IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306215|

Security and Privacy Concerns in Big Data

Isha Singh, Vaishnavi Joshi, Tanishq Vaidya

Department of MCA, Vivekanand Education Society's Institute of Technology, Mumbai University, Mumbai, India

ABSTRACT: Big Data's ascent has revolutionized numerous industries by facilitating sophisticated analytics and data-informed decision-making. But there are serious security and privacy issues because of the massive amount, diversity, and velocity of Big Data. These difficulties include insider threats, data breaches, illegal access, and privacy problems brought on by the anonymization and aggregation of data. This study conducts a thorough analysis of these issues as they stand today. Strong data governance frameworks, better anonymization techniques, greater access controls, and cutting-edge encryption techniques are some of the suggested remedies. The necessity of ongoing innovation is emphasized as the implementation and advantages of these approaches are examined. To guarantee the responsible and secure use of big data, future research topics are presented, with a focus on the development of effective cryptographic techniques and ethical data stewardship.

KEYWORDS: Big Data, Security, Privacy, Data Breach, Encryption, Data Anonymization, Data Masking

I. INTRODUCTION

Big Data's introduction has completely changed how industries function, spurring creativity and enabling sophisticated analytics that support data-driven decision-making. Big Data is the term used to describe vast amounts of organized and unorganized data that are produced at previously unheard-of speeds from a variety of sources, including financial transactions, social media, IoT devices, and medical information. There is a ton of opportunity in this data flood to find insights, streamline procedures, and open new business avenues. Nevertheless, the very qualities that add value to big data also pose significant privacy and security risks.

Big Data settings present a variety of security risks, such as ransomware and malware assaults, insider threats, illegal access, and data breaches. The deployment of strong security measures is frequently made more difficult by the distributed and heterogeneous nature of Big Data systems. For example, data is often held in many locations and accessed by multiple stakeholders, which raises the possibility of data breaches and illegal access. Notable events like the Facebook-Cambridge Analytical controversy and the Equifax hack highlight how important it is to have strict security procedures in place to safeguard private data.

Concerns about privacy are just as urgent, especially considering how sensitive a large portion of the data being gathered and processed is. Although crucial, data anonymization approaches frequently fall short in stopping re-identification attacks, in which identities are discovered through cross-referencing anonymized data with other datasets. Moreover, the combination of data from several sources may unintentionally reveal personal information, hence presenting serious privacy hazards to individuals. Complying with strict data protection laws like the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR) increases complexity and calls for careful data governance and stewardship.

An extensive examination of the situation of privacy and security concerns in big data contexts is given in this study. It suggests doable tactics to improve Big Data systems' security and privacy, guaranteeing their responsible and secure usage. Additionally, future goals for research are indicated, with an emphasis on developing effective cryptographic techniques, AI-driven security measures, and responsible data stewardship. The paper endeavours to make a thorough contribution to the ongoing endeavour of protecting confidential data and upholding public confidence in a society that is becoming more and more reliant on data.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306215|

II. CURRENT STATE OF SECURITY AND PRIVACY IN BIG DATA

A. Security Concerns

1. Data Breaches

Data breaches in large data environments are becoming more frequent. The enormous volume of data that businesses gather and retain is the main cause of the 45% annual increase in data breaches, according to recent figures. These security lapses may lead to large monetary losses, harm to one's reputation, and legal implications. Prominent events, like the 2017 Equifax hack that compromised the personal information of over 147 million people, emphasize how important it is for Big Data systems to have strong security safeguards.

Data breaches in large data environments are becoming more frequent. The enormous volume of data that businesses gather and retain is the main cause of the 45% annual increase in data breaches, according to recent figures. These security lapses may lead to large monetary losses, harm to one's reputation, and legal implications. Prominent events, like the 2017 Equifax hack that compromised the personal information of over 147 million people, emphasize how important it is for Big Data systems to have strong security safeguards.

2. Unauthorized Access

Because big data systems are distributed, there is always a risk of unauthorized access. Since traditional role-based access control (RBAC) systems do not take into consideration the dynamic and complex nature of big data environments, they frequently fail to provide sufficient security. For instance, to prevent unwanted access, cloud-based data storage solutions frequently need more detailed and context-aware access controls.

When sensitive information is exposed to unauthorized users, it is known as data leakage due to insufficient access controls. Access control is made more difficult by the growing usage of cloud services and outside vendors, since businesses need to make sure that these partners follow strict security guidelines. Additionally, strong access control mechanisms are required to protect data integrity due to the proliferation of IoT devices and mobile applications that generate large amounts of data.

3. Insider Threats

In big data environments, 34 percent of data breaches are caused by insider threats. These risks occur when privileged access holders abuse their login information to gain access to or steal confidential information. Prominent incidents like the Snowden revelations highlight the possible harm insider threats may cause. Companies must strike a balance between reducing the possibility of abuse and providing workers with enough access to enable them to do their jobs well.

Because insiders frequently have authorized access to data and systems, identifying and mitigating insider threats can be difficult. Continuous monitoring and behavioural analytics can be used to spot unusual activity that could be an indication of an insider threat. But putting these policies into action will cost a lot of money in terms of both experience and technology.

4. Malware and Ransomware

The threat to Big Data systems is severe from Malware and Ransomware attacks. For example, organisations around the world were affected by the WannaCry ransomware attack in 2017, leading to data loss and significant operational disruption. These malicious activities are attractive targets for large data environments because of their size and interconnected nature. It is therefore critical to ensure a high level of malware protection.

Large volumes of data can be encrypted by ransomware attacks, making it unusable until a ransom is paid. Big data systems are distributed, which can speed up the spread of malware and increase its impact. To defend against such attacks, advanced threat detection and response mechanisms must be put into place. Organizations can also recover from ransomware incidents without having to pay ransoms by implementing incident response plans and routine backups.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306215|

B. Privacy Concerns

1. Data Anonymization

Although crucial, anonymization techniques are not infallible. Attacks using re-identification can de-anonymize data, which presents serious privacy concerns. For instance, by cross-referencing anonymized data with other publicly available datasets, the renowned Netflix dataset challenge from 2006 resulted in the identification of multiple users. This demonstrates how ineffective conventional anonymization methods are at preventing privacy violations.

Organizations need to use more advanced anonymization strategies that fend off re-identification attempts in order to improve privacy protection. Finding a balance between data privacy and usefulness is still very difficult, though. A successful anonymization process should protect the data's analytical value while guaranteeing that no individual can be identified.

2. Data Aggregation

Private information may unintentionally be made public through data aggregation from several sources, particularly when merged with other datasets. In big data environments, where different data sources are often integrated for analysis, this concern is magnified. Strict privacy controls are required due to the possibility of sensitive information being revealed through data aggregation.

Although data aggregation raises the possibility of privacy violations, it can also highlight trends and insights that are hidden in isolated datasets. It is imperative to guarantee that aggregated data is suitably anonymized and safeguarded. Data aggregation poses privacy risks that require organizations to implement data governance frameworks.

3. Regulatory Compliance

Adherence to regulations like the CCPA and GDPR is imperative, albeit difficult. Strict data protection measures are required by these regulations, and noncompliance carries steep fines. It takes a lot of work and ongoing oversight to make sure big data practices comply with these laws. The severity of the financial penalties and legal ramifications associated with non-compliance underscores the necessity of following these regulations.

Businesses need to manage a complicated regulatory environment while making sure that their data practices adhere to the laws of several different jurisdictions. This frequently entails putting in place thorough data protection measures, carrying out frequent compliance audits, and keeping up with regulatory changes. There can be serious financial and reputational consequences from breaking data protection laws.

4. Data Ownership and Consent

In big data environments, it is imperative to understand the concept of data ownership and obtain informed consent from data subjects. People should oversee their personal data, and companies should make sure that consent is obtained and that data collection procedures are open and clear. Respecting people's right to privacy necessitates giving careful thought to the ethical implications of data usage.

Providing information about data collection, usage, and sharing practices in a clear and accessible manner is essential to ensuring informed consent. Organizations need to put in place consent management systems that make it simple for people to give, withhold, or change their consent. This fosters openness and confidence, both of which are necessary to maintain good relations with data subjects.

III. PROPOSED SOLUTIONS

A. Homomorphic Encryption

Homomorphic encryption ensures data privacy even while processing by enabling computations on encrypted data without first decrypting it. Despite being computationally demanding, this approach offers a reliable way to safeguard sensitive data all the way through its lifecycle. The goal of recent developments has been to lessen the performance overhead of homomorphic encryption, increasing its viability for real-world uses.

Data security can be greatly improved by implementing homomorphic encryption, particularly in settings where data is routinely processed by outside parties. Even though it is computationally complex, research is still being done to create

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306215|

hardware accelerators and more effective algorithms to increase performance. To safeguard sensitive data while it is being processed, organizations should think about incorporating homomorphic encryption into their security plans.

B. Attribute-Based Access Control (ABAC)
Compared to RBAC, ABAC offers more precise access control by taking user characteristics and external factors into account. This technique improves security in intricate large data environments by enabling more dynamic and context-aware access restrictions. ABAC offers a flexible and reliable access control by adapting to several variables, including the user's role, location, time of access, and the sensitivity of the data. Method.

A thorough understanding of user characteristics and environmental settings is necessary for ABAC implementation. To properly enforce ABAC, organizations need to create comprehensive policies and make use of cutting-edge technologies. This method can guarantee that data is accessible by authorized users only in the proper circumstances and drastically lower the danger of unauthorized access.

C. Multi-Factor Authentication (MFA)
By requiring several verification techniques, such as something the user knows (password), something they have (security token), and something they are (biometric verification), MFA improves security. The danger of unwanted access can be greatly decreased by putting MFA into practice.

To provide an additional layer of protection, organizations should require MFA for access to sensitive data and vital systems. MFA adds another degree of protection, increasing the difficulty for intruders to obtain unwanted access. Organizations can improve security without appreciably sacrificing user experience by combining various authentication elements. Using multi-factor authentication (MFA) to access sensitive data and critical systems can significantly reduce the possibility of credential-based attacks.

D. Differential Privacy
To prevent the identification of specific data points, differential privacy adds noise to data queries. Strong privacy guarantees are provided by this technique, which makes sure that the addition or deletion of a single data point does not materially alter the result of any analysis. Companies such as Apple and Google have embraced differential privacy extensively to protect user privacy without sacrificing data utility.

Organizations can gain important insights from data while safeguarding individual privacy thanks to differentiated privacy. Differential privacy guards against re-identification attacks and makes sure that individual data points are not compromised by analysis results by introducing controlled noise. To successfully strike a balance between data utility and privacy, organizations should think about implementing differential privacy.

E. K-Anonymity
K-Anonymity makes sure that, when it comes to some identifying qualities, every record in a dataset may be identified with at least k-1 other records. Adding l-diversity and t-closeness to this technique can improve privacy protection by adding more layers. Because K-Anonymity ensures that no record is unique inside the dataset, it can be very helpful in preventing re-identification attempts.

Careful evaluation of dataset characteristics and identifying attributes is necessary for the implementation of k-anonymity. By guaranteeing that sensitive attributes show diversity, and that the distribution of sensitive values stays close to the total data distribution, enhancing k-anonymity with l-diversity and t-closeness significantly lowers the chance of re-identification. These methods can preserve the usefulness of data while successfully protecting individual privacy.

F. Data Masking
Sensitive information is hidden in datasets used for training or testing that are not intended for production, thanks to data masking. This method makes sure that sensitive data is safeguarded even in the event that masked data is accessed. For example, to prevent unauthorized publication, names and Social Security numbers in a dataset holding client details can be substituted with fictional but plausible values.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306215|

While enabling the use of datasets for a variety of reasons, data masking safeguards sensitive information. This method guarantees that data that has been masked has its format and structure, which makes it appropriate for testing and development. Organizations can lower their exposure risk and continue to comply with data protection laws by hiding sensitive data.

IV. FUTURE RESEARCH

Future work should concentrate on improving differential privacy strategies to better balance privacy and usefulness, creating more effective homomorphic encryption algorithms to lower performance overhead, and combining AI-driven anomaly detection systems to proactively detect security concerns. Furthermore, it will be essential to investigate how quantum computing may affect existing cryptography techniques and to create quantum-safe cryptographic standards. In order to guarantee that big data is utilized in ways that advance society while preserving individual rights and privacy, it will also be crucial to look into the ethical implications of big data analytics and create frameworks for responsible data stewardship.

Improvements in homomorphic encryption have the potential to greatly increase data security, but further study is required to make it more effective. Enhancing differential privacy strategies can assist strike a balance between privacy and data value, allowing companies to gain insightful information without jeopardizing the privacy of individuals. Systems for anomaly detection powered by AI can proactively spot security risks, improving data security overall. Future-proofing data security against quantum computing threats requires research into quantum-safe cryptography standards.

V. CONCLUSION

To preserve confidence and guarantee data integrity, big data systems security and privacy are essential. Organizations can greatly reduce the risks associated with security and privacy by putting in place strong data governance, better access restrictions, improved anonymization techniques, and advanced encryption. In order to safeguard sensitive data and guarantee the moral application of big data, it will be imperative to continue study and innovate in order to create new methods and best practices. Organizations must maintain a culture of continuous development and compliance by being watchful and proactive in resolving security and privacy issues.

A thorough foundation for improving security and privacy in large data environments is provided by the suggested solutions. Data is safeguarded throughout its lifecycle by advanced encryption techniques including quantum-safe cryptography and homomorphic encryption. Insider risks and unwanted access are stopped by enhanced access controls like MFA and ABAC. Newer anonymization methods that preserve data value while protecting individual privacy include differential privacy and synthetic data generation. Strong data governance procedures, such as masking, lifecycle management, and data audits, guarantee regulatory compliance and encourage responsible data stewardship.

REFERENCES

1. A Big Data Security using Data Masking Methods. August 2017 Indonesian Journal of Electrical Engineering and Computer Science 7(2):449-456
2. Security and Privacy Challenges in Big Data Era. August 2016 International Journal of Control Theory and Applications 9(43):437-448
3. Hu, V. C., Ferraiolo, D. F., Kuhn, D. R., & Schnitzer, A. (2013). Attribute-Based Access Control. Computer, 48(2), 85-88.
4. Survey on Fully Homomorphic Encryption, Theory, and Applications. October 2022Proceedings of the IEEE PP(99):1-38

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details