

Big Data Mining Using Dataset Partition for Social Welfare

Kunwar Ambrish Singh¹, Pankaj Richhariya²

Department of Computer Science and Engineering

Bhopal Institute of Technology & Science, Bhopal

Corresponding Authors' email id: Kunwar.ambarish@gmail.com¹, richhariyap@gmail.com²

Abstract

Data exist all around the world. No one can measure the exact amount of data. As we know, we have a lot of data and we are struggling to store and analyze it. There is a solution to analyze the large volume of electronically stored data. The Technology in known as Big Data. These technique is capable to analysis the complex and real time data as well as complex hidden data, no matter what is the size of data. Today the population growth in very high and these leads to unemployment problem. Many people have tremendous skill but they don't get opportunity to show their skills. Partitioning Dataset and nearest data searching algorithm are used to get the useful data from the large data set. Here we highlight, how to mine the nearby data of employees for the users using Partitioning Dataset and nearest data searching algorithm.

Keywords: Big data, Data Mining, MapReduce, Hadoop, Big Data Analysis

INTRODUCTION

As we know the life is nothing without technology. Today every people uses different type's gadgets in their daily life. These gadgets generates a lot of data every seconds. Big Data is a technique to handle large set of data that cannot be processed using traditional data mining

technique or by common processing techniques used in traditional data mining. 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 [20].

Big Data is new term to analysis the huge amount of day to day operational and historical data. Hadoop is one the most popular technique used to extract the data from distributed server. Here we use the concept of MapReduce and “Frequent Neighbor Data Set Search Algorithm” to mine the data from the server. To implement this concept we develop a system Social Welfare which is very helpful for society. The concept of big data analysis has follow the following V's, which are very important while working with big data.

i. 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.

ii. 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. This task can be become an easy if data is in same format.

The real time or real world data comes from different heterogeneous sources with different format and to mining this type of data is extremely challengeable to organization. To mine these types of data we need the concept of Big Bata.

iii. 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 other organization also wants to fast processing and retrieving of data for business growth.

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.

iv. Veracity: It 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.

Map Reduce:

Hadoop provides a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by Map Reduce. In Map Reduce the programmer writes two functions: a map function and a reduce function, each of which defines a mapping from one set of key-value pairs to another. These functions are unaffected to the size of the data or the cluster that they are operating on, so they can be used unchanged for a small dataset and for a massive one.

i. 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.

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.

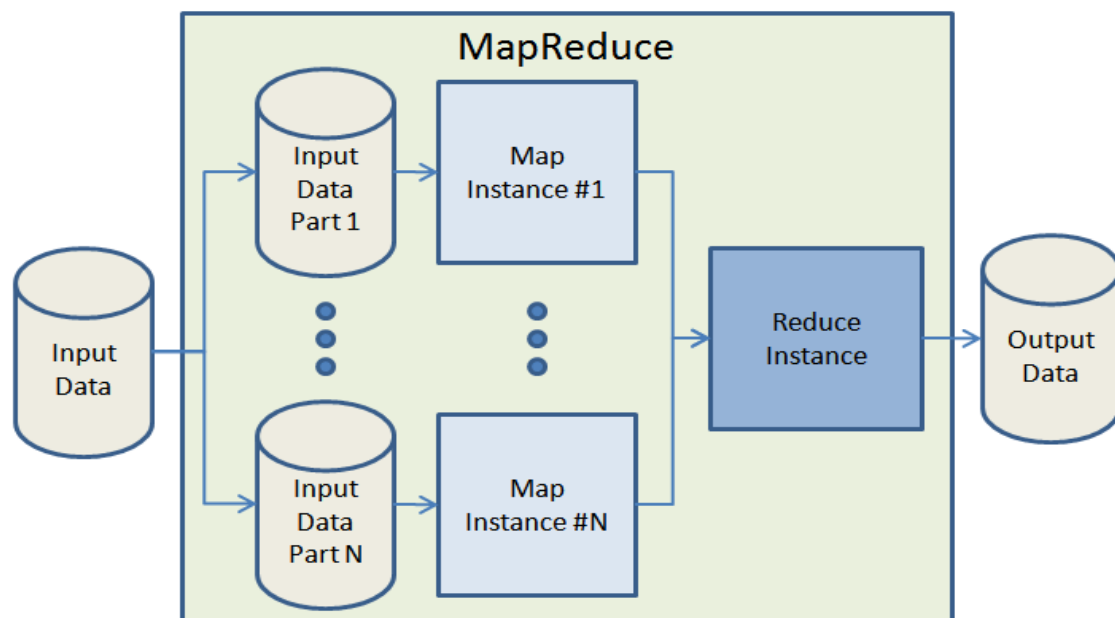


Figure 1: Map Reduce Processing model

ii. 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.

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

LITERATURE REVIEW

“Data Mining with Big Data” proposed by Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding [6] describes 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. Determining the key challenges for Big Data mining.

According to this, autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources are able to generate and collect information without involving any

centralized control. Complex and evolving relationships among data in Big Data and focus on how the social sites like twitter or facebook how establish the relationships between data. Elaborate the challenges with respect to Big Data Mining platform(Tier 1), Big Data Semantics and Application Knowledge(Tier 2) and Big Data Mining Algorithms(Tier 3).

“Harnessing Big Data for Social Good: A Grand Challenge for Social Work”(Grand Challenges for Social Work Initiative Working Paper) presented by Claudia J. Coulton, Robert Goerge, Emily Putnam-Hornstein, Benjamin de Haan [21]. The main focus in this paper facing the many critical social issues in day to day life. Include some related challenges for social works like Advance long and productive lives, Create social responses to a changing environment, Build financial capability for all, Ensure healthy development of all youth.

The advantage of digital evaluation is it produces a lot of data every second by using the social media, by many educational system and by many health agencies. Although there are evolutionary changes in technical field but the social sector are comparatively slow to incorporate to continuous flow of data

analysis because of its traditional data analysis system. To analysis these types of data the term Big Data analysis is used, but there are also some related security issues for social data with respect to human service, educational, and health records. Social sector must consider the ethical matter of using data for purposes of which individuals may not have been aware when providing personal information.

The data require careful cleaning and validation by data base administrative and specialists with a deep understanding of the data generation processes. In terms of big data mining the security and privacy is most important issue for social welfare.

“Mining Big Data: Current Status, and Forecast to the Future” given by Wei Fan, Albert Bifet [11] 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. Determine the future challenges of Big Data management and analysis that is very important like analytics drchitecture, statistical significance, distributed mining, Time evolving data, Compression, Visualization and Hidden Big Data. Describe the implementation of Big Data mining technique in various applications like business, technologies, health, Smart cities etc.

“Mining Big Data in Real Time” paper presented by Albert Bifet [1] focus on the Real time streaming of data to obtain the currents status and useful knowledge to help organization. Cost calculation based on per hour usage of application and how much memory used. Define new problem called structured pattern classification problem is defined as follows.

A set of examples of the form $(t; y)$ is given, where y is a discrete class label and t is a pattern. The goal is to produce from these examples a model $\hat{y} = f(t)$ that will predict the classes y of future pattern examples. In this paper, there also focus on the data mining for social media and twitter data stream analysis which provide the real time messages to the users as well as Sentiment analysis of all users, can be cast as a classification problem where the

task is to classify messages into two categories depending on whether they convey positive or negative feelings.

“Data Mining for Big Data” paper presented by Bharti Thakur, Manish Mann [17]. Describe the types of Big Data, structured and unstructured data. Define HACE theorem to model Big Data characteristics i.e. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These Characteristics make challenges for discovering useful knowledge for large set of data. Importance of three V’s of big data.

OBJECTIVE

The world is transforming and thus with each coming technological revolution data securing techniques are emerging as well. The developments in digital fields have begot various alternatives to store this massive accumulating data which goes even more complex to store with the growing time. These techniques have assisted the societies in exploiting such social, psychological and organizational data to cater to many of the societies' problems. New discovered techniques overtake the old data mining techniques in

terms of enhanced data managing capabilities. Data from computerized social service, education, and health records; open data portals; social media posts; web searches; and mobile GPS devices and sensors can prove effective in eradicating the problem of unemployment by conferring solutions to this alarming issue.

The complex data clusters are resolved using Big Data mining algorithms. The algorithms viz "Partitioning Algorithm" and "Nearest neighborhood Algorithm" have made it possible to employ the useful data for social welfare.

The motives include minimization in response time in order that process becomes speedier and also the increase in the efficiency of the process. The incorporation of these techniques can maintain a continuous flow of data analytics to restrict the loss of useful data. The MapReduce technique employed, resolves the difficulties to execute massive data. The facilitation to mine the requisite data to benefit both the worker and the employer persists.

A framework like Hadoop provides a platform to avail tools like MapReduce which can execute considerable data in

comparison to the tools like Weka which could execute data just up to 15MB. The technique helps employers to explore workers accordingly their requisiteness by searching in area included thus excluding the remaining areas leading to a quick response. This partitioning is based on city and even the categories of partitions may differ accordingly.

On the other hand the site displays the available unemployed workers leaving the ones already employed. This technique reflects idea of mutual assistance which is a social solution. The ineffectiveness of technological progresses to utilize the data analytics to propel effective social solutions is due to the absence of these data-driven algorithms.

Once these algorithms come into being thesis will demonstrate the long-term social welfare. The defy precludes in the provision of the fundamental cognizance of big data managing cum managing techniques with which a few are aquatinted to spread this perspective of social upheaval extirpating unemployment.

METHODOLOGY

To mine the data from the data set, firstly the process will create the partition on the basis of frequently searched area. This

partition is based on the user search behavior for particular area. After creating the partition the actual process will be executed. The primary task is to mine the actual data set based on users key word.

Here the partition will work as catalyst for the process. Because here the partition only store the data of particular area based on users behavior so it will only search within the partition if data not found in partition then it will go for entire warehouse or dataset.

If data found in partition then it will directly go to MapReduce phase for further analysis. If data found in main source (entire database or warehouse) then it will perform two task. Firstly it will go to MapReduce phase for analysis. And then the fetched data will be added to relevant partition.

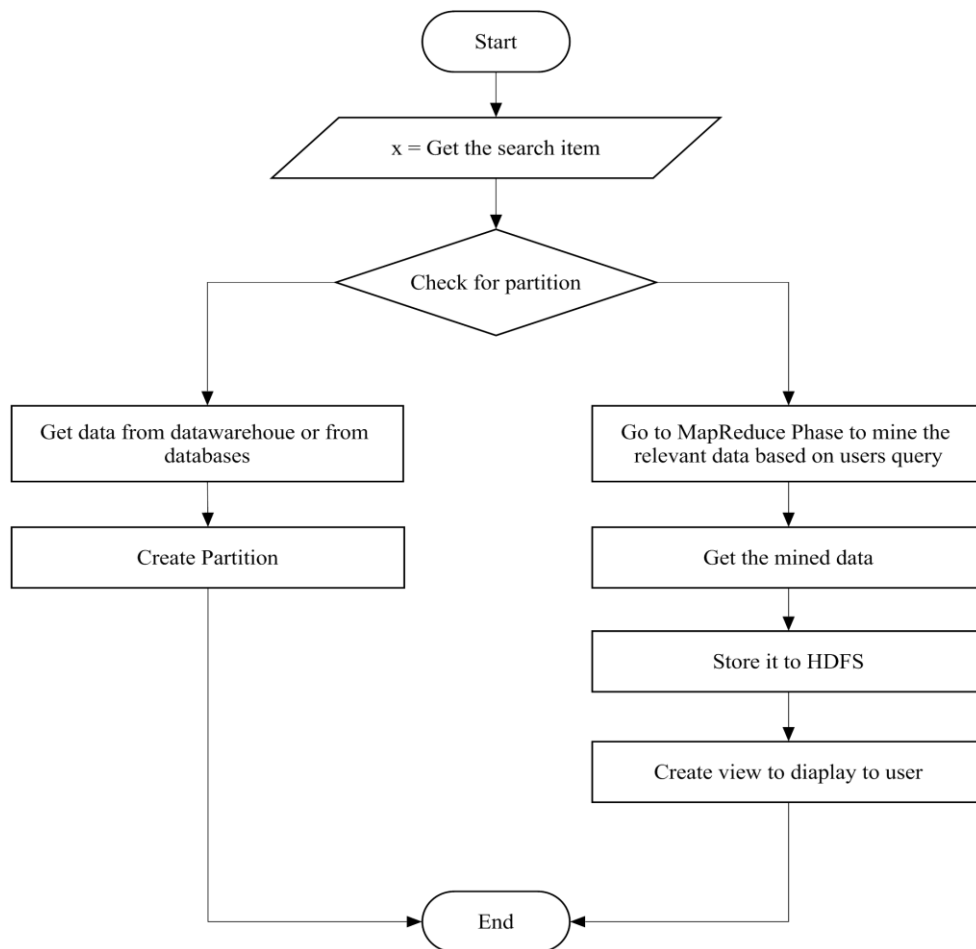


Figure 3: Block Diagram for dataset partition and MapReduce Processing

RESULT

Success and failure is depend on the. And the output is depends on the input, Processes and applying algorithm. The result analysis after applying the partition algorithm and with normal execution is shown below:

Table: 1

Execution Round	Execution Time in Milliseconds	
	Normal Data Extraction	Using Dataset Partition
Round 1	5900	6708
Round 2	5886	3912
Round 3	5882	3927
Round 4	5891	3903

After analyzing the above result, in first round the normal data extraction will take less time compared with the dataset partition. But after round 1 the normal data extraction will take always more time compared with dataset partition.

CONCLUSION

Today mostly people uses the different types of gadgets which generates a lot of data every seconds. Due to digital revolution the social sector are hided and not get much benefit with new technologies. The algorithms used in the technique make the process efficient and minimize the response time by restricting the search to the allotted city rather than the whole state. Community groups have also been slow to access the data that could individually or collectively advance the quality of their members' lives. If social work does not invest in building the capacity to fully make use of its existing big data, technologically savvy entities without a full appreciation of the data's social context could take up the task for the wrong reasons. Though there is a growing demand to prove what works, human service systems often lack the data management and analysis infrastructure necessary to efficiently evaluate programs. The wholesomeness includes removal of unemployment which cripples the society

owing to the improper use of data managing algorithms.. The project focuses on idea to derive social help from Big Data mining technique to mine the data of local person who have skill and knowledge.

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. Wikipedia, 'Apache Hadoop', 2015. [Online]. Available: https://en.wikipedia.org/wiki/Apache_Hadoop. [Accessed: 03- Dec- 2015].
10. www.tutorialspoint.com, 'Hadoop MapReduce', 2015. [Online]. Available: http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm. [Accessed: 03- Dec- 2015].
11. W. Fan and A. Bifet, 'Mining big data', SIGKDD Explor. Newsl., vol. 14, no. 2, p. 1, 2013.
12. 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.
13. J. Dittrich and J. Quijano-Ruiz, 'Efficient big data processing in Hadoop MapReduce', Proc. VLDB Endow., vol. 5, no. 12, pp. 2014-2015, 2012.
14. J. Lin and D. Ryaboy, 'Scaling big data mining infrastructure', SIGKDD Explor. Newsl., vol. 14, no. 2, p. 6, 2013.
15. 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.
16. S. Zhang, S. Zhang, X. Chen and X. Huo, 'Cloud Computing Research and Development Trend', 2010 Second International

Conference on Future Networks,
2010.

17. 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.

18. M. Chen, S. Mao and Y. Liu, 'Big Data: A Survey', Mobile Networks and Applications, vol. 19, no. 2, pp. 171-209, 2014.

19. 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.

20. N. Sawant and H. Shah, Big Data Application Architecture Q&A: A Problem-Solution Approach. Apress, 2013.

21. Claudia J. Coulton, Robert Goerge, Emily Putnam-Hornstein, Benjamin de Haan. 'Harnessing Big Data for Social Good: A Grand Challenge for Social Work. American Academy of Social Work and Social Welfare', 2015