RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
www.rspsciencehub.com
Vol. 07, Issue 02 February

Check for updates

http://dx.doi.org/10.47392/IRJASH.2025.011

Big Data Privacy and Security in Data Analytics: A Review On Issues, Challenges and Privacy Preserving Methods

Vinitha Mary[1], Merwin J[2], G Hemanth[3], Balarajesh R[4]

[1]Assistant professor, Dept. of IT, Manakula Vinayagar Institute of Technology, Pondicherry, India.

[2,3,4]UG Scholar, Dept. of IT, Manakula Vinayagar Institute of Technology, Pondicherry, India.

Emails: *jvinitha11@gmail.com[1], merwinofficial24@gmail.com[2], hemanthmvit26@gmail.com[3], dhanishkramesh@gmail.com[4]*

Article history

Received: 25 January 2025
Accepted: 04 February 2025
Published: 20 February 2025

Keywords:
Big data 5V features, security, privacy, CSA, K-anonymity

Abstract

In recent years, with the rapid development of technologies such as the internet, the internet of things and cloud computing, data production has increased in many areas such as business, education and economy. Big Data has emerged as a prominent concern of interest international, drawing huge attention throughout these fields. However, making sure the privateness and safety of Big Data stays a vital difficulty. The specific 5V traits of Big Data—Volume, Variety, Velocity, Value, and Veracity—demand stronger security measures to deal with those demanding situations efficiently. This studies paper highlights key security and privateness concerns related to Big Data as recognized by means of the Cloud Security Alliance (CSA). It also explores potential techniques and answers to beautify the safety of facts processing and computing infrastructures. Additionally, the paper gives a top-level view of the K-Anonymity method, a privacy-preserving technique designed to guard character identities and touchy records from being disclosed when datasets are shared or analyzed. Finally, the observe reviews Big Data safety solutions provided by using leading companies and outlines their features to ensure robust statistics safety.

1. Introduction

The rise of Internet communication, telecommunications, and the Internet of Things (IoT) has led to a data explosion, which is now often referred to as Big data [1]. Big data is characterized by being versatile, multi-format, and high-speed. It is defined by the 5V attributes: Volume, Velocity, Variety, Value, and Truth. These attributes present unique challenges at all stages of the big data lifecycle [2], especially with respect to privacy and five key security issues: confidentiality, efficiency, accuracy, availability, and fairness. Figure 1 shows how the stability dimension corresponds to the 5V characteristics of big data. When information is leaked, its value decreases. The value of large files can be lost if hackers destroy data by modifying it or accessing confidential information. Performance is particularly important for big data security and privacy because it requires high network connectivity. Authenticity is essential to ensure the reliability of data sources, data processors, and data

authorization requests. Authenticity can help prevent measurement errors and extract more value from big data. We need to be able to access big data whenever we need it. Otherwise, it will lose its value. Honesty is also very important to access useful and accurate information. If the information is wrong or incomplete, especially when the incomplete information is the most important and valuable information, we cannot confirm the truth. Big data is frequently used for analysis purposes today and is used in many areas such as health, public institutions, business, research and other organizations. These analyses often require data for reporting, research, etc. Big data also contains highly sensitive private information, and presenting this information directly to review can pose a threat to users' privacy. Therefore, privacy-preserving big data mining strategies are needed to prevent personal and sensitive information from being leaked from the database. Table 1 shows Security Viewpoints in Huge Information Life Cycle.

Table 1 Security Viewpoints in Huge Information Life Cycle

Big data 5V characteristics	Security Aspects				
	Confidentiality	Efficiency	Authenticity	Availability	Integrity
Volume		√		√	
Velocity		√		√	
Variety		√		√	
Value	√		√	√	√
Veracity	√		√		√

2. Security Versus Privacy

Security of Enormous Information are critical concerns that must be addressed to ensure its proper utilization. Security primarily deals with safeguarding data against malicious attacks and preventing unauthorized access or theft for financial gain [3]. Data privacy covers the management and control of personal data, emphasizing the development of policies that ensure individuals' Collect Personal Information, shared, and use it appropriately and responsibly.

3. Data Security and Privacy Issues & Challenges

Enormous information faces noteworthy challenges each day when managing with the protection and security of huge and assorted information. The information can be submitted from different people such as scientists, researchers, doctors, business people, important organizations, etc. and can become personal. Moreover, the security and privacy protections of existing technologies are weak and thus vulnerable to unauthorized or unintentional breaches. Recently, The Cloud Security Alliance (CSA) has published ten Data Security Thaub Privacy Guide [4]. The main purpose of these challenges is to refocus on supporting big data. Key researchers at the CSA Enormous Information Working Bunch have compiled a list of important issues within the setting of huge information security and security, which can be divided into four errands and after that encourage separated into 10 diverse security measures [5]:

- Infrastructure Security
- Data Privacy
- Data Management
- Integrity and Reactive Security

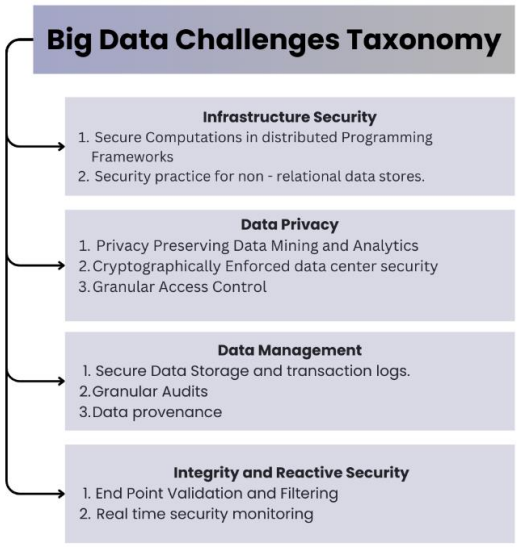

Figure 1 Scientific Classification of Best 10 Enormous Information Challenges

3.1. Included in The Security Classification Process

Distributed programming frameworks use the concepts of parallel computing and storage to process large datasets. The MapReduce framework is a prime example of this. It splits the input data into multiple chunks and then the mapper reads these chunks, performs computations, and places the elements in the form of key-value pairs. The Reducer then combines the results of each variable

into a single value and gives the result as output. The two main disadvantages here are protecting the mapper and protecting the data from malicious mappers.

3.2. Best security in storing irrelevant data

NoSQL (non-relational) databases are utilized to store huge sums of information and can solve many of the problems of analyzing huge sums of information without dealing with security issues. Developers working with NoSQL repositories often build security into their core architecture. NoSQL databases don't give express bolster for this in their documentation. Another challenge to the strength of these security hones is the integration of NoSQL storehouses [6].

3.3. Protecting The Confidentiality of Data Mining and Analysis

Big data can lead to privacy violations, business disruptions, reduced civil liberties, and increased government and business control. Client information collected by companies and public institutions is constantly monitored and reviewed by internal analysts and even external personnel. Malicious insiders or untrusted partners can misuse this information and obtain private data from clients. It is vital to have rules and proposals in put to anticipate security breaches.

3.4. Granular Access Control

According to CSA, control has two dimensions: Restrict user access and allow user access. The competition was organized and use strategies to make the right choice in each situation.

3.5. Security of Data Storage and Data Exchange

Data and data transactions are stored at different levels of data storage, and manually moving data between tiers allows IT administrators to control what data happens and when. However, as data continues to increase in size, scalability, and usability make it crucial to prevent big data storage management fatigue. Automated solutions fail to track where data is stored presents new challenges for data security [7].

3.6. Granular Audits

It may be a best hone to do quality testing. Real security monitoring, we can be alerted immediately when an attack occurs. But that’s not always the case. We need audit data to get to the bottom of every attack. This will help us understand what’s going on and is also important for compliance and monitoring. Analysis is not new, but the sources and data size may vary, for example, We need to create a large file of the deployed items.

3.7. Data Provenance

Metadata in big data usage authentication becomes more challenging due to the large number of certificates generated by the workstation supporting the authentication. Analyzing such large datasets to identify metadata dependencies from a security/privacy perspective is computationally intensive.

3.8. Endpoint Authentication Filtering

Huge information such as security data and occasion administration (SIEM), collect occasion logs from millions of equipment gadgets and applications across the network enterprise. One of the key challenges in data collection is: How can we trust this information? How can we be sure that products are not harmful and how can we kill hurtful thoughts from our items? Input approval and sifting are complex issues that arise from untrusted inputs, especially for a Bring Your Own Device (BYOD) mode.

3.9. Real Security Monitoring

Genuine security checking is continuously our request challenge with large-scale incident investigation data, largely due to the large number of reports generated by security devices. These alerts may or may not be relevant, leading to a large number of false alarms that can be clicked away or ignored due to the lack of capacity to process multiple alerts quickly [8].

4. Arrangement to Guarantee Huge Information Security and Security

Here are a few technologies designed to ensure data protection: Access control technology: In the big data environment, many users and intricate authority structures, adopting new technologies is essential to enable controlled data sharing. The Role-Based Access Control (RBAC) mechanism is a commonly utilized approach for managing access. By implementing access controls on data entries, this method ensures only a select group of users can access sensitive information. The main challenge is to prevent unauthorized people from accessing or misusing confidential information. Homomorphic Encryption Scheme (HES): Ensuring the confidentiality of big data, even in cases where data breaches occur, attackers should not be able to derive meaningful information from the

compromised data. This involves using an encryption Direct calculation method without decrypting encrypted data. in other language, with only the encrypted version of a message, it is possible to compute an encrypted output corresponding to a function applied to the original message, all while preserving the encryption. Secure Multi-Party Computing (SMC): It basically bargains with computational issues that have arrangements. The purpose of SMC is to handle tasks where both parties provide some input. In this contract, both parties have assurances such as confidentiality, correctness, etc. In terms of confidentiality, the security It should not leak any information other than the transaction output. Anonymization techniques: Anonymized or anonymous data is a very popular technique in both centralized and decentralized databases. In this way, sensitive data must be rendered incapable of being identified as personal data. Even if an attacker obtains this information, Since the key value is secret, access to the key information is not possible. The two fundamental protection objectives that got to be accomplished whereas anonymizing information are:

- Certain Personal Information: If products are made available, no personally identifiable information should be included.
- Good understanding of behavior: Attackers cannot learn about individual sensitive behavior from leaked data.

4.1. Data-Anonymization/De-Identification Innovation

- It is outlined to stow away a person's private and sensitive information. This is the privacy policy used when large amounts of information are provided to third parties. In general, there are 3 types of attributes of data in a dataset:
- Basic attributes are the characteristics that identify each person, such as identity, name, address, and phone number. Please delete before sending.
- Quasi-Identifier (QI) is a set of features that can be associated with other publicly available data to uniquely identify an individual's personal information. It can be used to associate anonymous data with other data. For example, age, gender, zip code, city, etc.
- Bad habits include sensitive information that a person wants to hide from others, such as income, salary, illness, medical history, etc.
- One of the foremost imperative benefits of sharing data namelessly is that in the event that the data is mysterious, it can be shared unreservedly between distinctive parties without limitation. There are three strategies to ensure the secrecy of mysterious information It is utilized to avoid assaults on the privacy of distributed data. These are K-anonymity, L-diversity and T-closure.

4.1.1. K-Anonymity

When the behavior is limited or generalized such that each push is continuously at slightest k-1 other lines, the strategy is called k-anonymity. It maintains the link between the exact data sources while also ensuring the accuracy of the published data. However, it has some limitations:

- It does not hide its identity.
- There is no way to avoid assaults based on chronicled information.
- K-anonymity cannot be utilized for high-level information.

4.1.2. L-Diversity

This method eliminates the disadvantages of k-anonymity, but cannot protect privacy from tilting and similar attacks.

4.1.3. T-Closeness

This approach can preserve privacy and prevent social and historical information. If the distance between the distribution of sensitive features in the same class and the distribution of the same feature in all words does not exceed a certain limit, it is called t-closeness. If all equal classes are t-closed, the table is said to be t-closed.

5. K-Anonymity

K-Anonymity gives a certain level of protection security by avoiding re-identification and guaranteeing exact and secure information recognizable proof. This protection show is utilized to protect against network security compromises [10] [11]. In the event that a tuple/individual within the distributed dataset cannot be recognized from at least k-1 other tuples/individuals within the dataset, then the dataset is K-anonymous. Therefore, an attacker who knows the value of a person's semi-identifying attribute cannot distinguish this person's data from the data of k-1 other individuals.

Two main ideas proposed to improve the K-anonymity of private data are [12]:

- Generalization is the process of replacing one value with a higher value. For example, instead of "man/woman" you can use "person".
- Bullying is hiding a esteem by not uncovering it at all. The esteem is supplanted with extraordinary characters (such as *, @).

5.1. K-Anonymity Algorithm

Input: Personal data set PT, partial identification number QI, input value A, anonymous number K.
Output: Set the RT table.

- Step 1: Select PT files from library.
- Step 2: Select the key attribute, semi-identifying attribute, and sensitive attribute from the attributes list
- Step 3: Select the most sensitive value setting A form the list of all values to be stored.
- Step 4: Sensitivity value for each tuple belonging to set A.
- Step 5: Find the statistics of the semi-characteristics in Table 1, that is, the difference between the character and all rows with that value.
- Step 6: The semi-descriptors in Table 1 are generalized to K-anonyms, which are RT output words ready for broadcast, Table 2 Important Medical Information, Table 3 Anonymous Data (After Removal of Important Attributes)

Table 2 Important Medical Information

Age	Sex	Zip Code	Disease
30	Male	12567	Fever
25	Male	13001	Cancer
26	Male	13001	Flu
34	Male	76512	Flu
23	Female	14599	Viral
35	Male	13057	Pneumonia
35	Female	17000	Fever
23	Female	32451	Cancer
27	Female	14560	Flu

To delete the first table, the main character must first be deleted from the table. The results are shown in Table2. The table below shows the external information that the opponent has access to, namely the voter roll. Table 4 shows Registration Information for Citizens to Vote

Table 3 Anonymous Data (After Removal of Important Attributes)

Key Attributes	QI Attributes		Sensitive Attribute	
Name	Age	Sex	Zip Code	Disease
Ravi	30	Male	12567	Fever
Sam	25	Male	13001	Cancer
Ramesh	26	Male	13001	Flu
Manav	34	Male	76512	Flu
Suhanti	23	Female	14599	Viral
Keshav	35	Male	13057	Pneumonia
Anita	35	Female	17000	Fever
Hema	23	Female	32451	Cancer
Reshma	27	Female	14560	Flu

Table 4 Registration Information for Citizens to Vote

Voter ID No	Name	Age	Sex	Zip Code
QDT2398452	Ravi	30	12567	12567
SAP2345918	Sam	25	13001	13001
QAC4982107	Ramesh	26	13001	13001
HTR4356723	Manav	34	76512	76512
RAP3412090	Suhanti	23	14599	14599
KDE7351898	Keshav	35	13057	13057
WRT2783261	Anita	35	17000	17000
WER253467	Hema	23	32541	32451
KYE3456123	Reshma	27	14560	14560

By analyzing Table 3 and Table 4, the attacker can see that Ramesh has a cold. This demonstrates that even when key identifiers are removed, an individual's identity can still be determined using publicly available data. The process of correlating the released table's data with publicly accessible datasets to identify individuals is referred to as a Linking Attack. To counter such attacks, this privacy model is employed. Essentially, when someone is trying to identify a person from a printed document, the as it were data they have is their age, sexual orientation, and zip code. Table 5 underneath appears a case of a 2-anonymous table with k = 2, which suggests that at slightest two tuples share the same esteem within the semi-identifier quality. As can be seen in this case,

t2[S]= t3[S]= t6[S] &t5[S]= t9[S]

Table 5 Anonymous Data

Age (equivalence class)	Sex	Zip Code(after suppression)	Disease
[20-30]	Male	125*	**Cancer**
[20-30]	Male	130*	**Cancer**
[20-30]	Male	130*	**Flu**
[30-40]	Male	765*	**Flu**
[20-30]	Female	145*	**Viral**
[30-40]	Male	130*	**Pneumonia**
[30-40]	Female	170*	**Fever**
[20-30]	Female	324*	**Cancer**
[20-30]	Female	145*	**Flu**

5.2. Attacks on K-Anonymity

- Feature Exfiltration Assault / Homogeneity Assault: This happens when there is no difference in the importance of sensitive features and the attacker only wants to know the importance of sensitive features.
- Foundation assault: Typically, moreover an assault that k-anonymity cannot ensure against. This demonstrate expect that the aggressor does not have any extra foundation information. Table 6 shows Big Data Security Solutions and Key Features from Leading Companies.

6. Big Data Security Solutions Provided by Various Companies and Their Basic Functions

Table 6 Big Data Security Solutions and Key Features from Leading Companies

COMPANY NAME	APPLICATION	KEY FEATURES
IBM	IBM QRadar Security Intelligence Platform [13]	• A comprehensive, integrated approach that connects to instantaneous continuous insights, analytics driven by large amounts of organized and unstructured information, and profound knowledge measurable capabilities. • Front-end graphics tools for visualizing and exploring large data sets.
INFOSYS	IIP-The Infosys Information Platform [14]	• Open information analytics stage. • Empower businesses to work on their important information and find modern openings for fast alter and development. • Using open innovation and internal development, provide an end-to-end data platform that seamlessly integrates into the business environment and can operate as a big data platform or as an add-on to existing equipment.
MICROSOFT	Big Data and Business Intelligence Solutions [15]	• Gives an advanced information administration chain of command that underpins all sorts of information. Yes. Inactive or energetic organized, semi-structured and unstructured information. MS makes it simple to coordinated, oversee and display information in genuine time, giving more prominent perceivability into the commerce, empowering quicker choices. • An extra layer that enhances our information through revelation, combines it with worldwide information, and upgrades it with progressed analytics. • HDInsight may be a new Hadoop-based benefit created by MS that's 100% consistent with Apache Hadoop.

Conclusion

This study highlights the critical challenges and opportunities in ensuring privacy and security within Big Data environments. By analyzing the existing methods and technologies, we identified the strengths and limitations of Privacy-preserving techniques such as k-anonymity, l-diversity, and t-closeness. The study underscores the importance of adopting advanced cryptographic measures, access control mechanisms, and real-time security monitoring to safeguard sensitive data. As Big Data continues to evolve, addressing these challenges will be paramount for leveraging its full potential while protecting individual and organizational privacy. Future research should focus on developing scalable and robust security frameworks that adapt to the ever-increasing volume, variety,

and velocity of data.

Acknowledgements

We would like to extend our heartfelt gratitude to all individuals and institutions that supported this study. We thank the Manakula Vinayagar Institute of Technology, Puducherry, for providing the necessary resources and facilities to conduct our research. Our sincere appreciation goes to the faculty and peers for their invaluable guidance and constructive feedback throughout the process. Additionally, we acknowledge the role of previous researchers whose work laid the foundation for this study. Lastly, we are deeply grateful for the encouragement and support from our families, which motivated us to complete this project.

References

[1]. Jha A, Dave M. and Madan, S. 2016. A Review on the Study and Analysis of Big Data using Data Mining Techniques, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol6, Issue 3.

[2]. Jha A, Dave M. and Madan, S. 2016. Quantitative Analysis and Interpretation of Big Data Variables in Crime Using R, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), Vol5, Issue7.

[3]. Q, etal, Jing. 2014. Security of the internet of things: perspectives and challenges. 20(8):2481–50. https:// cloudsecurityalliance. org/media/news/csa-big- data- releases-top-10-security-privacy challenges/.

[4]. A Cloud Security Alliance Collaborative research, Expanded Top Ten Big Data Security and Privacy Challenges. 2013.

[5]. Okman, L., Gal-Oz N., Gonen Y, Gudes E. and Abramov J. 2011. Security Issues in NoSQL Databases in TrustCom IEEE Conference on International Conference on Trust, Security and Privacy in Computing and Communications, pp 541-547.

[6]. Apparao, Yannam and Laxminarayanamma, Kadiyala. 2015. Security Issue on Secure Data Storage and Transaction Logs In Big Data" in International Journal of Innovative Research in Computer Science & Technology (IJIRCST).

[7]. Singh, Reena and Kunver Arif Ali. 2016. Challenges and Security Issues in Big Data Analysis, IJIRSET, Vol 5. Issue 1.

[8]. Sedayao, J. Enhancing cloud security using data anonymization, White Paper, Intel Corporation.

[9]. Kenig Batya and Tassa Tamir. 2011. A practical approximation algorithm for optimal k-anonymity, Data Mining Knowledge Discovery, Springer.

[10]. Sweeney, L. 2002. K-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty Fuzziness Knowledge based Systems.

[11]. Samarati. P. 2001. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Eng., 13:1010–1027.