A Survey on Big Data mining Applications and different Challenges

Dattatray Raghunath Kale, Satish R.Todmal

Abstract— **A Big Data is a new term applied to find out the datasets that due to their bulk size and involution. We cannot handle them with our current technique or data mining software tools. A Big data Composes large volume, difficult to analyze or understand and increasing in size data sets with multiple independent sources. Big Data mining is the ability of deducing helpful knowledge from these large datasets or streams of data, that due to its large volume, differentiability, and Fastness, it was not feasible before to do it. A big data are currently expanding into the every science and engineering fields. This survey paper shows all the characteristics and features of Big data, some applications of Big Data, challenging issues in Big data and its related work. In addition we focus on different articles studied and written by mostly talented scientist in the area of Big data mining. We hope our review will help to remodel the today's data mining technology for resolving the Big data challenges.**

Index Terms— **Data mining, Big Data, Independent sources, Data sets.**

I. INTRODUCTION

In the current years we all are giving a proof that there is a big increase in our ability to gather data from different resources in various formats, from autonomous or connected applications. Nowadays, the quantity of data that is created every two days is estimated to be 5 Exabyte's. This massive amount of data opens new challenging discovery tasks. This much more rising data has outpaced our capability to process, recover, analyze, store and understand these datasets. Data stream real time analytics are needed to handle the data recently generated, at an ever increasing rate, from different applications as: Email, blogging, Face book, Twitter sensor networks, log records etc. [1].

Manuscript received Nov, 2014.

Dattatray Raghunath Kale, *Department of computer engineering, imperial college of engineering & research, Pune, India, 9604386625.*

Satish R. Todmal*, Department of computer engineering, imperial college of engineering & research, Pune, India, 9960488094, 8805464743.*

Let us consider the Internet data. The web pages created by Google were around one million in 1998, but rapidly reached 1 billion in 2000 and have more increased 1 trillion in 2008.Each and every day whole world create 2.5 quintillion bytes of data. This is too much information and furthermore the 90% of the data in the whole world today has been created in the previous two years alone [1] This quickly growth is accelerated by the energetic increase in our well familiar social networking applications, such as Face book, Twitter, Weibo, etc., that gives permission to users to create different contents freely and modify the existing huge Web volume data. Then next, with different mobile phones becoming the easy way to acquire the real-time data, the very much amount of data that mobile can Implicitly processes to change our regular life has significantly outpaced our past call data record based processing for billing purposes. It can be seen that different Internet applications will increase the scale of data to the fresh level. People and different devices (from home to personal vehicles, buses, railway stations and airports) are all loosely connected. Trillions of such connected elements will generate a huge data, and important information must be discovered from the data to help to amend the quality of our life and make our world a better place. In all these above applications, we are facing the problem of system capabilities and how to solve the problems related to different business models. So we introduce Big Data mining.

II. BIG DATA MINING

The 'Big data' is originally found out due to the fact that we are generating a large volume of data each and every day. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [2]. The Big Data is nothing but a data, which is available at heterogeneous, independent sources, in very huge amount, which get updated in fractions of seconds. Big data is believed to play a captious role in the upcoming in years in all fields of our lives and our different organizations. For example, the data stored at the server of Facebook, as most of us, daily use the Facebook; we upload different types of information, upload photos, videos. All the information gets stored at the data warehouses at the server of Facebook. This data is called as the big data, which is so called due to its huge complications. Also another example is storage of photos and different information at flicart website. These are the very good real-time examples of the Big Data. Anew large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, and Twitter

ISSN: 2278 – 1323

are starting to look carefully to this data to find useful patterns to enhance user skills. The mismatch among the claims of the big data operations and the abilities that conventional Database management systems can provide has reached the historically high peak. We need new algorithms, and new tools to deal with all of this data. Doug Laney [3] was the first one in talking about 3 V's in Big Data management. The three Vs (volume, variety, and velocity) of big data each imply one distinct aspect of captious deficiencies of today's Database management systems. These threeVs (volume, variety and velocity) are three defining properties or dimensions of data. The "3V's" framework for understanding and dealing with "big data" has now become common.Currenltythere is two more V's are also present, one is variability and another is value. The variability shows changes in the structure of data. The value shows business value that gives organizations undeniable advantages.

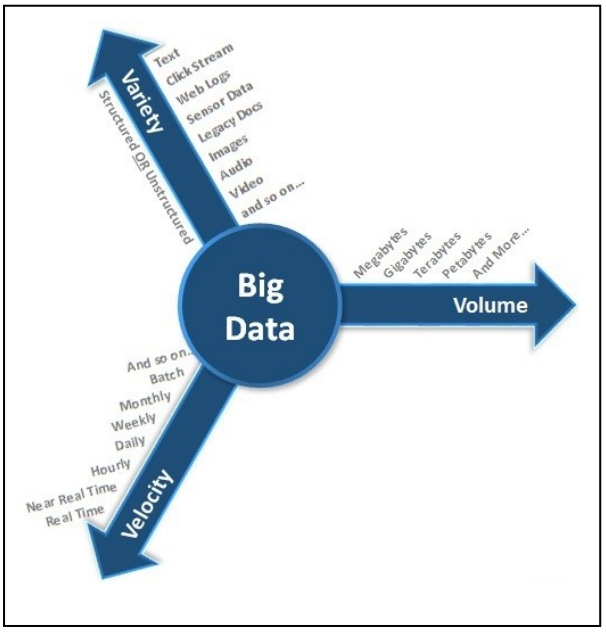

Fig.1: The three V's of Big Data

The 3Vs that define big data are Volume, Velocity and Variety.

a) Volume: Volume refers to the size of data that we are working with. With the advance of technology and with the invention of different social media, the amount of data is increasing very rapidly. This data is spread over the different places, in various formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and even more. Today, the data is not only generated by humans, but a large amount of data is being generated by machines and it surpasses human generated data. This size aspect of data is called as Volume in the Big Data world.
b) Variety: Variety refers to the various formats in which the data is being generated and then stored. Different applications generate and store the data in various formats. Now days, there are huge volumes of unstructured data being generated apart from the structured data getting generated in Business. Until the advancements in Big Data technologies, the industry didn't have any powerful and reliable tools/technologies which can work with such voluminous unstructured data that we see today. In today's world, organizations not only need to rely upon the structured data from enterprise databases, they are also forced to consume lots of data that is being generated both inside and outside of the enterprise like clickstream data, social media, etc. to stay competitive. Apart from the traditional flat files, spreadsheets, relational databases etc., we have a lot of unstructured data stored in the form of images, audio files, video files, web logs, sensor data, and many others. This aspect of varied data formats is called as Variety in the Big Data.
c) Velocity: Velocity can be taken as the speed at which the data is being generated. Various applications have different latency requirements and in currently competitive world, decision makers want the necessary information in the very small amount of time as possible. This speed aspect of data generation is referred to as Velocity in the Big Data world.

A. BIG *DATA APPLICATIONS IN VARIOUS FIELDS*

In this paper we want to show how big data is used today to add real value. Every aspect of our lives will be affected by big data. We have categorized the application of big data into different areas where we see the most widespread use as well as the highest benefits.

1. Optimizing big business Processes

Big data is also increasingly used to increase business processes. Traders are able to increase their stock based on predictions generated from social media information's, web search trends and weather forecasts. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here, geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic information.

2. Improving Healthcare and Public Health

The computing power of big data analytics allow us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. Just imagine of what happens when all the individual data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone! Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit. By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can now predict infections 24 hours before any physical symptoms appear. That way, the team can intervene early and save fragile babies in an environment where every hour counts.

3. Improving and Optimizing Cities and Countries

Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up.

4. Optimizing Machine and Device Performance

Big data analytics are also useful in optimizing Machine and Device Performance. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of

human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

B. BIG DATA CHALLENGES

1. Energetic Provisioning

The cloud computing provides service which is infrastructure as service in which it provides working out resources on requirement, many cloud related organizations are implementing this concept and to making it easy for clients to access these services. Recent frameworks do not have the property of the energetic provisioning. Here is issues that compute resources can be insufficient for the submitted job, some process may requires more resources. Another issue is scheduling and protection algorithm, current algorithms does not consider these aspects [4].

2. Mishandlings of Big in data mining

Challenges together with prospective misuse of big data are here, since information is power. Types of the data which people will produce in the future are unknown. To overcome these challenges we have to strengthen and increase our intent and capacity [5].

3. Security and Privacy of Big Data

Cloud protection alliance big data working group identify top protection and seclusion problems that need to confine for making the big data computing and infrastructure more secure. Most of these issues are linked to the big data storage and computation. There having some challenges which are related to secure data storage [6]. Different security challenges related to data security and privacy are discussed in [7] which include data breaches, data reliability, data accessibility and data support.

4. *Design of Big data mining algorithms*.

The Big Data is stored at various positions and also the data volumes may get enlarged as the data keeps on growing continuously. So, to gather all the data stored at various places is that much costly. Let us consider, if we use these classic data mining 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 difficulty for it.To protect the privacy is one of the main aims of data mining algorithms. Currently, to mine information from Big data, equivalent computing based algorithms such as MapReduce are used. In such algorithms, large data sets are separated 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 obtain the aim of Big Data mining. In this whole method, the privacy statements clearly break as we split the single Big Data into number of smaller datasets. While designing such algorithms, we face different challenges. Considering big data a group of complex and bulky data sets that are hard to process and mine for patterns and knowledge using traditional database management tools.

5. Constructing a universal unifying system related to big data mining.

There having a many methods which are planned for carrying out classification or grouping independently, but there is no theoretical background that unifies various responsibilities such as classification, grouping and association guidelines and so on. So building a universal unifying model for mining big data is a dynamic research field.

III. RELATED WORK

Big data having a huge, dissimilar, and dynamic features of application data involved in a circulated environment, so due to this Big Data is to perform computing on the petabyte (PB), even the Exabyte (EB)-level data with complex computing process. Therefore, utilizing an equivalent computing infrastructure, its conforming programming language support, and software models to proficiently analyze and mine the spread data are the critical goals for Big Data processing to change from "extent" to "excellence."

In the mining platform sector, at present, equivalent programming models like MapReduce are being used for the perseverance of study and mining of data, also there having a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented equivalent computing model. There is still a definite gap in performance with relational databases. Improving the performance of MapReduce and improving the real-time nature of large-scale data processing have accepted an important volume of attention, with MapReduce equivalent programming being applied to many machine learning and data mining algorithms. Data mining algorithms generally want to scan through the training data for gaining the statistics to explain or improve model.

The important work in Big data mining can be found in the major conferences like KDD, ICDM, ECMLPKDD, or other international journals. A Yizhou sun and Jiawei Han presents one paper in which they shows that mining heterogeneous information networks is a new and auspicious research leading edge in Big data mining. It considers interconnected, multy-typed data, with the usual relational database data, as heterogeneous data networks. A Jimmy Lin and Dmitriy Ryaboy present a scaling Big Data mining infrastructure. This paper presents insights about Big data mining infrastructure and the experience of doing analytics at Twitter. It shows that due to the recent status of the data mining tools, it is not simple to perform analytics.

A current paper on confidentiality protection in Big data [15] summarizes a various methods for giving the protection to the public release data, including aggregation, suppression, data exchange. In the different Big data applications, privacy concerns the main point on not including the third party from directly accessing the unique data. The ordinary solutions are to rely on some privacy-preserving approaches or encryption mechanisms to protect the information. A current effort by Lorch et al.[16] indicates that users " data access patterns "can also have strict data privacy issues and show the way to disclosures of geographically co-located users or users with general awareness.

A Big graph mining: Algorithms and discoveries by U kang and christors faloutstos gives an outline of mining big graphs, concentrating in the use of the PEGASUS tool, which shows certain results in the web graph and Twitter social network. The paper shows motivational future exploration ways for big graph mining.

To progress the fragile scalability of outdated study software and unfortunatescrutinyskills of Hadoopsystems,Das et al. [8] lead a study of the addition of R (open source statistical analysis software) and Hadoop. The in-depth assimilation shoves data computation to parallel processing, which allows dominant

ISSN: 2278 – 1323

profound analysis competences for Hadoop. Wegener et al. [9] accomplished the addition of Weka andMapReduce. Standard Weka tools can simply run on a particular machine, with constraint of 1-GB memory. After algorithm parallelization,Wekabreakdowns through the constraints and progresses performance by taking the improvement of parallel computing to handle more than 100-GB data on MapReduce clusters.Ghoting et al. [10] put forwarded Hadoop-ML, on which designers can simply build task-parallel or data-parallel machine learning and data mining algorithms on program blocks underneath the language runtime environment.

Some people, who expect to appoint a third party such as accountants to procedure their data, it is very essential to have effectual and operative contact to the data. In such cases, the privacy constraints of user may be faces like no limited copies or copying allowed, etc. So there is privacy-conserving free auditing mechanism planned for huge scale data storage.[11] This public key-based mechanism is used to allow third-party auditing, so users can securely permit a third party to evaluate their data without breaking the security settings or conceding the data confidentiality. In case of design of data mining algorithms, Knowledge progress is a public occurrence in actual world systems. But as the problematic statement varies, consequently the knowledge will change. Let us consider one example, when we go to the clinician for the treatment, that clinician's treatment program constantly changes with the situations of the patient. Likewise the knowledge. For this, Wu [12] [13] [14] suggested and started the theory of local pattern analysis, which has placed a groundwork for worldwide knowledge innovation in multisource data mining. This concept delivers a resolution not only for the problems of complete examine, but also for discovering worldwide prototypes that outmoded mining methods cannot find.

IV.CONCLUSION

In this paper we have studied the concept of Big data in data mining. Big Data is available to maintain highly increasing data during the subsequently years, and each data scientist will have to manage much more amount of data at every year. This data is going to be more miscellaneous, bigger, and quicker. We discussed some approaches about the subject, and what we think about are the major concerns and the major 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. We regard Big data as an emerging trend and the need for Big data mining is rising in all science and engineering domains. With Big data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

REFERENCES

[1] J. Gama. Knowledge discovery from data streams .Chapman & Hall/CRC, 2010.
[2] S. M. Weiss and N. Indurkhya. Predictive data mining practical guide. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.
[3] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.
[4] M. N. Vijayaraj, M. D. Rajalakshmi, and M. C. Sanoj, "Issues and challenges of scheduling and protection algorithms for proficient parallel data processing in cloud."
[5] U. G. Pulse, "Big Data for development: challenges & opportunities," NacionesUnidas, Nueva York, mayo, 2012.
[6] "Top ten big data security and privacy challenges," Cloud Security Alliance White paper, 2012.
[7] A. A. Soofi, M. I. Khan, R. Talib, and U. Sarwar, "Security Issues in SaaS Delivery Model of Cloud Computing," 2014.
[8] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J.McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998, 2010.
[9] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09),pp. 296-301, 2009.
[10] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc.Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.
[11] 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.
[12] 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.
[13] X. Wu, C. Zhang, and S. Zhang, "Database Classification for Multi-Database Mining," Information Systems,vol. 30, no. 1, pp. 71- 88, 2005
[14] 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.
[15] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.
[16] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13), 2013.

