Big Data Mining, Techniques, Handling Technologies and Some Related Issues: A Review

Gajendra Kumar[1] Prashant Richhariya[2]
[1,2]Department of Computer Science and Engineering
[1,2]Chhatrapati Shivaji Institute of Technology, Durg, Chhattisgarh

Abstract— The Size of the data is increasing day by day with the using of social site. Big Data is a concept to manage and mine the large set of data. Today the concept of Big Data is widely used to mine the insight data of organization as well outside data. There are many techniques and technologies used in Big Data mining to extract the useful information from the distributed system. It is more powerful to extract the information compare with traditional data mining techniques. One of the most known technologies is Hadoop, used in Big Data mining. It takes many advantages over the traditional data mining technique but it has some issues like visualization technique, privacy etc.

Key words: Big data, Data Mining, MapReduce, Hadoop, Big Data Analysis

I. INTRODUCTION

Today the world nothing without computer. Every day a lot of data is generated by the using of social media and by the many organizations. Till now the storing problem is not big issue but there is a lot problem comes while mining this huge amount of data. There are many question comes while mining the large set of data like processing speed, complexity, fault tolerance, capacity of data to mine etc. With the use of social site like facebook, twitter, linkedin and with use of ecommerce site like flipkart, snapdeal, jabong the size of the data is increasing rapidly worldwide. To handle these much amount of data is not possible, by the using of traditional data mining technique. So to overcome these types of problems the concept of Big Data is used. Big Data mining technique opening up new opportunities for enterprises to extract information from large volume of data in real time across multiple relational and non-relational data types [19].

The three V's are very important while working with Big Data [10]:

A. Volume:

It refers to the amount of data. The amount of data is varying according the organization. The growth in the data storage and processing technique is not limited only to the text data it is now more than the text data.

B. Variety:

It refers to types of data. Data can be stored in multiple formats. For example, database, excel, csv, access, it can be stored in a simple text file also. Sometimes the data is not even in the traditional format as we expect, it may be in the form of video, audio, forecast data, sensor data, pdf etc. It is the need of the organization to arrange it and get meaningful information.

C. Velocity:

It refers to the speed of data processing. The era where social media and e-commerce site are in pick position and there are many competitors. So they want to up to date in terms of data processing and information retrieval for users.

The above three are the most significant in but there is fourth aspect of big data which is also play a vital role in big data mining called Veracity.

Veracity also plays an important role to mining big data that defines the, how we are faithful with data and the quality of data. Is the data that is being stored, and mined meaningful to the problem being analyzed. In the data sources, there are two types of data i.e. structured and unstructured. With the structured data we need not to worry about the veracity while mining. But in case of unstructured data it is complex to mining and design a tool that is meaningful to the problem.

So what is Big Data: "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

II. TECHNIQUES

There are many data mining techniques which are also used in Big Data mining:

A. Regression:

Regression is a predictive data mining technique that predicts a number like Age, weight, distance, temperature, income, or sales. For example, a regression model can be used to predict a number of student in college, there average marks, their behavior. Regression analysis starts with set of data where target values are given.

B. Nearest Neighbor:

In this technique the values are predicted based on the predicted values of the records that are nearest to the record that needs to be predicted.

C. Clustering:

Clustering is a process in which the similar types of object or the object which have similar characteristics are grouped together as a cluster.

D. Classification:

Classification is a data mining technique where each item is classified into one of predefined the set of groups or classes.

E. Association Rule Mining:

Association Rule Mining is one of the best data mining techniques. In Association Rule Mining a pattern is discovered based on relationship between the items set by using given transaction.

III. TECHNOLOGIES

A. Hadoop:

Hadoop is open source software used for distributed computing. It provides Framework to store and process large amount of dataset in distributed system. Hadoop use the HDFS (Hadoop Distributed File System) system. HDFS is based on the GFS (Google File System). It provides the redundant storage for large set of data. In HDFS Data is distributed across all nodes at load time for efficient MapReduce processing.

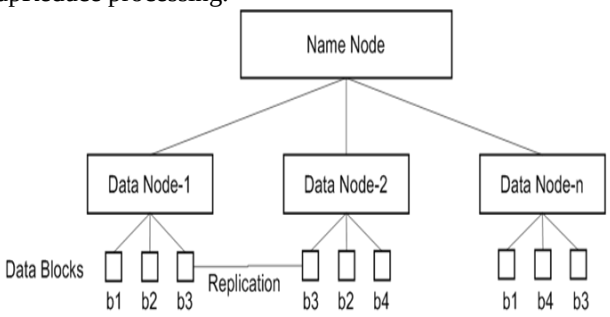

Fig. 1: HDFS Block Diagram

B. MapReduce:

MapReduce is a processing technique for generating large data sets with parallel distributed computing based on java. Basically MapReduce is method for distributing a task across multiple nodes. Each node processes data stored on the node. Automatic Distribution and Parallelization is the biggest advantage of MapReduce. MapReduce algorithm based on two important task:

- Map task
- Reduce task

Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples based on key value pair [9].

Map task syntax: map (key1, value) => list<key2, value2>

Means, for an input Map task returns a list containing zero or more (key, value) pairs:

- The output can be a different key from the input.
- The output can have multiple entries with the same key.

Reduce task takes the output from a map as an input and combines those data tuples into a smaller set of tuples based on key value pair. That is, it will generate a new list of reduced output [9].

Reduce task syntax: reduce (key2, list<value2>) => list<value3>

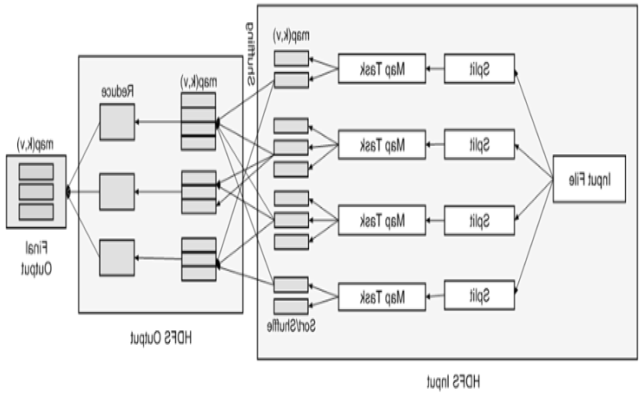

Fig. 1.2: MapReduce Working Block Diagram

C. NoSql:

NoSQL is a non-relational database management system, different from traditional relational database management systems. It is based on key-value store and document store, to storage and retrieval of large set of unstructured, semi-structured or structures data or data sets.

D. Hive:

Hive is data warehouse infrastructure tool to process large amount of structured data set in hadoop. Hive provides querying and managing large datasets residing in distributed storage. It allows SQL developers to write Hive Query Language (HQL) statements that are similar to standard SQL statements also called HiveQL

E. Pig:

Pig is a high level platform used with Apache Hadoop. It is high level high level scripting language. It creates MapReduce Program used with hadoop. To run the PIG scripts in Local mode, no Hadoop or HDFS installation is required. But in MapReduce mode, to run the scripts in mapreduce mode, Hadoop and HDFS installation is required.

IV. ISSUES RELATED TO BIG DATA MINING

When we work with big data there are different type of issue comes while working with it:

A. Security and Privacy:

Security and privacy are the most important issue in data mining. In data sources, the various forms of data, including structured data and unstructured data that comes from different resources. This data sources contains simple information of organization as well as it can include personally identifiable information, payment card data, intellectual property, health records, and much more. For these perspectives we need to strong security and privacy policy for unauthorized user.

B. Architecture:

To develop big data architecture is more complex compare to design a Business intelligence application and setting up data warehouse. Volume, velocity, and variety of the data make it difficult to extract information and business insight. The architectures for realizing big data solutions are composed of heterogeneous infrastructures, databases, visualization technique and analytics tools. Selecting the right architecture is the key to utilize the power of big data.

C. Real Time Data:

bigger data are not always better for analysis. It depends on the type of data, data is noisy or not and many other factors. In real time analysis data are frequently changed comes from different resources. In this case we need to design a powerful framework that manage data without delaying and provide a faster analysis tool for real time data.

D. Veracity:

Veracity defines the, how we are faithful with data and the quality of data. Is the data that is being stored, and mined meaningful to the problem being analyzed. In the data

sources there are two types of data i.e. structured and unstructured. With the structured data we need not to worry about the veracity while mining. But in case of unstructured data it is complex to mining and design a tool that is meaningful to the problem.

E. Visualization:

This is the hard part of big data. The key question for is, how to design visualization tool that easily manage the large amount of extracted data. How to design visualization tool that easily understandable by the user. How to design complex graphs and charts that include many meaningful variables of extracted data.

V. LITERATURE REVIEW

"Data Mining with Big Data" proposed by Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding: Find the fundamental challenge for the Big Data applications to explore the large volumes of data and extract useful information or knowledge for future actions. This paper also concerned about large-volume dataset, Complex data, growing data set that comes from different heterogeneous sources. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective and determining the key challenges for Big Data mining.

Wei Fan, Albert Bifet: study about "Mining Big Data: Current Status, and Forecast to the Future". Define the capabilities of Big Data, how the Big Data can mine the extremely large amount of data. Study about how the popular site like Google, facebook and twitter manage their data. Describe the importance of three V's of Big Data I.e. Volume, Variety, and the Velocity and Another Two V's of Big Data I.e. Variability: and Value. Brief description about the open source handling technologies: Apache Hadoop, Pig, Hive, Apache HBase etc.

"Mining Big Data in Real Time" paper presented by Albert Bifet: focus on the Real time streaming of data to obtain the currents status useful knowledge to help organization. Cost calculation Based on per hour usage and cost per hour and how much memory used

Xia Geng, Zhi Yang (Geng & Yang, 2013) study about the Data mining in cloud computing. Define cloud computing and how it overlaps some concept of distributed, grid and utility computing. Data mining with some technique like Association, Classification, Clustering etc. Data mining tools based on cloud like Weka4WS. Parallel programming model is a bridge between user needs and the underlying hardware system, it makes the parallel algorithm become more intuitive and more convenient for processing the large-scale data. Also focus on the concept of HDFS and MapReduce Infrastructure.

B R Prakash and Dr. M. Hanumanthappa Present th topic on, "Issues and Challenges in the Era of Big Data Mining", which describes the Defining Issues and challenges while working with big data. Define challenges like uncovering the hidden pattern between the different numerical parameters. Define difficulties in traditional data mining technique while handling unprecedented heterogeneous data.

VI. CONCLUSION

Mining the large data set is very tedious task with the using of traditional data mining technique. To overcome the problem, to mine the complex and large data set the Big Data technique is used. Big Data technique provide a Platform to organization to manage there their data smoothly. Hadoop is one of the popular techniques used in Big Data mining, based on Hadoop Distributed File System (HDFS). But It has some issues related to privacy, architecture, visualization tools and technique etc.

REFERENCES

[1] A. Bifet, 'Mining Big Data in Real Time', Informatica, vol. 37, no. 1, pp. 15-20, 2013.
[2] D. Che, M. Safran and Z. Peng, 'From Big Data to Big Data Mining: Challenges, Issues, and Opportunities', Springer, pp. 1-15, 2013.
[3] J. Lin and D. Ryaboy, 'Scaling big data mining infrastructure', SIGKDD Explor. Newsl., vol. 14, no. 2, p. 6, 2013.
[4] S. Ghuman, 'Big Data and its Handling Technologies-A Study', International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSSE), vol. 5, no. 6, pp. 29-32, 2015.
[5] J. Dean, Big Data, Data Mining, and Machine Learning. John Wiley & Sons, 2014.
[6] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, 'Data mining with big data', IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97-107, 2014.
[7] T. Rodrigues, '10 emerging technologies for Big Data - TechRepublic', TechRepublic, 2012. [Online]. Available: http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data. [Accessed: 03- Dec- 2015].
[8] Big Data - Made Simple, 'popular techniques for analysing Big Data'. [Online]. Available: http://bigdata-madesimple.com/26-popular-techniques-for-analysing-big-data. [Accessed: 03- Dec- 2015].
[9] www.tutorialspoint.com, 'Hadoop MapReduce', 2015. [Online].Available: http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm. [Accessed: 03- Dec- 2015].
[10] W. Fan and A. Bifet, 'Mining big data', SIGKDD Explor. Newsl., vol. 14, no. 2, p. 1, 2013.
[11] M. Goebel and L. Gruenwald, 'A survey of data mining and knowledge discovery software tools', SIGKDD Explor. Newsl., vol. 1, no. 1, pp. 20-33, 1999.
[12] J. Dittrich and J. QuianÃ©-Ruiz, 'Efficient big data processing in Hadoop MapReduce', Proc. VLDB Endow., vol. 5, no. 12, pp. 2014-2015, 2012.
[13] J. Lin and D. Ryaboy, 'Scaling big data mining infrastructure', SIGKDD Explor. Newsl., vol. 14, no. 2, p. 6, 2013.
[14] A. Gandomi and M. Haider, 'Beyond the hype: Big data concepts, methods, and analytics', International Journal of Information Management, vol. 35, no. 2, pp. 137-144, 2015.
[15] S. Zhang, S. Zhang, X. Chen and X. Huo, 'Cloud Computing Research and Development Trend', 2010

Second International Conference on Future Networks, 2010.
[16] B. Thakur and M. Mann, 'Data Mining for Big Data: A Review', International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 5, pp. 469-473, 2014.
[17] M. Chen, S. Mao and Y. Liu, 'Big Data: A Survey', Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, 2014.
[18] S. Suryawanshi and V. Wadne, 'Big Data Mining using Map Reduce: A Survey Paper', IOSR Journal of Computer Engineering, vol. 16, no. 6, pp. 37-40, 2014.
[19] N. Sawant and H. Shah, Big Data Application Architecture Q&A: A Problem-Solution Approach. Apress, 2013.

