G Vijaya Kumari* et al.
(IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH
Volume No.8, Issue No.1, December – January 2020, 9477-9481.

Evolution to Big Data Analytics Techniques and Challenging Issues in Data Mining With Big Data

G. VIJAYA KUMARI
Associate professor
Department of CSE, BVCITS,
Batlapalem.

Dr.P.VAMSI KRISHNA RAJA, M.Tech,Ph.D.
Director, Centre for Innovation Incubation &Startup.
R&D CSE, Swarnandhra College of Engineering and Technology(Autonomous), Narsapur,AP.

Abstract— **Big Data is another term used to recognize the datasets that because of their enormous size and multifaceted nature. Big Data are currently quickly growing in all science and engineering domains, including physical, natural and biomedical sciences. Big Data mining is the capacity of separating helpful information from these huge datasets or floods of data, that because of its volume, changeability, and velocity, it was impractical before to do it. The Big Data challenge is getting one of the most energizing open doors for the following years. In the present time of digitization, we take a shot at the variety of data. Colossal measure of data will be prepared by Google, Microsoft and Amazon. Regular routine these organization prepared huge measure of data. In such way we have to require some approach to adjust the innovation in with the end goal that every one of the data will be prepared adequately. Big Data is a developing concept that depicts imaginative systems and innovations to break down enormous volume of complex datasets that are exponentially produced from different sources and with different rates. Data mining procedures are giving extraordinary guide in the region of Big Data examination, since managing Big Data are big difficulties for the applications. Big Data examination is the capacity of removing valuable information from such colossal datasets. This paper exhibits a writing survey that incorporate the significance, difficulties and applications of Big Data in different fields and the various methodologies utilized for Big Data Analysis utilizing Data Mining procedures. The discoveries of this audit give important information to the analysts about the primary patterns in research and examination of Big Data utilizing diverse investigation domains. This examination paper incorporates the information about what is big data, Data mining, Data mining with big data, Challenging issues and its related work.**

Key words — **Big Data, Data mining, challenging issues, Datasets, Data Mining Algorithms**

I. INTRODUCTION

Today is the period of Google. The thing which is obscure for us, we Google it. What's more, in fractions of seconds we get the quantity of connections subsequently. This would be the better model for the preparing of Big Data. This Big Data isn't any unexpected thing in comparison to out normal term data. Simply big is a catchphrase utilized with the data to recognize the gathered datasets because of their enormous size and intricacy? We can't oversee them with our present philosophies or data mining programming instruments. Another model, the principal strike of Anna Hajare activated number of tweets inside 2 hours. Among every one of these tweets, the exceptional remarks that created the most discussions really uncovered the open interests. Such online discussions give another way to detect the open interests and create criticism progressively, and are for the most part engaging contrasted with conventional media, for example, radio or TV broadcasting.

This model demonstrates the ascent of Big Data applications. The data collection has developed massively and is beyond the capacity of commonly utilized programming instruments to catch, oversee, and process inside a passable time. Right now, investigators have gigantic measures of data accessible close by. Applications where data collection has developed colossally and is beyond the capacity of commonly utilized programming devices to catch, oversee, and process inside a "bearable slipped by time." The most key test for Big Data applications is to investigate the huge volumes of data and concentrate valuable information or information for future actions . Much of the time, the information extraction process must be extremely proficient and near ongoing in light of the fact that putting away all watched data is almost infeasible.

Data is being created at a consistently expanding rate. There has additionally been an acceleration in the proportion of machine-produced and unstructured data (photographs , recordings, web based life bolsters, etc) contrasted with organized data to such an extent that 80% or a greater amount of all data property are presently unstructured and new methodologies and advances are required to get to, connect, oversee and gain understanding from these data sets.

The commonly acknowledged definition of big data originates from Gartner who characterize it as high-volume, high-velocity as well as high-variety information resources that request financially savvy, creative types of information preparing for improved understanding, decision making, and procedure optimization. These are known as the "three Vs". A few experts likewise examine big data regarding esteem (the economic or political worth of data) and veracity (vulnerability presented through data quality issues). Government offices hold or approach a consistently expanding abundance of data including spatial and location data, just as data gathered from and by residents. Experience recommends that such data can be used in manners that can possibly change administration structure and conveyance with the goal that personalized and streamlined administrations, that precisely and explicitly address person's issues, can be conveyed to them in an auspicious way.

Big Data begins with huge volume, heterogeneous, autonomous sources with conveyed and decentralized control, and looks to investigate complex and developing relationships among data. These attributes make it an extraordinary test for finding helpful information from the Big Data.

Improved help conveyance could cover regions as differing as remote medicinal diagnostics, significant framework the board, personalized standardized savings benefits conveyance, improved specialist on call and crisis administrations, reduction of deceitful or crime across both government and private divisions, and the advancement of inventive new administrations as the development and accessibility of Public Sector Information (PSI) turns out to be increasingly pervasive.

The private division holds enormous measures of data about its clients and by and large leads the route in how this data is examined and used to make new plans of action and administrations. Offices have the chance to gain from the innovations happening in the private domain to work all the more proficiently and convey benefits all the more adequately while guaranteeing that protection and security matters are painstakingly considered.

Apache Hadoop - The Apache Hadoop venture creates open-source programming for dependable, adaptable, disseminated figuring. The Apache Hadoop programming library is a structure that takes into consideration the appropriated preparing of huge data sets across groups of PCs utilizing an a huge number of computational autonomous PCs and petabytes of data. Hadoop was gotten from Google's Map Reduce and Google File System (GFS).

HDFS (Hadoop Distributed File System)- The Hadoop Distributed File System (HDFS) is an appropriated document framework giving adaptation to non-critical failure and intended to run on item equipment. HDFS gives high throughput access to application data and is appropriate for applications that have enormous data sets. Hadoop gives an appropriated record framework (HDFS) that can store data across a huge number of servers and a methods for running work (Map/Reduce occupations) over those machines, running the work close to the data. HDFS has ace/slave engineering. Enormous data is consequently part into lumps which are overseen by various hubs in the hadoop bunch.

HBASE-HBase is a segment arranged database the executives framework that sudden spikes in demand for top of HDFS. It is appropriate for inadequate data sets, which are common in numerous big data use cases. In contrast to relational database frameworks, HBase doesn't bolster SQL.In reality, HBase is definitely not a relational database by any means. HBase applications are written In Java much like a run of the mill MapReduce application.

Guide Reduce - Map decrease is a product structure acquainted by Google in 2004 with help dispersed processing on enormous data sets on groups of PCs. Guide Reduce is a programming model for handling and creating huge data sets. Clients determine a guide function that procedures a key/esteem pair to produce a lot of moderate key/esteem sets and a lessen function that unions every middle of the road esteem related with a similar transitional key.

"Guide" step: The ace hub takes the information, partitions it up into littler sub-issues, and circulates them to laborer hubs. A specialist hub may do this again thusly, prompting a staggered tree structure. The laborer hub forms the littler issue, and passes the appropriate response back to its lord hub. Guide takes one sets of data with a sort in one data space, and returns a rundown of sets in an alternate domain: Map (k1, v1) → list (K2, v2)

"Diminish" step: The ace hub at that point gathers the responses to all the sub-issues and consolidates them here and there to shape the yield – the response to the issue it was initially attempting to unravel. The Reduce function is then applied in corresponding to each gathering, which thusly creates a collection of qualities in a similar domain: Reduce (K2, list (v2)) → list (v3)

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 Facebook, as most of us, daily use the Facebook; we upload various types of information, upload photos. All the data get stored at the data warehouses at the server of Facebook. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Flicker. 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.

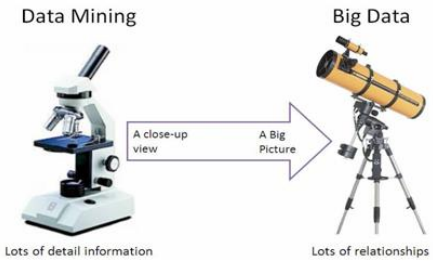

Fig.1 Data Mining with Big Data

III. KEY FEATURES OF BIG DATA

• 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.

IV. EVOLUTION TO BIG DATA ANALYTICS TECHNIQUES

The term 'Big Data' appeared for _rst time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [9]. Big Data mining was very relevant from the beginning, as the _rst book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [3] . However, the _rst academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8]. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [1] in his invited talk at the KDD BigMine'12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to find useful patterns to improve user experience.

Analyzing huge amounts of data allows analysts, researchers, and business users to make better and faster decisions using data that were previously not obvious before, inaccessible, or unusable. However, the dramatically increase of data amounts have made the well-known data mining algorithms unsuitable for such data sizes. Therefore, many studies have currently been directed towards the enhancements that can be introduced to data mining techniques in order to cope with big data, where big data analytics field has emerged. Big data analytic techniques are concerned with several data mining functions, where the most important functions are: association rules mining and classification tree analysis. In this section, we analyze the main data mining tasks that have been adopted to big data analytic techniques, clarifying the enhancements that have been introduced to achieve such adoption, in addition to the "V" dimension of big data that has been handled by such modifications. Table 1 represents our comprehensive summary of the analysis done for the evolution of data mining tasks to big data analytics. Techniques are grouped according to their data mining task. The table presents the status of each technique whether it has been developed to big data analytics and the dimension of big data that is handled by this developed technique. The following sub-sections describe the enhancements that have been introduced to the different data mining techniques to handle the dimensions of big data in order to evolve to big data analytic techniques.

Table 1: Evolution of Data Mining Technique to Big Data Analytics

S. No	Data Mining Task	Technique to be used	Developed to big data analytics	Dimensions covered
1	Classification	K- nearest neibhour	Y	Volume & Varacity
		Decision Tree	Y	Volume, Velocity & Variety
		Support Vector Machine	N	Volume, Velocity & Variety
		Naïve Bayes Classifier	N	Volume, Velocity & Variety
		Ripper	N	Volume, Velocity & Variety
		Neural Network	Y	Volume
2	Association Mining	Apriori	Y	Volume & Velocity
		FP Growth	Y	Velocity
3	Clustering	K-Means Clustering	Y	Volume
		K-Medoids	N	Volume
4	Optimization	Genetic Algorithm	N	-
		Sampling Techniques	N	-
5	Classifiers Ensembles	Bagging	N	-
		Random Forest	N	-
		Rotation Forest	N	-

V. BIG DATA MINING ALGORITHMS

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods [11], designing a data mining mechanism from a multisource perspective [11], [12], as well as the study of dynamic data mining methods and the analysis of stream data [7], [12]. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of gradual improvement of computer hardware functions, researchers continue to explore ways to improve the efficiency of knowledge discovery algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge discovery of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data stream or a

characteristic flow, a well-established mechanism is needed to discover knowledge and master the evolution of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multisource data provide essential differences between single-source knowledge discovery and multisource data mining. Wu et al. [11], [12], [10] 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. Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing.

Data streams are widely used in financial analysis, online trading, medical testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining [7]. Knowledge evolution is a common phenomenon in real world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by dynamics and context in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features.

VI. CHALLENGING ISSUES IN DATA MINING WITH BIG DATA.

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

• Mining platform.

• Privacy.

• Design of mining algorithms.

Basically, the Big Data is stored at different places and also the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored at different places is that much expensive. Suppose, if we use these typical data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of Big Data, and then it would become an obstacle for it. Because the typical methods are required data to be loaded in main memory, though we have super large main memory.

To maintain the privacy is one of the main aims of data mining algorithms. Presently, to mine information from Big data, parallel computing based algorithms such as MapReduce are used. In such algorithms, large data sets are divided into number of subsets and then, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to meet the goal of Big Data mining. In this whole procedure, the privacy statements obviously break as we divide the single Big Data into number of smaller datasets.

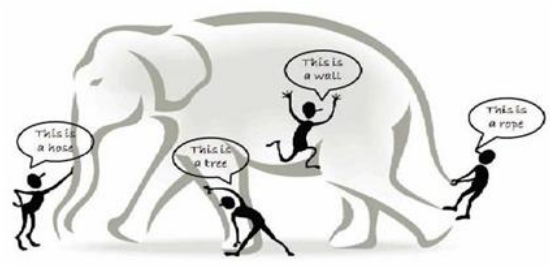

Fig. 2 Blind men and the giant elephant.

While designing such algorithms, we face various challenges. As shown in the figure 2 above, there are blind men observing the giant elephant. Everyone is trying to predict their conclusion on what the thing is actually. Somebody is saying that the thing is a hose; someone says it's a tree or pipe etc. Actually everyone is just observing some part of that giant elephant and not the whole, so the results of each blind person's prediction is something different than actually what it is.

Similarly, when we divide the Big Data in to number of subsets, and apply the mining algorithms on those subsets, the results of those mining algorithms will not always point us to the actual result as we want when we collect the results together.

VII. 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.The exponential growth in terms of capacity and complexity of data in last decade has led to substantial research in the field of big data technology. In this paper, we have made an attempt to summarize the recent literature review year wise in the area of Big Data & its analysis using different analytics approaches. Text analytics which is considered to be the next generation of Big Data, now much more commonly recognized as mainstream analysis to gain useful insight from millions of opinion shared on social media. The video, audio and image analytics technique has scaled with advances in machine vision, multi-lingual speech recognition and rules-based decision engines due to the intense interest existence of real time data of rich image and video content. They are the potential solutions to economical, political and social issues.

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.

[9] Zaki, Mohammed J., and Wagner Meira Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

[10] Washio, Takashi, and Hiroshi Motoda, "State of the art of graph-based data mining." ACM SIGKDD Explorations Newsletter 5.1 (2003): 59-68.

[11] Mohammed J. Zaki, Limsoon Wong, Data Mining Techniques, August 9, 2003 WSPC/Lecture Notes.

[12] Arinto Murdopo, "Distributed Decision Tree Learning for Mining Big Data Streams", July 2013.

[13] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

