International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

A HYBRID CLOUD APPROACH FOR SECURE AUTHORIZED DEDUPLICATION

Dr.B.MeenaPreethi[1], Assistant Professor

Aisvarya.S[2], Student,

Department of Software Systems,

Sri Krishna Arts and Science College,

Coimbatore

Abstract: *Data deduplication is a critical technique employed to optimize storage utilization and minimize redundancy in cloud environments proposes a novel hybrid cloud approach that combines the advantages of both private and public clouds to achieve secure and authorized data deduplication. This hybrid cloud approach introduces an innovative method for managing the deduplication index. Deduplication is to eliminate redundant data copies, has gained significant attention due to its potential to reduce storage requirements and enhance data transfer efficiency also it maintain data confidentiality, the proposed approach employs encryption technique convergent encryption, ensuring that data remains encrypted even during the deduplication process*.

Keyword: Data confidentiality, convergent encryption, deduplication, hybrid cloud

1.INTRODUCTION

Cloud computing offers scalable and cost-effective solutions for data storage and processing. Data deduplication has emerged as a promising solution by identifying and eliminating duplicate data, thereby reducing storage overhead and enhancing data transfer efficiency. By the strengths of private and public clouds, the HC-SAD framework aims to provide organizations with a versatile and secure solution for data deduplication in cloud environments. HC-SAD aims to offer organizations a pragmatic and effective strategy for managing data duplication, security, and storage demands in today's dynamic digital landscape. Traditional data deduplication methods,often neglect the security aspects necessary to safeguard sensitive information. Encrypting data before deduplication is one approach, but it raises concerns about the efficiency of deduplication algorithms when operating on encrypted data. To address these challenges, we propose a novel Hybrid Cloud Approach for Secure Authorized Deduplication. HC-SAD not

only enhances data confidentiality but also ensures that authorized users can efficiently access and manage their data. The purpose of the Hybrid Cloud Approach for Secure Authorized Deduplication is to provide organizations with a comprehensive solution that benefits of data deduplication with stringent security measures, while accommodating the diverse data needs of modern cloud computing. Through this purpose, HC-SAD strives to address the challenges associated with data duplication, privacy, and security, ultimately enhancing data management practices in hybrid cloud environments.

2.EXISTING METHODOLOGY

In data deduplication systems, the private cloud functions as an intermediary, facilitating secure duplicate checks for data owners and users with distinct access privileges. This architectural approach is both feasible and has garnered significant interest within the research community. Under this framework, data owners choose to delegate data storage to the public cloud, while retaining data management within a private cloud environment.

DRAWBACKS:

Cloud storage services face a significant obstacle in handling the continuously growing data volume.

While traditional encryption ensures data confidentiality, it clashes with the concept of data deduplication.

Instances of identical data copies owned by different users will result in distinct ciphertexts, rendering deduplication unfeasible.

2.1 PROPOSED METHODOLOGY

This protocol furnishes evidence that a user indeed possesses the identical file upon finding a duplicate. Subsequently, users with the same file are supplied a server pointer eliminating the need to upload the same file repeatedly. Traditional encryption which ensures data confidentiality, presents a compatibility issue when combined with data deduplication. This is attributed to the fact that traditional encryption mandates users to encrypt their data using individual keys. Consequently, identical data copies owned by different users yield distinct ciphertexts, rendering deduplication unattainable. A secure key management system is established to generate, distribute, and manage encryption keys. This ensures that data remains confidential and that authorized users can access the data when needed. The system is fortified with comprehensive security measures, including encryption, access controls, authentication, and secure communication protocols, to prevent unauthorized access and data breaches.

ADVANTAGES:

Duplicate checks are only accessible to authorized individuals.

The process contributes to data privacy by eliminating redundant copies of recurring data, making it a prevalent technique in cloud storage for optimizing storage space and conserving bandwidth.

The reduction of tag storage size is employed to bolster deduplication

security while safeguarding data confidentiality.

3.METHODOLOGY

The basic objective of this work is the problem of privacy preserving deduplication in cloud computing and a proposed System focus on these aspects.

[1].Encryption and Access Control Design: Select suitable encryption techniques, such as homomorphic encryption and convergent encryption, for different data categories. Determine the encryption keys and access control policies for each class of data. Develop mechanisms for secure data sharing and access control within and between the private and public clouds.

[2].Data Deduplication Implementation: Implement the data deduplication mechanisms specific to the hybrid cloud environment. Incorporate deduplication algorithms that work seamlessly with the chosen encryption techniques. Ensure that deduplication can be performed while maintaining data confidentiality and without compromising security.

[3].Authorized Duplicate Check: Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

[4].Access Control Check: Verify the user's credentials and privileges to ensure they have the authorization to perform the upload and duplicate check for the specific file.[5].Secure Upload to Private Cloud: If no duplicates are found, upload the encrypted data chunks to the user's private cloud. Utilize secure communication protocols (e.g., HTTPS) to ensure the data integrity transmission.

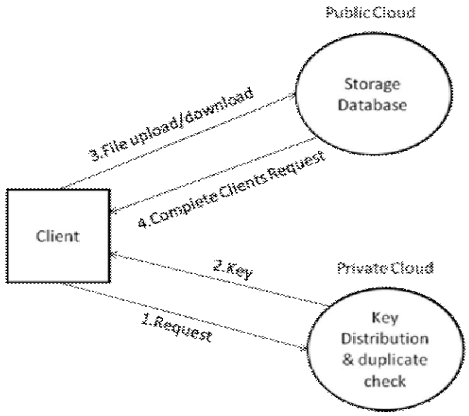

Fig1:Client architecture

[6].Secure Communication to Public Cloud: If the public cloud is used for storage, establish a secure connection and transmit the metadata about the uploaded file to the public cloud's management layer.

The diagram below illustrates the two keys involved: the token key and the user's password key.Demonstrate the implementation of privacy preservation within an authorized deduplication system. Utilize convergent encryption techniques to encrypt files, selecting a token generated from the file as the encryption key. This token is then transmitted to the administrator via the public network for duplicate checks, ensuring security. To safeguard the token during transmission over the public network, it is encrypted using the user's password.

International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

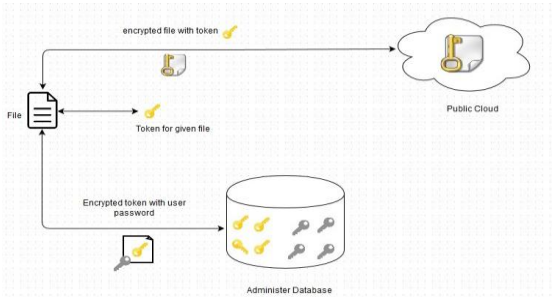

Fig2:File encryption with keys

4.HYBRID CLOUD FOR SECURE DEDUPLICATION

A hybrid cloud approach for secure deduplication combines the advantages of both private and public cloud environments to optimize data storage, sharing, and management while ensuring robust security measures. This approach addresses challenges related to data redundancy, confidentiality, and authorized access in a distributed and dynamic data landscape. In modern data-intensive landscapes, a hybrid cloud approach for secure deduplication offers a strategic solution that merges the strengths of private and public cloud infrastructures. This approach tackles challenges related to data redundancy, confidentiality, and authorized access, providing an efficient and secure framework for data management. The hybrid cloud model involves integrating a private cloud, which serves as a data operation manager, with a public cloud, which offers scalable storage solutions. This architecture ensures a balance between data security and cost-effectiveness. The system accommodates changes in data over time while maintaining security and deduplication integrity. Comprehensive metadata management is essential for tracking data chunks, encryption keys, access permissions. Metadata supports efficient deduplication and authorized access. The public cloud complements the private cloud by offering scalable and cost-effective data storage solutions. It caters to the storage needs of deduplicated data chunks and other non-sensitive data. The private cloud serves as a controlled data operation It manages sensitive tasks such as deduplication, encryption, and access controls. Sensitive data and critical operations are conducted within this secure enclave.

5.MESSAGE DIGEST ALGORITHM 5

The MD5 (Message Digest Algorithm 5) is a widely used cryptographic hash function. Divide the files into smaller data chunks, making it easier to handle and deduplicate. For each data chunk, calculate its MD5 hash value using the MD5 algorithm. The MD5 hash will act as a unique identifier for the data chunk. Before uploading a data chunk, check if its MD5 hash value already exists in your storage system. If it does, identified a duplicate data chunk. If the data chunk is not a duplicate, upload it to the appropriate storage environment (private or public cloud). It Maintain metadata that includes the MD5 hash value, encryption details, access controls, and other relevant information. This metadata facilitates efficient deduplication and authorized access. Implement a secure protocol if necessary, to establish file ownership and ensure authorized access to data chunks. For public cloud storage, you can generate pointers to uploaded data chunks to avoid redundancy.

6.COMPARISON OF ALGORITHM

AES (Advanced Encryption Standard), DES (Data Encryption Standard) and MD5 (Message Digest Algorithm 5) are three cryptographic algorithms that serve different purposes and have different characteristics. AES (Advanced Encryption Standard) typically refers to the process of removing duplicate data before encrypting it with AES. This can help reduce storage and processing overhead when dealing with large datasets that contain duplicate information. Here are the steps to calculate deduplication for the AES algorithm:

Data Source: You need a source of data that may contain duplicates, such as files, database records, or backup data.

Create a Deduplication Index: Maintain an index or database that stores unique data items and their corresponding keys . This index will help you keep track of which data items have been deduplicated.

Data Chunking: Divide your data into fixed-size or variable-size chunks. Common chunk sizes are 4KB, 8KB, or more, depending on your deduplication system.

Update Deduplication Index: Whenever you encounter a new data item, calculate its hash and check against the deduplication index. If the item is already in the index, use the existing reference/key for that item. If it's a new unique item, add it to the index and store the reference/key.

Hashing: Calculate the hash value for each data chunk.

Deduplication Process: Compare the hash values of data chunks to identify duplicates. When a duplicate is found, you can replace it with a reference to the original data chunk (deduplication) or simply discard it.

Storage Optimization: Store unique data chunks and maintain a mapping or index to reconstruct the original data when needed.

Data Encryption Standard (DES) is not typically used for deduplication directly, as DES is primarily an encryption algorithm, and deduplication is a separate data processing operation. Instead, you would perform deduplication on the plaintext data before encrypting it with DES or any other encryption algorithm. Below are the general steps for deduplicating data before encrypting it using DES:

Identify the Data Set: Determine the dataset or collection of data that you want to deduplicate. This could be files, database records, or any other type of data.

Choose a Deduplication Method:

1.Content-BasedDeduplication: Calculate a unique hash value for each piece of data and compare these hashes to identify duplicates. If two pieces of data have the same hash, they are considered duplicates.

2.File-Level Deduplication: If you are dealing with files, perform deduplication at the file level by comparing files based on their file size, filename, and potentially a hash of the file's content.

3.Block-Level Deduplication: For larger datasets, perform deduplication at a more granular level by dividing data into fixed-size or variable-size blocks and comparing blocks for duplication.

Calculate Hashes or Identifiers: If using content-based deduplication, calculate hash values for each piece of data in your dataset.If you're using other methods, generate unique identifiers or fingerprints for each data item.

Identify Duplicates: Compare the calculated hashes or identifiers to identify duplicate data. Duplicates are data items that have the same hash value or identifier.

Remove Duplicates: Depending on your specific requirements, you can either remove the duplicate data entirely or mark it for deletion. Be cautious when removing data, as it should align with your data retention policies.

Encrypt the Data with DES: Once deduplication is complete, you can proceed to encrypt the data using the DES algorithm. DES is a symmetric encryption algorithm, so you'll need a secret key to encrypt and decrypt the data. Ensure that you keep the encryption key secure.

Secure Key Management: Properly manage and secure the encryption keys used for DES. Use strong key management practices to protect the keys from unauthorized access.

Store or Transmit the Encrypted Data: The deduplicated and encrypted data can be safely stored or transmitted, knowing that duplicates have been removed, and the data is protected.

MD5 (Message Digest Algorithm 5) is a cryptographic hash function, and it is typically used for deduplication purposes and is designed to produce a fixed-size (128-bit) hash value that represents the input data uniquely. In other words, different inputs should produce different MD5 hashes, making it suitable for identifying duplicates.To deduplicate data, you would typically use methods that compare the actual content of the data rather than generating hash values like MD5. Here are the steps for content-based deduplication:

Identify the Data Set: Determine the dataset or collection of data that you want to deduplicate. This could be a set of files, database records, or any other type of data.

Choose a Deduplication Method:

1.Content-Based Deduplication: In this method, you compare the actual content of the data to identify duplicates.

Identify Duplicates: For each piece of data in your dataset, compare it to other pieces of data to identify duplicates. Duplicates are data items that have the same content.

Remove Duplicates: Depending on your specific requirements, you can either remove the duplicate data entirely or mark it for deletion. Be careful when removing data, as it should align with your data retention policies.

Calculate MD5 hashes for data, here are the steps:

Identify the Data Set: Determine the dataset or collection of data for which you want to calculate MD5 hashes.

Iterate Through the Data: For each piece of data in your dataset, apply the MD5 hashing algorithm.

Calculate the MD5 Hash: Use an MD5 hashing library or function to calculate the MD5 hash of the data. The resulting

hash will be a fixed-length hexadecimal string.

Store or Use the MD5 Hashes: store or use the MD5 hashes for various purposes, such as data integrity verification. MD5 hashes can be useful for checking whether the data has been tampered with.

It's important to note that MD5 is considered for cryptographic purposes that allow for collision attacks (two different inputs producing the same MD5 hash). Therefore, it should be used for security-critical applications.

By comparing algorithm I conclude that md5 algorithm is best for securing the data and it is suitable for data deduplication.

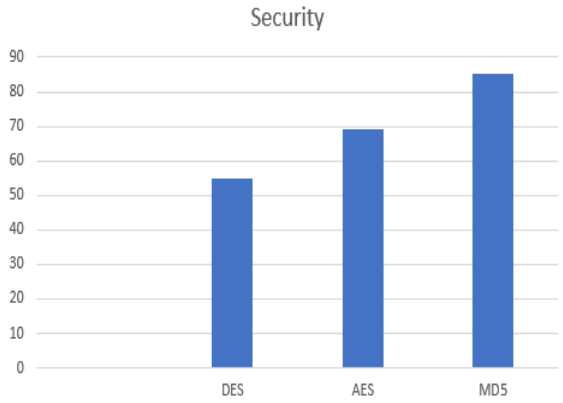

Fig3:Security Analysis

7.FLOW DIAGRAM

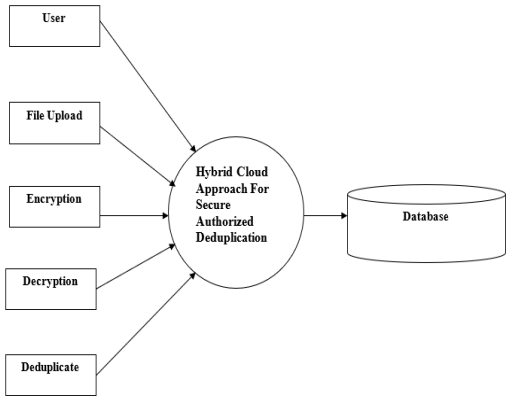

Fig4:Proposed framework

8.RESULTS

The deduplication process effectively eliminates duplicate data chunks, optimizing storage utilization across both private and public cloud environments. The hybrid architecture ensures that data operations are managed optimally, minimizing resource wastage and enhancing system performance. The integrated security measures decrease vulnerabilities, minimizing the potential for data breaches and unauthorized access. Deduplication results in reduced storage requirements, leading to cost savings for both private and public cloud storage. Deduplication leads to a significant reduction in data redundancy, optimizing storage resources across both private and public cloud domains. Retrieval of locally stored duplicate data chunks enhances data access speed, minimizing latency and improving overall user experience. Reduced data transfer time contributes to lower latency, further improving user satisfaction. Efficient resource scaling in the hybrid cloud architecture optimizes cost-effectiveness based on real-time demands. By harnessing the capabilities of private and public clouds, this approach not only streamlines data operations but also reinforces data security, offering a flexible, resilient, and secure data ecosystem for organizations to thrive in an ever-evolving digital landscape. The results affirm that the hybrid cloud approach

for secure authorized deduplication is a robust solution that efficiently addresses the complexities of modern data management.

The graph shows the results of three algorithm. By the results acquired it showing that the Md5 algorithm has received more accuracy compared to other two algorithm.

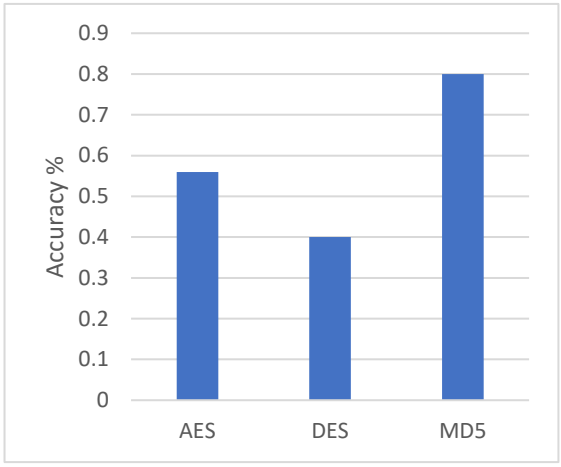

Fig5:Comparision of Deduplication Accuracy algorithms

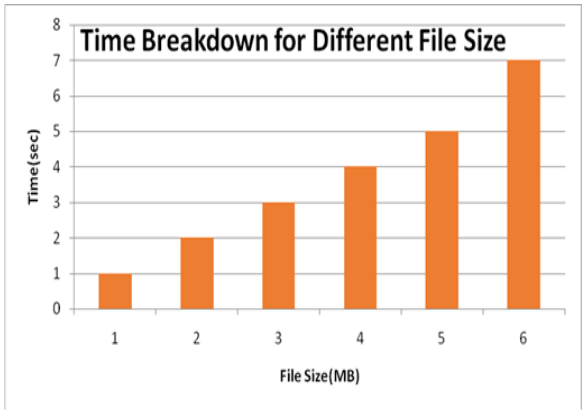

Fig6:Time breakdown for different file size

The above figure, evaluating the system by comparing overheads in various steps like token generation and file upload process. We are evaluating overhead by varying factors like 1) File Size 2) Number of Stored Files. A. File Size In order to evaluate the effect of file size, we upload 10 files of different size and we are recording the time for file break down.

9.CONCLUSION

A hybrid cloud approach for secure deduplication represents a powerful solution that combines the strengths of private and public cloud environments to address the challenges of data redundancy, confidentiality, and authorized access. By integrating these two cloud models, organizations can achieve optimal data storage, sharing, and management while ensuring robust security measure. Hybrid cloud model offers several key benefits. It enables efficient utilization of storage resources through advanced deduplication techniques, minimizing data redundancy and saving costs. By effectively segmenting data into smaller chunks, applying hash functions for unique identification, and employing convergent encryption for secure storage, the hybrid cloud approach streamlines data operations. The implementation of proof of ownership protocols adds an extra layer of security, allowing for controlled access and preventing unauthorized duplication. The system is designed with user-friendliness in mind, ensuring that individuals of varying technical backgrounds can easily navigate it. In a data-driven world where the volume of information continues to grow exponentially, a hybrid cloud approach for deduplication stands as a strategic cornerstone for efficient data management, optimal resource utilization, and strengthened security.

10.FUTURE WORK

Future work of hybrid cloud approach for secure authorized deduplication lays a strong foundation for efficient data

management and enhanced security. Enhanced Security Protocols that Develop and integrate advanced security protocols to fortify the authorization and authentication mechanisms. This could involve exploring multi-factor authentication, biometric recognition, or even blockchain-based identity verification to ensure only authorized users can access and manage data. Dynamic Key Management is to Implement dynamic key management techniques that enable automatic rotation of encryption keys. This adds an extra layer of security by minimizing the window of vulnerability in case of a compromised key. Conduct thorough performance benchmarking and analysis to assess the system's efficiency, scalability, and responsiveness under various conditions. This can guide further optimization efforts that involves a commitment to ongoing innovation, security enhancement, and alignment with evolving technological trends. By addressing these aspects, the hybrid cloud approach can continue to evolve as a robust solution for efficient and secure data management in the ever-changing landscape of cloud computing.

11.REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart.Messagelocked encryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

[2] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC2011), 2011.

[3] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[4] D. Ferraiolo and R. Kuhn, “Role-based access controls, ” in Proc. 15th NIST-NCSC Nat. Comput. Security Conf., 1992, pp. 554–563.

[5] W. K. Ng, Y. Wen, and H. Zhu, “Private data deduplication protocolsin cloud storage,” in Proc. 27th Annu. ACM Symp. Appl. Comput.,2012, pp. 441–446.

[6] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure datadeduplication scheme for cloud storage. In Technical Report, 2013.

[7] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS,pages 195–206, 2013.

[8] J. Yuan and S. Yu. Secure and constant to cost public cloud storage auditing with the deduplication. IACR Cryptology ePrint Archive, 2013:149, 2013.

[9] M. Bellare and A. Palacio, “Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks,” in Proc. 22nd Annu. Int. Cryptol. Conf. Adv. Cryptol., 2002, pp. 162–177.

[10] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS,Apr 2013.

[11] P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de-duplication,” in Proc. 24th Int. Conf.

Large Installation Syst. Admin., 2010, pp. 29–40.

[12] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.