Research issues in Big Data – A Review

G. Sandra

Assistant Professor, Sree Narayana Guru College of Advanced Studies, Cherthala, Alappuzha, Kerala, India. 688582.

Abstract.

Thanks to the development of modern information system and digital techniques, huge amount of data is generating at every moment, which is called Big Data. At present it is the major source of knowledge discovery. Different agencies, such as organizations, governments, academicians, politics, science, marketing companies, health sector, etc. are depending on it for their business prosperity and formulating policy decisions. Big data is both the source of benefits as well as many risks. For the realization of its benefits and solving the risks, big data becomes the center of research and development of both industries and academia. The basic objective of this paper to explore the major research issues in Big Data.

Key words: Big Data, Datasets, Data mining

I. INTRODUCTION

Ours is a digital community. The modern information systems generate terabytes of data at every moment [1]. This massive data is popularly known as Big Data, which is very complex and heterogeneous in nature [2]. It is the treasure house of knowledge, widely used by many organizations, such as the business firms, corporate, governments, banks, healthcare institutions, etc., for their business prosperity and planning [3]. To take the advantages from the massive data, they have to be collected, stored and mined in an appropriate way. Researchers and organizations are using many techniques, both traditional and modern, to fulfill the above task. The traditional techniques are framed to solve the problems of small and medium datasets. But they are ineffective to process and analyze the huge volume of datasets, which are growing at an exponential rate [4]. To solve this issue, researchers have developed many applications, but none of them have a general acceptability. So this paper tries to review the major research issues in Big data that are confronted by the scholars.

This study is organized in the following way. Section II gives the format of Big Data. The section III narrates the research problems in Big Data. Finally in section IV, we conclude the discussion with proposal for future work.

II. BIG DATA FORMAT

Big data consists of many formats, which can be classified broadly as structured data, semi-structured data and unstructured data. Chrisrine Taylor [5] defines the structured as the one whose characteristics make them easy to identify. It is pre-defined one and has rigid structure. They usually lie in the rational databases. To process and analyze this type of data, the data scientists are not facing much difficulty and different applications have effectively employed to analyze them. But in the case of unstructured data, it has no definite pre-defined model. It may be stored anywhere without any rational databases, but difficult to search. It can be of any format like images, texts, videos, etc., which are generated with the help of men and machines. Since it is not in rigid nature, processing and analyzing the unstructured datasets are very difficult. Hence new scientific processing techniques are the need of the time. The semi-structured data is self explanatory in nature and not based on any rigid structure. To identify it, they are governed by definite rules and maintain markers and tags. E-mail is the best example of semi-structured data. Like unstructured data, the semi-structured data is also not easy to process and analyze. Hence developing advanced techniques for data processing is a major issue that the scholars are facing [6]. Though these formats of big data are the source of knowledge discovery for many organizations and policy makers, but, they are also the abode of research challenges of big data and a review of them is given in the next section.

III. RESEARCH PROBLEMS IN BIG DATA.

The Big data challenges are the current focal points of research. To solve the big data challenges, researchers have made consistent efforts from time to time. This efforts end in the formulation of many approaches to solve the issues in big data. Because of the heterogeneous data and complexity in it, scholars are facing crucial research problems in big data analysis, especially in the domains of applications, algorithmic approach, speed, efficiency, complex data pattern and feature extraction and analysis of unstructured data format [7]. The data scientists have suggested a list of characteristics of big data as the sources of challenges in big data [8]. The massive data

collected from different sources have distinct style, which are initially called 3Vs and many other Vs have been added to it subsequently. The most popular among them are high volume, high velocity, high variety, low veracity, high value and high variability [9]. The high volume of data generated from different sources is a great challenge to big data processing, because to manipulate the voluminous data needs a lot of resources. Analyzing the huge volume of dataset is a tedious job and in most occasions the various technologies used are remained ineffective [10]. The speed at which the massive data generated is called velocity which is very high in big data. The main issue the data scientists confronting in this context is that they have to process the data and respond to the quarries at a speed equal to the data generated. This challenge cannot be solved easily because big data is multiplying exponentially and the technology used for data processing is not modifying at an equal tempo.

The different variety of datasets in big data, i.e., structured, semi-structured and unstructured, which are not integrated, is a great issue to data scientists. Most of the data mining tools are ineffective to process and analyze such a non-integrated data. So long as the big data structure remains not integrated, the degree to which they can be trusted is very low. It is referred as low veracity. The big data is the treasure house of knowledge and proper analysis of it confers very valuable information to the corporate and policy framers for their prosperity. Hence big data has high value to its stakeholders. Big data can be formatted and used in different ways, which is referred as variability of big data. The variability characteristics of big data aggravate the data mining challenges and make it too much complicated. All the above characteristics of big data make its analysis great challenging , which can be categorized as challenges to data infrastructure, computational complexities, privacy and security, scalability, data validation and management complexities [11]. To know how these issues affecting the processing and analyzing big data, a brief discussion about them is given below.

Sufficient and appropriate infrastructural facilities are required for storing, processing and analyzing big data. The large volume of data generating from different sources, such as social networks, internet, mobile applications, business transactions, etc, needed sufficient and reliable hardware and software systems but such systems can be dependable only for a time period. After the duration, because of its continuous and intensive use, it will result in malfunction. So there is the issue of maintaining the efficiency of the system always.

The data selection, feature extraction and data reduction are the major challenges to the data scientists. The heterogeneity and dimensional issues of big data have aggravated this challenge. The algorithms that are used to solve this problem have failed in their objectives. Different machine learning and AI techniques have been applied to mitigate these challenges [12]. But they have many limitations.

Selection and use of right techniques for big data analysis is a great challenge, which is called process challenge. It comprises of collection of data from different sources, filtering and featuring them, so as to make it suitable for analysis of data, developing suitable algorithms and analyzing the output for knowledge discovery [13]. Scholars have proposed many techniques for information retrieval and knowledge discovery and a few prominent among them are: (i) Rough set Theory [14], which is a new approach developed for data analysis and feature selection. (ii) Near set Theory [15]. This is also for feature selection based on proximity. (iii) Soft set Theory [16]. This is a mathematical tool to deal with uncertainty and vagueness of datasets. Principal Component Analysis is a multivariate technique to analyze the inter-correlated dependent variable. It is generally used in data processing and dimensionality reduction [17]. Formal Concept Analysis is employed for information retrieval and knowledge discovery [18]. But when data generated from different sources show an exponential growth, the data analysis with the help of existing tools is not efficient in handling the problems of uncertainty and data complexities.

To preserve information security is a big challenge to big data. The data made available to organizations comes from different sources, which to a greater extent cannot be trusted. The organizations are unaware of whether the data is compromised one or not. Any act of compromising security will give a chance to hackers to misuse it for their motives. This shows that the task of protection of individual's privacy is a delicate issue. Most of the organizations have their own security measures to protect the user's profiles and used to store the data information in cloud systems [9] because of its advantages of greater storage space and cost efficiency. But still to offer data security is remained as unsolved problem to all organizations. As a better solution to this, data scientists suggested the 'in house' big data system. In this technique data stored is encrypted with a personal key, hence others cannot access it without the permission of the user. For each time of the use of data, it must be decrypted and after the use it must be again encrypted. The other major applications suggested for the security issues of big data are the techniques of authentication and authorization, where the server

determines whether the client has the permission to use or access the data. In spite of the various security measures, the security problem of big data is a live issue and to solve it, multilevel security algorithms have to be developed.

Along with security problems, the big data faces the challenges from its scalability. Due to high velocity character of big data, the size of data is scaling at a higher speed than the speed of CPU, which demands modification in the processor technology. The efforts on the part of scientists in this way led the development of parallel computing [19].

Recently, the predictive algorithms for data streams received greater importance. Organizations are using many tools like big data analytic software, Hadoop, spark, NoSQL databases, horizontal scaling platforms like Peer to Peer networks, Apache Hadoop, Vertical scaling platforms like High Performance Computing Clusters, Multicore Processors, Graphic Processing Unit, Field Programmable Gate Arrays, Machine Learning Techniques, AI techniques etc., for big data processing and analysis. Though by employing the above mentioned applications the knowledge extraction is made possible, but to maintain security and privacy of user's profiles, data confidentiality protection, solving data multidimensionality problem, evaluation of data stream mining, data staging, data classification, clustering, feature extraction, development of approximation algorithms and computation complexities, etc., are the crucial open research issues remain in big data analysis

The above discussed various big data research issues are only one side of the big data challenges, i.e.; technological side and the other side is its management challenges. For the proper management of the application of techniques of data analysis, the organizations face many resistances from within. It needs sufficiently skilled professionals to solve various issues of techniques, privacy, security and governances of data processing and analysis in order to remain competitive in the industry [20]. Scholars have identified several management challenges related to big data. They grouped it into issues of privacy, security, knowledge dissemination, data management, data ownership and issues of operational efficiency [21]. How to preserve the privacy of user profiles is still an unsolved research issue. Organizations are making huge investment in managing privacy issues, since the failure to protect the user profile is made illegal [22]. Like the privacy issue, ensuring security is also a major challenge to organizations. Any failure in this will make the system vulnerable to attacks and threats to data security.

How the collected information is to be shared between the users is a crucial part of data management. The challenge here is the sharing of information without curtailing the individual privacy. This issue is related to data governance. Here the IT managers have to take care of data storing, mining, cleansing, categorizing, modeling, mapping and analyzing data. All these are activities incurring high cost, which is led to the management issue of cost efficiency. Besides all the above management challenges, the problem associated with the ownership of data is also a research issue, especially in social media like Face book, twitter, etc. Here the question is who owns the data. The users of these media while updating their status or tweets have an account also. So the question is who is the owner of the data . It is perceived that both the users and the service providers are the owners of the data. But this dichotomy is still going on and needs to be settled.

IV. CONCLUSION

It is true that emergence of big data has widen the horizon of knowledge discovery. It is the treasure house of benefits as well as challenges. We have reviewed the research issues in big data. They are assessed from the angle of data challenges, technical challenges and management challenges. The peculiar properties of big data create new challenges as and when the existing one gets solved. Hence researchers are facing new issues every time. To address it, multidisciplinary research approaches are needed. Hence we must encourage fundamental researches for finding out solutions to the above issues.

REFERENCES

[1] C Dobra and F Xhafa , "Intelligent services for big data science". Future Generation Computer Systems, 37, 2014, PP.267-281.

[2] 10 Challenges to Big Data security and privacy, Dataconomy.com, 2017,07.

[3] X Y , F Liu , J Liu and H Jin , "Building a network highway to big data: Architecture and Challenges", IEEE Network, 28{4}, 2014,PP.5-13.

[4] "Big data Analytics: Security and Privacy Challenges", 2016, IEEE Symposium on Computers and Communications.

[5] Christine Taylor, "Big-Data, Structured Vs Unstructured", Datamation, 2018.

[6] C.W Tsai, C.F Lai, H.C Chao, A.V Vasilakos, "Big Data Analytics: A Survey; Journal Big Data, 2015, 2(1): 21.

[7] Y Ma , H Wu , L Wang , B Huang , R Ranjan, , A Zomaya and W Jie ; "Remote sensing big data computing", Future Generation Computing Systems, Oct. 2015, Vol. 51, PP. 45-60.

[8] T Huang ,L Lan, X Fang , P An , J Min and F Wang , "Promises and Challenges of Big data computing in health science", Big Data Research, 2(1), 2015, PP.2-11.

[9] M.A Khan, M.F Uddin and N Gupta , " Seven Vs of Big data understanding", proceedings of 2014 Zone1 conferences of American Society for Engineering Education[ASEE Zone1]-IEEE, 2014, PP.1-5.

[10] Alexandru Adrian TOLE, "Big data Challenges", Database Systems Journal, Vol .IV. No.8.
[11] Kaislers,F Armour , J. A Espinosa, W. Money , "Big data: Issues and Challenges moving forward", 46th Hawaii International Conference on System sciences (HICSS), 2013, PP. 995-1004.
[12] C. L Philip, Q Chen and C. Y Zhang, "Data-intensive applications, challenges and technologies: A survey on big data", Information Sciences, 275, 2014, PP.314-347.
[13] V. Robert Zicari, "Big Data: Challenges and opportunities", Big Data Computing Pages, 2012, PP.104-128.
[14] Y. Qian , J Liang ,W Pedrycz and C Dang , "Artificial Intelligence:, Elsevier, 2010.
[15] J. F Peters ; "Near sets- An Introduction", Mathematics in Computer Science, March, 2013, Vol.7, Issue 1, PP.3-9.
[16] N Cagman S Enginoglu ; "Soft-set Theory and unit-int decision making", Elsevier, 2010, 05,004.
[17] HZou , Hastie Tand, R Tibshrani ; "Sparse Principal Component Analysis", Journal of Computational and Graphical Statistics, Vol. 15, 2006, Issues 2.
[18] Jonas Poelmans, Paul Elzinga, Stijn Viaene and Guido Didene; "Formal concept Analysis in knowledge discovery: A survey", ICCS, 2010, Springer Link, PP.139-153.
[19] A Jacobs ; "The Pathologies of Big data", communications of the ACM, 52(8), 2009, PP.36-44.
[20] I. A.I Hashem, I Yaqoob , N .B Anuar, A Gani , S.U. Khan; " The rise of Big data on clod Computing: Review and Open research Issues", Information System, 47, 2015, PP. 98-115.
[21] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani and Vishanth Weerakkody, " Critical analysis of Big Data challenges and analytical methods", Science Direct, Journal of Business Research, January 2017, PP. 263-286.
[22] A Machanavajjhala, and J. P Reiter, "Big privacy: protecting cinfidentiality in big data", XRDS: Crossroads, The ACM Magazine for students, 19 (1) 2012, PP.20-33.