IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DEDUPLICATION WITH DOUBLE ENCRYPTION ON SECURE CLOUD DATA SHARING

Jeyamurugan A[1], AravinthRaj N S[2], ArulSelvar S[3]
[1]Assistant professor Senior Grade, [2] student, [3] student
Department of Computer Science and Engineering,
Paavai Engineering College (Autonomous)
Pachal, Namakkal, Tamil Nadu., India.

Abstract : Cloud data storage has gained immense popularity in recent years due to its scalability and accessibility. However, the efficient management of data in the cloud, especially with the ever-increasing volume of data, remains a significant challenge. Data deduplication is a crucial technique to reduce storage costs and enhance data management efficiency. It also concerns the secure cloud data storage and sharing. This research proposes a novel approach to deduplication-based cloud data storage using chunk-based partitioning. In this approach, data is divided into fixed-size chunks, and a deduplication mechanism is applied to identify and eliminate redundant chunks. This process leading to improved storage efficiency and reduced data redundancy. This approach not only conserves storage space but also enhances data retrieval performance by reducing the volume of data to be transferred over the network. Proposed system also enhance the security of cloud data storage by employing Base64 encoding for data representation and implementing re-encryption techniques based on the Advanced Encryption Standard (AES) for secure data sharing. The Base64 encoding methodology is employed to transform binary data into a text format, rendering it less susceptible to unintended manipulations during data transmission and storage. This encoding mechanism ensures data integrity and facilitates its secure storage in cloud environments. By utilizing AES encryption, data shared among users is re-encrypted with unique keys, ensuring that only authorized recipients can access and decrypt the shared information. This approach significantly enhances data confidentiality and mitigates risks associated with unauthorized access, thereby fortifying the security of cloud-based data sharing.

Index Terms – Cloud Service Provider(CSP),Key Generation Center (KGC),Private Information Retrieval(PIR)

INTRODUCTION

Cloud computing is consistently growing and there are many main cloud computing providers including Amazon, Google, Microsoft, Yahoo and many others who are offering solutions including Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), Storage-as-a- Service and Infrastructure-as-a-Service (IaaS). In addition, considering the possibility to substantially minimizing expenses by optimization and also maximizing operating as well as economic effectiveness, cloud computing is an excellent technology. Furthermore, cloud computing can tremendously boost its cooperation, speed, and also range, thus empowering a totally worldwide computing model on the internet infrastructure. On top of that, the cloud computing has advantages in delivering additional scalable, fault tolerant services.Furthermore, the project integrates a re-encryption mechanism to fortify data security during sharing. Leveraging this approach, data owners can re-encrypt chunks of data with new encryption keys before sharing them with authorized users. This provides an additional layer of protection, ensuring that only designated individuals with the appropriate decryption keys can access and utilize the shared information. The re-encryption process not only contributes to data security but also facilitates controlled and secure data sharing practices, aligning with contemporary standards of privacy and confidentiality in cloud computing.

OBJECTIVES

The primary objective of this project is to develop a comprehensive and secure framework for cloud data storage, addressing critical challenges related to redundancy, security, and controlled access. The key points of objectives are given below,

RESEARCH METHODOLOGY

[1] Wu, Yunyun In their study, the propose a novel scheme that supports the search with the user-specified number of keywords contained in the search result. This number n can be used to customize the keyword relevancy. As a result, the data owner could obtain the desired search results containing any n keywords of a keyword set. The proposed scheme also supports the traditional disjunctive and conjunctive keyword searches when n equals 1 or the size of the keyword set, respectively. The keyword could be positive or negative. The proposed scheme borrows the idea of the attribute based encryption by representing the attributes as the keywords and the access policy as the search expression to enhance the search. In order to apply the idea in the proposed scheme, we first solve the problem of the keyword security (responding to the attribute security), and then improve the method to obtain the search results which contain any n keywords of a keyword set, rather than to get the boolean search results that the current schemes could support. In addition, the proposed scheme allows the keywords to be positive or negative, which enhances the search flexibility. A positive keyword means it is contained in the document, while a negative keyword means it is not contained in the document. At last, we define a security model based on PERKS that can defend not only against the adaptive chosen keyword attack but also against the offline keyword guessing attack. Most of the existing PEKS schemes are under the offline keyword guessing attack. Since the adversary can generate the encrypted index for the keywords, it may determine the relationship between keywords and the search token received. Under the proposed security model, the adversary can only learn the structure of the expression tree rather than the information about keywords.

[2] Yuan In their study, Propose a secure data deduplication scheme with efficient re-encryption based on the convergent all-or-nothing transform (CAONT) and randomly sampled bits from the Bloom filter. Due to the intrinsic property of one-way hash function, our scheme can resist the stub-reserved attack and guarantee the data privacy of data owners' sensitive data. Moreover, instead of re-encrypting the entire package, data owners are only required to re-encrypt a small part of it through the CAONT, thereby effectively reducing the computation overhead of the system. Proposed scheme is designed for enterprise or user groups in which multiple users want to outsource the data to a remote cloud service provider. The cloud service provider can conduct deduplication on ciphertexts and save abundant storage overhead. The system of our scheme contains three entities: cloud user, key server and cloud service provider (CSP). The CAONT mechanism has the property that one must decrypt the entire ciphertext before one can determine even one message block. If a data owner transforms a message into the t and C2 packages by using CAONT, the data owner is not required to re-encrypt the entire package but only a small part of the package C2, which saves excessive computation overhead. However, to protect the data privacy of data owners' sensitive data, not only do we need to prevent the revoked cloud user from accessing the original message, but also need to prevent the CSP from accessing the original message. In order to solve the above problems, we propose a Bloom filter-based location selection method and a secure data deduplication scheme with efficient re-encryption.

[3] Lidong, Junling Guo In this paper we Construct a fast and secure public key authenticated searchable encryption scheme with designed server. Our scheme can resist keyword guessing attacks, chosen multi-keyword attacks and multi-trapdoor attacks. The search function in our scheme achieves the logarithmic search time in number of keywords while most existing schemes required the linear time. It should be noted we does not discuss the technique how to encrypt the files since it is not within the scope of our research. The data owner extracts a set of keywords from data. He encrypts the files and keywords, and then builds a secure index for different keywords. Next, the index and the ciphertexts of data are uploaded by data owner to the cloud server. When a user does searching some data containing the queried keyword in the cloud, he constructs a keyword trapdoor and then sends it to the server. After receiving the search query, the cloud server utilizes the trapdoor to do search operation over the index and sends the search result to the user. Efficiency and security are two important aspects of evaluating the public key searchable encryption schemes. Sometimes they are one pair of irreconcilable contradictions. In order to make testing very fast, our scheme is based on the trust of a server and allows a server to distinguish the keyword ciphertexts and trapdoors. Despite it has weakness, the server cannot break the ciphertexts and trapdoors. It is hard to distinguish the keyword ciphertexts and trapdoors for any external adversary.

[4] Wei Yin In this paper, we Propose an ABMKS with only multiplication operations in encrypted keyword index generation. This scheme involves four entities: DO, DU, CSP and Authority. The DO will encrypt the data files set F as well as corresponding keyword sets with an access policy before uploading them to the CSP. The CSP provides the storage services and executes keyword search operations on behalf of the DU. When a DU wants to make a search query over the encrypted data, he/she generates a trapdoor by his/her specified query keywords and submits it to CSP. On receiving the trapdoor, CSP retrieves the appropriate data file by using the trapdoor, if the user's attributes satisfy the access policy and the trapdoor matches the encrypted keyword index I. The role of Authority is to issue credentials (PK/SK) to the data owners/users, the credentials are sent over secure communication channel. The threat model of our system is as follows: DO, Authority and the authorized data users are trusted, but the CSP is a trusted-but-curious entity which honestly executes the protocol but attempts to learn some sensitive information, e.g., the query keyword information. . The computation operations in the index generation are only multiplication, which is more efficient than modular exponentiation and pairing. In addition, the encrypted keyword indexes are aggregated into one item, being independent on the number of underlying keyword in a file. The formal security analysis shows that proposed scheme is secure.

PROBLEM DEFINITION

The problem we aim to address is the Cloud data storage poses significant challenges in terms of data security, redundancy, and efficient access control. Traditional storage methods often result in the duplication of data, leading to increased storage costs and potential security vulnerabilities. Furthermore, ensuring secure and controlled access to sensitive information stored in the cloud remains a critical concern. In this context, the lack of a comprehensive solution that addresses duplicate detection, data encoding, and secure access poses a substantial risk to the confidentiality and integrity of stored data. To overcome these challenges, the project introduces a holistic approach to cloud data management. It begins by implementing an advanced duplicate detection mechanism using chunk-based similarity checking, effectively identifying and eliminating redundant data. Data integrity

and compatibility are enhanced through the application of Base64 encoding before storage on the cloud, optimizing space utilization. Access requests are addressed through a secure process that involves re-encryption using the AES algorithm, ensuring that sensitive information remains protected during both storage and transmission. Users are provided with new decryption keys for secure data access, and local decryption using the shared key facilitates verification, ensuring only authorized users can access the stored data. This multifaceted solution addresses key concerns in cloud data storage, offering a robust framework that optimizes storage efficiency, enhances data security, and provides controlled and secure access to stored information.

OVERVIEW OF THE PROJECT

In essence, the combination of chunk-based similarity checking and re-encryption offers a comprehensive solution to the challenges of duplicate detection, efficient storage management, and secure data sharing in cloud environments. This project aims to advance the capabilities of cloud data systems, fostering a more resilient and streamlined approach to data management in an era where efficient resource utilization and data security are paramount. By implementing data deduplication techniques, the project seeks to reduce storage costs and enhance data management efficiency. Chunk-based partitioning and deduplication mechanisms are employed to identify and eliminate redundant data chunks, thereby optimizing storage utilization. By reducing data redundancy, the proposed approach aims to improve data retrieval performance, particularly in scenarios where large volumes of data need to be transferred over the network. This optimization can lead to faster access to data, benefiting users who rely on cloud storage for their data needs. The project prioritizes data security by employing Base64 encoding for data representation and Advanced Encryption Standard (AES) for secure data sharing. Base64 encoding helps protect data integrity during transmission and storage, while AES encryption ensures that shared data remains confidential and accessible only to authorized users. Through re-encryption techniques and unique encryption keys, the project enables secure data sharing among users in cloud environments. By implementing robust encryption mechanisms, it mitigates the risks associated with unauthorized access and strengthens the overall security posture of cloud-based data sharing.

CLOUD STORAGE FRAMEWORK

Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in either privately owned, or third-party data centers that may be located far from the user–ranging in distance from across a city to across the world. Cloud computing relies on sharing of resources to achieve coherence. In this framework, we can have two types of users such as data owner and data provider. The person or organization that legally owns a cloud service is called a cloud service owner. The cloud service owner can be the cloud consumer, or the cloud provider that owns the cloud within which the cloud service resides. Cloud service provider provides the storage space to the users. Storage space can be shared by multiple data owners. Data owners can be upload the files in storage system for future use.

DE-DUPLICATION CHECKING

In computing, data compression is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the compression process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. In this module, we can check the files using file name with file contents. Encrypted files are spilted into chunks. Service provider checks the chunks at the time of uploading files. Data owner only upload original file so save storage space in cloud system. We can compression in text file, document file and image files.

- **Chunk Based Similarity Checking**
- Input: a chunk data list of super-chunk S, {fp1, fp2, … , fpn}
- Output: a target node ID, i
- 1. Select the k data chunk {rfp1, rfp2, …, rfpk} and sent the chunk to candidate nodes with IDs {rfp1 mod N, rfp2 mod N, …, rfpk mod N} in the deduplicaton server cluster with N nodes;
- 2. In deduplication server cluster, obtain the count of the existing representative data chunk of the super-chunk in the candidate nodes by comparing the representative data chunks of the previously stored super-chunks in the similarity index. Index comparison was performed using Map-Reduce framework.
- 3. The returned k count values, one for each of the k candidate nodes, are denoted as {r1, r2, …, rk}, which are directly corresponding to the resemblances of S in these nodes;
- 4. Mapper find the duplicate content then reducer eliminates the redundant content.
- 5. Choose the deduplication server node with ID i that satisfies ri/wi = max{r1/w1, r2/w2, …, rk/wk} as the target node.

FILE ENCRYPTION

Encryption is the most effective way to achieve data security. To read an encrypted file, you must have access to a secret key or password that enables you to decrypt it. Base64 encoding transforms binary data into a text-based format, consisting of a set of ASCII characters, making it more suitable for storage and transfer across different systems. By converting binary information into a human-readable text representation, Base64 encoding facilitates compatibility with a wide range of devices and applications. This encoding process is particularly relevant in the context of cloud storage, where data is often transmitted over networks and stored in various formats. Base64-encoded data is not only platform-independent but also aids in the prevention of data corruption during transmission. Moreover, the utilization of Base64 encoding contributes to an added layer of security. While it is not a method of encryption, Base64 encoding obscures the original content, making it less susceptible to casual observation. This can be advantageous when sensitive data is being transferred or stored on cloud servers, providing a degree of obfuscation that adds a level of security, albeit not a replacement for robust encryption mechanisms.

Base64 Encoding

- Base64 is a widely used encoding scheme that plays a crucial role in data transmission and storage across various digital platforms. It provides a means to represent binary data, such as images, audio files, or any arbitrary binary information, in a human-readable format comprised of printable ASCII characters.

- The transformation of Base64 is one of the algorithms for encoding and decoding data into ASCII format, which is based on the number 64.

- The characters generated from Base64 consist of "A..Z", "a..z" and "0..9", and the last two characters are "/" and "+".

There are several steps to be done to complete the Base64 algorithm is:

- Look for the ASCII code of each text.
- Search binary number 8 bits of the ASCII code exist.
- Combine the last 8 bits to 24 bits.
- Then, broke a 24 bit earlier to 6 bits. It will produce four fractions.
- Each fragment is converted into a decimal value.
- Lastly, make value - the decimal value to an index to choose a character constituent of base64.

DATA ACCESS REQUEST

When a user logs into a cloud application and need to request access to specific data owned by another user, a streamlined and secure process is crucial. The request includes details such as the specific data sets or files needed, along with cloud verification for the access. Simultaneously, the cloud application generates a notification alert for the data owner, informing them of the access request. The data owner, upon logging into their account, can review the request, assessing the legitimacy and necessity of granting access. To ensure transparency, both parties receive real-time updates on the status of the request.

RE-ENCRYPTION AND DATA SHARING

In this module AES-based data re-encryption is employed, the process involves encrypting data with a unique key and subsequently re-encrypting it with a different key to enable secure sharing. When a user requests access to encrypted data, an AES-based data re-encryption mechanism comes into play. The data owner, who possesses the original encryption key, initiates the process by generating a new encryption key. Using this new key, the data is re-encrypted, creating a secure envelope around it. Simultaneously, the data owner securely shares the decryption key associated with the newly encrypted data with the requesting user. This key exchange occurs through a secure channel, ensuring confidentiality. Once the requesting user receives the decryption key, they can use it to decrypt the re-encrypted data and access the information securely. This approach provides an additional layer of security and control, allowing data owners to share information selectively while maintaining the integrity of the original encryption.

AES (Advanced Encryption Standard)

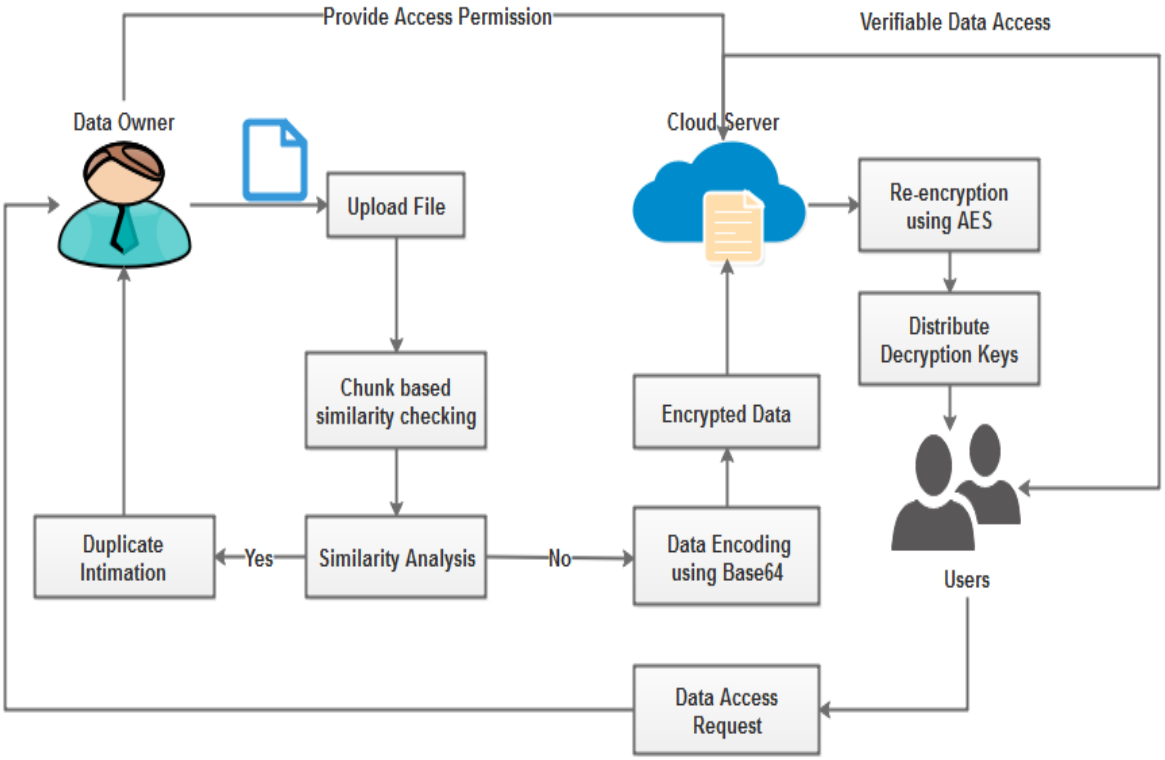
- The algorithm begins with an Add round key stage followed by 9 rounds of four stages and a tenth round of three stages.
- This applies for both encryption and decryption with the exception that each stage of a round the decryption algorithm is the inverse of its counterpart in the encryption algorithm. The four stages are as follows:

  1. Substitute bytes
  2. Shift rows
  3. Mix Columns
  4. Add Round Key
- The tenth round simply leaves out the **Mix Columns** stage. The first nine rounds of the decryption algorithm consist of the following:

  1. Inverse Shift rows
  2. Inverse Substitute bytes
  3. Inverse Add Round Key
  4. Inverse Mix Columns
- Again, the tenth round simply leaves out the **Inverse Mix Columns** stage. Each of these stages will now be considered in more detail.

SYSTEM ARCHITECTURE

A system architecture or systems architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviours of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behaviour) between them. It can provide a plan from which products can be procured, and systems developed, that will work together to implement the overall system. There have been efforts to formalize languages to describe system architecture; collectively these are called architecture description languages (ADLs).



CONCLUSION

For safe and effective data management, the proposed project offers cloud data storage with Base64 encoding and AES encryption, chunk-based similarity checking method for duplication detection, and re-encryption for data sharing. The use of AES encryption guarantees the privacy and security of data kept on cloud storage from unauthorised access. Modern data storage is incomplete without AES encryption because of its strong security features, which offer a strong defence against data breaches and privacy violations. The system's capacity to recognise and remove redundant data pieces lowers storage expenses and improves the cloud-based storage system's overall performance. Re-encryption is a crucial advance in data sharing since it allows for safe data interchange without sacrificing data integrity.

I. ACKNOWLEDGMENT

We are grateful to Dr.A.Jeyamurugan, Assistant Professor, CSE Department, Paavai Engineering College(Autonomous) for mentoring us to present this paper successfully

REFERENCES

[1] Wu, Yunyun, Jingyu Hou, Jing Liu, Wanlei Zhou, and Shaowen Yao. "Novel multi-keyword search on encrypted data in the cloud." IEEE Access 7 (2019): 31984-31996.

[2] Yuan, Haoran, Xiaofeng Chen, Jin Li, Tao Jiang, Jianfeng Wang, and Robert H. Deng. "Secure cloud data deduplication with efficient re-encryption." IEEE Transactions on Services Computing 15, no. 1 (2019): 442-456.

[3] Han, Lidong, Junling Guo, Guang Yang, Qi Xie, and Chengliang Tian. "An efficient and secure public key authenticated encryption with keyword search in the logarithmic time." IEEE Access 9 (2021): 151245-151253.

[4] Cui, Yuanbo, Fei Gao, Yijie Shi, Wei Yin, Emmanouil Panaousis, and Kaitai Liang. "An efficient attribute-based multi-keyword search scheme in encrypted keyword generation." IEEE access 8 (2020): 99024-99036.

[5] Yu, Xixun, Hui Bai, Zheng Yan, and Rui Zhang. "Veridedup: A verifiable cloud data deduplication scheme with integrity and duplication proof." IEEE Transactions on Dependable and Secure Computing 20, no. 1 (2022): 680-694.