Big Data Analytics : Existing Systems and Future Challenges – A Review

Poorvi Seth[1], Merin Meleet[2]

[1,2] *Department of Information Science and Engineering, R V College of Engineering, Bangalore, India*

Abstract – *In this data driven era with large amounts of information being generated, it has become quite difficult to handle this data and information through traditional methods using an acceptable number of resources. An efficient technology is required to extract value from these various sources of data, and this is where big data analytics comes into picture. Big data analytics makes use of advance analytical methods to handle data of all types – raw and unstructured, semi-structured or even structured data. The aim of this paper is to highlight the role of big data analytics in handling data efficiently in various fields, understand its advantages as well as the challenges it brings.*

Key Words: Big data, Big data analytics, Data handling, Data Lakes, Data Marts.

1.INTRODUCTION

Data is generated from almost every possible source available today, from industries and organizations to data from the digital world, that is, the internet and social media. The world now revolves around data, it has become the building block for any process. The findings in [1] suggest that the International Data Corporation (IDC) report predicts that the global data volume will grow 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025. This enormous quantity of data that is ever growing, is termed as big data. Lately, big data is also being used to refer to the datasets used to carry out research work as these datasets keep increasing constantly and special tools and technologies are required to handle and further analyze them.

Such vast amount of data comes with various characteristics which are broadly classified into five categories, known as the 5 Vs of big data. These are namely – volume, velocity, variety, veracity and value. **Volume** refers to the size and quantity of data present. If the volume of the data is large enough, it is considered as big data. As mentioned previously, data present currently is "enormous" which refers to the volume of data. Next comes Velocity which essentially determines the rate at which data gets generated and further grows. This characteristic of big data is very essential for organizations as they need the data flow to be constant and quick to help make the right decisions. While most of the data generated is useful, but not all of it is effective for an organization's decisions and growth. This is where Veracity comes into picture. It deals with the quality and correctness of the data being generated. Variety deals with the different types of data present. As mentioned previously, big data deals with all types of data which includes unstructured data that is the raw and unformatted data, which is generated from various sources, semi-structured data is a type that as some information attached to it that makes it more usable than unstructured data and lastly is structured data which is the most ideal for of data for effective data analysis. Data lakes constitute of unstructured data while data marts constitute structured data which is processed, filtered and formatted data from the data lakes. More about data lakes and data marts will be discussed in the later sections of this paper. Lastly, Value refers to the benefits that can be gained by the usage of data. With such enormous volume of data being generated every day, big data can play a big role in adding value to an organization or individual's performance.

Having discussed the characteristics of big data, it is now clear that such large volumes of data require special methodologies to help analyze it so that it can help add value to the organizations and industries. This is what big data analytics deals with – to handle and process big data in the right way so that the specific use case gains maximum value from it. For example, Amazon has gained massive value by analyzing the large data of its customers and their purchases and then building a recommendation system that allows the customers to easily view more products of their choice, which increases the chances of the customer buying products from their site, which in turn increases the profits and value of the organization [2]. However, this is not an easy task - larger the volume of data, tougher it gets to analyze it.

This paper first discusses the various existing systems that make use of big data analytics to increase data handling efficiency. The next section highlights the advantages of the big data analytics tools and methods. Lastly, the paper helps understand the challenges this methodology can bring with it.

2. EXISTING SYSTEMS

As discussed in the previous section, the right analysis of big data can add tremendous value to any organization or industry. This benefit of big data analysis has already been put into use in to increase the value and efficiency of various industries. In this section, a few such existing applications of big data analysis will be discussed.

2.1 Big data analytics in health care

Healthcare is one of the most essential industries in the world. As seen in the pandemic, the health care industry

was of prime importance, without which it was impossible to survive. Every day, the healthcare sector maintains a large quantity of data and the big data analysis helps this industry by providing tools and technologies to analyze the large datasets of all the patients which in turn can also help health care researchers understand the diseases and its symptoms with more accuracy.

The authors of [3] provided a research model that helps corelate how big data analytics can help fight against pandemics like COVID-19 and other epidemics as well because these diseases are highly contagious and infectious so appropriate and efficient tools and technologies are required to study the effects of these diseases which can help arrive at a solution for prevention and even cure. This model was based on researching and reviewing multiple data sources that have proved to help fight against serious diseases. The main intention of this work was to show that when there is such a large volume of data available, it should be analyzed and used in the right ways as it can contribute towards the health and well-being of the society. As a result of this research, it was also found out that majority of health care related data is generated from social media and search engines which make the data enormous and unstructured thus making it difficult to handle which is where big data analytics helps by providing methods for processing this large unstructured data and adding value to it in a way that benefits the well-being of the people.

2.2 Big data analytics meets social media

Now is the era of social media. Everything happens on social media because of which enormous volumes of data is being generated every minute. Of course, not all this data would be useful but majority of it can be of great value if analyzed in the right ways.

In [4] the authors aimed on providing intensive research on the existing work that has been carried out in social big data analytics. As shown in Fig -1, two main categories of social

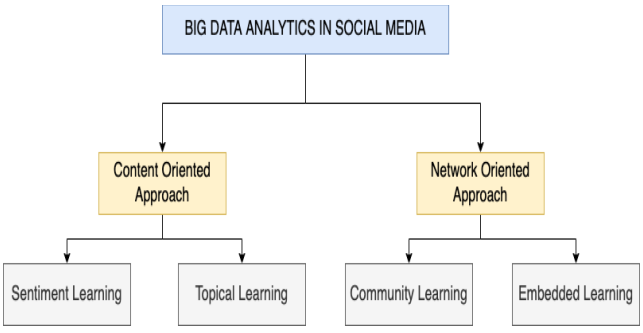

Fig -1: Social big data analytics [4]

big data analysis approaches were identified as: network - oriented and content – oriented approaches. The network - oriented approach took into consideration the relation between the users present on social media to carry out the analysis. Groups of social media users and their relationships as well as their interactions with users outside their groups are also considered. On the other hand, the content - oriented approach deals with the posts that are generated as a result of user activity, that is, the content that users create on social media. A large volume of data is generated from such content and big data analysis tools and strategies are used to analyze this content – generated data to help gain maximum value out of it. This can be seen in case of the content creators who are more famously known as influencers these days. Social media sites make use of big data analytics by providing its users with insights about the content they post. These insights are nothing but well-analyzed and structured data that help the content creators understand the number of people engaging with their content which would help them create better content. Also based on these insights, brands collaborate with these creators which adds value both to the brand and the creator. Thus, this shows how just by analysis data in the right way, individuals and businesses can have a great potential to grow. This aligns with the result of the research conducted by the authors of [4] who concluded that almost 51% of profits from social big data analysis come from content-based learning.

2.3 Big data analytics in intelligent transportation systems

The transportation system is another essential system globally as it is one of the most frequently used systems across the world. But the question here is, how can a transportation system generate data? The answer to this is the implementation of intelligent transportation systems which consist of devices like GPS trackers, video cameras, RFID tags, Vehicle to Infrastructure (V2I) sensors as well as Vehicle to Vehicle (V2V) sensors and many more devices. These devices help in the generation and collection of data. As mentioned previously, transportation is one of the most frequently used systems, hence huge volumes of data are generated everyday via the intelligent devices that are now installed in many parts of the world.

A very elaborate and easy to implement process of making use of big data analytics to make use of the data generated by intelligent transportation devices is explained by the authors of [5]. It is a three-phase process as illustrated in Fig -2. The first and most important phase is the collection of data for the analysis. As mentioned previously, data is generated from various devices like video cameras that record and capture details like the vehicle number and speed with which its moving, then there are RFID tags which have recently been introduced in India as FASTags which helps in making wireless toll payments also captures details of the vehicle and its user while registration. Once the data is collected, big data analytics techniques are used to store data using big data analytics technologies like Hadoop, Apache Spark, Kafka, etc., analyze it and further help with the data management. Lastly, big data analytics helps in adding value to the large volumes of data that is collected. With fast collection and analysis of road data from video cameras, the data can be analyzed to help control and manage the traffic in that

particular area hence aiming at reducing traffic congestion. Another very essential application could be to analyze the real – time

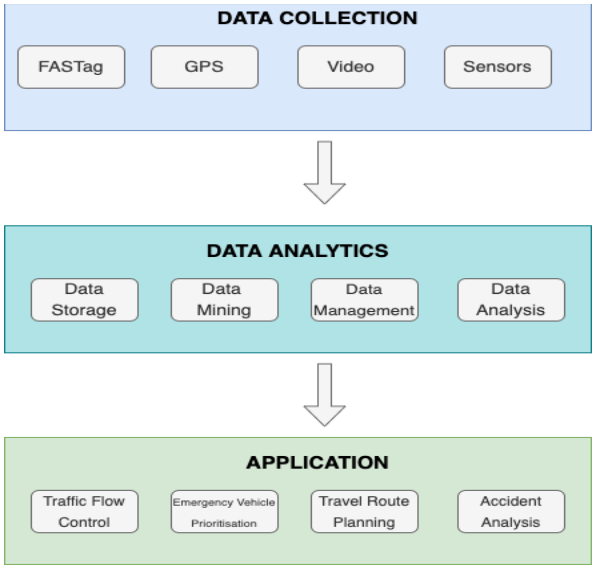

Fig -2: Process of big data analysis in transportation systems [5]

traffic data via V2I or V2V sensors which would help in prioritization of emergency vehicles like ambulance, fire trucks, etc. This collected data could help provide the vehicle drive the best suitable route so that traffic can be avoided. So, as explained above, the incorporation of big data analysis with intelligent transport devices can add big value to the global transportation system.

2.4 Weather forecasting and big data analytics

Weather forecasting is a concept which is an essential component of people's daily lives. The daily activities are planned according to the weather forecast of the day. As it is a daily process, it implies that data also gets generated daily which makes it fall under the category of big data and hence special techniques are required to analysis this big data. It may not seem like an essential system, but various research suggests that weather forecasting as an impact on many industries like tourism, transportation, agriculture, sports, construction, etc. Weather forecasting has an impact on these above-mentioned industries as these industries heavily rely on the weather to carry out their day-today proceedings. For example, weather plays a crucial role for the right growth of agriculture, tourism industry gets impacted as people plan their travels based on the weather and even the sports industry as the sports analysts studies the weather conditions before every game to understand how the pitch/field will play out during the game.

As understood from the above discussion, large amount of weather-related data is generated daily from different industries. It becomes important to formulate an approach to analyze this big data efficiently so that these industries can gain value from the big data generated. Three such approaches are proposed in [6] as shown in Fig -3. The technique – based approach deals with improving the quality of the weather predictions by using analysis techniques like machine learning. As a result, high accuracy is obtained but the time taken to achieve these results is considerably high. On the other hand, the technology – based analysis approach tries to overcome the drawback of the previous approach by speeding up the data analysis process to further enhance the forecasting process. As a result of this approach, it is found out that it provides quick results in term of processing time taken but also has a disadvantage of lesser reliability. As both the above – mentioned approaches have some drawbacks, a third approach is suggested which is a combination of both the approaches and various big data analysis techniques and hence it's called the hybrid approach. So as a result, this approach speeds up the big weather data analysis process as well as provides high quality and accurate results. But in application, currently majority of the industries follow the technique – based approach as it focuses on providing accurate results which is of at most importance.

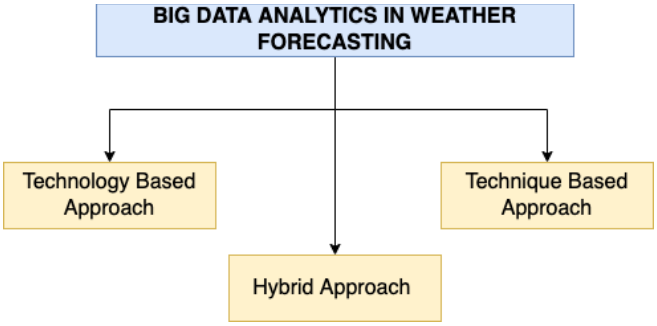

Fig -3: Approaches to analyze big weather data [6]

2.5 Big data analytics as a business concept

Businesses these days thrive on data. It has become the most essential aspect of a business. Again, data is being generated by businesses every day and hence classified as big data. This data is stored in the form of data lakes within the organizations. The data lakes act as a repository to store the big data in raw and unstructured format that are generated from various sources available within the organization.

As illustrated in Fig -4, the big data analytics process in handling business data is a three-phase process which includes: data ingestion, data processing and data usage.

The first phase is data ingestion which means collecting data from various sources which could be internal or external but relevant to the organization. As a part of transferring the data to the next phase, an ingestion API is used which stands for Application Program Interface which acts as a link between two applications hence making it easier to ingest data into the data lakes with efficiency. Here the next phase begins where the data that is ingested is processed within the data lakes. If required, data marts are created which consist of tables that contain

processed and filtered data from data lakes which makes it easier to analyze the data as the size of data and redundancy get reduced. This is a key concept of big data analytics where data marts are created from data lakes to make the analysis process faster and efficient. Finally, the ingested and

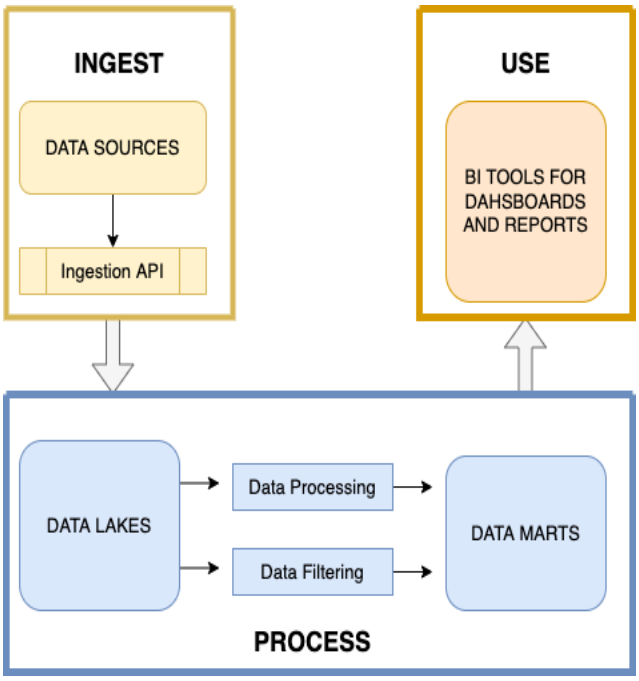

Fig -4: Big business data analytics process

processed big data is put into use with the aid of Business Intelligence (BI) tools which help in creating data jobs which query the data available to create various types of business reports and dashboards in the form of pie charts, big values, grouped charts, bar graphs, etc. These help the businesses in gaining in depth insights about their products, technologies and customers, hence helping the business grow and gain profits.

3. CHALLENGES

After discussing the existing systems, it is evident that big data analytics prove to be very advantageous in helping industries and organizations generate vale out of the data collected. But dealing with big data is not an easy task and brings along with it certain challenges. This section discusses the challenges faced in dealing with big data.

3.1 Uncertainty in big data

Uncertainty in big data relates to incomplete or at times incorrect data collection. [2] This can cause a huge impact on the big data analysis process as the analysis would be based on incomplete data which would most probably lead to wrong results. This would raise a major challenge for the systems and organizations making use of the big data analytics. For example, if a biased dataset is gathered for a medical research and analysis is performed on this dataset itself, it would lead to inaccurate results which could ruin the entire research study or if there is uncertainty in big weather data collected it can cause a negative impact on all the industries that depend on this day for their day – to – day functioning. There are a few theories that exist to deal with this uncertainty in big data, some of which include: Bayesian theory, Belief function theory, Fuzziness, Probability theory and many more [2].

3.2 Security and privacy of data

Handling such enormous volume of data is not an easy task. Few of the tools used for performing big data analysis make use of data disparate sources to do so which leads to a high risk of the data being exposed, making it vulnerable. When handling business data in this way it can lead to major security concerns as this data would confidential and private to the company. So, if the right big data analysis tools and technologies are not used, it can cause a breach of security and privacy leading to high amounts of risk. For example, while performing medical research, if the patient's dataset gets compromised during the analytics process, it can breach the privacy of the patients and land the researchers into serious trouble. The process of data handling must be done by keeping the risks associated with its mishandling in mind. Company security and confidentiality norms should always be followed while dealing with organizational data.

3.3 Scattered data storage

When the data is collected from different sources that reside in different databases which do not communicate with each other it hinders the analytics process as the initial access is limited to only certain data and the entire data view is not available. A lot of time gets wasted in raising access permissions. This is not a high priority concern as compared to the above - mentioned challenges, but this definitely slows down the analytics process and hence causes a delay in achieving the desired results.

3.4 Big data handling costs

Right from the data collection stage till the final analysis stage, the big data analysis process demands high amount of expenses. This is due the fact that large volumes of data are being dealt with and it is essential to use the best available tools and techniques for data handling, storage, management and analysis. For example, if an organization decides on using an on – premise solution for carrying out the process, then additional costs will be incurred for arranging the set up of the hardware devices. Or if the organization opts for a cloud – based solution then additional costs are required for hiring trained specialists in this field and for the development of cloud services and frameworks. Hence, depending on the organizations needs and goals, resources should be allocated accordingly.

4. CONCLUSION

In this study a rapidly growing technology has been discussed – big data. Having understood the five main characteristics of big data – Volume, Variety, Velocity, Value and Veracity, various existing systems have been discussed which generate big data and require new tools and technologies to handle this enormous volume of data. Here is where big data analytics comes into picture. The entire process starting from data collection to data storage, to data management and finally data analytics, is handled by big data analytics. It provides various data storage possibilities some of which are Hadoop, Kafka, Apache Spark and many more. It also provides various methodologies to process this stored data so that some value can be gained from the generation of big data.

With the application of big data analytics some of the most essential industries and systems of the world have gained huge profits – the health care and medicine industry, the transportation systems making use of intelligent devices, the big corporations and businesses, the weather forecasting system, and lastly the biggest industry of this era – social media. All of these make use of big data analytics to add value to large volume of data that they generate every day. But with so many advantages and such large volumes, challenges are bound to appear like the uncertainty of big data, scattered storage, security and privacy as well as big data handling costs.

In conclusion, almost everything in today's world generates large volume of data daily, and this is just the start. This is a rapidly growing concept which requires specialized techniques to extract value out of it for which analytics plays a key role – it helps generate value out of the large volumes of data generated. Hence, big data is the future of analytics as their interrelation helps gain maximum profits to both individuals and industries.

REFERENCES

[1] The Digitization of the World From Edge to Core, #US44413318, David Reinsel, John Gantz , John Rydning, 2018.

[2] H. Hariri, Reihaneh & Fredericks, Erik & Bowers, Kate, "Uncertainty in big data analytics: survey, opportunities, and challenges," Journal of Big Data, 6, Jun 2019, doi: 44. 10.1186/s40537-019-0206-3

[3] Corsi, A., de Souza, F.F., Pagani, R.N. et al, "Big data analytics as a tool for fighting pandemics: a systematic review of literature," J Ambient Intell Human Comput 12, 9163–9180, Oct 2020, doi: 10.1007/s12652-020-02617-4.

[4] Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Ebrahim Mahdipour, Seyed Mahdi Jameii, "Big data analytics meets social media: A systematic review of techniques, open issues, and future directions," Telematics and Informatics, Volume 57, 101517, ISSN 0736-5853, March 2021, doi: 10.1016/j.tele.2020.101517.

[5] L. Zhu, F. R. Yu, Y. Wang, B. Ning and T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 383-398, Jan. 2019, doi: 10.1109/TITS.2018.2815678.

[6] Fathi, Marzieh; Haghi Kashani, Mostafa; Jameii, Seyed Mahdi; Mahdipour, Ebrahim, "Archives of Computational Methods in Engineering,", Vol. 29 Issue 2, p1247-1275. 29p, Mar 2022.

[7] Blagoj Ristevski, Ming Chen, "Big Data Analytics in Medicine and Healthcare," Journal of integrative bioinformatics, Mar 2018.

[8] Galetsi, Panagiota & Katsaliaki, Korina & Kumar, Sameer, "Big data analytics in health sector: Theoretical framework, techniques and prospects," International Journal of Information Management, Elsevier, vol. 50(C), pages 206-216, 2020.

[9] Junliang Wang, Chuqiao Xu, Jie Zhang, Ray Zhong, "Big data analytics for intelligent manufacturing systems: A review," Journal of Manufacturing Systems, Volume 62, Pages 738-752, ISSN 0278-6125, Jan 2022, doi: 10.1016/j.jmsy.2021.03.005.

[10] Hassani H, Beneki C, Unger S, Mazinani M.T, Yeganegi M.R, "Text Mining in Big Data Analytics," Big Data Cogn. Comput, Jan 2020, doi: 10.3390/bdcc4010001.

[11] Chong, Dazhi & Shi, Hui, "Big data analytics: a literature review," Journal of Management Analytics 2, 175-201, Jul 2015, doi: 10.1080/23270012.2015.1082449.

[12] Tsai, CW., Lai, CF., Chao, HC. et al, "Big data analytics: a survey," Journal of Big Data 2, 21, Oct 2015 doi: 10.1186/s40537-015-0030-3

[13] Awotunde J.B., Jimoh R.G., Oladipo I.D., Abdulraheem M., Jimoh T.B., Ajamu G.J," Big Data and Data Analytics for an Enhanced COVID-19 Epidemic Management," Artificial Intelligence for COVID-19. Studies in Systems, Decision and Control, vol 358, Springer, Jul 2021, doi: 10.1007/978-3-030-69744-0_2.

[14] Alam T. Khan, M.A. Gharaibeh, N.K. Gharaibeh M.K., "Big Data for Smart Cities: A Case Study of NEOM City, Saudi Arabia," Smart Cities: A Data Analytics Perspective. Lecture Notes in Intelligent Transportation and Infrastructure, Springer, Dec 2021,doi:10.1007/978-3-030- 60922-1_11