Big Data Heterogeneity - A Short Review

Bhavana Hotchandani, Vishal Dahiya

Abstract**:** ***The use of electronic devices and the digital globalization has brought a lot of technical revolution in the world. People use many devices and several applications available on the cloud. This leads to generation of ample amounts of data which is termed as huge data, technically a big data. This big data is heterogeneous in nature which is a combination of structured and unstructured data. This review paper aims to show the general facts of big data and then narrowing it down we are showing types of heterogeneous big data. This paper will help the novel researchers to gain facts about big data and its heterogeneity.***

Keywords: 7Vs of Big data, Big Data, variety, heterogeneous data, heterogeneity.

I. INTRODUCTION

A. A Revolutionary Fact Of Data And Its Magnitude - Big Data

The world is continuously enhancing in its technological aspects and these advancements are generating a lot of data daily. This data is junk and very big and hence the term big data was coined a decade ago. Big data growth statistics reveal that data creation will be over 180 zettabytes by 2025. That was about 118.8 zettabytes more than in 2020. The reason for the spike is that the pandemic has triggered an increase in petition for remote learning, working, and entertainment. Storage for this data will grow at a Compound Annual Growth Rate (CAGR) of 19.2% during the forecast period. That's a big change in view of that users only stored 2% of the data in 2020. Figure 1 below shows the total volume of data generated, stored and processed worldwide. The figure explains that the amount of data generated in the year 2021 was 79 zettabytes which is expected to rise to 181 zettabytes by 2025. The figure is obtained from the work "Total data volume worldwide 2010-2025" [12]. To accomplish such voluminous data is the biggest challenge in the world today. Several researchers and community have geared up to handle this challenge with all their wit but still the concern to manage such data remains unsolved. Any traditional tools cannot manage, store and process such high voluminous data today. Hence upgrading in the traditional approach was very much necessary and is a domino effect as data volume is increasing day by day and year by year all over the world.

Manuscript received on 30 June 2022 | Revised Manuscript received on 05 July 2022 | Manuscript Accepted on 15 July 2022 | Manuscript published on 30 July 2022.
* Correspondence Author

Ms. Bhavana Hotchandani*, Department of Computer Science, Indus University, Ahmedabad (Gujarat), India. Email: bhavana.mca@gmail.com

Dr. Vishal Dahiya, Department of Computer Science, Indus University, Ahmedabad (Gujarat), India. Email: cs.hod@indusuni.ac.in

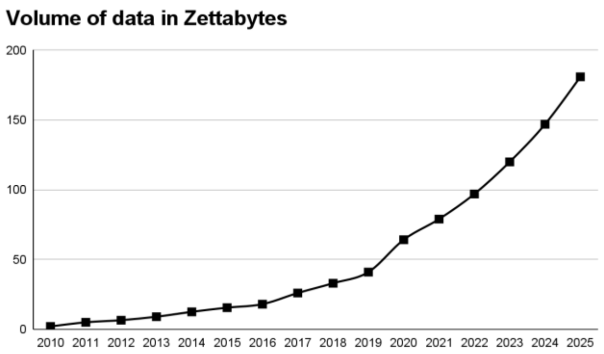

Fig. 1 Statistics of Data across the years from 2010 to 2025

There are basically three types of data which are commonly being stored and processed. These are Structured data, Semi-structured and Unstructured data. Structured data are all those which are stored in some predefined format like relations. While all those stored haphazardly and not in any particular relation are called as unstructured data. The mixture of both are all semi-structured by nature. These data differ in their nature and the type of storage, be it some in relation format or few in paragraph format while some with pictorial representation. Data which are mixed and with different varieties are usually referred to as heterogeneous data. Now let us dive down into the characteristics of big data. Big Data can be characterized by the 7 V's, which stands for Volume, Velocity, Variety, Variability, Veracity, Visualization and Value.

1.1 Volume: Volume is how much data we have – what is the size of data and data which was measured first in Gigabytes are now measured in Zettabytes and Yottabytes.

1.2 Velocity: Velocity is the speed at which data is processed and becomes accessible. I remember the days of nightly batches. Now, if it's not real-time, it's usually not fast enough.

1.3 Variety: Variety can be unstructured, and it can include so many different types of data, from XML to video to SMS.

1.4 Variability: Variability is different from variety. A coffee shop may offer 6 different blends of coffee, but if you get the same blend every day and it tastes different every day, that is variability.

1.5 Veracity: Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems. The simplest example is the contacts entering your marketing automation system with false names and inaccurate contact information. How many times have you seen Mickey Mouse in your database? It's the classic "garbage in, garbage out" challenge.

1.6 Visualisation: Visualisation is critical in today's world. Using charts and graphs to visualise large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.

1.7 Value: Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualisation – which takes a lot of time, effort and resources – you want to be sure your organisation is getting value from the data.

The 7Vs and their dynamics are explained in figure 2.

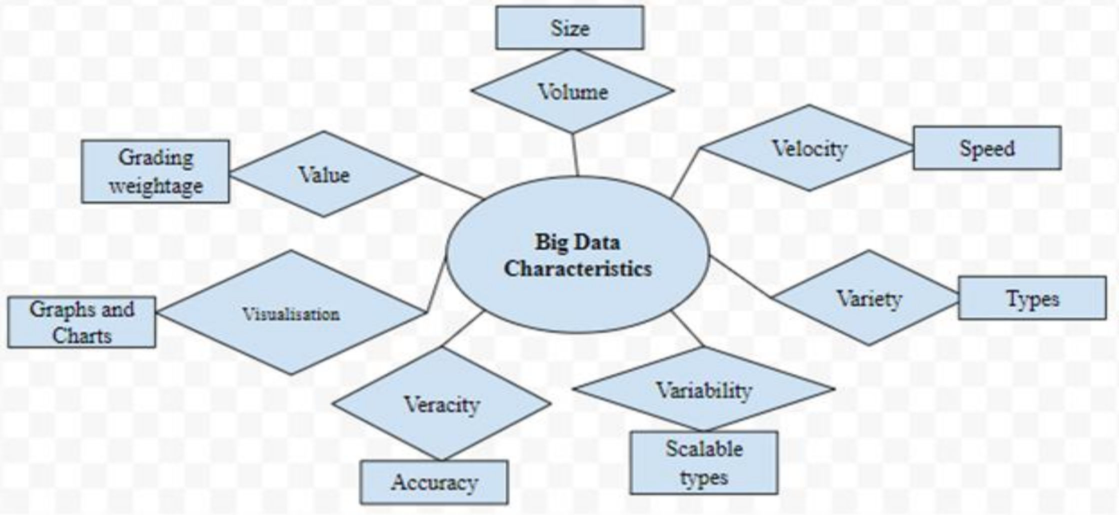

Fig. 2 Big Data Characteristics: 7Vs with their stated dynamics

Big Data has full-grown into becoming a requirement in this contemporary age. With the expansion and increase of apps and social media platforms, individuals and businesses are moving online. There's been an enormous increase in data. Social media platforms alone pull in over a million users daily. How this vast amount of data is handled, processed and stored is where Big Data comes into play.Big Data analytics has transformed the field of IT, augmenting and accumulating added advantages to organizations. It includes using analytics, new age tech like machine learning, mining, statistics and more. It empowers an organization to have well-organized resource management, progress customer services, enhances growth and development in their products, upsurge productivity and bring about intelligent decision making [21]. Out of the diverse characteristics of big data as deliberated above, 3Vs are very significant and construct a quantitative framework which can be analyzed statistically. Volume which refers to size, variety referring to different types and velocity which measures the speed of the data processing are the basic 3V's acting as a base pillar to the quantitative framework. Rest all 3Vs are modelled to build a qualitative framework and lastly the 7th V called Value is gained combining the qualitative framework or quantitative framework or both in hybrid. Narrowing down to the mentioned quantitative framework 3Vs, let us understand the interconnection and dynamics between these 3Vs with a figure 3 as shown here. This 3V model is interrelated to each other moving from voluminous data to different types of data and lastly the processing of the data with highest possible accuracy and speed. Thus, the volume represents the ever growing amount of data in Petabytes, Exabytes, Zettabytes which is about to reach yottabytes in a decade. This will further be generating a challenge in the current stage of storage systems [18]. While, on the other hand, the variety of data produced by the multiple magnitude in its structured, unstructured or semi-structured format from different devices like sensors, smart phones, social media in raw, tweets, Instagram reels and other rich media formats is further complicating the data processing and its storage. Finally, the velocity aspect comes into picture in order to retrieve and store such huge data and then process it further with as much high speed as possible. This is becoming like a bottleneck for the current traditional systems as they are not suited to deal with different formats, with such ample size, having varying processing time requirements [19]. Henceforth, from an information processing perspective, these three characteristics of big data together describe the new challenges present to the backend systems.

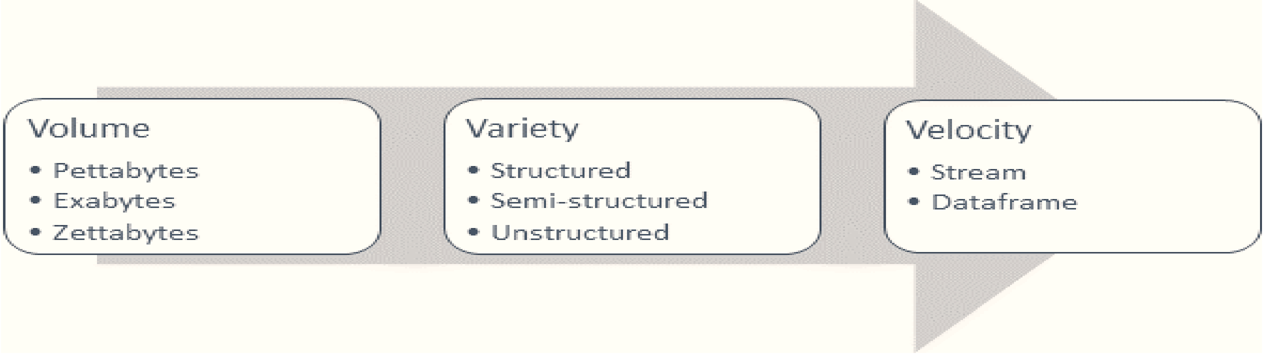

Fig. 3 3Vs leading to heterogeneous big data

The 3Vs model provides relatively a very abridged framework which is well tacit by today's researchers and practitioners working with big data. But the foremost concern is that the streamlined representation of the data processes intricate in it, can lead to architectural pitfalls on the design of big data platforms [20]. Considering the 3V framework which discourses particular cases of diverse requirements and how it can help in attaining saturation to the existing infrastructure by identifying and analyzing the value of data and analyzing it to a degree where decision making system is concluded, only these 3Vs will be targeted. This architectural paradigm made of Volume, Variety and Velocity is what is called Heterogeneity and we are further elaborating our study on the heterogeneous data within big data.

II. LITERATURE SURVEY

Data heterogeneity and its mitigation have been explored in few works. V. Jirkovskỳ [14], deliberate the several types of data heterogeneity existing in a cyber-physical system. The diverse groups of data heterogeneity are syntactic heterogeneity, terminological heterogeneity, semantic heterogeneity, and semiotic heterogeneity as conferred by [14]. They have also focused on the causes of heterogeneity in depth. The device heterogeneity is considered for smart localization using outstanding neural networks as shown by [2]. Though, heterogeneity mitigation in the data used for localization would have produced more consistency in the results. Device heterogeneity is addressed using a localization method and Gaussian mixture model by [15] . But the level and the worth of heterogeneity in terms of its value is missing. Zero-mean and unity-mean features of Wi-Fi (wireless fidelity) established signal strength is also used for localization promotion to mitigate device heterogeneity as shown by [7] and [9]. The tactic is still not used for data cataloguing or prediction purposes from multiple devices using neural networks. The effort presented by [1] goals at bringing interoperability in one mutual layer by using semantics to supply heterogeneous data streams produced by altered cyber physical systems. A mutual data model using related data technologies is used for the resolution. The notion of service oriented architecture is presented by [11] for extenuation of data heterogeneity. However, the adaptability of the data management system remains unmapped. In paper [8], an approach for gaining better-quality data gathering results has been conversed. The model is projected to minimise data in the primary phases. In this study, the recommended method focused exclusively on data reduction. It does not display the use of Machine Learning (ML) approaches and data processing. There is a paper [10] by Song et al. to test methods that can improve corporate decision-making capabilities. Through this investigation, multiple ML tactics were initiated to be used in data analysis. Therefore, the data evaluation process recommended focused chiefly on defining the best result for the open data according to the priorities of this study. The authors Jimenez-Marquez et al., of the paper [6], offered a methodical attitude for understanding of the approaches and techniques used for social media data in big data analytics. They analysed tourism data; in precise, reviews of hotel users, where user reviews are marked positively or negatively in the dataset. Results were used in many information combinations. Originally, learning classifiers with a multiclass classification were measured accessible for the five stars. The correctness result was about 57%. Subsequent, one star was then chosen as negative, two, three, four as neutral, and five as positive. It resulted in 67%. Moreover, the writers Tripathy et al., of the paper [13], used ML to gather feedback on a quantity of movie reviews. Several algorithms were used for the classification. By implementing a support vector machine (SVM) while implementing the Naive Bayes (NB) classifier, the accuracy was 72.9%, resulting in 73.7%.

The study by Dey et al., in the paper [3], grants a method for integrating heterogeneous and structured data into initiative analytics. In their method, structured data were handled in the form of time series that captured enterprise performance data such as daily progress reports, purchases figures, income, and stock prices, while heterogeneous data were taken from client reviews and remarks, discussion forums, blogs, and social media. This study utilises text processing and mining techniques to excerpt information from heterogeneous sources and allows multiple heterogeneous inputs to mechanize the knowledge discovery process through the correlation of data extracted components. Furthermore, in the study [5] the repetition of server consolidation for better programming aptitudes and lower power and cooling costs was implemented by a virtual machine migration-based algorithm to diminutive power consumption. The analysis depends on inferred complex resource planning based on three search algorithms mainly the sequential search, the optimum search and the random search.

The purpose of the algorithms was to competently use data center services. The results reveal that about 30% of energy savings have been achieved. In addition, the study by Gupta of the paper [4] presented the security problems with NoSQL databases MongoDB, HBase, and Cassandra and planned an outline to attain security for the web crawler applications using Cassandra, NoSQL. It used Amazon Web Services, an acquainted cloud platform, this led to the design of the ascendable architecture for the NoSQL datastore and Cassandra where the data is analysed and converted to a structured format suitable to evaluate the performance.

III. TYPES OF DATA HETEROGENEITY

Heterogeneity is defined as a dissimilarity between elements that comprise a whole [16]. When heterogeneity is present, there is diversity in the characteristic under study carried out by [17]. Let us closely look at the types of heterogeneous data. There is basically structured, semi-structured and unstructured data which is combined and prepared a heterogeneous data.

This heterogeneity in data represents its characteristics basically in four raw types mainly, Video, Audio, Textual Data and Images. These raw types are further preprocessed in a step where the extractions are made possible, which is again processed by implementing an algorithmic model of a researcher's choice and lastly we get a desired output with its accuracy measurement. A detailed view of the explanation is depicted in figure 4.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Usually at an initial stage of any of the basic raw types, preprocessing of data is done which is then followed by continuous processing using some sort of Machine Learning model and then an output which is nearly similar to its hypothetical data is generated. The steps vary but the core steps included are the mentioned ones. Moreover, once the feature extraction is completed at the primary level of data preprocessing, a dataset is bifurcated in usually 70:30 ratio or 80:20 ratios in training data and testing data. This split is referred to as data split. Once the split is carried out the processing via implementation of any machine learning model is performed which carries out the learning mechanism for a data split and generates a candidate model resulting in a comparative analysis between the algorithms used or implemented. This comparison helps to evaluate a basic machine learning model for a researcher, which can then be implemented further for any other set of data. This is how the basic structuring and processing of data is carried out by researchers. In this paper, we are focusing only on the types of heterogeneity where the processing of individual raw types has been depicted in figure 4. It also represents lastly the desired output that could be fetched for any of the raw data types implemented by the learner. Though the entire processing stages require a lot of understanding and the steps involved at every stage are out of scope of this paper.Thus the 3Vs, volume, variety and velocity leading to heterogeneous data can be processed in the mentioned way.

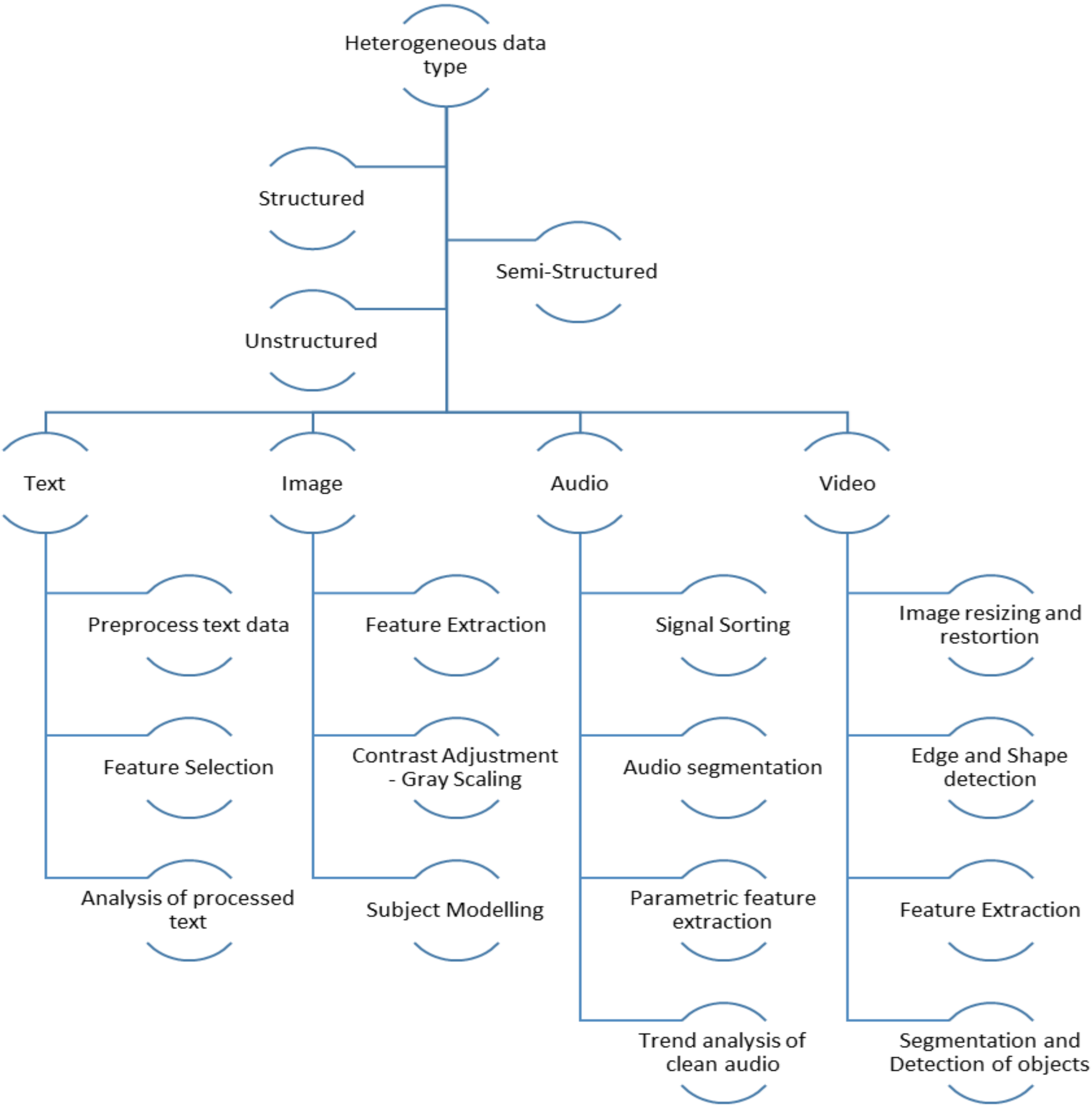

Fig. 4 Types of Heterogeneity and its processing stages for raw data types

IV. CONCLUSION AND FUTURE WORK

In this paper we have laid out the concepts of big data and its core facts. Moreover, we have also shown the types of heterogeneous data and how the process of moving towards big data from heterogeneity is also revealed. The work also compares several papers of researchers who have worked in different areas of big data and its heterogeneity as a whole. There are several applications using the heterogeneous characteristics of big data. In the future, we will show our model to carry out study with respect to stock market analysis and conclude the work by predicting the output of the stock market by implementing machine learning and deep learning algorithms. This paper will work as a base to the future paper we are targeting to publish with our model layout and implementation specifically for stock analysis.

REFERENCES

1. A. Kazmi, et al. “Overcoming the heterogeneity in the internet of things for smart cities.” Proceedings of the International Workshop on Interoperability and Open-Source Solutions. Springer, vol. 1, no. 1, 2016, pp. 20-35. [CrossRef]
2. A. Pandey, et al. “Residual neural networks for heterogeneous smart device localization in iot networks.” Proceedings of the 2020 29th International Con- ference on Computer Communications and Networks (ICCCN). IEEE, vol. 1, no. 1, 2020, pp. 1-9. [CrossRef]
3. Dey, et al. “A framework to integrate unstructured and structured data for enterprise analytics.” Proceedings of the 16th international conference on information fusion, Istanbul, Turkey, vol. 1, no. 12, 2013, pp. 1988–1995.
4. Gupta, et al. “Secure NoSQL for the social networking and e-commerce based bigdata applications deployed in cloud.” International Journal of Cloud Applications and Computing (IJCAC), vol. 8, no. 2, 2018, 113--129. [CrossRef]
5. Jeba, et al. “Towards green cloud computing an algorithmic approach for energy minimization in cloud data centers.” Research Anthology on Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing, vol. 1, no. 1, 2021, 846--872. [CrossRef]
6. Jimenez-Marquez, et al. “Towards a big data framework for analyzing social media content.” International Journal of Information Management, vol. 44, no. 1, 2019, pp. 1-12. [CrossRef]
7. Kumar, et al. “Target detection and localization methods using compartmental model for internet of things.” IEEE Transactions on Mobile Computing, vol. 19, no. 9, 2019, 2234--2249. [CrossRef]
8. Rehman, et al. “Big data reduction framework for value creation in sustainable enterprises.” International journal of information management, vol. 36, no. 6, 2016, 917--928. [CrossRef]
9. S. Kumar, and S.K. Das. “ZU-mean: fingerprinting based device localization methods for IoT in the presence of additive and multiplicative noise.” Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking. ACM, vol. 1, no. 1, 2015, p. 15.
10. Song, et al. “Decision tree methods: applications for classification and prediction.” Shanghai archives of psychiatry, vol. 27, no. 2, 2015, p. 130.
11. T. Fan, and Y. Chen. “A scheme of data management in the internet of things.” Proceedings of the 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content. IEEE, vol. 1, no. 1, 110-114. [CrossRef]
12. “Total data volume worldwide 2010-2025.” Statista, 23 May 2022, https://www.statista.com/statistics/871513/worldwide-data-created/. Accessed 24 June 2022.
13. Tripathy, et al. “Classification of Sentimental Reviews Using Machine Learning Techniques.” Procedia Computer Science, vol. 57, no. 1, 2017, 821--829. [CrossRef]
14. V. Jirkovskỳ, et al. “Understanding data heterogeneity in the context of cyber-physical systems integration.” IEEE Transactions on Industrial Informatics, vol. 13, no. 2, 2016, pp. 660-667. [CrossRef]
15. Yassine Laguel, et al. “Device Heterogeneity in Federated Learning: A Superquantile Approach.” Cornell University, arxiv, vol. 1, no. 1, 2020, pp. 1-7.
16. Chappelle D. Big Data & Analytics Reference Architecture, Oracle White Paper, Oracle Enterprise Transformation Solutions Series, September 2013, 1-39.
17. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2015 Feb 24; 2(1): 1. [CrossRef]
18. Yusuf Perwej, An Experiential Study of the Big Data, International Transaction of Electrical and Computer Engineers System, 2017, Vol. 4, No. 1, 14-25.
19. Almeida FL, Calistru C. The main challenges and issues of big data management. International Journal of Research Studies in Computing. 2013 Oct 9; 2(1). [CrossRef]
20. Zhang J, Yang X, Appelbaum D. Toward effective Big Data analysis in continuous auditing. Accounting Horizons. 2015 Jun; 29(2):469-76. [CrossRef]
21. Tak PA, Gumaste SV, Kahate SA, The Challenging View of Big Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, 5(5), May 2015, 1178-1181.

AUTHORS PROFILE

Ms. Bhavana Hotchandani, is working as an Assistant Professor with Indus University, Ahmedabad. She is passionate for research and carries an extensive ability to perform extraordinarily in research aspects. She is passionate about coding. She has cleared more then 15 online MOOCs. She has completed her MCA. Her area of interest lies with Big Data and data science. She is in the education field since 2010. She can be contacted on bhavana.mca@gmail.com or bhavnahotchandani.mca@indusuni.ac.in.

Dr. Vishal Dahiya, is working as a Professor and Head of Computer Science Department with Indus University. She has wide experience of 20+ Years. She has completed her Ph.D. in computer science from Sardar Patel University and currently is guiding more than 10 research scholars. She is acting as a chair person of research committee of Indus University for Computer Science and Engineering Department. She is a motivator and a mentor for students and faculties interested in research. Her research area focuses on Image Processing and Big Data. She has been widely renowned for her literally work on image processing. She can be approached at cs.hod@indusuni.ac.in.