A Review on Big Data Analytics

Dr Sonia Riyat, Assistant Professor

Department of Commerce & Management, Arka Jain University, Jamshedpur, Jharkhand, India

ABSTRACT: ***As more data is produced, conventional architectures and infrastructures are finding it more difficult to handle huge quantities of data in a timely and resource-efficient manner. Organizations must develop new tools and techniques specialized for big data processing in order to extract value from this data effectively. As a result, big data analytics has emerged as a critical component for businesses seeking to uncover hidden information and gain a competitive edge in the marketplace. Currently, the large number of papers on big data analytics makes it challenging for practitioners and academics to discover and keep up with subjects of interest. The purpose of this article is to provide an overview of the content, breadth, and location of big data analytics.***

KEYWORDS: Application, Big Data Analytics, Programming Model, Storage.

1. INTRODUCTION

Data is produced not just by businesses and governments in the information age, but also by each person in society. End consumers now have easy access to massive quantities of data. According to an International Data Corporation (IDC) study, worldwide data volume would increase from 130 exabytes to 40,000 exabytes between 2005 and 2016. While there is a growing interest in big data, there is no universal definition of the term. Big data is defined as a collection of huge data sets that cannot be handled using conventional database management software. Big data is defined as "datasets that are too large for traditional database software tools to collect, store, manage, and analyze." These definitions illustrate the difficulties that individuals encounter when dealing with large amounts of data. Some academics, on the other hand, describe big data in terms of its characteristics. Big data is defined as “a new generation of technologies and architectures that enable high-velocity collection, discovery, and/or analysis of extremely large quantities of a broad range of data in order to economically extract value from them”. Volume, diversity, velocity, and value are four important characteristics of big data, according to the authors. Their "4Vs" have been generally recognized as essential characteristics of big data [1][2].

For example, Forfás, Ireland, and the Expert Group on Future Skills Needs describe big data using comparable characteristics. Volume refers to the amount of the data, velocity to the pace at which the data changes, and diversity to the many forms of data as well as the various applications and methods of evaluating the data, according to their definition. The distinctions between conventional data and big data the main characteristics that may distinguish between big data and conventional data are volume, produced rate, structure, data source, data integration, data storage, and access, according to them. As more data is produced, conventional architectures and infrastructures are finding it more difficult to handle huge quantities of data in a reasonable amount of time and resources. The majority of the data is now unstructured, making conventional data-processing technologies ineffective.

Organizations must develop new tools and techniques specialized for big data processing in order to extract value from this data effectively. As a result, big data analytics has emerged as a critical component for businesses seeking to uncover hidden information and gain a competitive edge in the marketplace. Several methods have recently been suggested, and several results have been published in journals and conferences. However, the massive number of papers makes it difficult for practitioners and researchers to identify subjects of interest and keep up with current results and research trends. The goal of this study is to provide an overview of big data analytics' content, breadth, and results, as well as to analyze its future evolution [3][4].

To do this, we'll perform a systematic literature review to identify hotspots in big data analytics and develop a taxonomy to aid in-depth analysis in future research. The rest of the article is organized as follows: first, we'll go through the research technique that was used in this study. We analyze the results and the existing research needs after presenting the features of chosen studies. Finally, we describe the findings of the study and suggest future research [5][6].

1.1 Acquiring and storing large amounts of data:

Big data analytics, in contrast to conventional methods, deals with massive quantities of diverse and unstructured data. Traditional data methods and infrastructures cannot readily gather, integrate, or store this data. Many studies on large data collection and storage have been inspired by these difficulties [7].

1.2 Acquiring large amounts of data:

Data acquisition is the process of compiling data into a well-organized digital format for storage and analysis. Data acquisition is made up of three steps: data collecting, data transmission, and data pre-processing (X. Chen & Lin, 2014). Sensor-based data gathering, online data mining, and other data collection techniques are only a few examples. Sensor-based data gathering has become one of the most common data collecting methods, thanks to the rapid development of sensor technologies such as the Internet of Things and Radio-Frequency Identification (RFID) technology. To gather and transmit data, several data applications and research have used sensor-based data collecting methods [3]. However, in the age of big data, the high costs of initial installation and maintenance continue to limit the use of sensor-based data gathering. Some academics and businesses recommended crowd-sourced data gathering as an alternative to sensor-based data collection to solve these issues. The inclusion of crowd workers in the data collecting process allows for the reduction of noisy data as well as the gathering of new data kinds. Users may, for example, define the geographical position of any particular street on a map and report issues connected with the area using the Fix My Street platform [8].

The Ushahidi Platform allows for real-time data gathering via various channels such as email, social media, and so on. By rewarding mobile crowd workers, the EcoTop system lowers noise levels during data gathering. Overall, these methods encourage crowd workers to collaborate with one another to improve data availability and quality. Big data will be sent to the data center after collection for cleaning, processing, and integration. Many possible difficulties arise from the transfer of large quantities of data, including input/output (I/O) bottlenecks, network traffic delays, and data duplication. Researchers are using a variety of methods to enhance the efficiency of large data transmission in order to overcome these difficulties. Big data transfer across federated clouds is a problem. They suggest a Maximally Overlapped Bin-packing driven Bursting (MOBB) algorithm that wants to consider the time interact across computing nodes, data transfer delay, and calculation time in each computing node to address challenges such as node perseverance, synchronized completion, and data partition determination [9][10].

In comparison to current methods, they claim that their algorithm increases performance by up to 60%. Adopt Phase Change Memory (PCM) as a large data storage medium to try to address the I/O bottleneck issue. They propose a Content-Aware PCM that takes advantage of content locality in memory accesses by using a lightweight data duplication module. Their method effectively overcomes PCM's main flaws and substantially increases data transmission performance and reliability. They find that data replication is particularly costly for I/O-intensive large data tasks, and poor data replication may result in significant performance deterioration. Their solution can store task outputs across jobs and use them to reduce the work done during job re-computations by offering the option to switch to finer grained task scheduling granularity for recomputations. In large-scale data transmission, Representational State Transfer (REST)-ful systems may suffer latency and transmission bottlenecks. Their approach package includes the sender REST services' Transmission Control Protocol (TCP) connections into the payload of the UDT protocol, and then needs to send it to the recipient, by trying to combine the super-fast transition properties of User Datagram Protocol (UDP)-based Data Transport (UDT) with the encapsulation features of proxy. RESTful services may now bypass the bottlenecks of conventional techniques thanks to this technology. As a result, substantial savings in RESTful data transmission times will be shown. Integration of heterogeneous unstructured data gathered from many sources is another significant difficulty of big data analytics [11][12].

Three essential layers of data integration were identified: data accessibility, shared data platform, and consolidated data model. Many academics have suggested these levels as the foundation for various methods and platforms. Hadoop-GIS is a high-performance, scalable spatial query engine for geographical data integration. Hadoop-GIS offers an efficient spatial query engine for processing spatial queries, data- and space-based partitioning, and query pipelines that parallelize searches implicitly on MapReduce as a system built on MapReduce. The integrated geographic data warehouse can handle descriptive queries, spatial

relationship-based queries, distance-based queries, and spatial mining and statistics, among other things. They enhance Hadoop-GIS with spatial grouping and spatial regression capabilities based on this concept. The experiment demonstrates that their technology is a scalable and effective solution for spatial data integration across huge geographical data sets. Hadoop, on the other hand, is a distributed file system intended to operate on commodity hardware and is better suited to bulk data processing than interactive applications. It is inappropriate for online big data integration based on events. In order to address this issue, we created a data-integration platform that allows event processing systems to share data. Their solution shows how heterogeneity in publish or subscribe services may be handled by embedding a schema-matching mechanism inside the notification service by using automated schema mapping, Resource Description Framework (RDF) ontology, and semantic inference techniques [13][14][15].

Random access memory, magnetic disks and disk arrays, and storage class memory are all common hardware infrastructure technologies. Because each technology has its own set of performance characteristics, figuring out how to combine such characteristics to accommodate information permanently and reliably has become a major challenge in the development of large data storage systems. The transfer of huge amounts of data from hard disks to memory often reduces the performance of big data analytics. An architecture that can achieve high throughput by wanting to share large numbers of flash chips along a reduced, chip-to-chip backplane network managed by the flash controllers can reach high throughput by sharing huge numbers of memory modules across a low-latency, chip-to-chip backplane enhancing by the flash controllers.

The investigation found that the average delay for user software accessing the flash storage is less than 70 seconds, which is much faster than the conventional method. Storage infrastructure may be divided into three types based on networking architecture: direct-attached storage (DAS), network-attached storage (NAS), and storage area network (SAN). Due to the unique features of big data, these designs do not function well on big data analytics systems. The advancement of storage virtualization, on the other hand, makes it feasible to meet the demands of big data analytics. Storage virtualization is the process of combining many network storage devices into a single storage unit. Data may be readily found and connected with this technology from a single source. As a result, independent of the physical infrastructure, data may be transmitted in a uniform format. It will decrease storage costs and make it simpler to maintain data repositories needed for big data analysis from this point forward [2].

1.3 Data storage administration:

Infrastructure provides the foundation for the second component, data storage management. File systems, database technologies, and programming models are the three major layers (Hu et al., 2014). Other layers are built on top of file systems. Many companies and academics have used it as the foundation for their large data storage systems. The Google File System (GFS), for example, was created for big distributed data-intensive applications. It offers fault tolerance and excellent performance to consumers by using low-cost commodity hardware. GFS, on the other hand, does not function well with tiny files. Other systems, such as Hadoop Distributed File System and VERITAS Cluster File System, were created to address the limitations of GFS in order to meet the needs of large data storage. Traditional relational database systems face significant difficulties as a result of the diversity and size of big data. The None Structured Query Language (NoSQL) is a new kind of database that is built without utilizing tabular relationships (Cattell, 2011). Partition tolerance, high availability, and the capacity to handle large amounts of data make the NoSQL database a suitable choice for customers dealing with big data problems. Key-value stores, column-oriented databases, and document-based stores are the three major kinds of NoSQL databases. SimpleDB, Cassandra, HBase, Big, and MongoDB are examples of prominent NoSQL databases that fall into one of these categories. Many academics have suggested NoSQL databases as the foundation for their own large data database systems.

1.4 Programming model for big data:

Large data processing is the next step following big data storage. Rapid data loading, fast query processing, extremely effective storage capacity usage, and great adaptivity to changing workload patterns are four essential criteria of big data processing. Many programming models have been developed and implemented in response to these needs. The programming paradigm is intended to translate programs into a parallel environment. Big data demands scalability, flexibility, and fault tolerance, which traditional parallel architectures such as Open Multi-Processing and Message Passing Interface (MPI) lack. These difficulties

prompted the creation of a slew of innovative designs, including the map/reduce Simple Scalable Streaming System and Apache Storm (Apache Software Foundation, 2013). The majority of these designs fit into one of three categories: MapReduce and MapReduce-related models, graph processing models, and stream processing models.

1.5 Social-networking applications:

Aside from commercial applications, big data analytics has the potential to be useful in social domains. In general, big data analytics may help decision-makers by facilitating information exchange across different organizations in society, detecting connections between social events, and aggregating and analyzing data in real time. Many methods and frameworks for solving social issues have been suggested by academics recently. In the context of smart cities, city administrators may profit from big data analytics in terms of producing information intelligence and aiding decision-making. Big data may be used to reach out to people and evaluate their behavior. City administrators may use human behavior modeling to understand and forecast specific social phenomena like traffic distribution, civic unrest, and disease outbreaks. In the field of education, big data analytics can help instructors integrate and distribute online material more effectively.

1.6 Applications in science:

Big data sets are extensively used in scientific study. The growing quantity of data, on the other hand, has presented major difficulties to scientific study. Big data analytics gives scientists the ability to rapidly access huge quantities of data, make data gathering and sharing easier, and discover previously unseen patterns in large data sets. Big data analytics is now being used in a variety of academic fields. Big data analytics has become a strong tool in ecology disciplines for monitoring and evaluating ecosystem growth, life interactions and adaptations, and biodiversity in the environment. When using hyper spectral imaging in geosciences, big data analytics has been utilized to group spectral bands. Big data analytics has proved to be an effective technique for dealing with multi-wavelength, multi-messenger, and massive quantities of astronomical data in astronomy fields.

2. DISCUSSION

Big data analytics must discover ways to adapt to distribution applications' real-time needs. Currently, centralized computing is extensively utilized in big data analytics, which requires a significant amount of extra resources for data transmission and integration. At this moment, distributed big data computing may be a viable option for meeting real-time needs. The programming model is a useful tool for putting application logic into action. Current programming paradigms, on the other hand, are limited to particular tasks. In the future, a single programming paradigm may not be adequate to handle the difficulties of diverse, dispersed data sources. Future big data analytics will need a closer integration of various programming paradigms. The Apache YARN project is a leading example of this kind of integration. YARN is a cluster management solution that allows different programming paradigms to operate on the same cluster. This feature makes it simpler for consumers to benefit from low-cost, linear-scale storage and processing. Big data benchmark research is still in its early stages. Future benchmarks will have a significant problem in determining the most efficient data collection to reflect the variety and correlation of big data applications. More work is needed to make comparing the performance of different frameworks easier. Traditional security methods were created to address the security issues that arise when dealing with tiny amounts of data. Some large data security problems, such as safe computations in distributed programming frameworks and real-time security monitoring, have received little attention in the literature. In the big data age, future researchers will need to identify security and privacy issues and develop ways to strike a balance among data security and data processing efficiency.

3. CONCLUSION

This article covers different big data tools, techniques, and technologies in this academic field by analyzing 266 publications relevant to big data analytics. This article may assist academics in staying current with research results in big data analytics and identifying unanswered issues. This article will help practitioners use big data analytics successfully in the real world. One of the goals of this research is to classify the literature on big data analytics into research areas. Five research categories were determined after reading the chosen articles. Several data collecting, data transmission, and data integration methods, as well as data storage

technologies, were developed under the area of big data acquisition and storage. We examined the characteristics and evolution of three modeling techniques in the data programming model area. Three analytics – descriptive analytics, predictive analytics, and prescriptive analytics – were contrasted and explored in the data analysis area. We looked at two kinds of benchmarks in the data system benchmark category: component benchmarks and system benchmarks. Finally, the application category of big data analytics was addressed. Data volume, velocity, and diversity are all growing at a rapid pace right now. Big data analytics is complicated by the diverse, dispersed data.

REFERENCES

[1] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J. Big Data*, 2015, doi: 10.1186/s40537-015-0030-3.

[2] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electron. Mark.*, 2016, doi: 10.1007/s12525-016-0219-0.

[3] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *J. Big Data*, 2015, doi: 10.1186/s40537-014-0008-6.

[4] D. Chong and H. Shi, "Big data analytics: a literature review," *J. Manag. Anal.*, 2015, doi: 10.1080/23270012.2015.1082449.

[5] O. Kwon, N. Lee, and B. Shin, "Data quality management, data usage experience and acquisition intention of big data analytics," *Int. J. Inf. Manage.*, 2014, doi: 10.1016/j.ijinfomgt.2014.02.002.

[6] L. Duan and Y. Xiong, "Big data analytics and business analytics," *J. Manag. Anal.*, 2015, doi: 10.1080/23270012.2015.1020891.

[7] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *J. Parallel Distrib. Comput.*, 2014, doi: 10.1016/j.jpdc.2014.01.003.

[8] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, 2015, doi: 10.1186/s40537-014-0007-7.

[9] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, 2014, doi: 10.1109/ACCESS.2014.2332453.

[10] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *Biomed Res. Int.*, 2015, doi: 10.1155/2015/370194.

[11] N. Elgendy and A. Elragal, "Big Data Analytics in Support of the Decision Making Process," 2016. doi: 10.1016/j.procs.2016.09.251.

[12] Z. Xu, G. L. Frankwick, and E. Ramirez, "Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective," *J. Bus. Res.*, 2016, doi: 10.1016/j.jbusres.2015.10.017.

[13] E. G. Ularu, F. C. Puican, A. Apostu, and M. Velicanu, "Perspectives on Big Data and Big Data Analytics," *Database Syst. J.*, 2012.

[14] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[15] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and Systems*. 2014. doi: 10.1186/2047-2501-2-3.