## A MAJOR-PROJECT REPORT

### ON

**“TWITTER SENTIMENT AND DATA ANALYSIS”**

## Submitted to

**K.I.I.T. DEEMED TO BE UNIVERSITY**

## In Partial Fulfilment of the Requirement for the Award of

**DEGREE OF BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE ENGINEERING**

**ABSTRACT**

It is cumbersome to go through each of the hundreds of records of social media data manually. Most of the enterprise data warehouses dont have the ability to keep up with rapidly increasing social media data. There are several ways to define and analyze this data by performing different operation queries. But, the problem lies in the dealing with big-data of different types of data which is unstructured. Here, we try to solve it with the use of Hadoop and its packages.

An end to end data pipeline is to be designed that will enable us to analyze the nature and sentiment of the tweets. For example , we can find the relevant users and retweet counts for all tweets by analyzing the data. Using messages and retweets, we can go much further beyond and can find the followers of the person who send the original tweets. We try to engage users whose updates generate lots of retweets all the users for whom most retweets are generated and we sort them in descending order of most retweeted. To do this, we use HQL queries since twitter streaming API generates output tweets in JSON format which is semi-structured data of sev- eral types and therefore, complex and inconvenient. We use HIVE project in Hadoop ecosystem. HIVE provides a query interface to query data which resides in HDFS. Apache Flume will be used to aggregate and dump the streaming Twitter data into HDFS.

The twitter data and sentiments analysis is important to know the influential per- sonalities, network characteristics, nature of tweets at a given time, and so on, and it is possible free of cost thanks to the twitter API provided by the eponymous com- pany. The questions and their corresponding queries will be made unique as much as possible so that we can interpret something which was never discovered before. We will also use a dictionary lookup of certain kind of keywords to compare with the incoming tweets in order to know the sentiments behind them, and a web interface to visualize the analytics done on the Twitter data.

**Keywords:** Sentiment analysis, Hadoop, Apache flume, Twitter Streaming API, JSON, web interface

# Contents

1. Introduction 2
2. Literature Survey 4

[2.1 Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4](#_TOC_250018)

1. [Software Requirements Specification](#_bookmark4) 6

[3.1 Product Perspectiv](#_bookmark5)e . . . . . . . . . . . . . . . . . . . . . . . . . . 6

[3.2 User Classes and Characteristics](#_bookmark6) . . . . . . . . . . . . . . . . . . . 6

[3.3 Operating Environment](#_bookmark7) . . . . . . . . . . . . . . . . . . . . . . . . 6

[3.4 Software Interfaces](#_bookmark8) . . . . . . . . . . . . . . . . . . . . . . . . . . 6

[3.5 Hardware Interfaces](#_bookmark9) . . . . . . . . . . . . . . . . . . . . . . . . . . 7

* 1. [FUNCTIONAL REQUIREMENTS](#_bookmark10) . . . . . . . . . . . . . . . . . 7
     1. [EXTRACTION OF DATA FROM TWITTER WEBSITE .](#_bookmark11) . 7

[3.6.2 JSON SERDE FILE](#_bookmark13) . . . . . . . . . . . . . . . . . . . . . 8

[3.6.3 ANALYSIS OF TWITTER DATA](#_bookmark14) . . . . . . . . . . . . . . 8

* 1. [NON FUNCTIONAL REQUIREMENTS 10](#_TOC_250017)
     1. [Performance Requirements 10](#_TOC_250016)
     2. [Software Quality Attributes 10](#_TOC_250015)

1. Requirement Analysis 11
   1. [PROBLEM STATEMENTS 11](#_TOC_250014)
   2. [REQUIREMENTS 11](#_TOC_250013)
2. System Design 13
   1. [SYSTEM ARCHITECTURE 13](#_TOC_250012)
      1. [OVERALL ARCHITECTURE 13](#_TOC_250011)
      2. [FLUME COMPONENTS 13](#_TOC_250010)
   2. [WORKFLOW DIAGRAMS](#_bookmark25) 14
      1. [OVERALL W](#_bookmark27)ORKFLOW 14
      2. [INSTALLATION TO END](#_bookmark28) 14
      3. [USECASE DIAGRAM OF TWITTER DATA ANALYSIS .](#_bookmark34) 15
      4. SEQUENCE DIAGRAM OF TWITTER DATA ANALYSIS 15
3. [System Testing](#_bookmark37) 16
   1. [CONCLUSIVE TABLE 16](#_TOC_250009)
   2. [RESULTANT GRAPH 17](#_TOC_250008)
   3. [OUTPUTS 17](#_TOC_250007)
      1. [Most popular hashtags in tweets containing certain keywords 17](#_TOC_250006)
      2. Most followed users giving tweets related to certain keywords 18
      3. [Most retweeted users giving tweets related to certain key- words. 18](#_TOC_250005)
4. Project Planning 19
   1. [GANTT CHART 19](#_TOC_250004)
5. [Implementation](#_bookmark47) 20
   1. [SYSTEM IMPLEMENTATION](#_bookmark48) 20
   2. [Creating Twitter Application](#_bookmark49) 20
   3. [Getting data using Flume](#_bookmark51) 21
   4. [Querying using Hive Query Language](#_bookmark53) 22
   5. [SENTIMENT ANALYSIS](#_bookmark54) 24
      1. [STEPS FOR PERFORMING SENTIMENT ANALYSIS](#_bookmark55) 24
6. [Screenshots of Project](#_bookmark56) 27
   1. [HADOOP SNIPPETS](#_bookmark57) 27
   2. [Web Interface Snippets 32](#_TOC_250003)
7. Conclusion and Future Scope 41
   1. [CONCLUSION 41](#_TOC_250002)
   2. [PROPOSED FUTURE ENHANCEMENTS: 41](#_TOC_250001)

[References 42](#_TOC_250000)

**List of Figures**

3.1 **Twitter Logo** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

### System Architecture 13

* 1. **Components of Flume** 13

### Workflow of proposed system 14

* 1. **Installation to end** 14

### [Use case Diagram](#_bookmark33) 15

* 1. [**Sequence Diagram for Twitter Data Analysis**](#_bookmark35)15

[6.1 **Accuracy Graph**](#_bookmark39)17

[7.1 **Gantt chart**](#_bookmark46)19

### [Getting Consumer Key Access Token](#_bookmark50) 21

* 1. [**Configuration**](#_bookmark52)22

### [Twitter Data streaming Through Flume - commands](#_bookmark58) 27

* 1. [**Twitter Data streaming Through Flume**](#_bookmark59)27

### [Flume Data blocks of maximum 1000 tweets](#_bookmark60) 28

* 1. **Twitter data in JSON format** 28

### Twitter data streaming completed 29

* 1. **Table Creation query** 29

### Query Internally Converting to MR jobs 30

* 1. **Query showing most followed users** 30

### Query showing most popular hashtags 31

* 1. **Query for finding most retweeted users** 31

### HTML code for the index page 32

* 1. **JS code for showing change in number of tweets** 32

### [JS code for showing most popular hashtags](#_bookmark70) 33

* 1. [**JS code for showing most followed users**](#_bookmark71)33

### [Index Page](#_bookmark72) 34

* 1. [**Change in**](#_bookmark73) **number of tweeets** 34

### [Most popular](#_bookmark76) hashtags 35

### [Most followed users 35](#_bookmark77)

### [Hive table that will extract the](#_bookmark77) id and tweet text from the tweets

### using the Cloudera JSON serde 36

### Word table containing the array of strings 36

### use the split() function to split the text as words, it will return an array of values 37

### We will create another table called tweet word which can store

### id and word in general 37

### Dictionary table to load the contents of AFINN dictionary 38

### Contents of dictionary 38

* 1. **Table word join comprising the join the tweet***wordtableanddictionarytable* 39

### [Sentiment table comprising the average rating of each tweet by](#_bookmark84) [using each word of the tweet and arranging the tweets in the](#_bookmark84)

### [descending order as per their rating (1)](#_bookmark84) 39

### [Contents of sentiment table](#_bookmark85) 39

### [Total number of positive,negative and neutral tweets](#_bookmark86) 40

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 1 Introduction

Apache Hadoop is a better choice for twitter data and sentimental analysis as it works for distributed big data. Apache Hadoop is an open source tool for dis- tributed storage and high scale distributed processing of data-sets on Hadoop clus- ters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel form on different data nodes. In short, Hadoop tool is capable to develop applications capable for running on Hadoop clusters of data node and they could perform complete statistical and graphical based analysis for large amounts of data. Hadoop MapReduce is a software program for easy writing applications which can process ample of data in-parallel on high end clusters of commodity hardware in a reliable, fault-tolerant mode. Hadoop tool consists of different modules like MapReduce, Flume, Hive, Pig, Sqoop, Oozie, Zookeeper, Hbase for different func- tionality and features as shown in below diagram. I will be using FLUME and HIVE for twitter data and sentimental analysis.

Apache Hadoop Framework Hadoop use HDFS (Hadoop Distributed File System) file system. The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner. HDFS uses a master/slave architecture where master consists of a single NameNode that manages the file system metadata and one or more slave DataNodes that store the actual data. Benefit of using Hadoop is distributed stor- age, Distributed Processing, Security, Reliability, Speed , Efficiency, Availability, Scalability and lots more. This is the reason of using Hadoop for tweet processing.

III. FLUME Apache Flume is a distributed, reliable, and available service for effi- ciently 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. After the installation of VMWARE and Hadoop for single node

*School of Computer Engineering, KIIT, BBSR* 1

*TWITTER SENTIMENT AND DATA ANALYSIS*

next step comes the installation of FLUME. For this you need to log in to twitter. After that go to apps on twitter and create a new application. After you agree with all terms and conditions you will got new application. Then set Consumer Key, Consumer Secret, Owner Key and Owner Secret ID. Now access token needs to be created.

Sentiment analysis of twitter data is providing an effective way to expose user opinion which is necessary for decision making in various streams. Twitter allows the user to post real time short messages called as tweets For each hashtag, there may be lots of tweets new tweets are generated every minute. So in order to handle so many tweets, we are using hadoop framework. So using hadoop, we analyze twitter data where cluster of positive, negative neutral sense will be formed. The Hadoop framework was designed to solve problems which had huge amount of data for processing. Hadoop is distributed processing framework, which is work on large unstructured data. It is also work on semi structured structured data. Twitter data is relatively unstructured data. So it is store using hadoop. Hadoop is write ones read many times architecture. It is not suitable for online transaction. It is divides large data into 128 MB data chunks. In hadoop, code is distributed over the slave machine result is send to master machine. It increases efficiency decreases data failure or data lost. Data is distributed to nodes using HDFS. HDFS is Hadoops distributed file system. Hadoop provides MAP Reduce framework which contains name node, secondary name node, data node HDFS file system which contains job tracker, task tracker. Map Reduce is used for processing HDFS file system is used for storage. With the use of sentiment analysis any industry can know the feedback of people about their product and can improve their quality of product.

*School of Computer Engineering, KIIT, BBSR* 2

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 2 Literature Survey

## 2.1 Background

Mahalakshmi R, Suseela [4] (2015) Big-SoSA: Social Data Analysis and Data Visualization on Big Data. It tries to provide a way of Twitter Data Analysis using Hadoop and its ecosystem processing large volume of data. MapReduce function used to perform Twitter data analysis.

Praveen Kumar, Dr. Vijay Singh Rathore [5] (2014) : Tried to show efficient capability of processing of big data. We use Map- Reduce and several other solutions to big data problem. Problem of big data championed by Google. Map Reduce algorithms, Pig and other architecture like Hive, promised benefits of distributed processing.

Sunil B. Mane, Yashwant Sawant, Saif Kazi [3] (2014) Real Time Data Analy- sis of Twitter Data Using Hadoop. Gives a way of data analysis by using Hadoop processing huge amount of data in Hadoop cluster.

Manoj Kumar Danthala [6] (2015) Tweet Analysis: Twitter Data processing Us- ing Apache Hadoop . Analysing big data like social media data or twitter data using Apache Hadoop to process and analyse different tweets in Hadoop cluster. It in- cludes visualizing results into pictorial representations.

Manoj Kumar Danthala [7] (2015) Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using Big Insights. Twitter data is proposed here, which is one of largest social networking sites . Here data goes on increasing at high rates everyday. InfoSphere Big Insights tool brings the power of Hadoop to execute data in real time . We also visualize big data charts using big sheets.

*School of Computer Engineering, KIIT, BBSR* 3

*TWITTER SENTIMENT AND DATA ANALYSIS*

Judith Sherin Tilsha S, Shobha M.S [8] (2015) A Survey on Twitter Data Analysis Techniques to Extract Public Opinion. We are using certain machine learning algo- rithms , it is a feature vector which is constructed using emotion describing words from tweets and passed to the classifier to classify the sentiment or opinion. Various data analysis techniques are based on dictionary.

*School of Computer Engineering, KIIT, BBSR* 4

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 3

**Software Requirements Specification**

## Product Perspective

The product is an analysis of streaming twitter data.

## User Classes and Characteristics

As per current working structure, the data is streamed from TWITTER website using FLUME received in the form of JSON file. Then JSON Serde is used to convert the JSON file into Hive table . Hive Query statement is used to analyze the streamed twitter data.

## Operating Environment

Any system with the below mentioned software tools would be able to run the system on Hadoop Platform.

## Software Interfaces

* + - Name: Cloudera
    - Database: HDFS
    - Operating System: Virtual Machine Linux
    - Libraries: Mappers Reducers
    - Scripting Language: HiveQL , JSON Serde File
    - Web technologies: HTML, materializeCSS , canvasJS

*School of Computer Engineering, KIIT, BBSR* 5

## Hardware Interfaces

*TWITTER SENTIMENT AND DATA ANALYSIS*

Our application is totally based on internet and hence , devices like Modem, WAN, LAN, Ethernet cross cable shall require a hardware interface for the system for con- necting to Internet.

* + - Screen Resolution 1280\*720 (PC Website)
    - Desktop System

## FUNCTIONAL REQUIREMENTS

### EXTRACTION OF DATA FROM TWITTER WEBSITE

The data is collected from the website of twitter.com . The data eventually col- lected is in the form of semi- structured data i.e., in JSON(JavaScript Object Nota- tion) format which is a light weighted, data-interchange format.

* + - * Human can easily read and write. It is easy for machines to parse and generate.
      * It is based on a subset of the JavaScript Programming Language.

Raw data is the data generated from twitter in JSON format using twitter API 1.1. The data has fields such as:

* + - * Name
      * Screen
      * Date Time
      * Text
      * Hashtag

These fields are generated when a user tweets or retweets .

*School of Computer Engineering, KIIT, BBSR* 6

*TWITTER SENTIMENT AND DATA ANALYSIS*

There are many other fields in the data for a particular record, which are not required for the analysis.



**Figure 3.1: Twitter Logo**

### JSON SERDE FILE

* + - * SerDe stands for Serializer/Deserializer.
      * Hive uses an interface for IO known as SerDe interface.
      * A SerDe used to read the data using deserializer and to write the data using serializer.
      * Here we are using SerDe for row format. For JSON files, Amazon has provided a JSON SerDe.
      * Raw data and JSON SerDe files are the external data
      * Hive uses external data and JSON SerDe file to load external tables
      * These external files are transmitted from windows to Hadoop environment, us- ing a win SCP recommended by Hortonworks
      * It is a interface to access remote system from local machine, and store files and data from an external resource
      * Here remote system is hortonworks sandbox and external resource is the exter- nal data

### ANALYSIS OF TWITTER DATA

* + - * HIVEQL: HiveQL is the Hive query language.

*School of Computer Engineering, KIIT, BBSR* 7

*TWITTER SENTIMENT AND DATA ANALYSIS*

Hive makes easy to implement Map- Reduce and joins. Hive optimizes the execution using partitions and buckets. Hive supports multiple file formats.

Hive doesnt care about the structure of data and doesnt enforce the schema on write.

Examples: Create TABLE sample table (name String, age int);

LOAD DATA LOCAL PATH input/mydata/data.txt INTO TABLE mytable;

Insert into birthday Select firstname, lastname, birthday from customers where birthday is NOT NULL;

Select \* from myTable;

* + - * Map Reduce Functions in Hadoop: MapReduce is a distributed programming model for processing large datasets . Map- Reduce is conceived by Google

. Map-Reduce can be implemented in any programming language like Scala, Python, Java ,C .Map-Reduce is not a programming language . Map- Reduce implemented by Hadoop. It manages communications, data transfer and paral- lel execution across distributed servers or nodes.

* + - * + ”Map” step: This step has a program called Mapper. Mapper is a program which is invoked by Hadoop Framework once for every record in the input split. Mapper assigns each key to a partition. Node applies Map function to local data and writes output to temporary storage.
        + ”Shuffle” step: It is divided into 3 parts: Sort, Copy and Merge. Nodes redistribute data on the basis of output keys.
        + ”Reduce” step: This step has a program called Reducer. Reducer is a program which is invoked by Hadoop Framework once for every record in the input split. Nodes now process each group of output data, per key, in parallel.

*School of Computer Engineering, KIIT, BBSR* 8

*TWITTER SENTIMENT AND DATA ANALYSIS*

* + - * Flume: To fetch data from various services and transport them to central storage

, we use Flume .For fetching Twitter data and store it in HDFS, a web server generates Log data collected by Flume. It is a channel that buffers data to a sink, and finally pushing it to HDFS.

To fetch Twitter data, we will have to follow the steps given below

* + - * + Creating a twitter Application
        + Installing / Starting HDFS
        + Configuring Flume

## NON FUNCTIONAL REQUIREMENTS

### Performance Requirements

System performance depends on the response time and data submission speed. The systems response time is direct and it is a real-time application . For better per- formance , response time of the system should be fast depending on the implemented algorithms efficiency.

### Software Quality Attributes

The graphical interface of the system should be simple and user-friendly. All the features of the website can be easily understood by the users. Action have to be performed using a few clicks.

*School of Computer Engineering, KIIT, BBSR* 9

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 4 Requirement Analysis

## PROBLEM STATEMENTS

* + - Total number of tweets for 10 minutes.
    - Most popular hashtags in tweets containing certain keywords.
    - Most followed users giving tweets related to certain keywords
    - Most retweeted users giving tweets related to certain keywords.

## REQUIREMENTS

The significant issues engaged with big data are the following:

The principal challenge confronted is storing and getting to the information from the expansive tremendous amount of data sets from the clusters. We require a stan- dard computing platform to manage substantial data since the data is developing, and data stores in various data stockpiling areas in a centralized system, which will scale down the gigantic data into sizable data for computing. The second test is re- covering the data from the substantial online networking data sets. In the situations where the data is developing day by day, it’s fairly hard to getting to the data from the extensive networks in the event that we need to do particular activity to be per- formed. The third challenge is on the algorithm design for handling the problems raised due to the large data volume and the dynamic characteristics of data. This paper is based on three modules which is finding and performing operation on the basis of social media data sets. The main goal of the project is analyzing and fetch- ing the Twitter IDs of the users whose statuses have been retweeted by most of the user whose tweets are analyzed. At First the system collects the tweets from the

*School of Computer Engineering, KIIT, BBSR* 10

*TWITTER SENTIMENT AND DATA ANALYSIS*

social network using the twitter APIs. Second, it consists of standard platform as Hadoop to solve the challenges of big data through MapReduce framework where the complete data is mapped with frequent datasets and reduced to smaller size data to make easy handling . Finally analyze the collected tweets and fetch the Twitter IDs of those users whose statuses have been retweeted by most of the user whose tweets are being analyzed.

Twitter Data and sentimental Analysis is very useful in social media as it help us to gain an overview of the wider public opinion on various topics. Social media monitoring tools like Brand-watch Analytics make that processing faster and easier. The applications of sentiment analysis are broad and powerful. The change in sen- timent on social media have been shown to compare with the change in the stock market.

The core objective of the project are as follows:-

1. Content Retrieval: Using java Twitter streaming API large amount of data is collected.
2. Storage: This data is stored in a HDFS format to form key value pair which is required to feed to mapper in map-reduce programming. The data is stored in Hadoop Distributed File System.
3. Data Processing: Using java and distributed processing software framework de- veloped by Apache Hadoop and using map reduce programming model and Apache hive framework. Data is collected over a period of time is processed.
4. Data Analysis: The output obtained from reducer phase is being analyzed.
5. Data Representation: Data are represented in the form of pie charts.
6. The end outcome will be in the form of classified tweets that is how many Positive, Negative and Neutral tweets related to certain user given topics.

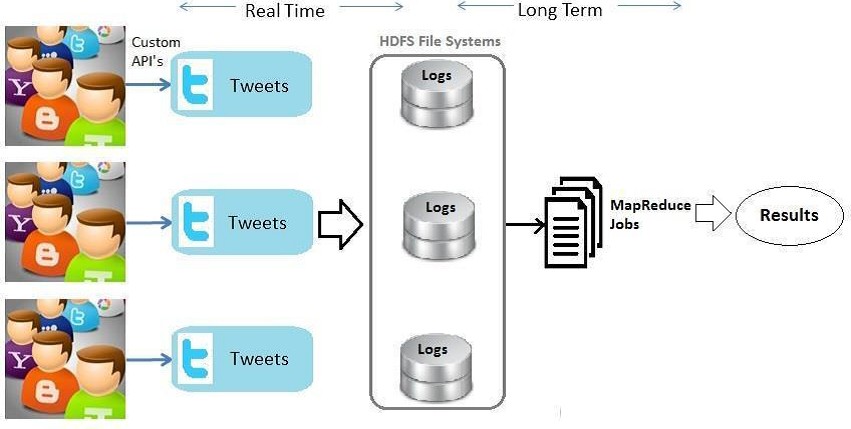
*School of Computer Engineering, KIIT, BBSR* 11

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 5 System Design

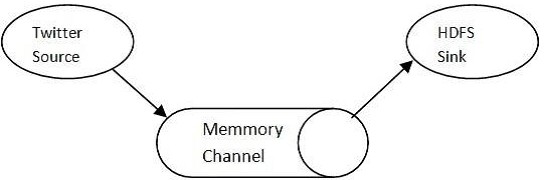
## SYSTEM ARCHITECTURE

### OVERALL ARCHITECTURE



**Figure 5.1: System Architecture**

### FLUME COMPONENTS



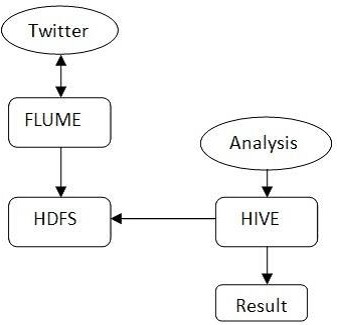
**Figure 5.2: Components of Flume**

*School of Computer Engineering, KIIT, BBSR* 12

*TWITTER SENTIMENT AND DATA ANALYSIS*

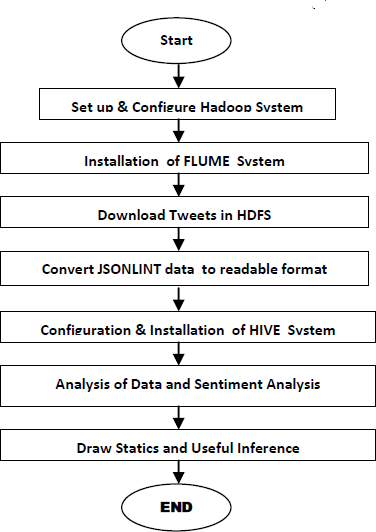
## WORKFLOW DIAGRAMS

### OVERALL WORKFLOW



**Figure 5.3: Workflow of proposed system**

### INSTALLATION TO END

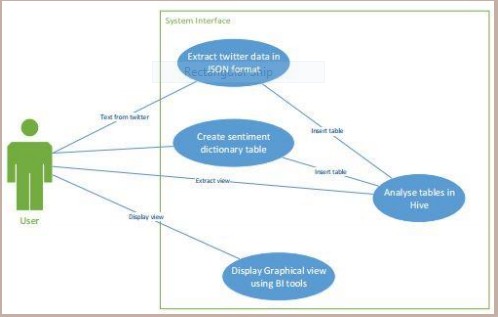


**Figure 5.4: Installation to end**

*School of Computer Engineering, KIIT, BBSR* 13

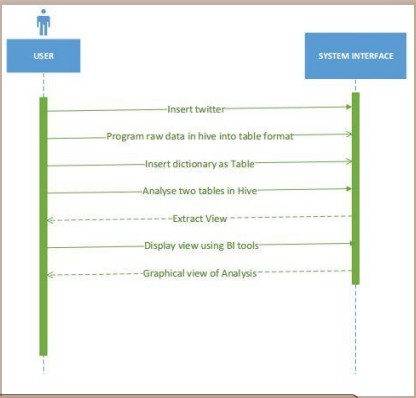
*TWITTER SENTIMENT AND DATA ANALYSIS*

### USECASE DIAGRAM OF TWITTER DATA ANALYSIS



**Figure 5.5: Use case Diagram**

### SEQUENCE DIAGRAM OF TWITTER DATA ANALYSIS



**Figure 5.6: Sequence Diagram for Twitter Data Analysis**

*School of Computer Engineering, KIIT, BBSR* 14

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 6 System Testing

## CONCLUSIVE TABLE

As per the experimental result analysis used the tweet in the sustem it gives the excellent result. Accuracy graph distinguishes between system 1 and system 2 by the percentage.

We have run the query in two different systems and compared the result accuracy.

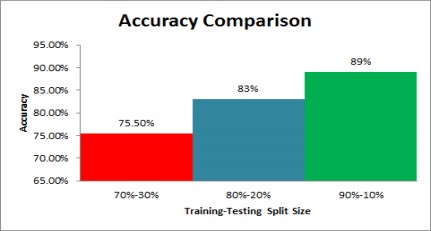
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **QUERY** | **TIME** | **SYSTEM 1** | **SYSTEM 2** | **COMPARISON** |
| No.of tweets | 10 am | 12866 | 12830 | Close match |
| Most popular hash-  tag | 10 am | ballandor | ballandOr | Almost match |
| Most followed user | 10 am | NDTV(10354577) | NDTV(10354577) | Exact match |

### Results are successful.

*School of Computer Engineering, KIIT, BBSR* 15

## RESULTANT GRAPH

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 6.1: Accuracy Graph**

## OUTPUTS

### Most popular hashtags in tweets containing certain keywords

**ballondor 206**

### datascience 187

**vols 185**

### datascientist 141

**hadoop 131**

### bigdata 119

**job 90**

### steelers 89

**fakeindian 77**

### probowlvote 72

**priyankachopra 65**

### machinelearning 60

**scientist 57**

*School of Computer Engineering, KIIT, BBSR* 16

*TWITTER SENTIMENT AND DATA ANALYSIS*

### Most followed users giving tweets related to certain keywords

|  |  |
| --- | --- |
| **Ndtv** | **10354577** |
| **Aajtak** | **6844209** |
| **IndiaToday** | **4587418** |
| **IndiaToday** | **4587062** |
| **CNNnews18** | **3809322** |
| **CNNnews18** | **3809254** |
| **ScienceNews** | **2701308** |
| **IndianExpress** | **2693886** |
| **businessinsider** | **2221203** |
| **businessinsider** | **2221202** |
| **businessinsider** | **2221188** |
| **KapanLagicom** | **2135951** |
| **nowthisnews** | **1557940** |
| **WorldAndScience** | **1504308** |
| **SAI** | **1446596** |
| **detikhot** | **1391350** |
| **GeorgeMentz** | **1269348** |
| **SAMAATV** | **1093712** |
| **SirJadeja** | **948815** |

### Most retweeted users giving tweets related to certain keywords.

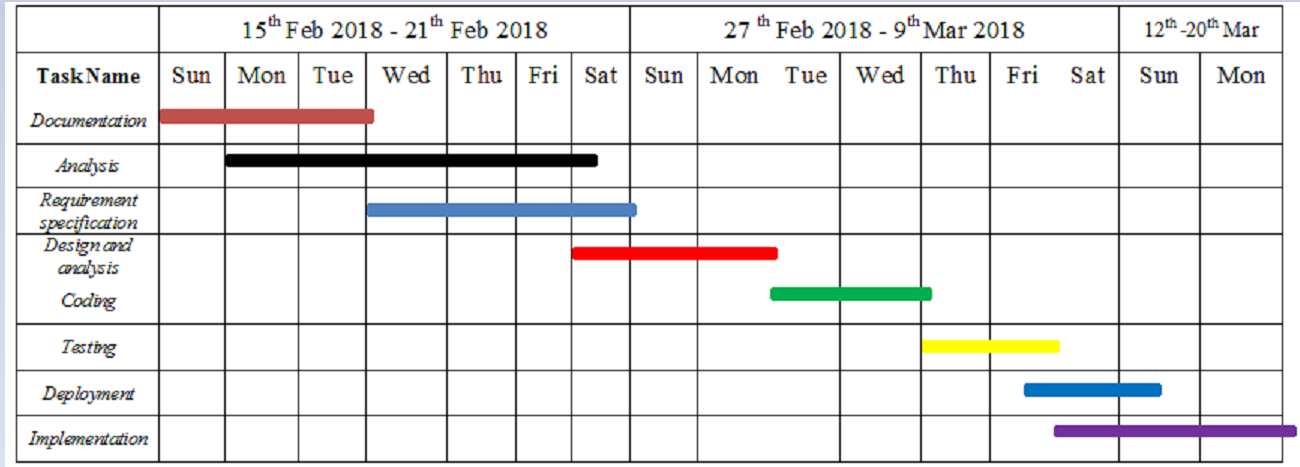
|  |  |  |
| --- | --- | --- |
| ukumar2008 | 5 | 6 |
| sree | 2 | 3 |
| reasonbruna | 1 | 3 |
| mini707070 | 0 | 2 |
| liberoj | 0 | 1 |
| jyotsana-khatri | 0 | 1 |

*School of Computer Engineering, KIIT, BBSR* 17

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 7 Project Planning

## 7.1 GANTT CHART



**Figure 7.1: Gantt chart**

*School of Computer Engineering, KIIT, BBSR* 18

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 8 Implementation

## SYSTEM IMPLEMENTATION

* + - Twitter is a social media platform where users can share their opinions by post- ing and sharing a variety of tweets.
    - Then all these tweets are being stored in twitter centralised database, there are millions of tweets that are posted everyday on twitter which is generating petabytes of data that are stored on data centre of the twitter.
    - For analysis we require these huge and extreme complex twitter and sentimen- tal data that contains variety of opinions that are being posted by different users, we use flume for fetch twitter data and that is being stored into HDFS, we can also create a twitter API using which we can extract real time twitter data from web and then store it into HDFS.
    - In the wake of storing these extensive and complex twitter JSON data we re- quire an examining tool to analyse these complex data, for these we uses hive which runs on top of the Hadoop and takes input from HDFS and its support SQL queries through which we would analysis be able to these data.
    - On the basis of hive analysis result, we can perform any desired queries.

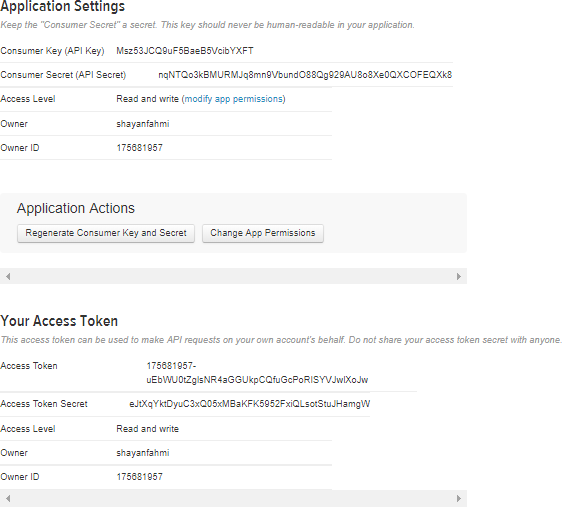
## Creating Twitter Application

To do opinion analysis on Twitter data, first we have to get Twitter data. For that, we have to create an account in Twitter developer mode. Now we login to our account, Go to Applications, Click on Create an application button provided by them which is shown below. Create the access tokens after creating a new application, so that we dont need to provide our authentication details there. It will be having one

*School of Computer Engineering, KIIT, BBSR* 19

*TWITTER SENTIMENT AND DATA ANALYSIS*

consumer key after creating application. It is used to access that application for getting Twitter data. To show clearly how the application data looks after creating the application, following is the figure shown. Here, the consumer details and also the access token details can be seen. We take these keys and token details and now want to set it in the Flume configuration file such that we can get the required data from the Twitter in the form of tweets.



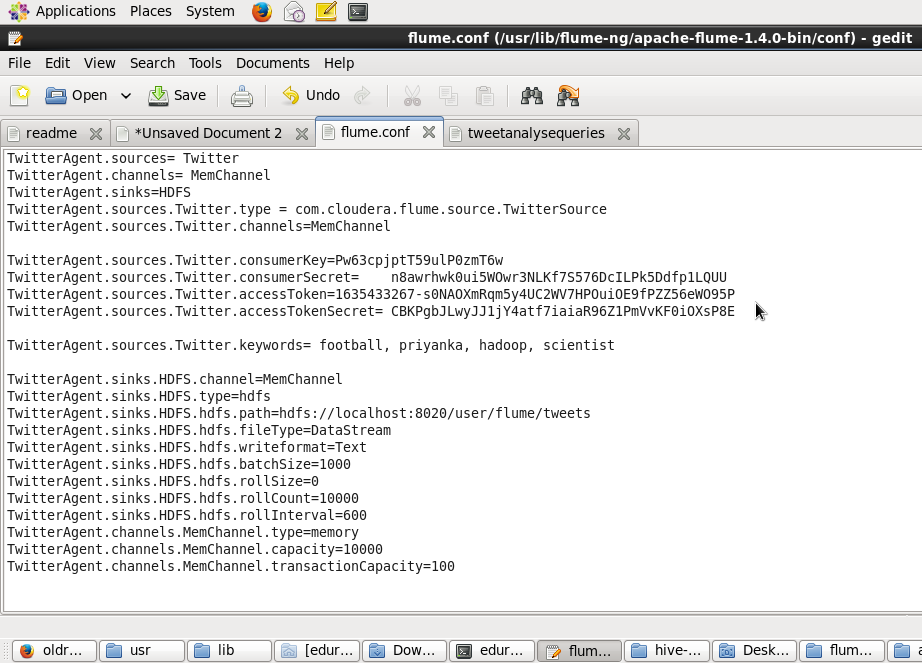
**Figure 8.1: Getting Consumer Key Access Token**

## Getting data using Flume

We want to use the consumer key and secret key along with the access token and secret values after creating an application in the Twitter developer site. Using them, we can access the Twitter and we can get the information that what we want and we will get everything in JSON format which gets stored in the HDFS in the location that we have given here. To get the Twitter data from the Twitter, the following is the configuration file that we want to use. We use flume.conf file shown in the figure to fill all the details.

*School of Computer Engineering, KIIT, BBSR* 20

*TWITTER SENTIMENT AND DATA ANALYSIS*

**Figure 8.2: Configuration**

While configuration of this file, configure sink as HDFS and set the path for stor- ing tweets in HDFS. Run the configuration file and tweets start downloading in HDFS in specified path. To do this, execute flume comments. Using the below command, we can start Flume :

* + - cd /usr/lib/flume-ng/apache-flume-1.4.0-bin/bin/
    - ./flume-ng agent -n TwitterAgent -c conf -f /usr/lib/flume-ng/apache-flume- 1.4.0-bin/conf/flume.conf

## Querying using Hive Query Language

To process structured data in Hadoop, we can use Apache Hive which is a data warehouse infrastructure tool. To summarize Big Data and makes querying and analyzing easy, it resides on top of Hadoop. For Analysis of data, Apache Hive

*School of Computer Engineering, KIIT, BBSR* 21

*TWITTER SENTIMENT AND DATA ANALYSIS*

(HiveQL) with Hadoop Distributed file System is used. SQL-like interface is pro- vided by Hive to process data stored in HDP. Hive is increasingly becoming the technology of choice for using Hadoop, beacause of its SQL-like interface. Build or download the JSON SerDe for setting up HIVE in Hadoop,. Before querying the data, we need to ensure that the Hive table can properly interpret the JSON data. Our Twitter data is in a JSON format but by default, Hive expects delimited row for- mat used by input files which doesnt work with the defaults. To specify that how to interpret what weve loaded, we can use the Hive SerDe interface. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process. From the root of the Git repository, we can build the hive-SerDes JAR:

* + - cd hive-serdes
    - mvn package cd ..

It will generate a file called hive-serdes-1.0- SNAPSHOT.jar in the target direc- tory. Now you need to create tweet table. Run hive, and execute the following commands:

ADD JAR /usr/lib/hive-0.13.1-bin/lib/hive-serdes-1.0-SNAPSHOT.jar;

1

CREATE EXTERNAL TABLE t w e e t s ( i d BIGINT ,

c r e a t e d a t STRING , s o u r c e STRING ,

f a v o r i t e d BOOLEAN, r e t w e e t c o u n t INT ,

r e t w e e t e d s t a t u s STRUCT*<* t e x t : STRING ,

‘ user ‘ : STRUCT*<*s c r e e n n a m e : STRING , name : STRING*>>*, e n t i t i e s STRUCT*<*

u r l s : ARRAY*<*STRUCT*<*e x p a n d e d u r l : STRING*>>*,

u s e r m e n t i o n s : ARRAY*<*STRUCT*<*s c r e e n n a m e : STRING , name : STRING*>>*, h a s h t a g s : ARRAY*<*STRUCT*<* t e x t : STRING*>>>*,

t e x t STRING ,

‘ user ‘ STRUCT*<*

s c r e e n n a m e : STRING , name : STRING ,

f r i e n d s c o u n t : INT ,

f o l l o w e r s c o u n t : INT , s t a t u s e s c o u n t : INT , v e r i f i e d : BOOLEAN,

u t c o f f s e t : INT ,

t i m e z o n e : STRING*>* ,

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

*School of Computer Engineering, KIIT, BBSR* 22

*TWITTER SENTIMENT AND DATA ANALYSIS*

24

i n r e p l y t o s c r e e n n a m e STRING )

ROW FORMAT SERDE ’ com . c l o u d e r a . h i v e . s e r d e . JSONSerDe ’ LOCATION ’ / u s e r / f lume / t w e e t s ’ ;

25

26

Now you have your data in relational form which can be easily analyzed. Actually this data looks in relational form bur is not. Data is analyzed using Map-Reduce form. Now Queries can be fired on this data for analysis as follows:

* + - Most popular hashtags in tweets containing certain keywords:

SELECT LOWER( h a s h t a g s . t e x t ) ,COUNT( ∗ ) AS t o t a l c o u n t FROM t w e e t s LATERAL

VIEW EXPLODE( e n t i t i e s . h a s h t a g s ) t 1 AS h a s h t a g s GROUP BY LOWER( h a s h t a g s . t e x t ) ORDER BY t o t a l c o u n t DESC LIMIT 1 5 ;

1

* + - Most followed users giving tweets related to certain keywords:

SELECT u s e r . screen name , u s e r . f o l l o w e r s c o u n t c FROM t w e e t s ORDER BY c d es c ;

1

* + - Most retweeted users giving tweets related to certain keywords:

SELECT t . r e t w e e t e d s c r e e n n a m e , sum ( r e t w e e t s ) AS t o t a l r e t w e e t s , c o u n t

( ∗ ) AS t w e e t c o u n t FROM ( SELECT r e t w e e t e d s t a t u s . u s e r . s c r e e n n a m e as

r e t w e e t e d s c r e e n n a m e , r e t w e e t e d s t a t u s . t e x t , max ( r e t w e e t c o u n t ) as r e t w e e t s FROM t w e e t s GROUP BY r e t w e e t e d s t a t u s . u s e r . screen name ,

r e t w e e t e d s t a t u s . t e x t ) t GROUP BY t . r e t w e e t e d s c r e e n n a m e ORDER BY t o t a l r e t w e e t s DESC LIMIT 1 0 ;

1

## SENTIMENT ANALYSIS

### STEPS FOR PERFORMING SENTIMENT ANALYSIS

* + - * For sentiment analysis,we need the tweet id and tweet text, so we will create a Hive table that will extract the id and tweet text from the tweets using the Cloudera Json serde:-

1

c r e a t e e x t e r n a l t a b l e l o a d t w e e t s ( i d BIGINT , t e x t STRING ) ROW FORMAT SERDE ’ com . c l o u d e r a . h i v e . s e r d e . JSONSerDe ’ LOCATION ’ / u s e r / f lume / t w e e t s 1 ’ ;

*School of Computer Engineering, KIIT, BBSR* 23

*TWITTER SENTIMENT AND DATA ANALYSIS*

* + - * We will split the text into words using the split() UDF available in Hive. If we use the split() function to split the text as words, it will return an array of values. So, we will create another Hive table and store the tweet id and the array of words:-

1

c r e a t e t a b l e words ( i d b i g i n t , words a r r a y *<*s t r i n g *>*) ;

i n s e r t o v e r w r i t e t a b l e words s e l e c t i d , s p l i t ( t e x t , ’ ’ ) from l o a d t w e e t s ;

2

* + - * We will create another table which can store id and word.n general, explode UDTF has some limitations; explode cannot be used with other columns in the same select statement. So we will add LATERAL VIEW in conjunction with explode so that the explode function can be used in other columns as well.The array of values has been converted into a string:-

1

c r e a t e t a b l e t w e e t w o r d ( i d b i g i n t , word s t r i n g ) ;

i n s e r t o v e r w r i t e t a b l e t w e e t w o r d s e l e c t i d as id , word from words LATERAL VIEW e x p l o d e ( words ) w as word ;

2

* + - * Lets use a dictionary called AFINN to calculate the sentiments. AFINN is a dictionary which consists of 2500 words rated from +5 to -5 depending on their meaning. We will create a table to load the contents of AFINN dictionary:-

1

c r e a t e t a b l e d i c t i o n a r y ( word s t r i n g , r a t i n g i n t ) ROW FORMAT DELIMITED

LOAD DATA LOCAL INPATH ’ / home / e d u r e k a / AFINN . t x t ’ i n t o TABLE d i c t i o n a r y ;

FIELDS TERMINATED BY ’ \ t ’ ;

2

* + - * We will join the tweet word table and dictionary table so that the rating of the word will be joined with the word.The rating column has been added along with the id and the word. Whenever there is a match with the word of the tweet in the dictionary, the rating will be given to that word else NULL will be present:-

1

c r e a t e t a b l e w o r d j o i n ( i d b i g i n t , word s t r i n g , r a t i n g i n t ) ;

i n s e r t o v e r w r i t e t a b l e w o r d j o i n s e l e c t t w e e t w o r d . id , t w e e t w o r d . word , d i c t i o n a r y . r a t i n g from t w e e t w o r d LEFT OUTER JOIN d i c t i o n a r y ON(

t w e e t w o r d . word= d i c t i o n a r y . word ) ;

2

* + - * We will perform the groupby operation on the tweet id so that all the words of one tweet will come to a single place. And then, we will be performing an Average operation on the rating of the words of each tweet so that the average

*School of Computer Engineering, KIIT, BBSR* 24

*TWITTER SENTIMENT AND DATA ANALYSIS*

rating of each tweet can be found. We will calculate the average rating of each tweet by using each word of the tweet and arranging the tweets in the descending order as per their rating.This entire output will be stored in the table called sentiment so that necessary analysis queries can be easily executed on it:-

c r e a t e t a b l e s e n t i m e n t ( i d b i g i n t , r a t i n g i n t ) ;

i n s e r t o v e r w r i t e t a b l e s e n t i m e n t s e l e c t id ,AVG( r a t i n g ) as r a t i n g from w o r d j o i n GROUP BY w o r d j o i n . i d o r d e r by r a t i n g DESC ;

1

2

* + - * We will then count the number of tweets ids with rating greater than 0 , rating less than 0 and rating equal to 0 to find the total number of positive,negative and neutral tweets related to the user-given topic/s respectively:-

1

s e l e c t c o u n t ( c a s e when r a t i n g *>*0 t h e n r a t i n g e l s e n u l l end ) p o s i t i v e , c o u n t ( c a s e when r a t i n g *<*0 t h e n r a t i n g e l s e n u l l end ) n e g a t i v e , c o u n t (

c a s e when r a t i n g =0 t h e n r a t i n g e l s e n u l l end ) n e u t r a l from s e n t i m e n t

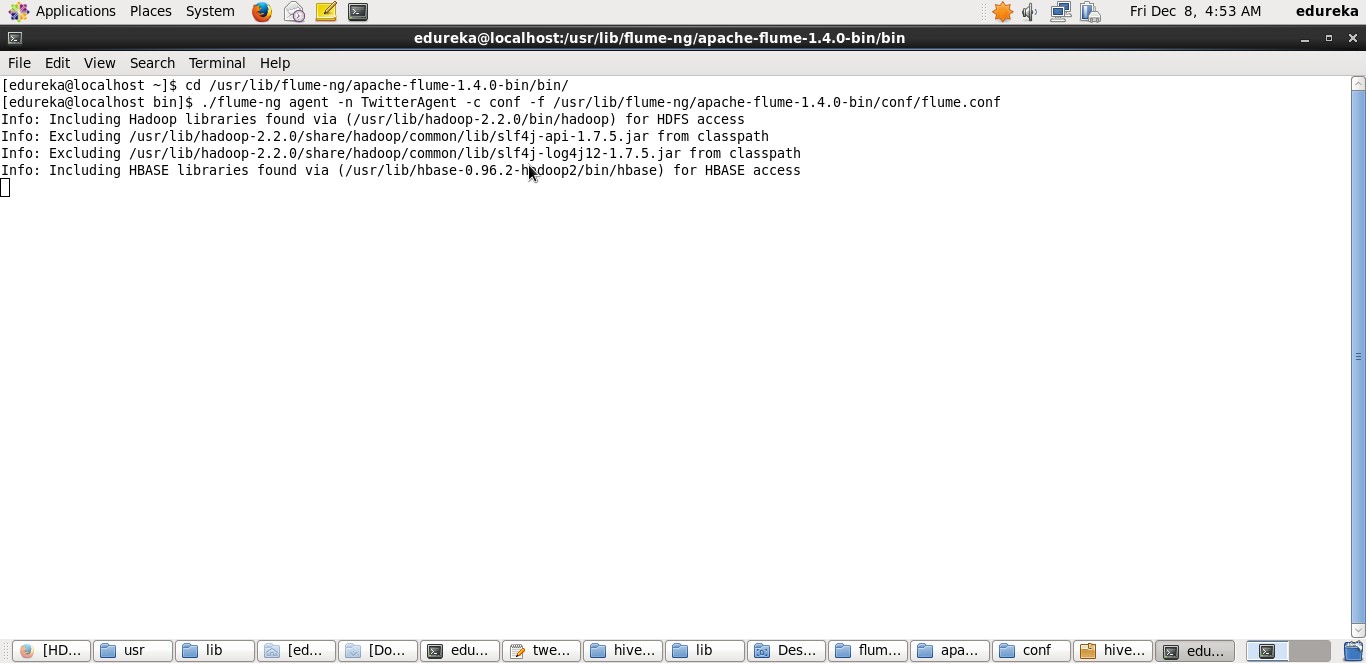
;

*School of Computer Engineering, KIIT, BBSR* 25

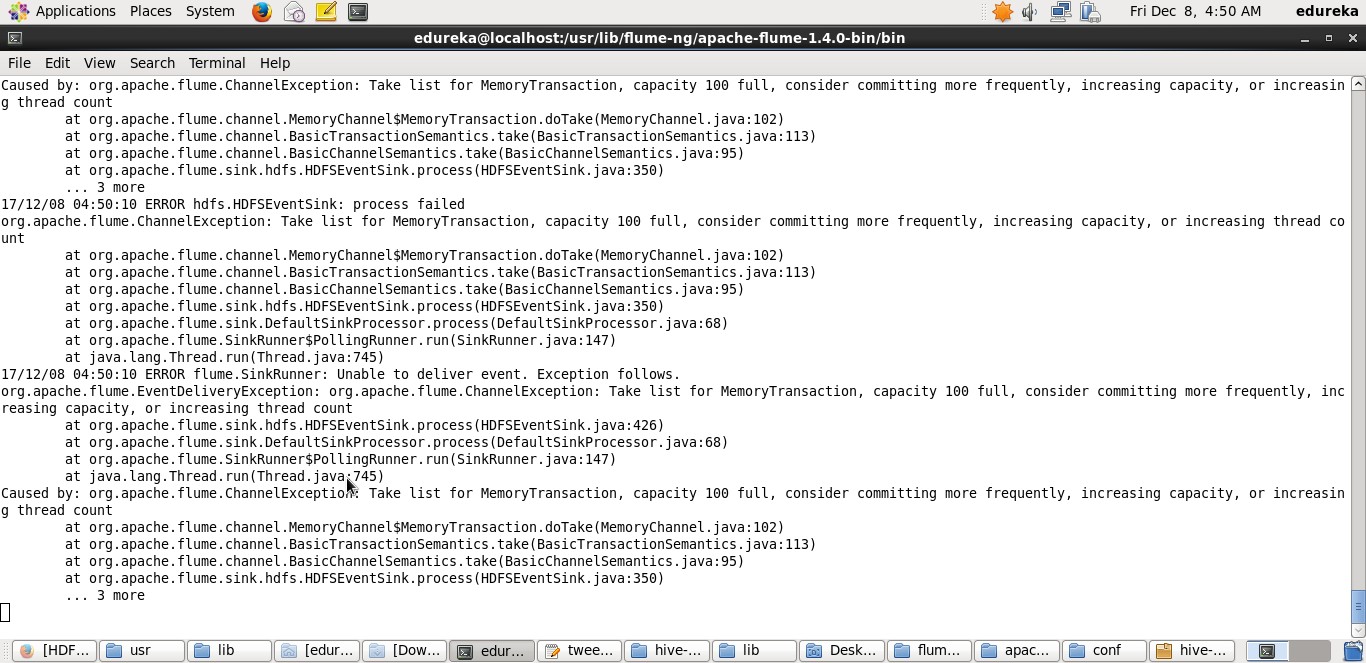
*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 9 Screenshots of Project

## HADOOP SNIPPETS



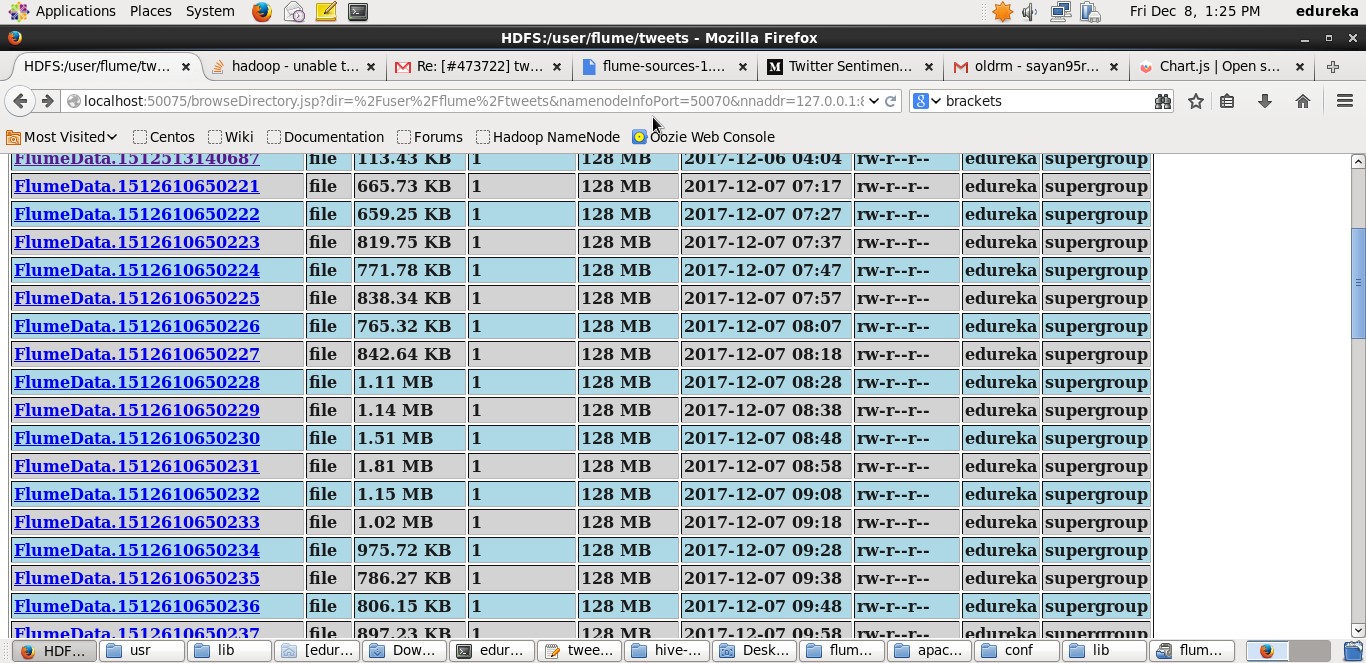
**Figure 9.1: Twitter Data streaming Through Flume - commands**



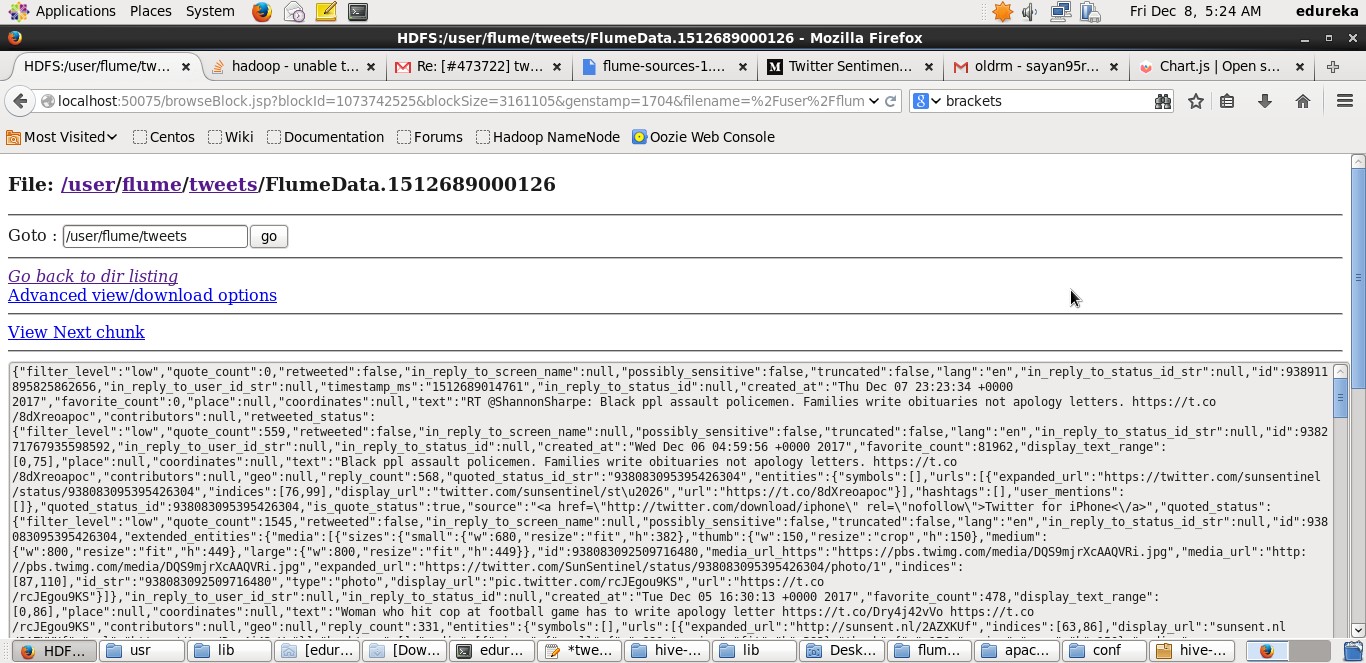
**Figure 9.2: Twitter Data streaming Through Flume**

*School of Computer Engineering, KIIT, BBSR* 26

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.3: Flume Data blocks of maximum 1000 tweets**



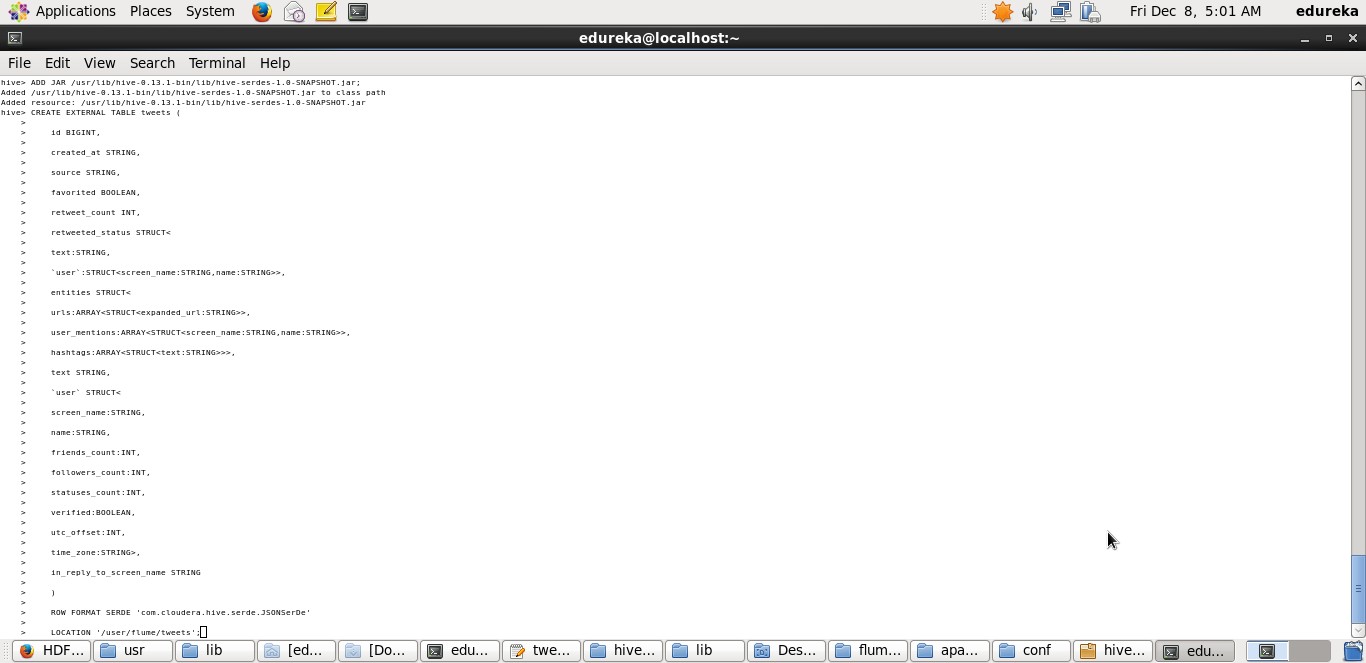
**Figure 9.4: Twitter data in JSON format**

*School of Computer Engineering, KIIT, BBSR* 27

*TWITTER SENTIMENT AND DATA ANALYSIS*



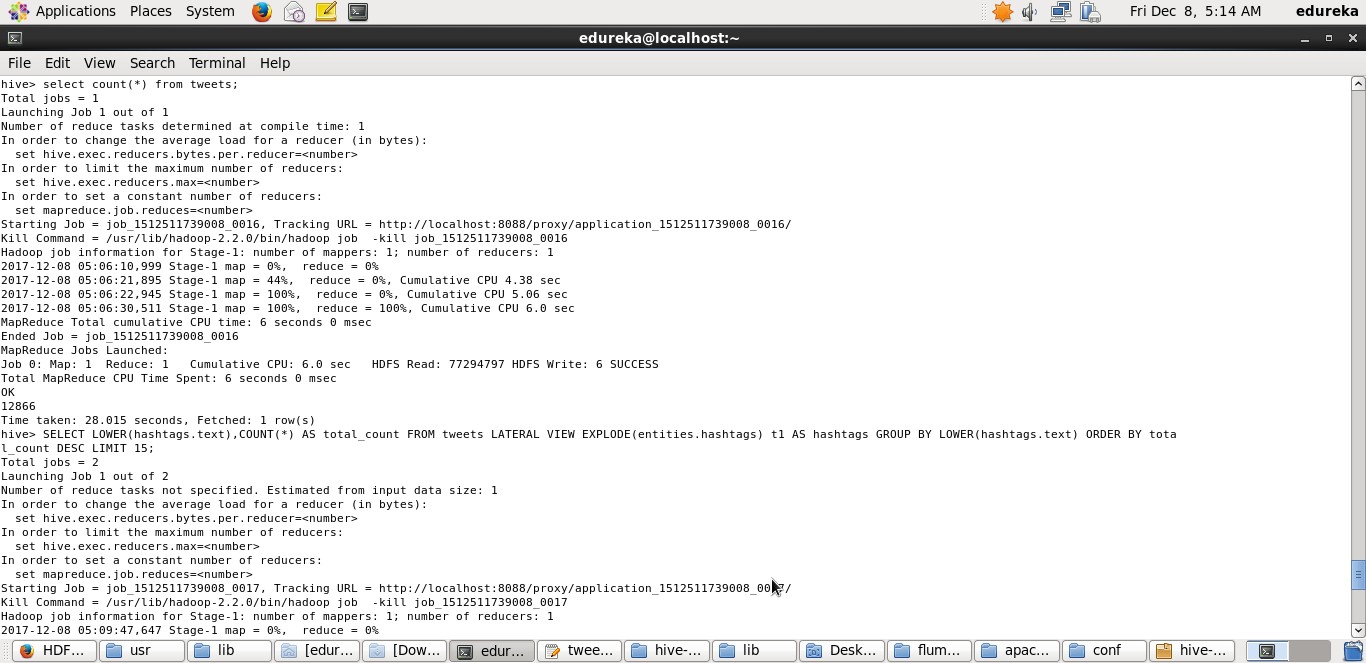
**Figure 9.5: Twitter data streaming completed**



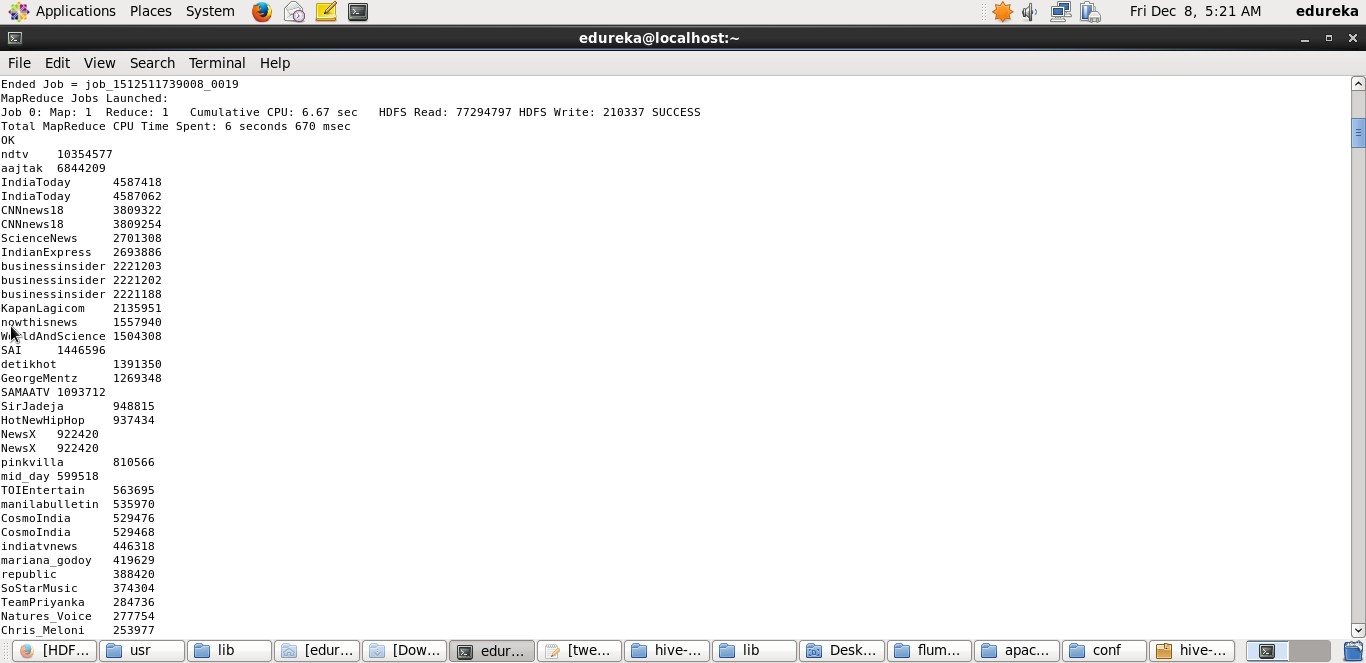
**Figure 9.6: Table Creation query**

*School of Computer Engineering, KIIT, BBSR* 28

*TWITTER SENTIMENT AND DATA ANALYSIS*



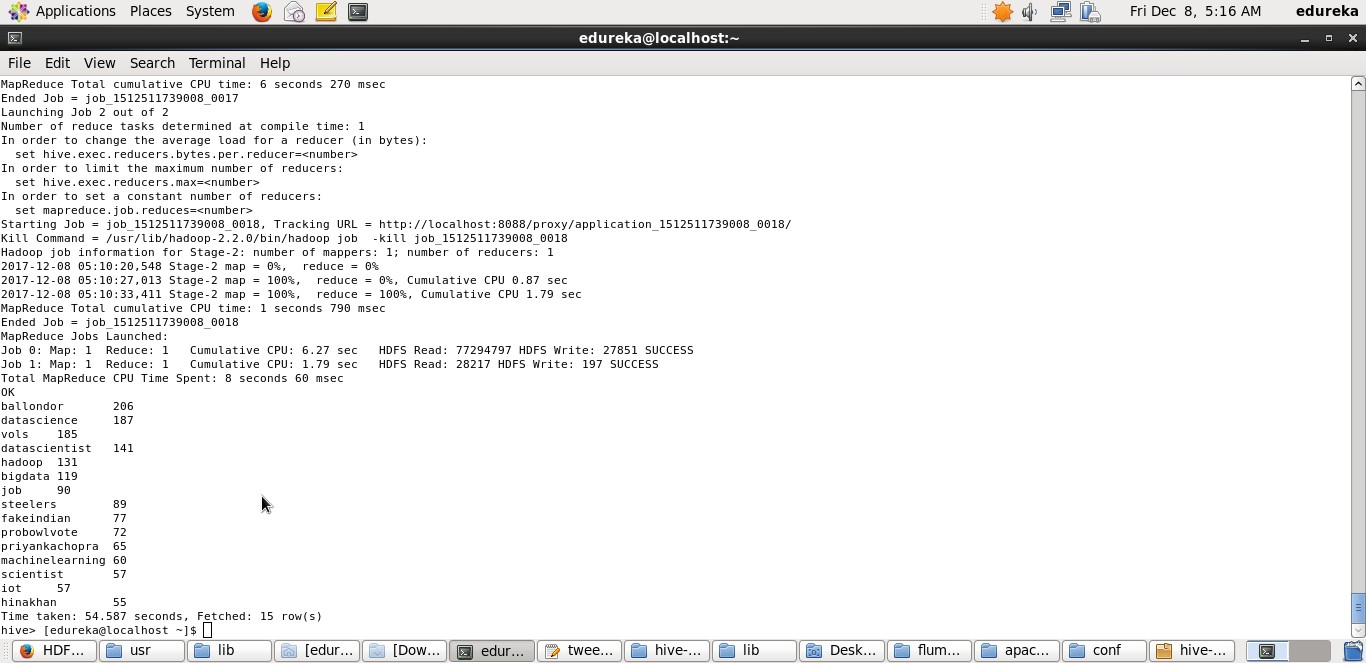
**Figure 9.7: Query Internally Converting to MR jobs**



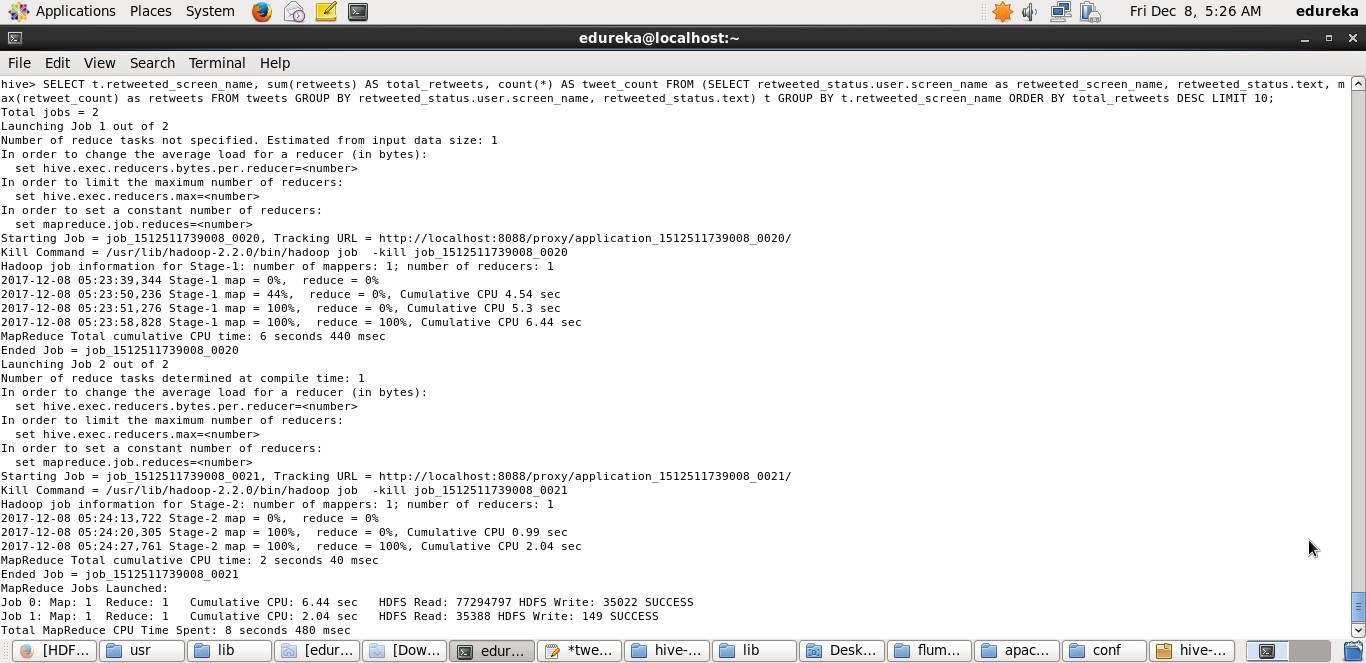
**Figure 9.8: Query showing most followed users**

*School of Computer Engineering, KIIT, BBSR* 29

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.9: Query showing most popular hashtags**

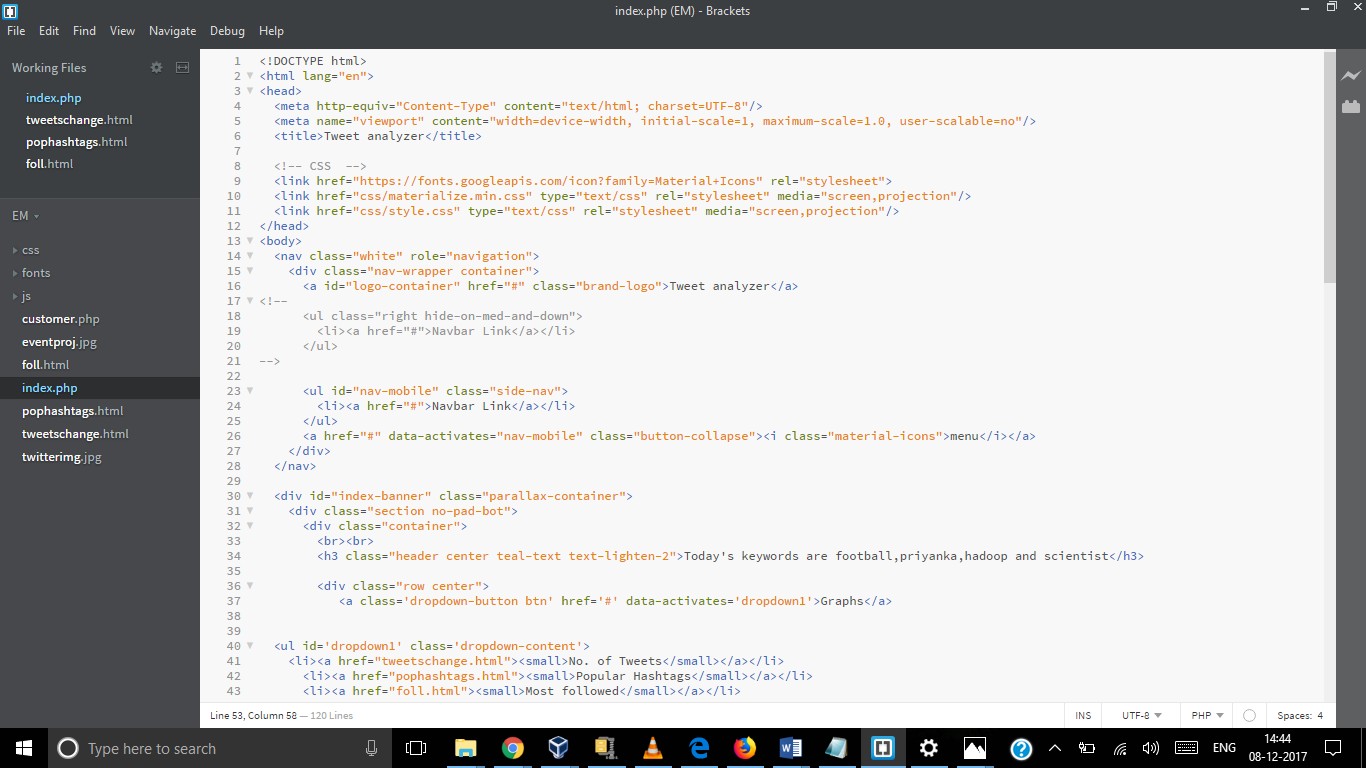


**Figure 9.10: Query for finding most retweeted users**

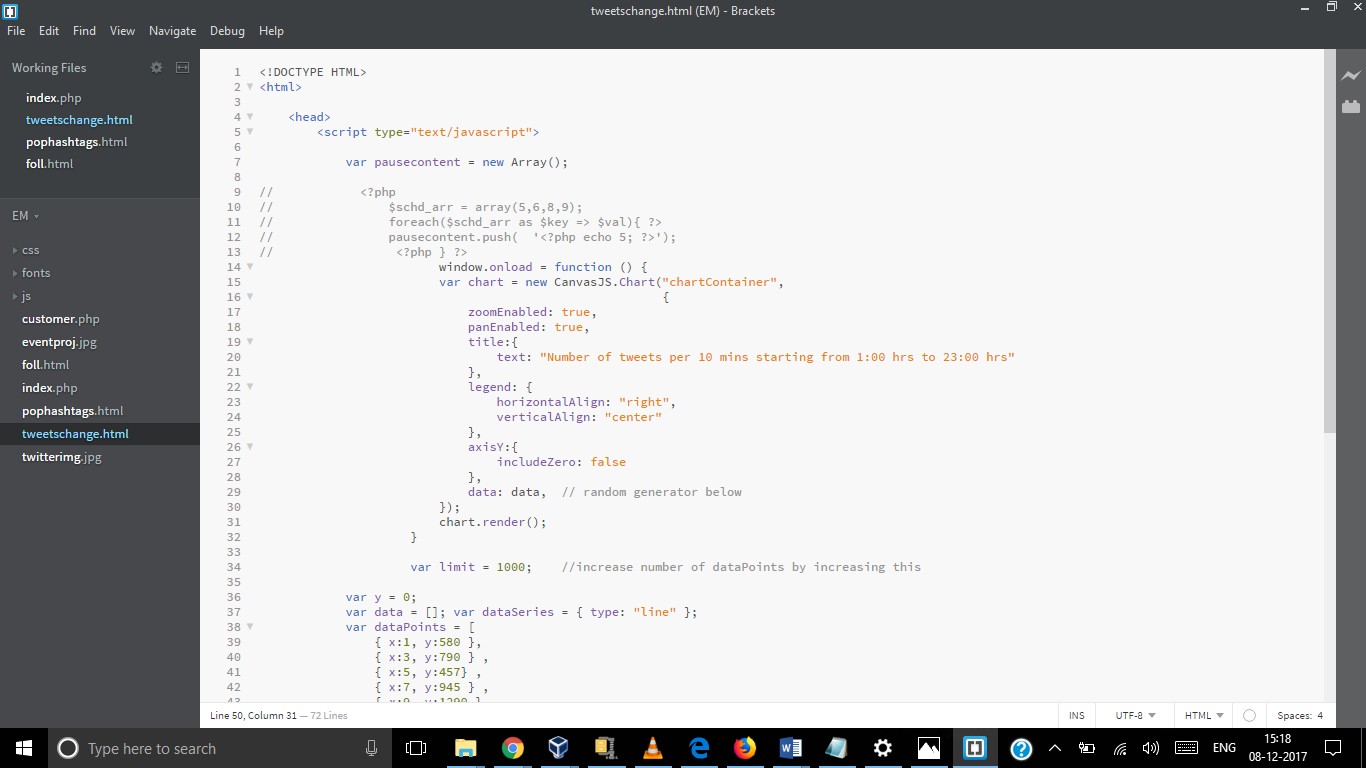
*School of Computer Engineering, KIIT, BBSR* 30

## Web Interface Snippets

*TWITTER SENTIMENT AND DATA ANALYSIS*



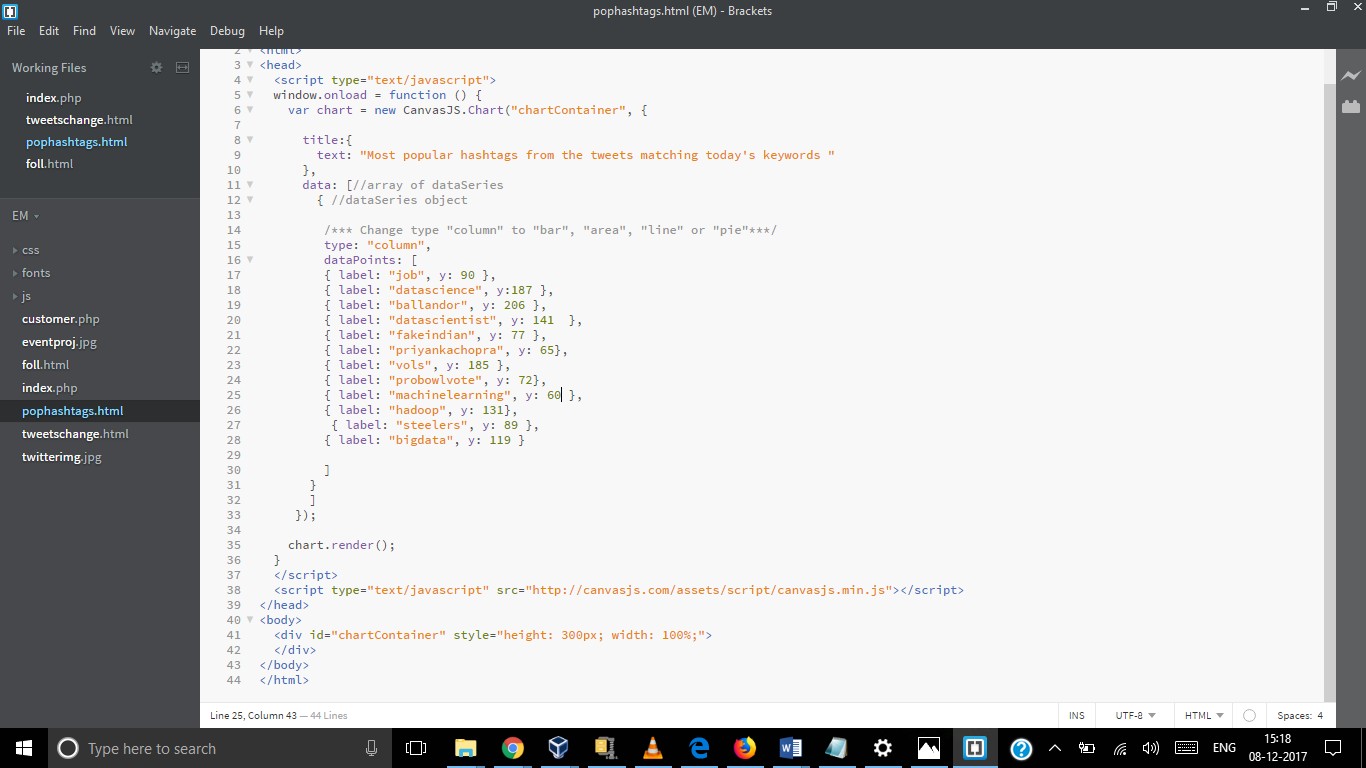
**Figure 9.11: HTML code for the index page**



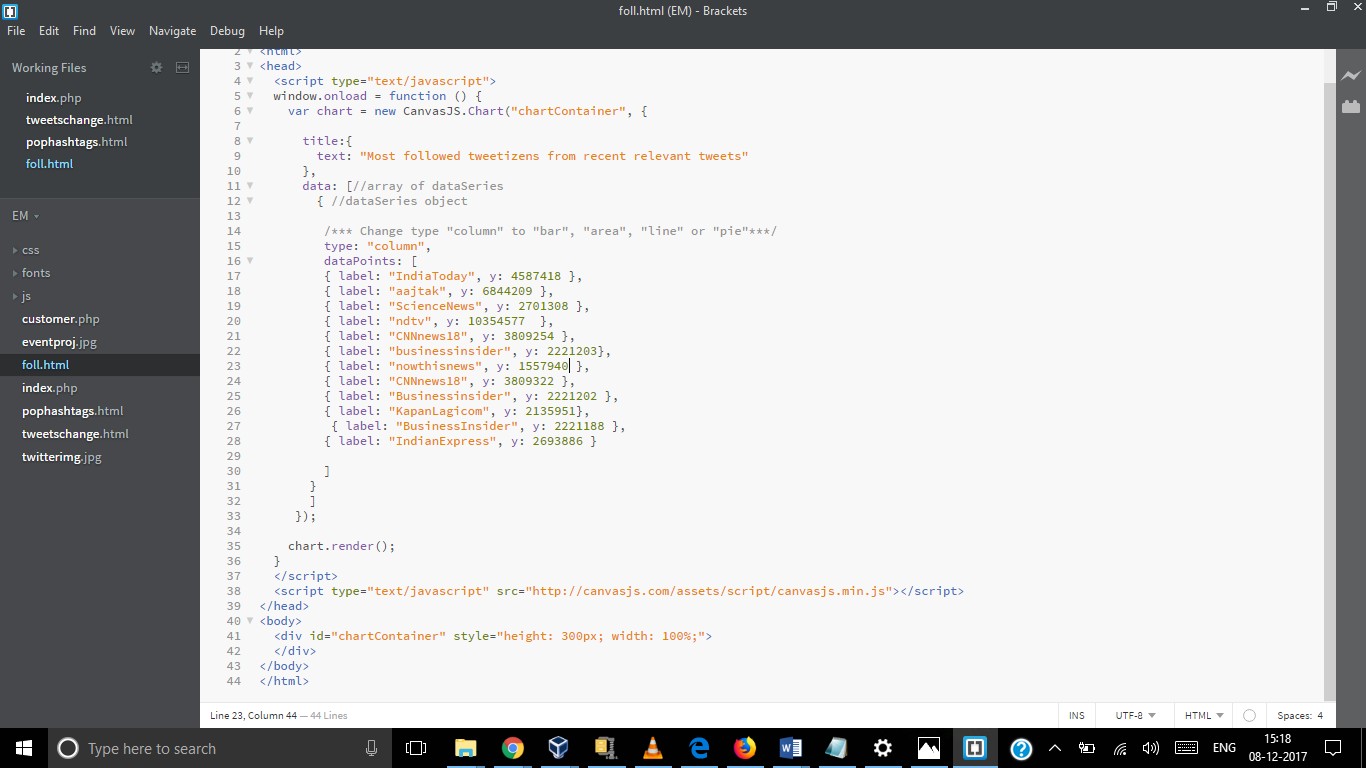
**Figure 9.12: JS code for showing change in number of tweets**

*School of Computer Engineering, KIIT, BBSR* 31

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.13: JS code for showing most popular hashtags**



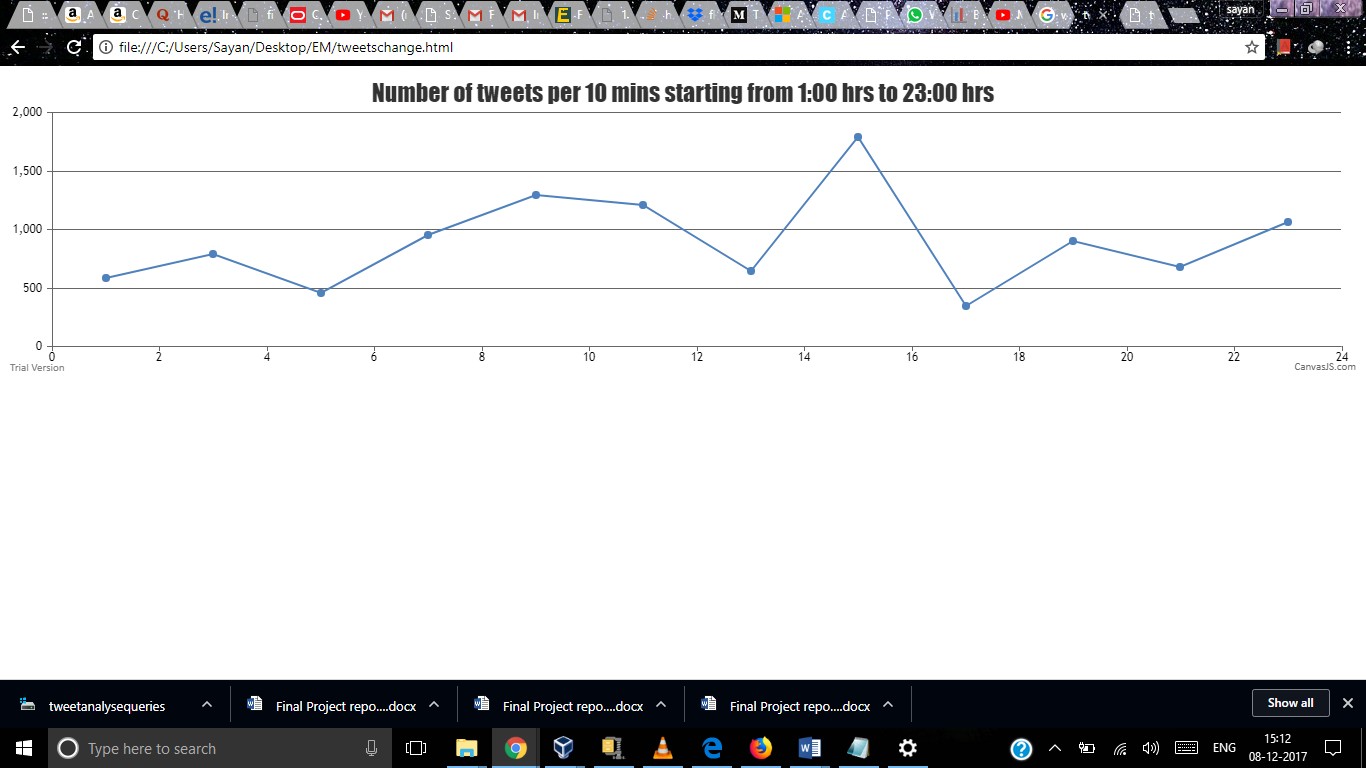
**Figure 9.14: JS code for showing most followed users**

*School of Computer Engineering, KIIT, BBSR* 32

*TWITTER SENTIMENT AND DATA ANALYSIS*



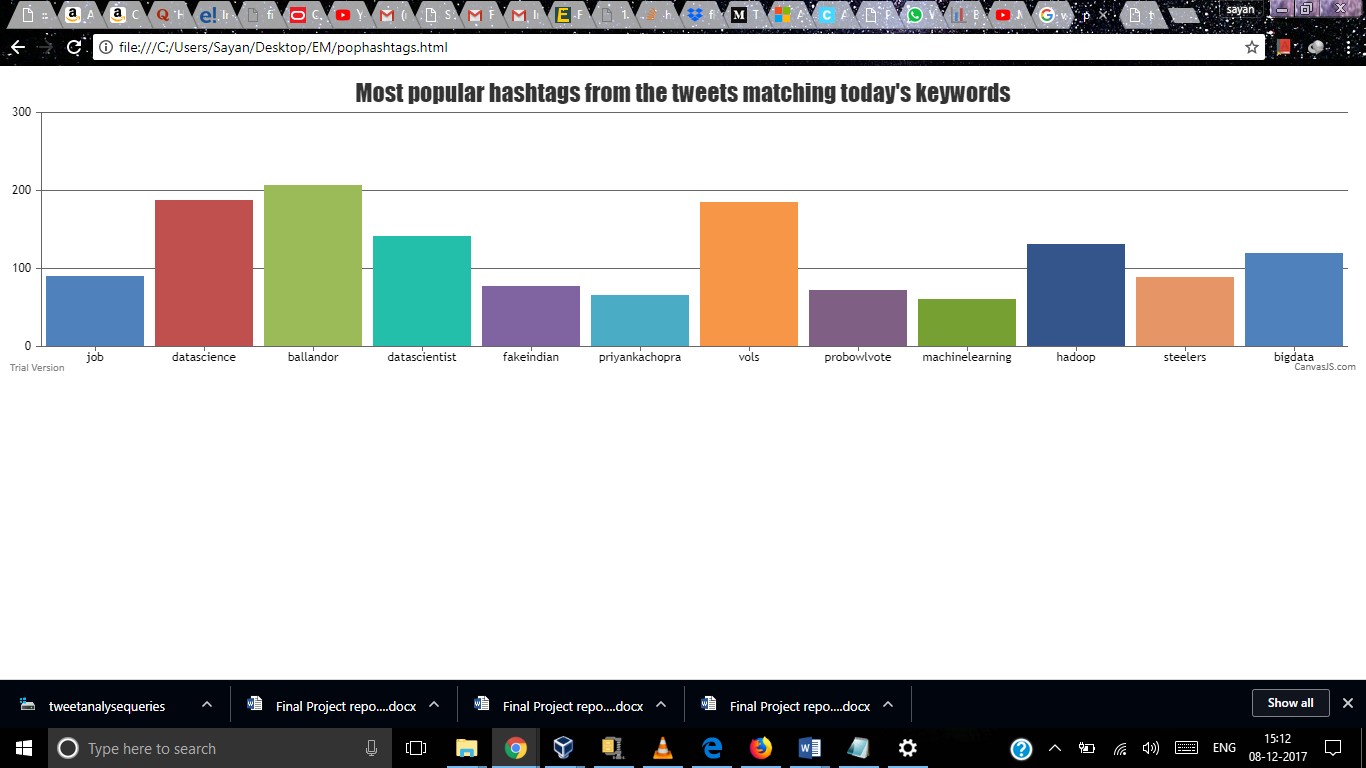
**Figure 9.15: Index Page**



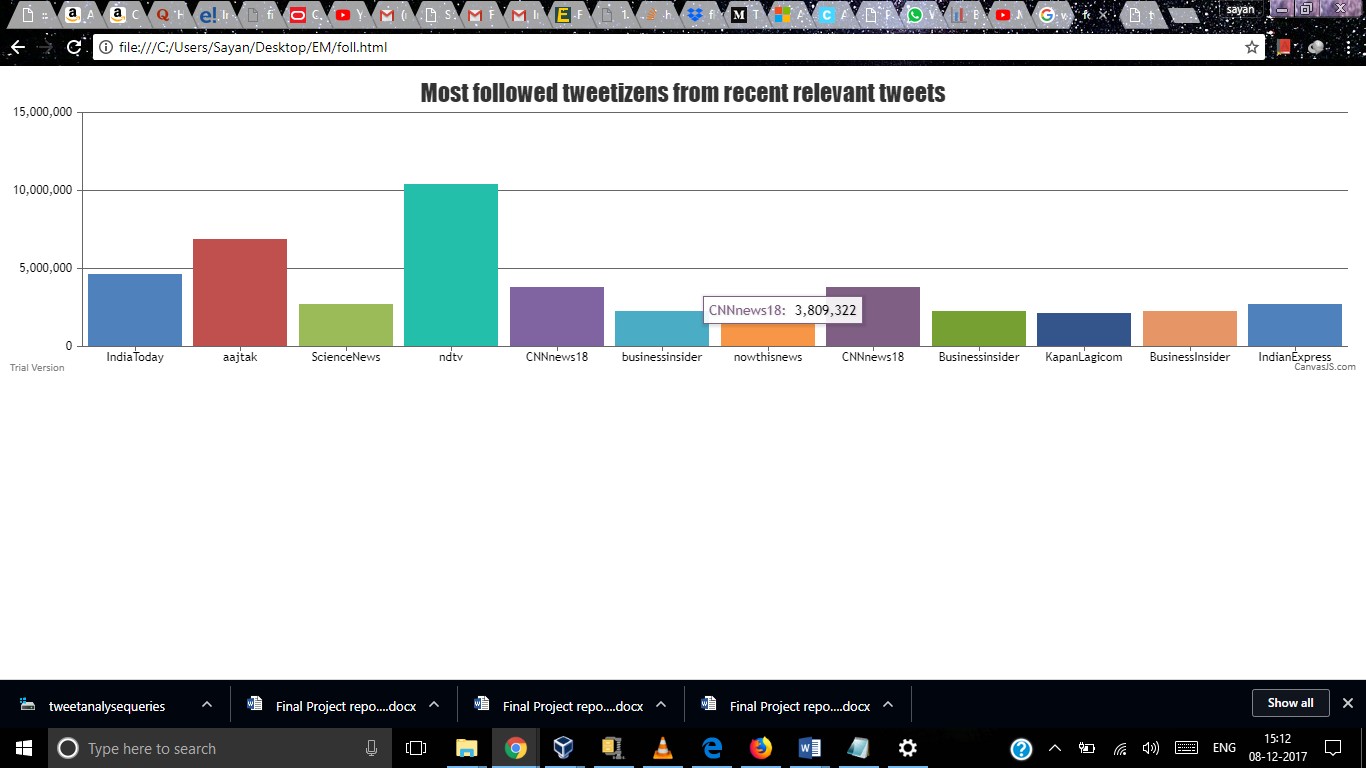
**Figure 9.16: Change in number of tweeets**

*School of Computer Engineering, KIIT, BBSR* 33

*TWITTER SENTIMENT AND DATA ANALYSIS*



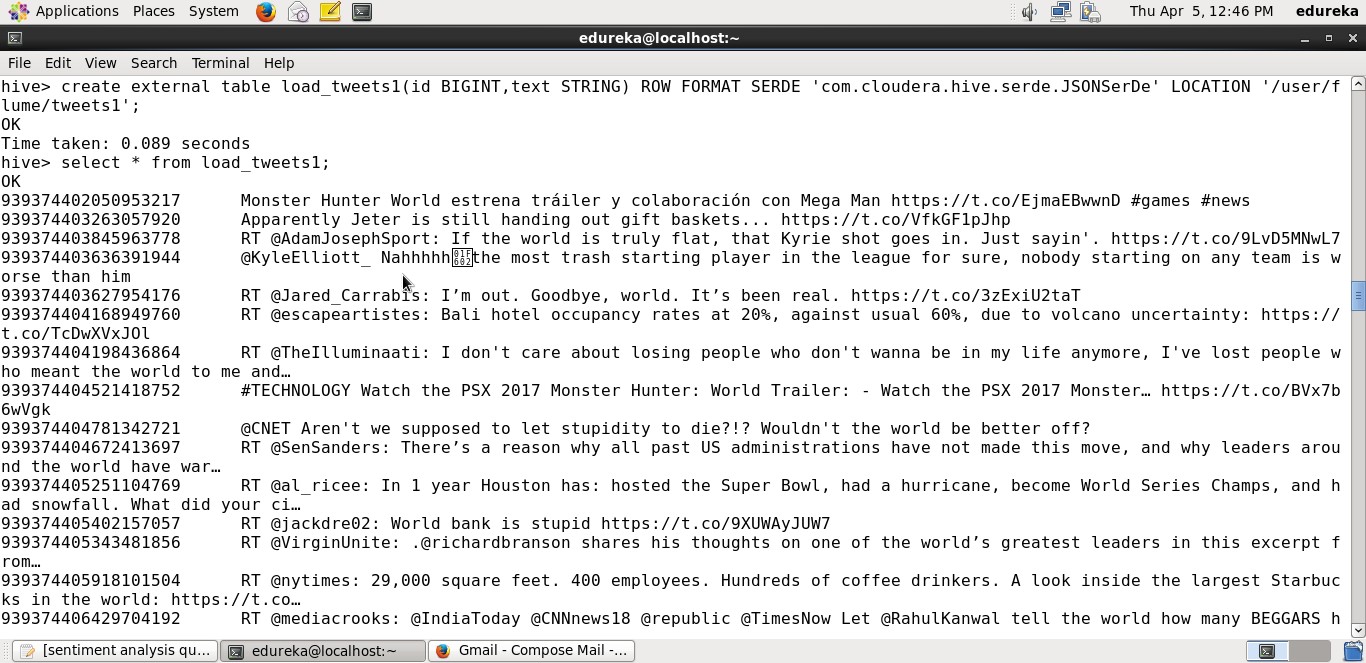
**Figure 9.17: Most popular hashtags**



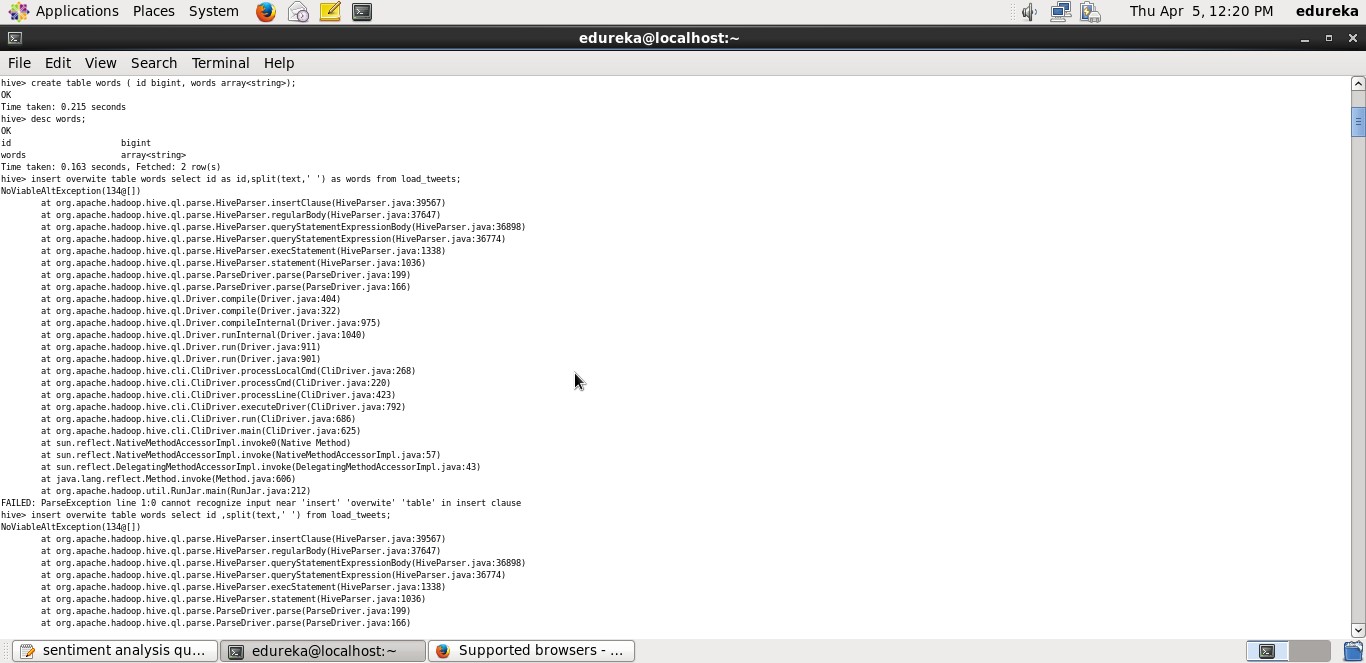
**Figure 9.18: Most followed users**

*School of Computer Engineering, KIIT, BBSR* 34

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.19: Hive table that will extract the id and tweet text from the tweets using the Cloudera JSON serde**



**Figure 9.20: Word table containing the array of strings**

*School of Computer Engineering, KIIT, BBSR* 35

*TWITTER SENTIMENT AND DATA ANALYSIS*



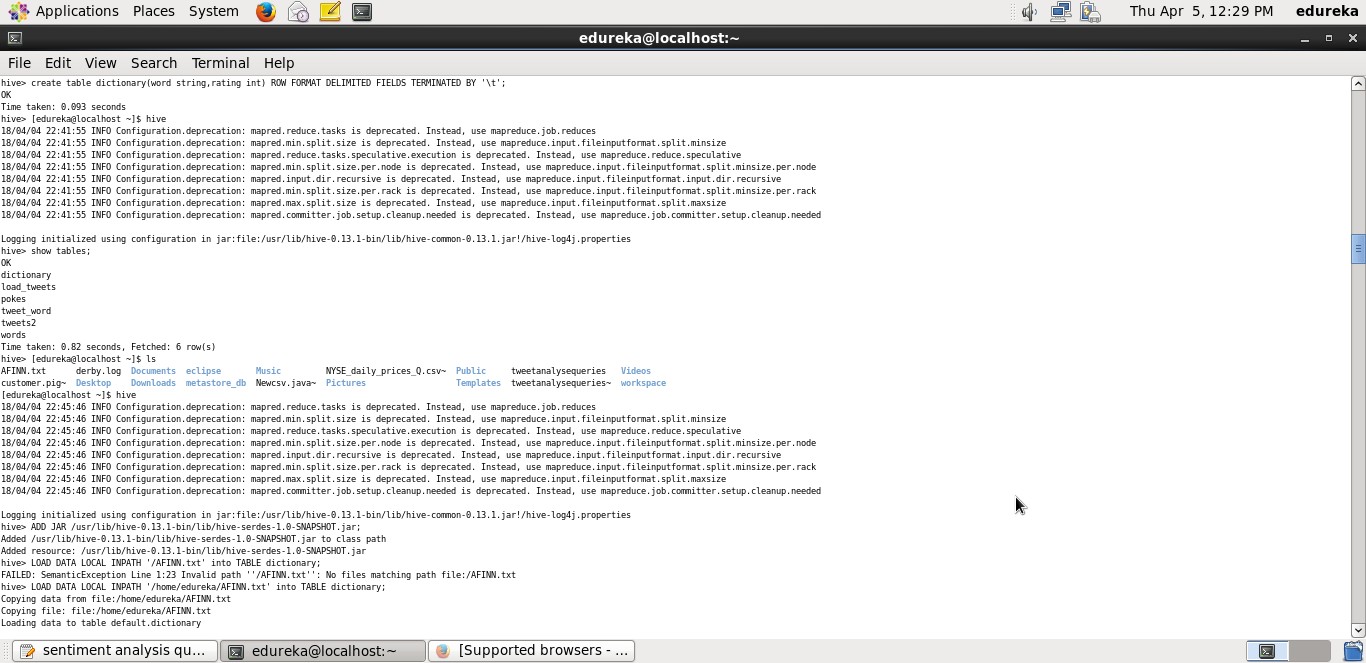
**Figure 9.21: use the split() function to split the text as words, it will return an array of values**



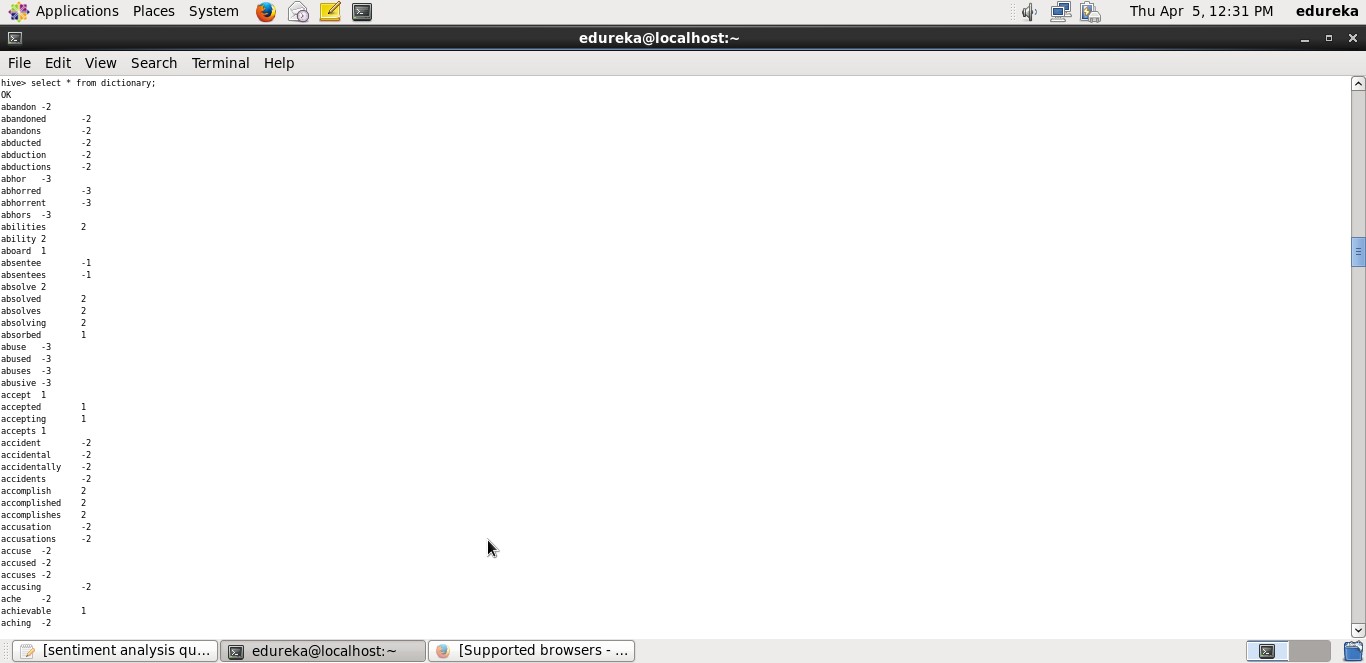
**Figure 9.22: We will create another table called tweet word which can store id and word in general**

*School of Computer Engineering, KIIT, BBSR* 36

*TWITTER SENTIMENT AND DATA ANALYSIS*



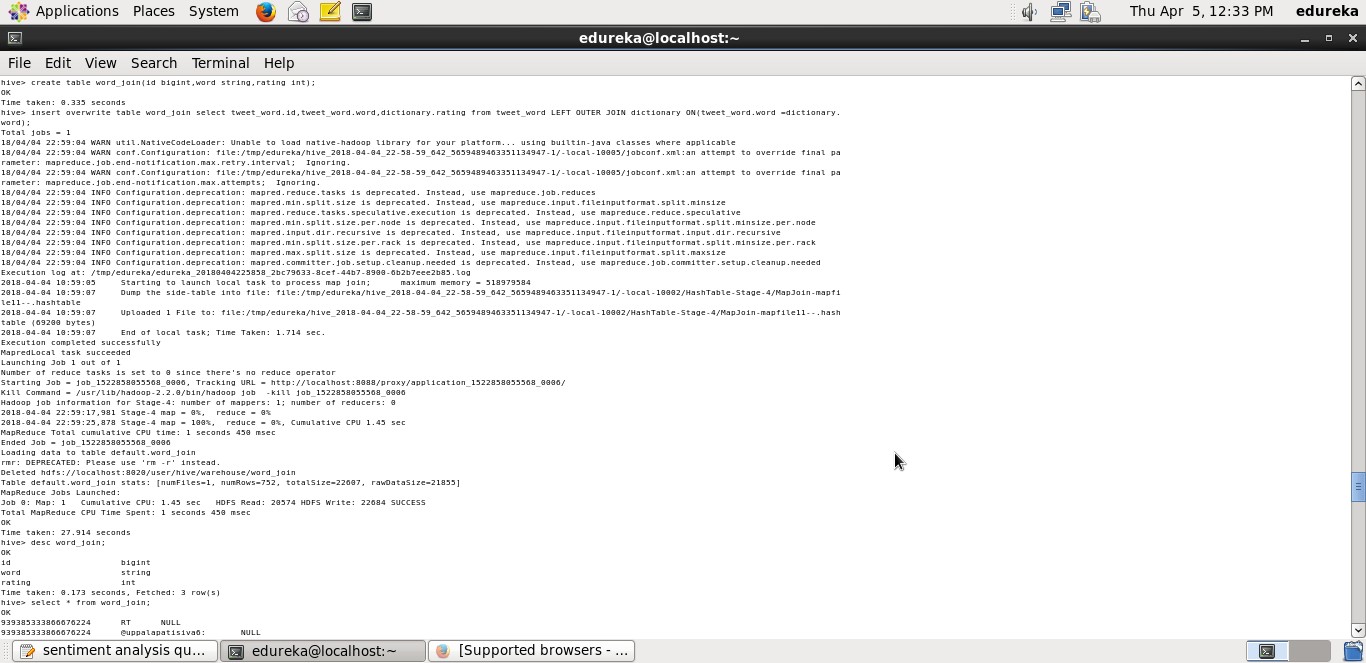
**Figure 9.23: Dictionary table to load the contents of AFINN dictionary**



**Figure 9.24: Contents of dictionary**

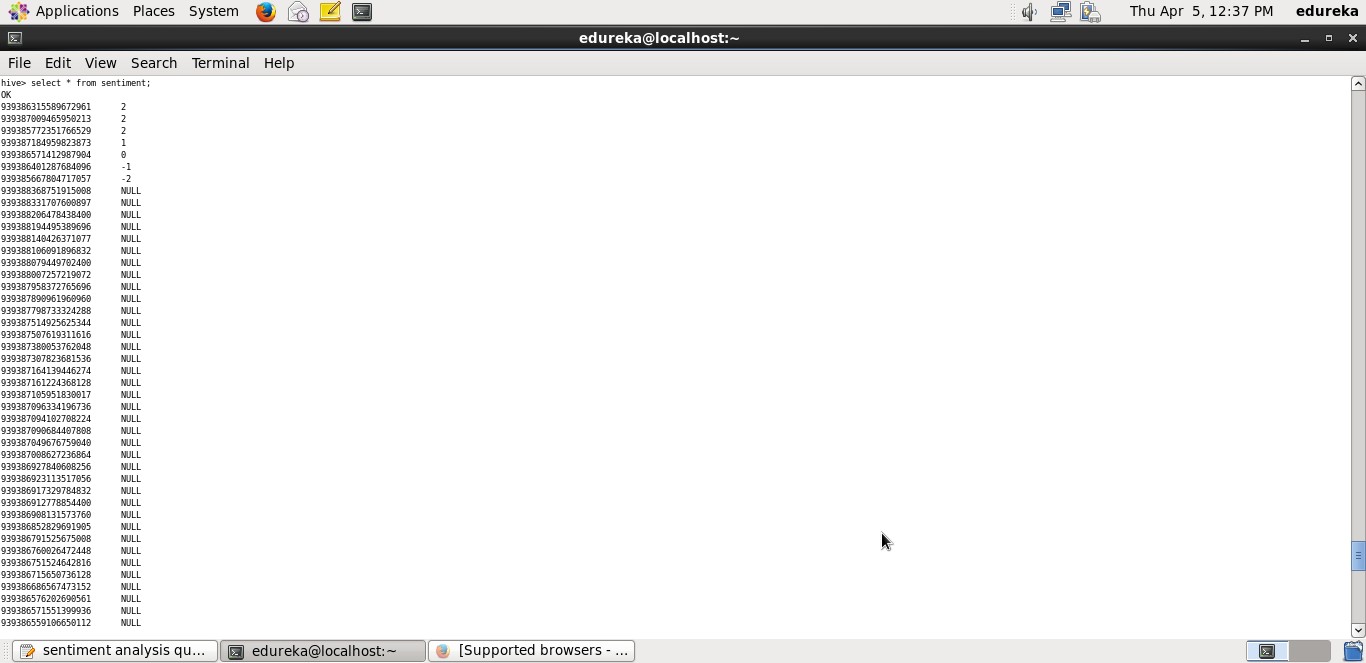
*School of Computer Engineering, KIIT, BBSR* 37

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.25: Table word join comprising the join the tweet***wordtableanddictionarytable*

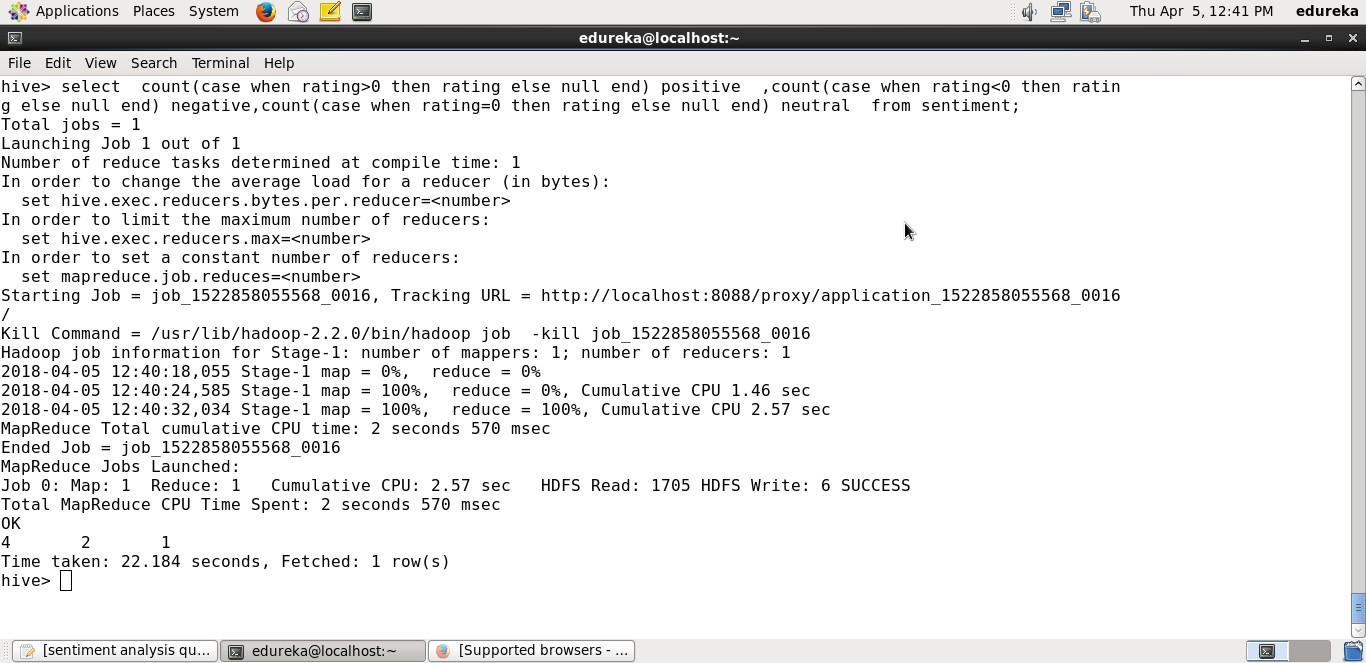
**Figure 9.26: Sentiment table comprising the average rating of each tweet by using each word of the tweet and arranging the tweets in the descending order as per their rating (1)**



**Figure 9.27: Contents of sentiment table**

*School of Computer Engineering, KIIT, BBSR* 38

*TWITTER SENTIMENT AND DATA ANALYSIS*



**Figure 9.28: Total number of positive,negative and neutral tweets**

*School of Computer Engineering, KIIT, BBSR* 39

*TWITTER SENTIMENT AND DATA ANALYSIS*

# Chapter 10

**Conclusion and Future Scope**

## CONCLUSION

Twitter posts are important source of opinion that deals on different issues and topics. It can give a keen insight on different topic and can be a good source for analysis. Analysis helps in decision making in various fields. Apache Hadoop is one of the best tool for twitter post analysis. Once the system is installed and set up using FLUME and HIVE, it helps in analysis of different topics by changing the query. We can also do real time analysis, so it is very much useful. The analysis what I did could be helpful in understanding peoples mood for election voting. And can be help in strategy planning making. Also sentimental analysis is done that data for finding polarity (Positive, Negative, Neutral) of tweets collected.

## PROPOSED FUTURE ENHANCEMENTS:

* + - These days big data has turned into the trendy expression in IT industry orga- nization. The need of analysing and preparing of data has grown a great deal. This paper actualized the analysing of big data (tweets) just for text. Further analysis should be possible to image and a wide range of interactive media doc- uments in light of record bolster. The consequence of Text mining and big data analysis would help in proposing related pages in view of various sorts of data.

*School of Computer Engineering, KIIT, BBSR* 40

*TWITTER SENTIMENT AND DATA ANALYSIS*

* + - Predictive Analytics: Predictive analysis gives the predictive scores which help the organizations in making smart decisions and make better business solutions. It reduces marketing campaigns and website behavior to improve customer re- sponses in business, conversions and meetings, and to decrease the stress on the people.
    - Web Clickstream data: Hadoop made easy way to keep track customers and their daily action in different issues like products purchasing and viewing etc. It makes analyzers to know customer behavior and interest and can able to show similar types and categories of products to the customers.

*School of Computer Engineering, KIIT, BBSR* 41

*TWITTER SENTIMENT AND DATA ANALYSIS*

# References

1. *Real Time Analysis of Twitter Data Using Hadoop*; Sunil B. Mane , Sunil B.

Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde

1. *”BigSoSA:Social Data Analysis and Data Visualization on Big Data, Interna- tional Journal of Advanced Research in Computer and Communication Engi- neering, Vol. 4*; Mahalakshmi R, Suseela S
2. *Potential Applications of Data Analysis in Educational Research and Practice Is SITE the Friendliest Conference*; Matthew Koehler, Spencer Greenhalgh,

Andrea Zellner

1. *Big Data Data Analysis using Hadoop , (IJIRST )International Journal for Innovative Research in Science Technology,Volume 1* ; Ramesh R, Divya G,

Divya D, Merin K Kurian

1. *Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets*; Peiman Barnaghi, Parsa Ghaffari, John G. Breslin
2. *Mining Data from Twitter*; AbhishangaUpadhyay, Luis Mao, Malavika Goda Krishna
3. *Sentiment Analysis and Opinion Mining: A Survey , International Jour- nal of Advanced Research in Computer Science and Software Engineering*;

G.Vinodhini , RM.Chandrasekaran

1. *”Addressing Big Data Problem Using Hadoop and Map Reduce”*; Aditya B.

Patel, Manashvi Birla, Ushma Nair

1. <http://www.michael-noll.com/tutorials/>

runninghadoop-on-ubuntu-linux-single-node-cluster/

1. <http://pingax.com/install-apache-hadoopubuntu-cluster-setup/>

*School of Computer Engineering, KIIT, BBSR* 42