Big Data: The Future of Information Management

AKASH SHARMA [1], ANUREET CHHABRA[2]
[1] Chandigarh College of Engineering and Technology, Chandigarh
[2] Chandigarh College of Engineering and Technology, Chandigarh

ABSTRACT Each piece of hardware and software in today's digital world produces quintillion bytes of data per day. For instance, Facebook produces 50 terabytes of data per day, while Fliker produces 1.8 billion photographs shared every day and Google processes hundreds of petabytes of data every month. Traditional methods of data management just aren't up to the task of handling this volume of information; enter big data. It stands for the collection of methods and systems used to manage massive amounts of information. Large amounts of data must be moved, processed, and stored. Researchers and major corporations alike are investing in innovative methods for processing massive data. The purpose of this article is to familiarize the reader with the current state of the area of Big data research, including its major trends and obstacles.

KEYWORDS Big Data, Blockchain, IoT, Security

I. INTRODUCTION

The concept of data processing and storing is introduced in 1970. But as the volume of data increase the processing of data at the single mainframe system is neither possible and nor economical. So in 1980, a new concept of parallel database system data processing method was proposed [1]. In this, data is stored and processed by a cluster of systems, named "Teradata system". The first database system based on Teradata is developed in 1986 and used to store 1 TB of data [2].

With the start of the new century, the role of the internet in everyone's life is increasing and due to this, the data generated by users is also increased. The three V's of big data were defined in 2001 [3]. The fourth V (value) is added to the big data definition in 2011 [4]. The fifth V (veracity) in the definition of big data was introduced in 2012 [5], [6], [7]. To handle the increased amount of data, search engine companies developed their own big data processing tools and technologies, like google developed GFS [8] and MapReduce [9].

Big data is huge and it is still growing. According to an estimate, by 2023 big data market reach $103 billion. In today's world, every user every day generates 2.5 quintillion bytes. In google have to manage 40000 search queries per second and 3.5 billion searches per day, which sum up to 1.2 trillion searches in a year, in all these searches 15% search data are searched the first time.

With the development of new-age digital technology, the number of devices increased which in turn increase the big data because a large number of users generates different formats of data, like GPS location, audio, video, text file. According to the estimate, 80% to 90% of total generated data is of unstructured form, which increases the challenge of big data analytics.

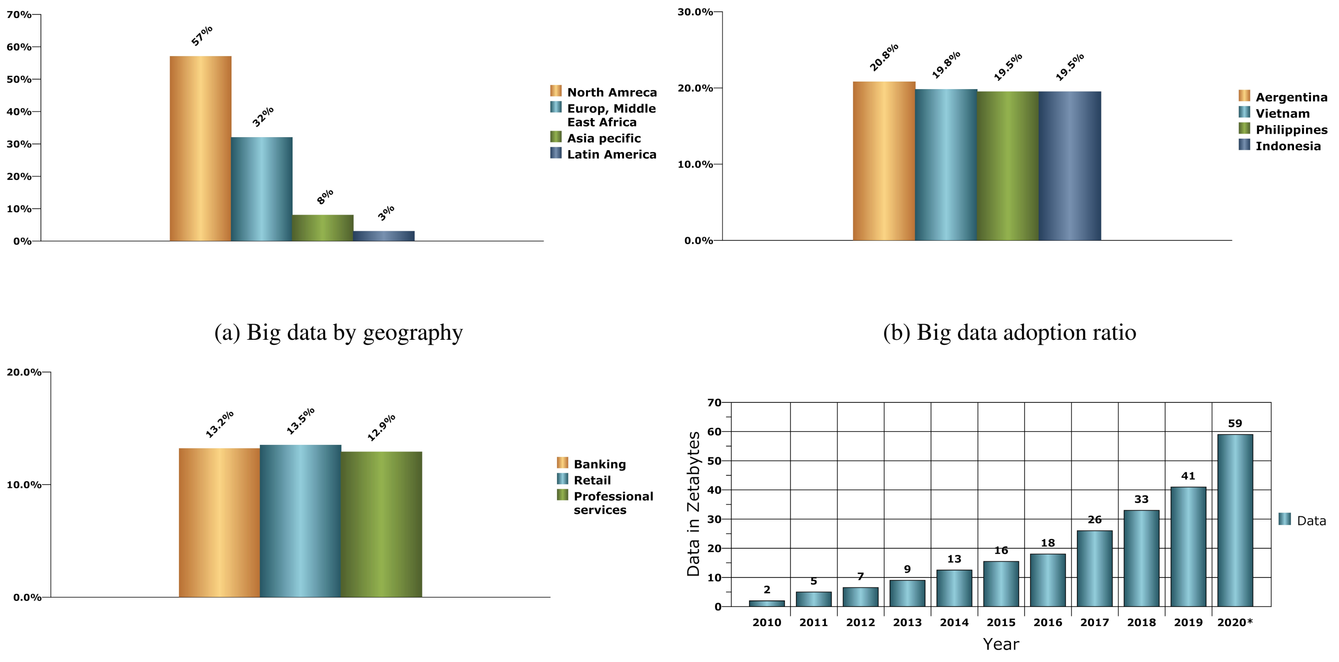
Large industrial sectors and service sectors generate big data. The banking sector is the major contributor to the generation of big data. In 2013, the global financial sector contributes 64% of total big data generation, due to this financial sector includes big data analytics into their infrastructure. Figure 1c represents the investment by different sectors in big data analytics. Different governments also see the potential of big data and start the projects in big data Figure 1b represents the adoption ration of governments.

In 2012, there was a total of 2.5 billion peoples connected to the internet. This figure reached 2..5 billion in two years and up to 2019 about 4.1 billion people were connected to the internet. So the big data generated by the users have been keeping on increasing, the Figure 1d represents the increase in the amount of big data in last ten years and Figure 1a represents big data generated by different countries.

II. LITERATURE REVIEW

Authors in [10] have surveyed the bigdata frameworks and the challenges faced by big data technologies. The author has does a comparative study of different big data frameworks like Hadoop, Spark, Strom, Flink, and Samza. In addition to theoretical analysis, the author has performs different experiments to analyze the performance of different big data frameworks.

Authors in [11] have analyzed security and privacy issues and their solutions related to big data Authors in [12]



(a) Big data by geography

(b) Big data adoption ratio

(c) Big data investment

(d) Big data change

FIGURE 1: Big data statistics

proposed joint computation offloading and task caching for multi-user and multi-task MEC systems. Authors in [13] proposed accelerating 3D medical volume segmentation using GPUs. In another work, authors [14] proposed accelerating compute-intensive medical imaging segmentation algorithms using hybrid CPU-GPU implementations. Authors in [15] proposed a prototype method to discover workflow violations and XSS vulnerabilities in PHP web applications. Author in [16] review XSS abuse and defense techniques. Author in [17] proposed an ISP level solution to combat DDoS attacks. Authors in [18] proposed a neural fuzzy framework for phishing detection. Author in [19] review DDoS attack detection techniques. Author in [20] proposed URL features-based phishing detection system using machine learning. Author in [21] proposed an identity-based authentication mechanism for the maritime transport system. Authors in [22] review recent advances in fog and mobile edge computing.

Authors in [23] have surveyed the application of big data technologies in the development of smart cities. Smart cities have been generating a huge amount of data from different applications like smart traffic lights, smart grid, smart governance, etc, and with the help of big data technologies analysis of this large amount of data is possible. The author reviewed the solutions provided by the big data techniques and he also reviewed the challenges faced by it to fulfill the needs of smart cities.

Authors in [24] have surveyed big data architecture and different big data technologies. The author has reviewed the cloud computing service model for big data processing. Authors in [25] have reviewed different big data tools and technologies. The author has reviewed both open source and commercial tools that are used in big data processing and he also given insight into the different big data applications.

Authors in [26] has surveyed the streaming big data processing techniques. The author has explained the different tools and frameworks used to analyze the streaming big data.

Machine learning and Deep learning are the two techniques that have a wide area of applications but the requirement of a large amount of training data is their primary limitation. All the data generating from the digital devices are not directly used as the training data because the generating data is in unstructured form or not labeled. Researchers proposed different models through which the big data can be usefull for Machine learning (ML) and Deep learning (DL) techniques. [27] and [28] have surveyed different ML and DL models for analyzing big data.

III. FIVE V'S OF BIG DATA

Due to the proliferation of digital gadgets, a vast amount of data is produced. Big Data is a phrase used to describe massive amounts of data that cannot be processed using traditional data management techniques. The five "Vs" of big data — volume, variety, velocity, value, and veracity — include all of the strategies used to extract useful information from massive amounts of data. According to **??**, the five "V's" of big data are shown in the following way:

- Volume – Big data is all about volume. There is no threshold by which we can set the limit of big data [29], [30]. Currently, datasets in the range of exabyte (EB) or zettabytes (ZB) are considered as big data [31], [32].

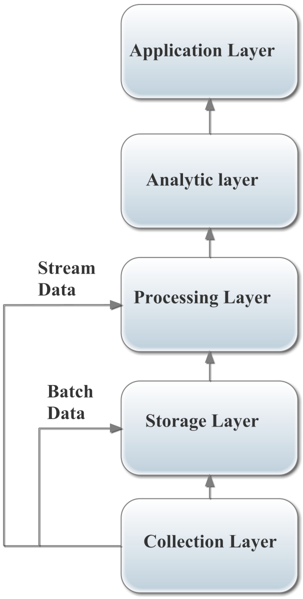


FIGURE 2: Architecture of big data analysicts and processing

The present processing tools always have an issue in processing this large amount of data [33]

- Variety – Big data can be in any format, it can be structured data, unstructured data, or semi-structured data. Structured data is stored in relational databases and generally in well-organized form, in contrast to unstructured data, which is random data i.e text, audio, video, etc. Semi-structured data are stored in NoSQL datasets [34], [35].
- Velocity – It is related to the generation and processing speed of the data. In big data, the speed of processing of data should be compatible with the speed of the generation of data. Stream data is generated in real-time, so it needs a processing method that can process it in real-time. In healthcare devices if there is a speed mismatch between the data generation and data processing then it may lead to patient death [36].
- Value – It is related to the output from the processor. Big data is processed from different tools and only relevant data is stored. Selecting relevant data is also a research topic [37].
- Veracity - The quality of data is also a big issue in the big data context. Due to the increasing sources of data generation, the quality of data is effected. We can differentiate data veracity [38] into good, bad, and undefined Data should not contain any noise and redundant terms. For example, if the data in a dataset which contains the health record of the patients consist of inconsistency then it is difficult to identify the disease pattern

IV. ARCHITECTURE OF BIG DATA

The stack of big data analytic is proposed by [39]. It is a five-layer stack. Each layer in the stack is fulfilling one of the big data analytic techniques. Each layer communicating with the upper and lower layer. The layered architecture is represented in Figure 2 and each layer is explained as follows:

- Layer 1- This layer collects the data for the upper layer. The collected data can be structured data, unstructured data, or semi-structured data, so the main task of this layer is to maintain the quality of data during the collection of data.
- Layer 2 – Different forms of data in big data analytics has different storage needs. Batch data can be off-line stored and processed and stream data need real-time processing. So the data storage layer stores the data as per the requirements and hence provides the scalability to big data analytics.
- Layer 3- This layer handles the data from the storage or real-time input and processes it for the upper layers. This layer is based on the scheduling module to increase the scalability of big data analytics.
- Layer 4 – This layer processed big data into big volume. The primary task of layer 4 is to perform data mining and user-customized tasks.
- Layer 5 – This layer fulfills the user's demands. It presents the data according to the requirements of the industrial needs.

V. CONCLUSION

Big data analysis methods are necessary because we live in a digital age where digital gadgets create billions of data every day. This article provides a high-level overview of the five V's of big data, with an examination of the state of the art in big data statistics. Then, we give the basic architecture of big data analytic. This paper will help the researchers to get better understanding of big-data analytics.

REFERENCES

[1] D. DeWitt and J. Gray, "Parallel database systems: the future of high performance database systems," Communications of the ACM, vol. 35, no. 6, pp. 85–98, 1992.
[2] T. Walter, "Teradata past, present, and future," UCI ISG lecture series on scalable data management, vol. 1, no. 1, pp. 44–48, 2009.
[3] D. Laney, "3d data management: controlling data volume, variety and velocity. meta group research, february 6," 2018.
[4] J. Gantz and D. Reinsel, "Extracting value from chaos. idc iview (2011)," 2015.
[5] A. Jain, "The 5 vs of big data," IBM Watson Health Perspectives. Dostupno na: https://www. ibm. com/blogs/watson-health/the-5-vs-of-big-data/.[30.05. 2017], 2016.
[6] I. B. Data and A. Hub, "Extracting business value from the 4 v's of big data," Retrieved July, vol. 19, p. 2017, 2016.
[7] D. Snow, "Dwaine snow's thoughts on databases and data management," 2012.
[8] C. Verma and R. Pandey, "Comparative analysis of gfs and hdfs: Technology and architectural landscape," in 2018 10th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2018, pp. 54–58.
[9] K. Kalia and N. Gupta, "Analysis of hadoop mapreduce scheduling in heterogeneous environment," Ain Shams Engineering Journal, 2020.

[10] W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, and E. M. Nguifo, "An experimental survey on big data frameworks," Future Generation Computer Systems, vol. 86, pp. 546–564, 2018.

[11] M. Singh, M. N. Halgamuge, G. Ekici, and C. S. Jayasekara, "A review on security and privacy challenges of big data," in Cognitive computing for big data systems over IoT. Springer, 2018, pp. 175–200.

[12] I. A. Elgendy and et al., "Joint computation offloading and task caching for multi-user and multi-task mec systems: reinforcement learning-based algorithms," Wireless Networks, vol. 27, no. 3, pp. 2023–2038, 2021.

[13] M. Al-Ayyoub and et al., "Accelerating 3d medical volume segmentation using gpus," Multimedia Tools and Applications, vol. 77, no. 4, pp. 4939–4958, 2018.

[14] M. A. Alsmirat and et al., "Accelerating compute intensive medical imaging segmentation algorithms using hybrid cpu-gpu implementations," Multimedia Tools and Applications, vol. 76, no. 3, pp. 3537–3555, 2017.

[15] S. Gupta and et al., "Php-sensor: a prototype method to discover workflow violation and xss vulnerabilities in php web applications," in Proceedings of the 12th ACM international conference on computing frontiers, 2015, pp. 1–8.

[16] B. Gupta, S. Gupta, S. Gangwar, M. Kumar, and P. Meena, "Cross-site scripting (xss) abuse and defense: exploitation on several testing bed environments and its defense," Journal of Information Privacy and Security, vol. 11, no. 2, pp. 118–136, 2015.

[17] B. B. Gupta, M. Misra, and R. C. Joshi, "An isp level solution to combat ddos attacks using combined statistical based approach," arXiv preprint arXiv:1203.2400, 2012.

[18] A. Almomani and et al., "Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing email," arXiv preprint arXiv:1302.0629, 2013.

[19] A. Mishra and et al., "A comparative study of distributed denial of service attacks, intrusion tolerance and mitigation techniques," in 2011 European Intelligence and Security Informatics Conference. IEEE, 2011, pp. 286–289.

[20] A. K. Jain and et al., "Phish-safe: Url features-based phishing detection system using machine learning," in Cyber Security. Springer, 2018, pp. 467–474.

[21] B. B. Gupta, A. Gaurav, C.-H. Hsu, and B. Jiao, "Identity-based authentication mechanism for secure information sharing in the maritime transport system," IEEE Transactions on Intelligent Transportation Systems, 2021.

[22] E. Ahmed and et al., "Recent advances in fog and mobile edge computing," p. e3307, 2018.

[23] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," Journal of Internet Services and Applications, vol. 6, no. 1, p. 25, 2015.

[24] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," Journal of Internet Technology, vol. 21, no. 2, pp. 393–405, 2020.

[25] Y. Arfat, S. Usman, R. Mehmood, and I. Katib, "Big data tools, technologies, and applications: A survey," in Smart Infrastructure and Applications. Springer, 2020, pp. 453–490.

[26] A. H. Ali and M. Z. Abdullah, "Recent trends in distributed online stream processing platform for big data: Survey," in 2018 1st Annual International Conference on Information and Sciences (AiCIS). IEEE, 2018, pp. 140–145.

[27] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," IEEE Transactions on Knowledge and Data Engineering, 2019.

[28] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146–157, 2018.

[29] X. Wang, L. T. Yang, L. Kuang, X. Liu, Q. Zhang, and M. J. Deen, "A tensor-based big-data-driven routing recommendation approach for heterogeneous networks," IEEE Network, vol. 33, no. 1, pp. 64–69, 2019.

[30] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International journal of information management, vol. 35, no. 2, pp. 137–144, 2015.

[31] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," Journal of Big Data, vol. 6, no. 1, p. 44, 2019.

[32] N. R. Vajjhala, K. D. Strang, and Z. Sun, "Statistical modeling and visualizing open big data using a terrorism case study," in 2015 3rd International Conference on Future Internet of Things and Cloud. IEEE, 2015, pp. 489–496.

[33] D. Saidulu and R. Sasikala, "Machine learning and statistical approaches for big data: issues, challenges and research directions," International Journal of Applied Engineering Research, vol. 12, no. 21, pp. 11 691–11 699, 2017.

[34] P. Martins, M. Abbasi, and F. Sá, "A study over nosql performance," in World Conference on Information Systems and Technologies. Springer, 2019, pp. 603–611.

[35] J. Pokornỳ, P. Škoda, I. Zelinka, D. Bednárek, F. Zavoral, M. Kruliš, and P. Šaloun, "Big data movement: a challenge in data processing," in Big Data in Complex Systems. Springer, 2015, pp. 29–69.

[36] D. Dolezel and A. McLeod, "Big data analytics in healthcare: Investigating the diffusion of innovation," Perspectives in health information management, vol. 16, no. Summer, 2019.

[37] A. Nargundkar and A. J. Kulkarni, "Big data in supply chain management and medicinal domain," in Big Data Analytics in Healthcare. Springer, 2020, pp. 45–54.

[38] A. P. Reimer and E. A. Madigan, "Veracity in big data: How good is good enough," Health informatics journal, vol. 25, no. 4, pp. 1290–1298, 2019.

[39] J. Y. Zhu, B. Tang, and V. O. Li, "A five-layer architecture for big data processing and analytics," International Journal of Big Data Intelligence, vol. 6, no. 1, pp. 38–49, 2019.