International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

Twitter Data Analysis – A Review

Surbhi Choudhary[1], Neha Kaurav[2]

[1]M.Tech, [2]Asst. Professor, Department of Computer Science & Engineering, Oriental College of Technology, Bhopal, India

***Abstract*— Social networking over Internet has become popular in the last years, which is also justified with the increased data volumes. New challenges appeared in relation to data storage architectures with scalability features and effective processing algorithms. Twitter has big potential for data mining as its users produce Big Data that can be processed. In addition, there are requirements for architecture development that can scale to continuous new-streamed tweets and also ability to integrate with advanced machine learning algorithms. Knowing what users think or how they feel about products is valuable proposition for companies. So the twitter data is very important to analyse for various companies to find the public opinions and trends. Hadoop is one of the best tool options for twitter data analysis and hadoop works for distributed Big data. This paper we introduce an open source solution for analysis twitter data which is hadoop and its various ecosystems.**

***Index Terms*-- Hadoop, twitter, data mining, social analysis, hadoop ecosystem.**

I. INTRODUCTION

Over the last several years there has been an explosion of growth and new activity in social networking. Various companies such as Facebook, LinkedIn, Reddit, Pintrest, and Twitter have grown exponentially in recent years. The amount of data exchanged between users on these sites is staggering. On Facebook alone on an average day in 2014 there are 4.75 billion items being shared, 4.5 billion items "liked", and 300 million photographs being uploaded [1]. That translates to over 500 terabytes of data generated by Facebook users on a single day. There is an incredible amount of useful information [5] about individual opinions, feelings, and relationships contained in these transactions, but the loosely structured nature of human communication makes harnessing this data a challenge. In order to make sense of the large portion of this data which is text-based, Natural Language Processing tools can be used to rigorously categorize user generated text. One of these tools for determining useful information from massive data sources such as Twitter is a sentiment analysis.

Sentiment analysis [11][13] focuses on determining the opinion of a speaker on the particular topic about which he is speaking.

The most basic structure for sentiment analysis is a single word, unfortunately based on sentence structure and words with context dependent meanings, techniques that ignore sentence structure or bag of words models often fail on smaller texts. A solution to this is constructing parse trees which identify the structure of a sentence as a binary tree by separating distinct phrases. In this case using the sentiment of each word in the tree can take into account clause structures and the possibility of multiple meanings. In cases where a larger text must be analyzed, it can be treated as a collection of smaller phrases, or as a larger bag of words. Opinions are usually classified somewhere between positive or negative often with some stratification between the two. This can be done numerically or categorically. When division is categorical, it usually distinguishes between positive, negative, and sometimes neutral sentiments, otherwise the numerical classification falls somewhere on a continuum between positive and negative. These classifications can be used to determine and aggregate the sentiment of a large number of authors on a given topic.

II. HADOOP

The Apache Hadoop project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using a thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

III. HADOOP ECOSYSTEMS

Hadoop Ecosystem[16] is a framework of various types of complex and evolving tools and components which have proficient advantage in solving problems. Some of the elements may be very dissimilar from each other in terms of their architecture; however, what keeps them all together under a single roof is that they all derive their functionalities from the scalability and power of Hadoop. Hadoop Ecosystem is alienated in four different layers: data storage, data processing, data access, data management. All the components of the Hadoop ecosystem, as explicit entities are evident. The holistic view of Hadoop architecture gives prominence to Hadoop common, Hadoop YARN, Hadoop Distributed File Systems (HDFS) and Hadoop MapReduce of the Hadoop Ecosystem. Hadoop common provides all Java libraries, utilities, OS level abstraction, necessary Java files and script to run Hadoop, while Hadoop YARN is a framework for job scheduling and cluster resource management. HDFS in Hadoop architecture provides high throughput access to application data and Hadoop MapReduce provides YARN based parallel processing of large data sets.

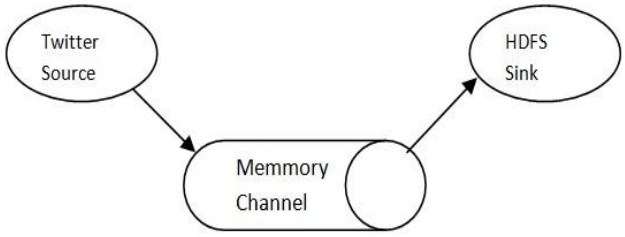

Fig 1. Hadoop ecosystem for real time data fetching

IV. LITERATURE REVIEW

In [1] With rapid innovations and growing Internet population, petabytes of information are being generated every second. Processing these enormous data and analysing is a tedious process now-a-days. The amount of data in real-time is growing tremendously. Nearly 80% of the data is in unstructured format. Analysis of unstructured data in real-time is a very challenging task. Existing traditional business intelligence (BI) tools perform best only in a pre-defined schema. Most of the real-time data are logs and dont have any defined schema. Doing queries over these large datasets takes long time. During streaming of realtime data, much unwanted information is extracted from the data source causing overhead in the system. This results in an increase in the cost of construction and maintenance. Each and every second, new data streams keeps accumulating in the system consistently about whats going on in the world. Gathering these data and processing is an essential skill to know, for preparing a vital report.

According to [2], Hadoop is Java based programming framework for distributed storage and processing of large data sets on commodity hardware. It is developed by Apache Software Foundation as open source framework. Hadoop basically has two main components. First one is Hadoop Distributed File System (HDFS) for distributed storage and second part is MapReduce for distributed processing. HDFS is a file system which builds on the existing file system. It is Java-based sub project of Apache Hadoop. HDFS provides scalable and reliable data storage on commandity hardware. The MapReduce framework consists of two process which are JobTracker and TaskTracker. The JobTracker manages the resources that are TaskTracker. The TaskTracker is a processing node in the cluster. It accepts several tasks like map reduce and shuffle from a Job Tracker. Twitter4J is an unofficial Java library for the Twitter application programming interface. It is integrated Java application with the all Twitter services.

In [3], the author describes that Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom form big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical [17] capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform ,HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

Praveen Kumar, Dr Vijay Singh Rathore [07] (2014) Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce Proposes, several solutions to the Big Data problem have emerged which includes the Map Reduce environment championed by Google which is now available open-source in Hadoop. Hadoops distributed processing, Map Reduce algorithms and overall architecture are a major step towards achieving the promised benefits of Big Data.

Manoj Kumar Danthala [04] (2015) Tweet Analysis: Twitter Data processing Using Apache Hadoop . This paper provides a way of analyzing of big data such as twitter data using Apache Hadoop which will process and analyze the tweets on a Hadoop clusters.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 7, July 2017)

This also includes visualizing the results into pictorial representations of twitter users and their tweets.

Judith Sherin Tilsha S, Shobha M.S [06] (2015) A Survey on Twitter Data Analysis Techniques to Extract Public Opinion. Using machine learning algorithm ,a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. It said that various twitter data analysis techniques that are based on dictionary and that are using the machine learning approaches.

V. Observation

Hadoop [4] and its Ecosystems, for getting raw data from the Social Network, we may use Hadoop online streaming tool. By utilizing this tool only, we are going to configure everything, which we wanted to get (data) from the Social Network. Mainly we want to set the configuration model and also want to define what information that we want to collect form Social Network. All these will be stored into our HDFS (Hadoop Distributed File System) in our own prescribed format. From this unrefined data we are going to create the table and filter the information that is needed for us and sort them into the Hive Table [14]. And from this, we are going to perform the Sentiment Analysis by using some UDF's (User Defined Functions) by which we can perform sentiment analysis.

VI. Problem Definition

The paper focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis.[6] The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. It will be shown how to automatically collect a twitter data for data analysis and data mining purposes and then perform linguistic analysis of the collected twitter data. All public tweets posted on twitter are freely available through a set of APIs provided by Twitter.

Twitter sites receives petabytes of data every day and these data is nothing but a collection of tweets so these data is very important in real life to analyse different scenario through which its helps us in decision making. The analysis of twitter data gives real view or different user opinions regarding what they think and to analysis these data provide a better way for making any decision.

VII. Proposed Methodolgy

For analysing these large and complex data required a power tool, we are using hadoop[10] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

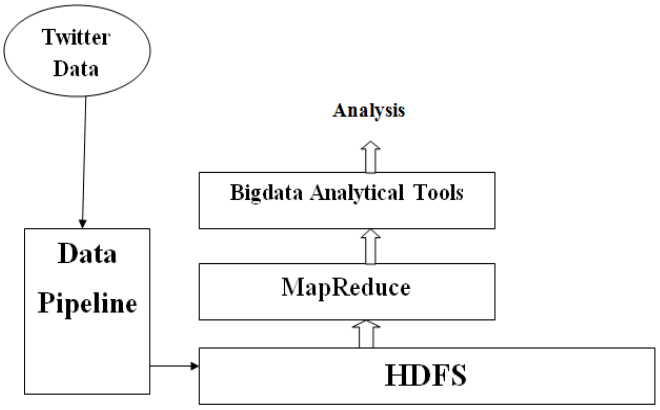

Figure 2. Workflow Diagram

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on social media data sets. For analysing first we used standard platform as hadoop on single node ubuntu machine to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling ,after this we integrate hadoop ecosystems[15]. Hadoop ecosystem[16] consists of different level layers, each layer performing different kind of tasks like storing your data, processing stored data, resource allocating and supporting different programming languages to develop various applications in Hadoop ecosystem.

Our Steps or Algorithm Steps will follow:

1. In first step We are creating a twitter app using a twitter streaming API [9] for fetching real time twitter data.
2. For doing twitter data analysis first data is uploaded using hadoop data pipelined ecosystems in local HDFS [8]. The twitter API through which all the tweets are directly fetch from the twitter site and stored it into the HDFS. Data comes from the twitter site is in un-structure form called JSON data [7].
3. After storing all twitter data into the HDFS we are performing the analysis part for these we use hadoop and its ecosystems through which we can convert the un-structure complex data in to readable or understandable structure form.
4. Tweets are preprocesses for removing noise and meaningless symbols, and than we perform various analysis activities on these twitter data through which we can find the real opinions and trends.

VIII. CONCLUSION

On analysing complete scenario regarding the analysis of social data we say that using traditional analytical tool we can not perform analysis on such huge and complex data, so we uses a new powerful tool which is designed for deep analysis called hadoop and also integrate with its ecosystems. And using these various ecosystems we can easily analyse the large and complex twitter data through which we can get the real opinions and trends.

REFERENCES

[1] Nikitha Johnsirani Venkatesan, Earl Kim, Dong Ryeol Shin, "PoN: Open Source solution for Real-time Data Analysis" in IEEE, ISBN: 978-1-4673-9379-9 ©2016 IEEE.

[2] Can Uzunkayaa, Tolga Ensaria, Yusuf Kavurucu, "Hadoop Ecosystem and Its Analysis on Tweets" in World Conference on Technology, Innovation and Entrepreneurship, Elsevier 2015.

[3] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced Analytics Tools", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[4] Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015, pp 94-102.

[5] White Paper Big Data Analytics Extract, Transform, and Load Big Data with Apache Hadoop-Intel corporation.

[6] Judith Sherin Tilsha S , Shobha M S, "A Survey on Twitter Data Analysis Techniques to Extract Public Opinion", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.

[7] Praveen Kumar, Dr Vijay Singh Rathore," Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014, pp 7123-7126.

[8] http://www.json.org/ECMA-404 The JSON Data Interchange Standard.

[9] http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS.

[10] "Application Programming Interface." *Wikipedia*. Wikimedia Foundation, 23 Oct. 2014. Web. 24 Oct. 2014.

[11] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[12] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More---Matthew A. Russell.

[13] Brian Dickinson, Wei Hu, "Sentiment Analysis of Investor Opinions on Twitter" in Social Networking, 2015, 4, 62-71

[14] Manish Wankhede, Vijay Trivedi, Vineet Richhariya, "Location based Analysis of Twitter Data using Apache Hive" in International Journal of Computer Applications (0975 – 8887) Volume 153 – No 10, November 2016.

[15] Sneha Mehta , Viral Mehta, "Hadoop Ecosystem: An Introduction" in International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[16] Manish Wankhede , Vijay Trivedi , Dr.Vineet Richhariya, "Analysis of Social Data Using Hadoop Ecosystem" in International Journal of Computer Science and Information Technologies, Vol. 7 (6) , 2016, 2402-2404.

[17] Dr. Vineet Richhariya , Mr. Jay Prakash Maurya , Sakshi Agrawal, "ANALYSIS OF SOCIAL DATA USING BIGDATA ANALYTICAL TOOLS" in IJARSE vol 6, issue no. 5, may 2017.