ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

AN IMPLEMENTATION FOR SECURE DATA DEDUPLICATION ON END-TO-END ENCRYPTED DOCUMENTS

VANGALA SWATHI [1], PALLA ALEKHYA REDDY[2], AILI SIDDARTHA[3], VENEPALLY VIVEK RAO[4], SOMASANI KOWSHIK[5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— In the realm of data storage and management, secure data deduplication represents cornerstone technology for optimizing storage space and reducing redundancy. Traditional client-side deduplication approaches, while efficient regarding storage and network traffic, expose vulnerabilities that allow malicious users to infer the existence of specific files through traffic analysis. Even using a Proof of ownership scheme does not guarantee protection from all attack scenarios, specific to data deduplication. This paper introduces a novel secure data deduplication framework employing a deduplication proxy that operates on-premise, effectively mitigating the risk of such inference attacks. By leveraging convergent encryption, for proof of ownership, our solution ensures that data deduplication does not compromise data privacy or security. The deduplication proxy acts as an intermediary, performing deduplication processes on-premise. This approach not only preserves the efficiency benefits of deduplication but also enhances security by preventing external visibility into data traffic patterns. Our implementation end-to-end encryption while maintaining data deduplication storage-saving advantages. The proposed framework is suitable for organizations aiming to safeguard their data while optimizing storage resources.

Index Terms – Secure Data Deduplication, End-to-End Encryption, Cloud Storage, Data Confidentiality, Cryptography, Storage Optimization, Secure File Sharing, Privacy Preservation, Information Security, Cybersecurity.

I. INTRODUCTION

Driven by the rapid development of computer technologies and Internet applications, the amount of data is increasing explosively. So, more and more individuals and enterprises have to outsource their data to the cloud service provider (CSP) to alleviate the storage burden and share them with others who need them. At the same time, in order to prevent data content disclosure, shared data is encrypted before uploading. Now, a new problem comes. After data is encrypted, the use of data is limited. Who can access and how to access these shared data? Searchable encryption (SE) is introduced to address this problem. Song et al. first put forward symmetric searchable encryption on the basis of stream cipher, they independently encrypted each word in a file and allowed the server to determine whether a single queried keyword is presented in the file without knowing the exact word. Due to the key distribution difficulties, this proposal lacks practicality. Boneh et al. proposed public-key encryption with keyword search (PEKS), as for query expressiveness, the scheme achieved single-keyword queries. i.e. queries matched by documents indexed by a single chosen keyword. Since their pioneering work, several PEKS schemes were proposed in the literature, including fuzzy keyword search, multi keyword search, result ranking and subset keyword search.

II. RELATED WORK

A. Blockchain-based public auditing and secure deduplication with fair arbitration

Data auditing enables data owners to verify the integrity of their sensitive data stored at an untrusted cloud without retrieving them. This feature has been widely adopted by commercial cloud storage. However, the existing approaches still have some drawbacks. On the one hand, the existing schemes have a defect of fair arbitration, i.e., existing auditing schemes lack an effective method to punish the malicious cloud service provider (CSP) and compensate users whose data integrity is destroyed. These redundant data inevitably increase management overhead and computational cost during the whole data life cycle. Compared with the existing schemes, our scheme can effectively reduce the computational cost of tag verification and data storage costs. We give a

comprehensive analysis to demonstrate the correctness of the proposed scheme in terms of storage, batch auditing, and data consistency.

B. *Achieving efficient secure deduplication with user-defined access control in cloud*

Cloud storage as one of the most important services of cloud computing which significantly facilitates cloud users to outsource their data to the cloud for storage and share them with authorized users. In cloud storage, secure deduplication has been widely investigated as it can eliminate the redundancy over the encrypted data to reduce storage space and communication overhead. our scheme can maximally eliminate duplicates without violating the security and privacy of cloud users. Furthermore, extensive simulations demonstrate that our scheme outperforms the existing competing schemes, in terms of computational, communication and storage overheads as well as the effectiveness of deduplication.

C. *Novel Android malware detection method based on multi-dimensional hybrid features extraction and analysis*

In order to prevent the spread of Android malware and protect privacy information from being compromised, this study proposes a novel multidimensional hybrid features extraction and analysis method for Android malware detection. This method is based primarily on a multidimensional hybrid features vector by extracting the information of permission requests, API calls, and runtime behaviors. Finally, the unknown samples are detected and identified by using the obtained classification model. Our experiment is conducted based on 359 malicious and 500 benign applications as experimental samples, and the results indicate that our proposed method performs better in the accuracy rate of Android malware detection compared with those methods using static methods alone.

III. PROPOSED METHODOLOGY

I Even using a Proof of ownership scheme does not guarantee protection from all attack scenarios, specific to data deduplication. This paper introduces a novel secure data deduplication framework employing a deduplication proxy that operates on-premise, effectively mitigating the risk of such inference attacks. By leveraging convergent encryption, for proof of ownership, our solution ensures that data deduplication does not compromise data privacy or security. The deduplication proxy acts as an intermediary, performing deduplication processes on-premise Our

implementation end-to end encryption while maintaining data deduplication storage-saving advantages.

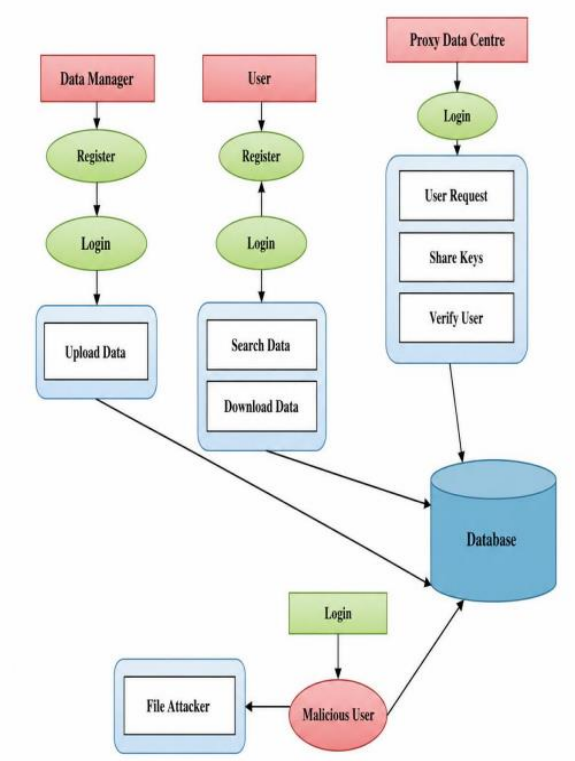
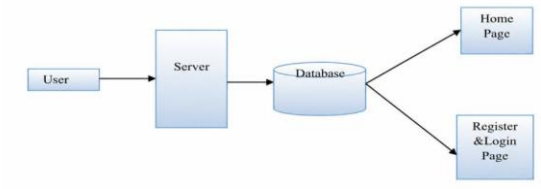


Fig. 1: System Overview

1. User Interface

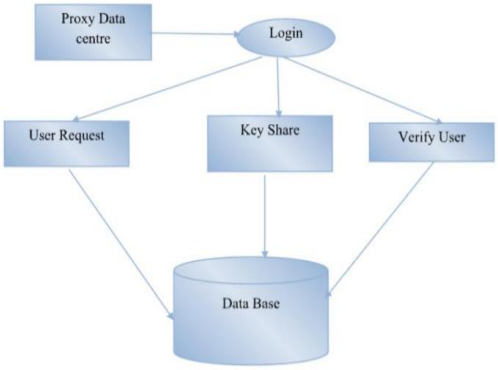
In this module we design the windows for the project. These windows are used for secure login for all users. To connect with server user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password and Email id, into the server. Server will create the account for the entire user to maintain upload and download rate. Name will be set as user id. Logging in is usually used to enter a specific page. After successful authentication, users are redirected to the dashboard, where they can upload encrypted documents, view previously uploaded files, and download authorized documents. During the upload process, the system automatically encrypts the file before storing it in the cloud server. The deduplication mechanism checks whether an identical encrypted file already exists. If a duplicate file is detected, the system avoids storing another copy and instead updates the ownership information, thereby reducing storage consumption while maintaining data integrity.



2. Proxy Data Center

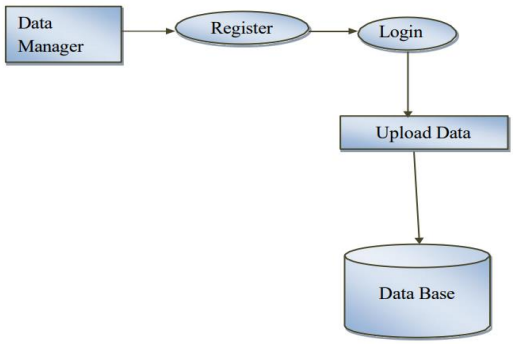
The Proxy Data Center acts as an intermediate trusted entity between the users and the cloud server. It is responsible for handling user authentication, secure key management, access control, and communication between the client and the cloud storage system. The proxy ensures that only authorized users can upload, access, or share encrypted documents. The Proxy Data Center provides a secure login interface where the administrator logs in using a valid

user ID and password. After successful authentication, the administrator can monitor all registered users, verify user requests, and manage secure communication within the system.



3. Data Manager

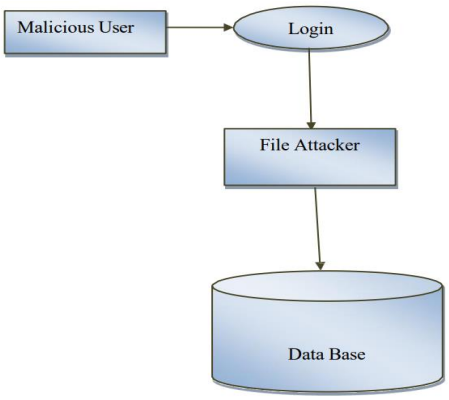
The Data Manager (DM) is one of the core modules of the proposed system. It is responsible for managing document storage, verifying user requests, controlling data access, and ensuring that all uploaded documents are securely maintained in the cloud environment. The Data Manager plays an important role in maintaining the confidentiality, integrity, and availability of the stored data. Initially, the Data Manager must register with the system by providing the required details. After successful registration, the Data Manager can log in securely using a valid username and password. Only authenticated Data Managers are authorized to perform administrative operations on the stored documents. Once logged in, the Data Manager can upload and store data, such as text documents and other supported file formats, into the cloud server. Before storage, the documents are encrypted using end-to-end encryption techniques to ensure that the contents remain confidential and protected from unauthorized access.



4.Malicious User

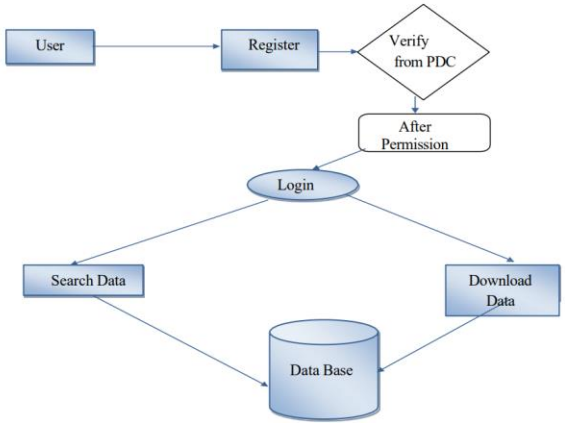
The Malicious User module represents unauthorized users or attackers who attempt to gain illegal access to the secure cloud storage system. This module is included to demonstrate how the proposed system protects encrypted documents against unauthorized access, data tampering, and other security threats. A malicious user may attempt to log into the system using invalid or stolen user credentials. However, the system verifies every login request through secure authentication mechanisms. If the provided username or password is incorrect, the login request is rejected, and the unauthorized user is denied access to the system. The malicious user may also attempt to attack stored files by modifying, deleting, or accessing confidential documents without proper

authorization. Since all files are protected using end-to-end encryption, the attacker cannot understand or modify the actual content of the documents without possessing the correct decryption key.



5.User

The User module represents the end users who interact with the secure cloud storage system to upload, search, access, and download encrypted documents. This module ensures that only authenticated and authorized users can perform operations on the stored data, thereby maintaining the security and privacy of the system. Initially, a new user must register by providing the required details such as username, password, email address, and other necessary information. After successful registration, the user's account is stored securely in the system database. To access the system, the user must log in using valid credentials.



IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results demonstrated that the system successfully identified and eliminated duplicate encrypted documents without compromising end-to-end security, resulting in significant storage optimization and reduced network bandwidth consumption.

The analysis showed that the proposed deduplication mechanism maintained strong data confidentiality and integrity while introducing only minimal computational and communication overhead.

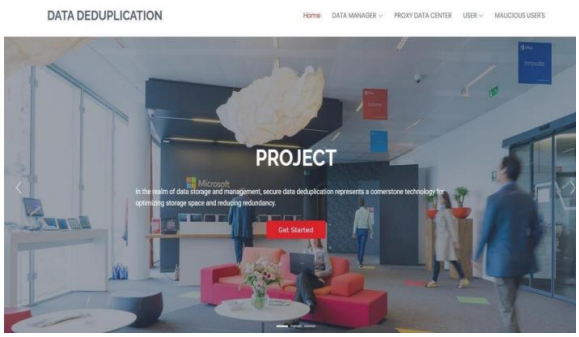


Fig. 2: Homepage

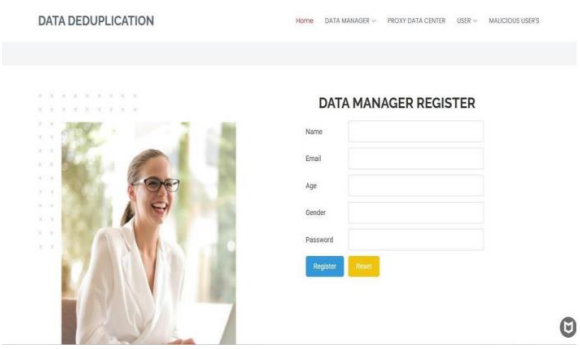


Fig. 3: Data Manager Register

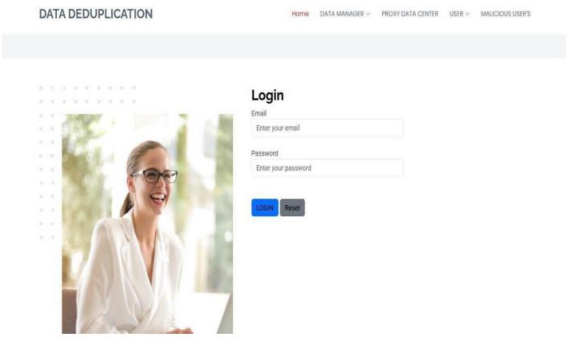


Fig. 4: Login

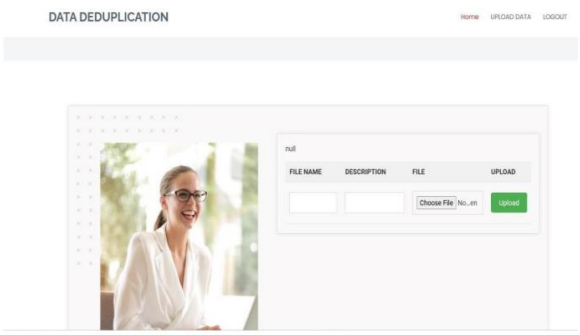


Fig. 5: Upload

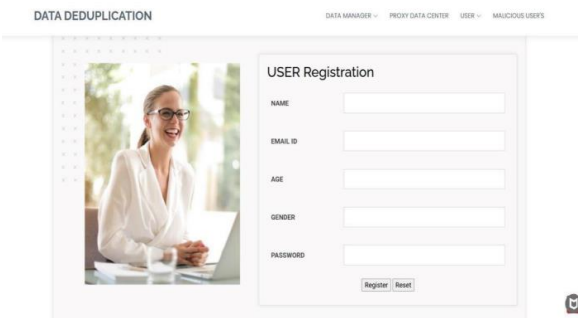


Fig.6: User Registration

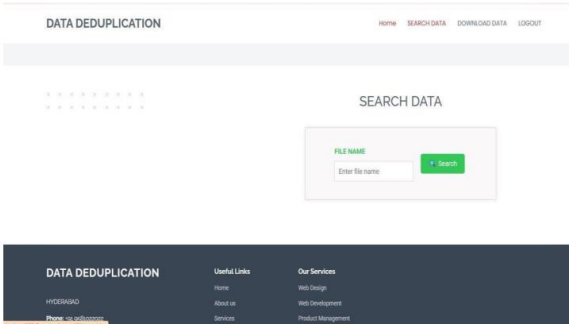


Fig.7: Search Data

V. CONCLUSION

In this paper, we propose a deduplication scheme for encrypted data, named Data Dedupulication. It does not relay on any online trusted third party, and it allows dynamic ownership updating. We use cryptographic primitives such as proxy server convergent encryption. Compared with previous schemes, the security of Data Dedupulication is significantly enhanced. In our design, the proxy server can work as an intermediary, which focuses on three core functionalities, in particular, updating the ownership list when a user updates his data, assisting an initial uploader to validate subsequent uploaders, and delivering data encryption with keys. The optimized Proxy server in Data Dedupulication can easily and effectively perform deduplication on encrypted data. Meanwhile, it is enforced that a user is allowed to access the data only if he can provide a valid key. Simulation experiments show that our scheme is

applicable and efficient. Overall, the proposed secure data deduplication system successfully achieves its objectives by providing secure cloud storage, efficient elimination of duplicate encrypted files, reliable access control, and strong protection against unauthorized access and malicious attacks. The project demonstrates that secure deduplication can effectively reduce storage costs while maintaining the confidentiality, integrity, and availability of sensitive documents in cloud computing environments, making it a practical and reliable solution for modern cloud-based data storage systems.

REFERENCES

[1] X. Yang, R. Lu, J. Shao, X. Tang, and A. Ghorbani, ‘‘Achieving efficient and privacypreserving multi-domain big data deduplication in cloud,’’IEEETrans. ServicesComput., earlyaccess,Nov.13,2018,doi:10.1109/TSC.2018.2881147.

[2] M.Bellare, S.Keelveedhi, and T.Ristenpart, ‘‘Message-locked encryption and secure deduplication,’’ in Advances in Cryptology Euro crypt. Berlin, Germany: Springer, 2013, pp. 296–312.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart, ‘‘DupLESS: Server-aided encryption for deduplicated storage,’’ in Proc. 22nd Usenix Conf. Secure. Usenix Assoc., 2013, pp. 179–194.

[4] P. Puzio, R. Molva, M. Onen, and S. Loureiro, ‘‘ClouDedup: Secure deduplication with encrypted data for cloud storage,’’ in Proc. IEEE 5thInt. Conf. Cloud Compute. Technol. Sci., Dec. 2013, pp. 363–370.

[5] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, A Secure Data Deduplication Scheme for Cloud Storage. New York, NY, USA: IBM Corporation, 2014, pp. 99–118.

[6] P.Puzio, R.Molva, M.Önen, and S.Loureiro, ‘‘Perfect Dedup: Secure data deduplication,’’inProc.10thInt.Workshop DataPrivacyManage.,Secur.Assurance(DPM) 4thInt.WorkshopQASA,inLecture NotesinComputerScience, vol. 9481. Vienna, Austria: Springer, Sep. 2015, pp. 150–166.

AUTHORS Profile

Mrs. VANGALA SWATHI has received her B.Tech from JNTUH, Hyderabad in 2015, M.Tech degree in Computer science from JNTUH, Hyderabad in 2017. She is dedicated to teaching field from the last 7 years. Her research area Deep Learning. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.

Ms. PALLA ALEKHYA REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. AILI SIDDARTHA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. VENEPALLY VIVEK RAO pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. SOMASANI KOWSHIK pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.