**A Big Data Analytics Survey: Challenges, Unresolved Research Issues, and Tools**

Dr. Rajul Jain,drrajuljainmgmm@gmail.com

Department of Computer Science, Mata Gujri Mahila Mahavidyalaya, Jabalpur

Harsharan Kaura

Department of Computer Science, Mata Gujri Mahila Mahavidyalaya, Jabalpur

Abstract

Every day, current information systems and digital technologies such as the Internet of Things and cloud computing generate terabytes of data. To extract knowledge for decision making, huge amounts of data must be analyzed at different levels. As a result, big data analysis is a contemporary research and development topic. The primary goal of this study is to investigate the possible influence of big data concerns, open research issues, and associated tools. As a result, this article provides a framework for investigating large data at various phases. Furthermore, it provides up a new avenue for academics to develop solutions based on the challenges and open research issues.

Keyword: Big data analytics, Hadoop, Massive data, structured data, unstructured data.

I. INTRODUCTION

Data are generated from numerous sources in the digital age, and the rapid transition from digital technologies has resulted in the creation of big data. With the acquisition of massive datasets, it provides evolutionary breakthroughs in many domains. It is a collection of huge and complex datasets that are difficult to analyse using typical database administration tools or data processing programmes. These are available in petabytes and beyond in structured, semi-structured, and unstructured formats. Formally, it ranges from 3Vs to 4Vs. 3Vs is an abbreviation for volume, velocity, and variety. Volume refers to the massive amount of data generated every day, whereas velocity relates to the rate of increase and how quickly data is acquired for analysis. Variety offers information on The fourth V stands for veracity, which involves accessibility and responsibility. The primary goal of big data analysis is to handle large amounts of data with high velocity, variety, and veracity utilizing a variety of classical and computational intelligent techniques [1]. Gandomi and Haider [2] described several of these extraction strategies for acquiring useful information. The definition of big data is depicted in Figure 1. However, the precise meaning of big data is unknown, and it is thought to be problem-specific. This will assist us in improving decision making, insight finding, and optimization while remaining inventive and cost-effective.

It is anticipated that the expansion of big data would reach 25 billion by 2015 [3]. Big data, from the standpoint of information and communication technology, is a ro-bust stimulus to the next generation of information technology,industries [4], which are broadly based on the third platform and include big data, cloud computing, the internet of things, and social business. Data warehouses are commonly used to manage huge datasets. The extraction of exact knowledge from available huge data is a critical issue in this scenario. Most data mining approaches given are incapable of successfully handling huge datasets. The main issue in big data analysis is a lack of coordination across database systems as well as analysis techniques such as data mining and statistical analysis. These difficulties typically arise when we wish to undertake knowledge discovery and representation for practical applications. A basic issue is determining how to quantify the .Furthermore, research on big data complexity theory will aid in understanding the key properties and development of complex patterns in big data, simplify its representation, improve knowledge abstraction, and lead the design of big data computing models and algorithms [4]. Various academics have conducted extensive research on big data and related tendencies [6, 7, 8].

It should be highlighted, however, that not all data available in the form of big data is appropriate for analysis or decision making. Industry and academics are both interested in spreading big data results. This study focuses on big data difficulties and available methods. In addition, we discuss open research issues in big data. So, in order to elaborate, the paper is separated into the following sections. Section 2 addresses the issues that arise during the fine tuning of huge data. Section 3 presents open research questions that will aid us in processing huge data and extracting relevant insights from it. Section 4 delves into big data tools and methodologies. Section 5 concludes with observations that summarise the results.

II. DIFFICULTIES IN BIG DATA ANALYTICS

Big data has accumulated in various fields in recent years, including health care, public administration, retail, biochemistry, and other interdisciplinary scientific projects. Big data is regularly encountered by web-based applications, such as social computing, online text and documents, and internet search indexing. Social networking analysis, online communities, recommender systems, reputation systems, and prediction markets are examples of social computing, whereas internet search indexing includes ISI, IEEE Xplorer, Scopus, and Thomson.

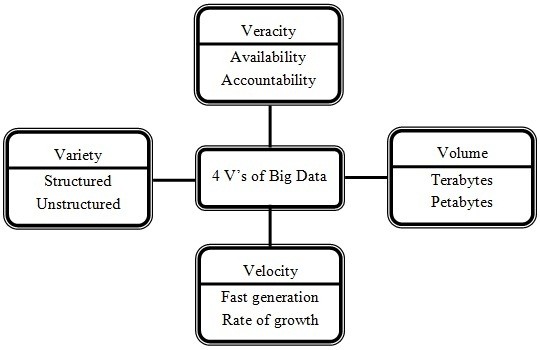


Fig. 1: Characteristics of Big Data

Reuters, for example. Given the benefits of big data, it opens up new chances in knowledge processing activities for aspiring researchers. However, opportunities always follow some difficulties.

To meet the problems, we must be familiar with varied computational complexities, information security, and computational methods for analyzing big data. Many statistical procedures, for example, that perform well with little amounts of data do not scale to large amounts of data. Similarly, many computational techniques that work well with modest amounts of data confront major hurdles when analysing large amounts of data. Many scholars have studied the various issues that the health sector faces [9], [10]. Big data analytics difficulties are divided into four basic categories: data storage and analysis; knowledge discovery and computational complexities; data scalability and visualization and information security. These topics are covered briefly in the sections that follow.

A. Data Retention and Analysis

The size of data has expanded tremendously in recent years due to numerous ways such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers, and so on. These data are stored at great expense, yet they are eventually disregarded or erased due to a lack of storage capacity. As a result, storage mediums and faster input/output speeds are the first challenges for big data analysis. In such instances, data accessibility must be a major concern for the organization. in order to uncover and represent knowledge. The primary reason is that material must be simply and quickly accessible for subsequent investigation. Analysts used hard disc drives to store data in previous decades, but they had poorer random input/output performance than sequential input/output. To solve this constraint, the solid state drive (SSD) and phrase change memory (PCM) concepts were proposed. However, available storage solutions lack the performance required for massive data processing.

Another difficulty with Big Data analysis is the variety of data. Data mining jobs have grown in importance as datasets have grown in size. Additionally, it is crucial to perform data reduction, data selection, and feature selection, especially when working with huge datasets. For researchers, this poses an unprecedented challenge. It is because, when dealing with these high dimensional data, traditional algorithms may not always reply in a timely manner. A significant difficulty in recent years has been automating this procedure and creating new machine learning algorithms to guarantee consistency. Additionally to all of these .The primary aim is the clustering of enormous datasets that aid in the analysis of big data [11]. It is now possible to gather a sizable amount of semi-structured and unstructured data in a respectable amount of time thanks to modern technologies like hadoop and mapReduce. How to effectively analyse these data in order to gain greater understanding is the main engineering challenge. To do this, a common procedure is to convert semi-structured or unstructured data into structured data, after which data mining methods are used to extract knowledge. Das and Kumar talked about a framework for data analysis [12]. Das et al. also included a similar, in-depth overview of data analysis for public tweets in their study [13]. The main difficulty in this situation is to give more thought to creating storage systems and to elevating effective data analysis tools that offer assurances on the result when the data arrives from many sources. Additionally, designing machine learning algorithms is crucial for increasing efficiency and scalability while analysing data.

B. Complexities of Knowledge Discovery and Computation

A major problem with huge data is the finding and representation of knowledge. It comprises a number of related subfields, including representation, management, archiving, and information retrieval. For example, principal component analysis [19], fuzzy set [14], rough set [15], soft set [16], near set [17], formal concept analysis [18], and soft set [16] are a few tools for knowledge discovery and representation.

To process challenges in real life, other hybridized techniques are also being created. These strategies are all based on the problem. Furthermore, with a sequential computer, some of these strategies might not be appropriate for enormous datasets. In addition, some of the techniques have good parallel computer scalability properties. Big data processing technologies may not be effective due to the exponential growth in big data size, making it difficult to extract useful information from these data. Data warehouses and data marts are the most common method for managing large datasets. Data mart is built on a data warehouse and facilitates analysis, whereas data warehouse is primarily responsible for storing data that are supplied from operational systems.

Large dataset analysis necessitates increasingly complicated computational tasks. Dealing with the uncertainties and discrepancies in the datasets is the main challenge. In most cases, computational complexity is modelled systematically. Establishing a complete mathematical framework that is universally applicable to Big Data may be challenging. However, if the relevant complexity are understood, domain-specific data analytics can be completed quickly. Such developments could be used to emulate big data analytics in various fields. The least memory-intensive machine learning algorithms have been used in a lot of research and surveys in this area. The main goal of these studies is to reduce the complexity and expense of computation [20], [21], and [22]. However, the performance of current big data analysis tools is subpar when addressing computational complexity, ambiguity, and contradictions. The development of methods and tools that can effectively cope with computational complexity, ambiguity, and inconsistencies presents a significant challenge.

III. BIG DATA ANALYTICS OPEN RESEARCH ISSUES

In both academia and industry, big data analytics and data science are taking centre stage in research. Big data and information extraction from data are the subjects of data science research. Information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing are examples of applications for big data and data science. The future drift of events can be predicted with the help of an efficient integration of technology and analysis. This section's primary goal is to present current open research questions in big data analytics. The internet of things (IoT), cloud computing, bio-inspired computing, and quantum computing are the three broad categories into which the research questions relevant to big data analysis are divided. But it is not constrained. not just to these problems. In the work of Husing Kuo et al. [9], more research concerns with big data in healthcare are discussed.

A. Big Data Analytics using IoT

The internet has transformed a staggering array of personal traits, economic practises, cultural revolutions, and international relations. At this time, machines are joining the action to construct Internet of Things (IoT) and control countless autonomous devices online. As a result, appliances are starting to use the internet in the same way that people do with web browsers. Recent academics are paying close attention to the Internet of Things because of its most exciting potential and difficulties. It has a crucial societal and economic impact on how information, network, and communication technology will be built in the future. The new rules for everything will eventually be interconnected and intelligently controlled in the future. Due to the advancement of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics, the idea of IoT is becoming more applicable to the real world. IoT also poses difficulties when volume, velocity, and variety are combined. In a broader sense, the Internet of Things, like the internet, allows for the existence of devices in a variety of locations and supports applications ranging from the trivial to the essential. On the other hand, it continues to be baffling to fully comprehend IoT, including definitions, substance, and differences from other related ideas. Big data and computational intelligence, among other diverse technologies, can be combined to enhance large-scale automation's data management and knowledge discovery. Mishra, Lin, and Chang have conducted a significant amount of study in this area [27].

The largest challenge that big data professionals are experiencing is learning from IoT data. Therefore, building infrastructure for IoT data analysis is crucial. A continuous stream of data is generated by an IoT device, and researchers can create tools to extract useful information from this data using machine learning techniques. Big data analytics is a necessary solution for understanding these streams of data produced by IoT devices and evaluating them to obtain useful information. From an Internet of Things (IoT) perspective, only machine learning algorithms and computational intelligence techniques can handle large data. Numerous academic publications also cover important IoT-related technologies [28]. An overview of the IoT big data and knowledge discovery process is shown in Figure 2.

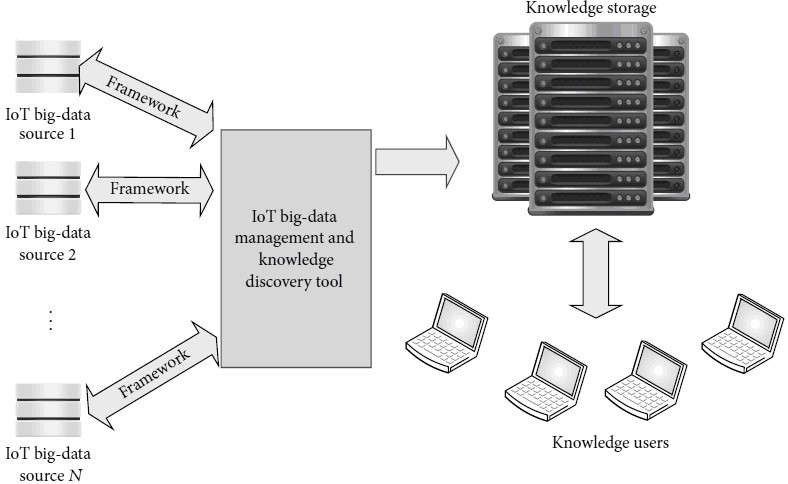


Fig. 2: IoT Big Data Knowledge Discovery

B. Using the Cloud for Big Data Analytics

The advancement of virtualization technologies has increased the affordability and accessibility of supercomputing. Systems behave like real computers thanks to computing infrastructures disguised in virtualization software, but with more freedom in terms of specification specifics like processor count, disc space, memory, and operating system. Cloud computing, one of the most effective big data techniques, is the utilisation of these virtual machines. Big Data and cloud computing technologies are created with the goal of creating scalable and on-demand resource and data availability. By providing on-demand access to reconfigurable computer resources via virtualization techniques, cloud computing harmonises enormous amounts of data. Offering resources when there is a need and paying solely for the resources required to build the product are two advantages of using cloud computing. It increases availability while lowering costs simultaneously. Numerous researchers have highlighted in-depth open challenges and research concerns of big data and cloud computing, highlighting the difficulties in data management, data variety and velocity, data storage, data processing, and resource management [29], [30]. Therefore, using infrastructure and technologies, cloud computing aids in creating a business model for all types of apps.

IV. BIG DATA PROCESSING TOOLS

There are many different tools available to process big data. In this section, we go over some of the current methods for analysing big data with a focus on MapReduce, Apache Spark, and Storm, three crucial new tools. The majority of the tools that are now available are focused on batch processing, stream processing, and interactive analysis. The majority of batch processing technologies, like Mahout and Dryad, are built on the Apache Hadoop architecture. Real-time analytics is where stream data applications are most frequently deployed. Strom and Splunk are two examples of large-scale streaming platforms. Users can immediately interact in real time for their own analysis using interactive analysis.

A. MapReduce and Apache Hadoop

The most well-known software platform for large data analysis is comprised of MapReduce and Apache Hadoop. Hadoop distributed file system (HDFS), map-reduce, the hadoop kernel, and apache hive are some of its components. Divide and conquer is the foundation of the map reduce programming model, which is used to process big datasets. The map step and reduce step are the two steps that make up the divide and conquer strategy. Master nodes and worker nodes are the two different types of nodes that make up Hadoop. In the map stage, the master node divides the input into smaller subproblems before sending them on to the worker nodes. The outputs for all of the subproblems are then combined by the master node in the reduction stage. Additionally, Hadoop and MapReduce function as a strong provides a strong software framework for handling big data issues. High throughput data processing and fault-tolerant storage both benefit from it.

B. Apache Mahout

The goal of Apache Mahout is to offer scalable and profitable machine learning methods for sophisticated and large-scale data processing applications. Mahout's fundamental algorithms, including as clustering, classification, pattern mining, regression, dimension reduction, evolutionary algorithms, and batch-based collaborative filtering, are implemented using the map-reduce architecture on top of the Hadoop platform. Mahout aims to create an active, receptive, and varied community to support conversations about the project and prospective use cases. Apache Mahout's main goal is to offer a tool for overcoming significant obstacles. Companies like Google, IBM, Amazon, Yahoo, Twitter, and Facebook have all implemented scalable machine learning algorithms [31].

C. Apache Drill

Another distributed system for interactive large data analysis is Apache Drill. It can accommodate a wider range of query languages, data formats, and data sources thanks to its increased flexibility. It is also specifically made to take use of nested data. Additionally, it aims to expand up to at least 10,000 computers and achieve the capacity to process petabytes of data and trillions of records in a single second. Drill employs map reduce for batch analysis and HDFS for storage.

V. ADVICE FOR FUTURE WORK

The latter approach is not recommended because it can occasionally result in information loss. This raises numerous research concerns in the scientific community and industry regarding efficient data collection and access. Another difficulty is the need for quick processing that nevertheless achieves high speed and high throughput, as well as effective data storage for later use. Programming for large data analysis is another crucial and difficult subject. There is an urgent need to express data access needs for programmes and provide programming language abstractions to take advantage of parallelism [32]. In order to facilitate meaningful results from these concepts, machine learning concepts and technologies are also becoming more and more popular among researchers. Data processing, algorithmic implementation, and optimisation have been the main areas of research in the field of machine learning for big data. Many of the recently developed machine learning technologies for big data require significant adjustment to be adopted. We contend that even if each tool has benefits and drawbacks of its own, more effective tools can be created to address big data's inherent issues. The effective tools that need to be created must be able to deal with data that is noisy and unbalanced, as well as with uncertainty, inconsistent results, and missing numbers.

VI. CONCLUSION

In recent years, data production has increased dramatically. For a regular man, analysing these facts presents a challenge. In order to do this, we examine the many research problems, difficulties, and analytical techniques in this study. It is clear from this poll that each big data platform has a distinct focus. While some of them excel in real-time analytics, others are better suited for batch processing. Additionally, each big data platform offers particular features. Statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing are some of the several analytical methods used. We believe that in the future, researchers will focus more on these strategies to successfully handle big data difficulties.

REFERENCES

1. M. K.Kakhani, S. Kakhani and S. R.Biradar, *Research issues in big data analytics*, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.
2. A. Gandomi and M. Haider, *Beyond the hype: Big data concepts, meth- ods, and analytics*, International Journal of Information Management, 35(2) (2015), pp.137-144.
3. C. Lynch, *Big data: How do your data grow?*, Nature, 455 (2008), pp.28-29.
4. X. Jin, B. W.Wah, X. Cheng and Y. Wang, *Significance and challenges of big data research*, Big Data Research, 2(2) (2015), pp.59-64.
5. R. Kitchin, *Big Data, new epistemologies and paradigm shifts*, Big Data Society, 1(1) (2014), pp.1-12.
6. C. L. Philip, Q. Chen and C. Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on big data*, Infor- mation Sciences, 275 (2014), pp.314-347.
7. K. Kambatla, G. Kollias, V. Kumar and A. Gram, *Trends in big data analytics*, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.
8. 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.
9. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, *Health big data analytics: current perspectives, challenges and potential solutions*, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.
10. R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, *A look at challenges and opportunities of big data analytics in healthcare*, IEEE International Conference on Big Data, 2013, pp.17-22.
11. Z. Huang, *A fast clustering algorithm to cluster very large categorical data sets in data mining*, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.
12. T. K. Das and P. M. Kumar, *Big data analytics: A framework for unstructured data analysis*, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.
13. T. K. Das, D. P. Acharjya and M. R. Patra, *Opinion mining about a product by analyzing public tweets in twitter*, International Conference on Computer Communication and Informatics, 2014.
14. L. A. Zadeh, *Fuzzy sets*, Information and Control, 8 (1965), pp.338- 353.
15. Z. Pawlak, *Rough sets*, International Journal of Computer Information Science, 11 (1982), pp.341-356.
16. D. Molodtsov, *Soft set theory first results*, Computers and Mathe- matics with Aplications, 37(4/5) (1999), pp.19-31.
17. J. F.Peters, *Near sets. General theory about nearness of objects*, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.
18. R. Wille, *Formal concept analysis as mathematical theory of concept and concept hierarchies*, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.
19. I. T.Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.
20. O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and

K. Taha, *Efficient machine learning for big data: A review*, Big Data Research, 2(3) (2015), pp.87-93.

1. Changwon. Y, Luis. Ramirez and Juan. Liuzzi, *Big data analysis using modern statistical and machine learning methods in medicine*, International Neurourology Journal, 18 (2014), pp.50-57.
2. P. Singh and B. Suri, *Quality assessment of data using statistical and machine learning methods*. L. C.Jain, H. S.Behera, J. K.Mandal and

D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.

1. A. Jacobs, *The pathologies of big data*, Communications of the ACM, 52(8) (2009), pp.36-44.
2. H. Zhu, Z. Xu and Y. Huang, *Research on the security technology of big data information*, International Conference on Information Technology and Management Innovation, 2015, pp.1041-1044.
3. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, *Survey of research on information security in big data*, Congresso da sociedada Brasileira de Computacao, 2014, pp.1-6.
4. I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, *Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives*, BioMed Research International, 2014, (2014), pp.1-13.
5. N. Mishra, C. Lin and H. Chang, *A cognitive adopted framework for iot big data management and knowledge discovery prospective*, International Journal of Distributed Sensor Networks, 2015, (2015), pp. 1-13
6. X. Y.Chen and Z. G.Jin, *Research on key technology and applications for internet of things*, Physics Procedia, 33, (2012), pp. 561-566.
7. M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, *Big data computing and clouds: Trends and future directions*, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.
8. I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and

S. Ullah Khan, *The rise of big data on cloud computing: Review and open research issues*, Information Systems, 47 (2014), pp. 98-115.

1. L. Wang and J. Shen, *Bioinspired cost-effective access to big data*, International Symposium for Next Generation Infrastructure, 2013, pp.1- 7.
2. C. Shi, Y. Shi, Q. Qin and R. Bai *Swarm intelligence in big data analytics*, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise,

B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, 2013, pp.417-426.