Volume 1 Issue 1 August -2024 ISSN (online) XXXX-XXXX (Applied For)

A COMPREHENSIVE SURVEY ON BIG DATA ANALYTICS

Vaidehi Patel

Assistant Professor,
Department of Computer Engineering,
LDRP Institute of Technology & Research,
KSV University, Gandhinagar – 382015, Gujarat, India
vaidehi_ce@ldrp.ac.in

Abstract – Data is being generated data speed and scalene even seen before thanks to the amount of information that has come with the modern era. This deluge of information, also known as "Big Data" has completely changed the field of data analytics. Big Data analytics is transforming how businesses and individuals use data for insights and decision-making. It's not just an evolution, but a revolution. In order to provide a comprehensive overview of the big-data analytics field's methods, applications, and challenges, this in-depth survey article delves deeply into the field. This survey is a valuable tool for academics, professionals, and hobbyists interested inbig-data analytics since its main objective is to make clear the basic concepts and procedures that underpin the paradigm. Volume, velocity, and variety—the three primary components of big-data—are highlighted in the survey's first section, which explores the fundamental concepts of thefield. Analyzing the several big-data sources—structured,semi-structured, and unstructured data—highlights theheterogeneous nature of the information ecosystem.

Key Words: Big Data, heterogeneous nature, Data analytics,Revolution

1. INTRODUCTION

Consider a society in which there is absolutely no data storage. Every transaction conducted, every piece of information about a person or business, and every element that may be captured would all be erased right away afteruse. Organizations would be unable to compile crucial information, carry out in-depth analyses, or present novel advantages and opportunities as a result. Information on customer names and addresses, products available, purchases made, recruited workers, etc., is increasingly critical for day-to-day operations. Data is the cornerstone of any successful business. Now think about the amount of information and the explosion of facts that are available now because of the internet and technological advancements. Large volumes of data are now readily available due to advancement sin data collection techniques and storage capacity. Data is being created every second, and in order to extract value, it must be stored and evaluated. Moreover, the cost of storing data has decreased, so businesses must get the maximum benefit from the massive volumes of data they have saved. These kinds of data are so large, diverse, and dynamic that they call for new kinds of big-data analytics in addition to alternative techniques for data storage and processing. Such amount of data needs attention for the concerns such as deliberate analysis and subsequently, related details should be process [1].

2. LITERATURE SURVEY

Authors of [1] proposed paper on Big Data Analytics: a literaturere view paper in which the many big data tools, techniques, and technologies that are applicable were succinctly discussed. Developers can make decisions based on such understanding and in turn they can propose more sophisticated solutions. Users can learn the technology that they need. Hence, decision making is made smooth through big-data analytics.

[2] proposed paper on Big Data: An Overview with Legal Aspects and Prospects provided a concise explanation of the big-data idea, along within formation on its applications, problems ,tools, and methodologies. Large and complicated datasets that are difficult to handle and analyze with conventional technique sere referred to as "big-data. "Big data's four V's volume,variety,velocity,and veracity—highlight its

features and the requirement for specialized methods.

[3] proposed paper on A Comprehensive Survey On Big Data Analytics and Techniques, where in they provided a quick explanation of thelately observed surge in interest stemming from its feasible advantages and special opportunities. The modern world produces a wide range of high-velocity data every day, and within those data are hidden patterns and intricacies that may be discovered and used. Therefore, by using sophisticated analytical tools on large data sets and uncovering important information and hidden patterns, big- data analytics may be used to halt corporate transformation and improve decision-making. There exist several obstacles for upcoming Big Data research.

[4] proposed paper on Big Data Analytics: A Literature Review Paper that provided a quick overview of the cutting- edge subject of big-data, which has at tracted a lot of attention lately because of its apparent unheard-of advantagesand potential. We currently live in an information era when vast variety of high velocity data are produced on a daily basis. These data include hidden patterns and intrinsic nuances that should be discovered and used. The process of decision making can be smoothen for commercial purposes through the usage of cutting-edge analytical approaches and by learning concealed information by big- data analytics.

[5] proposed a review on big-data analytics wherein they define big-data and elaborate various sources of big-data generation. Big data has complexity in terms of volume, diversity, and velocity. In big-data analytics, these three words are more difficult to understand. Our review of the literature demonstrates the industries' exponential grow thin data starting in 2005. Whether data is in text, audio, video, or image form, there may be variances when creating and storing it. Researchers separated the generated data in big-data analytics into different applications, including text, web, multimedia, mobile, structured, and text analytics.

[6] proposed paper on Big Data Analytics: A Literature Review Paper in which they briefly outlined the significance of the big-data analytics principles under investigation for decision-making and provided an analysis of them. As a result, big-data was covered together with its features and significance. Authors studied various tools and technology related big-data analysis along with the aspects of storage and processing.

3. BIG DATA ANALYTICS

3.1 Notion of Big Data

"Big data" refers to large and complex datasets that are challenging to handle, process, and analyses using traditional data processing methods. It encompasses the size of the dataset, its production rate, the variety of data types and sources, and the data's reliability or correctness. Big data's ability to produce insightful information and support informed decision-making when properly analyzed is what distinguishes it most. [2]. Here, we'll go over the four main traits and categories:

3.2 Volume,Variety,Velocity,and Veracity:

Four fundamental characteristics is help to characterize big-data. They cover the primary components of the data challenges resulting from huge and intricate datasets. They are explained below:

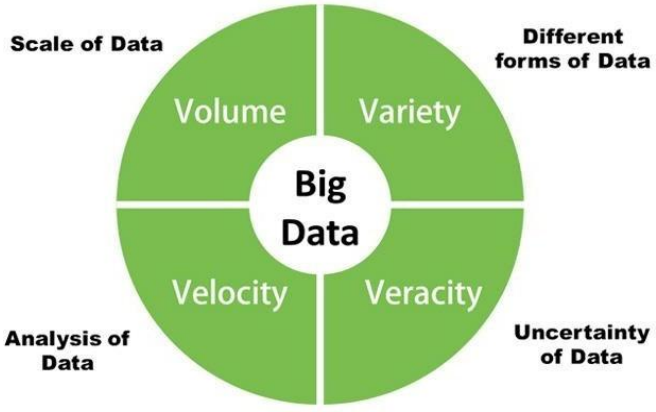

Fig.1.Characteristicsof Big Data

Volume: The term "volume" describes the vast amountof data that is produced and gathered. Conventional data processing and storage methods are frequently unable to handle these enormous data volumes.

Exabytes, zettabytes, and petabytes are standard units of measurement for largedata quantities.

Variety: "Variety" refers to the vast array of data kinds and sources that make up big-data. This includes unstructureddata (such as text documents, social media postings, photos, videos, and sensor data) as well as semi-structured data (such as XML and JSON files) and structured data (like databases and spreadsheets). The challenge is in combining, managing, and analyzing these disparate data kinds to take meaningful notes information.

Velocity: The term "velocity" describes the rate at which data is produced, analyzed, and assessed. The growing volume of real-time data streams from sources like social media, IoT devices, and financial activity has made the capacity to manage and analyses data almost instantaneously essential. With the help of velocity, organizations may react quickly to new possibilities and trends and gain insightful information.

Veracity: The term "veracity" describes the data's reliability, quality, and correctness. Big data is frequently composed of noisy, inconsistent, and incomplete information. Making sure that data is accurate is crucial to preventing poor judgements based on inaccurate or insufficient information. Large data's correctness and dependability are increased bythe application of tactics like validation procedures and assessments of data quality. [2]

3.3 Structured, Semi- structured , Quasi- structured, andUnstructured Data

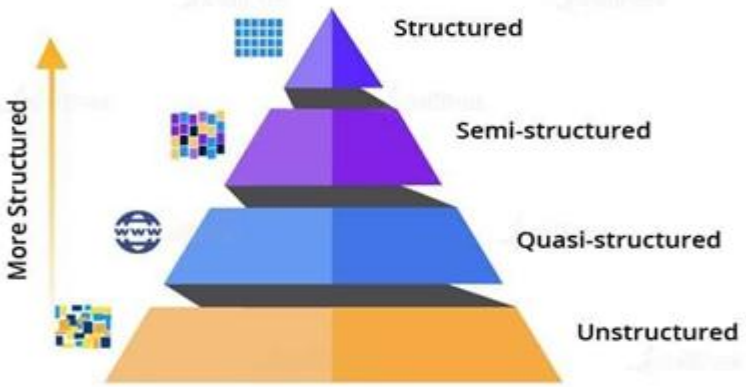

Fig.2.Categories of BigData

Structured Data: Information that follows a sets chemaor format and is well-organized and defined is referred to asstructured data. Conventional relational databases can be used to process and handle this kind ofdata with ease. Data elements like numbers, dates,names,addresses,and category variables are all included instructured data. Financial statements, inventory databases,and customer transaction records are a few examples ofstructureddata.

Semi-structuredData: Even though semi-structured data does not adhere to a strict standard; it is somewhat organized. It has structured tags or information to aid in better organization and search functions. A few instances of semi-structured data include JSON documents, XML files, and log files. This kind of data is frequently used by web apps, social media feeds, and data transfers betweendifferent systems.

Unstructured Data: Information without a predetermined structure that does not cleanly fit into traditional databases is referred to as unstructured data.Included are textual data, multimedia data, emails, posts on social media, sensor data, and more. Becauseunstructureddata lacks intrinsic organization and is largeand complicated, analyzing it can be difficult. However, itcontains insightful information that can be uncovered with the use of sophisticated analytics methods.

Quasi-structured Data: Unstructured textual data with inconsistent formatting makes up the special category known as quasi-structured data. To organize this kind of data properly, a significant amount of work and specialized technologies are required. Web server logs, which are log files created and kept up to date by

servers, are one example. These logs provide information on events and activity on the server, such as timestamps, IP addresses, HTTP requests, and response codes.

To effectively handle and extract value from large datasets, one must grasp the basic principles of volume, variety, velocity, and veracity as well as the differences between organized, semi-structured, and unstructured data. These features influence the potential and difficulties involved in big-data analysis as well as the choice of methods and instruments that are best suited for handling and examiningthese enormous and varied datasets [3].

4. BIG DATA ANALYTICS TOOLS AND METHODS

The need for faster and more effective methods of analyzing data is rising as a result of technological advancements andthe constant inflow of data into organizations. Making fastand successful judgments now requires more than just having a lot of data easily accessible.

For the purpose of effectively analyzing enormous datasets, traditional infrastructures and methods for data administration and analysis are no longer adequate. As a result, new, specialized tools and approaches that are suited to the difficulties of big-data analytics are becoming increasingly necessary. The architectures required fororganizing and storing this data have also grown in importance. As such, the emergence of big-data has significant ramifications that extend beyond the acquisition and manipulation of data to the final choices made us ignite.

In response to these evolving demands, the Big Data, Analytics, and Decisions (B-DAD) paradigm was presented. This framework integrates big-data analytics technologies and methodologies into the decision-making process. It synchronizes several components, including big-data architecture and storage, analytics and data processing tools, and visualization and evaluation tools, with distinct stages of the decision-making process. The architecture and storage of big-data, the processing of data and analytics, and the application of big-data analyses for knowledge discovery and well-informed decision-making are therefore the three primary domains where big-data analytics brings about changes. In this part, each of these sectors will be covered in more detail. Nonetheless, it's crucial to remember that as big- data continues to evolve, new findings, and instruments are consistently produced; hence, this section offers a broad perspective rather than a comprehensive catalogue of all possible options and technology [4].

4.1 BIG DATA STORAGE AND MANAGEMENT

Organizations using big-data have an initial challenge in deciding how and where to store the gathered data. Standard database management technique assists storage and retrieval of data. With the usage of Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) technologies, data is extracted from external sources, changed to meet operational requirements, and then loaded into the data repository. This gives surety about cleaned, transformed, and organized datafor the purpose of data mining and analysis.

The idea of big-data provides a new set of requirements known as Magnetic, Agile, Deep (MAD) analytical capabilities, which are different from standard Enterprise Data Warehouse (EDW) installations. Contrasting to EDW which often blocks the incorporation of new database till the time it is cleaned and unified, the big-data has to be compelling, and it needs to pull all the data from various sources irrespective of data quality. Given the variety of data sources and the complexity of data analysis, big-data storage must facilitate analysts' ability to produce and modify data quickly. Because of this, using an elastic database that can swiftly adapt to changes in the logical and physical data is required. A big-data store house also has to be profound and it should function as a cutting-edge computational device since analysts need to be able to work with huge information and use intricate statistical methods to analyses data.

Consequently, many tactics have been developed to deal with these problems. Databases and distributed systems using massively parallel processing (MPP) offer highly improved performance and scalability. Unstructured or non- relational data may also be managed with product such as NoSQL. Data administration and data storage are treated separately in NoSQL. It emphasizes large-scale, flexible data models, as well as easier application development and deployment.

On the other hand, disc input/output (I/O) is not necessary with in-memory databases as they store data in server memory and allow for immediate database answers. Deploying the major database into silicon-based

main memory, as opposed to mechanical disc drives, enables substantial speed gains and opens up new application development opportunities. Scalability and performance encourage the usage of in-memory databases to achieve data-analytics for huge datasets.

The scalable, robust, and controllable Hadoop architecture offers an alternative huge data analytics platform. It puts into practice the MapReduce paradigm, which combines analytics and storage. MapReduce handles massive data analytics, whereas Hadoop Distributed File System (HDFS) manages big-data storage. HDFS divides data into blocks and distributes them among cluster nodes to offer redundant, reliable distributed file storage that is appropriate for big files. By securing data between nodes, replication guarantees data availability and dependability even in the case of node failures. HDFS nodes come in two different varieties: (a) Name Nodes and (b) Data Nodes. Data Nodes store data in duplicated file blocks across several nodes, while Name Nodes serve as a bridge between clients and Data Nodes, guiding clients towards Data Node with requested results.[6].

4.2 BIG DATA ANALYTIC PROCESSING

Analytical processing is the next stage after the initial data storage phase. Four essential requirements are listed in the reference for handling vast amounts of data. The first prerequisite is the quick loading of the data. It is critical to reduce data loading time to minimize interference from disc and network traffic. Fast query processing is the second prerequisite. Because of the needs of huge workloads and urgent requests, many inquiries have a time constraint.

As a result, the data storage structure must continue processing requests swiftly as the number of queries climbs quickly. Moreover, the third requirement for handling large amounts of data is the incredibly efficient use of storage capacity. Scalable computer and storage resources may be required in response to rapid increases in user activity, which highlights the necessity of effective data storage management and space-saving strategies. The fourth criterion emphasizes adaptability in the face of highly varied work patterns. Large data is analyzed by many users and applications for various purposes, thus the underlying system must be extremely robust towards data processing and workloads.

This parallel programming technique was inspired by the "Map" and "Reduce" functions present in functional languages. It works well when managing massive amounts of data. It manages analytics and data processing duties and is the central component of Hadoop. EMC claims that the Map Reduce paradigm focuses more on expanding the number of computers or resources than it does on increasing the computational/storage ability of a single system. In essence, it encourages scaling downward rather than upward. The fundamental principle of Map Reduce is breaking down a task into smaller steps and completing each step simultaneously in order to reduce the total time required to complete the task.

A Map Reduce process starts with mapping input values to a set of output key/value pairs. The "Map" function assigns the proper key/value pairs to each chunk after breaking complicated computations up into digestible pieces. Unstructured data, like text, is converted into structured key/value pairs using this procedure. A word in the text may, for instance, be represented by a key, and the value would indicate how frequently the phrase occurs. After utilizing this output as the input for the "Reduce" function, which combines and averages the output by merging values with the same key, the final result of the computing job is then achieved.

Hadoop's Map Reduce function requires two different types of nodes: (a) Task Tracker and (b) Job Tracker nodes. Job Tracker nodes monitor the results, bifurcate the mapper and reducer among Task Trackers. A section of the input file is received from the Job Tracker by the Hadoop Distributed File System (HDFS) map work in progress on a node, which initiates the Map Reduce job. Conversely, the Task Tracker nodes are responsible for completing the tasks and informing the Job Tracker with the results. Typically, files and folders in HDFS are used for inter-node communication to reduce direct connections between nodes [5].

5. CONCLUSION

The present study delves into the nascent domain of big- data, which has attracted substantial interest owing to its seeming abundance of prospects and benefits. Massive amounts of high-velocity data are produced every day in the current era of information overload, hiding important in sights and patterns that must be discovered and used. Asa result, using big-data analytics becomes a potent instrument for accelerating corporate change and

enhancing judgment. We can reveal knowledge that has been hidden and insights that have been overlooked by using sophisticated analytical methods on massive datasets.We did athorough literature analysis as part of this investigation to explore the ideas behind big-data analyticsand its importance when making decisions. This included talking about the qualities and significance of big- data. Furthermore,we examined certain tools and techniques used in big-data analytics, delving into topics like dataprocessing, administration, and storage. Additionally,we looked at a variety of use full advanced data analytics methods.

By utilizing these analytics on large data sets, we are able to extract important information that may be used to improve decision-making and enable well-informed choices. We then looked into many areas where big-data analytics may help and facilitate the decision-making process. It became clear that supply chain management, fraud detection, consumer intelligence, and other various applications and sectors may benefit greatly from big-data analytics. Numerous industries,including healthcare, retail, communications, manufacturing,and more, can profit from it.

6. REFERENCES

1. *Prof. (Dr.) T. MUTHUKUMAR Big Data Analytics: A Literature Review Paper , International Journal of Scientific Development and Research (IJSDR) ISSN: 2455-2631 April 2023 IJSDR, Volume 8 Issue 4.*

2. *Mohammad Nazmul Alam, Vakil Singh,Ms. Ripendeep Kaur, Md. Shahin Kabir on Big Data: An Overview with Legal Aspects and Future Prospects, Journal of Emerging Technologies and Innovative Research (JETIR) © 2023JETIR May 2023, Volume 10, Issue 5 www.jetir.org (ISSN-2349-5162.*

3. *K.Naresh Babu1, Dr.Suneetha Manne A ComprehensiveSurvey On Big Data Analytics And Techniques , Vol. 14 ICETCSE 2016 Special Issue International Journal of Computer Science and Information Security (IJCSIS) ISSN1947-5500 [https://sites.google.com/site/ijcsis/]*.

4. *Nada Elgendy ,Ahmed Elragal Manne , "Big Data Analytics: A Literature Review Paper P. Perner (Ed.)", ICDM 2014, LNAI 8557, pp. 214–227, 2014. © Springer International Publishing Switzerland 2014.*

5. *Ankita S. Tiwarkhede , Prof. Vinit Kakde, " A Review Paper on Big Data Analytics", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14, Impact Factor (2013): 4.438 Volume 4 Issue 4, April 2015*.

6. *Abdian, S., Hosseinzadeh, M., 2 , S., & Khadivar, A. (n.d.). " A Bibliometric Analysis of Research on Big Data and Its Potential to Value Creation and Capture", In Iranian Journal of Management Studies (IJMS) (Vol. 2023, Issue1).*

7. *Munawar, H. S., Ullah, F., Qayyum, S., & Shahzad, D. (2022), " Big Data in Construction: Current Applications andFuture Opportunities", In Big Data and CognitiveComputing(Vol.6,Issue1).MDPI.*

8. *Ananda Kumar, K. S., Muleta Hababa, S., Worku, B., Tadele, G., Gebru Mengistu, Y., & Y, P. A. (2021). "Big Data Characteristics, Classification And Challenges-A Review", In Turkish Journal of Computer and Mathematics Education (Vol. 12, Issue 12).*

9. *Ajah, I. A., & Nweke, H. F.(2019), "Big data and businessanalytics: Trends, platforms, success factors and applications", In Big Data and Cognitive Computing (Vol 3, Issue2, pp.130).*

10. *Batko, K., & Ślęzak, A. (2022), "The use of Big Data Analytics in healthcare", Journal of Big Data, 9(1).*

11. *Cravero, A., Pardo, S., Sepúlveda, S., & Muñoz, L. (2022), "Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review", In Agronomy(Vol.12,Issue3).MDPI.*

12. *Anwar, M. J., Gill, A. Q., Hussain, F. K., & Imran, M. (2021), " Secure big data ecosystem architecture: challenges and solutions", In Eurasip Journal on Wireless Communications and Networking (Vol. 2021, Issue 1). Springer Science and Business Media Deutschland GmbH.*

13. *Hydén, H. (2020), " AI, norms, big data, and the law", Asian Journal of Law and Society, 7(3), 409-436.*

14. *Assistant Professor, A. B., & Babu, D. (n.d.). "Big Data Technologies For E-Business-Future Opportunities, Challenges Ahead And Growing Trends", International Journal Of Advanced Research In Computer Science, 9(2).*

15. *Simranjot Kaur, Er. Sikander Singh Cheema, "A review paper on big data tools and techniques," IJARCCE ISSN (Online) 2278-1021 ISSN (Print) 2319 5940 International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 6, June 2017*.

16. *Lundberg, L., & Grahn, H. (2022), " Research Trends, Enabling Technologies and Application Areas for Big Data.Algorithms,15(8).https://doi.org/10.3390/a15080280*.

17. *Alam, M. N., Kaur, K., Kabir, M. S., Susmi, N. H., & Hussain, S. (2023), "Uncovering Consumer Sentiments And Dining Preferences: A Legal And Ethical Consideration To Machine Learning-Based Sentiment Analysis Of Online Restaurant Reviews" In International Journal of Creative Research Thoughts (Vol. 11, Issue 5).*

18. *Avci, C., Tekinerdogan, B., & Athanasiadis, I. N. (2020), " Software architectures for big data: a systematic literature review", Big Data Analytics, 5(1).* https://doi.org/10.1186/s41044-020-00045-1.

19. *Basu, S. (n.d.). "Cloud Computing and Big Data for Genomics: A Review", International Journal of Advanced Research in Computer Science, 8(3).*

20. *Ahmad, F., & Tripathi, M. M. (2018), "Approaches Of Big Data In Healthcare: A Critical Review", International Journal of Advanced Research in Computer Science, 9(2).*

21. *García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016), " Big data preprocessing: methods and prospects", Big Data Analytics, 1(1).*

22. *Singh, B., & Kaur, M. (2016) " A Survey on Big Data: Challenges, Tools and Technique", In International Journal of Advanced Research in Computer Science (Vol. 7, Issue 6).*