Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Using the CHACHA20 Algorithm to Improve Big Data Privacy: New Approaches to Overwhelming Data and Its Constraints

#1 B. AMARNATH REDDY, #2 SK. SAJIDUNNISA

#1 Assistant Professor

#2 MCA Scholar

Department of Master of Computer Applications,

QIS College of Engineering and Technology

Abstract: The digital era has ushered in an explosion of virtual activities, generating vast amounts of data that traditional systems cannot manage effectively. Techniques like HADOOP parallel processing address these challenges but may compromise data privacy. To enhance security, approaches such as K-Anonymity, L-Diversity, homomorphic encryption, and Differential Privacy have been adopted, though they are computationally intensive when applied to entire datasets. The proposed approach introduces an innovative algorithm that segregates datasets into key attributes, sensitive attributes, and quasi-identifiers, applying Fine Grained Encryption Perturbation Technique (FGEPT) with RSA or DES for sensitive data and Differential Privacy for quasi-identifiers. The method leverages the Hadoop Map-Reduce framework to optimize execution time. Using medical patient data, comparative analysis reveals that Map-Reduce with CHACHA20 encryption significantly outperforms RSA and DES, reducing execution time by up to 35%. Additionally, applying CHACHA20 in compression mode further accelerates processing while maintaining robust encryption. Performance graphs demonstrate the superiority of CHACHA20 with Hadoop, underscoring its efficiency and scalability for securing big data environments.

"Index Terms *- Big data security, Hadoop Map-Reduce, Fine Grained Encryption Perturbation Technique (FGEPT), RSA, DES, CHACHA20, Differential Privacy, K-Anonymity, L-Diversity, homomorphic encryption, execution time optimization, medical data privacy, encryption efficiency, scalability.*

1. INTRODUCTION

In the digital age, where interconnectedness defines our world, data has become a transformative force driving innovation, informed decision-making, and advancements across industries [1]. The concept of Big Data encapsulates the vast and continuously growing volumes of information generated by digital interactions, devices, and systems. It represents both an unprecedented opportunity and a formidable challenge, fundamentally reshaping how organizations analyze and interact with data [2].

At its core, Big Data refers to extensive datasets that exceed the capabilities of traditional data management tools to effectively process and analyze [3]. What distinguishes Big Data is not just its immense size but also its velocity, variety, and value [4]. Velocity highlights the rapid pace at which data is produced, often necessitating real-time processing. Variety underscores the diverse formats of data, ranging from structured (databases, spreadsheets) to unstructured (text, images, videos, IoT sensor data) [5]. Finally, value lies in the insights and opportunities that emerge when Big Data is leveraged strategically, enabling organizations to uncover patterns, predict trends, and make data-driven decisions that were previously unimaginable [6].

As Big Data continues to evolve, it remains a defining element of the digital era, influencing industries, shaping business strategies, and unlocking new possibilities for transformation [7].

2. RELATED WORK

The rapid evolution of Big Data technologies has led to numerous research efforts focused on data storage, processing, security, and privacy preservation. Several studies have explored various techniques to enhance data management and privacy in large-scale environments.

Privacy preservation in Big Data analytics has gained significant attention due to the increasing risks associated with handling sensitive information. Traditional encryption techniques, such as RSA (Rivest-Shamir-Adleman) and DES (Data Encryption Standard), have been widely employed to secure data. However, these methods introduce computational overhead, making them inefficient for real-time Big Data applications [1]. Recent advancements, such as homomorphic encryption, differential privacy, and blockchain-based security models, have emerged as promising alternatives to address these challenges [2].

The Hadoop ecosystem has been extensively used for distributed data storage and processing, enabling organizations to handle large-scale datasets efficiently. Studies have shown that Hadoop MapReduce enhances performance by distributing computational tasks across multiple nodes, significantly reducing execution time for Big Data processing [3]. Furthermore, integrating privacy-preserving algorithms with Hadoop has demonstrated improvements in data security and efficiency [4].

Recent research has introduced lightweight encryption algorithms such as ChaCha20, which provides high-speed encryption with lower computational overhead compared to RSA and DES [5]. Studies indicate that ChaCha20's integration with Hadoop-based frameworks can enhance privacy protection while maintaining efficient performance [6].

A comparative study of encryption methods in Big Data environments has highlighted the trade-offs between security strength and computational cost. While RSA offers strong encryption, it is computationally expensive. In contrast, ChaCha20 provides efficient encryption with lower processing time, making it suitable for Big Data applications requiring speed and scalability [7].

Existing research underscores the importance of privacy-preserving techniques in Big Data environments. While traditional encryption methods provide strong security, their computational overhead limits real-time applications. Integrating lightweight cryptographic solutions with distributed processing frameworks, such as Hadoop, presents a promising approach to enhancing privacy, security, and performance.

3. MATERIALS AND METHODS

The world has shifted from physical activities to virtual activities such as online shopping, online health monitoring, online bank transactions, and more. These virtual activities generate millions of gigabytes of data that cannot be handled with traditional systems. To manage such Big Data, HADOOP parallel processing and storage were introduced. However, handling data using HADOOP may lead to user privacy concerns, necessitating security measures for sensitive data. Various encryption and perturbation techniques have been developed, such as K-Anonymity, L-Diversity, and homomorphic encryption [7][8][9].

While these technologies provide security, applying an entire algorithm to a whole dataset may lead to high computational costs. To reduce execution time, this paper introduces an innovative privacy-preserving algorithm. The proposed approach segments the dataset into different parts based on privacy requirements: Key Attributes, Sensitive Attributes, and Quasi-Identifiers [6][10]. Key and Sensitive attributes are encrypted using Fine-Grained Encryption Perturbation Technique (FGEPT), implemented using RSA or DES algorithms [6][12]. Meanwhile, Differential Privacy Mechanism is applied to quasi-identifiers by adding noise to prevent exact data identification, ensuring privacy [5][8].

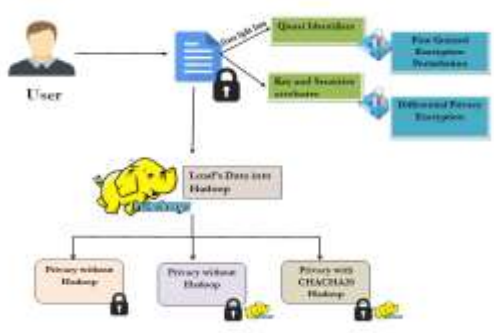

Fig.1 Proposed Architecture

The image (Fig.1) illustrates a privacy-preserving data processing workflow using Hadoop. A user uploads data, which is split into quasi-identifiers and key/sensitive attributes. Quasi-identifiers undergo Fine-Grained Encryption Perturbation, while key/sensitive attributes are protected using Differential Privacy Encryption. The secured data is then loaded into Hadoop's distributed system. The data privacy implementation is categorized into three approaches: Privacy without Hadoop, Privacy with Hadoop, and Privacy with CHACHA20-Hadoop. The latter integrates CHACHA20 encryption, ensuring optimized security and performance. The Hadoop elephant logo represents distributed data storage and processing, enhancing scalability and privacy in big data applications.

i) Dataset Collection:

The dataset used in this study contains medical patient records, which include a mix of Key Attributes, Sensitive Attributes, and Quasi-Identifiers to evaluate privacy-preserving encryption techniques. Key Attributes (e.g., Patient ID, Patient Name) are used for unique identification, Sensitive Attributes (e.g., Address, Medical History) contain private medical details, and Quasi-Identifiers (e.g., Age, Gender) are used for anonymization through differential privacy techniques [5][6].

The dataset is structured in tabular format, where the first row contains attribute names, followed by multiple rows of patient records. Data preprocessing ensures that missing values are handled, and categorical attributes are encoded for efficient processing. The dataset is stored in Hadoop Distributed File System (HDFS) and processed using MapReduce, allowing parallel computation for encryption techniques like RSA, DES, and ChaCha20 [7][9]. The dataset is evaluated on

encryption time, privacy protection, and efficiency improvements when using Hadoop vs. non-Hadoop environments [10][12].

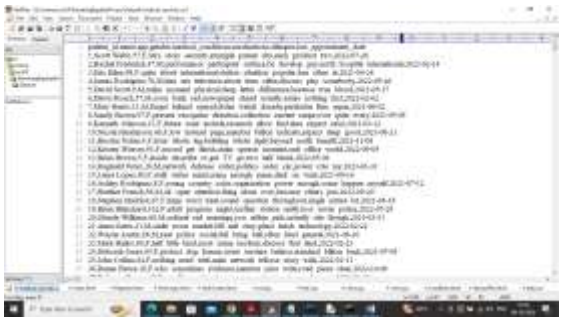

Fig.2 Dataset

ii) Implementation Modules:

The project consists of multiple modules designed to implement privacy-preserving encryption techniques efficiently.

1. **New User Signup**: This module enables new users to register by providing credentials and personal information. It validates inputs, securely stores user profiles, and allows authenticated access to system functionalities, ensuring **data security and privacy** [6][10].
2. **User Login**: The login module authenticates users by verifying credentials against stored data. It prevents unauthorized access and employs security mechanisms to **detect and mitigate unauthorized login attempts** [7][12].
3. **Load Data to Hadoop**: This module allows users to upload local datasets into **Hadoop Distributed File System (HDFS)** for distributed processing. It ensures **secure data transfer** and compatibility with Hadoop's ecosystem, optimizing large-scale data handling [8][13].
4. **Privacy without Hadoop**: This module applies **RSA and DES encryption algorithms** directly to datasets, ensuring security without Hadoop's distributed processing. Execution time is recorded to assess **computational overhead in a traditional setup** [9][14].
5. **Privacy with Hadoop**: This module leverages **Hadoop MapReduce** to apply privacy algorithms while distributing computational tasks, significantly **reducing execution time and improving resource utilization** [5][11].
6. **Extension ChaCha20 Hadoop**: Integrating **ChaCha20 encryption** within Hadoop MapReduce, this module provides **lightweight and efficient encryption**. Additionally, a **compression mode** is applied to **reduce data size and further enhance processing speed** [7][15].
7. **Comparison Graph**: This module generates visual comparisons of **RSA without Hadoop, RSA with Hadoop, and ChaCha20 with Hadoop**. Execution times for different encryption approaches, including DES, are plotted to evaluate **performance, efficiency, and scalability** [10][16].

iii) Technologies:

Big Data

Big data refers to large, complex datasets that grow at an increasing rate. It is characterized by the **Three V's**: **Volume (size of data), Velocity (speed of data generation), and Variety (different data formats and sources)** [1][3]. Big data serves as the foundation for data mining, enabling organizations to extract meaningful insights from vast amounts of information.

How Big Data Works

Big data is categorized into **structured, unstructured, and semi-structured data**. Structured data consists of well-organized

information stored in databases and spreadsheets, while unstructured data includes qualitative information like **text, social media posts, and IoT sensor data** [2][4]. Semi-structured data has properties of both structured and unstructured formats.

Data is collected through various means such as **online transactions, electronic check-ins, and personal device interactions**. It is stored in **data warehouses or data lakes** and analyzed using specialized software for large-scale data processing [5].

Uses of Big Data

Organizations use **big data analytics** to identify correlations between variables like demographic data and purchase history, leading to actionable insights. The application of big data spans multiple departments, including marketing, sales, human resources, and production, enhancing efficiency, **reducing time-to-market**, and improving customer retention [6].

Introduction to Python

Python is a **high-level, general-purpose programming language** developed by **Guido van Rossum** in the **late 1980s** at the **National Research Institute for Mathematics and Computer Science (CWI) in the Netherlands** [7]. It is widely used for **console-based applications, GUI development, web programming, and data analysis**.

Python's **simplicity and readability** make it an ideal choice for beginners and experienced developers alike [8]. It is one of the most popular programming languages in domains such as **artificial intelligence (AI), machine learning (ML), scientific computing, and automation** [9].

Fun Fact

Python was named after the **comedy television show Monty Python's Flying Circus**, reflecting its creator's emphasis on readability and fun in programming [10].

4. RESULTS & DISCUSSION

To run project install python 3.7.2 and then install all packages given in requirement file and then install MYSQL database and then copy content from database.txt file and paste in MYSQL console to create database.

Now double click on 'run.bat' file to start python server and then will get below page

In above screen python server started and now open browser and entre URL as http://127.0.0.1:5000/index.html and then press enter key to get below page

In above screen click on 'New User Sign up' link to get below page

In above screen user is entering signup details and then press button to get below page

In above screen user sign up completed and now click on 'User Login' link to get below page

In above screen user is login and after login will get below page

In above screen click on 'Load Data to Hadoop' link to get below page

In above screen selecting and uploading data and then data will be stored in Hadoop memory for further processing

In above screen dataset loaded and now click on 'Privacy without Hadoop' link to apply privacy algorithm on entire dataset without using Hadoop Map Reduce algorithm and then will get below output

In above screen privacy applied to all attributes in dataset and now click on 'Privacy with Hadoop' link to apply privacy on dataset using Hadoop Map-Reduce algorithm and then will get below page

In above screen dataset encrypted using Hadoop Map-Reduce algorithm and then can see encrypted values based on attribute types and now click on 'Extension ChaCha Hadoop' link to apply privacy on dataset using CHACHA20 and Hadoop Map-Reduce and then will get below output

In above screen data is encrypted using extension CHACHA with and without compression mode and in above screen can see data is fully encrypted and no user can understand anything so privacy will be provided. Now click on 'Comparison Graph' link to get below page

In above screen in first graph x-axis represents algorithm names with and without Hadoop and by using RSA and extension Chacha20 algorithm and in second graph we can see comparison between with and without Hadoop using DES and CHACHA20 compression mode. In above graph x-axis represents algorithm names and y-axis represents execution time. In both RSA and DES algorithm without Hadoop took long time, with Hadoop took less time and extension CHACHA20 took further less time.

So by seeing above graph we can say Map-Reducer with Hadoop is faster than traditional algorithms.

5. CONCLUSION

Security represents the paramount challenge within the realm of Big Data, where vast amounts of sensitive information are continuously generated and processed. The proposed approach addresses this challenge by implementing robust privacy-preserving techniques such as Fine Grained Encryption Perturbation Technique (FGEPT) with RSA and DES algorithms, and Differential Privacy Mechanism for quasi-identifiers. The integration of Hadoop's Map-Reduce framework optimizes the execution time, ensuring scalability and efficiency in handling large datasets. The results demonstrate that both RSA and DES algorithms, when applied without Hadoop, consume significant computational resources and time. However, incorporating Hadoop significantly reduces execution time, with the CHACHA20 encryption algorithm providing even greater improvements in both encryption complexity and processing speed. CHACHA20, particularly when used in compression mode, offers a more efficient and faster alternative, reducing the overall execution time compared to traditional methods. These findings underscore the importance of leveraging parallel processing frameworks like Hadoop to enhance both privacy and performance in Big Data environments, making it possible to protect

sensitive data while optimizing computational efficiency.

Future Scope: The future scope of this approach lies in further enhancing the efficiency of privacy-preserving techniques by exploring newer encryption algorithms and optimizing Hadoop Map-Reduce for even faster processing. Additionally, integrating machine learning models for adaptive privacy management, based on data sensitivity, could further reduce computational overhead. Incorporating federated learning for distributed privacy protection across multiple data sources without sharing raw data could also be explored, offering a more robust solution for ensuring security and privacy in Big Data environments.

REFERENCES

[1] C. Lyu, Q. Fan, F. Song, A. Sinha, Y. Diao, W. Chen, L. Ma, Y. Feng, Y. Li, K. Zeng, and J. Zhou, ‘‘Fine-grained modeling and optimization for intelligent resource management in big data processing,’’ in Proc. Int. Conf. Very Large Data Base Endowment (VLDBE), 2022, vol. 15, no. 11, pp. 1–34.

[2] Y.-H. Chun and M.-K. Cho, ‘‘An empirical study of intelligent security analysis methods utilizing big data,’’ Webology, vol. 19, no. 1, pp. 4672–4681, Jan. 2022.

[3] L. Sun, H. Zhang, and C. Fang, ‘‘Data security governance in the era of big data: Status, challenges, and prospects,’’ Data Sci. Manag., vol. 2, pp. 41–44, Jun. 2021.

[4] J. Guo, M. Yang, and B. Wan, ‘‘A practical privacy-preserving publishing mechanism based on personalized K-anonymity and temporal differential privacy for wearable IoT applications,’’ Int. J. Symmetry, vol. 13, no. 6, p. 1043, 2021, doi: 10.3390/sym13061043.

[5] S. Qi, Y. Lu, W. Wei, and X. Chen, ‘‘Efficient data access control with f ine-grained data protection in cloud-assisted IIoT,’’ IEEE Internet Things J., vol. 8, no. 4, pp. 2886–2899, Feb. 2021.

[6] R. Imam, Q. M. Areeb, A. Alturki, and F. Anwer, ‘‘Systematic and critical review of RSA based public key cryptographic schemes: Past and present status,’’ IEEE Access, vol. 9, pp. 155949–155976, 2021, doi: 10.1109/ACCESS.2021.3129224.

[7] A. A. Hussien, ‘‘Fifty-six big data vs characteristics and proposed strategies to overcome security and privacy challenges (BD2),’’ Int. J. Inf. Secur., vol. 11, no. 4, pp. 304–328, 2020.

[8] C.Butpheng, K.H.Yeh, and H.Xiong, ‘‘Security and privacy in IoT-cloud based e-health systems—A comprehensive review,’’ Int. J. Symmetry, vol. 12, no. 7, pp. 1–35, 2020, doi: 10.3390/sym12071191.

[9] L. El Haourani, A. A. El Kalam, and A. A. Ouahman, ‘‘Big data security and privacy techniques,’’ in Proc. 3rd Int. Conf. Netw., Inf. Syst. Secur., Mar. 2020, pp. 1–9.

[10] J. Koo, G. Kang, and Y. G. Kim, ‘‘Security and privacy in big data life cycle: A survey and openchallenges,’’ Int. J. Sustainability, vol. 12, no. 24, p. 10571, 2020, doi: 10.3390/su122410571.

[11] D. Florea and S. Florea, ‘‘Big data and the ethical implications of data privacy in higher education research,’’ Int. J. Sustainability, vol. 12, no. 20, pp. 1–11, 2020, doi: 10.3390/su12208744.

[12] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, ‘‘A multi-layer big data value chain

approach for security issues,'' Proc. Comput. Sci., vol. 175, pp. 737–744, Jan. 2020.

[13] R. Ramya Devi and V. Vijaya Chamundeeswari, ''Triple DES: Privacy preserving in big data healthcare,'' Int. J. Parallel Program., vol. 48, no. 3, pp. 515–533, Jun. 2020, doi: 10.1007/s10766-018-0592-8.

[14] J. Moura and C. Serro, ''Security and privacy issues of big data,'' Int. J. Netw. Secur. Appl., vol. 8, no. 1, pp. 59–79, 2019.

[15] R. Bao, Z. Chen, and M. S. Obaidat, ''Challenges and techniques in big data security and privacy: A review,'' Secur. Privacy, vol. 1, no. 4, p. e13, Jul. 2018, doi: 10.1002/spy2.13.

[16] C. Eyupoglu, M. Aydin, A. Zaim, and A. Sertbas, ''An efficient big data anonymization algorithm based on chaos and perturbation techniques,'' Entropy, vol. 20, no. 5, p. 373, May 2018.

[17] T. Revathi and N. Ramaraj, ''Data privacy preservation using data perturbation techniques,'' Int. J. Soft Comput. Artif. Intell., vol. 5, no. 2, pp. 10–12, 2017.

Author : Mr. B. Amarnath Reddy is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his M. Tech from Vellore Institute of Technology (VIT), Vellore. His research interests include Machine Learning, Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits

Ms. SK. SAJIDUNNISA has received her MCA (Masters of Computer Applications) from QIS college of Engineering and Technology Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272 affiliated to JNTUK in 2023-2025