ISSN No:-2456-2165

An Information Security Risk Assessment Framework for Big Data Environment

Keerti Dixit, Dr. Umesh Kumar Singh, Dr. Bhupendra Kumar Pandya
Institute of Computer Science Vikram University, Ujjain

Abstract:- Big data has been utilized by numerous businesses to produce straightforward goods and offer improved customer insights via predictive technology like artificial intelligence. Big data is a discipline that primarily focuses on the extraction and systems analysis of enormous data sets to assist organizations in identifying trends. Big Data is now widely used by businesses to boost development across a range of functional areas and increase their capacity for managing sizable client databases. Due to the increased demand for information management specialists brought on by big data, a growing number of software corporations are making investments in companies that focus on data management and analytics. Nevertheless, big data management is threatened by the problem of data protection or privacy. This research paper presents security Challenges and Privacy Issues of Big Data. Further, we created a framework for assessing information security risks in the big data world.

Keywords:- *Information Security Risk Assessment, Privacy, Security, Big Data.*

I. INTRODUCTION

Big data refers to the various sources of information that generate data, including social media accounts, email communications, closed-circuit television (CCTV) sensors, Internet of Things (IoT) devices, and various platforms that gather client information. [1,2] claimed that industrialized nations are using big data technologies more and more to analyze extensive consumer datasets and generate insightful data. Big Data has many benefits for various organizations, including enhancing corporate or healthcare decision-making processes and enhancing consumer experiences through customer data analysis[3]. Identifying and producing knowledge that can be used to address problems is the primary objective of Artificial Intelligence (AI) technology in the field of Big Data [4]. [5] asserts that there are countless possibilities opened up by big data, some of which include the incorporation of instruments that improve the effectiveness of company operations.

Artificial intelligence and data mining are used as techniques in Big Data analytics to speed up the acquisition of customer insights aimed at enhancing their brand loyalty [6].

Big Data poses privacy and security issues, according to [7], which have an impact on how organizations manage their data and whether or not customers are ready to give their data for free services. In this paper, we discussed the issues related to big data privacy and security.

II. PRIVACY AND SECURITY ISSUES IN BIG DATA

In terms of big data, privacy and security are critical concerns. Data can be easily compromised in the absence of a Big data security model. As a result, this segment concentrates on privacy and security concerns.

- **Privacy:** Information privacy is the benefit of having some obligation on the way personal information is assembled and used. Information privacy indicate to a person's or a group's capability to store personal information sharing with persons other than those to whom they disclose it. The recognition of user information at the time of transmission over the Internet is one important privacy issue [8].
- **Security:** Information and information assets by help of training and technology from the various sources can be defended by the process of Security.
- **Privacy vs. security:** Data privacy is concerned with how personal data is utilised and regulated, including setting policies to guarantee that customers' private information is collected, shared, and used appropriately. Data protection and financial gain from stolen data are the main goals of security [9]. Security is crucial for data protection, but it falls short when it comes to protecting personal information.
- **Privacy and Security issues:** Big data technology's security procedures are frequently lacking. Finding reliable security solutions for features like autotiering, parallelism, and other characteristics has proven to be a difficult task. Issues such as invasion of privacy, disk drive storage complexity, invasive marketing, and others have made deploying Big Data Analytics tools for Big data solutions and applications difficult [10].

A. Insecure Computation:

Attackers utilise untrusted computational programmes to extract and output vulnerable information through data sources. Insecure computing can pollute your data and lead to erroneous analysis or prediction findings in addition to causing information leaks. It can result in a Denial of Service (DoS) that prevents your big data solution from using a massively parallel programming language.

B. Input Validation and Filtering:

Data validation is crucial and required because big data requires input from many different sources. Finding reliable and unreliable data sources is required for this. It should also be able to identify authentic data from malicious or erroneous data. These problems are historical and can be found in conventional databases.

The input validation data filtering on the incoming batch of data is tremendously challenging as there is huge volume of continuous data flow in the gigabytes and terabytes in big data. Signature-based data filtering has drawbacks as well, such as the inability to filter rogue or malicious data with a behavioral component. When a substantial volume of fraudulent data is injected into a dataset, it has a huge impact on the final result. Individual specialized algorithms must be devised to cope with such scenarios because signature-based data filtering is not capable of monitoring such attacks.

C. Granular Access Controls:

Big data has earlier been built for performance and scalability, with less regard for security. Traditional databases include extremely detailed table, row, and cell access controls, which have mostly vanished in big data solutions. Ad-hoc searches, in which users can extract sensitive information from data using ad-hoc queries, present another additional difficulty to big data solutions. Access control is deactivated by default, despite the fact that it is offered by a big data solution.

D. Insecure Data Storage:

Data collection across numerous nodes makes it challenging to authenticate, authorise, and encrypt data at those nodes. Real-time data encryption may not be advantageous if a solution offers it because it can have an impact on performance. Because it is disabled by default, secure communication between nodes, middlewares, and end users must always be enabled.

E. Privacy Concerns in Data Mining and Analytics:

Data mining and analytics are used to monetize Big Data, and releasing the analytical results poses a number of issues, including invasion of privacy, invasive marketing, and unintended information disclosure. Several examples include AOL's publication of anonymized search logs that may easily identify individuals, which is quite problematic.

III. PRIVACY-PRESERVING APPROACHES FOR BIG DATA

In this segment, we look at few of the available privacy-preserving strategies for large data and discuss the issues that may occur when these approaches are used to huge data. Encryption, de-identification, aggregation, perturbation, and suppression are some of the common privacy-preserving strategies we'll look at.

A. De-identification:

De-identification is a generic method for preventing a person's identity from being linked to other data. To protect user privacy, the data should be sanitized (to eliminate noise and inconsistent data, to manage missing data fields, and so on) and then integrated (to merge data from numerous sources) before being published. De-identifying datasets often involves masking or erasing identifiers (IDs) like name, passport ID, and SSN, as well as generalizing quasi-identifiers like age and date of birth. In some cases, de-identification may entail preserving identifying information which can be re-linked by a trustworthy third party. De-identification [11] can make data analytics and mining more successful and versatile than other privacy approaches such as encryption, aggregation, and suppression, synthesis and swapping. Traditional de-identification approaches, such as k-anonymity, l-diversity, and t-closeness, have been used to preserve privacy in huge data once their functionality has been improved [11]. Because an entity (person) has numerous qualities, an attacker can acquire a lot of information about it in the big data era, which usually provides deep background knowledge and increases the risk of re-identification. As a result, due to various restrictions, de-identifications are insufficient to protect large data privacy. Using these strategies to de-identify genetic information is difficult [12].

B. Aggregation:

In big data, data aggregation identifies relevant search query data and individual data discoveries in a consolidated format that is comprehensible and valuable to the end user or application. Data aggregation is a popular strategy for gathering data through a data source in the privacy context, which is based on some homomorphic encryptions [13]. Data aggregation lowers disclosure risks by converting atypical records [14], which are the most vulnerable, into typical records. It typically works with massive data or data marts that don't provide much information value overall. Aggregation can preserve privacy during the data collecting and preservation phases, but it does not play a significant role during the data posting phase [15]. In privacy, there are two fundamental issues: preserving aggregation and keeping privacy. The first issue is a misunderstanding of inference, which implies that aggregation is common but not applicable at the individual level. The second issue is obstinacy, which implies that aggregation is intended for a precise purpose and is therefore inflexible for generic applications. Such a biassed technique is unsuited for large-scale knowledge extraction.

C. Suppression:

Suppression is the practice of replacing some values with a specific value and, in extreme cases, erasing entire tuples or cell values [16]. Suppression operations are used in agencies to remove sensitive variables from modified datasets. Cell, record, and value suppressions are common suppression schemes. The use of compression in large data can substantially decrease data quality and alter overall statistics, effectively leaving the data unusable. The fundamental issue with cell suppression is that it causes information in the table to be distorted by deliberately suppressing cells. As a result, data providers can derive false and biassed inferences from the cell values that are replicated. [14].

D. Perturbation:

Perturbation is the procedure of transforming original data values with some febricated data values, which is the substitution of original data values with values generated from probability distributions when there is a significant danger of exposure. The probability distribution approach is the first method, which collects the data and changes it with data within the distribution sample and the distribution itself. The value distortion approach is a method of perturbing data using multiplicative or additive noise or other randomized processes. Privacy protection in the case of multiplicative or additive noise is determined by the type of the noise distribution, with large variance providing the best security. Swapping values among records in the dataset is part of the perturbation family. When perturbation methods are used in a big data context, they have some limitations. For providing real-time reaction in large-scale datasets, velocity is a significant challenge, and perturbation over large-scale datasets is typically complex and time expensive. As a result, some perturbation methods are unable to respond quickly to users or consumers. When perturbation is used to cluster large datasets, a misclassification problem is discovered [16].

E. Encryption:

Encryption encrypts a communication so that it can only be viewed by authorized persons. Encryption does not restrict access; rather, it prevents the raw message from being transmitted. Data (image/text) transit [17] and data storage [18] are the two types of encryption utilised in applications [18]. Operations on encrypted data are typically complex and difficult in big data analytics. High volume is a big difficulty in big data, because processing large amounts of encrypted data takes time and might be a real-time response barrier in big data analytics.

IV. SECURITY APPROACHES FOR BIG DATA

The following is a list of common big data security techniques:

A. Encryption:

Big data encryption solutions can be used to safeguard data in transit and at rest thanks to large amounts of data. Additionally, businesses want to encrypt data produced by both humans and machines. Finally, different types of big data storage, such as NoSQL databases and Hadoop distributed file systems, should be compatible with encryption technologies.

B. User Access Control:

User access control is a fundamental tool for network security. Lack of appropriate access control methods may be detrimental for huge data systems. A robust user control policy needs automated role-based settings and policies. Policy-driven access control shields big data platforms from insider threats by automatically managing complex user control levels, such as multiple administrator settings.

C. Intrusion Detection and Prevention:

Intrusion attempts benefit from the distributed design of big data. An Intrusion Prevention System (IPS) can protect large data platforms from vulnerability assaults by keeping track of network traffic. Commonly put directly behind the firewall, the intrusion prevention system (IPS) isolates the intrusion before it can do any harm.

D. Centralized Key Management:

Key management is the practise of protecting cryptographic keys from theft or unauthorised usage. Centralised key management is more effective than distributed key management or application-specific key management. In centralised management systems, it is possible to access rules, audit logs, and secure keys from a single location. A trustworthy key management system is essential for companies that deal with sensitive data.

V. RESEARCH ISSUES

Big data is still a new technology, and people aren't fully aware of the security concerns it poses. Finally, the security of Big data must be given top priority. Initiatives in any field should be completely aware of the features of privacy issues as well as the need of ensuring data privacy. In this section, we'll look ahead to the future of Big Data privacy and security.

- **Enhance the supervision of sensitive data:** The accumulation of enormous amounts of data increases the risk of sensitive data being exposed. Meanwhile, the uncontrolled use of big data has raised the potential of information leakage. As a result, authorities are expected to speed up the adoption of Big data security development plans and strong security systems. As a result, it is critical to accelerate the classification and protection of information systems, as well as to explain the scope of data security in critical sectors.
- **Research and develop Big data security key technologies:** Develop secure encryption and trusted computing approaches for distributed storage, segregated management, multiplatform sharing, and other Big data characteristics. Decentralization, de-trust, and breaking information asymmetry are potential directions for cloud storage, such as the BlockChain scheme, which is used to store bitcoin transactions. In addition, striking a balance between safety and efficiency is a critical issue that must be addressed in the future. One of the trends in the coming years will be how to increase the efficiency of the security scheme in the process of analysis and search, based on the premise of ensuring safety.

VI. AN INFORMATION SECURITY RISK ASSESSMENT FRAMEWORK FOR BIG DATA ENVIRONMENT

Three phases of activity and a practical strategy for securing the Big Data environment are included in the Information Security Risk Assessment Framework. In order to reduce the possibility of security breaches, the proposed Framework calls for identifying what makes a system vulnerable.

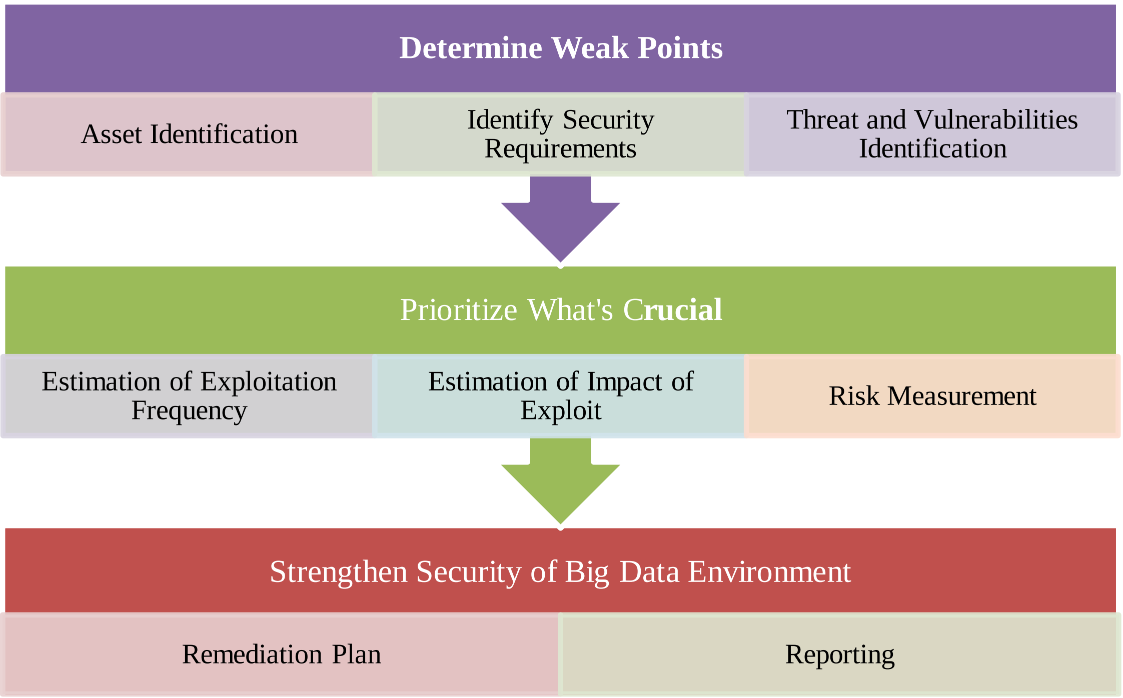
Determine Weak Points
Asset Identification
Identify Security Requirements
Threat and Vulnerabilities Identification
Prioritize What's Crucial
Estimation of Exploitation Frequency
Estimation of Impact of Exploit
Risk Measurement
Strengthen Security of Big Data Environment
Remediation Plan
Reporting

Fig. 1: Information Security Risk Assessment Framework for Big Data Environment

A. *Phase 1: Determine Weak Points*

Finding weak spots is the aim of this stage. The Risk Assessment Framework requires that the assets be precisely defined. In this stage of the suggested architecture, the parameters and content of the asset to be appraised are set. In the proposed framework, information is regarded as a resource. Then, the scope of the risk assessment effort is established, and the data required to demonstrate the risk is given. Data concerning the system interface, network connections, data and information, hardware, software, and data must be provided as input for this stage. The most common combinations of attack vector, attack goal, and attackers that potentially lead to the compromise of an asset will be listed, and these combinations will be used to create threat scenarios.

B. *Phase 2: Prioritise what is Crucial*

Finding the locations that present the highest risk is the emphasis of the second phase. This stage involves calculating the likelihood that a vulnerability will be exploited by an attacker. The frequency of exploits will be established.

The effect measures how much a targeted system could be negatively impacted by the exploitation of a configuration defect as well as how much confidentiality, integrity, and availability would be compromised. Exploit frequency and impact can be combined to produce a quantitative security risk level.

C. *Phase 3: Strengthen security of Big Data Environment*

Creating a repair strategy and, ultimately, creating thorough reporting to monitor recurrent risk measurement efforts are the main goals of the third phase.

VII. CONCLUSION

Big data approaches' ultimate goal is to be able to find meaningful and usable information in a timely manner. To develop effective analytics, there is a demand for the ability to evaluate a large amount of diverse data as well as the ability to identify, store, access, and retrieve large amounts of data. Data is becoming more vulnerable to cyber attacks since it grows exponentially every day.

In this work an overview of big data technology is being given. We have identified a number of criteria that have an impact on big data privacy and security. Following that, we explored a variety of challenges, issues, and potential solutions in the domain of big data privacy and security. Further, to protect the Big Data environment, we have created an information security risk assessment framework.

REFERENCES

[1.] Wu Y, Huang H, Wu N, Wang Y, Bhuiyan MZA, and 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

[2.] Suoniemi S, Meyer-Waarden L, Munzel A, Zablah AR, and Straub D (2020). Big data and firm performance: The roles of marketdirected capabilities and business strategy. Information and Management, 57(7): 103365. https://doi.org/10.1016/j.im.2020.103365

[3.] Atoum I and Al-Jarallah NA (2019). Big data analytics for valuebased care: Challenges and opportunities. International Journal of Advanced Trends in Computer Science and Engineering, 8(6):

3012-3016. https://doi.org/10.30534/ijatcse/2019/55862019

[4.] Reis T, Bornschlegl MX, and Hemmje ML (2020). Big data analysis, AI, and visualization workshop: Road mapping infrastructures for artificial intelligence supporting advanced visual big data analysis. In the International Conference on Advanced Visual Interfaces, Association for Computing Machinery, Salerno, Italy: 1-2. https://doi.org/10.1145/3399715.3400860

[5.] Sardi A, Sorano E, Cantino V, and Garengo P (2020). Big data and performance measurement research: Trends, evolution and future opportunities. Measuring Business Excellence. https://doi.org/10.1108/MBE-06-2019-0053

[6.] Davenport T, Guha A, Grewal D, and Bressgott T (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(1): 24-42. https://doi.org/10.1007/s11747-019-00696-0

[7.] Martin KD, Borah A, and Palmatier RW (2017). Data privacy: Effects on customer and firm performance. Journal of Marketing, 81(1): 36-58. https://doi.org/10.1509/jm.15.0497

[8.] Porambage P, et al. “The quest for privacy in the internet of things”. IEEE Cloud Comp. 2016;3(2):36–45.

[9.] Jing Q, et al. “Security of the internet of things: perspectives and challenges”. Wirel Netw. 2014;20(8):2481–501.

[10.] Renu Bhandari, Vaibhav Hans and Neelu Jyothi Ahuja “Big Data Security – Challenges and Recommendations” International Journal of Computer Science and Engeneering vol 4(1), pp (93-98) Jan 2016

[11.] Wang, J., Li, H., Guo, F., Zhang, W., & Cui, Y.. 2019, “D2D big data privacy-preserving framework based on (a, k)-anonymity model”, Mathematical Problems in Engineering.

[12.] McGuire, A. L., & Gibbs, R. A. (2006) “No longer de-identified. Science”, 312, 370–371.

[13.] Magkos, E., Maragoudakis, M., Chrissikopoulos, V., & Gritzalis, S. (2009), “Accurate and large-scale privacy-preserving data mining using the election paradigm”, Data & Knowledge Engineering, 68, 1224–1236.

[14.] Zhu, D., Li, X. B., & Wu, S. (2009), “Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining”, Decision Support Systems, 48, 133–140.

[15.] Xiao, X., & Tao, Y., 2007, “M-invariance: Towards privacy preserving re-publication of dynamic datasets”, Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing China, pp. 689–700.

[16.] Oliveira, S., & Zaiane, O, “Data perturbation by rotation for privacy-preserving clustering”, 2004.

[17.] Huang, X., & Du, X.. 2014, “Achieving big data privacy via hybrid cloud”, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS); Toronto, Ontario, Canada, pp. 512–517.

[18.] Yang, Y., Zheng, X., et al (2019) “Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system”, Information Sciences, 479, 567–592.