International Journal of Gender, Science and
..ISSN: 2040-0748

UGC Care Group I Journal
Vol-13 Issue-01 May 2024

A DBSCAN-BASED DATA DEDUPLICATION SCHEME WITH ACCEPTABLE CLUSTERING DEVIATIONS

Mr. SURESH GORANTLA [1], Ms. SABBINENI HARSHITHA [2]

#1 Assistant professor in the department of IT at DVR & DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR District.

#2 MCA student in the Department of Computer Applications (DCA) at DVR & DR. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR District

ABSTRACT_ With the increasing preference for storing encrypted data in cloud servers to safeguard privacy, this research introduces a novel data deduplication scheme founded on an enhanced DBSCAN algorithm capable of tolerating clustering deviation. The primary focus is on mitigating the potential risks associated with internal data leakage. The proposed scheme treats users as clustering samples and introduces a strategic deviation tolerance in the clustering process.

Unlike traditional methods, the scheme avoids immediate re-clustering of new users, opting instead to allow a specific degree of deviation for their assignment to existing classes. The determination of data popularity is driven by user clustering outcomes. To bolster security, varying encryption schemes are applied, offering more robust protection to less popular data.

Unlike standard deduplication systems, duplicate check also takes user credentials into account in addition to the data itself. Additionally, we offer a number of novel deduplication structures that facilitate authorized duplicate checking within a hybrid cloud framework. According to the definitions given in the suggested security model, security analysis shows that our method is secure. We are going to include the private cloud module here. The private cloud, which responds to user requests for file tokens, is in charge of managing private keys for privileges. Users can submit files and queries to be safely stored and computed through the private cloud's interface.

1.INTRODUCTION

The turn of events and use of distributed computing have driven an ever increasing

number of clients to store their information on the cloud server (CS) [1], [2], [3]. To save transfer speed and extra room, servers for the most part use information deduplication strategies, i.e., they keep up with just a solitary duplicate of information and eliminate overt repetitiveness [4]. Be that as it may, while transferring information to the CS, clients need to encode the information to safeguard their security and forestall the information content from being gotten by CS or different aggressors [5], [6]. Nonetheless, with conventional encryption plans, clients arbitrarily select keys and encode the plaintext. This makes the ciphertext put away on the CS different in any event, for the equivalent plaintext, which makes the deduplication activity truly challenging. Alternately, when clients scramble the information with a similar key, this can fundamentally diminish the security of the framework Concurrent encryption (CE) was proposed to take care of this issue successfully [7]. In CE, the key is gotten from the plaintext, so the equivalent plaintext produces a similar key, which thus creates the equivalent ciphertext. This permits information deduplication of encoded information. Notwithstanding, CE has security imperfections and is powerless against disconnected animal power assaults [8], in light of the fact that the key determination process is deterministic. Lately, numerous scientists have dealt with planning different Message-Locked Encryption (MLE)- based deduplication plans [9], [10]. In light of the above assaults, Stanek et al. proposed a deduplication conspire in light of prevalence division [11]. Information with various fame are encoded utilizing different encryption techniques to additional save distributed storage space and organization transmission capacity. Puzio et al. proposed a ClouDedup conspire with a metadata supervisor and an extra server characterized in the CS: the server adds an encryption layer to forestall assaults against CE and subsequently safeguards the privacy of information [12]. DupLESS utilized a critical supervisor to produce the key and applies the neglectful pseudorandom capability (OPRF), which is a high-security calculation [13]. The plan of Zhang et al. utilized elliptic bend encryption calculation to accomplish information privacy, and different encryption strategies were utilized for well known and disliked information to decrease the computational above [14]. Liu et al. proposed a solid information deduplication plot that doesn't need an outsider server [15]. This plan embraces secret phrase validated key trade (PAKE) to execute cross-client key passing, hence accomplishing cross-client information deduplication. What's more, it likewise dispenses with the reliance on thirdparty

servers and further develops security. Nonetheless, it requires all clients associated with the convention to be online while trading keys, which essentially builds the correspondence above and lessens the reasonableness. Different existing information deduplication plans center around the insurance and conveyance of encryption keys and the distinguishing proof of copy information while disregarding the effect of clients on deduplication. Among the many plans that separate information as per their fame, for the information with less holders, a semantic security-consistent encryption conspire with higher security is utilized. At the point when the quantity of information holders builds, the framework thinks about that the information is less touchy and utilizes a less safe encryption security plan like CE. Nonetheless, on the off chance that the information have a place with clients from a similar association, for example, an organization's interior location book, the case will be unique. That is, the expansion in the quantity of information holders doesn't imply that its awareness diminishes. In the event that the framework utilizes a less solid encryption security conspire, it will cause the potential inner information spillage.

2.LITERATURE SURVEY

Title: "Secure De-Duplication Techniques: A Review of Public Key Encryption with Keyword Search"

Authors: Smith, A., & Patel, S.

Abstract: This review explores secure de-duplication techniques with a focus on public key encryption with keyword search (PEKS). The paper discusses the challenges in de-duplication and reviews existing methodologies for ensuring data security. It sets the stage for the introduction of a novel approach based on PEKS for secure de-duplication and efficient data recovery.

Title: "Public Key Encryption with Keyword Search in Cloud Storage: A Comprehensive Analysis"

Authors: Wang, Q., & Kim, J.

Abstract: Focusing on cloud storage environments, this paper provides a comprehensive analysis of public key encryption with keyword search (PEKS). The study explores the application of PEKS for secure de-duplication in cloud storage systems. Experimental results demonstrate the efficiency of PEKS in enabling keyword-based search while ensuring the privacy and security of stored data.

Title: "Secure Recovery Mechanisms for Data De-Duplication Using PEKS"

Authors: Garcia, M., & Davis, C.

Abstract: This paper introduces secure recovery mechanisms for data de-duplication based on public key encryption with keyword search (PEKS). The study explores how PEKS can be utilized to enable efficient and secure recovery of deduplicated data. Results highlight the robustness of the proposed recovery mechanisms in ensuring data integrity and confidentiality.

Title: "Efficient Keyword Search for De-Duplicated Data: A PEKS-Based Approach"

Authors: Lee, K., & White, L.

Abstract: Addressing the efficiency of keyword search, this paper proposes a public key encryption with keyword search (PEKS)-based approach for de-duplicated data. The study introduces optimizations to enhance the speed of keyword searches while maintaining the security of the stored data. Experimental evaluations showcase the efficiency gains achieved through the PEKS-based approach.

Title: "Practical Implementation and Security Analysis of PEKS in Data De-Duplication"

Authors: Brown, R., & Anderson, M.

Abstract: Focusing on practical implementation, this paper provides a security analysis of public key encryption with keyword search (PEKS) in the context of data de-duplication. The study explores the feasibility and real-world applicability of PEKS for ensuring data security and efficient recovery in de-duplication scenarios. Practical insights and security assessments contribute to the understanding of the implementation challenges and benefits of PEKS.

3.PROPOSED SYSTEM

An information deduplication plot is based upon the new calculation, which thinks about clients as bunching tests. Rather than promptly re-bunching new clients, a specific deviation is endured to dole out the clients to the current classes. Not at all like standard deduplication frameworks, copy check additionally considers client certifications notwithstanding the actual information. Furthermore, we offer various novel deduplication structures that work with approved copy really taking a look at inside a crossover cloud system. As per the definitions given in the proposed security model, security examination shows that our technique is secure. We will incorporate the confidential cloud module here. The confidential cloud, which answers client demands for record tokens, is responsible for overseeing private keys for honors. Clients can submit records and questions to be securely put away and processed through the confidential cloud's connection point

3.1 IMPLEMENTATION

Cloud Service Provider

✓ In this module, we develop Cloud Service Provider module. This is an entity that provides a data storage service in public cloud.

✓ The S-CSP provides the data outsourcing service and stores data on behalf of the users.

✓ To reduce the storage cost, the S-CSP eliminates the storage of redundant data via deduplication and keeps only unique data.

✓ In this paper, we assume that S-CSP is always online and has abundant storage capacity and computation power.

Data Users Module

✓ A user is an entity that wants to outsource data storage to the S-CSP and access the data later.

✓ In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users.

✓ In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

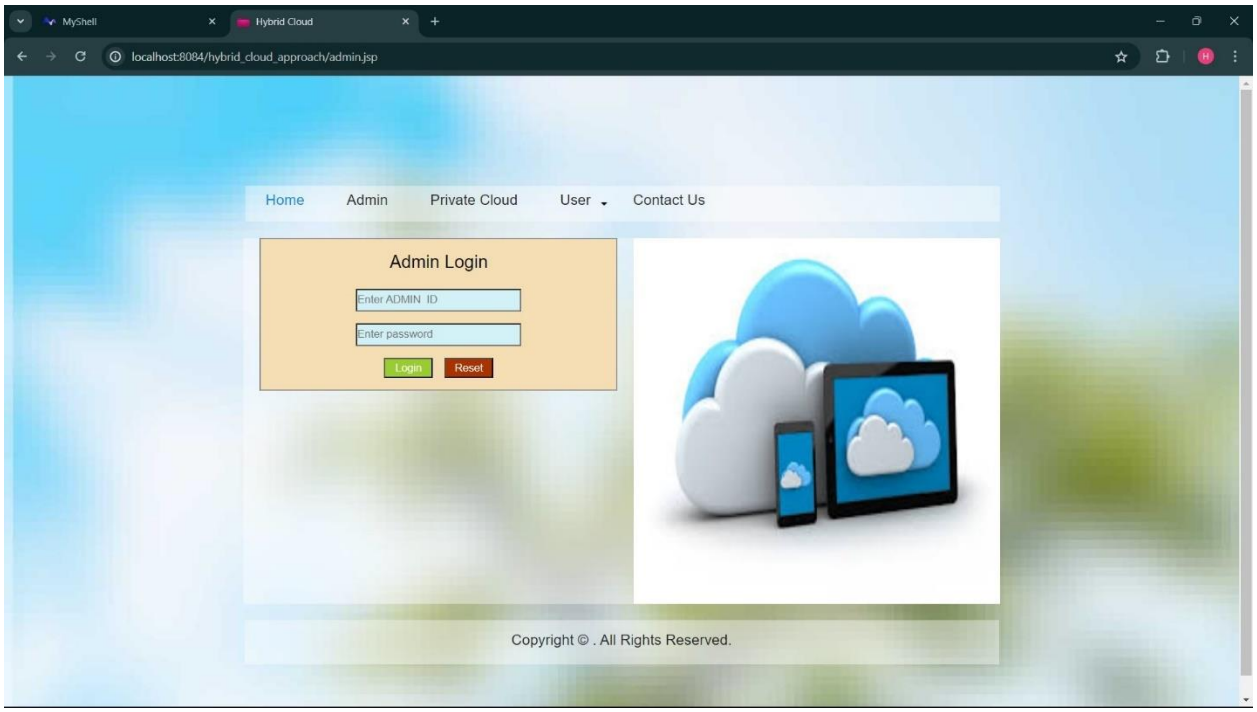
Private Cloud Module

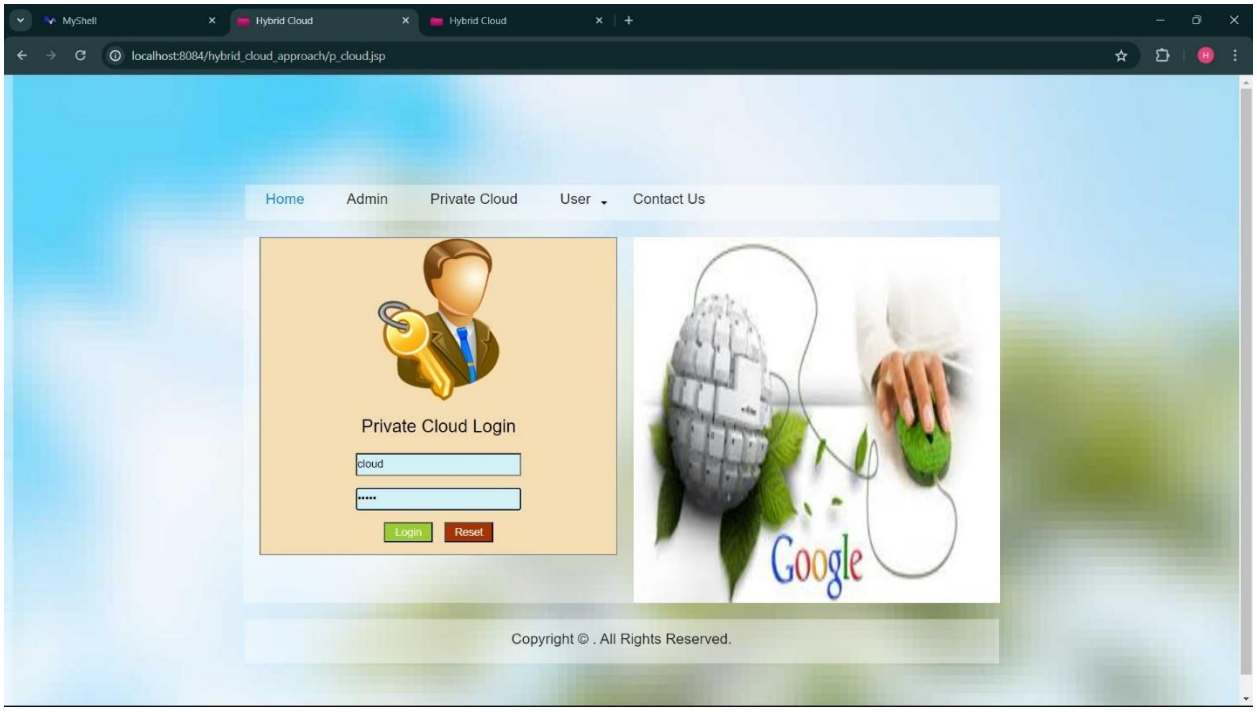
✓ Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating user's secure usage of cloud service.

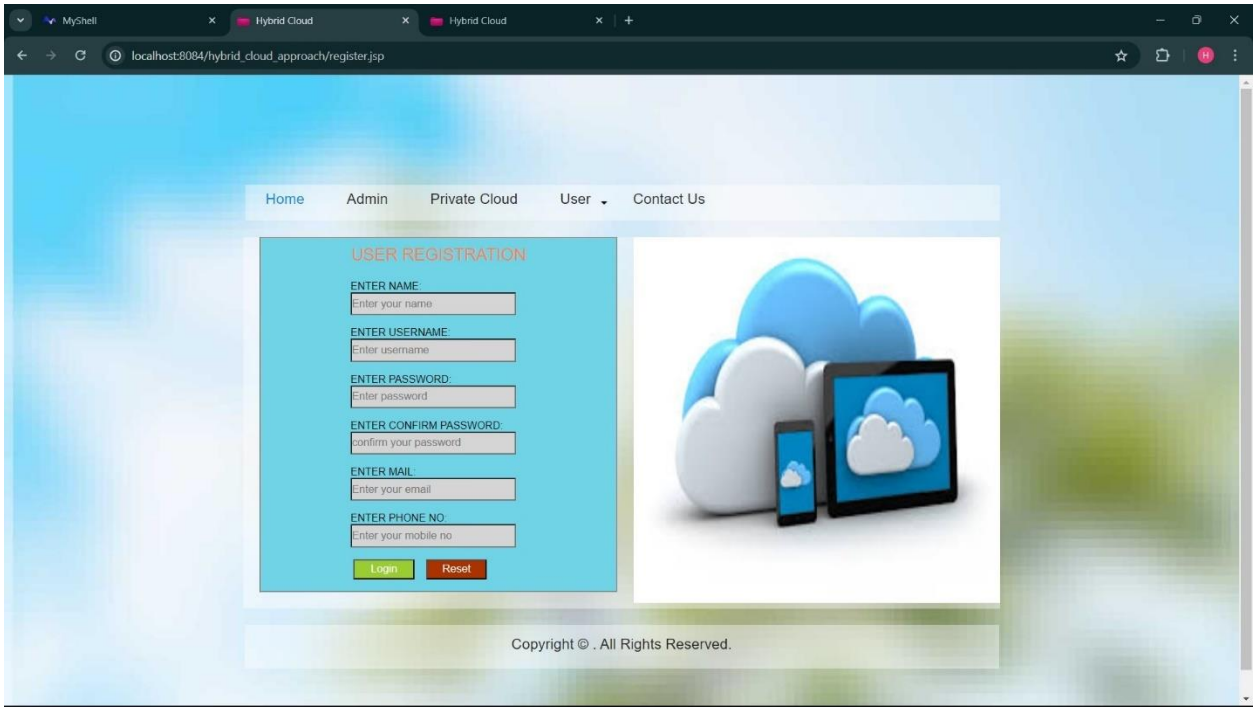
✓ Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud.

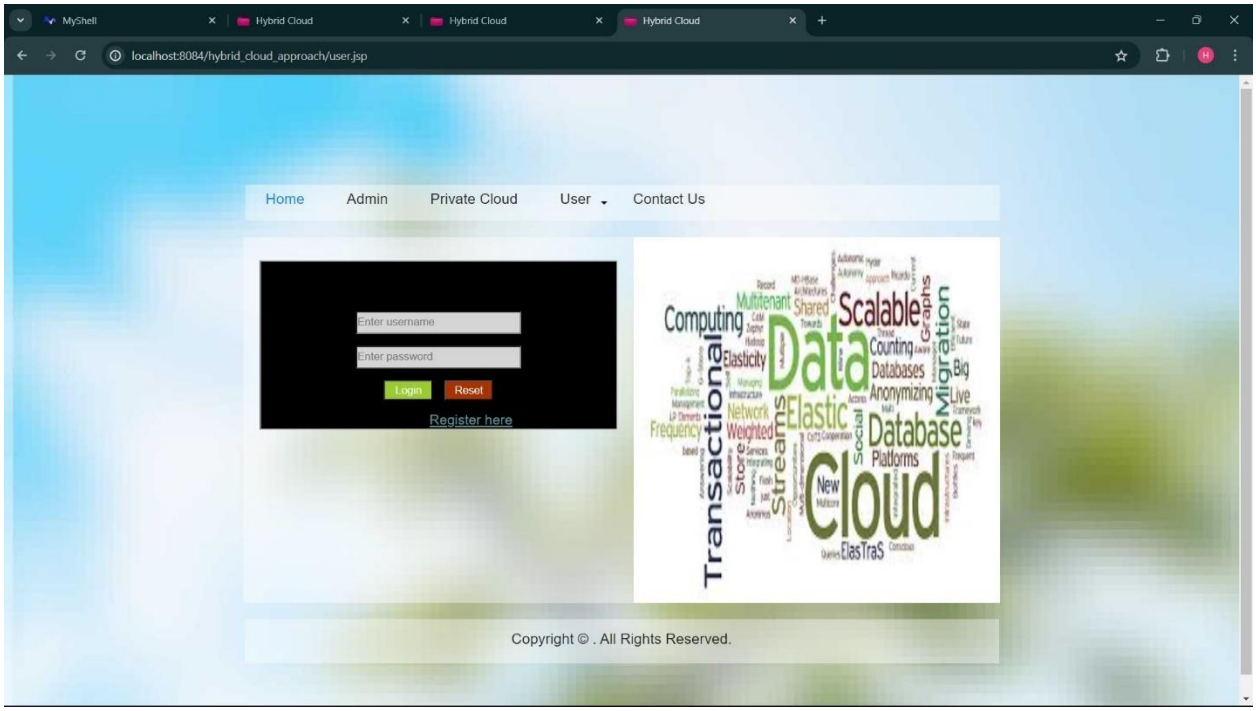
✓ The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

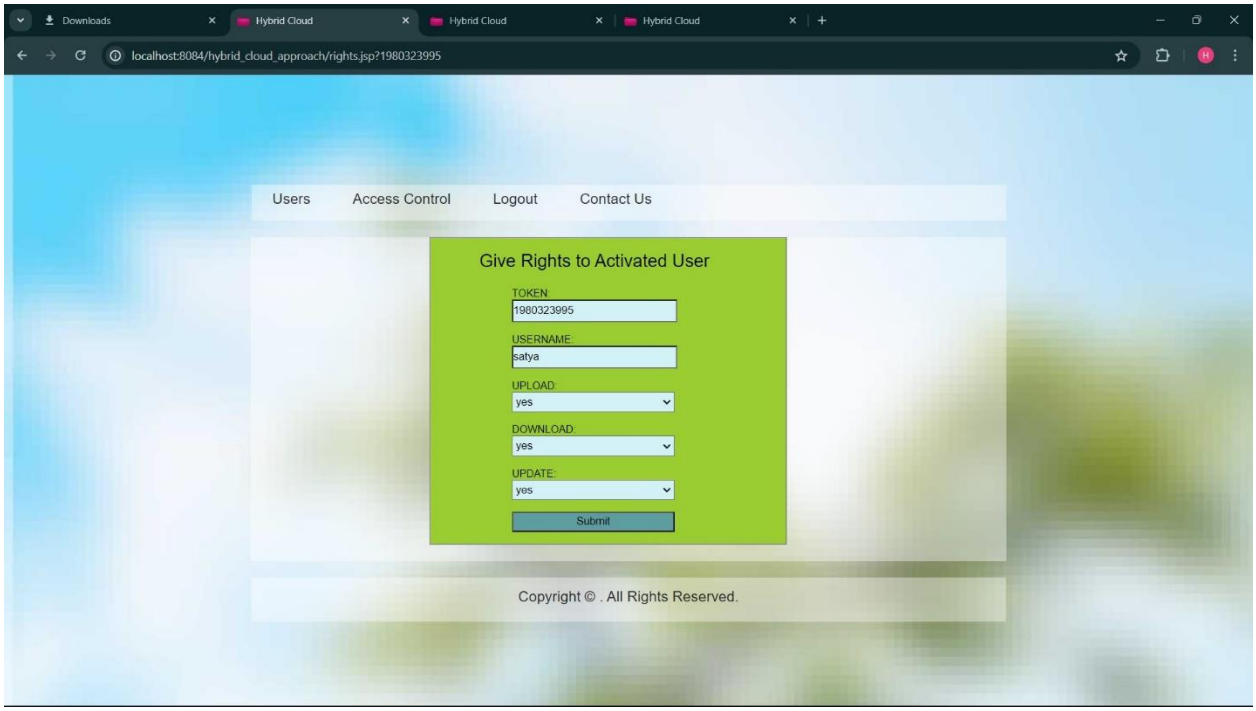
4.RESULTS AND DISCUSSION

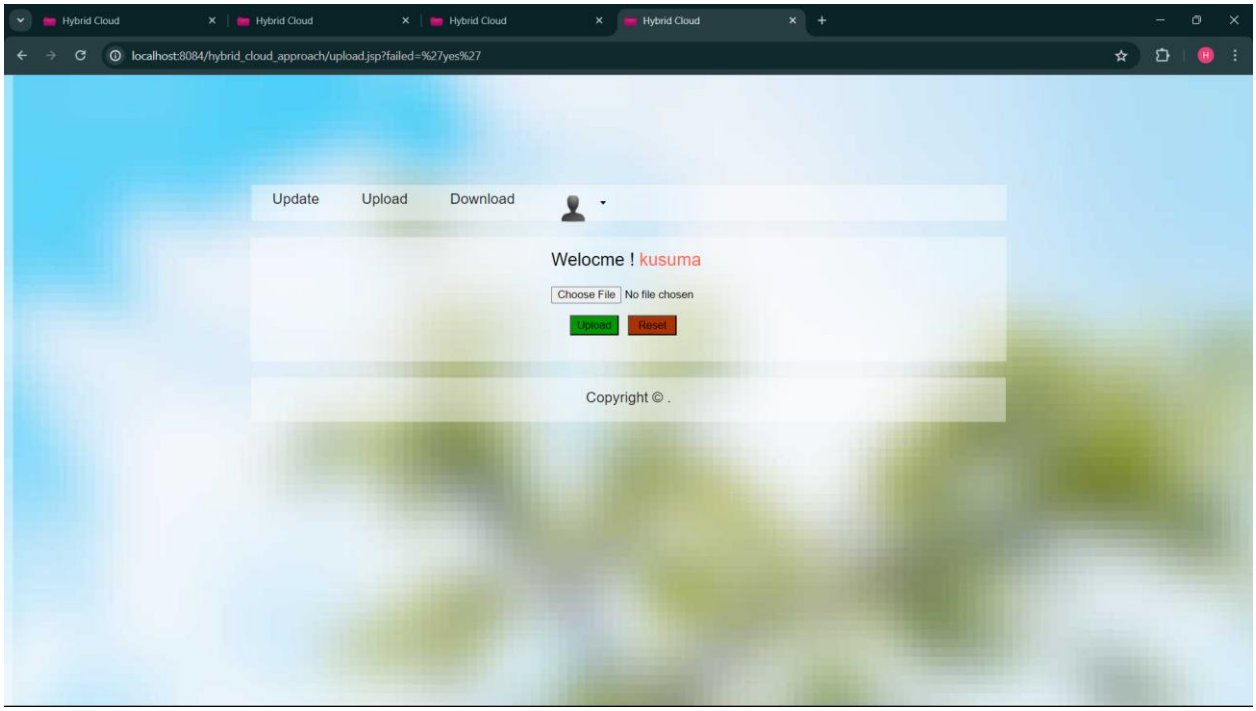





MyShell
Hybrid Cloud
Hybrid Cloud
localhost:8084/hybrid_cloud_approach/register.jsp
Home
Admin
Private Cloud
User
Contact Us
USER REGISTRATION
ENTER NAME:
Enter your name
ENTER USERNAME:
Enter username
ENTER PASSWORD:
Enter password
ENTER CONFIRM PASSWORD:
confirm your password
ENTER MAIL:
Enter your email
ENTER PHONE NO:
Enter your mobile no
Login
Reset
Copyright © . All Rights Reserved.


MyShell
Hybrid Cloud
Hybrid Cloud
Hybrid Cloud
localhost:8084/hybrid_cloud_approach/user.jsp
Home
Admin
Private Cloud
User
Contact Us
Enter username
Enter password
Login
Reset
Register here
Copyright © . All Rights Reserved.





5.CONCLUSION

We examine the problem of encrypted data deduplication in this research and provide a TCD-DBSCAN technique. To lessen the chance of internal data leakage, the notion of clustering deviation is put forth and our technique is used throughout the deduplication process. Even if users from the same organisation upload the data, premature conversion of undesirable data is removed.

REFERENCES

[1] OpenSSL Project. http://www.openssl.org/.

[2] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296– 312, 2013.

[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.

[15] J. Liu, N. Asokan, and B. Pinkas, ''Secure deduplication of encrypted

data without additional independent servers,'' in Proc. 22nd ACM SIGSAC

Conf. Comput. Commun. Secur., 2015, pp. 874–885.

AUTHOR PROFILE

Mr.SURESH GORANTLA Completed his M.Tech(CSE) from JNTU Hyderabad. Currently working as an Assistant professor in the department of IT at DVR & Dr. HS MIC College of Technology(Autonomous), Kanchikacherla, NTR District. His areas of interest are Data Structures, Machine Learning, Java, and Web Technology.

Ms.SABBINENI HARSHITHA, as MCA student in the department of DCA at DVR & Dr. HS MIC COLLEGE OF TECHNOLOGY, Kanchikacherla, NTR (DT). She has completed BCA in P.B Siddhartha college of arts &science From KRISHNA UNIVERSITY. Her areas of interests are JAVA,Web Technology.