BIG DATA SECURITY CHALLENGES: A SYSTEMATIC REVIEW

Elira Hoxha

Department of Statistics and Applied Informatics, Faculty of Economy, University of Tirana, Rr. Arben Broci, Tirana, Albania

Abstract

The large amount of data generated minute after minute from the everyday activities of the different companies or institutions and also from every individual using social media apps, entertainment apps, financial apps, and so on, gives the possibility to use lots of information to support and improve operational tasks and strategic decisions. This paper presents the concept of big data and the different fields where it is applied. Big data comes also with a lot of security challenges that is needed to investigate, coupled with the different ways that they can be addressed. The systematic review methodology is used to search for scientific papers in the most known online databases, through the addressed topics of interest. Then a qualitative analysis is performed to evidence the techniques that support managing and securing large data volumes, to be used by the researchers in the field.

Keywords: *big data, infrastructure security, data management, data availability, data integrity, data confidentiality*

1. INTRODUCTION

Business, industries, and governments, among other things, rely their success on accurate information and good decision-making. This information is gathered from different data sources and the amount of data generated increases with every activity done within the business, or toward the business. This makes it very difficult for the organizations to catch the relevant information and use it for their own benefit to support operational, tactical, and strategical decisions. So, the main aim of the big data concept is to identify the correct information in a timely fashion. The systems that enable the big data processes should be able to discover, access, store, retrieve and analyze these huge amounts of structured and unstructured data [1]. The definition of the big data concept is "a multidisciplinary and evolutionary fusion of new technologies in combination with new dimensions in data storage and processing, a new era of data source variety and the challenge of managing data quality adequately" [2]. It highlights the fact that big data has a multidisciplinary aspect, and can be ascertained that nowadays this concept is spread and used almost in every sector of our life by adding extra advantages to the organizations using it. Some of the industries benefiting from the big data analytics and other opportunities that come with this concept, are Education, Healthcare, Military, Government, Banking, Entertainment and Media, Transportation, Agriculture, Telecommunication, Manufacturing, Insurance, Sports, Energy, E-commerce, Etc. [3], [4], [5]. Some of the key benefits businesses realize by leveraging the potential of big data include [6], [7]:

- An improvement in understanding the customer behavior and motivations at different stages, by taking into consideration unanalyzed large amounts of data, which leads to better products and tailored marketing (e.g., industries of entertainment, e-commerce, and tourism).
- Automation of application performance management, network monitoring, and other network functions (e.g., industries that want to spend less time and effort on network security and operations practices, and more time on strategic projects).
- Improvement of fraud detection, risk management, and general compliance (e.g., industries of finance, banking, government, healthcare).
- Improvement of the supply chain efficiencies by tracing the products' path and conditions to attract customers' interest.
- Scanning and analyzing social media feeds, to better reach the sentiment of customers.

However, despite the many potentialities offered by big data, there are also challenges and vulnerabilities related to security over the whole lifecycle of storing, acquiring, and using the information obtained from big data sources. These security challenges evolve over time and are addressed differently by different researchers and security experts. Taylor [8] clearly defined the mission of big data security to keep out unauthorized users and intrusions by using firewalls, strong user authentication, end-user training, intrusion protection systems, intrusion detection systems, and data encryption. The idea shown in the study is to use a mix of traditional security tools, newly developed toolsets, and intelligent processes for monitoring security throughout the life of big data platforms. Other studies focus their work only on data privacy, as a crucial part of security issues related to big data, and give the readers the data encryption process, as a solution to bypass this challenge [9]. Kumar [7] organized the issues of big data security into the seven following categories: Data Storage, Fake Data, Data Privacy, Data Management, Data Access Control, Data Poisoning, and Employee Theft, with respective solutions such as the use of on-premise and Cloud Data Storage, employing cybersecurity experts, validating the data sources by periodic assessments and finding anomalies by evaluating Machine Learning models with different datasets, safeguarding the network against unauthorized access, assessing risks on a regular basis, training the users on Data Privacy and Data Security, encrypting the data, implementing a trusted server, monitor data sharing, etc. To have a better view and understanding of the most actual challenges and solutions regarding big data security, this study covers a systematic review of the latest research articles related to this topic.

2. METHODOLOGY

The methodology used in this paper is the systematic review which mostly consists of a protocol-driven comprehensive review and synthesis of data, focusing on a topic or on related key questions [10]. It involves a critical and reproducible summary of the results of the available publications on the chosen topic [11]. The review process will be completed to reduce biases and eliminate duplicate, irrelevant, and low-quality studies. The steps followed for this systematic review will include the formulation of the research question, the development of an appropriate protocol containing the inclusion and exclusion criteria for filtering the results, carrying out an exhaustive literature search based on relevant search strings in the most known databases of scientific papers, and in the end the qualitative screening of the abstracts and full papers of the most relevant studies identified during the previews steps.

The research question of our study is: “Which are the recent big data security challenges and the respective solutions”. The search strings that we will use for this systematic review are “big data” AND “security” AND “challenges”.

The databases taken into consideration for the literature search are

1. ACM digital library (ACM DL) [https://dl.acm.org/],
2. Springer Link [https://link.springer.com/],
3. Science Direct [https://www.sciencedirect.com/], and
4. IEEE Xplore [https://ieeexplore.ieee.org/].

These well-known databases were selected because they come up with a large number of relevant scientific papers for the review process. Google Scholar and Research Gate databases were excluded from the literature search because they did not provide a simple solution or user-friendly mechanism for extracting the search results in a timely fashion.

The inclusion criteria will be related to the publishing year. The filters of the database searches will be set to research articles published in the last 5 years (2017-2022).

The exclusion criteria are related to the language of the papers, where possible. The papers written not in English will not be accepted. Another exclusion will be related to the type of content of the search results. So, content types such as books, short papers, abstracts, etc. will be excluded from the list of research articles that will be examined in the steps that follow the systematic review. The last exclusion

criteria is the process of removing duplicate articles that resulted in the searches of the four different databases.

3. RESULTS AND DISCUSSION

Table 1 shows the number of results in each search database with the same search string: "big data" AND security AND challenges. The ACM DL which searches in a full-text collection of 660985 records, resulted in 2777 records after adding the filters for showing only "research articles" published in the "last 5 years" (2017-2022). The Springer Link database resulted in 7,887 records after adding the filters "article", within "English", and between "2017-2022" as the publishing year. Science Direct results with the highest number of "research articles" published in the interval "2017-2022" with 9,643 records, and IEEE Xplore with the lowest number (232 records) based on the filters "journal" and publishing year "2017-2022".

Search databases: / **Search string:**	**ACM DL**	**Springer Link**	**Science Direct**	**IEEE Xplore**
Big data security challenges	2777	7887	9643	232

Table 1. The number of records obtained from the online databases used

The total number of downloaded records from the four databases is 4234 because ACM DL allowed only the 2000 most relevant articles to be visualized in the search results, and Springer Link and Science Direct only the 1000 most relevant articles. The next step is removing the duplicates from the articles found in different databases. This resulted in 35 duplicates based on the title, authors, and publishing year, and another 7 duplicates based only on the title of the paper. As a result, remained 4191 unique research articles related to the research topic "big data security challenges". The next step is looking more closely at the title, abstract, and keywords of these records to find relevant studies for the qualitative analyses. The most interesting papers will be downloaded and scanned in detail to conclude the latest challenges and issues related to big data security and the different ways of avoiding or solving the respective problems.

Table 2 show some relevant studies related to our research topic. A qualitative analysis was made of the full papers to show an overview of the challenges and the different ways of addressing them.

Study number	**Title**	**Year**	**Authors**
st_1	Secure big data ecosystem architecture: challenges and solutions [12]	2021	M. J. Anwar, A. Q. Gill, F. K. Hussain, M. Imran
st_2	Security Challenges for Big Data and IoT [13]	2020	S. Venkatraman, A. Overmars, K. Fahd, S. Parvin, S. Kaspi
st_3	Big Data: Trade-off between Data Quality and Data Security [14]	2019	M. Talha, A. Abou El Kalam, N. Elmarzouqi
st_4	Secure Big Data using QKD protocols [15]	2019	H. Amellal, A. Meslouhi, A. El Allati
st_5	Encryption as the cornerstone of big data security [16]	2017	C. Tankard
st_6	A Review on Big Data security and handling [17]	2017	N. Bhogal, Sh. Jain
st_7	Securing Big Data: A Survey on Security Solutions [18]	2022	H. Habbak, K. Metwally, A. M. Mattar

st_8	The 10 Vs, Issues and Challenges of Big Data [19]	2018	N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. Abbasi, S. Salehian, Solmaz
st_9	Big Data Analytics – Security and Compliance Challenges in 2019 [20]	2019	M. Small

Table 2. List of papers relevant to the research topic, obtained from the qualitative search.

Source: Author, June 2022.

Anwar et al. [12] identified through a systematic literature review the following big data privacy and security challenges (st_1):

- Inference
- Lack of consent
- Social misuse of knowledge
- Diverse data sources
- Multiple uses of data
- Storage and processing
- Technology gap
- Agreed data usage
- Unauthorized Access
- Lack of governance
- Data provenance

and adapted them in a theoretical framework (Figure 1) to match each category of challenge with the proper solution:

- User Validation
- Skills
- Access Control
- User Education
- Audits
- Communications Security
- Threat Modelling
- Risk Assessment
- Data Classification
- Data-driven privacy-preserving
- Privacy at Social Networks
- Endpoint validation and encryption
- Secure Queries
- Differential Privacy
- Latent Data Privacy
- Secure Data Collection
- Storage & Transformation
- Machine Learning Algorithms
- Intrusion Recognition
- Data Anonymization
- Distributed sources for Data Backup and Recovery
- Storage of Encrypted Header Information
- Governance
- Legal Support

Venkatraman et al. [13] recommend using NoSQL databases over relational ones when dealing with big data and consider the security challenges based on this type of database (st_2). The main solution offered in this study is the use of key distribution encryption for preserving the data. A similar solution based on encryption (st_5), is shown by Tankard [16], meanwhile, Amellal et al. [15] studied big data security based on quantum information theory by discussing the opportunities offered by quantum key distribution protocols and quantum algorithms to increase the security level of big data (st_4). Talha et al. [14] and Small [20] considered the data security and privacy challenges of big data in four directions (st_3) (st_9):

- Confidentiality: treated with access control and cryptographic mechanisms.
- Integrity: addressed through data reliability assurance and the establishment of mechanisms for detection and prevention of data loss.
- Availability: addressed by applying multiple data replications.

- Privacy: addressed through detection and prevention of re-identification processes, preservation of analytical results, compliance with regulatory and legal constraints, etc.

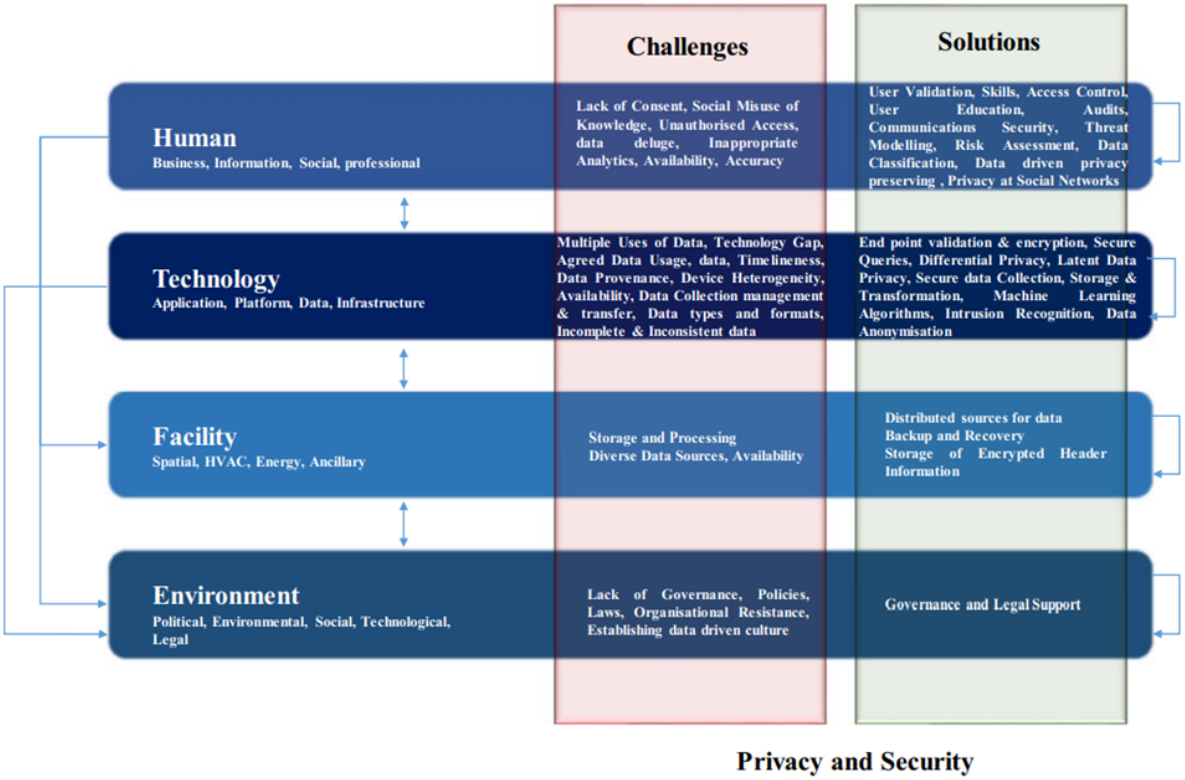


Fig. 1. Adaptive enterprise architecture theoretical framework [12].

Source: Anwar et al. J Wireless Com Network 2021, 130 (2021)

Small [20] recommended these three key pillars to be followed:

- Ethics and compliance: The security and compliance controls must be data-centric – they must secure the data wherever it is stored and processed.
- Data management: The lifecycle of Big Data must be properly managed.
- Infrastructure: The infrastructure for processing Big Data provides the foundation for its security and accurate analysis.

Bhogal and Jain [17] analyzed a long and detailed list of security challenges by focusing on four main pillars: Infrastructure security, Data privacy, Data management, Integrity, and reactive security (st_6). Each one of these categories was then expanded into finer challenges and explained in detail related to several controls for Application Software Security; Maintenance, Monitoring, and Analysis of Audit Logs; Secure Configurations for Hardware and Software; and Account Monitoring and Control.

Habbak et al. [18] focused their work on the confidentiality, integrity, and availability of big data (st_7). Apart from access control and encryption techniques that are suggested by many researchers they mention neural network-based cryptography techniques, quantum cryptography, and a blockchain trust-management approach as new proposals to address big data security.

Khan et al. [19] explored the security issues and challenges focusing on the 10Vs characteristics of big data: Volume, Value, Velocity, Veracity, Viscosity, Variability, Volatility, Viability, Validity, and Variety (st_8).

From a qualitative analysis on the evidenced studies the challenges of security in big data can be grouped on those related to distributed data, the usage of databases, infrastructure vulnerabilities, data mining algorithms and access control. The security solutions suggested start from the use of encryption, user access control techniques, intrusion detection and prevention, to more recent ones such as quantum cryptography, blockchain technology, or the use of artificial intelligence, to identify patterns and trends to predict future attacks on data.

4. CONCLUSIONS

The main aim of this study was to investigate the literature for summarizing the current state of big data security challenges by using a systematic review methodology. From the searches in four well-known databases, and by using inclusion and exclusion criteria, 4191 unique research articles, published in the last five years (2017–2022), were extracted. Then a qualitative analysis was performed, where 9 studies are identified to explore the existing challenges of privacy and security in the big data world. In particular, it can be highlighted that the focus of most of the studies is challenges related to confidentiality, integrity, and availability, and their suggested solutions such as the various methods of data encryption, the different techniques for access control, and dealing in particular with infrastructure security.

The results of this research work should be considered with respect to some limitations of the study. So, there were used only four databases to research the topic. Although these scientific databases are well known for their extensive coverage of the existing literature and are still other valuable resources for research, such as Google Scholar, Research Gate, ProQuest Science and Technology, JSTOR, Crossref, etc. Another limitation is related to the few search strings that were used. This study has a large scope, and some relevant findings might have been excluded from the results due to the limited number of search strings. The third limitation is related to the fact that we did not examine the results divided into the theoretical and practical perspectives, which might have been useful for the researchers of the field.

However, the field of big data can still be considered emerging, and a lot of work needs to be done in the research area to report empirical solutions for many of the mentioned challenges and issues of privacy and security, and to evidence more practical security and privacy implementations in big data environments.

REFERENCES

1. Almeida, F., 2018. Big Data: Concept, Potentialities and Vulnerabilities. Emerging Science Journal, 2(1).
2. Buhl, H., Röglinger, M., Moser, F. and Heidemann, J., 2013. Big Data. Business & Information Systems Engineering, 5(2), pp.65-69.
3. Techlabs, M., 2017. Top 10 sectors making use of Big Data analytics. [online] Medium. Available at: <https://towardsdatascience.com/top-10-sectors-making-use-of-big-data-analytics-be79d2301e79> [Accessed 1 June 2022].
4. Vidvan, T., 2017. Top 15 Big Data Applications Across Industries. [online]. Available at: <https://techvidvan.com/tutorials/big-data-applications-across-industries/> [Accessed 1 June 2022].
5. Shah, D., 2022. Top 20 Big Data Applications In Different Sectors (2022). [online] Software Test Tips. Available at: <https://www.softwaretesttips.com/top-big-data-applications/> [Accessed 1 June 2022].
6. Hiter, Sh., 2021. Big Data Market Review 2022 | Big Data Definition & Analytics. [online] Datamation. Available at: <https://www.datamation.com/storage/big-data-market/> [Accessed 1 June 2022].

7. Kumar, D., Sarin, P. and Varshney, H., 2021. Top 7 Big Data Security Issues and Their Solutions - Learn | Hevo. [online] Learn | Hevo. Available at: <https://hevodata.com/learn/big-data-security/> [Accessed 1 June 2022].
8. Taylor, Ch. 2021. What is Big Data Security? Challenges & Solutions in 2022. [online] Datamation. Available at: <https://www.datamation.com/big-data/big-data-security/> [Accessed 2 June 2022].
9. Kszczanowicz, L. and Walat, P., 2021. Big data security – what are the best solutions for business? | DS Stream. [online] Data Analytics. Available at: <https://dsstream.com/the-best-solutions-for-big-data-security/> [Accessed 2 June 2022].
10. Russell, R., Chung, M., Balk, E., Atkinson, S., Giovannucci, E., Ip, S., Mayne, S., Raman, G., Ross, A., Trikalinos, T., Keith P West, J. and Lau, J., 2009. Systematic Review Methods. [online] Ncbi.nlm.nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK44088/> [Accessed 4 June 2022].
11. Linares-Espinós, E., Hernández, V., Domínguez-Escrig, J., Fernández-Pello, S., Hevia, V., Mayor, J., Padilla-Fernández, B. and Ribal, M., 2018. Metodología de una revisión sistemática. *Actas Urológicas Españolas*, (Engl Ed). 42(8), pp.499-506.
12. Anwar, M., Gill, A., Hussain, F. and Imran, M., 2021. Secure big data ecosystem architecture: challenges and solutions. *EURASIP Journal on Wireless Communications and Networking*, 2021(1).
13. Venkatraman, S., Overmars, A., Fahd, K., Parvin, S. and Kaspi, S., 2020. Security Challenges for Big Data and IoT. *Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology*.
14. Talha, M., El Kalam, A. and Elmarzouqi, N., 2019. Big Data: Trade-off between Data Quality and Data Security. *Procedia Computer Science*, 151, pp.916-922.
15. Amellal, H., Meslouhi, A. and Allati, A., 2019. Secure Big Data using QKD protocols. *Procedia Computer Science*, 148, pp.21-29.
16. Tankard, C., 2017. Encryption as the cornerstone of big data security. *Network Security*, 2017(3), pp.5-7.
17. Bhogal, N., Jain, Sh., 2017. A REVIEW ON BIG DATA SECURITY AND HANDLING. *International Research Based Journal*, Vol(6)-Issue(1), ISSN 2348-1943, March 11, 2017.
18. Habbak, H., Metwally, K., Mattar, A. M., (2022). Securing Big Data: A Survey on Security Solutions. *TechRxiv*. Preprint.
19. Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. and Salehian, S., 2018. The 10 Vs, Issues and Challenges of Big Data. *Proceedings of the 2018 International Conference on Big Data and Education*, 2018, pp. 52–56.
20. Small, M., 2019. Big Data Analytics - Security and Compliance Challenges in 2019. *CuppingerCole Report*. Whitepaper, April 2019.