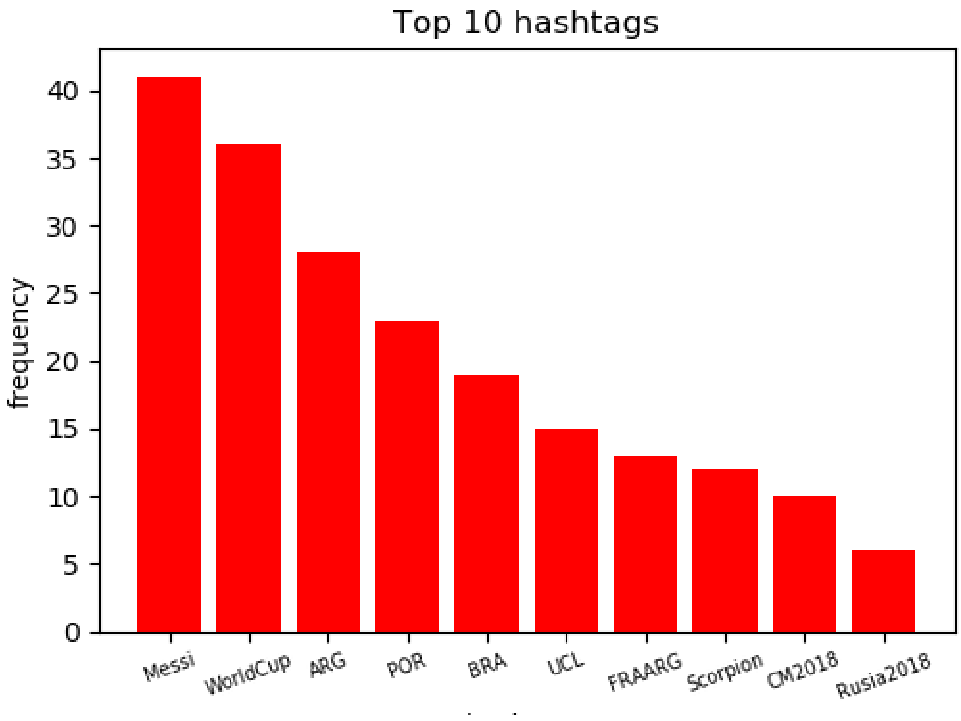
1.Save anaysed results to HDFS diretory

2. Retrieve the saved data from HDFS into local system

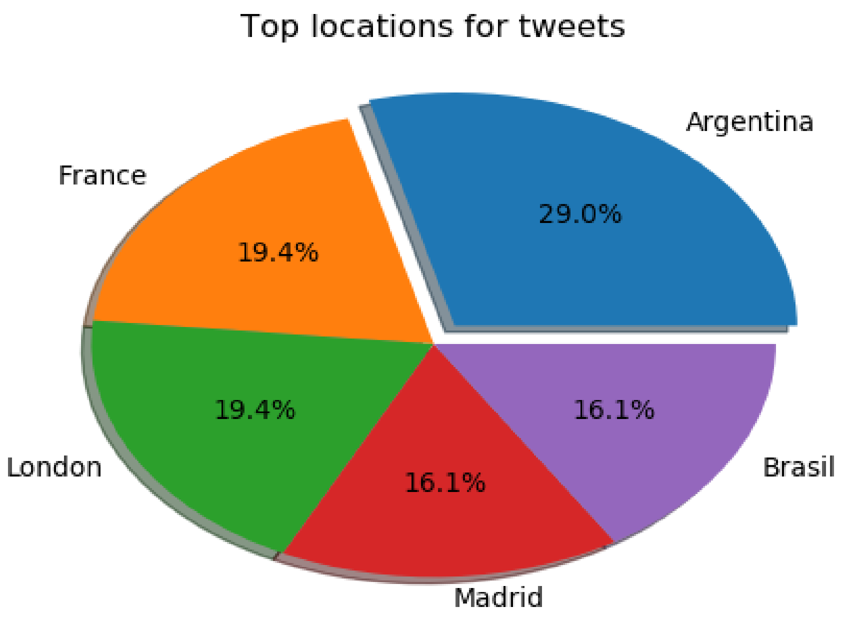
3. Use python to visualise

VI. RESULTS

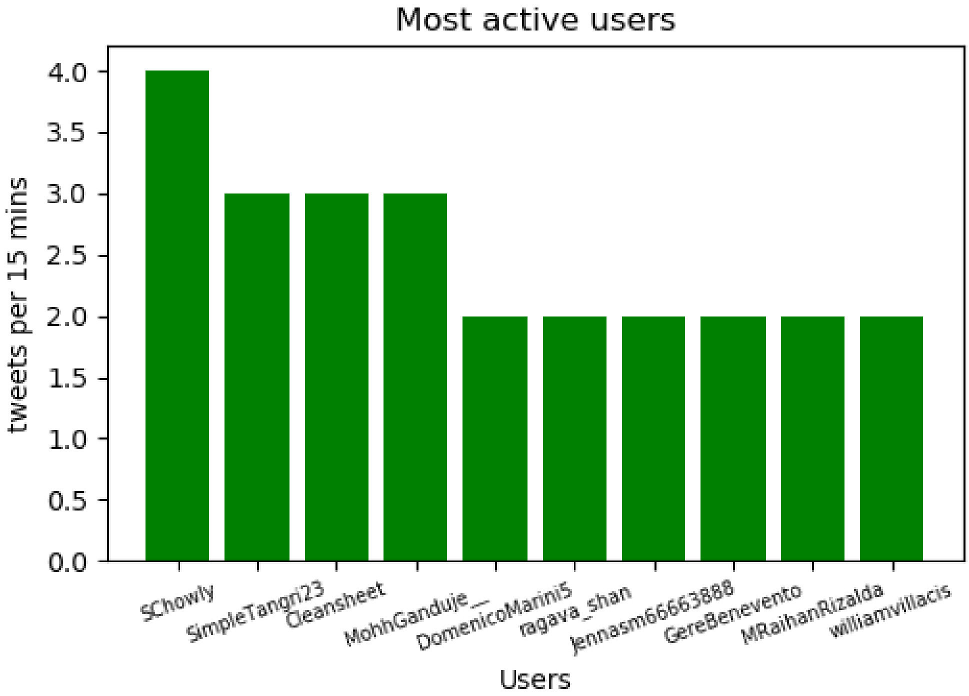
1. **Top 10 Hashtags-** the topics on which most people are expressing opinion



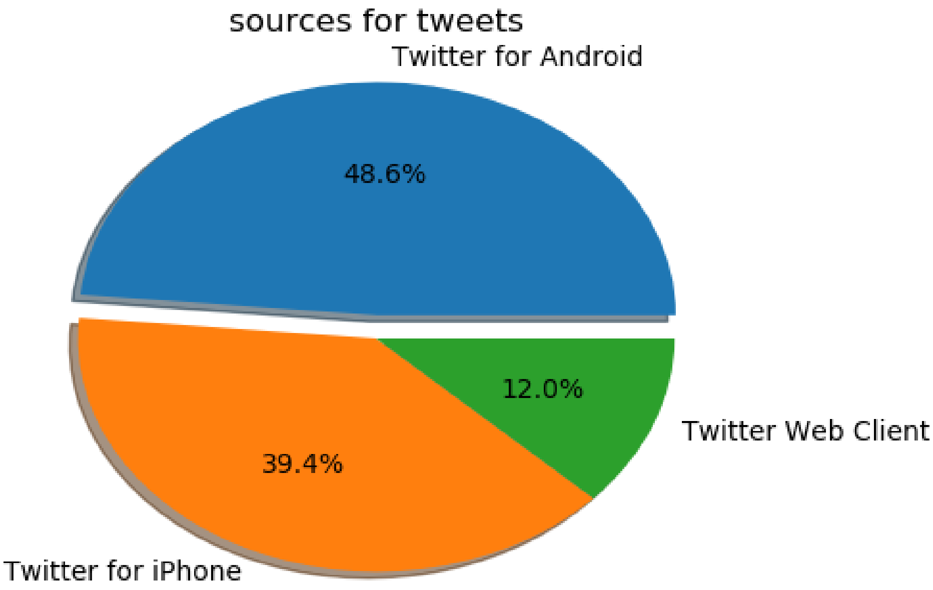
2. Top 5 locations with most tweets



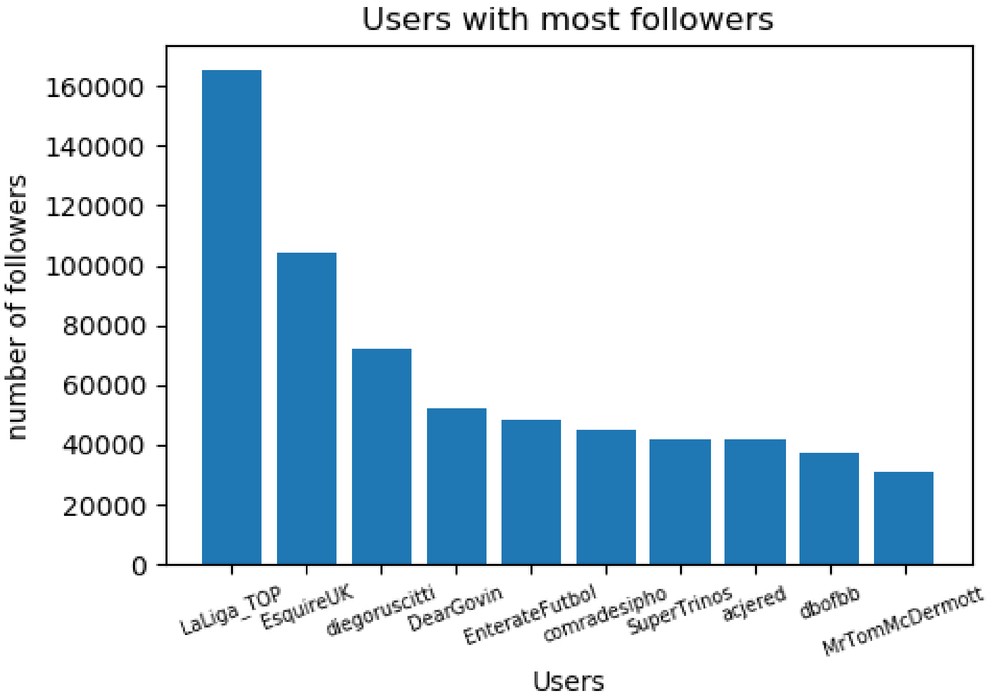
3. Top 10 active users



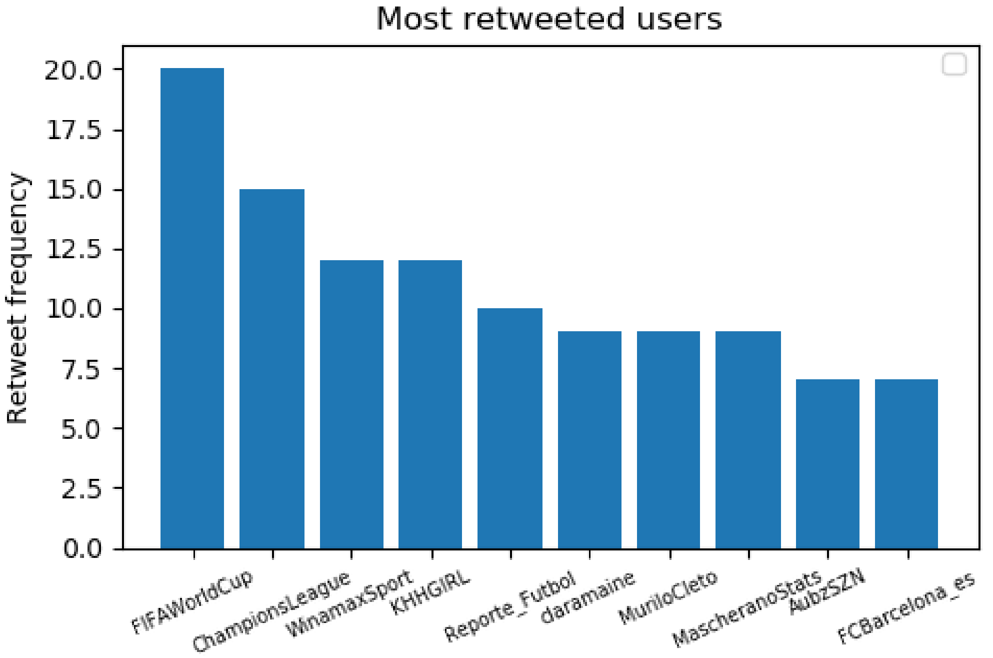
4. Top sources used for tweeting



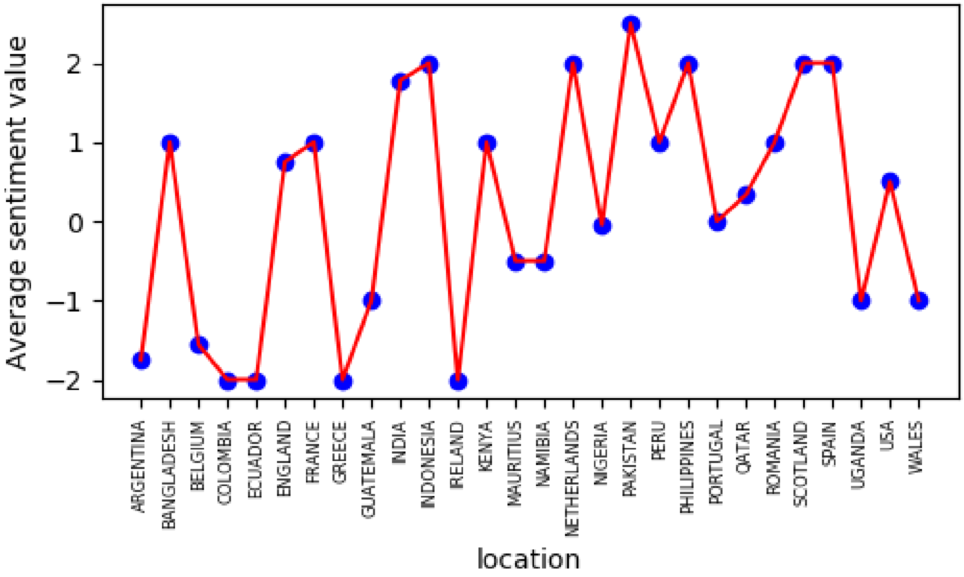
5. Top 10 most followed users



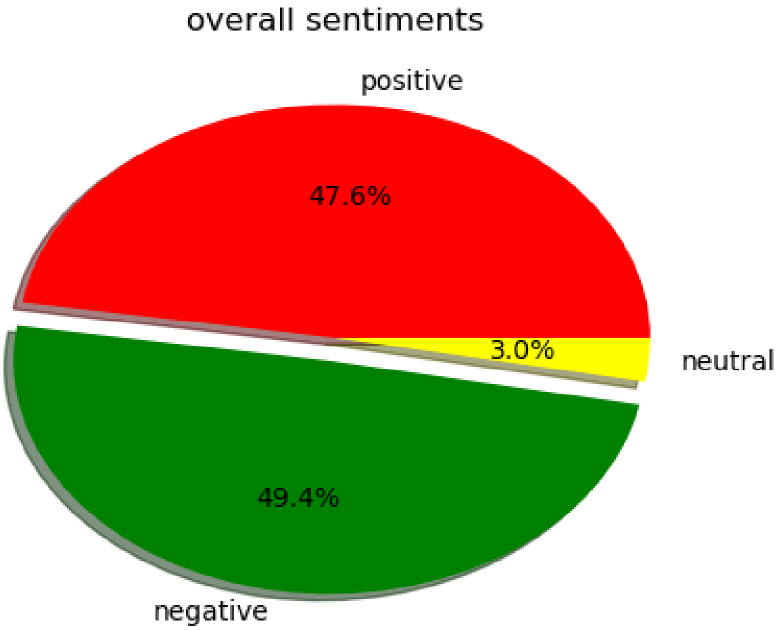
6. Top 10 users with maximum retweets



7. Country wise sentiment analysis



8. Overall sentiment analysis among users



VII. CHALLENGES FACED

1. Retrieving the twitter data was difficult as twitter have not made their data public. Data is provided only to twitter apps.
2. Flume setup needed a lot of complex classpath JARs and configurations.
3. Creating a customized JSON SerDe for conversion of unstructured JSON data to structured Hive tables.
4. Calculation of Country wise sentiment value.
5. Inefficient JVM runtime memory in the system

VIII. CONCLUSION

Successfully streamed data from twitter using Flume, handled unstructured data in Hive , cleaned twitter data, removed spam from it , performed analysis :country wise polarity of sentiments,overall polarity, most active users, etc using Hive and visualised it using python libraries.

IX. OTHER LEARNINGS

1) SQOOP

Sqoop stands for "SQL to Hadoop and Hadoop to SQL" .Sqoop is a command-line interface application for transferring data between relational databases and Hadoop. Sqoop internally uses Map Reduce jobs to import the data and spread it across the cluster**.** The connection to the Oracle database should be made correctly else, the tool would not work . For the connection to be established, the user should pass the username and the password of the database that they would be using in the Orace DB. Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported.

2) CREATION OF HIVE SCRIPT

The datatypes in HQL are different from the datatypes in Oracle DB. So, to create a table in Hive, same as that specified in the Oracle DB , one has to manually note down all the datatypes on a sheet of paper, convert the datatypes to Hive compatible types and then wite the DDL command in Hive to create the table having the same schema as that in Oracle DB, which becomes a cumbersome task. So, I made a program in Java that reads all the tables in Oracle DB, converts all the datatypes in the schema of all the tables to make it compatible with HQL and automatically generate a Hive DDL command that lets the creation of the table successful. The code in Java is scalable and I had successfully tested it on 153 Oracle DB Tables.