IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Data Deduplication Using DBSCAN With Admissible Clustering Variance

D.SRINIVASU
Assistant Professor
Department of CSE , Guru Nanak Institute of Technology , Hyderabad ,Telangana ,India

Abstract : To protect privacy, users prefer to store encrypted data on cloud servers. Cloud servers reduce storage costs and network bandwidth by eliminating duplicate copies. To solve the problem of possible internal information flow, the concept of cluster deviation was proposed for the first time. We improve the DBSCAN algorithm to tolerate cluster change. The data duplication model is built on a new algorithm that treats users as instances of clusters. Instead of immediately regrouping new users, certain deviations are allowed when assigning users to existing categories. We determine the popularity of data based on user grouping results and use different encryption methods to protect the security of unpopular data more effectively. The performance of the algorithm is analyzed and compared with experiments with other methods, and the results confirm the feasibility and effectiveness of the proposed duplication method.

IndexTerms - Text Recognition, Character Recognition, Document Image Analysis.

I. INTRODUCTION

The development and implementation of cloud services has led more and more users to store their data in a cloud server (CS). To save bandwidth and storage space, servers usually use data reduction techniques, ie. keeps only one copy of data and eliminates redundancy. However, when sending data to CS, users want to encrypt the data to protect their privacy and prevent the data content from being obtained by CS or other attackers. In traditional encryption methods, users choose keys and encrypt plaintext randomly. This makes the ciphertext stored in CS different from even the same plaintext, making decryption very difficult. On the other hand, if users encrypt data with the same key, it can significantly weaken the security of the system. Convergent encryption (CE) has been proposed to effectively solve this problem. In CE, the key is derived from the plaintext, so the same plaintext produces the same key, which in turn produces the same ciphertext. This enables the duplication of encrypted data. However, CE has security flaws and is vulnerable to offline brute force attacks because the key derivation process is deterministic. In recent years, many researchers have worked on developing different deduplication methods based on Message Lock Encryption (MLE). In response to the above attacks, Stank et al. proposed a popular distribution based deduplication system. Information of different popularity is encrypted using different encryption methods to further save cloud storage space and network bandwidth. Puzio et al. The proposed Clouded up model with a metadata manager and an additional server defined in CS: the server adds an encryption layer to prevent attacks against the CE, thus protecting data confidentiality. Dup LESS used key management and OPRF (Forgotten Pseudorandom Function), which is a high security algorithm, to generate the key. Zhang et al. used an elliptic curve encryption algorithm to achieve data confidentiality, and different encryption methods were used for popular and unpopular data to reduce computational cost. Liu et al. proposed a secure data replication system that does not require a third-party server. This system uses a password - And it also removes the dependency on third-party servers and improves security. However, this requires all users participating in the protocol to be online when exchanging keys, which greatly increases communication costs and reduces practicality. Several existing data duplication systems focus on the

protection and transmission of encryption keys and the detection of duplicate data, neglecting the impact of users on duplication. Among the many systems that separate data by popularity.

II. LITERATURE SURVEY

[1] introduced data deduplication as a crucial method for enhancing storage efficiency within cloud computing. By consolidating redundant files into a single copy, cloud service providers substantially decrease both storage space and data transfer expenses. However, the conventional deduplication approach, widely adopted, poses a significant risk to data confidentiality due to cloud computing's data storage models. Addressing this challenge, we propose a secure deduplication scheme based on a Trusted Execution Environment (TEE). In our scheme, each cloud user is allocated a specific privilege set, enabling deduplication only when users possess the correct privileges. Additionally, our scheme reinforces convergent encryption with user privileges and leverages TEE for secure key management, bolstering the cryptosystem's resilience against chosen plaintext and cipher text attacks. Security analysis demonstrates that our scheme effectively supports data deduplication while safeguarding the confidentiality of sensitive data. Furthermore, we implement a prototype and evaluate its performance, with experiments confirming the practicality of our scheme's overhead in real-world environments.

[2] explored the challenge of deduplicating encrypted big data in the cloud. Cloud computing, offering service provision through Internet resources, notably features data storage. To maintain data holder privacy, data are typically encrypted when stored in the cloud. However, encrypted data pose hurdles for cloud data deduplication, crucial for big data storage and processing. Traditional deduplication methods are ineffective on encrypted data, and existing encrypted data deduplication solutions exhibit security vulnerabilities and lack flexibility in supporting data access control and revocation. Hence, practical deployment remains limited. To address this, a scheme for deduplicating encrypted cloud data using ownership challenge and proxy re-encryption, integrating data deduplication with access control is proposed. Through extensive analysis and computer simulations, we demonstrate the scheme's superior efficiency and effectiveness, particularly in the context of big data deduplication in cloud storage, paving the way for practical deployment.

[3] introduced R-dedup, a secure client-side deduplication approach for encrypted data that eliminates the need for third-party entities. Deduplication is a widely utilized technology in various applications, including cloud computing services, to enhance storage performance by storing only one copy of identical data. However, when encryption is employed to ensure data confidentiality in the cloud, deduplication becomes challenging due to the use of different encryption keys for the same content. In our paper, we propose R-Dedup, a randomized, secure, cross-user deduplication scheme. R-Dedup operates independently of third-party entities or assistance from other users. In this scheme, randomization ensures that users sharing identical copies of a file use the same random value via ElGamal encryption. Through security analysis and experimental evaluations, we demonstrate that R-Dedup is lightweight, ensuring both data privacy and integrity.

.

[4] presented an innovative approach to tag deduplication in cloud storage, focusing on resistance against side channel attacks. Tag deduplication is a promising technique for reducing redundancy in cloud storage, involving the signing of integrity tags with content-associated keys rather than user-associated secret keys. However, existing solutions often reintroduce the linkage between cloud users and their integrity tags to achieve public auditability, potentially exposing a side channel for malicious third-party auditors to compromise file privacy. This vulnerability poses a significant obstacle to the widespread adoption of tag deduplication. To address this challenge, we propose a secure aggregation-based tag deduplication scheme (ATDS) that prioritizes resistance against side channel attacks during public verification. Our scheme defines a user-associated integrity tag based on a content-associated polynomial and employs a Lagrangian interpolation-based aggregation strategy for tag deduplication. By utilizing content-associated public keys instead of user-associated ones for auditing, our approach ensures that third-party auditors can only verify data correspondence to a group of users rather than specific owners. Security analysis and experimental results confirm the effectiveness of our scheme in resisting side channel attacks and its superior efficiency compared to existing methods.

[5] explored reclaiming space from duplicate files in a serverless distributed file system, specifically focusing on the Farsite distributed file system, which ensures availability by replicating files across multiple desktop computers. Given the substantial storage space consumed by this replication, reclaiming used space is essential. Analysis of over 500 desktop file systems reveals that almost half of all consumed space is attributed to duplicate files. To address this, we propose a mechanism to reclaim space from incidental duplication, making it available for controlled file replication. Our approach involves convergent encryption, enabling duplicate files to be consolidated into the space of a single file, even when encrypted with different users' keys. Additionally, we introduce SALAD, a Self-Arranging Lossy Associative Database, for aggregating file content and location information in a decentralized, scalable, and fault-tolerant manner. Through large-scale simulation experiments, we demonstrate that the duplicate-file coalescing system is scalable, highly effective, and fault-tolerant.

III. PROPOSED METHODOLOGY

This paper proposes a blockchain-based information exchange platform to ensure the manufacturing quality of aerospace suppliers. First, the paper presents the possibility of integrating manufacturing supply chain quality management with block chain technology. Second, the architecture of the quality and information-sharing platform for the production process of new aviation suppliers is presented based on high-quality national and island aviation suppliers. Then, a detailed method to implement quality and information security sharing is proposed to support the real-time and regular operation of the sharing platform. Build critical technologies on this foundation, such as production-level data block compression models, data storage security sharing, and vendor evaluation models. Finally, depending on the data collection of the production processes of the supplier's products, common implementation practices based on the specific aircraft industry fleet under the control of platform architecture and technology. The application platform integrates data transfer and query components, providing practical and intelligent sharing solutions for airline product quality data.

Advantages- Facilitating the seamless and efficient operation of the platform in real-time.
Establishing a robust framework for secure data storage and sharing, alongside supplier evaluation models.
Offering airlines tailored and intelligent sharing solutions that are both practical and effective.

a. SYSTEM ARCHITECTURE

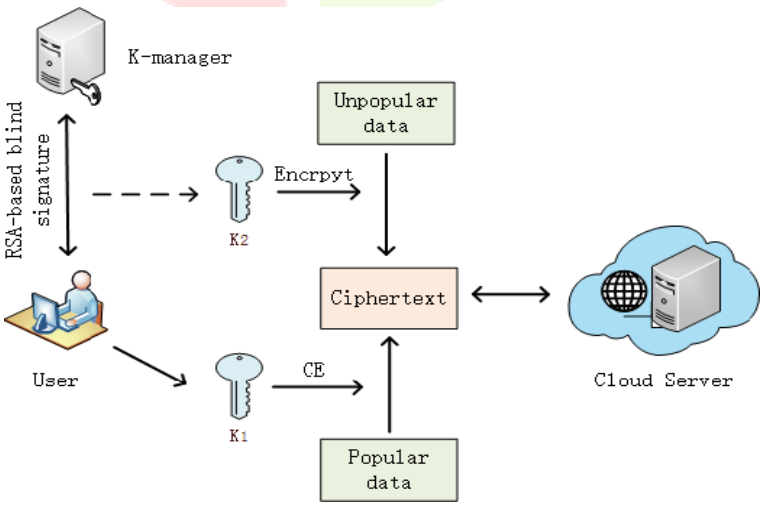

Fig. 3.1 Architecture

The methodology involves a user registration process followed by login authentication. Once logged in, the user can utilize the K-Manager platform to upload documents. Additionally, users have the ability to send requests to the K-Manager and perform queries based on uploaded documents. The downloaded files are encrypted for security purposes. Users can also send requests to the cloud server, which requires login authentication. Upon login, the cloud server reviews and approves access requests. It has visibility into all data and user information stored within it. Furthermore, the cloud server is responsible for managing key requests from users, which are subsequently approved.

The K-Manager then facilitates the transmission of secret keys to the cloud server, enabling users to securely download files.In the event of incorrect key input by the user, a warning is issued, and repeated offenses result in permanent account blocking. Additionally, measures are in place to safeguard against potential attacks on the files.

IV. RESULTS AND DISCUSSIONS

User Authentication:- Upon accessing the login page, users enter their credentials for validation against a user database.Successful authentication redirects users to the home page, displaying a personalized welcome message and access to application functionalities.

Profile Management:- Users can view and update their profile information, including name, email, and password, with possible additional functionalities like password confirmation.

File Management:-Users can upload files, with the system generating a unique identifier (hash) for each file and employing DBSCAN clustering to detect and handle duplicate files.Uploaded files are listed with details such as file name, size, and upload date, offering options for file management (e.g., delete, download).

Administrator Access:-Admins access a separate login page with dedicated credentials, granting them special privileges for user and file management.Admins can view and edit user profiles, deactivate accounts, and perform administrative tasks like viewing registered users and uploaded files.

System-Level File Management:-Admins have access to a comprehensive list of all uploaded files, with options to manage files system-wide, including deleting duplicates, analyzing usage patterns, and setting access permissions, with additional information such as the uploader's username and storage location.

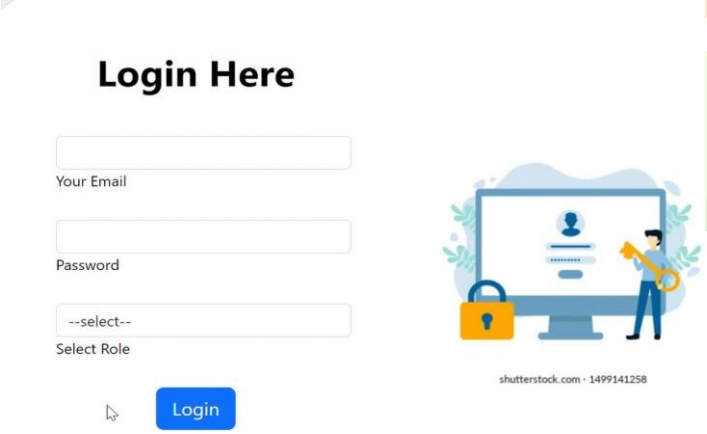

Fig 4.1 User Login

Fig. 4.2 Home Layout

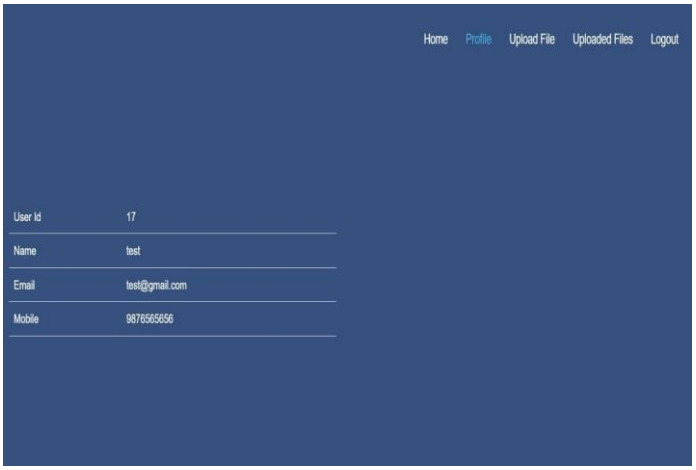

Fig. 4.3 Profile Layout

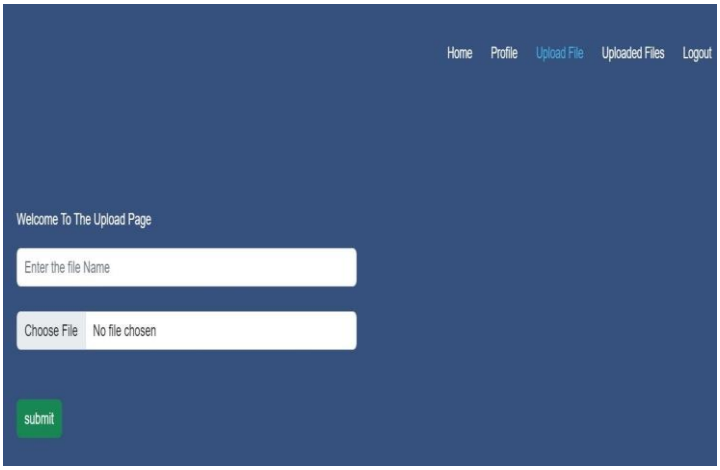

Fig. 4.4 File Upload

Fig. 4.5 View Files

Fig. 4.6 View Users

V. CONCLUSION

This paper has addressed the critical issue of deduplicating encrypted data and proposed a novel TCD-DBSCAN algorithm. By introducing the concept of clustering deviation, our approach significantly reduces the risk of inadvertent data leakage during the deduplication process, even when data originates from the same source. Utilizing symmetric encryption and blind signature protocols enhances data security without requiring users to transmit encryption keys online, thereby enhancing deduplication efficiency. Security and performance analyses confirm the effectiveness and practicality of our proposed scheme.

VI. REFERENCES

[1] Y. Fan, X. Lin, W. Liang, G. Tan, and P. Nanda, ‘‘A secure privacy preserving deduplication scheme for cloud computing,’’ Future Gener. Comput. Syst., vol. 101, pp. 127–135, Dec. 2019.

[2] Z. Yan, W. Ding, X. Yu, H. Zhu, and R. H. Deng, ‘‘Deduplication on encrypted big data in cloud,’’ IEEE Trans. Big Data, vol. 2, no. 2, pp. 138–150, Jun. 2016.

[3] Y. Zhai, M. Ibrahim, Y. Qiu, F. Boemer, Z. Chen, A. Titov, and A. Lyashevsky, ‘‘Accelerating encrypted computing on Intel GPUs,’’ in Proc. IEEE Int. Parallel Distrib. Process. Symp. (IPDPS), May 2022, pp. 705–716.

[4] X. Yang, R. Lu, J. Shao, X. Tang, and A. A. Ghorbani, ‘‘Achieving efficient and privacy- preserving multi-domain big data deduplication in cloud,’’ IEEE Trans. Services Comput., vol. 14, no. 5, pp. 1292–1305, Sep. 2021.

[5] C. Guo, X. Jiang, K.-K.-R. Choo, and Y. Jie, ‘‘R-dedup: Secure client-side deduplication for encrypted data without involving a third-party entity,’’ J. Netw. Comput. Appl., vol. 162, Jul. 2020, Art. no. 102664.

[6] X. Tang, L. Zhou, B. Hu, and H. Wu, ‘‘Aggregation-based tag deduplication for cloud storage with resistance against side channel attack,’’ Secur. Commun. Netw., vol. 2021, pp. 1–15, Feb. 2021.

[7] J. R. Douceur, A. Adya, W. J. Bolosky, P. Simon, and M. Theimer, ‘‘Reclaiming space from

duplicate
files in a serverless distributed file system,'' in Proc. 22nd Int. Conf. Distrib. Comput. Syst., 2002, pp.
617–624.

[8] L. Wang, B. Wang, W. Song, and Z. Zhang, ''A key-sharing based secure deduplication scheme in cloud
storage,'' Inf. Sci., vol. 504, pp. 48–60, Dec. 2019. [9] Y. Zhao and S. S. M. Chow, ''Updatable block-
level message-locked encryption,'' in Proc. ACM Asia Conf. Comput. Commun. Secur., Apr. 2017, pp.
449– 460.

[10] H. Yuan, X. Chen, J. Li, T. Jiang, J. Wang, and R. H. Deng, ''Secure cloud data deduplication with efficient re-encryption,'' IEEE Trans. Services Comput., vol. 15, no. 1, pp. 442–456, Jan. 2022.