4th International Conference on Science, Technology and Management
India International Centre, New Delhi (ICSTM-16)
15th May 2016, www.conferenceworld.in ISBN: 978-81-932074-8-2

BIG DATA MINING TO SEARCH FREQUENT NEAREST DATASET FOR SOCIAL WELFARE

Gajendra Kumar[1], Prashant Richhariya[2]

[1,2] *Department Of Computer Science and Engineering, Chhatrapati Shivaji Institute of Technology, Durg, Chhattisgarh(India)*

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 Frequent 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, Big Data mining.

I. INTRODUCTION

Today every people uses different types of 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 [1].

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 the concept of MapReduce has been used to implement the "Dataset Partition Algorithm" and "Frequent Neighbor Data Set Search Algorithm" to mine the data from the server. To implement this concept we are developing a system Social Welfare which is beneficial for the society. The concept of big data analysis has follow the following V's, which are very important while dealing with big data.

- **Volume:** It refers to the amount of data. The amount of data is varying according the organization [2]. 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.

- **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.
- **Velocity:** It refers to the speed of data processing [2]. 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:** 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.

MapReduce:

Hadoop provides a reliable shared storage and analysis system. The storage is provided by HDFS and analysis by Map Reduce[3]. 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[3]. 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.

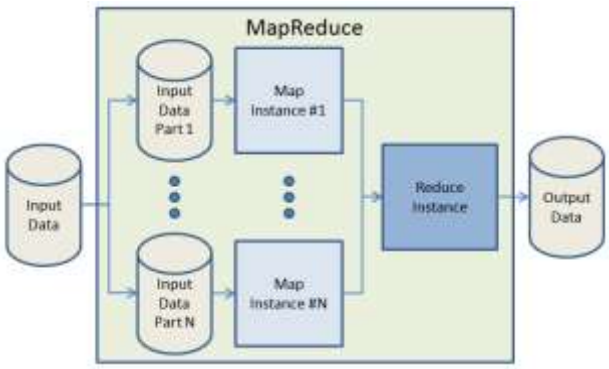

Figure 1 : MapReduce Processing model

- **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.
- **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>

II. LITERATURE SURVEY

"Data Mining with Big Data" proposed by Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding [4] 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. 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).

"Mining Big Data: Current Status, and Forecast to the Future" given by Wei Fan, Albert Bifet [2] 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.

"Mining Big Data in Real Time" paper presented by Albert Bifet [5] 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 solving pattern 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.

"Data Mining for Big Data" paper presented by Bharti Thakur, Manish Mann [6]. 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.

III. OBJECTIVES

Often a technology launch leaves people unemployed, thus it must be a technology to fetch them employment. Big data mining for social welfare is one such initiative towards employment in which the required data is mined. Here the system is developed to efficient use of MapReduce framework for mining the day to day data to provide the information to user based on the searching keywords. The Social Welfare system provides a platform to use skills of the person who have knowledge about household problem.

IV. 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 user's keyword.

Here the partition will act as a 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.

Pseudo Code: Complete process

1. get the search keyword.
2. apply Partitioning Algorithm
3. get the data for search item from partitioned data set.
4. get the actual searched and nearest data for searched location from portioned data set.
5. exit.

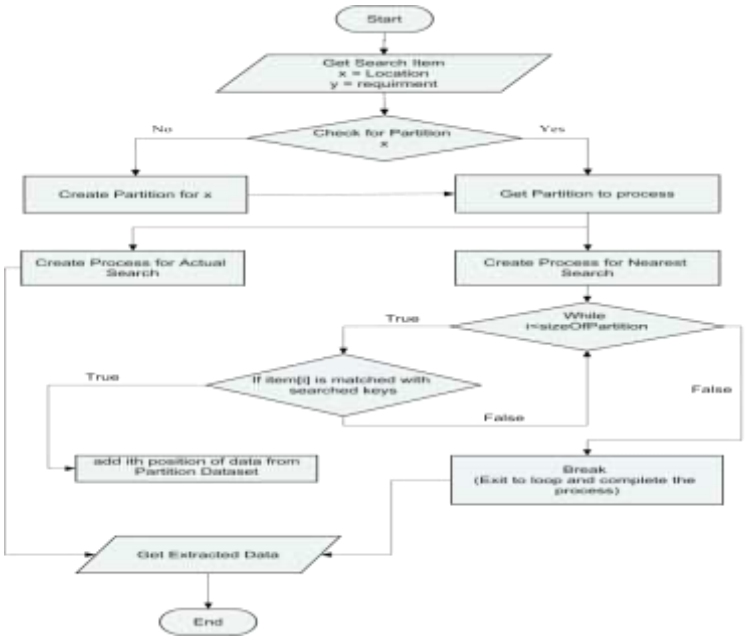

Figure 2 : Block Diagram for dataset partition and frequent nearest dataset searching to extract the data .

V. RESULT

Success or failure of every project is depend on the result and the result is depend on the applied algorithm, input, and the output of the project. If the algorithm is errorless and input is compatible to the applied algorithm then the

out must be accurate. Here we discussed the result of project when we applied the algorithm "Dataset Partition" and "Frequent Nearest Dataset Search".

Execution Round (Searched & Nearest Data extraction)	Execution Time in Milliseconds	
	Normal Data Extraction	Using Dataset Partition
Round 1	6100	6901
Round 2	5987	4012
Round 3	5984	4021
Round 4	5993	4002

Table 1: Comparision between normal execution & using dataset partition.

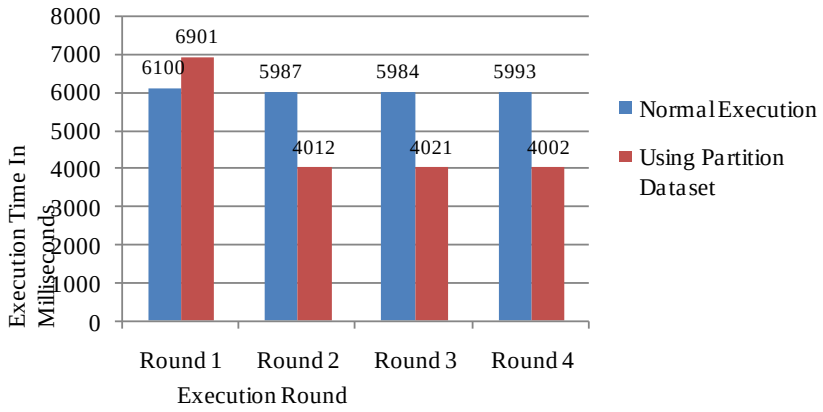

Figure 3:Graphical representation of comparison between normal execution & using dataset partition.

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.

VI. CONCLUSION

Today most of the people uses different types of gadgets which generates a lot of data every seconds. Here the system Social Welfare is aimed to provide a platform to both one who have a skill and one who don't have the knowledge about the local household workers. This system provide a platform to skilled person to utilize their skill so that they can get a part time job and earn money. This 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. To mine day to day

data for social welfare here the system used the MapReduce processing technique, "Dataset Partition Algorithm" and "Frequent Neighbor Data Set Search Algorithm"

REFERENCES

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

[2] W. Fan and A. Bifet, 'Mining big data - Current Status, and Forecast to the Future', *SIGKDD Explor. Newsl.*, vol. 14, no. 2, p. 1-5, 2013.

[3] www.tutorialspoint.com, 'Hadoop MapReduce', 2015. [Online]. Available :http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm. [Accessed: 03- Dec- 2015].

[4] 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.

[5] A. Bifet, 'Mining Big Data in Real Time', *Informatica*, vol. 37, no. 1, pp. 15-20, 2013.

[6] 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.

[7] D. Che, M. Safran and Z. Peng, 'From Big Data to Big Data Mining: Challenges, Issues, and Opportunities', *Springer*, pp. 1-15, 2013.

[8] J. Dean, *Big Data, Data Mining, and Machine Learning*. John Wiley & Sons, 2014.

[9] S. Zhang, S. Zhang, X. Chen and X. Huo, 'Cloud Computing Research and Development Trend', *2010 Second International Conference on Future Networks*, 2010.