JCBI
Journal of Computing & Biomedical Informatics Volume 05 Issue 02
ISSN: 2710 - 1606 2023

Review Article
https://doi.org/10.56979/502/2023

Big Data Privacy Issues and Challenges: A Survey

Benish Khalid[1], and Irshad Ahmed Sumra[1*]

[1]Department of Information Technology, Lahore Garrison University, Lahore, Pakistan.
[*]Corresponding Author: Irshad Ahmed Sumra. Email: irshadahmed@lgu.edu.pk

Received: June 11, 2023 **Accepted:** August 26, 2023 **Published:** September 17, 2023

Abstract: Big data, known for its intricate and diverse nature, entails handling vast amounts of information. This poses substantial challenges in storing, analyzing, processing, and extracting meaningful insights. The surge in big data's significance has led to its extensive exploration as an academic field. However, with the increasing volume of data comes a heightened risk of privacy infringement. Consequently, numerous privacy-preserving methodologies have been developed and implemented across various stages of data handling. This paper provides a comprehensive and review of research dedicated to preserving privacy in the realm of big data. Notably, it concentrates on talking about privacy strategies designed to protect data throughout critical processes including storage, production, and processing. The major goal of this paper is to examine prior studies in big data privacy.

Keywords: Extensive Data, Data Handling, Data Storage Management, Governance of Data, Data Integrity, Data Sharing and Protection, Privacy and Security.

1. Introduction

When dealing with complicated data sets that are excessively large for typical database systems to manage, big data is used [1-3]. It's a collection of Data that has a large volume- both structured and unstructured data. due to recent advancements in technology, the volume of data generated by social networking platforms, the internet, sensors, and various businesses is growing incessantly. The term "big data" encompasses a new era of technologies that facilitate the rapid capture, exploration, and analysis of data, enabling cost-effective extraction of value from extensive and diverse datasets [4, 5]. In accordance with the principles outlined in this definition, the characteristics of big data, often referred to as the 3Vs—velocity, volume, and variety—are identified. However, it has been recognized through subsequent research that these 3Vs alone are inadequate in elucidating the vast and intricate nature of the data we presently encounter. To offer a more comprehensive understanding of big data, attributes such as veracity, validity, value, variety, venue, and terminology have been introduced [6]. This study delves into the three primary aspects of big data, commonly known as the three Vs: volume, velocity, and variety. Volume pertains to the sheer quantity of data generated, and we have observed a significant increase in data size since the advent of social networking sites. On the other hand, velocity describes the rate at which new data is generated. A frequent element in the field of big data is the diversity of data, which can include text, audio, photos, or videos. This diversity is embodied by the concept of variety. To manage the massive influx of data coming in at unusually high rates from multiple sources and managing the varied aspects of big data in terms of volume, velocity, and variety, it is essential to design reliable and effective frameworks. The life cycle of big data necessitates the traversal of several stages.

Big data can be transformed into valuable insights that can be put to use if it is collected and analyzed in a timely manner. Through data analysis, it can assist companies and organizations in strengthening their internal decision-making capacity and opening new prospects. By changing established business structures and scientific principles, it can also support economic growth and scientific research [7].

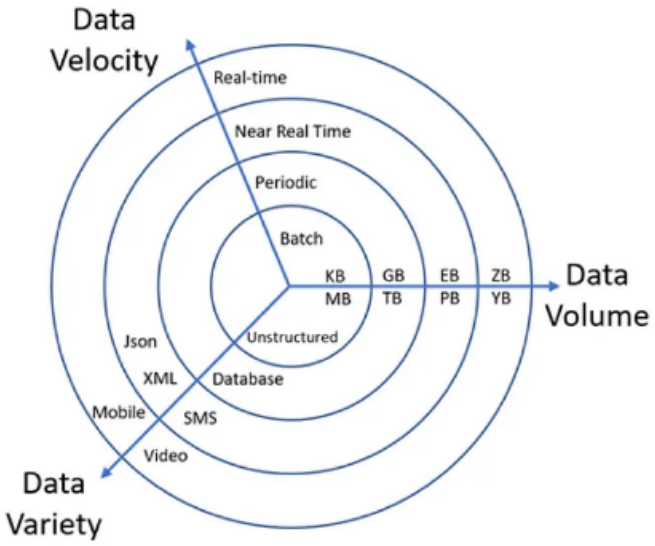

Figure 1. 3 Vs representation of Big Data [45].

Numerous tactics have been developed in recent years to protect the privacy of large data sets. The stages of the Big Data life cycle, as shown in Figure 2, which include data generation, storage, and processing [8], can be used to classify these mechanisms.

Figure 2. Stages of Life Cycle of Big Data.

During the data generation phase, privacy is upheld through the implementation of access control and data falsification techniques. In the context of data storage, privacy is safeguarded through the utilization of encryption methods. To protect data privacy during processing, anonymization techniques like generalization and suppression are employed.

The year 2021 is projected to witness a staggering daily data production of 2.5 quintillion bytes, reflecting the continued growth of large data volumes and rapid data generation [9]. In this contemporary digital era, where substantial information is housed within big data repositories, the study of databases provides a means to address significant issues, including those related to healthcare and beyond.

This study provides a thorough analysis of the methods used to protect privacy in the Big Data space. We thoroughly cover the important facets of privacy and security issues regarding big data. In the parts that follow, it will review into detail about the privacy requirements that are particular to different stages, including data generation, storage, and processing.

Additionally, we compare and contrast current research on protecting privacy in the context of big data to show the developments in this quickly developing area.

2. Infrastructure of Big Data

To analyze the enormous amounts of data arriving quickly from various sources, efficient and effective solutions are required to manage the multi-dimensional features of big data, spanning volume, velocity, and variety. As shown in Figure 2, there are various stages in the life cycle of massive data. Due to the present dissemination of data, modern technology is developing to support the storage and management of enormous data warehouses. Big data processing and storage options on the cloud, such as Hadoop Map Reduce, are being investigated. This section will delve into the big data lifecycle, shedding light on how

big data leverages cloud technology and the challenges that arise when storing and handling large datasets in such environments.

2.1. Big Data's Lifecycle

2.1.1. Data generation

Data is capable of being produced by a variety of scattered sources. Data produced by both machines and people has increased recently. For instance, 90% of the 2.5 quintillion bytes of data produced daily on the internet was just recently created. One social networking site, Facebook, alone produces 27TB of fresh data each day. The resulting information is frequently vast, varied, and complex. As a result, handling them is difficult for conventional systems. The material produced frequently has to do with a certain subject, such business, the internet, research, etc.

2.1.2. Data processing

The operations of data collecting data transfer, pre-processing, and information extraction are commonly referred to as the data processing phase. Data gathering is important since information can come from a variety of sources, including websites containing text, photographs, and videos. In the data collection phase, information is collected from a designated data production environment through specialized data collection technologies. Once raw data is acquired from this specific production environment, an efficient high-speed transmission mechanism becomes necessary to transfer it to a suitable storage infrastructure, catering to various analytical applications. To optimize storage resources, the pre-processing stage is dedicated to the removal of redundant and unnecessary data elements, thereby conserving storage space.

2.1.3. Data storage

Large datasets are managed and stored during this stage. Both physical infrastructure and data management components are included in the data storage system. Distributed storage is one of the components of the physical infrastructure, or information and communications technology (ICT) resources, which performs a variety of tasks. To facilitate the management and querying of large data collections, a suite of software is implemented atop the hardware infrastructure. This software must provide diverse interfaces for interacting with and analyzing the stored data.

3. Big Data Privacy and Security Issues

Regarding **Privacy**, we mean the freedom you have over how your personal information is utilized. The restriction of who you wish to share your data with is related to privacy. A serious privacy risk is the identification of personal information as it is being transmitted over the internet [10].

Security refers to how your personal information is protected. Security includes protecting your data from attacks that may destroy data. Security is a practice of defending your data from unauthorized access and Destruction.

Security vs. Privacy: While data security focuses on securing data against threats, data privacy emphasizes using data responsibly. Additional distinctions between privacy and security are highlighted in Table 1.

Table 1. Difference between Security and Privacy

Security	Privacy
Data storage is protected from unauthorized use and access.	The responsible handling of user information is privacy.
Security involves ensuring the "confidentiality, integrity, and accessibility" of data.	Using and disclosing personal information in authorized manner.
Confidentiality may be supported by security. Most security systems are designed to safeguard an organization or agency from outside threats.	The issue of privacy often pertains to an individual's entitlement to safeguard their personal information.

4. Privacy Needs in the Context of Big Data

While big data analytics are appealing to many organizations, a sizable portion choose not to use them since there are no standardized security and privacy safeguards in place. In these sections, we examine potential methods for adding privacy protection features to large data platforms. This involves elucidating the fundamental principles and expansion strategies of an architecture that facilitates:

1. Establishing privacy guidelines that dictate user access to data within the designated big data platforms.
2. Creating effective enforcement mechanisms to uphold these policies.
3. Integrating the Target Analytic Platforms with the resulting monitoring systems.

4.1. Big data privacy during the data generating phase

Active data generation and passive data production are the two types of data generating. When data is generated passively, it is a result of the data owner's online behavior, frequently without the data owner's knowledge that a third party is collecting this data [11]. Conversely, active data generation occurs when the data owner intentionally provides their data to a third party. Several strategies are employed to mitigate the risk of privacy breaches, including:

4.1.1. Access Restriction

The data owner refrains from disclosing the data due to concerns that it may unveil sensitive information that should remain confidential. Several steps, including encryption, might be implemented to preserve privacy if the owner provided the data. One can properly restrict access to sensitive data by using these methods. Most of these utilities are created as browser extensions for simplicity of usage.

4.1.2. Falsifying Data

In certain cases, it may be impractical to completely block access to sensitive information. In such scenarios, data can be altered using tools before it is accessed by a third party. Manipulating the data makes it challenging to discern the true information. Two tools for data falsification are the "socket puppet" and "Mask ME."

4.2. Big data privacy during the data storage phase

Thanks to advances in data storage technology, notably the ascent of cloud computing, the storage of vast volumes of data is no longer a formidable challenge [12]. Data privacy is difficult to maintain, though. A distributed environment presents the difficulty of privacy protection because an application may require many datasets from different data centers [13].

Traditional security measures employed to protect data can be categorized into four groups: application-level encryption schemes, database-level encryption schemes, media-level encryption schemes, and file-level encryption techniques [14]. Given the nature of the 3Vs in big data analytics, the storage infrastructure must exhibit scalability and dynamic configuration support for various applications. Storage virtualization, facilitated by the expanding cloud computing paradigm, emerges as a promising technology to fulfill these requirements [15]. Multiple network storage devices are combined into what seems to be a single storage unit through storage virtualization. But using a third-party provider's cloud service necessitates handing the organization's data to an external party, which raises questions regarding data privacy. Therefore, this paper will specifically address the privacy of data stored within the cloud.

In a distributed environment, an application may need many datasets from different data centers, which provides a challenge for privacy protection.

The four categories of traditional security measures for data protection are as follows. These include media-level encryption techniques, database-level encryption techniques, file-level data security techniques, and media-level security techniques.

The necessity for scalability in the storage infrastructure arises from the workings of big data analytics characterized by the 3Vs. To accommodate diverse applications, this infrastructure must also offer dynamic customization. Storage virtualization is a technique that amalgamates multiple network storage devices, creating the illusion of a single storage unit. However, opting for a cloud service from an external provider entails the outsourcing of an organization's data, potentially impacting data privacy. Consequently, this paper will focus specifically on the privacy of data stored in the cloud. The ensuing methods are employed to ensure user privacy when data is stored within cloud environments.

4.2.1 Attribute Based Encryption

In cloud storage systems, complete big data privacy is guaranteed thanks to a type of encryption called attribute-based encryption. According to the access regulations set out by the data owner, data is encrypted in ABE. Only users whose attributes comply with the access criteria established by the data owner can decrypt the data. Big data frequently necessitates changing data access rules because the owner of the data may need to share it with numerous businesses.

4.2.2 Homomorphic Encryption

Public clouds' inherent multi-tenancy and virtualization features make them more susceptible to privacy violations. The potential for sharing physical infrastructure with other cloud users heightens the risk of data leakage. One approach to protect data in the cloud involves encrypting it before storage and allowing the cloud to process encrypted data. Fully homomorphic encryption is a type of encryption that enables computations to be performed on encrypted data.

4.2.3 Identity Based Encryption:

IBE and ABE encryption techniques do not support the updating of the cypher text recipient. There are several techniques to update the cypher text's recipient. Owners of data can, for instance, re-encrypt after decrypting. However, due to computation overhead, it may be extremely time-consuming and expensive to decrypt and re-encrypt large data sets, which is often the situation when working with big data. Additionally, to use this option, the data owner must always be online. Assigning this task to a trustworthy third party who is aware of the data owner's decryption key is another way to update the receiver of the cypher text. This method has a few flaws, including the fact that it depends entirely on the third party's confidence and that it makes it impossible for the recipient of the cipher text to remain anonymous because the third party requires this knowledge to proceed with the re-encryption.

4.2.4 Storage Path Encryption:

Big data is first broken up into many sequential parts, and each part is then saved on a unique storage media owned by several cloud storage providers. The data must first be gathered from numerous data centers, restored to its former state, and then provided to the data owner to be accessed. This architecture separates the categories for private and public data in the cloud-based big data storage. Anyone can access public data without restriction and no additional security measures are necessary. On the contrary, sensitive information is consistently safeguarded and inaccessible to unrelated individuals and organizations. It involves a computation process that is easily performed in one direction but challenging to reverse without supplementary information, a characteristic often used in cryptographic applications known as trapdoor features. In the proposed method, the entirety of massive data is not encrypted; instead, only the storage path, sometimes referred to as the cryptographic virtual mapping of extensive data, is subjected to encryption. Additionally, certain designated data considered confidential is encrypted within this method. To enhance the availability and resilience of big data, the strategy involves storing duplicate copies of each data piece in cloud storage. In case of data loss, this method allows for the retrieval of backup copies. The storage index data will be maintained by the owner of the big data.

4.3. Big data privacy preserving in data processing

The big data processing paradigm categorizes processing systems into batch, stream, graph, and machine learning processing [16, 17]. To ensure privacy protection, the data processing aspect can be divided into two phases. Given that the acquired data may include sensitive information owned by the data proprietor, the first phase aims to safeguard the data from unauthorized disclosure. The second phase's objective is to extract valuable insights from the data while preserving privacy.

4.3.1 Anonymization Techniques

In accordance with the designated privacy standards, the initial table is modified before its publication. The data underwent one of the following anonymization processes to uphold privacy.

Generalization: operates by substituting a specific Quasi-Identifying (QID) feature's value with a broader and less precise description. In this technique, certain attribute taxonomy values are replaced with their parent values. For instance, one might use the term "artist" to represent the job attribute instead of specifying "singer" or "actor." There are various approaches to generalization, including whole domain generalization, sub-tree generalization, multidimensional generalization, sibling generalization, and cell generalization, among others.

Suppression: Some values in suppression are replaced with a special character, like "@," to signify that the replacement value is not revealed. Record suppression, value suppression, and cell suppression are a few examples of suppression schemes.

Anatomization: In two independent tables, the information on QID and SA is made available in this way. While one table includes quasi-identifiers, both tables have sensitive characteristics. Group ID is a feature that regularly appears in both tables. The same Group ID value will serve as the connection point for all the sensitive numbers in a group.

Permutation: A dataset is segmented into groups, and within each group, the sensitive values are shuffled randomly. This process is executed to break the linkage between a quasi-identifier and a numerically sensitive attribute.

4.3.2 Privacy-Utility Trade-Off

Data anonymization is a positive indicator of robust privacy protection. However, it can potentially impact the data's utility, leading to reduced inferential value. Consequently, it is crucial for big data applications to strike a balance between privacy and utility. Various methods for assessing information loss have been proposed in the literature, with some focusing on minimal distortion. Privacy and utility trade-off challenges are often addressed by Privacy-Preserving Data Publishing (PPDP) algorithms that employ a greedy strategy to identify the optimal trade-off point. Using established metrics for privacy preservation and information loss, these algorithms generate multiple tables during the anonymization process, each satisfying privacy model criteria. The table with the least amount of information loss is the result of the greedy algorithm.

The task of calculating privacy is incredibly difficult. Imagine a situation where a data item is obtained from a data owner. The data owner has complete discretion over the quantity and kind of information that is shared with a third party. There may be some privacy loss when data is shared to a third party. The same data may be sent to the third party by various data owners. However, when it is released, certain people who take privacy seriously can suffer a bigger loss than those who don't.

5. Security and Privacy Aspects in Healthcare Big Data

Healthcare organizations deal with substantial data volumes for the purpose of enhancing the provision of effective and tailored treatment. However, they face challenges such as inadequate technological support and insufficient security measures. Adding to the complexity is the healthcare sector's susceptibility to highly publicized data breaches. Data breaches can occur if attackers employ data mining techniques to locate sensitive information and disclose it to the public. Despite the ongoing efforts to implement security measures, the risks are escalating as security controls become increasingly difficult to circumvent [18]. The Big data offers a thorough analysis of the many methods and strategies employed in ubiquitous healthcare in a disease-specific way. It covered a significant problem or disease that may be subtly identified and treated using technology, such as stress, cardiovascular disease, deadly and non-fatal falls.

1. Embracing a value-based business model through healthcare analytics necessitates prioritizing data governance as the initial step in overseeing and managing healthcare data.
2. The objective is to establish a unified data representation framework that incorporates industry standards along with regional and local standards.

“Privacy preserving analytics.”

1. The encroachment upon patient privacy looms as an escalating concern within the realm of big data analytics.
2. Privacy-preserving encryption strategies, which enable the execution of predictive algorithms on encrypted data while safeguarding patients' identities, are indispensable for propelling healthcare analytics forward [19].

“Data quality”

1. Health data typically originates from disparate sources with widely varying setups and database designs, resulting in complex, unclean data rife with gaps and disparities in coding standards for identical fields.
2. The era of problematic handwriting in Electronic Health Record (EHR) systems is over. Data collected via these systems primarily serves clinical purposes, leading to issues such as missing data, inaccuracies, miscoding due to clinicians' heavy workloads, non-user-friendly interfaces, and a lack of human validation checks [20].

“Data sharing and privacy”

1. Personal health information (PHI) inherent in health data poses legal obstacles to data access due to privacy invasion risks.
2. Anonymizing health data through masking and de-identification techniques facilitates disclosure to researchers under legally binding data sharing agreements [21].

“Relying on predictive models.”

1. It is essential to maintain realistic expectations regarding the performance of constructed data mining models, recognizing that every model has inherent accuracy limitations.
2. Importantly, relying solely on predictive models for critical decisions directly impacting patients' lives is perilous and should not be an anticipated outcome.

"Variety of methods and complex math's"

For data analysts, familiarity with diverse techniques and accuracy metrics is imperative to apply multiple methods effectively when analyzing specific datasets.

6. Recent Studies in Big Data Privacy

As a result of the proliferation of big data in almost every industrial area over the last decade, new methodologies for data analytics have been developed. Even though many organizations are aware of the promise of big data analytics, many are still in the early stages of reaping its advantages because of the large and fast-growing amount of big data. The Cloud computing, on the other hand, comes with security and privacy concerns. To secure the integrity of data in cloud and IoT contexts, researchers conducted research. Security and privacy concerns in big data systems are getting more difficult as dispersed and heterogeneous settings become more prevalent, requiring additional study. Increasingly lately, a lot of study has focused on the health business since big data is becoming more popular in electronic health care services. However, with the rise of social and internet-of-things (IoT) networks, the issue of large data security is now a top priority for many governments. The Literature evaluation of recent big data publications from the perspectives of security and privacy was recently conducted using Scopus database articles from the major scientific peer-reviewed journals. According to the findings of this research, the computational elements of big data security and privacy are receiving considerable attention. For the sake of this study, privacy, data analytics, and confidentiality are the most significant research categories and subjects being examined. There have been just a few studies that have presented an overview of big data, and they have proposed new avenues of study in the field of security and privacy. Big data security problems have not been considered by these organizations.

Table 2 presents a summary of recent methodologies and strategies related to privacy and security in the context of big data.

Table 2. Overview of Recent Studies on Privacy in the Big Data Context

Paper	Year	Focus	Limitations
[22]	2014	How location-based services and social networks, two developing applications, implement privacy-preserving data publication (PPDP).	Instead of illustrating the entire application chain, concentrate solely on depicting common issues.
[23]	2014	Reviews and assessments of privacy-preserving data mining techniques.	The study also provides an analysis and comparison of privacy-preserving techniques in the domains of clustering and association rule mining.
[24]	2014	The proposed technique for privacy-preserving data mining in Hadoop aims to address privacy breaches without compromising data utility.	The execution time of the proposed technique is influenced by the magnitude of noise.
[25]	2014	Investigate the difficulties and insights encountered while merging privacy protection through anonymization with big data approaches to analyze usage data while preserving individual identities.	Utilizes the K-anonymity technique, which exhibits susceptibility to correlation attacks.
[26]	2014	Suggested a scalable method employing a two-phase top-down specialization (TDS)	Deploys an anonymization method that is susceptible to collision attacks.

		strategy for anonymizing extensive datasets.	
[27]	2014	Share an algorithm for anonymizing big data streams with enhanced speed.	Additional research is necessary to develop and execute FAST within a distributed cloud-based framework, aiming to harness the computational capabilities of the cloud and attain substantial scalability.
[28]	2015	Suggested an approach for privacy-preserving data mining on extremely large datasets using the MapReduce framework.	Data utility is decreased by generalization's difficulties with high-dimensional data as well as by disturbance.
[29]	2015	Put focus on key management for Hadoop, cloud security, monitoring and auditing, and anonymization.	Discuss all limitations and issues in these security techniques.
[30]	2015	Suggested a mechanism for privacy-preserving ciphertext multi-sharing.	It's possible to establish delegation rights between two parties even if they have not previously reached an agreement on the delegation process.
[31]	2015	The training data is spread, and each component shares a fraction of the large dataset, in this ground-breaking approach for privacy-preserving machine learning.	Unable to attain distributed feature selection.
[32]	2015	Create a proximity-aware clustering problem to describe the difficulty of maintaining proximity privacy in local large data recording, and then propose a scalable two-phase clustering approach as a solution.	For the research to integrate over approach with Apache
[33]	2016	Identified multiple privacy concerns.	Customer segmentation have the potential to inadvertently result in discrimination based on factors such as age, gender, ethnic background, health condition, social status, and more.
[34]	2016	Suggested a pragmatic scheme for handling encrypted big data in the cloud, integrating deduplication while addressing ownership challenges and incorporating Proxy Re-Encryption.	Convergent encryption possesses an inherent security limitation, specifically its vulnerability to offline attacks.
[35]	2016	Emphasize the security management platform, the information security system, and the pertinent laws and regulations.	Assemble the massive data Information security system setup, law and regulatory implementation, and security management platform setup.

[36]	2018	This paper introduces key concepts related to sensitivity and privacy budget within the context of differential privacy. It discusses noise-based techniques employed in differential privacy, composition aspects, and methods to attain it.	This study does not explore efficient approaches for DP implementation in real-world scenarios, cryptography techniques to DP, or a variety of other topics.
[37]	2019	This study examines the function of differential privacy in big data correlated datasets, focuses on an enhanced differential privacy method that may be used with correlated datasets, and suggests and evaluates novel models and algorithms.	This paper did not emphasize the external correlation between the datasets. Additionally, in the context of the k-means algorithm, selecting the appropriate k-value can be challenging, prompting the exploration of alternative optimization clustering algorithms to address big data challenges.
[38]	2020	The primary focus of this study revolves around the novel Privacy Preservation Algorithm for Big Data Using Optimal Geometric Transformations, referred to as PABIDOT.	Model's efficiency and scalability needs to be checked on big datasets.
[39]	2020	This paper primarily tackles security and privacy concerns arising from data virus propagation within big data networks. A Protection and Recovery Strategy (PRS) with the goal of minimizing infections and enhancing immunity within the network is also presented.	Reliance on a simplified infectious disease model, complexity of the incentive mechanism, lack of real-world deployment analysis, and a narrow focus on virus propagation without addressing broader security aspects.
[40]	2021	This paper reviews the literature on COVID-19-based big data analysis, highlighting research contributions and proposing a taxonomy of applications for managing and controlling the pandemic.	The paper suggests further exploration of its scope, patient perspective, and ethical considerations for a comprehensive understanding of COVID-19 data analysis applications.
[41]	2021	This paper proposes an efficient perturbation algorithm using optimal geometric transformation to address privacy concerns and ensure big data utility.	Lack of thorough exploration of utility-privacy trade-offs and limited discussion on the method's adaptability to diverse datasets or scenarios.
[42]	2022	To enhance the security of big data during transmission over a network by integrating compression, splitting, and encryption	Lacks in-depth quantitative analysis or experimental validation of the

		mechanisms while maintaining performance and reliability.	proposed mechanisms, and it does not address potential challenges or drawbacks associated with compression and splitting methods for big data security.
[43]	2022	To discuss and propose various methods for preserving privacy in the context of deep learning (DL) applied to big data analysis. The paper ultimately aims to present effective solutions for enhancing privacy preservation within DL models for big data analysis.	The paper does not explicitly outline its own limitations or potential challenges, which could include the need for addressing practical implementation constraints, scalability issues, and the trade-off between privacy and utility in differential privacy preservation methods for deep learning.
[44]	2023	With an emphasis on privacy and security issues as well as DP's advantages over alternative data privacy protection solutions, this article examines and analyses differential privacy (DP) methodologies used in edge computing-based smart city applications.	Lacks discussion on the practical challenges and potential obstacles that may be encountered when implementing the identified future directions for applying differential privacy (DP)

7. Conclusion

The use of big data technology in predicting future trends has gained popularity. However, privacy considerations are crucial in this data-driven environment. This survey carefully discusses privacy issues across the many phases of the big data life cycle. Notably, the study also explores the privacy issues that arise in the crucial field of healthcare. It thoroughly discusses the numerous privacy measures used during the phases of big data collection, processing, and storage. It gives a comprehensive survey, neatly presented in tabular form, to provide insights into the emerging landscape of privacy in the context of big data through a thorough study of the available big data privacy research.

References

1. Protection of Big Data Privacy ABID MEHMOOD1, IYNKARAN NATGUNANATHAN1, YONG XIANG1, (Senior Member, IEEE), GUANG HUA2, (Member, IEEE), AND SONG GUO3, (Senior Member, IEEE).
2. Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003; 12(2):120–39.
3. Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015; 2(4):155–65.
4. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1. 137
5. Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1. 12.
6. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big Data Springer Open J. 2015.
7. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.
8. B. Matturdi, X. Zhou, S. Li, and F. Lin, ''Big data security and privacy: A review, China Commun., vol. 11, no. 14, pp. 135–145, Apr. 2014.
9. Qin Y, et al. When things matter: a survey on data-centric internet of things. J Netw Comp Appl. 2016; 64:137–53.
10. Porambage P, et al. The quest for privacy in the internet of things. IEEE Cloud Comp. 2016; 3(2):36–45.
11. 11. Xu L, Jiang C, Wang J, Yuan J, Ren Y. Information security in big data: privacy and data mining. IEEE Access. 2014; 2:1149–76.
12. Liu S. Exploring the future of computing. IT Prof. 2011; 15(1):2–3.
13. Sokolova M, Matwin S. Personal privacy protection in time of big data. Berlin: Springer; 2015.
14. Cheng H, Rong C, Hwang K, Wang W, Li Y. Secure big data storage and sharing scheme for cloud tenants. China. Communication. 2015; 12(6):106–15.
15. Mell P, Grance T. The NIST definition of cloud computing. Natl Inst Stand Technol. 2009; 53(6):50.
16. Xu K, et al. Privacy-preserving machine learning algorithms for big data systems. In: Distributed computing systems (ICDCS) IEEE 35th international conference; 2015.
17. Zhang Y, Cao T, Li S, Tian X, Yuan L, Jia H, Vasilakos AV. Parallel processing systems for big data: a survey. In: Proceedings of the IEEE. 2016.
18. Abouelmehdi, K., Beni-Hessane, A. & Khaloufi, H. Big healthcare data: preserving security and privacy. J Big Data 5, 1 (2018). https://doi.org/10.1186/s40537-017-0110-
19. Hill K. How target figured out a teen girl was pregnant before her father did. New York: Forbes, Inc.; 2012. [Online]. http://www.forbes.com/sites/kashmirhill/2012/02/16/ howtarget- figured-out-a-teen-girl-was-pregnant-before-herfather- did/.
20. Violán C, Foguet-Boreu Q, Hermosilla-Pérez E, Valderas JM, Bolíbar B, Fàbregas-Escurriola M, Brugulat-Guiteras P, Muñoz-Pérez MÁ. Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multi morbidity. BMC Public Health. 2013;13(1):251.
21. Emam KE. Guide to the de-identification of personal health information. Boca Raton: CRC Press; 2013.
22. K. Hu, D. Liu, et al., Research on Security Connotation and Response Strategies for Big Data[J], Telecommunications Science, vol.2, pp.112-117, Feb. 2014.
23. R. Lu, H. Zhu, et al., Toward Efficient and Privacy-Preserving Computing in Big Data Era[J], IEEE Network, Aug. 2014.
24. Jung K, Park S, Park S. Hiding a needle in a haystack: privacy preserving Apriori algorithm in Map Reduce framework PSBD'14, Shanghai; 2014. p. 11–17.
25. Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. Big Data Congress; 201441. Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. Big Data Congress; 2014.
26. Zhang X, Yang T, Liu C, Chen J. A scalable two-phase top-down specialization approach for data anonymization using systems, in Map Reduce on cloud. IEEE Trans Parallel Distribute. 2014;25(2):363–73.
27. Mohammad Ian E, Noferesti M, Jalili R. FAST: fast anonymization of big data streams. In: ACM proceedings of the 2014 international conference on big data science and computing, article 1. 2014.
28. Sumra, I.A. (2022). Security Issues and Challenges in Vehicular Big Data Network (VBDN): A Survey. Engineering Software for Modern Challenges. ESMoC 2021. Communications in Computer and Information Science, vol 1615. Springer, Cham. https://doi.org/10.1007/978-3-031-19968-4_3.
29. D. S. Terzi, R. Terzi, et al., A Survey on Security and Privacy Issues in Big Data[C], in Proc. 2015 IEEE International Conference on Internet Technology and Secured Transactions (ICITST' 2015), 2015.
30. Liang K, Susilo W, Liu JK. Privacy-preserving cipher text for big data storage. In: IEEE transactions on informatics and forensics security. vol 10, no. 8. 2015.
31. Xu K, Yue H, Guo Y, Fang Y. Privacy-preserving machine learning algorithms for big data systems. In: IEEE 35th international conference on distributed systems. 2015.

32. Zhang X, Dou W, Pei J, Nepal S, Yang C, Liu C, Chen J. Proximity-aware local-recoding anonymization with Map Reduce for scalable big data privacy preservation in cloud. In: IEEE transactions on computers, vol. 64, no. 8, 2015.
33. Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S. Protection of big data privacy. In: IEEE translations and content mining are permitted for academic research. 2016.
34. Yan Z, Ding W, Xixun Yu, Zhu H, Deng RH. Deduplication on encrypted big data in cloud. IEEE Trans Big Data. 2016;2(2):138.
35. M.Yang, X. Zhou, et al., Challenges and Solutions of Information Security Issues in the Age of Big Data[J]. China Communications Magazine, Mar. 2016.
36. Jain, P., Gyanchandani, M. & Khare, N. Differential privacy: its technological prescriptive using big data. J Big Data 5, 15 (2018). https://doi.org/10.1186/s40537-018-0124-9.
37. Lv, D., & Zhu, S. (2019). Achieving correlated differential privacy of big data publication. Computers & Security, 82, 184-195.
38. M.A.P. Chamikara, P. Bertok and D. Liu et al., Efficient privacy preservation of big data for accurate
39. data mining, Information Sciences, https://doi.org/10.1016/j.ins.2019.05.053
40. Wu, Y., Huang, H., Wu, N., Wang, Y., Bhuiyan, M. Z. A., & Wang, T. (2020). An incentive-based protection and recovery strategy for secure big data in social networks. Information Sciences, 508, 79-91. https://doi.org/10.1016/j.ins.2019.08.064.
41. Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Alshahrani, M. S. (2021). Applications of Big Data Analytics to Control COVID-19 Pandemic. Sensors, 21(7), 2282.
42. https://doi.org/10.3390/s21072282
43. Haoxiang, W., & Smys, S. (2021). Big data analysis and perturbation using data mining algorithm. Journal of Soft Computing Paradigm (JSCP), 3(01), 19-28.
44. Bansal, B., Jenipher, V. N., Jain, R., Dilip, R., Kumbhkar, M., Pramanik, S., Roy, S., & Gupta, A. (2022). Big Data Architecture for Network Security. https://doi.org/10.1002/9781119812555.
45. Vasa, J., & Thakkar, A. (2022). Deep Learning: Differential Privacy Preservation in the Era of Big Data, 608-631. https://doi.org/10.1080/08874417.2022.2089775
46. Yao, A., Li, G., Li, X., Jiang, F., Xu, J., & Liu, X. (2023). Differential privacy in edge computing-based smart city Applications: Security issues, solutions, and future directions. https://doi.org/10.1016/j.array.2023.100293
47. K2 Data Science. (n.d.). Big Data Systems 101. K2 Data Science & Engineering. Retrieved [Month Day, Year], from https://blog.k2datascience.com/big-data-systems-101-e9226a691d6b.