JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Data Mining with Big Data: It's Issues and Challenges

Swati Namdev

Department of Computer Science Career College, Bhopal, India

Corresponding Authors: swati.tailor@gmail.com

Available online at: www.ijcseonline.org

Accepted: 06/Dec/2018, Published: 31/Jan/2019

Abstract— Big Data could be a new term wont to determine the datasets that because of their large size and complexity. Big Data are currently speedily increasing altogether science and engineering domains, as well as physical, biological and medical specialty sciences. Big Data processing is that the capability of extracting helpful data from these large datasets or streams of data, that because of its volume, variability, and velocity, it absolutely was unfeasible before to try to to it. The Big data challenge is turning into one in every of the foremost exciting opportunities for the following years. This paper includes the knowledge concerning what is Big Data, Data Mining, Data Mining with Big Data, Challenging issues and its related work.

***Keywords* —** Big Data, Data mining, Datasets, Data Mining Algorithms

I. INTRODUCTION

Google now a day is our need. On the other hand, today is the age of Google. The thing which is unknown for us, we see it on Google. And in fractions of seconds we get the number of links as a result. This would be the better example for the processing of Big Data. This Big Data is not any totally different thing than out regular term data. Simply big may be a keyword used with the data to identify the collected datasets due to their massive size and complexity? We have a tendency to can't manage them with our current methodologies or data mining software tools.

Another example, the first strike of Pakistani unit of measurement Hajare triggered varies of tweets at intervals a pair of hours. Among of those tweets, the special comments that generated the foremost discussions actually discovered the final public interests. Such on-line discussions provides a fresh suggests that to sense the final public interests and generate feedback in period, and are mostly appealing compared to generic media, like radio or TV broadcasting. This instance demonstrates the rise of large data applications. The data assortment has mature hugely and is on the way facet the ability of usually used software package tools to capture, manage, and technique at intervals a tolerable time

II. BIG DATA AND DATA MINING

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored at the server of WhatsApp, as most of us, daily use the WhatsApp; we upload various types of information, upload photos. All the data get stored at the data warehouses at the server of WhatsApp. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Instagram. These are the good real-time examples of the Big Data. Another best example of Big data would be, the readings taken from an electronic microscope of the universe. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining.

So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information. As shown in fig 1 below.

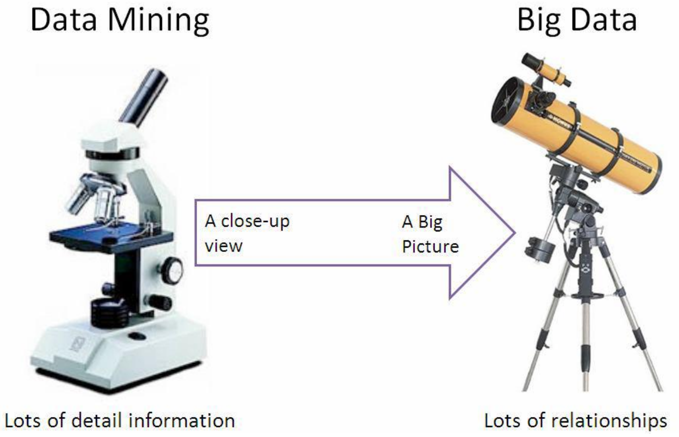

Fig.1 Data Mining with Big Data

III. BIG DATA KEY FEATURES

The features of Big Data are:

- It is huge in size.
- The data keep on changing time to time.
- Its data sources are from different phases.
- It is free from the influence, guidance, or control of anyone.
- It is too much complex in nature, thus hard to handle.

It’s vast in nature as a result of, there's the gathering of data from varied sources along. If we tend to contemplate the instance of WhatsApp, plenty of numbers of individuals are uploading their data in varied sorts like text, pictures or videos. The people as well keep their data changing continuously. This tremendous and instantly, time to time dynamic stock of the data is keep during a warehouse. This huge storage of data needs large space for actual implementation. Because the size is simply too massive, nobody is capable to manage it oneself. The Big Data must be controlled by dividing it in teams.

Due to largeness in size, decentralized management and completely different data sources with differing kinds the Big Data becomes a lot of troublesome and tougher to handle. we tend to cannot manage them with the native tools those we tend to use for managing the regular data in real time. For many necessary Big Data-related applications, like Google, Flicker, Facebook, an oversized variety of server farms are deployed everywhere the globe to confirm nonstop services and fast responses for native markets.

IV. CHALLENGING ISSUES

There are three sectors at which the challenges for Big Data arrive. These three sectors are:

- Design of mining algorithms
- Privacy
- Mining platform

Basically, the Big Data is hold on at totally different places and additionally the Data volumes might get multiplied because the data keeps on increasing incessantly. So, to gather all the data hold on at totally different places is that a lot of high-priced. Suppose, if we have a tendency to use these typical data mining strategies (those strategies that are used for mining the little scale data in our personal computer systems) for mining of Big Data, so it might become associate obstacle for it. as a result of the everyday strategies are needed data to be loaded in main memory, although we've got super massive main memory.

To maintain the privacy is one among the most aims of Data mining algorithms. Presently, to mine information from Big Data, parallel computing primarily based algorithms like MapReduce are used. In such algorithms, large data sets are divided into variety of subsets so, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to fulfill the goal of Big Data mining. During this whole procedure, the privacy statements clearly break as we have a tendency to divide the single Big Data into variety of smaller datasets.

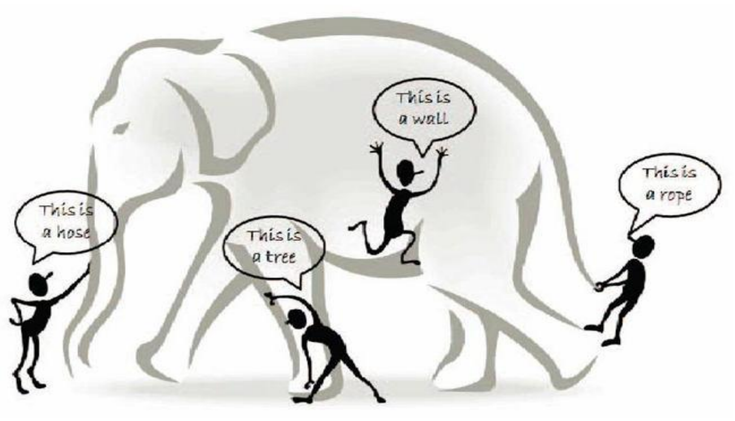

Fig. 2 Blind men and the giant elephant.

While planning such algorithms, we have a tendency to face varied challenges. As shown within the figure a pair of on top of, there are blind men perceptive the large elephant. Most are attempting to predict their conclusion on what the factor is truly. Someone is voice communication that the factor could be a hose; somebody says it’s a tree or pipe etc. Truly most are simply perceptive some a part of that big elephant and not the entire, that the results of every blind person’s prediction are some things totally different than truly what it's.

Similarly, after we divide the Big Data in to variety of subsets, and apply the mining algorithms on those subsets, the results of these mining algorithms won't continually purpose us to the particular result as we would like after we collect the results along.

V. RELATED WORK

On the extent of mining platform sector, at present, parallel programming models like MapReduce are being employed for the aim of research and mining of knowledge.

MapReduce may be a batch-oriented parallel computing model. There's still a particular gap in performance with relative databases. Rising the performance of MapReduce and enhancing the time period nature of large-scale data processing have received a big quantity of attention, with MapReduce parallel programming being applied to several machine learning and data mining algorithms. Data Mining algorithms typically must scan through the coaching data for getting the statistics to unravel or optimize model.

For those people, who intend to hire a third party such as auditors to process their data, it is very important to have efficient and effective access to the data. In such cases, the privacy restrictions of user may be faces like no local copies or downloading allowed, etc. So there is privacy-preserving public auditing mechanism proposed for large scale data storage.[1] This public key-based mechanism is used to enable third-party auditing, so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ. For example, when we go to the doctor for the treatment, that doctor's treatment program continuously adjusts with the conditions of the patient. Similarly the knowledge. For this, Wu [2] [3][4] proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find.

VI. CONCLUSION

Big Data is going to continue increasing throughout the next years, and everyone information someone can comprise to pander to weigh more amounts of data each year. This data goes to be a lot of various, larger, and faster. We tend to mention some insight regarding the topic; additionally to what we tend to think about are the most issues and also the main challenges for the longer term. Big Data is turning into the new Final Frontier for scientific data analysis and for business applications. We tend to are at the start of a brand new era wherever huge data processing can facilitate United States of America to find data that nobody has discovered before.

ACKNOWLEDGEMENT

I would like to thank to my friends for providing feedback and suggestions for improving my ideas. I also wish to thank my husband and parents for their support and encouragement throughout my study. Finally thanks to all the authors which are used in my research paper as references.

REFERENCES

[1] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Secure Cloud Storage” IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[2] X. Wu and S. Zhang, “Synthesizing High-Frequency Rules from Different Data Sources,” IEEE Trans. Knowledge and Data Eng., vol. 15, no. 2, pp. 353-367, Mar./Apr. 2003.

[3] X. Wu, C. Zhang, and S. Zhang, “Database Classification for Multi-Database Mining,” Information Systems,vol. 30, no. 1, pp. 71- 88, 2005

[4] K. Su, H. Huang, X. Wu, and S. Zhang, “A Logical Framework for Identifying Quality Knowledge from Different Data Sources,” Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[5] E.Y. Chang, H. Bai, and K. Zhu, “Parallel Algorithms for Mining Large-Scale Rich-Media Data,” Proc. 17th ACM Int’l Conf. Multimedia, (MM ’09,) pp. 917-918, 2009.

[6] D. Howe et al., “Big Data: The Future of Biocuration,” Nature, vol. 455, pp. 47-50, Sept. 2008.

[7] A. Labrinidis and H. Jagadish, “Challenges and Opportunities with Big Data,” Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.

[8] Y. Lindell and B. Pinkas, “Privacy Preserving Data Mining,” J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.