ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Twitter, Facebook and Google Data Analysis Using Open Source Hadoop Technologies Hive

Y.sirisha, N.Arun Jyothi, K.S Parimala

Assistant Professor, Dept. of CSE, Bharat Institute of Engineering and Technology, Hyderabad, Telangana, India

Assistant Professor, Dept. of CSE, Bharat Institute of Engineering and Technology, Hyderabad, Telangana, India

Assistant Professor, Dept. of CSE, Bharat Institute of Engineering and Technology, Hyderabad, Telangana, India

ABSTRACT: In the recent years huge amount of unstructured data has been generated from the social media websites like twitter, facebook and google.The data that is generated from these social media websites and forums is in the form of text, images, videos and documents. Such huge amount of data that is generated is called as big data. Using RDBMS only structured data can be processed and stored. In this paper we can analyze all unstructured, semistructured and structured data using hadoop.Hence HDFS file system is used for storing the data and map reduce can be used for processing the data. The analysis is done on large amount of datasets using flume and hive.

KEYWORDS: Hadoop, HDFS, Mapreduce, Semistructured, Unstructured, Flume, Hive

I.INTRODUCTION

In every day social media websites generate vast amount of data. The data that is found on the web social media, remote sensing data, medical and industrial applications is termed as Big data. Big data generates large amount of data sets which leads to data explosion[1].All the traditional processing systems are inadequate for the big data sets .

The challenges for the big data are

- analysis
- capture
- data creation
- search
- storing
- transfer
- visualization
- querying and information privacy.

1.1 what is meant by big data?

Big Data contains high volume, velocity and high variety of data.

As a knowing all above terms, Big Data can be more complex, like cloud technology because it contains divergent technologies and information. The incoming data to Big Data system are collected from different social networks, satellites , web server logs, banking transactions, scans of government document, MP3 rock music, GPS tracking, airplane tracking, and stock market and so on[2].The whole data that is generated from all these social networking sites is termed as big data.

To knowing this entire concept there are three Vs of Big Data i.e. volume, variety and velocity are generally used to categorizing dissimilar feature of Big Data.[3]

Following details of Big Data 3 V's Challenges:

Velocity: How fast the data is entering the systems.

Variety: It includes all kinds of structured and unstructured data

Volume: The data complexity of terabytes to peta bytes

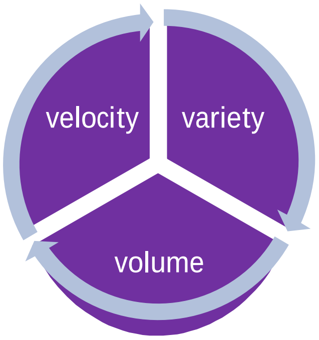

Figure 1.1:Bigdata 3v's Challenges

social media networks generates unstructured and semi structured data which is difficult to analyze. such large data sets are difficult to store process and analyze. Hence in this paper for storing and processing the massive amount of data can be done using apache Hadoop architecture.

HDFS: The Hadoop Distributed File System is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks.

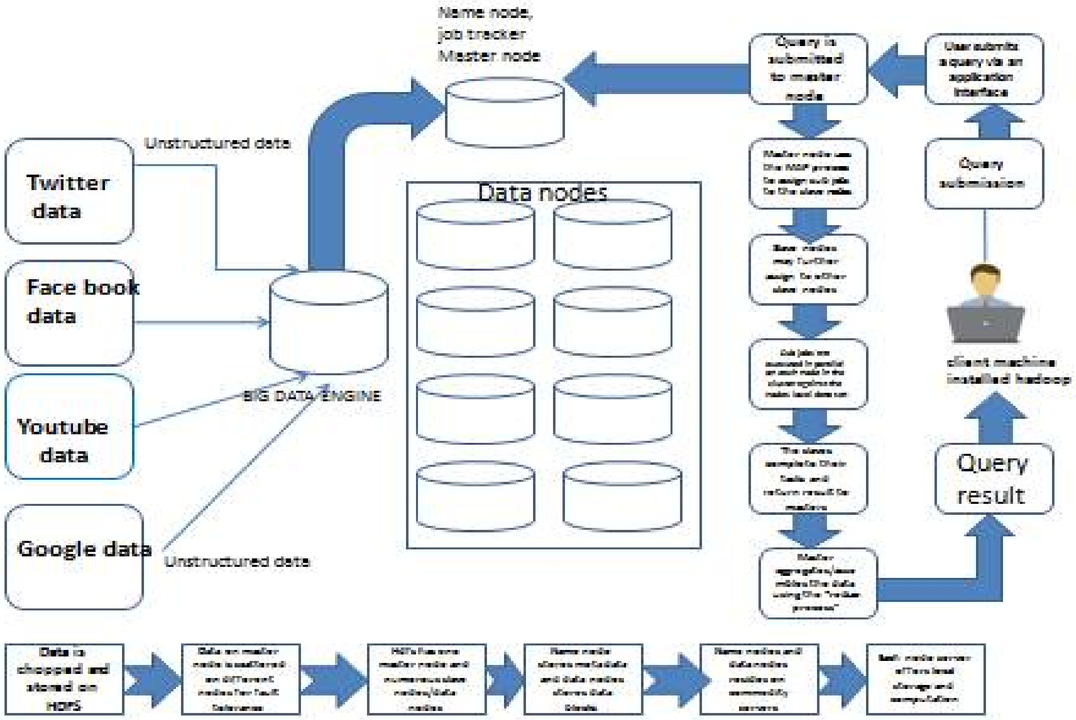
Name node,
job tracker
Master node
Unstructured data
Twitter data
Face book data
Youtube data
Google data
Unstructured data
BIG DATA ENGINE
Data nodes
Query is submitted to master node
User submits a query via an application interface
Master node use the MAP process to assign sub jobs to the slave nodes
Slave nodes may further assign to other slave nodes
The slaves complete their task and return result to master
Query submission
client machine installed hadoop
Query result
Data is chopped and stored on HDFS
Data on master node is scattered on different nodes for fault tolerance
HDFS has one master node and numerous slave nodes/data nodes
Name node stores meta data and data nodes stores data blocks
Name nodes and data nodes resides on commodity servers
Each node server offers local storage and computation

Figure 1.2:HDFS File System Architecture

II. RELATED WORK

2.1 Social media data analysis using RDBMS

For managing small amounts of data relational database management system can be used .In RDBMS structured query language can be used (SQL) to define, create and modify the database tables. Now the size of the data is increased in that way that we cannot store the data in a table. The size of the data is increased up to petabytes i.e 1PB=1024TB.This is a challenging task to handle the large amount of data[4]

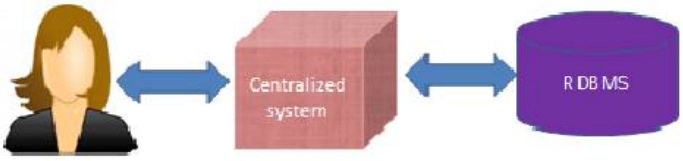

FIGURE 2.1 DATA ANALYSIS USING RDBMS

The maximum amount of data is from audio,video,text,images,documents,emails and social media which is in semi structured and unstructured form.RDBMS cannot manage such unstructured data. It can manage only structured data like tables, records and files. But Big data consists of massive volume of data which require high velocity of processing where RDBMS cannot handle such high volumes of data.

2.2 Social media data analysis using Hadoop:

Analysing the social media data is very complex and challenging.Hadoop framework architecture deals with large volume and variety of data but not small volumes of data. It requires fast processing for such high volumes of data. The complicated task is to manage the high velocity of data in big data. The data from facebook,twitter,google is in very large amount of data which cannot handled by RDBMS.

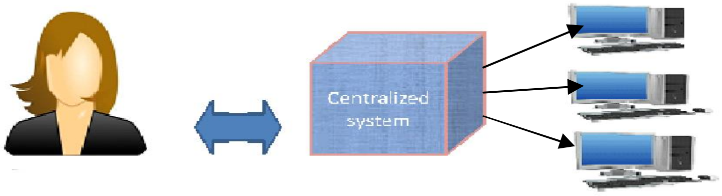

Figure 2.2 Data Analysis Using Hadoop

we need to analyze the system with online streaming i.e the number of text messages video, audio data that is generated per second is termed as online streaming data. For processing this online streaming data more computations and calculations are required.

2.3 social Data Analytics using Proposed System

Hadoop technology is one of the powerful tool which can be utilized for large and massive data set analysis[10].Map reduce is one of the open source tool which is designed for conversion and extreme analysis of very massive data.

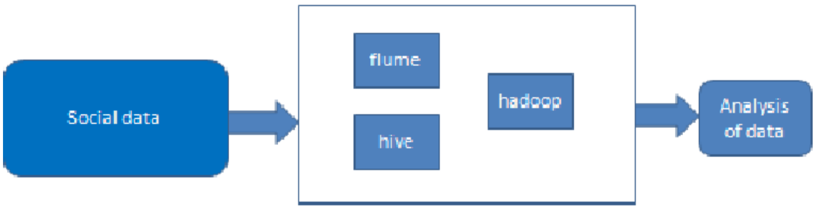

Figure 2.3 System Architecture

In this paper we designed algorithm for handling problems raised by wide extent of datasets. For the analysis of operations that are performed on the social media data sets first we used standard platform as hadoop which is installed on ubuntu to solve the challenging tasks using map reduce frame work. Here the complete data is mapped with frequent datasets[5] and this can be reduced to smaller sized data so that hive can handle the data which is on the top of the hadoop.The preconditions for the flume and hive is hadoop should be pre-install. Flume can be used for retrieving the real time twitter data and stored in HDFS.After the data storage we are executing the analysis of these complicated data using hive.

III. PROPOSED METHODOLOGY

step1:For retrieving the realtime twitter data we are creating twitter app using twitter streaming API.
step2:For the analysis of twitter data first data is uploaded using flume in local HDFS.All the tweets from twitter site are directly stored in the HDFS.Data retrieved from the twitter is in unstructured form known as JSON data.
step3:Next step is the analysis of the twitter data stored in the HDFS can be performed using hive.using hive we can convert the unstructured form of data into readable structured form of data.
step 4:The preprocessing of data can be done for removing noise and meaning less symbols.Now the data is schema oriented and using hive we can analyze the data by writing different queries.

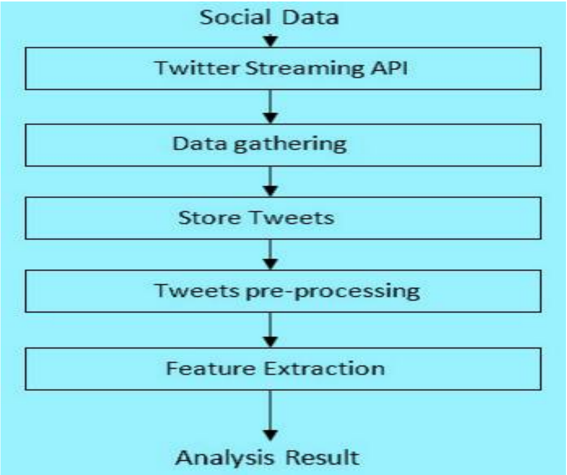

Figure 3.1 proposed methodology

IV. SIMULATION RESULTS

In this paper social data analysis is done using hive on twitter data,facebook data and google data .In the figures we have shown how the tweets can be analysed i.e it can be positive , negative tweets or neutral tweets.In the second diagram we have shown how the positive,negative and neutral tweets can be represented in different countries.In the same way facebook data we have shown how many likes and comments can be analysed for a particular page.In the google data analysis we have shown the number of population are interested in browsing for one particular website.

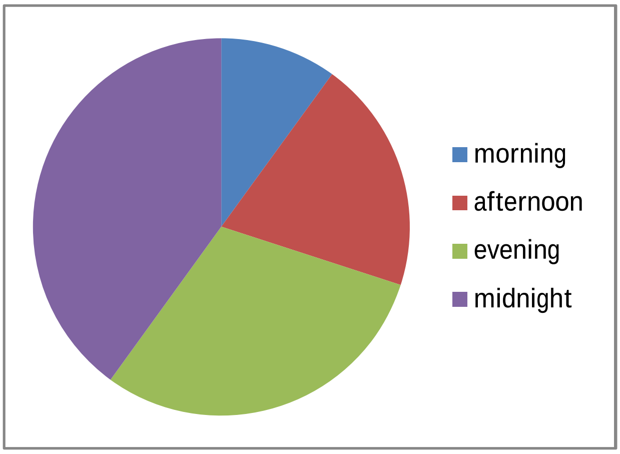

Figure1:No Of Tweets Counted During Time Period

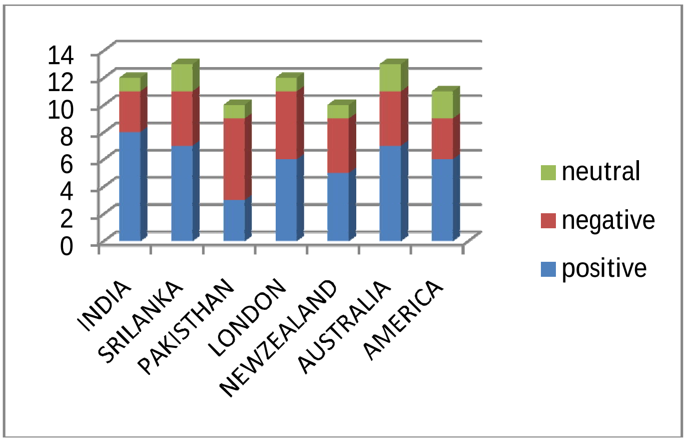

Figure 2:No of positive,negative and neutral Tweets

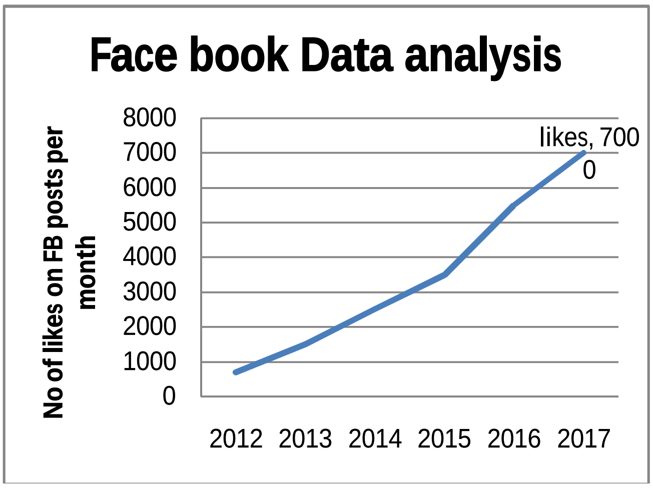

Figure 3:Face book data Analysis

V. CONCLUSION AND FUTURE WORK

Using the traditional RDBMS we cannot perform analysis on such large and complex data sets.But now using hadoop tool an intense analysis of the data can be done using FLUME,HIVE.Both the echosystems runs on the top of the hadoop and the we can use hive for analyzing these massive amount of data.Here we performed on twitter data because it gives different opinions and different types of topics helps us in decision making.In future the same analysis can be done on youtube data also using any of these hadoop technologies

REFERENCES

1 Ms.Sulochana Panigrahi Social Data Analysis Using Big-Data Analytic Technologies-APACHE FLUME, HDFS, HIVE Journal of Computer Science and Engineering Volume-2 , Issue-5 , May 2016.

2Amita Jajoo1, Kuldeep U. Karpe2, Nilesh R. Parade3, Pratik D. Powar Real-Time Big Data Analytics using Hadoop International Journal of Emerging Technology and Advanced Engineering Volume 6, Issue 1, January 2016

3Ms.Sulochana Panigrahi Social Data Analysis Using Big-Data Analytic Technologies-APACHE FLUME, HDFS, HIVE Journal of Computer Science and Engineering Volume-2 , Issue-5 , May 2016.

4Amita Jajoo1, Kuldeep U. Karpe2, Nilesh R. Parade3, Pratik D. Powar Real-Time Big Data Analytics using Hadoop International Journal of Emerging Technology and Advanced Engineering Volume 6, Issue 1, January 2016

5 Manish Wankhede1 , Vijay Trivedi2, Dr.Vineet Richhariya Analysis of Social Data Using Hadoop Ecosystem International Journal of Computer Science and Information Technologies, Vol. 7 (6) , 2016, 2402-2404

6 Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

7.Hadoop Map-Reduce Tutorial at http://hadoop.apache.org/common/docs/current/mapred_tutorial.html.

8.Hadoop HDFS User Guide at http://hadoop.apache.org/common/docs/current/hdfs_user_guide.html

9. Judith Sherin Tilsha S*, Shobha ,A Survey on Twitter Data Analysis Techniques to Extract Public Opinion , International Journal of Advanced Research in Computer Science and Software Engineering Research Paper , Volume 5, Issue 11 , November 2015 .

10 Munesh Kataria1, Ms. Pooja Mittal,Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql, International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology . IJCSMC, Vol. 3, Issue 7, July 2014.

BIOGRAPHY

Y.sirisha working as Assistant Professor in CSE Department in Bharat Institute of Engineering and Technology.she received master of Technology (M.tech) degree in 2010 from vignans Engineering College,Vadlamudi,Guntur.Her research interests are Big data analysis, cloud computing,adhoc sensor networks etc.