e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 12, December 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Data Deduplication Scheme Based on DBSCAN with Tolerable Clustering Deviation

Shaikh Mohamad

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, India

ABSTRACT: To protect data privacy, users prefer to store encrypted data in cloud servers. Cloud servers reduce the cost of storage and network bandwidth by eliminating duplicate copies. To address the potential internal data leakage problem, the concept of clustering deviation is proposed for the first time. We improve the DBSCAN algorithm to tolerate clustering deviation. A data deduplication scheme is built upon the new algorithm, which considers users as clustering samples. Instead of immediately re-clustering new users, a certain deviation is tolerated to assign the users to the existing classes. We determine the popularity of the data according to user clustering results and apply different encryption schemes to protect the security of unpopular data more effectively. The performance of the algorithm is analyzed and compared with other methods through experiments, and the results verify the feasibility and efficiency of the proposed deduplication scheme.

KEYWORDS: Deduplication scheme, DBSCAN with tolerable clustering deviation.

I. INTRODUCTION

The rise of cloud computing has prompted a growing number of users to store data on cloud servers (CS). To optimize bandwidth and storage, servers commonly use data deduplication techniques, maintaining only a single instance of each data set to eliminate redundancy. However, users also seek to encrypt their data to protect their privacy and prevent unauthorized access by the CS or other attackers. Traditional encryption methods, where users randomly select keys to encrypt plaintext, result in varied ciphertexts for the same plaintext, complicating deduplication. Conversely, using the same key for encryption reduces system security.

Convergent Encryption (CE) addresses this by deriving the encryption key from the plaintext itself, ensuring that identical plaintexts produce identical ciphertexts, thus enabling data deduplication of encrypted data. However, CE has security vulnerabilities and is prone to offline brute-force attacks due to its deterministic key derivation process. In response, researchers have developed various Message-Locked Encryption (MLE)-based deduplication schemes. For instance, Stanek et al. proposed a scheme based on data popularity, using different encryption methods for data of varying popularity to optimize cloud storage and bandwidth. Puzio et al. introduced ClouDedup, incorporating an additional encryption layer to enhance data confidentiality. DupLESS employs a high-security key management.

II. LITERATURE REVIEW

Y. Fan & X. Lin dicussed that “Data deduplication is a crucial technique to enhance storage efficiency in cloud computing. By directing redundant files to a single copy, cloud service providers significantly curtail their storage space and data transfer expenses. Despite the widespread adoption of the traditional deduplication approach, it poses a considerable risk of compromising data confidentiality due to the prevalent data storage models in cloud computing. To address this challenge in cloud storage, we initially propose a Trusted Execution Environment (TEE) based secure deduplication scheme. In our approach, each cloud user is assigned a set of privileges; deduplication is permissible only when cloud users possess the correct privileges. Additionally, our scheme supplements convergent encryption with user privileges and relies on TEE for secure key management, thereby bolstering the system's resilience against chosen plaintext and chosen ciphertext attacks. A security analysis attests that our scheme adequately supports data deduplication while safeguarding the confidentiality of sensitive data.”

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Z. Yan & W. Ding explained that "Cloud computing revolutionizes service provision by reallocating various resources via the Internet, with data storage emerging as its cornerstone service. To safeguard data holder privacy, data are frequently stored encrypted in the cloud. However, encrypted data pose challenges for cloud data deduplication, critical for big data storage and processing in the cloud. Traditional deduplication schemes are ineffective on encrypted data, while existing encrypted data deduplication solutions suffer from security vulnerabilities, lacking flexibility in supporting data access control and revocation, thus hindering practical deployment. In this study, we propose a scheme to deduplicate encrypted data stored in the cloud using ownership challenge and proxy re-encryption. This scheme seamlessly integrates cloud data deduplication with access control. Through extensive analysis and computer simulations, we evaluate its performance, revealing its superior efficiency and effectiveness, particularly for potential practical deployment in big data deduplication within cloud storage."

X. Jiang, K.-K.-R. Choo, and Y. Jie conclude that" Deduplication technology is widely utilized across various applications, notably in cloud computing services, to optimize storage performance. This involves cloud service providers consolidating duplicate data into a single copy. However, the integration of encryption for ensuring data confidentiality in cloud storage poses significant challenges for deduplication, primarily due to the diverse encryption keys used for similar content. In this study, we present a novel randomized, secure, cross-user deduplication scheme (R-Dedup). Unlike traditional methods, R-Dedup operates autonomously, devoid of third-party entities like additional cloud servers, and does not necessitate assistance from other users. Under the R-Dedup framework, a randomized approach facilitates users in sharing identical file copies through ElGamal encryption, thereby ensuring both data privacy and integrity. Rigorous security analysis and experimental evaluations substantiate R-Dedup's lightweight nature and its ability to provide robust protection for data confidentiality and integrity."

III. METHODOLOGY

1.EXISTING SYSTEM

Traditional encryption methods lead to varying ciphertexts for identical plaintexts, hindering deduplication. Conversely, using a single key compromises security. Convergent encryption (CE) addresses this by deriving keys from plaintexts, enabling consistent ciphertexts for identical data. However, CE is susceptible to offline attacks due to its deterministic key derivation. Researchers have developed Message-Locked Encryption (MLE)-based deduplication schemes to counter these vulnerabilities. Stanek et al. introduced a popularity-based deduplication scheme in response to security concerns.

2.PROPOSED SYSTEM

This research paper introduces a blockchain-based platform for sharing aviation supplier manufacturing process quality data. Initially, it discusses the potential integration of manufacturing supply chain quality management with blockchain technology. Subsequently, it outlines the architecture of the quality and data sharing platform for new aviation suppliers, categorized based on quality state and types. The paper then proposes a detailed method for implementing quality and data security sharing to ensure real-time and orderly platform operation. Critical technologies such as manufacturing quality data block packaging models, data storage security sharing, and supplier assessment models are developed. Finally, the platform facilitates data collection from supplier product production processes, enabling shared application practices within a specific aircraft industrial park under platform supervision.

3.SYSTEM ARCHITECTURE

In this system, users must register with all their details before logging in. The K-Manager can upload documents, and users can send requests to the K-Manager. Users can search queries related to uploaded documents, which are encrypted for download. Additionally, users can send requests to the cloud server, where they log in and await key approval. The cloud server has access to all data and user information, approving key requests. Upon receiving a user request, the K-Manager sends a secret key to the cloud server. Users can then download files, with incorrect key attempts resulting in permanent blocking and potential security breaches.

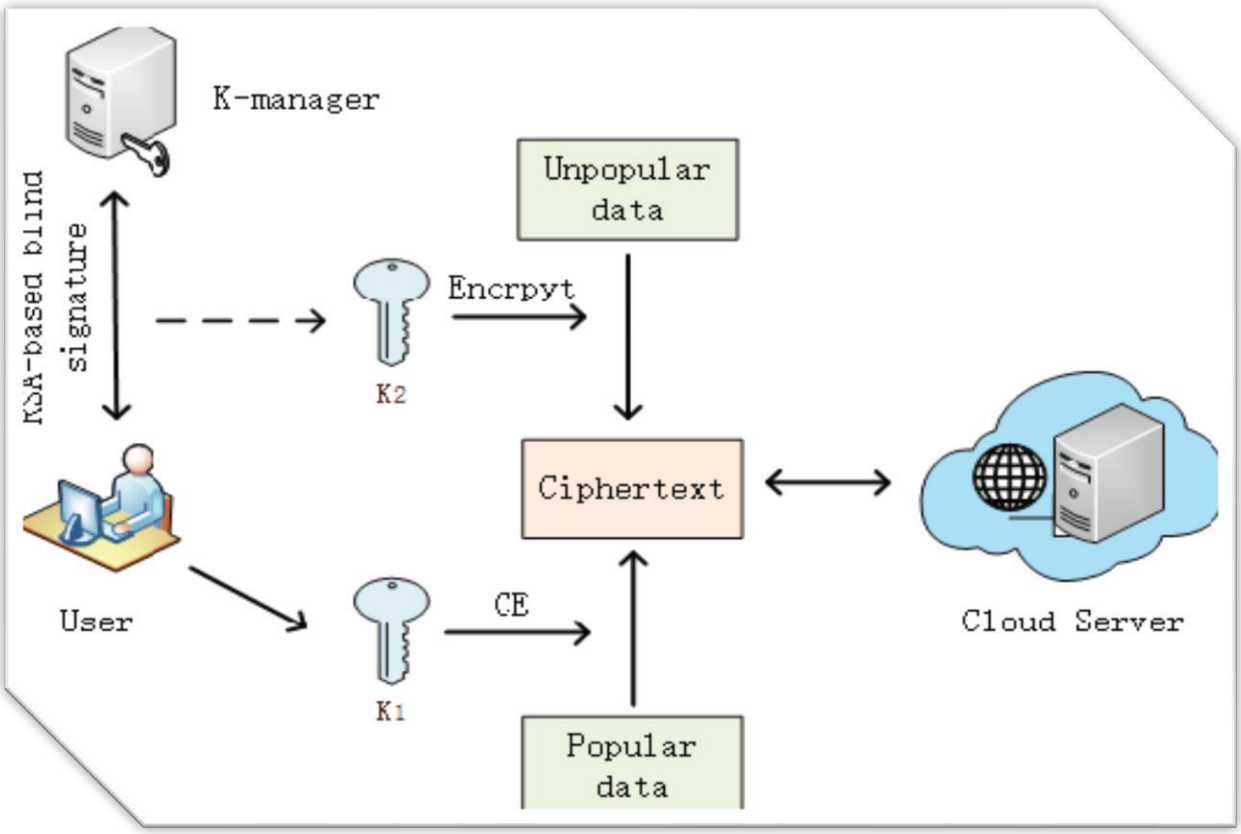
© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IMPLEMENTATION



DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised clustering algorithm designed for handling non-linear datasets effectively. Unlike the k-means clustering method, DBSCAN does not necessitate the upfront specification of the number of clusters. Instead, it operates by constructing nearest neighbor graphs and forming clusters with arbitrary shapes within datasets, including noise or outliers, contrasting with k-means clustering, which typically produces spherical clusters. DBSCAN identifies clusters from dense regions, separated by regions with low or no densities. The algorithm relies on two parameters: the radius of neighborhoods (eps or ε) around a given data point (p) and the minimum number of data points (minPts) required within an ε-neighborhood to form clusters. It is an unsupervised machine learning algorithm capable of identifying clusters of varying shapes within datasets, even in the presence of noise and outliers. Epsilon defines the radius of the neighborhood around each point, while minPts specifies the minimum number of points necessary to constitute a cluster.

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

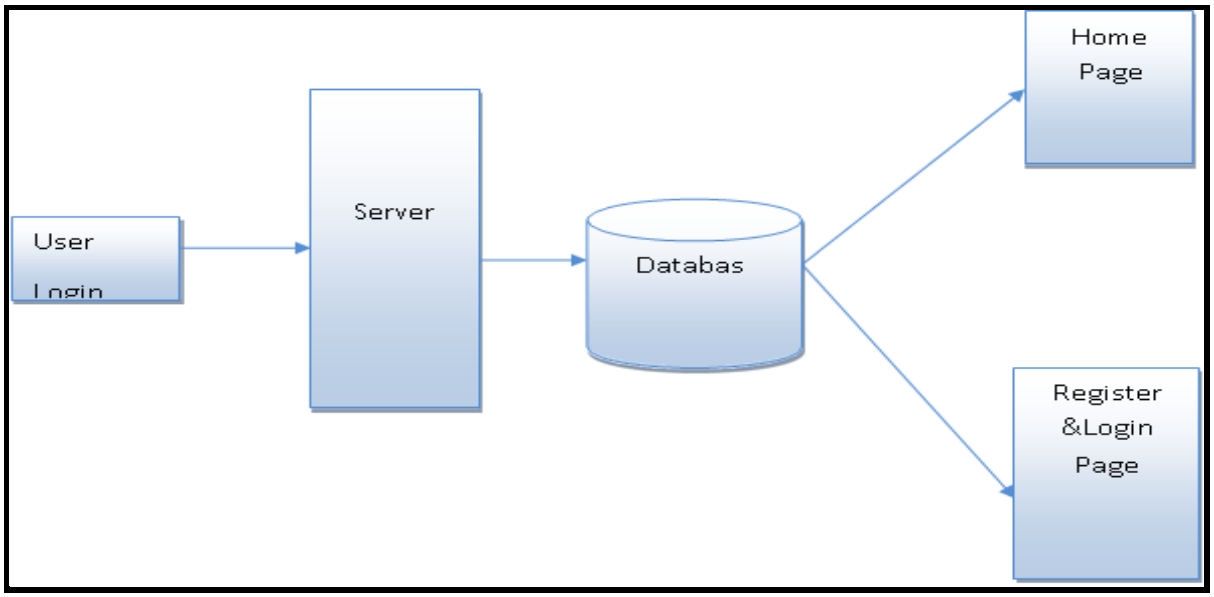
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

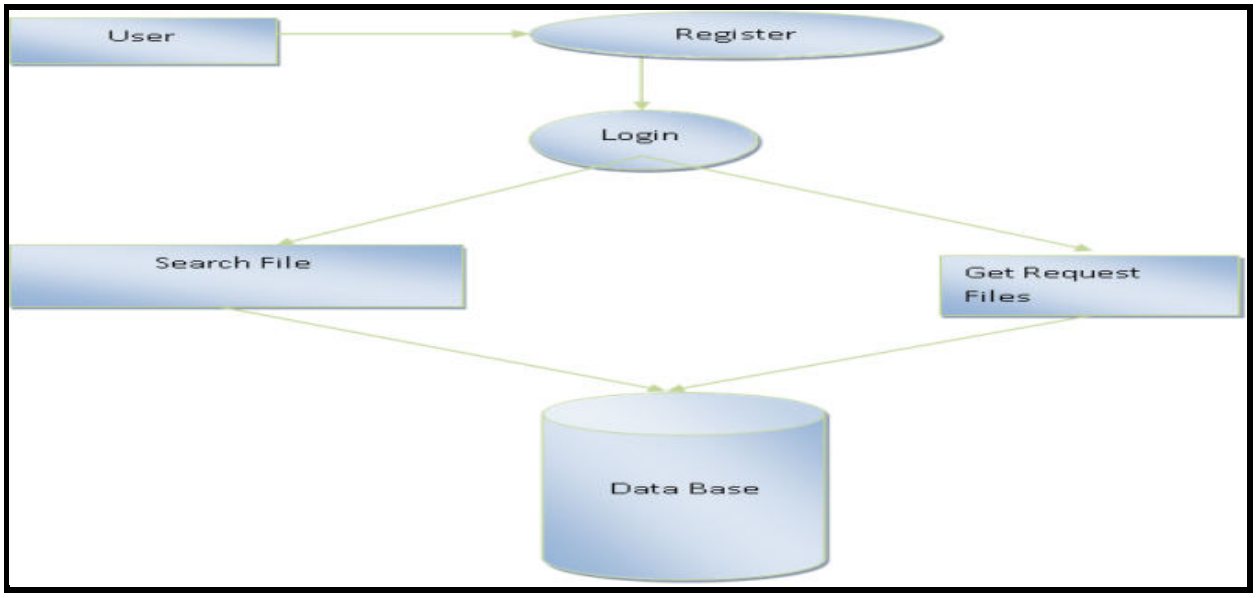
Modules:

1.User Interface Design Module:



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.

2.User Module:



© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

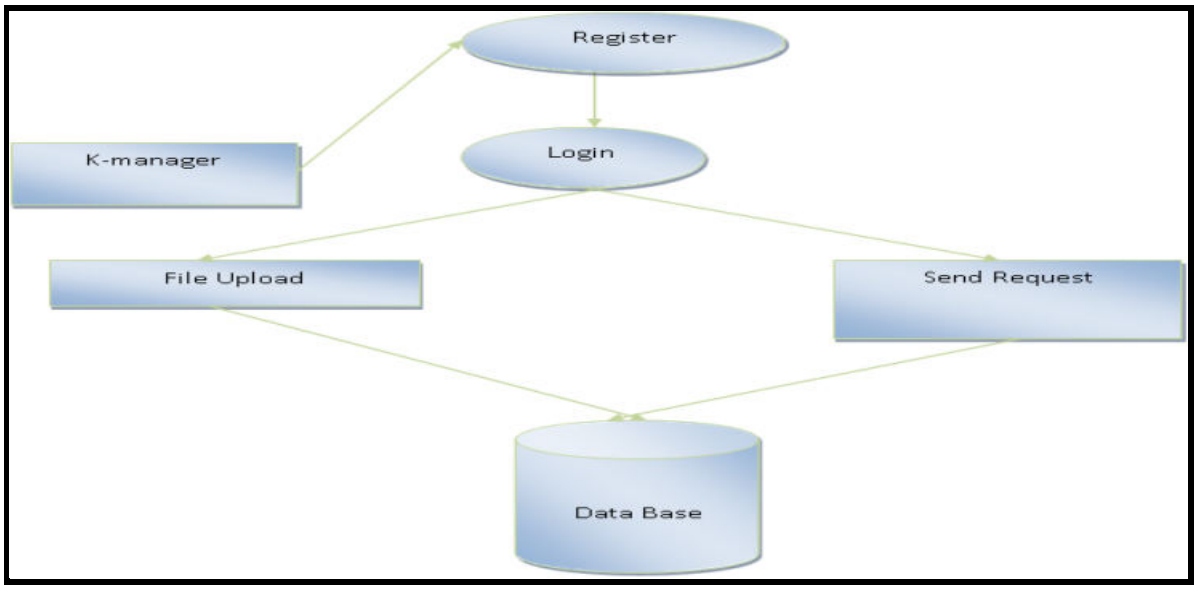
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

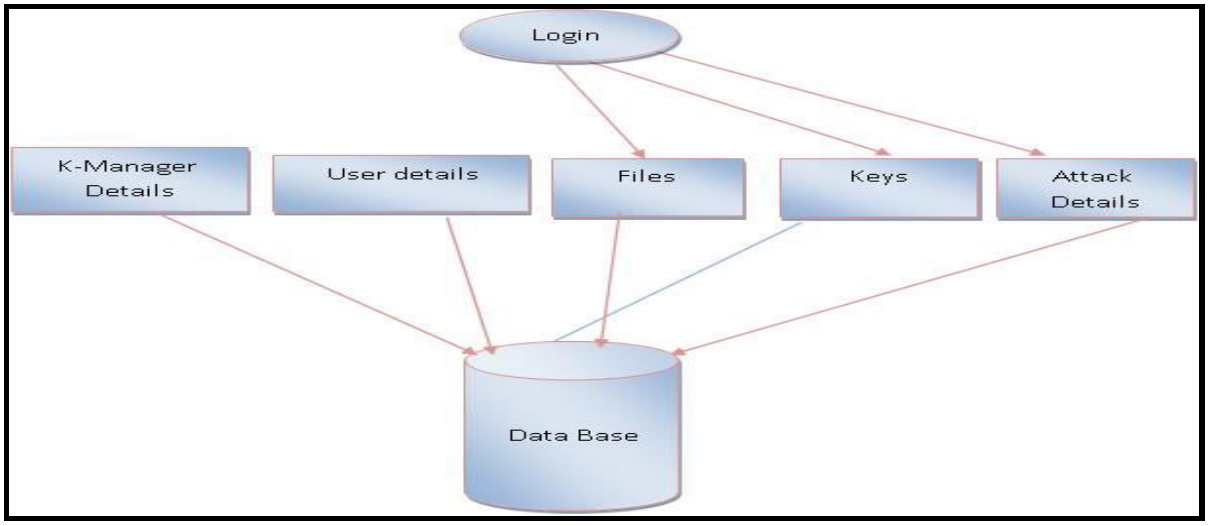
In this module data User can register and Login. After login Data User have an option of searching the files as a file name. Data user can also have a download file it will show an encrypted data. Data user can also send a trapdoor request to the server. Server can accept the request and then data user can takes permissions from the owner then the file it will downloaded in plain text.

3.K-Manager Module:

In this module Data Owner should register and Login. Data Owner will Uploads the files into the database. Data owner can also send request to the data user.



4.Cloud Server Module:



In this module Cloud Server can login. After login it will see all data owners' information. Cloud server can see all users' information. Cloud server can see an all Stored data files. Cloud server can give keys request to the user. Cloud server can also see an attacker information of file.

© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. EXPERIMENTAL RESULTS

1. User Login:

2. Home Page:

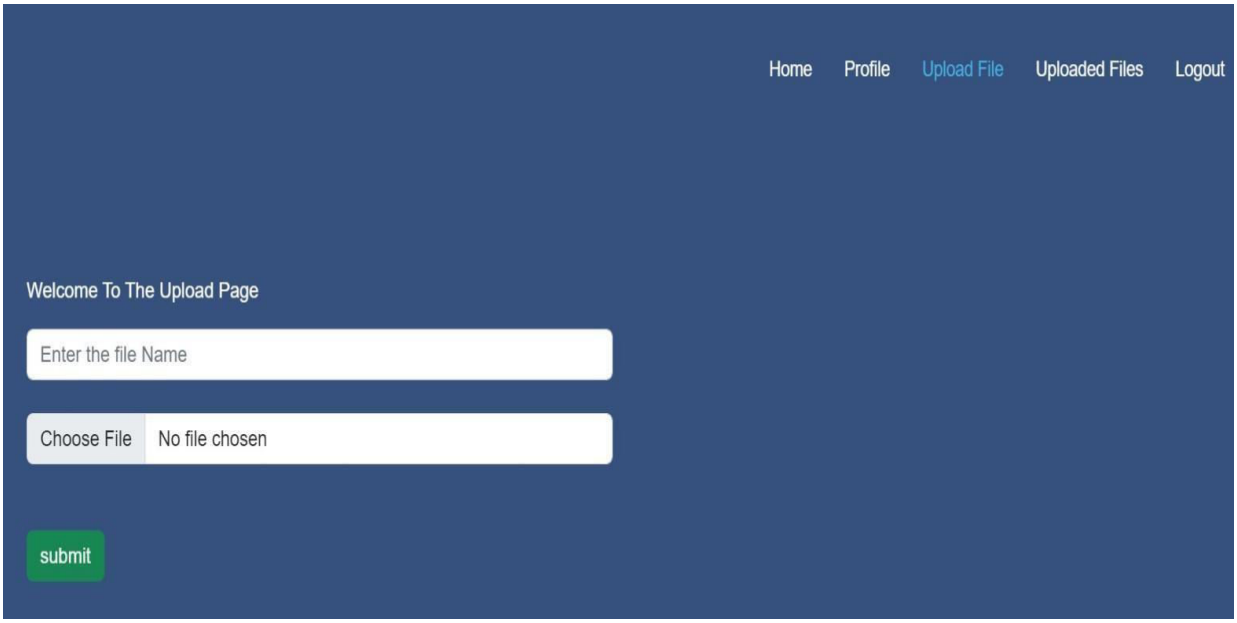
© 2024 IJMRSET | Volume 7, Issue 12, December 2024| DOI: 10.15680/IJMRSET.2024.0712036

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3. Upload File:



4. Uploaded Files:

V. CONCLUSION

This research paper addresses the challenge of deduplicating scrambled information and introduces a TCD-DBSCAN algorithm. We propose the concept of bunching deviation and incorporate our algorithm into the deduplication process to mitigate the risk of internal information leakage. Additionally, untimely modification of disagreeable information is prevented, even when transferred by users from the same organization. For disagreeable data, symmetric encryption is employed, with the encryption key obtained through a blind signature protocol between the key manager and users. This approach enables deduplication of disagreeable data without requiring users to transmit their keys online, thereby

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

enhancing deduplication efficiency. Security analysis and performance evaluation demonstrate the proposed scheme's robustness and significant practical value.

VI. FUTURE ENHANCEMENT

In the realm of hierarchical management, as the technology proposed in this paper gains broader acceptance, various data processing modules associated with supplier product strategies, pricing, energy consumption, and other factors can be integrated to further enhance the digital aviation supply quality management system. Introduce machine learning for adaptive cluster deviation. The existing scheme likely employs a predefined threshold for allowable deviation. Machine learning can analyze historical deduplication decisions and dynamically adjust the deviation threshold based on data attributes, thus potentially enhancing accuracy by enabling tighter clustering for highly similar data and looser clustering for data with more acceptable variations.

Additionally, consider a hybrid approach: Combine DBSCAN with other techniques such as hashing to expedite the deduplication process for highly similar data. Implement adaptive deviation: Dynamically adjust the permissible deviation based on data characteristics or user preferences. Incorporate semantics: Integrate domain-specific knowledge to enhance cluster quality and deduplication accuracy. Develop incremental deduplication methods: Efficiently manage new data streams and update existing clusters.

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.
[11] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, A Secure Data Deduplication Scheme for Cloud Storage. Berlin, Germany: Springer, 2013.
[12] P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE 5th Int. Conf. Cloud Comput. Technol. Sci., Dec. 2013, pp. 363–370.
[13] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Server-aided encryption for deduplicated storage," in Proc. Usenix Conf. Secur., 2013, pp. 1–17.
[14] S. G. Zhang, H. Q. Xian, Y. Z. Wang, H. Y. Liu, and R. T. Hou, "Secure encrypted data deduplication method based on offline key distribution," J. Softw., vol. 29, no. 7, pp. 1909– 1921, 2018.
Etc.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com