International Journal of Scientific and Technical Advancements
IJSTA
ISSN: 2454-1532

A Survey on Social Data Processing Using Apache Hadoop, Map-Reduce

Sulochana Panigrahi[1], S Mohan Kumar[2]
[1, 2]Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India
Email address: [1]sulochanap01@gmail.com, [2]drsmohankumar@gmail.com

***Abstract*—**'BIG DATA' has been getting much importance in different industries over the last year or two, on a scale that has generated lots of data every day. Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a reasonable amount of time. It has tremendous potential to transform business and power in several ways. Here the challenge is not only storing the data, but also accessing and analyzing the required data in specified amount of time. One of the popular implementation to solve the above challenges of big data is using Hadoop. Hadoop is well-known open-source implementation of the MapReduce programming model for processing big data in parallel of data-intensive jobs on clusters of commodity servers. It is highly scalable compute platform. Hadoop enables users to store and process bulk amount which is not possible while using less scalable techniques. Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. 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. This also includes visualizing the results into pictorial representations of twitter users and their tweets.

***Keywords*—** Big data; Hadoop; MapReduce.

I. INTRODUCTION

Industries and organizations don't need to store and perform much operations and analytics on data of the customers. But around from 2005, the need to transform everything into data is much entertained to satisfy the requirements of the people. So Big data came into picture in the real time business analysis of processing data. The term big data refers to the data that is generating around us everyday life. It is generally exceeds the capacity of normal conventional traditional databases. For example by combining a large number of signals from the user's actions and those of their friends, Facebook developed the large network area to the users to share their views, ideas and lot many things.

The value of big data to an organizations falls into two categories: analytical use and enabling new products based on the existing ones. Big data can reveal the issues hidden by data that is too costly to process and perform the analytics such as user's transactions, social and geographical data issues faced by the industry. The major characteristics and challenges of big data are Volume, Velocity, and Variety. These are called as 3V's of big data which are used to characterize different aspects of big data. Here the storing of these huge amounts of data will require high clusters and large servers with high bandwidth. And here the problem is not only storing the information but also the processing at much higher speed. This became the major issue nowadays in most of the companies.

Variety

In the distributed environment there may be the chances of presenting various types of data. This is known as variety of data. These can be categorized as structured, semi structured and unstructured data. The process of analysis and performing operations are varying from one and another. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are. So this is the major issue to process information from different sets of data.

Veracity

Big data Veracity refers to the biases, noise and abnormally in data. It is the data that is being stored, and mined meaningful to the problem being analyzed. In other words, Veracity can be treated as the uncertainty of data due to data inconsistency and incompleteness, ambiguities, latency, model approximations in the process of analyzing data.

Hadoop

In the distributed environment, the data should be available to users and capable of performing different analysis from databases in a specified amount of time. If we use the normal approaches, it will be difficult to achieve big data challenges. So these types of data required a specialized platform to deal in such cases. Hadoop is the one of the solution to solve big data problems. Hadoop is the open source flexible infrastructure for large scale computation and data processing on a network commodity of hardware systems. Here it deals with both structured and unstructured data and various challenges of big data are solved. The main components of Hadoop are commodity hardware and MapReduce.

MapReduce

MapReduce is the heart of Hadoop and each job is performed by this technique only. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The term MapReduce refers to two tasks in Hadoop i.e. Map and Reduce. In the first step, it takes the data and converts it into another set of data. Here the each word is referred as key and

Sulochana Panigrahi and S Mohan Kumar, "A survey on social data processing using apache Hadoop, Map-Reduce ," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 2, pp. 121-123, 2016.

the number of occurrences is treated as value. So MapReduce tuple consists of key value pairs. Then second step consists of reduce operation where it takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

Related Research Works

- Rui LI , Kin Hou Lei, Ravi Khadiwala , Kevin Chen-Chuan Chang, TEDAS: a Twitter Based Event Detection and Analysis System[1]. It propose a Twitter-based Event Detection and Analysis System (TEDAS), which helps to detect new events, to analyze the spatial and temporal pattern of an event, and to identify importance of events. In this demonstration, the overall system architecture, explain in detail the implementation of the components that crawl, classify, and rank tweets and extract locations from tweets, and present some interesting results of system.
- Andrea Vanzo and Danilo Croce and Roberto Basili, A context-based model for Sentiment Analysis in Twitter [2]. It modeled the polarity detection problem as a sequential classification task over streams of tweets. A Markovian formulation of the Support Vector Machine discriminative model as embodied by the SVMhmm algorithm has been here employed to assign the sentiment polarity to entire sequences. The experimental evaluation proves that sequential tagging effectively embodies evidence about the contexts and is able to reach a relative increment in detection accuracy of around 20% in F1 measure. These results are particularly interesting as the approach is flexible and does not require manually coded resources.
- Mohit Tare, Indrajit Gohokar, Jayant Sable, Devendra Paratwar, Rakhi Wajgi, 2014 Multi-Class Tweet Categorization Using Map Reduce Paradigm [3]. It propose the use of one of the classification algorithm called Naïve Bayes for the categorization of tweets which has been discussed in this paper. It then proposes how the Map – Reduce paradigm can be applied to existing Naïve Bayes algorithm to handle large number of tweets.
- Mahalakshmi R, Suseela 2015 Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data [4]. It proposed a method of sentiment analysis on twitter by using Hadoop and its ecosystems that will process the large volume of data on a Hadoop and the MapReduce function will perform the sentiment analysis.
- Dipak Gaikar, Bijith Marakarkandy 2015 Product Sales Prediction Based on Sentiment Analysis Using Twitter Data [5]. This research paper uses a survey approach for movie sales prediction. This paper analyses, impact of the positive, negative, strongly positive and strongly negative online reviews of movies on the audience. It should be noted that the user feedback is given prior to watching the movie only on the basis of the online reviews. The result of this research will help the film industry to effectively address and meet the expectations of customers and stakeholder. This paper also investigates techniques for twitter data extraction using an API key.
- Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Luigi Grimaudo, Xin Xiao, Analysis of Twitter Data Using a Multiple-Level Clustering Strategy [6]. This paper proposes a data analysis framework to discover groups of similar twitter messages posted on a given event. By analyzing these groups, user emotions or thoughts that seem to be associated with specific events can be extracted, as well as aspects characterizing events according to user perception. To deal with the inherent sparseness of micro-messages, the proposed approach relies on a multiple-level strategy that allows clustering text data with a variable distribution. Clusters are then characterized through the most representative words appearing in their messages, and association rules are used to highlight correlations among these words. To measure the relevance of specific words for a given event, text data has been represented in the Vector Space Model using the TF-IDF weighting score. As a case study, two real Twitter datasets have been analysed.
- Supraja.G.S, Dr Jharna Majumdar, Shilpa Ankalaki 2015 A Big Data Methodology for Sentiment Analysis of Twitter Data [7]. In this paper it is proposing a methodology to collect and store live twitter data and perform sentimental analysis using machine learning techniques and provide some prediction. To store the live data fetched we are using MongoDB a NoSQL database, the output of the analysis will be trend analysis with different sections that is positive, negative and neutral.
- Jasmeen Gill RIMT-IET, Mandi Gobindgarh, Shaminder Singh GGI, Khanna, Devdutt Baresary GGI, Khanna 2015 Big Data: Big Innovations in Healthcare [8]. This research article highlights various aspects of big data like usability, security and reliability in healthcare services. Apart from this, it provides various analytical tools of big data used in healthcare.

Observation

Today, there is demand of quick and accurate result. So need to store data in proper manner with power of easily retrieval. With explosion of data in recent scenario only traditional database is not enough to handle it. With high rate of changing data on web applications there is need of database which can perform to provide consistency as well as partition tolerance. Present database system work with vertical enhancement that give scale-in facility for system. That is not enough for huge database like LinkedIn, face book, Amazon etc. That huge amount of data needs to have horizontal enhancement that give scale-out property. By this enhancement any number of node can be added with system. For Big data there is use of MapReduce programming model that perform operation on single large file so that there is no need to split data. Companies like facebook, twitter, linkedin etc start using Hadoop Hadoop Ecosystem include MapReduce, Apache Hive, PigLatin, Sqoop, flume, zookeeper and HBase. HBase is column oriented database. Its structure consist column family in which different columns are defined

Sulochana Panigrahi and S Mohan Kumar, "A survey on social data processing using apache Hadoop, Map-Reduce ," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 2, pp. 121-123, 2016.

with unique row id. In this paper, part1 describe the Introduction of the basics of MapReduce and Bloom filter. Part 2 gives the basic information of Hadoop components and detail description of HBase. Part 3 shows the proposed implementation on HBase with Bloom filter.

II. CONCLUSION

Hadoop with its efficient DFS & programming framework based on concept of mapped reduction, is a powerful tool to manage large data sets. With its map-reduce programming paradigms, overall architecture, ecosystem, fault- tolerance techniques and distributed processing, Hadoop offers a complete infrastructure to handle Big Data. Users must leverage the benefits of Big-Data by adopting Hadoop infrastructure for data processing. However, the issues such as lack of flexible resource management, application deployment support, and multiple data source support pose a challenge to Hadoop"s adoption. Proper skill training is also needed for achieving large scale data analysis. These challenges must be overcome so that we can tap the full potential of Hadoop data management power.

REFERENCES

[1] R. LI, K. H. Lei, R. Khadiwala, and K. C. Chuan Chang, "TEDAS: A twitter based event detection and analysis system," Computer Science, University of Illinois at Urbana-Champaign.

[2] A. Vanzo, D. Croce, and R. Basili, "A context-based model for sentiment analysis in twitter," Department of Enterprise Engineering University of Roma Tor Vergata.

[3] M. Tare, I. Gohokar, J. Sable, D. Paratwar, and R. Wajgi, "Multi-Class tweet categorization using map reduce paradigm," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 9, no. 2, pp. 78, 2014.

[4] R. Mahalakshmi and Suseela, "Big-SoSA: Social sentiment analysis and data visualization on big data," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 6, issue 3, pp. 2303-2313, 2015.

[5] D. Gaikar and B. Marakarkandy, "Product sales prediction based on sentiment analysis using twitter data," *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 6, issue 3, pp. 2303-2313, 2015.

[6] E. Baralis, T. Cerquitelli, S. Chiusano, L. Grimaudo, and X. Xiao, "Analysis of twitter data using a multiple-level clustering strategy," Dipartimento di Automatica e Informatica Politecnico di Torino - Torino, Italy.

[7] G. S. Supraja, J. Majumdar, and S. Ankalaki, "A big data methodology for sentiment analysis of twitter data," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 3, issue 7, 2015.

[8] J. Gill, M. Gobindgarh, S. Singh, and D. Baresary, "Big data: Big innovations in healthcare," Imperial Journal of Interdisciplinary Research (IJIR), vol. 1, issue 6, 18, 2015.

[9] Y. Revathi and Dr. S Mohan Kumar, "Review on importance and advancement in detecting sensitive data leakage in public network," *International Journal of Engineering Research and General Science*, vol. 4, issue 2, 2016.

[10] Y. Revathi and Dr. S Mohan Kumar, "A survey on detecting the leakage of sensitive data in public network," *International Journal of Emerging Technology and Advanced Engineering*.

[11] D. Babu and S. Mohan Kumar, "A Survey on secure communication in public network during disaster,".

Sulochana Panigrahi and S Mohan Kumar, "A survey on social data processing using apache Hadoop, Map-Reduce ," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 2, pp. 121-123, 2016.