Secure Big Data Processing in the Cloud: An Access Pattern Matching-Based Security Assurance Framework

HRIDAY KUMAR GUPTA [1,3] RAFAT PARVEEN[2]

[1,2]Department of Computer Science
Jamia Millia Islamia
New Delhi, India
[3]KIET Group of Institutions, Ghaziabad
Corresponding Author Email: hridaykumargupta@gmail.com

Abstract— Access Pattern Matching (APM) secures and authenticates Big Data (BD) access requests. It compares the Data User (DU) access pattern with the Data Owner (DO) access structure defined during data upload. The proposed approach generates an access structure from BD reduced blocks. This structure defines data access patterns. The Cloud Server (CS) compares the access pattern in a DU's access request to the DO's access structure. The access pattern generally includes the DU's desired data blocks or subsets of BD, their order, and any DO-defined criteria or constraints. The access structure checks this pattern to see if the DU can access the data. System design determines access pattern matching algorithms. These algorithms compare the access pattern and structure to the DO's approved patterns and circumstances. The DU gets data if the access pattern meets the access structure. Access pattern matching restricts BD content to authorized users with valid access patterns. It prevents unauthorized access and enforces DO access control regulations to safeguard data.

Keywords: Access Pattern Matching., Big data security, Privacy protection, timestamp, XOR Public and Private Key , Double Elliptic Curve Cryptography .

1. INTRODUCTION

Big data security, Privacy protection, timestamp, XOR Public and Private Key based Double Elliptic Curve Cryptography (XPPDECC), access pattern [1]. The objective of access pattern matching in the cloud for big data is to maximize the efficacy of data storage and processing by analyzing and recognizing patterns in data access behavior. Understanding access patterns become critical for optimizing resource utilization and boosting overall performance in the setting of big data when huge volumes of data are stored and processed in distant cloud environments. Data access patterns include the frequency, volume, and data items that are accessed by applications or users [2]. Cloud systems may make educated choices regarding data placement, caching, replication, and resource allocation by analyzing these trends. To minimize latency and increase availability, data may be duplicated over numerous cloud servers or placed in a high-performance storage tier. On the other hand, seldom-used data may be preserved or stored in a less expensive location [3]. Techniques for matching access patterns use algorithms and statistical analysis to learn about and anticipate access behavior. Access patterns may be classified as sequential, random, skewed, or concentrated in a few areas (hotspots). Organizations can optimize their infrastructure, save costs, and guarantee that the most frequently used data is always accessible for quicker processing and a better user experience by employing access pattern matching in the cloud for big data [4]. To protect user privacy using access pattern matching, encryption techniques, differential privacy, and anonymization techniques can be implemented. Encryption algorithms and secure key management practices ensure integrity, while differential privacy enhances user anonymity by introducing controlled randomness. Anonymization techniques like generalization, suppression, and perturbation effectively anonymize data. Organizations should adhere to best practices in access control, data governance, and security protocols, and conduct periodic assessments and audits to identify and mitigate potential vulnerabilities.

2. LITERATURE REVIEW

Cloud computing, as defined by researchers [6–9] is "a paradigm for delivering IT services to end users wherein they pay only for the resources they really use." In this model, consumers have access to resources like computer power, data storage, and application software through the Internet without putting in any money or putting in any effort to set up the necessary physical infrastructure themselves. By enabling users to consume resources on-demand and pay only for what they need, the cloud allows flexibility, scalability, and cost-effectiveness. The ability to delegate infrastructure management and upkeep to third-party cloud providers has contributed to this model's meteoric rise in popularity. [10] described Secured MapReduce (SMR) is a privacy layer developed for the purpose of ensuring the Hadoop Distributed File System (HDFS) and the MapReduce layer. The SMR model required data collection, followed by HDFS uploading. The HDFS data was encrypted after being processed in the MapReduce layer. The HybrEx concept used a vertical partitioning method, with encryption taking place in the private cloud and decryption taking place in the public cloud [5]. The experimental results proved the viability of the approach in improving privacy and security in Big Data settings.

This paper [11],authors explore how the AES encryption technique may be used to make data transfers in the cloud more secure, especially when dealing with huge datasets. This approach gives consumers more leeway in picking cloud service providers that meet their unique needs. Despite its efficacy as an encryption method, AES does not deal with additional security issues including insider attacks, collusion, or data integrity breaches. When it comes to protecting data in the cloud, it's important to take into account and handle all of these extra security concerns.

G. Viswanath and P. Venkata Krishna devised a secure framework that prevented insider attacks [12] The authors of the paper presented an encryption technique designed specifically for storing enormous amounts of data across multiple cloud storage locations. The framework included numerous operations, such as data uploading, segmentation, indexing, encryption, distribution, retrieval, and combination. Before storing vast amounts of data in multiple cloud environments, the hybrid encryption method was devised to offer a higher level of data security. The authors conducted a simulation study with real-time cloud storage settings and discovered that the algorithm outperformed existing algorithms in terms of performance. The suggested approach is especially well-suited for huge data sets, since we have intended it to encrypt in terms of multiple-character blocks. Moreover, the procedure is applicable to any script with established ASCII values; it is not limited to the English alphabet. Big Data Technology is very much adequate for the accurate analysis of our big data which yields strong conclusions and prediction [13]

3. MOTIVATION

The purpose of doing an analysis of access patterns in big data cloud security is to strengthen the security and maintain the integrity of data stored in environments that are hosted in the cloud. Organizations are able to identify possible security issues, discover abnormalities, and put in place suitable security measures if they understand and monitor access patterns. Access pattern analysis is a useful tool for identifying unauthorised access, insider threats, and aberrant behaviour involving the accessing of data. It makes it possible to identify trends of data consumption and contributes to the process of assuring compliance with privacy legislation. The objective is to boost the entire security posture in big data cloud settings by proactively identifying and responding to security events, enhancing access controls, and protecting sensitive data from breaches or unauthorized disclosure. This will be accomplished through identifying and responding to security incidents.

4. CHALLENGES

The complexity and scalability of data access, dynamic and evolving access patterns, privacy considerations, data governance, and compliance requirements, and the detection of insider threats are the challenges of access patterns in big data cloud security. Managing and analyzing access patterns across various data sources, applications, and users can be complex, necessitating customized security controls and strategies. The scalability of big data systems presents challenges for effectively managing concurrent access requests with a high throughput. Concerns regarding privacy must be addressed when analysing access patterns while ensuring regulatory compliance. In addition, monitoring access patterns alone may not be enough to detect internal threats, necessitating the use of advanced mitigation techniques. To address these challenges, a comprehensive strategy is required, including scalable access control mechanisms, advanced analytics, techniques that protect privacy, and continuous security updates. the encryption algorithm's significance in ensuring cloud security cannot be emphasized enough. Our research focused on examining the effectiveness of well-known encryption methods like AES, DES, triple DES, and Blowfish algorithms. Through a series of experiments involving files of varying sizes and diverse data types, we evaluated their performance and effectiveness in safeguarding sensitive information within the cloud environment. [14]

5. SOLUTION AND PROPOSED METHODOLOGY

The best way to deal with problems with access patterns in big data cloud security is to take a wide range of steps. This includes setting up access control systems that can handle large amounts of data and multiple calls for entry at the same time. Based on access trends, advanced analytics and machine learning can be used to find and deal with anomalies and insider risks. Techniques that protect privacy, like data anonymization and encryption, help keep private information safe while still letting it be analysed well. Data control and legal standards are met when entry trends are checked and monitored regularly. Also, security measures are always being updated and new technologies are being used, which helps keep entry patterns secure in big data cloud settings.

```
function authenticateUser(userCredentials):
if userCredentials are valid: then
    return true
else
    return false
end
function authorizeUser(userRole, requestedResource):
```

```
Authorize the user based on role and requested resource
if userRole has access to requestedResource: then return
true
else return false
end function
accessPatternMatching(accessRequest):
    Main function to secure and authenticate BD access requests using APM if
    authenticateUser(accessRequest.userCredentials) is true: then
    authorizeUser(accessRequest.userRole, accessRequest.requestedResource) is true:
    Access request is authenticated and authorized logAccessRequest(accessRequest)
    grantAccess(accessRequest.requestedResource)
    return ”Access granted”
    else if User is not authorized to access the requested resource then
    logAccessRequest(access request) return ”Access denied: User is not authorized” else
    User authentication failed logAccessRequest(accessRequest)
    return ”Access denied: User authentication failed”
    end function
    logAccessRequest(accessRequest): // Log
access request details for auditing purposes
log(accessRequest) function
grantAccess(resource): // Grant access to the
requested resource log(grantAccess)

accessRequest = { userCredentials: { username:
        "example_user", password:
        "example_password"
        },
        userRole: "admin", requestedResource:
        "big_data_set_1"
}

result = accessPatternMatching(accessRequest) print(result)
```

5.1. Attack level analysis

The purpose of doing an attack level analysis in a big data environment that is hosted in the cloud is to get an understanding of the many levels and kinds of assaults that may take place in such a setting and the possible dangers that they could pose. There are a number of different attack vectors that might put a cloud environment for big data at risk of compromising the data's availability, integrity, and security. Infrastructure assaults, also known as distributed denial of service attacks or server breaches, are directed against the cloud's underlying infrastructure, which consists of servers, networks, and storage systems [15]. Data breaches occur when unauthorised users get access to sensitive data that has been stored in the cloud and exploit holes in access controls or use authentication methods that are inadequate. There is a substantial danger posed by insider threats, which occur when trusted employees abuse the credentials they have been granted in order to steal information or undermine the integrity of data. Attacks that include data modification or tampering try to change the integrity or dependability of large amounts of data, which might possibly lead to inaccurate analytical findings or the introduction of harmful material. Attacks on service availability have the effect of disrupting cloud services, which in turn affects the accessibility and operation of big data platforms. Malware and advanced persistent threats may also undermine the data integrity of a system, steal information from it, or disrupt its operations. In order to protect against these assaults, you will need to implement a multi-layered security strategy. This strategy should include stringent access restrictions, encryption, intrusion detection systems, continuous monitoring, security audits, vulnerability assessments, and staff training. When it comes to protecting the security and integrity of massive data stored in cloud settings, preventative steps and an all-encompassing defense plan are absolutely necessary.

This research is motivated to address the attack level by the increasing complexity and variety of cyber threats and attacks designed to compromise the security of big data in cloud computing environments. These attacks can take various forms, including unauthorised access, data interception, data manipulation, denial of service, and malware implantation, posing significant threats to the privacy and integrity of big data. The objective of the research is to develop countermeasures, detection mechanisms, and defensive strategies to identify and mitigate potential cloud-based big data attacks. This involves the development of robust encryption techniques, intrusion detection systems, access control mechanisms, and threat intelligence systems that can detect and respond to assaults, thereby enhancing the overall security of big data in cloud environments.

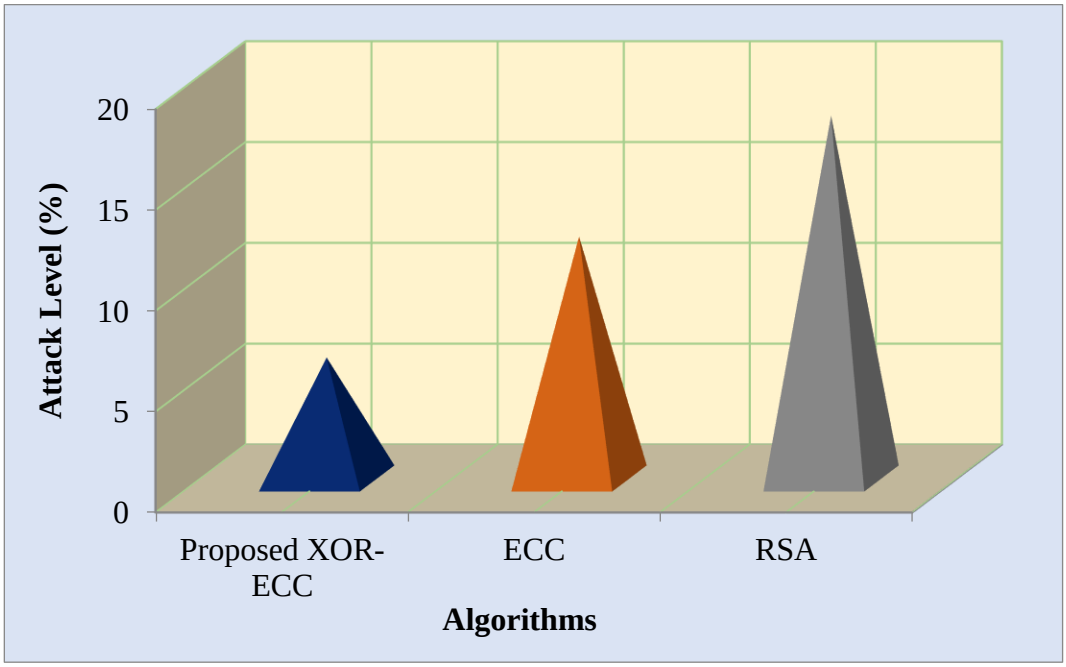

Figure 1: Comparison Attack Level

The degree of attacks is the focus in Figure 1 visual representation of the comparative analysis of the XOR-ECC. Malware and intrusions that allow an attacker to get access to a system or network are analyzed. A lower assault level, therefore, is indicative of a more successful approach. When compared to the current ECC and RSA algorithms, the suggested XOR-ECC has an attack level that is 50% and 66.67% lower, respectively. As a result, the suggested method considerably reduces the chance of assaults compared to existing systems.

5.2. Security level analysis

Analysis of the security mechanisms and controls in place to protect data in a big data over cloud environment is known as a security level analysis. To protect the privacy, authenticity, and accessibility of large data, it incorporates a thorough evaluation of many layers of security. When it comes to limiting who can see what and do what with the data, access controls play a vital role. Multi-factor authentication and role-based access control are two examples of strong authentication and authorization systems that may be used to guard against data breaches. Second, encryption is a must for safeguarding data at both rests and in motion. Data confidentiality must be protected using rigorous encryption methods and key management procedures. Third, firewalls, intrusion detection systems, and other network security measures must be implemented to protect against malicious attacks and unauthorized access. As a fourth point, validation, and verification techniques for data assure its correctness and consistency at all times. The integrity of large data may be checked for and tampered with using methods like digital signatures and hash functions. Also helpful in quickly identifying and reacting to security breaches is the use of sophisticated security monitoring and incident response systems. Tools like security information and event management (SIEM), log analysis, and real-time monitoring make it easier to spot suspicious behavior and take corrective measures in a timely manner. If you want to find and fix security holes, you need to conduct regular security audits, vulnerability assessments, and penetration testing. Maintaining a high degree of security in large data cloud systems also benefits from employee training and awareness programs on proper security practices. By examining the security of their big data installations in the cloud, businesses may find weak spots, strengthen them with the right controls, and keep their data safe.

It is of the uttermost significance to ensure the security of large data stored and processed in cloud environments [16]. The objective of this research is to improve the overall security of large data in the cloud through the development of methodologies, frameworks, and techniques. These efforts aim to protect data from unauthorized access, maintain data integrity, mitigate the risk of data breaches and disclosures, and guarantee that authorized users have uninterrupted access to the data [17]. This research endeavors to develop solutions that effectively reduce the risks associated with storing and processing sensitive data in cloud systems by placing a premium on security.

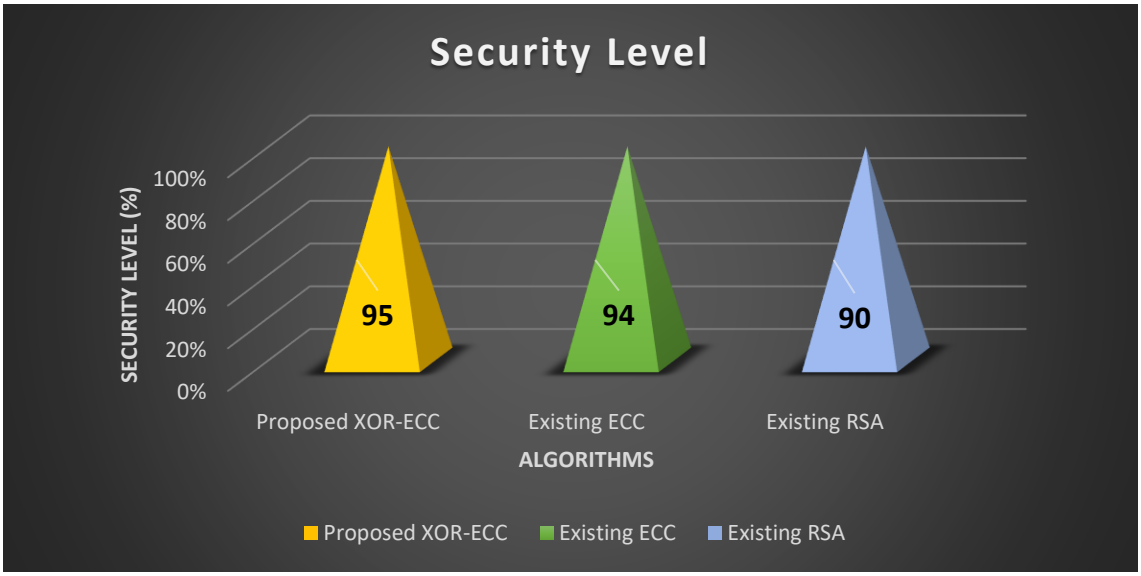

Figure 2: Comparison Security Level

A comparison of the XOR-ECC algorithm's security to that of the current ECC and RSA algorithms is shown in Figure 2 based on the findings of the experimental investigation. An attack's efficacy depends on whether or not the attacker can deduce the attributes of the plaintext or key from the ciphertext. Since there is little to no difference between data encrypted by an ordinary user and that decrypted by an attacker, the findings show that the suggested technique is extremely safe. By comparison, the suggested XOR-ECC method has a security level of 96%, 5.5% higher than RSA and 3.2% higher than ECC. This proves that the suggested XOR-ECC method for data uploading offers a far greater degree of security than the alternatives.

6. CONCLUSION AND FUTURE WORK

Compared to the current ECC (10%) and RSA (18%) algorithms, the XOR-ECC method has a lower threat level of 6%. This study shows that the suggested method does an adequate task of keeping big data (BD) in the cloud safe and private. Future work on enhancing user privacy through access pattern matching ought to employ a multidisciplinary approach, incorporating advancements in cryptography, privacy-preserving techniques, and data governance to provide stronger guarantees of privacy while enabling valuable analysis of access patterns. Future efforts to improve user privacy in access pattern matching may concentrate on advanced privacy-preserving techniques such as secure multiparty computation (MPC) and homomorphic encryption. Exploring new methods for access pattern obfuscation and perturbation can also provide additional privacy protection. Integrating privacyby-design principles into system development, evaluating existing techniques in real-world scenarios, and optimizing parameters to achieve a balance between privacy and utility is essential. Continuous research and innovation in these areas will contribute to enhancing the secrecy of users' access patterns and preserving their privacy throughout the matching process.

REFERENCES

[1] S. W. Kareem, "Hybrid public key encryption algorithms for e-commerce," *Erbil: University of Salahaddin–Hawler*, 2009.
[2] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa, "Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study," *IEEE network*, vol. 30, no. 2, pp. 54–61, 2016.
[3] H. Wang, C. D. Adenutsi, C. Wang, Z. Sun, Y. Zhang, Y. Li, Y. Zhang, and J. Wang, "Construction and application of a big data system for regional lakes in coalbed methane development," *ACS omega*, 2023.
[4] Y. Yang, Y. Lin, Z. Li, L. Zhao, M. Yao, Y. Lai, and P. Li, "Goosebt: A programmable malware detection framework based on process, file, registry, and com monitoring," *Computer Communications*, vol. 204, pp. 24–32, 2023.
[5] A. Majeed, S. Khan, and S. O. Hwang, "Toward privacy preservation using clustering based anonymization: recent advances and future research outlook," *IEEE Access*, vol. 10, pp. 53066–53097, 2022.
[6] P. Li, J. Li, Z. Huang, C.-Z. Gao, W.-B. Chen, and K. Chen, "Privacy-preserving outsourced classification in cloud computing," *Cluster Computing*, vol. 21, no. 1, pp. 277–286, 2018.
[7] K. Yang, X. Jia, K. Ren, B. Zhang, and R. Xie, "Dac-macs: Effective data access control for multiauthority cloud storage systems," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1790–1801, 2013.
[8] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, "Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption," *IEEE transactions on parallel and distributed systems*, vol. 24, no. 1, pp. 131–143, 2012.
[9] J.-J. Yang, J.-Q. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment," *Future Generation computer systems*, vol. 43, pp. 74–86, 2015.

[10] P. Jain, M. Gyanchandani, and N. Khare, "Enhanced secured map reduce layer for big data privacy and security," *Journal of Big Data*, vol. 6, no. 1, pp. 1–17, 2019.
[11] H. Matallah, G. Belalem, and K. Bouamrane, "Towards a new model of storage and access to data in big data and cloud computing," *International Journal of Ambient Computing and Intelligence (IJACI)*, vol. 8, no. 4, pp. 31–44, 2017.
[12] G. Viswanath and P. V. Krishna, "Hybrid encryption framework for securing big data storage in multi-cloud environment," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 691–698, 2021.
[13] H. K. Gupta and R. Parveen, "Comparative study of big data frameworks," in *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, vol. 1. IEEE, 2019, pp. 1–4.
[14] H. K. GUPTA and R. PARVEEN, "Security algorithms in big data overcloud computing environment," 2022.
[15] H. K. Gupta and R. Parveen, "An efficient cluster by cluster head selection approach in big data," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2022, pp. 1–6.
[16] J. B. Awotunde, R. G. Jimoh, R. O. Ogundokun, S. Misra, and O. C. Abikoye, "Big data analytics of iot-based cloud system framework: Smart healthcare monitoring systems," in *Artificial intelligence for cloud and edge computing*. Springer, 2022, pp. 181–208.
[17] M. N. Ramachandra, M. Srinivasa Rao, W. C. Lai, B. D. Parameshachari, J. Ananda Babu, and K. L. Hemalatha, "An efficient and secure big data storage in cloud environment by using triple data encryption standard," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 101, 2022.