INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

CHALLENGES AND ISSUES IN BIG DATA ARCHITECTURE FOR ANALYTICS

KAPPERA RAMESH
Assistant professor, Malla Reddy university, Maisammaguda, Dulapally, Hyderabad, Telangana 500043
ramesh.reddy531@gmail.com

Abstract - These days, a lot of the data we possess comes from digital technologies and contemporary information systems, like the cloud and the internet of things. Even if we receive a lot of data, it can be very challenging to analyse it and draw conclusions that will help us make decisions. These days, analysing such data—what we refer to as big data analysis—is crucial. It's a big field for study and advancement. This paper provides an overview of big data, including its definition, problems, unresolved research questions, and related tools. It also assists researchers in creating solutions based on problems and unresolved research concerns.

Keywords: Hadoop, big data analytics, structured and unstructured data, massive data

1 INTRODUCTION

Big data is growing as a result of the rapid advancement of digital technologies, which generates data. It is challenging to use conventional database administration tools or data processing applications to handle the collection of massive datasets from these devices. These data typically have sizes of petabytes or more. Structured, unstructured, or semi-structured ones are possible. Such data are technically defined by the 3Vs, or volume, velocity, and variety. While velocity refers to the rate of growth and the speed at which the data are gathered for analysis, volume alludes to the enormous amount of data that are generated every day. Variety tells us about the several kinds of data—structured, unstructured, semi-structured, etc. Veracity, which encompasses availability and responsibility, is defined by the fourth V. Processing data with high volume, velocity, diversity, and veracity through a combination of conventional and computationally intelligent techniques is the main goal of big data analysis [1]. A significant amount of data available nowadays is not natively structured. For instance, tweets and blogs are poorly structured textual documents, whereas images and videos are structured for storage and presentation purposes only, not for semantic content or search. Converting such content into a structured format for subsequent analysis is a difficult task.

Big data is a stronghold for the upcoming information technology sectors [2]. Numerous IT sectors that discuss big data, cloud computing, the internet of things, and social media can resolve a number of intentional problems. Large datasets cannot be handled by data mining, even when data warehouses are used to manage datasets. Extracting knowledge data from the available bigdata is an extremely laborious procedure. The main issue with big data analysis is the absence of communication between database systems and analytical tools like statistics and data mining. In these kinds of circumstances, discovering new information is difficult. Another significant requirement is the representation of data for real-world applications. The epistemological implications of the data revolution must also be considered [3]. Research on the complexity theory of big data will also aid in comprehending the

ISSN: 2057-5688

fundamental traits and complex pattern formation of big data, simplify its representation, improve knowledge abstraction, and direct the development of big data computing models and algorithms. It should be remembered, though, that not all data that is available in the big data format is helpful for analysis or decision-making. The dissemination of big data insights is of importance to both industry and academia. This paper focusses on large data difficulties and available methods. We also list open research questions related to big data. To further explain, the article addresses the difficulties that come up while fine-tuning large data, offering open research questions that will facilitate the processing of big data and the extraction of valuable knowledge from it, as well as offering an overview of big data tools and methodologies. The final section includes concluding remarks that provide a summary of the results

2 CHALLENGES IN BIG DATA ANALYTICS

Big data isn't related to just one domain. Boundaries extend beyond fields such as biochemistry, retail, public administration, health care, and other multidisciplinary scientific studies. The primary sources of big data are internet text and document repositories, search indexing on the internet, and social computing. Social computing comprises online communities, recommender systems, reputation systems, prediction markets, and social network analysis. With these benefits in mind, big data offers up new avenues for knowledge processing tasks for future academics. Researchers work mostly in the health industry, which presents a number of obstacles. [4] [5] Opportunities, however, inevitably follow challenges.

It will be challenging to overcome the obstacles that big data presents. The amount of data is already massive and keeps growing daily. Its creation and expansion are happening more quickly now, partly due to the rise in internet-connected gadgets. Additionally, organisations are producing a growing variety of data, and their capacity to collect and handle this data is constrained. To fully utilise big data, organisations will need to adapt the way they plan, control, manage, process, and report on data. Current technology, architecture, management, and analysis methodologies are insufficient to handle the deluge of data.

2.1 Data Storage and Analysis:

Through a variety of methods, including mobile devices, aerial sensory technologies, remote sensing, etc., the amount of data is growing very quickly. As a result, some valuable data may be erased since there is not enough room to keep such large amounts of data. Therefore, faster input/output speeds and storage mediums are the initial obstacle for large data processing. In these kinds of situations, the finding and representation of information must place the accessibility of the data first. The main justification is that it needs to be quickly and easily accessible for additional examination. Hard disc drives were once used by analysts to store data, however their random input/output performance was inferior to that of sequential input/output. Phrase change memory (PCM) and solid state drives (SSD) were developed as solutions to this drawback. However, the performance needed for processing large amounts of data cannot be achieved by the current storage technology. The amount of data mining tasks has greatly expanded due to the constantly expanding datasets. Another significant obstacle for big data is this. Data reduction, data selection, and

ISSN: 2057-5688

feature selection are applied when working with huge datasets. For researchers, this poses an unprecedented challenge. When working with very high dimensional data, current algorithms might not always react quickly enough. One of the biggest challenges in recent years has been automating this procedure and creating new machine learning algorithms to guarantee consistency. Furthermore, clustering of large datasets is a major problem as it aids in the analysis of big data [6]. Large volumes of semi-structured and unstructured data may be gathered quickly thanks to Hadoop and MapReduce. How to efficiently analyse these data to gain more knowledge is the main engineering challenge. Converting semi-structured or unstructured data into structured data and then using data mining techniques to extract knowledge is a common procedure to achieve this goal. Das and Kumar have talked about a framework for data analysis [7]. Das et al. included a similar in-depth overview of data analysis for public tweets in their study [8]. In this situation, the main task is to focus more on the architecture of storage systems and to advance effective data analysis tools that offer output assurances when the data originates from various sources. Furthermore, increasing scalability and efficiency requires carefully designing machine learning algorithms for data analysis.

2.2 Computational Complexities and Knowledge Discovery:

In big data, finding and representing knowledge is a major problem. It encompasses several subfields, including information retrieval, representation, archiving, management, and authentication. A number of technologies are available for knowledge representation and discovery, including formal concept analysis [13], principal component analysis [14], fuzzy set [9], rough set [10], soft set [11], near set [12], and so on. Several of these methods might not be appropriate for sequential computers with big datasets. Simultaneously, certain methods exhibit favourable scalability over parallel computing systems. Big data is growing exponentially in size, and the technologies that are currently available may not be able to process this data well enough to yield information that is useful. Data marts and warehouses are the most widely used method for managing huge datasets. While data marts are built on top of data warehouses and enable analysis, data warehouses are primarily in charge of storing data that comes from operational systems. Large dataset analysis necessitates higher computational complexity. Managing the uncertainties and inconsistency in the datasets is the main problem. Generally speaking, computational complexity is modelled systematically. Developing a thorough mathematical framework that can be used with Big Data may be challenging. However, if the relevant complications are understood, domain-specific data analytics can be completed with ease. A sequence of these developments could serve as a model for big data analytics in several domains. In this direction, a great deal of study and survey work has been done utilising machine learning methods that use the least amount of memory. Reducing computing costs and processing difficulties is the main goal of this research [15], [16], [17].

2.3 Scalability and Visualization of Data:

Scalability and security of big data analysis tools are two of the biggest obstacles. Researchers have focused on accelerating data analysis and how it speeds up computers in recent decades, which has led to Moore's Law. The development of

ISSN: 2057-5688

sampling, online, and multi-resolution analysis techniques is required for the former. In terms of big data analysis, incremental approaches have high scalability properties. With more cores integrated into processors, there is a natural and drastic transition in processor technology as data sizes are growing considerably quicker than CPU speeds. It is this change in processors that gives rise to parallel computing. Parallel computing is necessary for real-time applications such as online search, social networking, banking, navigation, and timeliness, among others. The goal of data visualisation is to use some graph theory techniques to convey the data more effectively. The connection between data and appropriate interpretation is made possible by graphic visualisation. Nonetheless, millions of people utilise online marketplaces like Flipkart, Amazon, and eBay every month, and they sell billions of goods. It produces a large amount of data. Some businesses utilise Tableau, a platform for large data visualisation, to do this. It can convert vast and complicated data into clear visual representations. This aids in the visual representation of search relevance, the tracking of recent customer comments, and sentiment analysis for staff members. Unfortunately, the majority of big data visualisation technologies available today perform poorly in terms of functionality, scalability, and reaction time. We can see that the development of hardware and software that supports distributed computing, cloud computing, parallel computing, visualisation, and scalability has faced numerous obstacles as a result of big data. We must correlate more mathematical models with computer science in order to solve this problem.

2.4 Information Security:

Large volumes of data are correlated, examined, and mined for interesting patterns in big data analysis. Every organisation has a particular set of rules in place to protect sensitive data. Sensitive data preservation is a significant problem for big data analysis. Big data comes with a significant security risk. As a result, big data analytics is starting to focus on information security. Authentication, authorisation, and encryption approaches can be used to improve large data security. Big data applications phase in a number of security measures, including network scale, device diversity, real-time security monitoring, and absence of intrusion system. Information security professionals have been interested in the security threat posed by big data. Consequently, it is necessary to focus on creating a layered security policy model and preventative mechanism.

3 RESEARCH AREAS IN BIG DATA ANALYTICS

Without a doubt, everything we decide to do with big data has the potential to become a very important catalyst for value generation and innovation. Three dimensions can be used to group the difficulties posed by big data: data, process, and management.

3.1 Data Challenges

Volume: Managing enormous volumes of information is the primary difficulty.
Variety: Managing the multitude of data sources, forms, and types is a challenge.
Velocity: One of the main issues is knowing how to respond to the deluge of data in the amount of time that the application demands.
Veracity: comprises both data availability and data quality.

ISSN: 2057-5688

Validity: It is implied by validity that the data is accurate and correct for the intended application. Clearly, having reliable data is essential to making wise decisions. additionally the instruments they provide to support data correctness and authenticity.
Volatility: Two key questions in big data can be answered by the concept of volatility.

Data validity, or how much data is valid, is one of them. The other is how long data should be stored. In the age of real-time data, you must ascertain whether a piece of data is no longer pertinent to the study at hand. It is evident that big data management addresses concerns including validity, volatility, and veracity in addition to volume, diversity, and velocity. To learn about further big data trends, presentation, and applications, click here.

In what ways can we deal with ambiguity, imprecision, missing numbers, statements that are lacking, or lies? To what extent is the data accurate? To what extent is the coverage comprehensive? What is the sample resolution's fineness? How current do the readings seem? To what extent are the sampling biases understood? Is there any data available?
Data discovery presents a formidable challenge: how to sift through the enormous amounts of data available on the Internet to locate high-quality data?
The difficulty lies in assessing the quality of data sets and their applicability to specific problems (i.e., does the data set contain underlying assumptions that make it biassed or uninformative for a certain subject).
Data completeness: Which are the uncovered areas? What ramifications result?
Personally identifiable information: Is it possible to obtain sufficient data to assist individuals without jeopardising their privacy in the process?

3.2 Process Challenges

How to process the massive amounts of data in order to extract the important information is the main difficulty. One of the process issues is gathering data. Sync up the information gathered from various sources. converting the data into the format needed for analysis. modelling the information gathered. Recognising, displaying, and disseminating the output.

3.3 Management challenges

Data privacy, security, and governance are the three primary management challenges. Making sure that data is used appropriately is the primary difficulty. Private information is among the sensitive data kept in data warehouses. The availability to such data raises legal concerns. Therefore, it is imperative that organisations ensure the safe and secure usage of big data technologies.

4 SUGGESTIONS FOR FUTURE WORK

It is anticipated that the volume of data gathered from several applications spread over the globe and a wide range of industries would double every two years. It is useless unless these are examined to provide insightful data. This calls for the creation of methods that will make large data analysis easier. The implementation of these strategies leading to automated systems is made possible by the advancement of powerful computers. With high performance large-scale data processing, turning data into knowledge is by no means a simple undertaking. This includes taking advantage

ISSN: 2057-5688

of the parallelism of existing and future computer architectures for data mining. Furthermore, there could be a wide range of uncertainties in these data. Many models have been shown to be effective in expressing data, including fuzzy sets, rough sets, soft sets, neural networks, their generalisations, and hybrid models created by merging two or more of these models. Additionally, these models lend themselves quite well to analysis. Big data are typically condensed to just include the crucial features required from the perspective of a certain study or based on the application area. Thus, methods for reduction have been created. The gathered data frequently contains missing values. Before analysis, these values must be generated or the tuples containing these missing values must be removed from the data set. More significantly, the performance, effectiveness, and scalability of the specialised data-intensive computing systems may be compromised by these additional difficulties, and in certain cases they may even worsen. The latter strategy is not recommended since it can occasionally result in information loss. This raises a number of research questions for efficiently collecting and obtaining data that are of interest to both the scientific community and industry. Another problem is processing data quickly without sacrificing performance or throughput, and then storing it effectively for later use. Furthermore, one significant and difficult problem is programming for big data analysis. There is an immediate need to express application data access needs and create programming language abstractions to take advantage of parallelism. In addition, machine learning concepts and techniques are becoming more and more popular among researchers in order to enable these notions to yield meaningful results. Machine learning for big data research has primarily concentrated on data processing, algorithm implementation, and optimisation. Since many machine learning methods for big data are still new, adopting them will need significant changes. We contend that although each tool has benefits and drawbacks, more effective tools might be created to address issues that arise from working with large amounts of data. The effective tools that need to be created must be able to deal with missing values, ambiguous and inconsistent data, and noisy and imbalanced data.

5 CONCLUSION

Data generation has accelerated dramatically in recent years. This data analysis is difficult for a general individual to do. In order to do this, we review the many research questions, difficulties, and instruments for analysing these large data sets in this study. It is clear from this poll that each big data platform has a unique focus. While some of them are well suited for real-time analytics, others are made for batch processing. Every big data platform offers features that are unique as well. Statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing are some of the several approaches utilised in the analysis. We think that in the future, more research will focus on these methods to effectively and efficiently handle large data problems.

6 ACKNOWLEDGMENTS We would like to express our gratitude to the specialists who helped build the template.

7 REFERENCES

ISSN: 2057-5688

[1] Bag S, Wood LC, Xu L, et al.: Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. Resour Conserv Recycl. 2020; 153: 104559.
[2] Begenau J, Farboodi M, Veldkamp L: Big data in finance and the growth of large firms. J Monetary Econ. 2018; 97: 71–87.
[3] Belhadi A, Zkik K, Cherrafi A, et al.: Understanding big data analytics for manufacturing processes: Insights from literature review and multiple case studies. Comput Ind Eng. 2019; 137:
[4] Cockcroft S, Russell M: Big data opportunities for accounting and finance practice and research. Aust Account Rev. 2018; 28(3): 323–333.
[5] Diebold FX, Ghysels E, Mykland P, et al.: Big Data in Dynamic Predictive Econometric Modeling. Elsevier: Amsterdam, The Netherlands. J Econom. 2019;212(1): 1–3.
[6] Dubey R, Gunasekaran A, Childe SJ, et al.: Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. Int J Prod Econ. 2020; 226: 107599.
[7] Garoufallou E, Gaitanou P: Big data: opportunities and challenges in libraries, a systematic literature review. Coll Res Libr. 2021; 82(3): 410.
[8] Grover P, Kar AK: Big data analytics: A review on theoretical contributions and tools used in literature. Glob J Flex Syst Manag. 2017; 18: 203–229.
[9] Huang L, Wu C, Wang B: Challenges, opportunities and paradigm of applying big data to production safety management: From a theoretical perspective. J Clean Prod. 2019; 231: 592–599.
[10] Ji W, Yin S, Wang L: A big data analytics based machining optimisation approach. J Intell Manuf. 2019; 30: 1483–1495.
[11] 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.
[12] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and. Taha, Efficient machine learning for big data: A review, Big Data search, 2(3) (2015), pp.87- 93.
[13]]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. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2(2014),
[14] 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.
[15] 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.
[16] R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.
[17] 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.