e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

TOOLS AND CHALLENGES IN BIG DATA ANALYTICS

Khushbu H Zambre[*1], Dr.Dinesh D.Patil[*2], Assit. Prof. Dhiraj G. Patil[*3]

[*1,2,3]Shri Sant Gadge Baba College Of Engineering & Technology Bhusawal, Dist Jalgaon, India.

DOI : https://www.doi.org/10.56726/IRJMETS60111

ABSTRACT

Big data analytics techniques for data-driven industries are becoming more and more popular in corporate intelligence, industrial operations, and research, and they are also quickly altering how people view industrial revolutions. Regarding the appropriate information and big data analytics methods to extract knowledge from massive, important industrial data that could aid in handling big data formats, academics and practitioners are significantly lacking in expertise. Despite the numerous study initiatives and academic studies on big data analytics procedures for enhancing industrial performance that have been put forth recently. thorough examination, methodical data-driven analysis, comparison, and exacting evaluation of techniques, data sources

Keywords: Big Data Analytics, Machine Learning, Big Data Analytics Tools, Review, Big Data Management.

I. INTRODUCTION

We live in a world where there is an abundance of data generated in every aspect of our life due to the quick growth of new technologies like social media, cloud computing, Internet of Things (IoTs), mobile and smart devices, and the Internet. Additionally, the public, private, and social sectors never stop Mauro Gaggero was the associate editor who oversaw the manuscript's evaluation and gave it the go-ahead to be published. generate enormous volumes of data from various sources in a range of formats. The higher education ecosystem, which includes various platforms and systems like learning management systems (LMS), massive open online courses (MOOC), open courseware (OCW), open educational resources (OER), and social media sites like Facebook, Twitter, and Instagram, is where educational data are quickly gathered and generated.

1.1 BIG DATA ANALYTICS

The dataset is generated at enormous scales and has high velocity, variety, and volume. Using a variety of methods, including machine learning, predictive analytics, data mining, statistics, text analytics, and deep learning analytics, the generated big data assists analysts, researchers, and businesspeople in making wise decisions.

The two main goals of big data analysis are to create efficient techniques that can reliably predict outcomes and to obtain understanding of the relationships between features. As the digital world develops, unstructured big data of all shapes and sizes is produced. These data present a number of difficulties in the current.

II. METHODOLOGY

2.1 TOOLS IN BIG DATA ANALYTICS

Developed as part of Apache projects, some of these tools are open source software components that help organizations handle structured, semi-structured, and unstructured data. The primary studies that made use of these tools were examined and evaluated critically in this section. The tools were divided into a number of themes, including intelligent computing techniques, batch processing, stream processing, and fusion processing models. The section also covered the ecosystem for Hadoop big data processing and outlined the advantages and disadvantages of the different big data analytic tools. Hadoop is an essential software distribution technique for managing and processing large amounts of data.

Based on a column-oriented value data model, HBase [44] is a distributed, non-relational database that offers scalable and effective big data storage. Furthermore, HBase leverages real-time, random read/write access on top of Hadoop and the Hadoop Distributed File System to store and process large amounts of data with features akin to Bigtables. Name, replicas, and other Namenode Metadata; Customer Metadata Operations Examine the Datanodes Blocks in Racks 1 and 2 Block operations for replication Nodes for data Write Fig. 3 down. General architecture of Hadoop 15 (b) Built in C++, MongoDB [45] is an open-source, cross-platform project. Moreover, it is a document-oriented database with high performance, high availability, and ease of scalability that offers a

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

data model based on JSON documents and supported by BSON. Cassandra is a distributed database system [46, 47].

2.2 CHALLENGES IN BIG DATA ANALYTICS

The implementation of effective data-driven industries and decision-making processes has been challenged in recent years by the massive accumulation of data [135]–[137]. Numerous difficulties that are inherent in big data analytics for data-driven industries were discovered after a thorough analysis of the chosen studies [6,], [9], and [138]. These difficulties include the following: (i) big data management; (ii) data storage system scalability; (iii) data veracity; (iv) infrastructure readiness; (v) skill shortage; (vi) culture; (vii) data complexity; (viii) infringement on intellectual property rights; (ix) data aggregation and cleaning; and (x) privacy and security. Infrastructure Readiness: When developing new, independent platforms or updating current ones, organizations often struggle to supply all the infrastructures required. Infrastructure continues to be a major obstacle to data efficiency since more infrastructure means more facilities and policies that will be available to support infrastructure in the future [6].

Availability, integrity, scalability, and secrecy Cloud computing [6] technologies like virtualization, distributed processing, and storage have made it possible to complete tasks that were difficult to complete with traditional data processing systems. However, there are a number of problems with the power and capability of cloud storage privacy. Organizations and industries are now reluctant to move sensitive data to the cloud unless the cloud storage system has strong security measures in place.

The Apache Foundation created Hadoop, an open source distributed data processing system infrastructure. It makes it possible to process massive data sets in parallel and distributed over numerous computer clusters. It has excellent fault tolerance, scalability, dependability, efficiency, and low cost. MapReduce, the HDFS distributed file system, and a number of general-purpose tools make up Hadoop. Security and privacy: To be clear, privacy [11] has caused a great deal of anxiety in big data operations. Data collection, processing, transmission, storage, and administration present security risks [142]. Even with cloud computing's clear benefit of lower capital and operating costs, organizations still find it difficult to outsource their data storage to third-party clouds because doing so usually results in less physical control over their data. In particular, the ability to handle multiple cloud platforms and multi-tenancy as well as massive computation. Enhancement of professional skills and training: Lifelong learning and professional development are essential for skilled workers, data analysts, and other professionals. For example, companies and academic institutions should work together to match the curriculum in schools with the needs of the big data industry. In addition, business and academic institutions ought to collaborate to offer hands-on training to fill in the skills gaps in the big data and data analytics fields [6].

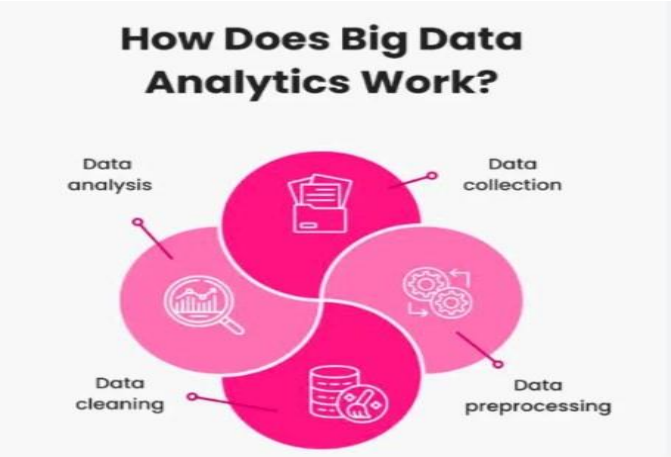
The vast and growing amounts of online education data present a number of technological opportunities as well as challenges for using big data in learning analytics and education. A high-performance computational infrastructure that can manage massive volumes of data for capture, storing, processing, and visualization would be necessary for big education systems.

Introduction to Cost Management: Provide an overview of the financial challenges associated with big data analytics. Cost Components: Hardware and Storage: Discuss the costs associated with acquiring and maintaining the hardware and storage infrastructure for big data. Software and Licensing: Outline the expenses related to software licenses and proprietary analytics tools. Human Resources: Analyze the costs of hiring and retaining skilled data scientists and engineers. Cost Reduction Strategies: Open Source Solutions: Discuss the role of open-source tools in reducing software costs. Optimizing Storage: Explain strategies for optimizing storage costs, such as data compression and tiered storage solutions. Automation: Describe how automation can help in reducing operational costs.

Information from users is gathered and used, frequently without the users' knowledge, to expand an industry or offer greater value to businesses. The concept of user privacy is frequently violated by synonyms of data sharing between organizations for investigative purposes, which aim to prevent data reutilization. For example, there is frequently the problem of social media privacy and policies (WhatsApp, Yahoo, Twitter, and Facebook) that are not made clear to users at the time of registration. In 2017 and 2018, Facebook faced allegations of data breaches, while Twitter faced allegations of selling a substantial amount of tweets to a big data dealer [143].

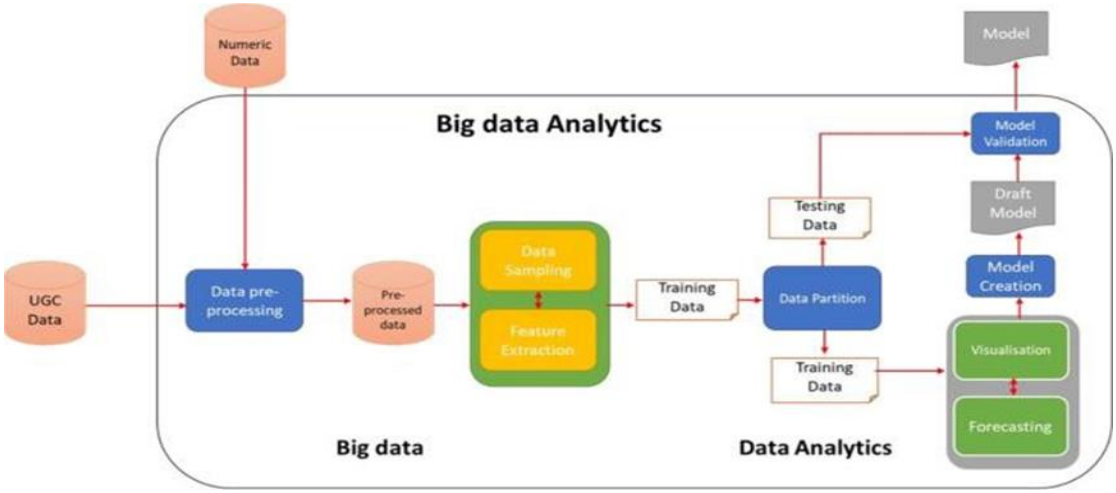
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

The concern over user privacy grows as a result. Big data is sourced from a multitude of sources, including traffic streams, GPS, location trackers, log data, Blockchain, ERC-20 token project Omise Go, and other cryptocurrency projects like Bitcoin, Ripple, and Binance [147], [148]. Such data collection and sharing may result in privacy violations and gaps. Furthermore, the transmission of personal data over the internet raises concerns about user privacy related to identity. Additional difficulties with data collection and integration, including disruption, modification, inspection, disclosure, unauthorized access, recording, and destruction [11], have made privacy and security concerns even worse. Describe your business rules for data aggregation, cleaning.



Diag.1 Working of Big Data

Diag.2 Big data tools



Diag.3 Big data Analytics Processing

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:07/July-2024 Impact Factor- 7.868 www.irjmets.com

III. CASE STUDY

Netflix:The online entertainment company's 148 million subscribers give it a massive BI advantage.

Netflix has digitized its interactions with its 151 million subscribers. It collects data from each of its users and with the help of data analytics understands the behavior of subscribers and their watching patterns. It then leverages that information to recommend movies and TV shows customized as per the subscriber's choice and preferences.

As per Netflix, around 80% of the viewer's activity is triggered by personalized algorithmic recommendations. Where Netflix gains an edge over its peers is that by collecting different data points, it creates detailed profiles of its subscribers which helps them engage with them better.

IV. FUTURE SCOPE

Mobile cellular networks have an excellent opportunity to improve performance through big data analytics. The amount of data generated by the uncontrollably rapid expansion of mobile sensing applications may surpass the processing capacity of the server. As a result, big data analytics is an appropriate technique for handling massive amounts of data. Big data on the internet, however, also creates a bottleneck for real-time data, which is needed for integrated mobile sensors, video surveillance, visual maps, and video games. As a result, the fth generation of network standards is suggested, which will increase network speed tenfold [101]. Furthermore, it's critical to compute the intricate details of deep learning techniques given the prevalence of smartphones among today's youth. The deep learning approach's ability to create deep networks allows for the development of intricate conceptual hierarchies.

V. CONCLUSION

In this paper, we present a summary of the definition of big data from a number of recent studies, wherein the term "big data" is limited to three variables: volume, variety, and velocity. In order to better understand the significance of big data, other researchers added three more Vs: value, variability, and veracity, in addition to complexity. We also talk about the difficulties posed by big data in terms of the complexity factor and the six V's. Furthermore, we concentrate on the advantages of big data analytics, which are divided into five categories: text, voice, video, network, and geospatial analytics. A few examples are provided to highlight the true benefits of big data analytics.

VI. REFERENCES

[1] G. I. Glosarry, The importance of big data: A de nition, Tech. Rep., Gartner, Stamford, CT, USA, Jun. 2012.

[2] N. Mohamed and J. Al-Jaroodi, Real-time big data analytics: Appli cations and challenges, in Proc. Int. Conf. High Perform. Comput. Simulation (HPCS), Jul. 2014, pp. 305310.

[3] Z. Sun, K. Strang, and S. Firmin, "Business analytics-based enterprise information systems," J. Comput. Inf. Syst., vol. 57, no. 2, pp. 169–178, 2016.

[4] J. Li, F. Tao, Y. Cheng, and L. Zhao, "Big data in product lifecycle management," Int.J.Adv.Manuf.Technol., vol.81, no.1–4, pp. 667–684, 2015.

[5] L. Da Xu and L. Duan, "Big data for cyber physical systems in industry 4 . 0 : a survey," Enterp. Inf. Syst., vol. 7575, pp. 0–22, 2018.

[6] D. M. Shah, N. D., Steyerberg, E. W., & Kent, "Big data and predictive analytics: recalibrating expectations," Jama Netw., vol. 320, no. 1, pp. 27–29, 2018.

[7] E. Graham-Harrison and C. Cadwalladr, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach," Guard., pp. 1–5, 2018.

[8] A. Taylor, "The 5 worst big data privacy risks (and how to guard against them)," 2017. [Online]. Available: https://www.csoonline.com/article/2855641/privacy/the-5-worst-big-data-privacy-risks-and-how-to-guard against-them.html. [Accessed: 26-Dec-2010].

[9] R. J. González, "Hacking the citizenry?," Anthropol. Today, vol. 33, no. 3, pp. 9–12, 2017.

[10] A. Oussous, F. Benjelloun, A. Ait, and S. Belfkih, "Big Data technologies : A survey," J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 4, pp. 431–448, 2018

[11] K. Vidhya and R. Shanmugalakshmi, "Modified adaptive neuro - fuzzy inference system (M - ANFIS) based multi - disease analysis of healthcare Big," J. Supercomput., no. 0123456789, pp. 1-22., 2020.

[12] R. Yang, L. Yu, Y. Zhao, H. Yu, G. Xu, and Y. Wu, "Big data analytics for financial Market volatility forecast based on support vector machine," Int. J. Inf. Manage., vol. 50, pp. 452–462, 2020.

[13] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Management Text mining of industry 4 . 0 job advertisements," Int. J. Inf. Manage., vol. 50, pp. 416–431, 2020.

[14] R. Bao, Z. Chen, and M. S. Obaidat, "Challenges and techniques in Big data security and privacy : A review," Secur. Priv., vol. 1, no. 4, p. e13., 2018.