RELEVANCE OF BIG DATA RESEARCH AND RELATED CHALLENGES: A SURVEY

Ms. Anushree Negi
Computer Science Engineering, University Institute of Engineering
Chandigarh University, Gharaun
anushree.e9392@cumail.in

ABSTRACT: Throughout the age of technology, huge amounts of data are accessible to decision-makers on hand. Besides, decision-makers need to gain relevant knowledge from these diverse and constantly changing data, from everyday transactions to user interactions and networking site data. It can be delivered using Big Data Analytics, which is the implementation of data analysis methods to Big Data. Big Data corresponds to data which are not only huge, as well as broad in variety and size, making them difficult to process using typical methodologies. The application of these data involves a great deal of work for successful decision-making at several levels of information extraction. We discuss the meaning of big data in this paper including its characteristics, and importance. Then we recognize the meaning and possibilities Big data brings to us from diverse perspectives. First, we're introducing descriptive big data projects from all around the world. We identify the significant challenges in big data and its analysis, as well as possible alternatives to such difficulties and also bring out overview of few tools for Data processing. Lastly, we summarize the paper by putting forth some proposals on the execution of big data initiatives.

Keywords: bigdata, datamining, analytics, decision making, Hadoop.

I. INTRODUCTION

Well into the digital world, information is developed from numerous technological resources that have contributed to big data growth. The development of massive datasets offers evolutionary breakthroughs in many fields. The approach of the selection of large and complicated resources is difficult to manage using typical data management methods or programs for information processing. Such are accessible in structured, semi-structured, and unstructured format in petabytes & beyond. Big data analysis's main purpose is to manage large volume, velocity, variety, and veracity information utilizing numerous conventional and analytical intelligent techniques[1]. Figure 1 below corresponds to the concept of big data.

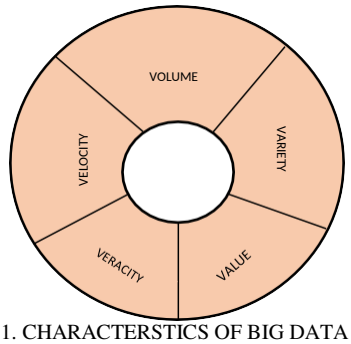

FIG 1. CHARACTERSTICS OF BIG DATA

The features of big data can be described by 5V compared to standard data, notably huge volume, high speed, high variety, low veracity, and high value. The biggest issue in working with big data is not simply its tremendous scale, although we can often minimize this challenge by enhancing or increasing our processing systems fairly. In particular, the real issues are also the differentiated information types (variety), speedy reply requirements (speed), and data inconsistencies (veracity). Due to the various data types, an application also needs to handle not only typical structured data but also semi-structured or unstructured data (including text,

images, video, and voice). Timely replies are also difficult even though in a reasonable time frame, there will not be enough resources to obtain, process and analyze the big data. Ultimately, it is extremely difficult to discriminate between true and false or reliable and inaccurate data, including with the best techniques of data cleaning to remove any underlying information unpredictability.

Imagine an environment lacking data storage; a world in which every information regarding an individual or an entity, every process conducted, or every feature which can be registered is lost instantly upon use. In this way, companies may risk the opportunity to gather relevant information and expertise, conduct comprehensive analyzes and develop new possibilities and benefits. Everything that ranges between customer names and addresses to items present, to transactions made, to hired workers, etc. are becoming necessary for consistency now and then today. Data is the foundation on which every company flourishes.

So think about the magnitude of the knowledge and the explosion of data and information generated by technology and internet developments nowadays. With the increase in storage capacities and data collection techniques, large quantities of data have easily become usable. More and more data is being generated every second and needs to be processed and analyzed to extract value. Additionally, data has become easier to store, and companies need to extract as much value from the large volumes of stored data as possible.

IDC projected that the third IT platform's market size will hit US$ 5.3 trillion by the end of 2020; and the third IT platform will drive 90 percent of the growth in the IT industry from 2013 to 2020. Big data is the central connotation and vital support of the so-called second economy from a socio-economic point of view, a term promoted by the American economist W.B. Arthur in 2011[2], which applies to the processor, connectors, sensors, and executors working on economic activities. It is estimated that by 2030, the scale of the second economy exceeds that of the first economy (the conventional physical economy, in particular). The second economy's key help is big data because it's an inexhaustible tool that continually enriches. In the future, the skill of the second economy would not be that of labor efficiency, but of information productivity, thanks to big data.

II. IMPORTANCE OF BIG DATA

Big data has fundamentally changed and transformed the way we live, function and think, because of its tremendous value[3]. In what follows, we explain in depth the value of big data from various views.

2.1. Importance for national growth:

The planet is now truly entering the era of the age of information. The widespread use of the Internet, the Internet of Things, cloud computing, and other new IT technologies has made numerous data sources expand at an exponential pace while increasing the complexity of data structures and forms. Depth research and the use of big data can play an important role in promoting countries' sustainable economic growth and improving market competitiveness. Big data would remain a key component of economic development in the future.Big data would allow companies to upgrade and turn to the Analysis as a Service (AaaS) model, thus changing the ecology of the IT and other industries. In this sense, the global IT business giants (such as IBM, Google, Microsoft, and Oracle) have already begun their preparation for technology growth in the big data age. At the national level, the ability to collect, process and use large

quantities of data will become a new symbol of the strength of a country. Apart from land, sea, air, and outer spaces, a country's data sovereignty in cyberspace would be another great power game space.

2.2 Importance to industrial upgrades:

Big data is a common problem faced by many businesses, and it presents significant challenges to the digitization and information production of these industries. Research on popular big data issues, particularly on core technology breakthroughs, will allow industries to leverage the uncertainty induced by data interconnection and to master uncertainties caused by redundancy and/or data shortage. Everybody hopes to mine information, expertise and even insights from demand-driven big data and eventually take full advantage of big data's great value. This means that data is no longer an industrial-sector by-product but has become a central nexus of all aspects. And such sense, the focus of the new generation of IT and its implementations will be the analysis of common problems and core Big data technologies. Not only will it be the next catalyst to support the technology industry's fast growth but it will also be the latest resource for businesses to boost their competitiveness.

2.3. Importance of scientific science :

Big data has forced the scientific world to re-examine its scientific science methodology[4] and sparked a revolution in scientific thought and methods. It is well known that experiments were used to establish the earliest scientific work in human history. Theoretical science emerged later and was characterized by the study of various rules and theorems. Nonetheless, since theoretical research is too complex and not feasible to solve practical problems, people began to search for approaches based on simulation, which led to computational science. The advent of big data has given rise to a new research paradigm; that is, with big data, researchers may only need to find or mine the information, knowledge and intelligence needed from it.

They don't even need to have direct access to the items being examined. In 2007, Jim Gray, the late Turing Award recipient, portrayed the fourth model of data-intensive scientific research in his last speech[4], which distinguishes data-intensive science from computational science. Gray claimed the fourth model could be the only structural way of addressing some of today's toughest global challenges. The fourth paradigm is not only a shift in the way scientific work is done but also a change in the way that people think [3].

2.4 Importance to allow users accurately forecast the future:

By successful incorporation and accurate analysis on heterogeneous multi-source big data, better predictions of future event patterns can be achieved. Big data research can also facilitate sustainable social and economic transformations and further give birth to new data processing companies. Big-network data technology has been highly developed and implemented successfully in the security and military fields. Big data-based predictive modeling was used to solve social challenges like public health and economic growth. Ginsberg, et al. found that if the amount of queries submitted to Google and with keywords such as "flu symptom" and "flu care" increases in a region, then after a few weeks, the number of influenza patients in hospital emergency rooms in the area concerned will increase accordingly[5]. We will be able to forecast influenza outbreaks with this knowledge and apply countermeasures in advance. The United Nations recently unveiled a new initiative on economic growth, called Global Pulse[6], which aims to use big data evidence to support sustainable economic growth.

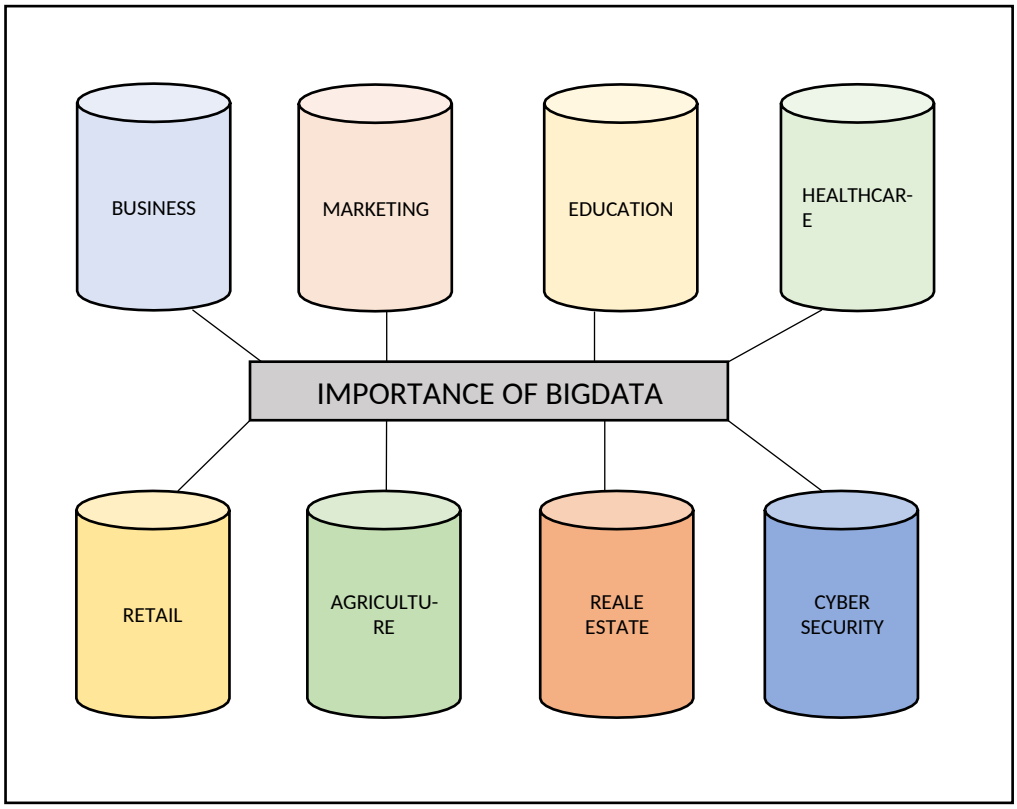

FIG 2. IMPORTANCE OF DATA IN VARIOUS FIELD

III. CHALLENGES IN BIG DATA AND ITS ANALYSIS

There are several obstacles to leveraging the potential of big data today, ranging from developing processing systems at the lower layer to evaluating methods at the higher layer, as well as several open science research issues. We'll briefly describe the major issues and challenges in this segment.

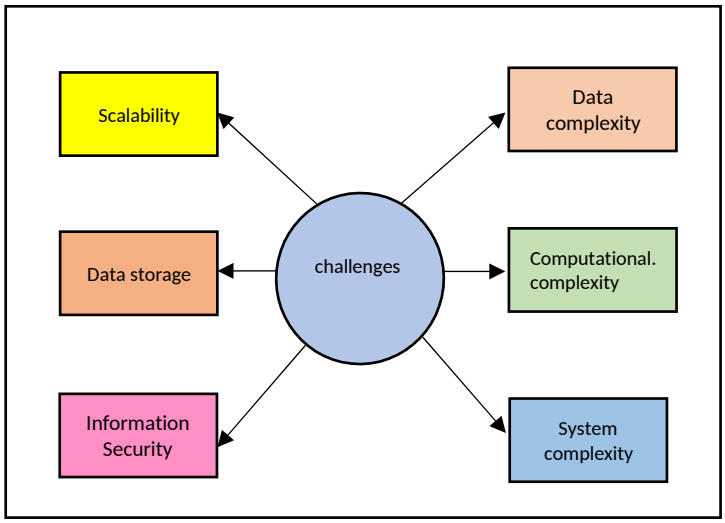

FIG 3. CHALLENGES IN BIG DATA

3.1. The complexity of data :

The advent of big data has provided us with unparalleled large-scale datasets when dealing with computational issues, but we now face even more complex data objects. Big data's inherent complexity (including complex forms, complex structures, and complex patterns) makes their interpretation, representation, comprehension, and processing much more difficult and contributes to sharp increases in computational complexity relative to conventional processing models based on total data. Standard data processing and mining activities, such as collection, exploration of subjects, semantic processing and analysis of emotions become incredibly difficult when using big data. We don't have a clear understanding of the the moment about tackling the nature of big data. For example, we lack information on the laws of big data distribution and association relationships. We lack an in-depth understanding of the inherent relationship- ship between data complexity and big data computational complexity, as well as domain-oriented methods of big data processing. All of these significantly hinder our ability to develop highly efficient computational models and methods for solving Big Data problems.

3.2 Computational complexity:

Three of the main features of big data, namely multi-sources, huge volume, and rapid change, make it difficult for conventional computing methods (such as machine learning, extraction of information, and data mining) to effectively support big data collection, analysis, and computation. Such estimates cannot simply rely on past data, analytical methods, and iterative something- rhythms used in conventional approaches to managing small amounts of knowledge. New methods would need to break away from conventional statistical assumptions focused on independent and equivalent data distribution and sufficient sampling to produce accurate statistics. New methods to big data technology would need to tackle broad content-oriented, innovative, extremely efficient computational concepts, provide ground breaking methods for processing and analyzing big data, and help value-driven applications in different domains. New features of large data processing, such as inadequate samples, transparent and ambiguous data relationships, and unbalanced value volume distribution, not only provide great opportunities, but also pose great challenges, researching big data computability and building new computational concepts.

3.3.System complexity:

Big data processing systems which are ideal for managing several data types and applications are the key to supporting big data science research. Its processing is faced with high computational complexity, long service cycle, and real-time requirements for data of huge volume, complex structure, and sparse meaning.

These specifications not only present new challenges for the design of system architectures, computational structures, and processing systems but also place strict limits on their operating performance including energy usage. The main problem to be tackled in system complexity is the design of device architectures, computational structures, pro-ceasing modes, and standards for highly energy-efficient big data processing platforms. The concepts for designing, implementing, evaluating and optimizing big data processing systems can be laid down to address these problems. With power-optimized and effective distributed storage and

processing, their solutions can form a significant foundation for evolving hardware and software system architectures.

3.4 Data storage :

The size of data has increased exponentially in recent years through various means such as mobile devices, sensor technology, remote sensing, radio frequency readers, etc. These data are processed at the expense of investing a great deal when they are eventually overlooked or removed because there is not enough room to store them. The first challenge for large-data analysis is therefore storage media and higher input/output speed. In these instances, the highest priority for the exploration and representation of information must be the data accessibility. The prime explanation is that for further study, it must be readily and promptly obtained. However, storage technologies available cannot possess the required performance.

3.5 Optimization and Data Visualization:

Its scalability and protection are the most critical problems for big data analysis techniques. Researchers have paid attention in recent decades to improve data processing and speed up processors led by the law of Moore. Sampling, on-line, and multi-resolution research techniques should be developed for the former. Incremental techniques have a strong property of scalability in the big data analysis field. As the data size is rising much faster than CPU speeds, the processor technology is embedded in a natural dramatic change with an increasing number of cores[7]. A change in processors allows parallel computation to evolve. Parallel computation is needed for real-time technologies such as mapping, social media, business, web search, timeliness, etc. We may note that big data has provided many challenges for hardware and software technologies that contribute to parallel computing, cloud computing, centralized computation, process simulation, scalable. We have to combine more statistical models with computer science to overcome this problem.

3.6 Cybersecurity:

Large volumes of data are combined, evaluated, and manipulated for useful trends in big data analysis. All companies have various policies to ensure that confidential information is secured. Preservation of confidential information is an important a concern in the study of big data. Big data is correlated with an unprecedented security risk[8].

Consequently, network security becomes a big data analytics problem. Big data protection can be improved by using encryption, authorization, and authentication methods. Special security challenges faced by big data systems are system size, range of multiple devices, real-time safety control and risk of attack system[9],[10]. The security issue posed by big data also introduced cyber protection to the focus. Furthermore, importance must be paid to creating a multilevel system of information security and a method of protection. While a great deal of work has been done to protect big data[9], it does need a lot of improvement. The major obstacle for big data was the creation of multi-level protection, security-preserved information system.

IV. TOOLS USED FOR BIG DATA PROCESSING

TABLE1 : VARIOUS BIG DATA TOOLS

TOOLS	DESCRIPTION	LANGUAGE	PURPOSE	ENVIRONMENT	OPEN-SOURCE	REFERENCE
HADOOP	It is a framework for storing large set of data by distributing it on computer clusters.	Java, Scala	Clustering, classification	Backend independent	yes	[11]
CLOUDERA	CDH builds data hub for business and help organization with better access to data	Apache framework	data engineering, data warehousing, machine learning and analytics	Cloud	yes	[12]
DATA CLEANER	It cleans the semi-structured data and transform into clean readable data that can be utilized further for analysis.	Java	ad-hoc analysis, recurring cleansing ,Data Management solutions.	Cloud	yes	[13]
RAPID MINER	It is an open source data science platform to streamline predictive analytics process.	Java	Business processes, predictive analysis, text mining	Visual workflow, Built-in templates	Partially	[14]
ORACLE DATA MINING	It helps in discovering the future insights, making predictions and also provide access to Oracle Data.	C Fortan	Classification Prediction Feature extraction Specialized analytics	Backend dependent	No	[15]

Throughout this analysis the big data methods are introduced to variables and criteria depending on both the functioning of each method and its variations are acknowledged. It offers user comfort for people employed in large corporations and business-related clients to learn every tool's functionality in a simplified and improved manner[16]. Table 1 describes the various tools used in big data along with its various features.

V. CONCLUSION

Information is produced at such a remarkable speed in past years, and examination of such information is difficult for a specific person. Towards this aim, we review the numerous research questions, difficulties and methods that had to evaluate these analytics. Big data but also its features and meaning were thus addressed. Out of this study, it is assumed that many of them are built for batch processing with each big data application having their unique emphasis while others are good at real-time analytics.

Every big data framework already has features unique to it. In this environment of information overload, we assume that big data analytics is of major importance and could provide unintended observations and advantages to decision-makers in different fields. Big data analytics seems to have the ability, if successfully implemented and enforced, to get a foundation for developments at the science, technical and social levels. Future studies should concentrate on creating a path or structure for big data planning which can address the aforementioned challenges.

REFRENCES

1. M. K. Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation inEngineering &Management, 2(8) (2015), pp.228-232.
2. W.B. Arthur, The second economy, available at: http://www.images-et-reseaux. com/sites/default/files/medias/blog/2011/12/the-2nd-economy.pdf, 2011.
3. V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.
4. T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, 2009.
5. J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 7232 (2009) 1012–1014.
6. [20] Big data for development: challenges & opportunities, available at: http:// www.unglobalpulse.org/projects/BigDataforDevelopment, May 2012.
7. A. Jacobs, The pathologies of big data, Communications of the ACM, 52(8) (2009), pp.36-44.
8. H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, International Conference on Information Technology Management Innovation, 2015, pp.1041-1044.
9. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, Congress da sociedada Brasileira de Computacao, 2014, pp.1-6.
10. I. Merelli, H. Perez-sanchez, S. Gesing and D. Agostino, Managing, analyzing, and integrating big data in medical bioinformatics: open problems and future perspectives, BioMed Research International, 2014, (2014), pp.1-13.

11. Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A review paper on big data and hadoop." International Journal of Scientific and Research Publications 4.10 (2014): 1-7.
12. Pol, Urmila R. "Big data and hadoop technology solutions with cloudera manager." International Journal 4.11 (2014).
13. Kumar, Ajay, et al. "A big data MapReduce framework for fault diagnosis in cloud-based manufacturing." International Journal of Production Research 54.23 (2016): 7060-7073.
14. Rangra, Kalpana, and K. L. Bansal. "Comparative study of data mining tools." International journal of advanced research in computer science and software engineering 4.6 (2014).
15. Berger, Charlie. "Oracle Advanced Analytics: Oracle R Enterprise & Oracle Data Mining." Product Presentation(2012): 1-58.
16. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, on the use of mapreduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.