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

Trend Analysis Using Hadoop and Its Ecosystems

Deepak Ranjan, Dr. Tripti Arjariya, Dr. Mohit Gangwar

Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

Prof. & Head, Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

Principal, Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

ABSTRACT*:* Social media generates massive amounts of data every minute, which is caused by its mainstream adoption over the past years. Innovations in the industry have enabled new ways of communications between people and created many business opportunities. Big Data in social media require effective and advanced processing technologies. Purpose of data mining analyses is to find valuable patterns and insights from Twitter data. Thus, analysis for twitter data is meaningful for both individuals and organizations to make decisions. Due to the huge amount of data generated by twitter every day, a system which can store and process big data is becoming a problem. In this paper, we present a method to collect twitter data sets, and store and analyze the data sets on Hadoop platform.

KEYWORDS: Hadoop, data mining, data analysis ,social datas, bigdata analytical tools.

I. INTRODUCTION

Over the past years, the amounts of data generated from the Internet services have increased significantly. Innovations in the field of ICT have enabled new business opportunities for creating services capable of handling vast data volumes. Technology reached the level, where people are interconnected with social media on daily basis and are able to share their life over social networks. 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.

Data mining analysis has a great potential in finding meaningful insights within social networks data. Twitter social network is a service developed in order to enable communication between people by sending short messages. According to research [1], Twitter as the second largest social media platform, right behind Facebook, generates around 350, 000 tweets each minute, or 21 million per hour. Such volumes of data present challenges for engineers to develop innovative solution for effective data architecture and processing capabilities in order to apply data mining. The importance of Big Data implementations within enterprises in various sectors, such as health industry, retail, telecom or social networks plays crucial scenario in optimizing business processes and creating new value propositions for revenue streams.

According to Accenture research [1], 87% of enterprises believe that Big Data will reshape the industry in the next years. Therefore, the early adoption of this trend may create additional value to enterprises in sense of data mining of customer's opinions about company in the Twitter use case. Knowing what customers think about the services or how services can be innovated in the future gives companies insights and strategies for development. Furthermore, survey also found that 89% of respondents said that companies who do not adapt to this Big Data trend would risk losing market share.

Start by searching Twitter. Check replies to your association's tweets and the accounts of other leading industry organizations to learn what hashtags they are using. You can also search common industry terms. Compare hashtags by using Topsy (topsy.com/analytics). This site keeps you up to date with what's trending and what people are following.

DOI:10.15680/IJIRCCE.2017. 0505260

HADOOP

The Apache Hadoop[9] 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 [8] 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.

II. LITERATURE REVIEW

In [1] This paper author investigates the problem of predicting Twitter hashtags popularity level. A data set of more than 18 million tweets containing 748 thousand hashtags has been prepared by using Twitter's rest API. Early adoption properties including profile of tweet authors and adoption time series are used to predict a tag's later popularity level. The followers count and tweets count are two such characteristics related to adopters' profile. On the other hand, two types of frequency domain analyses are used to augment the simple mean and standard deviation characteristics of the adoption time series. Fourier transform (FT) spectrum and wavelet transform (WT) spectrum are considered in this study. Experimental results show that WT spectrum improves the prediction result of viral hashtags while FT spectrum does not.

Online social networks provide communication channels to spread an idea, behavior, style or usage throughout the world village. Twitter is a special online service that provides both social network and microblog functions. Posting tweets through devices from desktop to mobile is the main activity of the microblog function, while following and retweeting offer the social network function. Users post tweets by encoding topics in the form of hashtags, which are summarized by Twitter to make a list of current trending tags.

In [2], 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 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.

III. OBSERVATION

Hadoop and bigdata analytical tools, for getting raw data from the Social Network, we may use Hadoop online streaming tool such as Apache Flume, apache kafka. 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 refined the data using analytical tools and than start analysing these social data to predict or to help in decision making.

IV. PROBLEM DEFINITION

The project focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. 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. Because these tweets generates the huge information related to different area like government, election, etc. millions of tweets is generated every day and which is very useful in decision making because every one is share their view and opinions on issues or variety of topics. The analysis of twitter data gives real view on trends or different user opinions regarding what they think and to analysis these data provide a better way for making any decision.But the twitter site generates petabyte of data per day which is difficult to handle with traditional tools and technique for this we need a new powerful technique to handle bigdata.

V. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[6] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

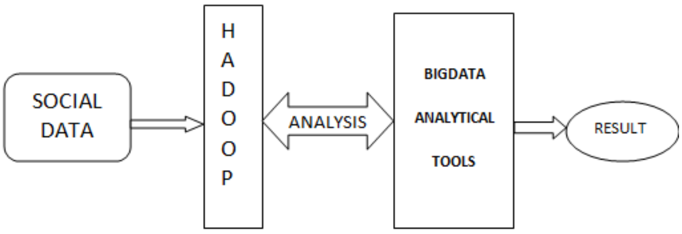

Figure1. Workflow Diagram

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 that we can use bigdata analytical tools to refine such unstructured data and analyse the social data using bigdata analytical tools.

VI. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow:

1. First we can create a social account and than we can use social API's for fetching real time social data and store it into HDFS.
2. For fetching social data we can use bigdata tools such as apache flume through which we can authenticate our keys and start fetching fetching data from social sites.
3. After fetching data, the data is store into HDFS (Hadoop Distributed File System), which is very reliable for storing such huge amount of data.
4. After storing data into HDFS, we can pre-process the data because from the social sites an unstructured raw data is coming, which is very difficult to analyze such kind of unstructured data, so we can first pre-process the data and convert it into some structure form.
5. After pre-processing we can start analyzing such huge amount of social data.

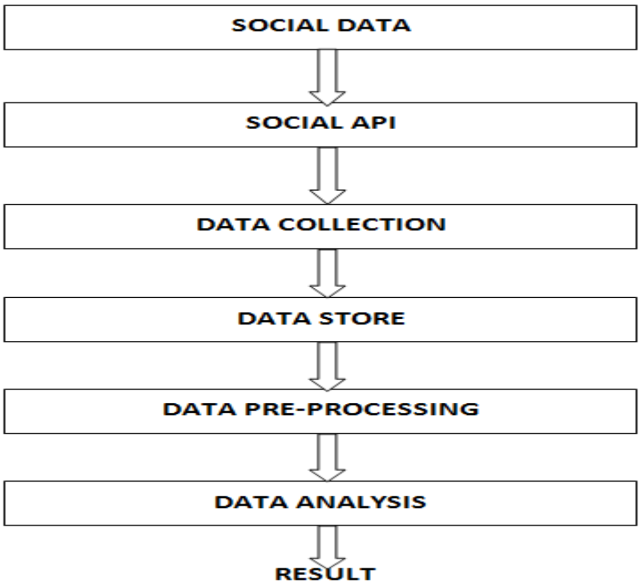

Figure 2: Analysis Steps

VI. CONCLUSION

In this paper, we introduced bigdata analytical tools through which we can do sentiment analysis of social data. It will give us an effective output which is easy to understand. This application is very useful for decision making in various domains. And because of HADOOP it becomes easy to process the data in less time.

REFERENCES

[1] Shing H. Doong, "Predicting Twitter Hashtags Popularity Level", in 2016 49th Hawaii International Conference on System Sciences, IEEE, DOI 10.1109/HICSS.2016.247.

[2] 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.

[3] Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015.

[4] 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.

[5] Ramesh R, Divya G, Divya D, Merin K Kurian , "Big Data Sentiment Analysis using Hadoop ", (IJIRST)International Journal for Innovative Research in Science & Technology,Volume 1 , Issue 11 , April 2015 ISSN : 2349-6010

[6] 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.

[7] G.Vinodhini , RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

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

[9] 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/