Journal of Network Communications and Emerging Technologies (JNCET)
Volume 10, Issue 9, September (2020)

An Overview on Big Data: Characteristics, Security and Applications

Bihter Das
Department of Software Engineering, Technology Faculty, Firat University, 23119, Elazig, Turkey.
bihterdas@gmail.com

Abstract – Big data refers to enormous data sets that are complex, multidimensional, and in different formats. Big data analytics is the extraction or discovery of meaningful patterns by using advanced analytical techniques against very large, diverse data sets containing structural, semi-structural and unstructured data in different sizes from different sources and from terabytes to zettabytes. The analysis process of big data, on the other hand, is a very laborious and long way that requires great experience. In this review paper study, the important characteristics and challenges of big data are presented in general. Additionally, security and privacy issues were discussed and the problems experienced were expressed. Moreover, sectoral studies on the application areas of big data are presented.

Index Terms – Big Data Analytics, Big Data Characteristics, Data Security, Cloud Computing, Big Data Applications.

1. INTRODUCTION

The concept of data has been important and valuable in every period of human history, and the storage of data has come to this day by being shaped by various methods in different periods. In the past, data were on ancient inscriptions, stones, embroidered pieces of leather, paper, etc. Today it is mostly stored in relational or non-relational databases. The way data is stored, namely the storage technology, has been constantly changing from the past to the present [1].

Nowadays, the use of the internet has increased and the areas stored in the cloud have increased excessively. While users worry about the security of their data, they also experience the convenience of storing data in the cloud. For this reason, the demands for data storage in the cloud environment are increasing day by day. Thus, it becomes difficult to control, manage and effectively use the extremely large data volumes collected in the cloud environment.

In today, many large software companies (Facebook, Google, Amazon etc.) develop their own big data projects. Big data is one of the priority areas in many developed states and public institutions and organizations of our country. The term big data finds its place in many areas, with governments allocating significant financial resources for big data research. In addition, there are so many studies about big data with other related studies.

The most popular fields of study such as Internet of Things (IoT) [2-4], Industrial Internet of Things (IIoT) [5], Machine-to-Machine (M2M) Communication [6,7], Cloud Computing [8,9], Social Networks [10,11], Web Mining [12,13], Smart Grid [14,15], Artificial Intelligence (AI) [16,17], and Intrusion Detection and Prevention Systems (ID/IPS) [19,20] for cyber-security are directly related to the big data subject and their applications are intertwined.

The purpose of this study is to review in detail the issues of processing big data and data security on cloud computing. In the study, besides the difficulties of big data, the comparison of big data with traditional data analysis and the big data life cycle are also examined.

2. BIG DATA CHARACTERISTICS

In our digitalizing world, big data is used in many different areas in our daily life. There are great difficulties in collecting, storing, managing and analyzing the data used. Traditional data management and analysis systems are based on relational database management systems. However, relational database management systems are only valid for structured data other than semi-structured or unstructured data. Also, relational database management systems tend to use more expensive hardware. However, traditional relational database management systems fall short due to the large amount of data and heterogeneity. As a solution, cloud computing is used to support big data infrastructure requirements. Cloud computing offers cost, flexibility and easy update advantages. Distributed file systems and NoSQL databases are used in cloud systems for permanent storage and management solutions of large-scale irregular data sets.

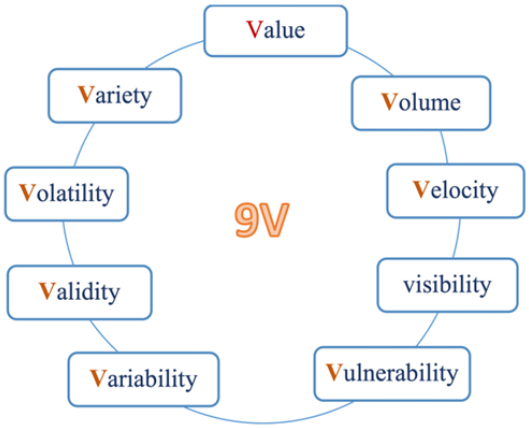

Figure 1 Characteristics of Big Data

Data-related challenges present challenges based on the characteristics and properties of the data. In the literature, among the studies of many different researchers, difficulties in the concept of big data have been suggested as 3V, 4V, 6V, 7V, 9V. Figure 1 demonstrates the several characteristics of Big Data, known as the 9V's, and mentioned in the literature [21].

- *Variety* is the variance of data types and data sources in different formats in digital environments. These are examined under three different categories: Structured Data, semi-structured data, and unstructured data.
- *Velocity* is the fast processing of data and its flow or transport to different *locations*.
- *Volume* is the amount of data that is continuously generated from different and diverse sources every day.
- *Value* is the identification of data that is more valuable than transforming and analyzing data.
- *Visibility* provides the visualization to solve and understand complex big data problems.
- *Vulnerability* is the evaluation of security vulnerabilities against cyber-attacks of excessively increased data and the provision of solutions.
- *Validity* is about the accuracy rates in the predictions made in data science applications.
- *Volatility* is about how long a large amount of data is kept in databases, data loss, and suddenly manipulation of data with different characters.
- *Vulnerability* is the evaluation of security vulnerabilities against cyber-attacks of excessively increased data and the provision of solutions.
- *Variability* deals with different types of data and varieties in many different sources in data science applications. So, it is about different types of big data received from different internal and external sources such as documents, emails, text messages, video, still images, audio, graphs, and the output from all types of machine-generated data from sensors, devices, RFID It is related to tags, machine logs, cell phone GPS signals, DNA analysis devices**.**

When the literature is examined, this number of V may be less or more in various sources. That is, it may vary according to the researcher's approach style. In fact, it is clear that they are all intertwined with each other when the details are entered.

3. BIG DATA SECURITY AND PRIVACY

Security and privacy in big data has always been the main concern, especially in terms of data management. Nowadays, this issue includes a much more sensitive process as big data is present in every sector [5, 22,23]. There is big data in cloud-based applications of Amazon, Google, Microsoft and Apple companies that almost everyone in the world use actively. These data are important and valuable personal data. Although the security and privacy of these data in data centers are under the guarantee of the company authorities, sometimes uncontrolled situations may occur. Many examples of this have been experienced especially on social media platforms. For this reason, the establishment of national data centers is on the agenda of every country. The multi-dimensional nature of big data brings with it many difficulties other than security and privacy. These can be summarized as follows.

- Due to the huge data volume, information with excessive value is difficult to identify and extract.
- High data rate in real-time applications increases dynamics, but requires powerful supercomputers and servers in analysis processes.
- Diversity of data sources ensure traceability of users. In addition, the variety of data types allows data owners to create more complex and rich user profiles. Moreover, this diversity leads to the diversity of business plans that make big data more attractive on a larger level.
- Developing extra security mechanisms for the elements or parameters that require privacy in big data brings additional costs.
- Laws and regulations on the protection of personal data are enacted all over the world regarding the privacy of information or data. This has been a mandatory requirement.

Security measures can also be examined under 3 different general headings. These can be listed as the physical security of the system room in the data processing center, the security on the network where the server is located, the security of the operating system and software running on the server. Network and system managers working in the system room have to take precautions for these security measures. It should be configured with current and new generation network intrusion detection and prevention systems. Network security methods and approaches here are the same as known network and system security issues [19,20] . However, what is different is that the database servers where big data are located can remain safe and working 24/7 against network attacks. In this context, there should be more than one server that can work with each other in sync and backup. Software configurations of database servers should be shut down against all kinds of injection attacks. For developers who develop software applications on servers, account definitions and authorizations should be made carefully and recorded with a signature. Log records of all transactions made by different users on the servers should be kept [12,13,24]. Taking these precautions secures the relevant unit managers.

There are different and reliable technologies for long-term permanent storage of big data kept in the servers of data centers

and for digital media security in the cloud environment. However, researchers and expert technical personnel of companies working in this field are constantly working to bring these technologies to a better point. We see how fast the technologies of many permanent storage units such as hard drive, CD, DVD and USB memory have changed from past to present. It is obvious that the existing systems and products are changing due to the difficulties experienced due to the increase in internet usage and the demands for big data. Because with the use of industry 4.0 solutions such as IoT, IoE, IoNT, smart systems, the size of big data that may occur in the future cannot be fully predicted. It has been the focus of attention of people from all walks of life, as institutions and organizations carry workflows and trade on digital media. For this reason, many new professions have been born in the internet world, and there will be more new ones in the future.

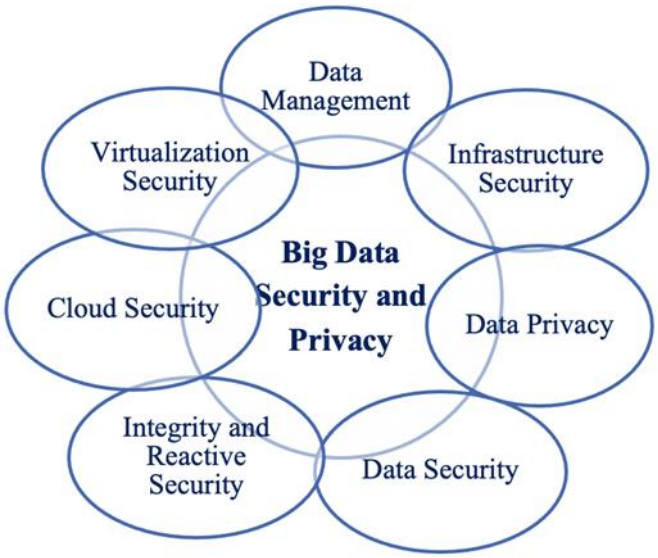

Figure 2 Big Data Security and Privacy Concepts

4. BIG DATA APPLICATONS

Big data analytics is the analysis of data in different formats and containing different types of content using advanced analytical and parallel techniques. It is a very difficult and problematic process to perform these analysis processes correctly and accurately. Because, rapidly changing and large amount of structural, semi-structural and non-structural data are analyzed as a whole. As a result of this process, it aims to extract valuable information from big data. For this purpose, powerful super computers that can perform fast processing in big data analytics operations and new generation software tools and technologies are used. For this reason, data is stored and processed in the cloud. Data is stored in powerful discs with high capacity and very fast processing. These servers in the cloud environment have powerful multi-core microprocessors and high-capacity memory. These servers are also located in system rooms in data processing centers in many public institutions and organizations. The security of the servers is directly related to the security of their data. It is of vital importance to take all kinds of security measures so that valuable and very important data is not manipulated, deleted or changed [25-27].

Among the main application areas of big data, it has a place in application areas in many sectors such as banking, communication, media, entertainment sector, health services, education, production, government services, social networks, media, smart grid systems, insurance, retail, trade, transportation. With the active and widespread use of the Internet, the number and type of these application areas are increasing day by day.

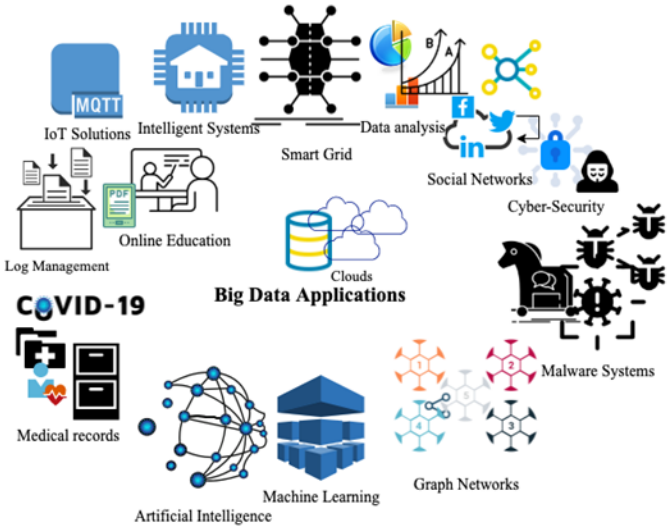

Figure 3 A Section of the Application of Big Data

When the literature is examined, there are hundreds of application examples in these sectors. In addition, there are dozens of solution approaches used in these applications. Moreover, there are many methods and algorithms in these solution approaches [28-30]. A few examples of various application sectors where big data is used can be listed as follows.

- *In the business sector;* It is used in situations such as customer analysis, distribution and logistics optimization, customer personalization.
- *In the retail sector;* It is used in situations such as employee income optimization, store behavior analysis, customer relationship analysis, product variety and price optimization.
- *In public institutions and organizations;* It is used in cases such as providing accessibility to data, creating confidentiality and transparency, taking action for appropriate products and services, reducing risk and fraud.

- *In technology sectors;* Intelligent web technologies are used in situations such as real-time analysis, rapid response generation, reduction of processing time, control and automation systems, data centers, automatic systems and artificial intelligence-based decision making to reduce risks.
- *In the education sector;* It is used in cases such as the programming of education and training systems, student analysis, lesson planning, online education systems.
- *As personal location data;* It is used in cases such as ad targeting, smart routing, emergency response by region.
- *In the health sector;* It is used in cases such as disease detection, patient control, personal DNA analysis [31,32]. The application areas and sectors can be increased much more.

5. CONCLUSION

Today, the big data topic is one of the hottest and most popular topics. Almost all of the states of the world are constantly moving their public services to the cloud with software within the scope of digital transformation. Thus, the scope of usage areas of big data is constantly increasing. Therefore, the analysis of big data stored in the cloud environment, the common use of these data in different applications, and the critical importance of privacy and security issues constantly keep the big data issue on the agenda. In this article, data analytics across big data, big data characteristics, big data security and privacy, big data applications in daily life are discussed. Investigations and researches show us that at least for a long time, the big data subject will continue to be popular and the solution approaches of private commercial companies will continue. In addition, the current and very popular Internet of Things, Industrial Internet of Things, Internet of NanoThings, Machine to Machine Communication, Internet of Behavior, Cloud Computing, Artificial Intelligence are also intertwined with big data. Moreover, finding and extracting the desired and necessary from the Internet ocean is difficult and involves very different processes. In parallel with this, academic studies increasingly continue their studies on methods, algorithms and new solution approaches within this scope.

REFERENCES

[1] D. Demirol, R. Das, and D. Hanbay, "Büyük veri üzerine perspektif bir bakış," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Inonu University, Malatya, Turkey., Sep. 2019, pp. 1–9, doi: 10.1109/IDAP.2019.8875902.

[2] M. Z. Gündüz and R. Daş, "Internet of things (IoT): Evolution, components and applications fields," *PAJES*, vol. 24, no. 2, pp. 327–335, 2018, doi: 10.5505/pajes.2017.89106.

[3] E. Ahmed *et al.*, "The role of big data analytics in Internet of Things," *Computer Networks*, vol. 129, pp. 459–471, Dec. 2017, doi: 10.1016/j.comnet.2017.06.013.

[4] S. T. Demirel, M. Demirel, I. Dogru, and R. Das, "InterOpT: A new testing platform based on oneM2M standards for IoT systems," in *2019 International Symposium on Networks, Computers and Communications (ISNCC)*, Istanbul, Turkey., Jun. 2019, pp. 1–6, doi: 10.1109/ISNCC.2019.8909198.

[5] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (IIoT): An analysis framework," *Computers in Industry*, volume. 101, pp. 1–12, October 2018, doi: 10.1016/j.compind.2018.04.015.

[6] G. Tuna, R. Das, B. Ramakrishnan, and Y. Kilicaslan, "Big data analysis for M2M networks: Research challenges and open research issues," *International Journal of Computer Networks and Applications*, vol. 4, no. 1, pp. 27–34, Feb. 2017, doi: 10.22247/ijcna/2017/41308.

[7] R. Daş and G. Tuna, "Machine-to-machine communications for smart homes," *International Journal of Computer Networks and Applications*, vol. 2, no. 4, pp. 196–202, 2015.

[8] L. Rajabion, A. A. Shaltooki, M. Taghikhah, A. Ghasemi, and A. Badfar, "Healthcare big data processing mechanisms: The role of cloud computing," *International Journal of Information Management*, vol. 49, pp. 271–289, Dec. 2019, doi: 10.1016/j.ijinfomgt.2019.05.017.

[9] G. Aceto, V. Persico, and A. Pescapé, "Industry 4.0 and Health: Internet of Things, Big Data, and Cloud Computing for Healthcare 4.0," *Journal of Industrial Information Integration*, vol. 18, p. 100129, Jun. 2020, doi: 10.1016/j.jii.2020.100129.

[10] Y. Bürhan and R. Daş, "Co-author link prediction from academic databases," *Gazi University, Journal of Polytechnic*, vol. 20, no. 4, pp. 787–800, Dec. 2017, doi: 10.2339/politeknik.368989.

[11] Y. Bürhan, M. Baykara, and R. Daş, "Sosyal ağ analizi ve veri görselleştirme araçlarının incelenmesi ve uygulamalı karşılaştırılması," in *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Inonu University, Malatya, Turkey., Sep. 2017, pp. 1–5, doi: 10.1109/IDAP.2017.8090295.

[12] R. Daş, I. Turkoglu, and M. Poyraz, "Web kayıt dosyalarından ilginç örüntülerin keşfedilmesi," *Firat University, Journal of Science and Engineering*, vol. 19, no. 4, pp. 493–503, 2007.

[13] R. Das and I. Turkoglu, "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6635–6644, Apr. 2009, doi: 10.1016/j.eswa.2008.08.067.

[14] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," *Computer Networks*, vol. 169, p. 107094, Mar. 2020, doi: 10.1016/j.comnet.2019.107094.

[15] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big Data Analytics for Dynamic Energy Management in Smart Grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, Sep. 2015, doi: 10.1016/j.bdr.2015.03.003.

[16] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-017-0111-6.

[17] M. Aledhari, M. Di Pierro, M. Hefeida, and F. Saeed, "A Deep Learning-Based Data Minimization Algorithm for Fast and Secure Transfer of Big Genomic Datasets," *IEEE Transactions on Big Data*, pp. 1–1, 2018, doi: 10.1109/TBDATA.2018.2805687.

[18] R. Zuech, T. M. Khoshgoftaar, and R. Wald, "Intrusion detection and Big Heterogeneous Data: a Survey," *Journal of Big Data*, vol. 2, no. 1, Dec. 2015, doi: 10.1186/s40537-015-0013-4.

[19] M. Baykara and R. Das, "A novel honeypot-based security approach for real-time intrusion detection and prevention systems," *Journal of Information Security and Applications*, vol. 41, pp. 103–116, Aug. 2018, doi: 10.1016/j.jisa.2018.06.004.

[20] M. Baykara and R. Das, "A novel hybrid approach for detection of web-based attacks in intrusion detection systems," *International Journal of Computer Networks and Applications*, vol. 4, no. 2, pp. 62–76, Apr. 2017, doi: 10.22247/ijcna/2017/48968.

[21] S. Sami and N. Sael, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data," *ijacsa*, vol. 7, no. 3, 2016, doi: 10.14569/IJACSA.2016.070337.

[22] S. E. Seker, "Büyük Veri ve Büyük Veri Yaşam Döngüleri," *YBS Ansiklopedi*, vol. 2, no. 3, p. 8, 2015.

[23] E. Aktan, "Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu," *Bilgi Yönetimi*, vol. 1, no. 1, pp. 1–22, Jun. 2018, doi: 10.33721/by.403010.

[24] R. Daş, D. Demirol, and G. Tuna, "A novel tool for mining access patterns efficiently from web user access logs," in *The International Conference on Engineering and Natural Sciences (ICENS) 2016*, Sarajevo, May 2016, pp. 2836–2843.
[25] M. M. Fouad, N. E. Oweis, T. Gaber, M. Ahmed, and V. Snasel, "Data Mining and Fusion Techniques for WSNs as a Source of the Big Data," *Procedia Computer Science*, vol. 65, pp. 778–786, 2015, doi: 10.1016/j.procs.2015.09.023.
[26] C. Kacfah Emani, N. Cullot, and C. Nicolle, "Understandable Big Data: A survey," *Computer Science Review*, vol. 17, pp. 70–81, Aug. 2015, doi: 10.1016/j.cosrev.2015.05.002.
[27] A. Duque Barrachina and A. O'Driscoll, "A big data methodology for categorising technical support requests using Hadoop and Mahout," *Journal of Big Data*, 1(1), p. 1, 2014, doi: 10.1186/2196-1115-1-1.
[28] C. Roy, S. Swarup Rautaray, and M. Pandey, "Big Data Optimization Techniques: A Survey," *IJIEEB*, vol. 10, no. 4, pp. 41–48, Jul. 2018, doi: 10.5815/ijieeb.2018.04.06.
[29] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.
[30] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, Dec. 2015, doi: 10.1186/s40537-014-0007-7.
[31] B. Daş, S. Toraman, and İ. Türkoğlu, "A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks," *Turkish Journal of Electrical Engineering & Computer Sciences*, 28(4), pp. 1932–1948, 2020.
[32] B. Das and I. Turkoglu, "A novel numerical mapping method based on entropy for digitizing DNA sequences," *Neural Computing and Applications*, 29(8), pp. 207–215, Apr. 2018, doi: 10.1007/s00521-017-2871-5.

Author

Bihter Das graduated B.S. and M.S. degrees from the Department of Computer Science at the Firat University in 2004 and 2007 respectively. Then she received Ph.D. degree at the Department of Software Engineering at the same university in 2018. She also worked between September 2017 and June 2018 as a visiting scholar at the Department of Computing Science at the University of Alberta, Edmonton, Canada. Her current research areas include data science, big data, data analytics, bioinformatic, digital signal processing, Genome data analysis.