DATA SANITIZATION AND PRIVACY PRESERVATION METHOD ON EDGE NODES

[1]LISA BISWAS, [2]RASHMI AGARWAL, [3]VINAY BY

[1,2,3]Reva Academy for Corporate Excellence, Reva University, Bangalore, India
E-mail: [1]lisa.cs02@reva.edu.in,[2]rashmi.agarwal@reva.edu.in, [3]vinay.cs02@reva.edu.in

Abstract - With more edge devices becoming internet-connected, the demand for real-time, no-delay data processing on edge nodes has increased, as it allows systems to process data effectively while consuming less internet bandwidth, and a lot of data can be processed at or near the source. Edge computing is a distributed open architecture that enables systems to process data closer to the information source instead of sending it to the cloud. Edge is defined as the communication that unites the physical and digital worlds. Security risks occur when data is sent to the cloud through IoT devices. Every change the data undergoes might lead to a security lapse. Information transmission across borders is being restricted by tightening privacy restrictions. Edge computing, which stores and analyzes data at the point of origin, is a valuable method for avoiding security problems. A strict governance plan might inhibit innovation, while inadequate or insufficient data governance could expose the company to market disruptions. By minimizing data disruption and analyzing data to improve privacy, decrease data breaches, and promote usability using edge computing for data management. An organization's primary goal is to safeguard the personal information of users, clients, and workers. Data masking has become essential for many firms to protect sensitive information. Companies must have policies for identifying sensitive data, selecting appropriate data masking techniques, and conducting frequent data assurance audits.

Keywords - Edge Computing, IoT, Data integrity, Edge Nodes, PII.

I. INTRODUCTION

There are billions of IoT devices in businesses, oil rigs, hospitals, vehicles, residences, and tens of thousands of other locations. As the number of devices increases, technologies are needed to connect, collect, store, and analyze data. The Internet of Things (IoT) devices are used in conjunction with cloud infrastructures, big data platforms, and machine learning to speed up processing. Over the past decade, the amount of processing and storage space needed in the cloud to handle the high volume, velocity, and integrity of data from IoT sensors has increased substantially, but edge network capacity has not. The user's requirement for openness and information sharing is growing swiftly due to technological innovation. Because of its nature, data sharing cannot meet all of the needs and demands of its consumers. A secure method of transferring data between domains is essential, and a safe way to share data across fields is critical. IoT gadgets and systems provide the best of both worlds as part of edge computing. With the potential for data loss or leakage, user data security and confidentiality become a worry, potentially putting their privacy at risk. In this edge computing environment, many infrastructure providers are in charge of various trust domains, and consumers have no means of knowing which one they should trust[1].

Governments, companies, and organizations are constantly accumulating personal data, either for research or other purposes, and this data is made available to the public. When a table containing personal data is disclosed, sensitive personal information shouldn't be made available to the general public. If a hospital wishes to publish complete patient records for any research or scientific study, it would be simple to mask specific identities from a table such as names of individuals' Aadhaar numbers, Addresses, or Social Security numbers, as shown in Figure 1 to prevent problematic publication features[2]. This study suggests an edge computing paradigm, which considers edge computing nodes and controlled edge servers as one domain; all are connected through the cloud as a safe mechanism to share and transfer data between different environments. The edge computing model securely exchanges data between its many areas and domains [3].

Id	Name	DOJ	Email id	Cell	Salary
1	xxxx	01-01-1980	PXXX@XXX.com	9p0	100000
2	xxxx	09-08-2010	RXXX@XXX.com	9p1	450000
3	xxxx	18-03-2001	KXXX@XXX.com	9j9	897621
4	xxxx	27-09-2003	KXXX@XXX.com	9p1	344587
5	xxxx	13-09-1999	RXXX@XXX.com	9s8	230009
6	xxxx	11-04-1988	JXXX@XXX.com	9p0	930846
7	xxxx	03-11-1977	RXXX@XXX.com	9d6	930846
8	xxxx	01-10-1980	EXXX@XXX.com	9i30	678202
9	xxxx	25-01-2013	RXXX@XXX.com	9p0	313233
10	xxxx	20-08-1992	DXXX@XXX.com	9p15	324355

Figure 1. Displays data masking technique applied to a data set

Every business is responsible for keeping sensitive customer information private. Personal information is valuable to our customers and critical to maintaining our company's reputation as a secure place to do business. By exploring a few bits of PII data, thieves can create false accounts in our names, sell or buy falsified evidence, and sell it to criminals via the dark web. Putting or uploading as little personal

information as possible is the best approach to safeguard information.

As the demands for a better degree of protection, the law regulation has become stricter; such companies try to ensure they protect themselves by all means. The two most common and necessary measures are encryption and data masking. Data-masking (also known as-data anonymization, data de-identification, or data obfuscation) has become mainstream in IT functions of Health care, financial, educational government, and other organizations carrying out business dealing with sensitive personal data. Proposed a model for Privacy Preservation and Data Sanitization on Edge nodes in Edge computing to preserve the data owner's private data and information being uncovered without authorization.
Data masking safeguards a variety of data types; a typical example is:

- IP: Intellectual property
- PII: Personally identifiable information
- e-PHI: Protected health information
- PCI-DSS: Payment card information

The most intricate and safe method of data obscuration is encryption. It employs a cryptographic process to conceal the data, which must be unmasked using the encryption key. This way, the data will be protected and safe if the key is with only authorized individuals who have access to it. If an unauthorized party compromises our system, The rogue user can decrypt the data and view the actual data using a key. Therefore, proper encryption key management is essential

II. RELATED WORK

Some of the recent works on privacy preservation and secure data transfer on edge nodes that are published are reported below:

This article provides a quick overview on the characteristics, security, and uses of edge computing enabled by IoT, as well as its security implications in our data-driven world. Mainly concentrating on developing a scalable, dependable, and distributed edge computing system[4].

This research provides a comprehensive analysis and summary of the issues that the edge computing paradigm presents for maintaining data security and privacy. Besides that, possible security mechanisms and cryptographic-based technologies for solving data security are summarized[5].

Proposed algorithm to attain high reliability for processing and filtering data at the edge nodes. Which could help beat the optimization technique effectively through edge computing [6].

This paper proposes some methods of constructing a secure public-key encryption structure against attacks[2].
This proposal used a light-weight and aggregation-optimized encryption strategy to encrypt the data before off-loading it to the Edge, enabling efficient anomaly detection on encrypted data[7].

There are various models for data publishing and data mining. However, not much is found in the literature that addresses data sanitization to achieve the goal of protecting PII data. It is also discovered in the literature that several authors altered current machine learning approaches to incorporate privacy to publish privacy-preserving findings, such as classification results and data set graphs. These techniques work well for publishing results but not for secure exchange and preservation of data sets.

III. APROACH AND SETUP

A. Methodology
At the network's Edge, the Internet of Things collects enormous volumes of data, but not all are relevant or useful. The majority of this data tends to be "heartbeat" data. If the information does not change significantly, we consider everything to function in order. For example, transferring hours of gathered data to a remote far away to display or verify that the device's vital signs have not changed doesn't make sense. Organizations send all monitoring data to the cloud or enterprise data centers for processing, analysis, and storage. However, as the number of IOTs grows, the volume of data makes this method impractical. In this situation, edge computing becomes useful. Data collected at the edge nodes could fall into three categories:

- It does not require any further processing and storage.
- It may be reserved for record-keeping.
- It requires an immediate response or action.

Edge computing aims to differentiate between distinct data types, determine the amount of reaction necessary, and act or reply accordingly. Most of the time, carrying out these tasks directly at the Edge, where the data is being gathered, is far more effective. Due to its proximity and low latency, edge computing can instantly respond to local events. As a result, data won't be required to be sent back and forth from the Edge to the cloud. Reducing network traffic also significantly saves bandwidth, especially for wireless cellular connections, reducing network costs[8].

In this case, the data is reserved for later analysis and storing securely for further processing. A secure encrypted communication channel has been established between the Server and edge device for

sending and getting requests back. The connected nodes will send data/files which may contain sensitive personal information to the Server.

The script will be able to detect the files if they are mainly in pdf, CSV, or image formats. Once the model detects the particular file, masking or anonymizing PII data will be performed to protect individual users' data identity. The masked data will be encrypted using the AES encryption method, and the AES key will be encrypted using the public key generated by the RSA algorithm. Later, the encrypted key and the masked data can be sent to the destination. The destination node will be able to decrypt the encrypted AES key using the private RSA key. And the same key can be used to decrypt the masked file. This is how the secure data transmission would occur using an encrypted channel[9].

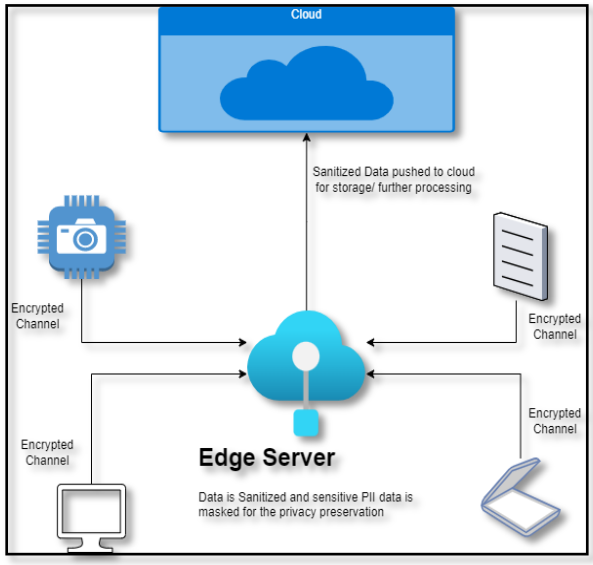


Figure 2. The architecture of the Edge Computing Model

B. Resource Requirement

A minimal software installation is required to run the script on the user's computer with the most recent Operating System and storage needs; the user must already have pre-installed programs like Anaconda, Jupyter Notebook, and Python 3.7. Socket Programming is also used here to send a message over the network. And for communication set up with the Edge(client) and Server mode.

IV. IMPLEMENTATION

The implementation of software design has been broken down into the following steps:

A. Setting up the Environment

For the communication process, using an interface built with the ktinter library using Python interface, where the data can be uploaded and received securely. The nodes must be connected using the same communication ID to exchange the data.

B. Identification of File type

Based on the extension provided, the script will be able to recognize the file type and type of document to handle and will be able to perform analysis further accordingly. A Secure communication is set up, and once the file is sent from the edge node, it confirms the file transfer to the server location.

C. Encrypting the Data

Once the information is classified, the next step is masking data based on the file type. An appropriate data masking technique will be applied to sensitive PII data of users. Then the masked data will be encrypted using the AES algorithm. For encryption – here, a library called crypto is used and using the public key generated by the RSA algorithm, encrypting the AES key. By invoking the RSA library into the script, can generate Public and private key (.pem files) between node and Server through the same communication setup. Once the file reaches the destination node, it will be decrypted using the available private key. Using the method, encrypted data will be shared with the destination node only. Using the public AES key, the encrypted data can decrypted. If the intruder tries to access the data or keys, they will be unsuccessful because the private key is available with the destination node only, and the communication is protected. Once the user enters the credentials, it will redirect to the communication window where data exchange can occur. The below Figure.3 depicts the architecture of the working model in detail and how the data is encrypted and transferred from the Edge to the destination location.

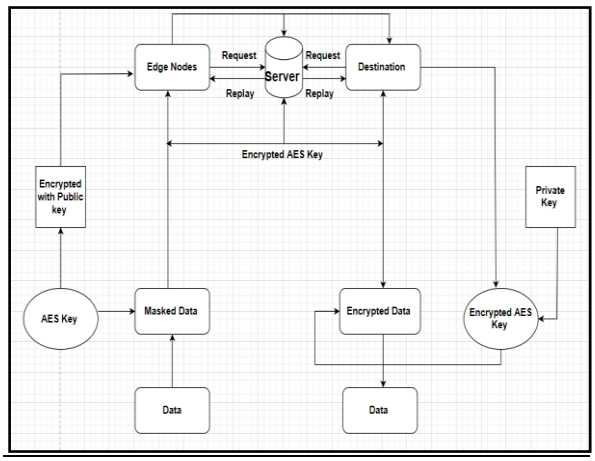


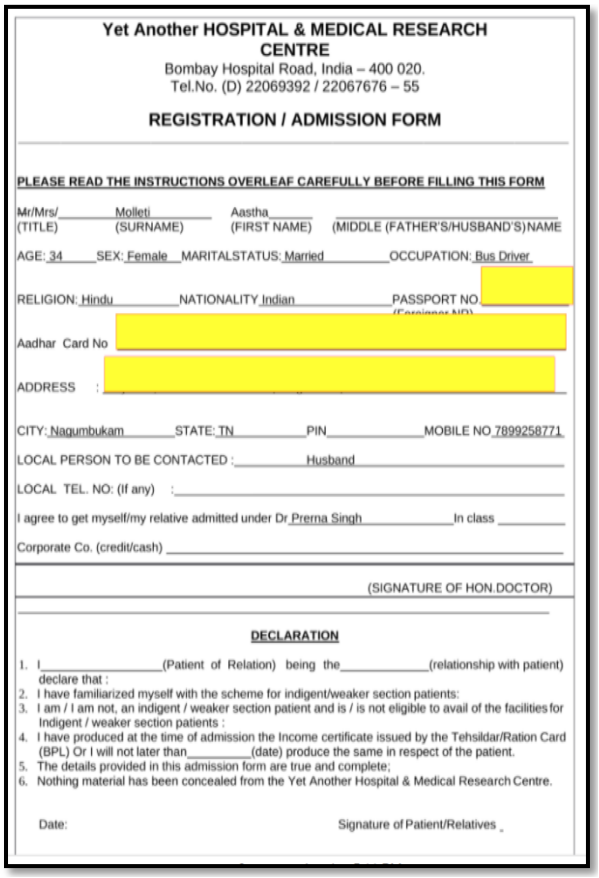
Figure 3. The Architecture of the Data Flow Model Implemented

D. Results

To safeguard data from unauthorized exposure, anonymization is frequently required in test and development databases. Data sharing isn't simply a technological issue. Data transfer across regional and international boundaries might exacerbate worries about data security, privacy, and other legal matters. Edge computing is employed to keep data close to its point of origin and within the parameters of the

GDPR, which governs how data should be stored, processed, and made available in the European Union. This allows raw data to be processed close by, obscuring or securing sensitive data before sending anything to the cloud or primary data center, which can be in other jurisdictions or locations.

Users provide a lot of PII (Personal Identifiable Information), which will be required during the verification process in many places. But not necessarily; this data needs to be stored for further use. The sensitive PII data fields must be masked by either changing the values or using image masking, making them invisible. The goal/objective is to create a version that is impossible to be deciphered or reverse engineered. All this data will be masked before uploading to the cloud for later storage or further processing, as depicted in Figure 4. Later gets saved in Cloud location selected by the individual for storage. During the verification process, random files have been manually verified to confirm the presence of any sensitive personal data and confirmed that the algorithm applied has worked and can mask the sensitive fields in the file[10].



Yet Another HOSPITAL & MEDICAL RESEARCH CENTRE
Bombay Hospital Road, India – 400 020.
Tel.No. (D) 22069392 / 22067676 – 55

REGISTRATION / ADMISSION FORM

PLEASE READ THE INSTRUCTIONS OVERLEAF CAREFULLY BEFORE FILLING THIS FORM

Mr/Mrs/ Molleti Aastha
(TITLE) (SURNAME) (FIRST NAME) (MIDDLE (FATHER'S/HUSBAND'S) NAME

AGE: 34 SEX: Female MARITALSTATUS: Married OCCUPATION: Bus Driver

RELIGION: Hindu NATIONALITY Indian PASSPORT NO.

Aadhar Card No

ADDRESS :

CITY: Nagumbukam STATE: TN PIN MOBILE NO 7899258771

LOCAL PERSON TO BE CONTACTED : Husband

LOCAL TEL. NO: (If any) :

I agree to get myself/my relative admitted under Dr Prerna Singh In class

Corporate Co. (credit/cash)

(SIGNATURE OF HON.DOCTOR)

DECLARATION

1. I (Patient of Relation) being the (relationship with patient) declare that :
2. I have familiarized myself with the scheme for indigent/weaker section patients:
3. I am / I am not, an indigent / weaker section patient and is / is not eligible to avail of the facilities for Indigent / weaker section patients :
4. I have produced at the time of admission the Income certificate issued by the Tehsildar/Ration Card (BPL) Or I will not later than (date) produce the same in respect of the patient.
5. The details provided in this admission form are true and complete;
6. Nothing material has been concealed from the Yet Another Hospital & Medical Research Centre.

Date: Signature of Patient/Relatives

Figure 4. Displays Uploaded File After Applying the Data Masking on PII Data

V. CONCLUSION

Information sharing is commonplace nowadays; thus, sensitive data like employee or customer information that people or authorities will share for commercial purposes or with an organization has to be protected. Malicious intent people may readily access the servers where data is hosted and exploit the sensitive data. This could hurt the organization or business, as there is no protection on the shared data, giving an unfavorable picture of the organization. Therefore, protecting the information is essential and could be done using information masking techniques on the data. This research presents a methodology for learning a linear regression model over data from remote

REFERENCES

[1] J. Lee, H. J. Ko, E. Lee, W. Choi, and U. M. Kim, "A data sanitization method for privacy preserving data re-publication," Proc. - 4th Int. Conf. Networked Comput. Adv. Inf. Manag. NCM 2008, vol. 2, pp. 28–31, 2008, doi: 10.1109/NCM.2008.203.

[2] P. Liu, "Public-key encryption secure against related randomness attacks for improved end-to-end security of cloud/Edge computing," IEEE Access, vol. 8, pp. 16750–16759, 2020, doi: 10.1109/ACCESS.2020.2967457.

[3] G. Qiu, X. Gui, and Y. Zhao, "Privacy-Preserving Linear Regression on Distributed Data by Homomorphic Encryption and Data Masking," IEEE Access, vol. PP, p. 1, Jun. 2020, doi: 10.1109/ACCESS.2020.3000764.

[4] M. Alrowaily and Z. Lu, "Secure edge computing in IoT systems: Review and case studies," Proc. - 2018 3rd ACM/IEEE Symp. Edge Comput. SEC 2018, pp. 440–444, 2018, doi: 10.1109/SEC.2018.00060.

[5] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data Security and Privacy-Preserving in Edge Computing Paradigm: Survey and Open Issues," IEEE Access, vol. 6, no. Idc, pp. 18209–18237, 2018, doi: 10.1109/ACCESS.2018.2820162.

[6] V. D. A. Kumar, A. Kumar, R. S. Batth, M. Rashid, S. K. Gupta, and M. Raghuraman, "Efficient data transfer in edge envisioned environment using artificial intelligence based edge node algorithm," Trans. Emerg. Telecommun. Technol., vol. 32, no. 6, 2021, doi: 10.1002/ett.4110.

[7] S. Mehnaz and E. Bertino, "Privacy-preserving real-time anomaly detection using edge computing," Proc. - Int. Conf. Data Eng., vol. 2020-April, pp. 469–480, 2020, doi: 10.1109/ICDE48307.2020.00047.

[8] F. Y. Rao and E. Bertino, "Privacy Techniques for Edge Computing Systems," Proc. IEEE, vol. 107, no. 8, 2019, doi: 10.1109/JPROC.2019.2918749.

[9] W. Wu, Q. Zhang, and H. J. Wang, "Edge computing security protection from the perspective of classified protection of cybersecurity," Proc. - 2019 6th Int. Conf. Inf. Sci. Control Eng. ICISCE 2019, pp. 278–281, 2019, doi: 10.1109/ICISCE48695.2019.00062.

[10] A.N.K.Zaman, C. Obimbo, and R. A. Dara, "An improved data sanitization algorithm for privacy preserving medical data publishing," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10233 LNAI, no. April, pp. 64–70, 2017, doi: 10.1007/978-3-319-57351-9_8.

★★★