International Journal of Engineering Research and Sustainable Technologies
Volume 1, No.1, September 2023, P 38-49
ISSN: 2584-1394

A REVIEW ON BIG DATA PRIVACY AND SECURITY IN HEALTH CARE

N. Prabu Sankar[a], D.Usha[b]
[a]Research Scholar, [b]Associate Professor
[b]Dept.of CSE, Dr. M.G.R. Educational and Research Institute, Chennai, India
*** Corresponding author email:** n.prabusankar81@gmail.com

DOI: https://doi.org/10.63458/ijerst.v1i1.61 | ARK: https://n2t.net/ark:/61909/IJERST.v1i1.61

Abstract

Due to the proliferation of the Internet, IoT, and Cloud Computing, there is now an abundance of virtual data in every industry, field of study, and government agency. Big data has quickly become a topic of intense interest, garnering media coverage and commentary from all over the world. Data privacy and security in Big Data is a pressing concern. The 5Vs of big data—size, velocity, value, veracity, and variety—lower the bar for adequate protection. This paper aimed to draw attention to security and privacy issues and challenges associated with Big Data in healthcare, the resolution of which could result in a more secure data processing and computing infrastructure. This paper also provides a high -level overview of the K-Anonymity technique for protecting the privacy of large datasets before they are released for analysis, with the goal of preventing the disclosure of personally identifiable information. In conclusion, this paper summarizes the functions and features of the best big data security solutions offered by industry leaders.

Keywords: Big Data, Analytics, Security, Privacy, Confidentiality

1. Introduction

The proliferation of online resources like social media and the IoT has resulted in a surge in data production, or "big data," in recent years [1]. Given that big data is generated rapidly from numerous locations using a wide variety of file formats, it is crucial to consider the 5Vs of big data: volume, velocity, variety, value, and veracity. New challenges arise for Big data throughout its entire life cycle [2] due to privacy and the five major security aspects: confidentiality, efficiency, authenticity, availability, and integrity. In the era of big data, data privacy and security depend on keeping data safe and secure. Information loses value the moment it is exposed. If hackers can alter the data or obtain confidential information, the value of the big data will be destroyed. Big data security and privacy are especially dependent on efficiency due to the high network bandwidth required. Valid data sources, valid data processors, and valid data requesters must insist on authenticity. Incorrect analysis results can be avoided, and high potential value can be realized, with the help of authenticity. Whenever we need it, we should be able to access large amounts of data. If not, it may lose its worth. Finding trustworthy sources is also crucial for gathering useful information. Lacking the most sensitive and useful information, we cannot conduct an accurate analysis without it.

Consequently, these institutions must effectively oversee patient data, guaranteeing swift access to vital health information while upholding the highest standards of data security and privacy. Healthcare, government, businesses, researchers, and other organizations all use big data today for analysis purposes. Their information is often needed for various research and publication projects. Given that big data often includes details about specific people, its direct release for analysis can put users' privacy at risk. Therefore, there is a need for privacy-preserving big data mining techniques that aim to prevent the disclosure of personal identities and other sensitive data. Insufficient cyber security measures can expose patient data to breaches, jeopardizing the confidentiality of sensitive medical information. Incorporating big data and security in healthcare is not just a technological hurdle; it stands as an absolute imperative. Striking the right equilibrium between the potential advantages of big data analytics and the implementation of strong security measures is indispensable for upholding the privacy and confidentiality of patient data.

2. Data Security and Privacy

In our modern, highly interconnected world, the practice of gathering and scrutinizing immense datasets has become commonplace. The utilization of big data analytics has brought about substantial transformations in multiple sectors, including healthcare, finance, marketing, and government. Nevertheless, as entities leverage the capabilities of big data, they are confronted with the intricate challenge of striking a harmonious equilibrium between safeguarding data security and upholding individual privacy.

The term "big data" encompasses vast amounts of structured and unstructured data that organizations gather and examine with the aim of gaining insights, making well-informed decisions and refining their operations. This data is distinguished by its substantial volume, rapid generation pace, diverse sources, and the need for high accuracy. These data streams emanate from a variety of origins, including social media, Internet of Things (IoT) devices, sensors, and online transactions. Big data analytics relies on sophisticated algorithms and machine learning techniques to unearth underlying patterns, trends, and correlations within this extensive pool of information. The emergence of the big data revolution carries the potential to stimulate innovation, augment competitiveness, and propel advancements in scientific research across numerous sectors. In healthcare, for instance, big data analysis can play a pivotal role in forecasting disease outbreaks, optimizing treatment strategies, and enhancing patient outcomes.

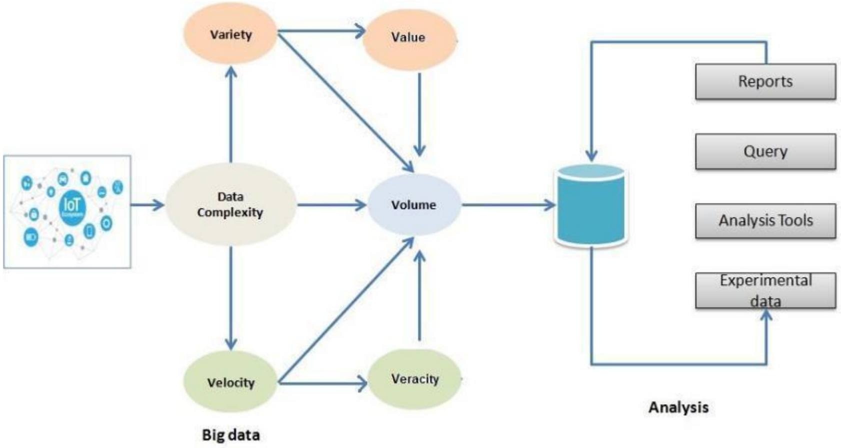

Fig. 1 Proposed IoT enabled Big data analysis in Healthcare

Figure 1 shows the proposed IoT enabled big data analysis of healthcare. Big data is based on the principle that data sets and databases should expand indefinitely. Additionally, there is a corresponding increase in issues relating to privacy, law, and security. What privacy, legal, and security challenges arise from the acquisition and processing of massive amounts of data? is a natural research question to ask.

2.1 Challenges in Data Security and Privacy

Concerns about data security and privacy are paramount when dealing with large datasets. To make effective use of big data, it is necessary to resolve concerns about privacy and data protection. When it comes to data, security is all about keeping it safe from malicious attacks and data thieves who only care about making a buck off of you [3]. The term "data privacy" describes the process of keeping information secure, particularly sensitive information such as that which identifies an individual person. Security and privacy are two intertwined yet distinct aspects of big data that require careful consideration. Security primarily centers on safeguarding data from unauthorized access, breaches, and cyber attacks, while privacy focuses on protecting individuals' personal information and ensuring it is not misused or disclosed without consent. These two dimensions often conflict, as strengthening security measures to protect data can sometimes compromise individual privacy, and vice versa. A central challenge in big data security is the sheer volume of data being collected and stored. Big data environments typically handle vast datasets that are continuously growing and evolving. This poses a unique challenge for security professionals tasked with implementing measures to protect data at rest, during transmission, and while being processed. The more data there is, the greater the potential attack surface for malicious actors.

1.2 Security Measures in Big Data

One common security measure is encryption, which involves encoding data to render it unreadable without the appropriate decryption key. While encryption effectively protects data from unauthorized access, it also adds complexity and can slow down data processing. This is particularly concerning in big data environments that demand real-time or near-real-time analysis. Another substantial challenge in big data security [4] is the diversity of data sources and formats. Big data often originates from various sources such as sensors, social media, IoT devices, and more. These diverse data streams may have distinct security requirements, making it challenging to apply uniform security measures across the entire dataset. Additionally, data can be unstructured, semi-structured, or structured, further complicating the security landscape.

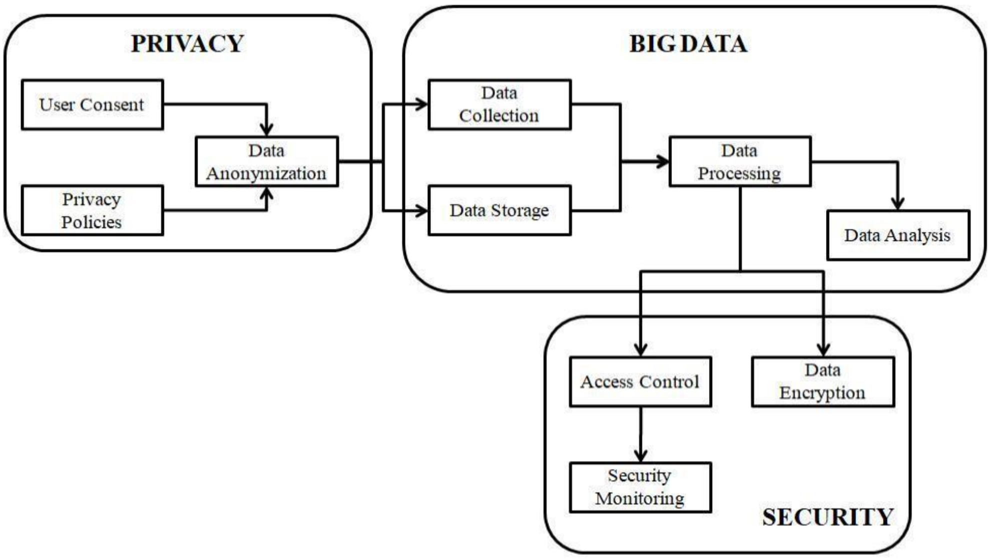

Fig. 2 Privacy and security in big data

In Figure 2, "Big Data" encompasses data collection, storage, processing, and analysis. "Security" includes access control, data encryption, and security monitoring to protect data. "Privacy" focuses on user consent, data anonymization, and privacy policies to protect individuals' privacy. Privacy concerns within the realm of big data are equally significant. Given the vast amount of data being collected and analyzed, there is a growing risk of individuals' personal information being exposed or exploited. This raises ethical and legal questions about how data should be collected, used, and retained. A primary privacy concern is data anonymization [11], which entails removing or altering personally identifiable information (PII) in datasets to protect individuals' privacy. However, achieving true anonymization[15] is challenging, as data can often be re-identified through various means, like cross-referencing with other datasets or using advanced machine learning techniques. The risk of re-identification underscores the delicate balance between data utility and privacy protection[16].

In some cases, organizations may opt to de-identify data by removing or altering identifiers, such as names or social security numbers. While this approach reduces privacy risks, it may also diminish the data's usefulness for analysis. Striking the right balance between data utility and privacy is an ongoing challenge in the realm of big data[36] analytics. Data breaches can have severe consequences, affecting both security and privacy[5]. When a breach occurs, sensitive information may be exposed, leading to identity theft, financial fraud, and damage to an entity's reputation. Organizations must find a balance between their efforts to prevent breaches and their responsibility to protect individuals' privacy. Balancing big data security and privacy is an ongoing challenge, and there are inherent trade-offs. Organizations must continuously evaluate and adapt their strategies to address emerging threats and changing privacy regulations.

1.3 Concern about Big Data Security and Privacy

Big Data is constantly pushed to its limits by the need to protect the privacy and security of vast, often unstructured datasets. People from all walks of life, including researchers, scientists, doctors, business executives, government agencies, and so on, share data on a massive scale. While these massive data sets are becoming increasingly important, to deal with them, we lack adequate resources, such as tools and technologies. Moreover, breaches, both accidental and intentional, are constant with today's technologies due to their inadequate security and privacy [12] maintenance capabilities. Figure 3 gives a brief overview of the security and privacy issues in big data technology.

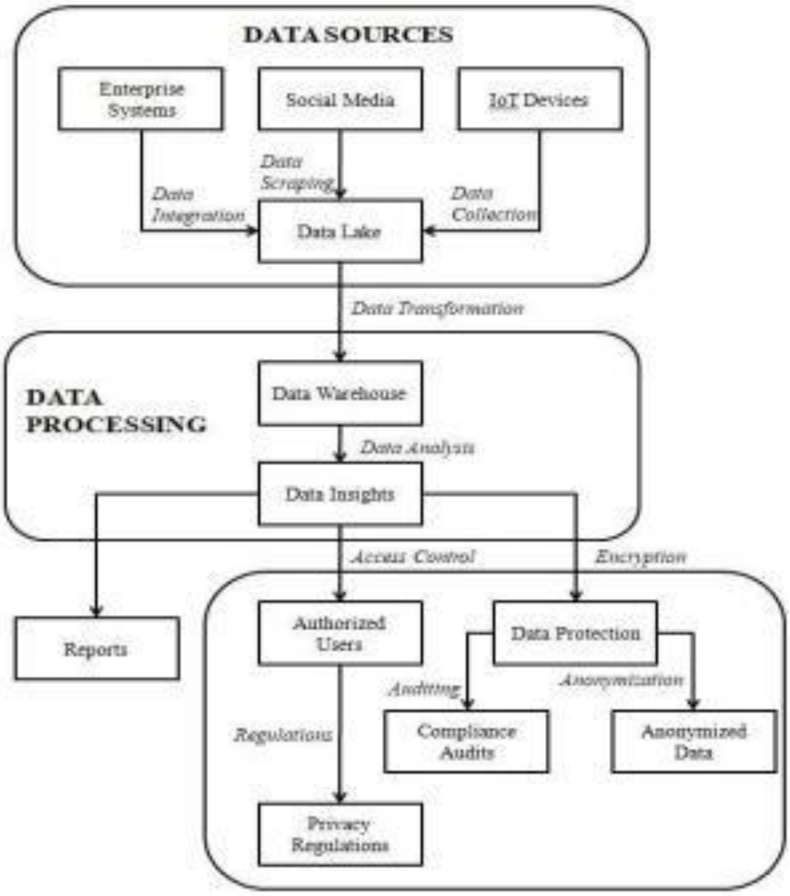

Fig. 3 Concerns in security and privacy of Big Data

The Cloud Security Alliance (CSA) published a paper outlining the top ten threats to the privacy[13] and security of big data [7,8]. The point of bringing attention to these issues is to refocus the community's efforts on fostering big data infrastructures. Ten unique security concerns in the context of big data have been identified by the CSA's Big Data working group, and these challenges fall into one of four broad categories. Table 1 shows a brief category about the classification of top big data problems.

Table 1

A Hierarchy of the Top Ten Big Data Problems

BIG DATA TAXANOMY CHALLENGES			
I. Infrastructure Security	**II. Data Privacy in healthcare**	**III. Healthcare data management**	**IV. Reactive security and integrity**
1.Integrity and confidentiality of data in distributed systems 2. Methods that have proven effective for protecting non-relational databases	1. Mining and analyzing data while protecting privacy 2. Data centric cryptographic enforcement 3. Granular Access Control	1. Protected database and Activity Records 2. Granular audits 3. Provenance of data	1. Verifying and sifting at the endpoint 2. The need for Constant Monitoring to Ensure Security

To process such large datasets, distributed programming frameworks often resort to the idea of parallel computation and storage. The best illustration of this is the MapReduce framework. The mapper takes the chunks of the input file, reads them, performs the computations, and then returns the results as a set of key/value pairs. The values associated with each unique key are then merged by the Reducer, and the resulting value is returned. The security of the mappers and the security of the data from a malicious mappers are the two main attack vectors here. Large data sets can be stored in NoSQL (non-relational) databases without too much worry about security. Most often, NoSQL database security is implemented by the middleware used by developers. In general, to enforce such rules within the database itself is not a feature offered by NoSQL databases. But the fact that NoSQL databases are typically organized in clusters makes it more difficult to implement effective security measures [9]. Potential downsides of big data include increased state and corporate control, less respect for individual liberties, and more intrusive marketing. Inside analysts and possibly external contractors are constantly mining and analysing user data collected by businesses and government agencies. Customers' personal information can be stolen from these datasets if used by a malicious insider or untrustworthy partner. Preventing privacy[14] breaches due to carelessness requires the establishment of guidelines and recommendations.

To guarantee end-to-end security and restrict access to only authorized parties, depending on the rules set forth by the system administrators, sensitive data may need to be encrypted. We need to improve the depth, efficiency, and scalability of this kind of research to advance technologies like attribute-based encryption (ABE). Distributed entities need a cryptographically secure communication framework to ensure authentication, agreement, and fairness. The CSA states that there are two primary aspects of access control: limiting user access and granting user access. The difficulty lies in developing and enforcing a policy that can pick the best option under all conditions. The it manager has complete say over what and when data is moved between the various levels of storage media used to house data and transaction logs. For big data storage management, however, auto-tiring is a necessity due to the exponential growth in data set size. As a result, secure data storage is made more difficult by auto-retirement solutions [10].

Many data security[27] strategies have been created, including Because of the sheer volume, we need to switch to a new technology. of users and the complexity of authority to implement the controlled sharing of data in a big data environment. One common approach to security is the use of predefined roles (RBAC). To accomplish this, we can implement permissions on the individual data records, such that only specific user groups have access to sensitive information. The biggest challenge is preventing unauthorized parties from mishandling sensitive information. Big data can be kept private by preventing an attacker from knowing how much the data is actually worth, even if sensitive information is leaked. A form of encryption that does not necessitate decryption in order to perform computations on encrypted material. So, given only the encryption of a message, it is possible to immediately compute the encryption of a function of the message.

Table 2

A comparison of recent privacy and security measures for big data

Paper	Focus	Limitations
[12]	Examines the trials and tribulations of using anonymization, Big Data techniques and methods for protecting users' anonymity during data analysis are combined to keep users' personal information safe.	Even now, it employs the insecure K-anonymity method, which is open to correlation attacks.
[13]	Developed privacy-protecting data mining methods for Hadoop, i.e., eliminate privacy breaches without diminishing benefits	The time it takes to run depends on the amount of noise present.
[14]	We have: Presented the Cosine Similarity Computing Protocol, a Privacy-Preserving Algorithm	Applications of big data analytics raise novel privacy concerns that require substantial more study.
[15]	Discussed and Suggested the Applicability of an Existing Method, Differential Privacy, to Big Data;	Noise levels are estimated only by the curator using this approach. Therefore, the entire system is at risk if the curator is hacked. Therefore, the entire system is at risk if the curator is hacked.
[16]	It was suggested to use the cloud-based MapReduce architecture and implement a top-down specialization strategy in two stages.	Uses a form of anonymization that can be exploited via correlation attacks.

[17]	(TDS) method was suggested for anonymizing massive amounts of data.	Discrimination of any kind, whether it be based on age, gender, race, religion, disability, socioeconomic status, or any other characteristic revealed by customer segmentation and profiling, is possible.
[18]	We propose a method dubbed FAST to quickly anonymize massive data streams.	In direction to take use of the vast computational resources of the cloud and to achieve great scalability, more study is needed to develop and execute FAST in such a framework.
[19]	This paper presents a unique framework for implementing machine learning in a way that protects users' privacy.	Due to the dispersed nature of the training data and the massive volumes of data that must be exchanged, distributed feature selection cannot be achieved.
[20]	Access control for malevolent insiders, as well as forward and backward access control, are all features of the proposed solution.	Reducing how much you can put your faith on a crypto server

3. Data Ownership

Encryption of data is a cornerstone of healthcare's big data security initiatives. Any sensitive data transmitted or stored by healthcare organizations should be encrypted. Those who are not permitted to view the data should be unable to do so. If a data breach occurs, the perpetrators won't be able to access the encrypted files because they don't have the decryption keys. Depending on how a healthcare organization operates, healthcare providers can decide for themselves which encryption methods are required. Data privacy and security [6] in healthcare are strengthened when only authorized individuals have access to private patient information and related applications. Passwords and other forms of access control should be implemented for an extra layer of security. Users must first authenticate before being granted access to restricted areas, therefore limiting access to only those who need to see confidential patient information. Multi-factor authentication is another form of authentication that can be implemented by businesses; it requires users to verify their identity using a combination of factors. Hence, security teams may simply locate unauthorized users and identify the source of privacy issues[30] in healthcare by restricting and tracking access to protected information.

As a result of the rapid digital revolution being wrought by new technologies, IoMT devices are increasingly being deployed in healthcare facilities. These tools pave the way for better patient data accessibility, remote monitoring, and more. Yet, this raises a wide range of additional confidentiality concerns in the medical field. Thus, one of the key things to develop big data security in healthcare is to secure IoMT devices by periodically updating the devices and providing a rigorous authentication process. Careful evaluation of the compliance of business partners and third-party vendor risk is essential due to the daily transmission of healthcare information between providers and across covered organizations to deliver superior treatment and to smooth payments. Your business will be held responsible for any data breaches that occur within its walls, regardless of whether or not they were caused by the carelessness of outside vendors. As a result, it is crucial to keep a close eye on all the various business associates in the healthcare industry.

As important as cyber security is this is one of the most neglected areas in the medical field. Most healthcare data privacy problems can be traced back to untrained personnel. Employees' homes have become a prime target for hackers, particularly as more and more businesses adopt remote work models. For this reason, it is essential to provide workers with education on cyber security and data privacy. To identify vulnerabilities in the event of a cyber attack, businesses should conduct frequent testing and drills. This helps lessen the chances of a breach occurring and gives employees the tools they need to respond appropriately in the event of a data breach. Therefore, it is essential for the success of Big Data Security in healthcare to foster a culture that places a premium on protecting sensitive information. Proactively integrating Big Data Security in healthcare guarantees important patient data is being properly safeguarded from ever-evolving threats, especially as the number and sophistication of cyber attacks against healthcare companies and privacy issues[28] develop simultaneously.

3.1 Maintaining Data Confidentiality

To protect the confidentiality[24] of sensitive information included in large healthcare databases, many different methods have been implemented. Common technologies include: To authenticate something means to prove that a statement made by or about the subject is genuine. It's crucial for every business since it helps prevent unauthorized access to sensitive information and networks, keeps user identities safe, and verifies that a user is indeed who he claims to be. Man-in-the-middle (MITM) attacks can be a challenge when it comes to authenticating data. Most cryptographic protocols have built-in endpoint authentication measures meant to foil man-in-the-middle attacks. Examples of cryptographic protocols that ensure the safety of online communications include transport layer security (TLS) and secure sockets layer (SSL). TLS and SSL both provide end-to-end encryption for transport layer network communications. The protocols are used in many different contexts, including online browsing, electronic mail, Internet faxing, instant messaging, and voice-over-IP, with multiple versions being widely deployed (VoIP). Secure Sockets Layer (SSL) and Transport Layer Security (TLS) allow users to verify the server's identity through a common, mutually trusted certification authority. Kerberos, a system based on Ticket Granting Ticket or Service Ticket and hashing algorithms like SHA-256 can also be used for authentication.

Data encryption[38] is an effective method of keeping private information safe from prying eyes. From the hospital's servers to the doctors' and administrators' mobile devices to the cloud, all points in between are protected by the company's solutions, ensuring that the data always remains under the control of the organization. Protection[21] against attacks like packet sniffing and stolen storage media is facilitated by encryption. The encryption technique used by healthcare organizations or providers must be robust, user-friendly for both patients and healthcare professionals, and capable of being expanded to accommodate emerging forms of electronic health data. When data is masked, it is replaced with an anonymous value to protect privacy. Since it is not a true encryption method, decrypting the masked value does not reveal the original. Data sets are de-identified by masking personal identifiers like names and social security numbers, and pseudo-identifiers like birth dates and postcodes are either removed or made generic. As a result, data masking has become one of the most widely used methods for protecting sensitive information in real time.

The difficulty in masking high-dimensional data sets is a challenge shared by these approaches [25, 26]. As a bonus, this method makes massive data deployments cheaper to secure. If data is encrypted before it is transferred onto the platform, masking can eliminate the need for extra security controls once the data is in the platform. Users can gain access to an information system after authentication, but their actions within the system are subject to the terms of an access control policy, which is often based on the privileges and rights of each practitioner allowed by patient or a trusted third party. Hence, it is an effective and adaptable means of conferring privileges on users.

3.2 Monitoring Security Breaches

Monitoring for security breaches entails collecting and analyzing data about unusual activity on a network in order to identify potential threats. By keeping track of each user's access and modifications to data, an audit ensures that the healthcare system is operating as intended. There are two additional security metrics that can be used to monitor and guarantee the integrity of a healthcare system [29]. It's not easy to implement intrusion prevention measures across an entire network's worth of traffic. The proposed fix makes use of data correlation methods to allow for dispersed storage and processing of data. At this point, we can tell if a domain name, packet, or flow is malicious thanks to three determined likelihood metrics. As a result of this calculation, either an alert is generated by the detection system, or the procedure is terminated by the prevention system. The goal of network security[31] systems for large amounts of data is to swiftly detect anomalies and correctly flag them as threats. That's why it's important to have a model for a big data security[35] event monitoring system [32] that includes the four components of data collection, integration, analysis, and interpretation. Logs and event data[22] from security and network devices are collected. Filtering and classifying the data is an integral part of the data integration process. The data analysis section is where we find the patterns and rules that allow us to detect the occurrence of certain events. Finally, knowledge databases that make decisions, forecast network behavior, and respond to events receive visual and statistical outputs from data.

4. Review Outcome

By reducing the likelihood of re-identification, K-Anonymity[23] helps safeguard individuals' privacy while yet allowing for in-depth, protected data analysis. To avoid connection assaults, use this privacy-preserving technique [33] [34] When a tuple or person in a dataset cannot be discriminated from at least k-1 other tuples or people in the dataset, the dataset has been k-anonymized. For this reason, if an enemy learns the values of a person's quasi-identifier qualities, he will not be able to tell his record apart from the records of k-minus-one other people. When working with a confidential dataset, the two most popular methods for boosting K-anonymity are [35]:

- The process of substituting a specific value with a more general one is known as generalization. Examples include replacing male/female with human.

- The term "suppression" refers to the process of keeping something from being revealed. A wildcard character, such as * or @, is substituted for the value.

Data integrity is preserved by both aforementioned techniques.

Input: A proprietary data set (PT) includes some sensitive information (A) and certain pseudo-identifiable attributes (QI) for privacy.

Step 1. Choose the Data set PT from the database.

Step 2: Choose the Key, Quasi-identifier, and Sensitive attributes from the available options.

Step 3: The list of all sensitive values that must be preserved, select set A containing the most sensitive values.

Step 4: When the sensitive value of a tuple is in the set A. We must move these tuples to Table 1 if t[S] is true.

The number of distinct values for each characteristic, as well as the total number of rows carrying that value, can be determined by calculating the statistics of the quasi-attributes in Table 3. Use generalization on output table IDs; complete.

Table 3

Sample Dataset

Key Attribute	**QI Attribute**			**Sensitive Attribute**
Name	**Age**	**Sex**	**Zip code**	**Disease**
Ravi	30	M	12567	Fever
Sam	25	M	13001	Cancer
Ramesh	26	M	13001	Flu
Manav	34	M	76512	Flu
Suhari	23	F	14599	Viral
Keshav	35	M	13057	Pneumonia
Anita	35	F	17000	Fever
Hema	23	F	32451	Cancer
Reshma	27	F	14560	Flu

Table 4

Unidentified Data Set (After removing Key Attributes)

Age	**Sex**	**Zip code**	**Disease**
30	M	12567	Fever
25	M	13001	Cancer
26	M	13001	Flu
34	M	76512	Flu
23	F	14599	Viral
35	M	13057	Pneumonia
35	F	17000	Fever
23	F	32451	Cancer
27	F	14560	Flu

Releasing the original table requires first removing the key attributes. Table 4 displays the final tabulation. Yet, table 5 can be linked to other data that the attacker has access to. The following table displays external data obtained by the attacker from a Voter Registration List.

Table 5

Publicly available Voter Registration Dataset

Voter Id No	**Name**	**Age**	**Sex**	**Zip code**
QDT2398452	Ravi	30	M	12567
SAP2345918	Sam	25	M	13001
QAC4982107	Ramesh	26	M	13001

HTR4356723	Manav	34	M	76512
RAP3412090	Suhari	23	F	14599
KDE7351898	Keshav	35	M	13057
WRT2783261	Anita	35	F	17000
WER254367	Hema	23	F	32451
KYE3456123	Reshma	27	F	14560

The assailant will learn that Ramesh has the flu after comparing Tables 4 and 5. Hence, a person can be re-identified using publicly available data even if the essential identifiers have been removed. To link the information in the published table with the information in the public table is known as a Linkage Attack. As a result, linking attacks can be avoided by employing this privacy paradigm. The only information we must go on when trying to identify someone from a release is their age, gender, and possibly their address. One example of a 2-anonymous table is shown in Table 6, where k=2, indicating that at least two tuples share the identical values for the quasi-identifier characteristics. The formulae t2[S]= t3[S]= t6[S]&t5[S]= t9[S] are established here.

Table 6

Unidentifiable Data Set

Age	**Sex**	**Zip code**	**Disease**
[20-30]	M	125*	Fever
[20-30]	M	130*	Cancer
[20-30]	M	130*	Flu
[30-40]	M	765*	Flu
[20-30]	F	145*	Viral
[30-40]	M	130*	Pneumonia
[30-40]	F	170*	Fever
[20-30]	F	324*	Cancer
[20-30]	F	145*	Flu

Homogeneity/Attribute Disclosure Attacks: This occurs when the attacker is only concerned with a single value of a sensitive property and there isn't enough variance in those values. Table 4's second, third, and sixth tuples share identical values for the sensitive characteristics Age, Sex, and Zip code, suggesting that Ramesh, too, suffers from either Cancer or Pneumonia. The Background Attack: There is yet added form of assault that cannot be defended against using k-anonymity. This model supposes the attacker has no prior information.

Table 7 below provides an overview of the uses and features of top companies' big data security solutions.

Table 7

Prominent corporations Essential Functions of Big Data Security Apps

COMPANY NAME	APPLICATION	KEY FEATURES
IBM	Platform for Security Intelligence by IBM called QRadar [38]	▪ An all-inclusive, holistic strategy that employs forensic capabilities for deep visibility, custom analytics on massive amounts of structured and unstructured data for real-time correlation, and continuous insight. Together, they can be an asset in the fight against advanced persistent threats, fraud, and even insider attacks. ▪ Security intelligence data queries performed quickly. ▪ An intuitive interface for browsing and viewing massive amounts of data.
INFOSYS	Infosys Information Platform, or IIP for short [39]	▪ To cite an example: A free and public database analysis tool. ▪ In other words, it helps companies put their data to work so they can accelerate innovation and growth. ▪ Provides a full-featured data platform that can either stand alone as a big data solution or be used as an addition to existing proprietary tools by leveraging open source developments and custom

		developments.
HP	HPE Security, by Hewlett Packard Enterprise [41]	▪ In order to ensure the safety of your data, Security offers state-of-the-art tokenization and encryption. ▪ Both organized than unstructured data can be encrypted and tokenized with utmost security using Security. ▪ Secure analytics, reduced scope, and PCI compliance at a low cost. ▪ Global industry leaders rely on it to safeguard their reputations and cut costs.
ORACLE	Oracle's Database Security Analysis Tool (DBSAT) [42]	▪ Find database security configuration issues quickly. ▪ Encourage the use of sound security methods. ▪ If you use Oracle Databases[37], you should take steps to strengthen their security immediately. ▪ Conceal more of your system's vulnerable areas and reduce your risk exposure. ▪ Assist businesses in realizing the full potential of big data analytics.
VORMETRIC	Data Security Platform with Vormetric Analysis [43]	Provides fine-grained controls, strong encryption, and all-encompassing coverage that businesses require to protect their massive data. ▪ Provides the ability for security teams to make use of a unified set of controls to improve productivity and policy observance. ▪ Features include encryption, key management, and authorization for large datasets.

5. Conclusion

Data privacy and security concerns are major roadblocks for scientists studying big data. A few examples of internationally acclaimed work in this area have been briefly discussed. In the context of large healthcare data privacy and security, we have also discussed challenges related to privacy and security at each stage of the big data lifecycle. Recent healthcare privacy preservation approaches were explored, including encryption and anonymization's application and limitations for protecting sensitive patient information. Yet more methods exist, such as "hiding a needle in a haystack," "attribute-based encryption," "access control," "homomorphic encryption," "storage path encryption," and so on. The catch is that the difficulty is always imposed. Thus, we can see that the successful solutions in privacy and security in the era of large healthcare data are the future path we should be heading towards. There is also a need to improve privacy protections. As the Internet of Things expands rapidly, there is a corresponding decline in quality as more devices are connected. So, researchers shouldn't have to compromise on data quality too much to use privacy-preserving methods.

References

1. Jha A, Dave M. and Madan, S. A Review on the Study and Analysis of Big Data using Data Mining Techniques, International Journal of Latest Trends in Engineering and Technology (IJLTET), vol.6(3), 2016.
2. Jha A, Dave, M. and Madan, S. 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(7), 2016.
3. Q, Jing. Security of the internet of things: perspectives and challenges. Vol.20(8), pp.481–50., 2016.
4. De la Torre, Isabel; García-Zairian, Begoña; and López-Coronado, Miguel. Analysis of Security in Big Data Related to Healthcare," Journal of Digital Forensics, Security and Law. Vol.12(3), 2017. DOI: https://doi.org/10.15394/jdfsl.2017.1448
5. Saurabh Pandey, Rashmi Pandey. Medical (Healthcare) Big Data Security and Privacy Issues International Journal of Scientific & Engineering Research, vol.9(2), 2018.
6. W. Nicholson Price, II, JD, PhD and I. Glenn Cohen, JD. Privacy in the Age of Medical Big Data" Nat Med. 2019 January; vol.25(1), pp. 37–43, 2019.https://cloudsecurityalliance.org/media/news/csa-big-data-releases-top-10-security-privacy challenges/.

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

8. Okman, L., Gal-Oz N., Gonen Y, Gudas E. and Abramov J. Security Issues in NoSQL Databases in Trust Com IEEE Conference on International Conference on Trust, Security and Privacy in Computing and Communications, pp.541-547,2011.
9. Apparao, Yannam and Laxminarayan Amma, Kadiyala. Security Issue on Secure Data Storage and Transaction Logs In Big Data" in International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 2015.
10. Singh, Reena and KunverArif Ali. 2016. Challenges and Security Issues in Big Data Analysis, IJIRSET, vol 5(1), 2016.

11. Sedayao J, Bhardwaj R. Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. In: Big data congress. 2014.
12. Jung K, Park S, Park S. Hiding a needle in a haystack: privacy preserving Apriori algorithm in MapReduce framework PSBD'14, Shanghai. pp. 11–17, 2014.
13. Lu R, Zhu H, Liu X, Liu JK, Shao J. Toward efficient and privacy-preserving computing in big data era. IEEE Netw. vol.28, pp. 46–50,2014.
14. Microsoft differential privacy for everyone. 2015. http://download.microsoft.com/…/Differential_Privacy_for_Everyone.pdf.
15. Zhang X, Yang T, Liu C, Chen J. A scalable two-phase top-down specialization approach for data anonymization using systems, in MapReduce on cloud. IEEE Trans Parallel Distrib. vol.25(2), pp. 63–73, 2014.
16. 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.
17. Mohammadian 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.
18. 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.
19. Wei L, Zhu H, Cao Z, Dong X, Jia W, Chen Y, Vasilakos AV. Security and privacy for storage and computation in cloud computing. Inf Sci. 2014;258:371–86, 2014
20. Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: IEEE 3rd international conference on cloud computing, 2010.
21. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowle Based Syst, vol.;10, pp.571–588,2002
22. Samrati P. Protecting respondents' identities in micro data release. IEEE Trans Knowle Data Eng. vol.13, pp.10–27, 2001.
23. Truta TM, Vinay B. Privacy protection: p-sensitive k-anonymity property. In: Proceedings of 22nd international conference on data engineering workshops. pp. 94, 2006.
24. Spruill N. The confidentiality and analytic usefulness of masked business microdata. In: Proceedings on survey research methods. pp. 602–607,1983.
25. Chawala S, Dwork C, Sheny FM, Smith A, Wee H. Towards privacy in public databases. In: Proceedings on second theory of cryptography conference. 2005.
26. Science Applications International Corporation (SAIC). Role-based access control (RBAC) Role Engineering Process Version 3.0., 2004.
27. Zhou H, Wen Q. Data security accessing for HDFS based on attribute-group in cloud computing. In: International conference on logistics engineering, management, and computer science (LEMCS 2014). 2014.
28. Linden H, Kalra D, Hasman A, Talmon J. Inter-organization future proof HER systems—a review of the security and privacy related issues. Int J Med Inform. 2009; 78:141–60, 2009.
29. Marchal S, Xiuyan J, State R, Engel T. "A big data architecture for large scale security monitoring", Big Data (BigData Congress), Anchorage, AK. pp. 56–63,2014.
30. Duygu ST, Ramazan T, Seref S. A survey on security and privacy issues in big data. In: The 10th international conference for internet technology and secured transactions (ICITST-2015). 2015.
31. Liu L, Lin J. Some special issues of network security monitoring on big data environments. Dependable, Autonomic and Secure Computing (DASC), Chengdu. pp.10–5, 2013.

32. KenigBatya andTassaTamir. A practical approximation algorithm for optimal k-anonymity, Data Mining Knowledge Discovery, Springer. 2011.
33. Sweeney, L. K-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty Fuzziness Knowledge based Systems. 2002.
34. Samarati. P. Protecting respondents' identities in microdata release. IEEE Trans. on Knowledge and Data Eng., vol.13, pp.1010–1027, 2001.
35. https://www-01.ibm.com/software/se/security/bigdata/
36. https://msdn.microsoft.com/en-us/library/dn749804.aspx
37. https://docs.oracle.com/cd/E76178_01/
38. https://www.thalesesecurity.com/products/data-encryption/vormetric-data-security-platform