Analysis of Big Data Algorithms

Vecha.Harshini[1], **Nelluru.Anjaneyulu**[2]

[1]**M.Tech (CSE), Malineni Lakshmaiah Women's Engineering College ,Pulladigunta, Vatticherukur, Prathipadu Road, Guntur, Andhra Pradesh 522017**

[2]**Assistant. Professor, Malineni Lakshmaiah Women's Engineering College ,Pulladigunta, Vatticherukur, Prathipadu Road, Guntur, Andhra Pradesh 522017**

Abstract: Data means collection of information. Every day there lots is data is coming from all over the world. Everywhere in the world we can find the data. These data can generate from various sources and used for various purpose. Data mining is the research domain which is used to generate the analysis, computations etc is done by data mining. This meaningful information can be derived using some data mining tasks. In short we can call big data as an ‟asset‟ and data mining is a „handler‟ that is used to provide beneficial results. To perform these analysis data mining algorithms can be used and also the big data methods.

Keywords: Big data, Data Mining, HACE theorem, structured and unstructured.

I. INTRODUCTION

Big data refers to the enormous amount of structured and unstructured data that overflow the organization. If the overflowed data is used in a proper way it leads to meaningful information. When big data is compared to traditional databases it includes a large number of data which requires more processing in real time. It also provides opportunities to discover new values, to understand an in-depth knowledge from hidden values and also provides space to manage those data effectively. Big Data concern large-volume, complex, growing datasets with multiple data sources. With the fast development of networking, data storage and data collection capacity, big data are now expanding in all science and engineering domains, including physical, biological and biomedical sciences.[1]. Data Mining is a task of identifying relevant and significant information from large data set.

II. BIG DATA WITH DATA MINING

Generally big data refers to a collection of large volumes of data and these data are generated from various sources such as internet, social media, business organizations etc., with these data some useful information can be extracted with the help of data mining. Data mining is a technique for discovering interesting patterns as well as descriptive, understandable models from large scale data [2].

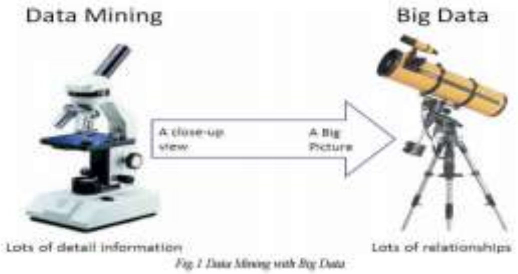

Fig 1 Data Mining with Big Data

KEY FEATURES OF BIG DATA

The features of Big Data are:

- It is huge in size.
- The data keep on changing time 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 huge in nature because, there is the collection of data from various sources together. If we consider the example of Facebook, lots of numbers of people are uploading their data in various types such as text, images or videos. The people also keep their data changing continuously. This tremendous and instantaneously, time to time changing stock of the data is stored in a warehouse. This large storage of data requires large area for actual implementation. As the size is too large, no one is capable to control it oneself. The Big Data needs to be controlled by dividing it in groups. Due to largeness in size, decentralized control and different data sources with different types the Big Data becomes much complex and harder to handle. We cannot manage them with the local tools those we use for managing the regular data in real time. For major Big Data-related applications, such as Google, Flicker, Facebook, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets.

V. RELATED WORK

On the level of mining platform sector, at present, parallel programming models like MapReduce are being used for the purpose of analysis and mining of data. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve 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.

III. BIG DATA CHARACTERISTICS -HACE THEOREM :

Big Data starts with large volume, heterogeneous autonomous sources with distributed and decentralized control and seeks to explore complex and evolving relationships among data [1].These characteristics makes it an extreme challenge for discovering useful information from big data. In connection with this scenario, let us imagine a scenario where blind people are asked to draw the picture of an elephant. The information collected by each blind people will be such that they may think the trunk as a „wall", leg as a „tree", body as a „wall" and tail as a „rope". In this case one blind men can exchange information with other which may be biased.

i. Vast data with heterogeneous and diverse sources One of the fundamental characteristics of big data is the large volume of data represented by heterogeneous and diverse dimensions. For example in the biomedical world, a single human being is represented as name, age, gender, family history etc., For X-ray and CT scan images and videos are used. Taking the example heterogeneity refers to the different types of representations of same individual and diverse refers to the variety of features to represent single information [1].

ii. ii. Autonomous with distributed and de-centralized control These are the main characteristics of big data. Since the sources are autonomous, i.e., automatically generated, it generates information without any centralized control. We can compare it with World Wide Web (WWW) where each server provides a certain amount of information without depending on other servers.

iii. Complex and Evolving relationships As the size of data becomes infinitely large, the complexity and relationships of data also becomes large. In the early stages when data are so small, there is no difficulty in establishing relationships among data. As the size of data become larger in the current scenario, data are generated from social media and other sources, so there arise complexity in establishing relationships. Such a complication is becoming part of the reality for big data applications, where the key is to take complex data relationships, along with the evolving changes into consideration to discover useful patterns from big data collections [1].

Conclusion:

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed some insights about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited to participate in this intrepid journey.

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.