ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 8, Issue 1,January 2022

A NOVEL DEDUPLICATION TECHNIQUE ON ENCRYPTED BIG DATA IN CLOUD USING SECURE DEKEY METHODOLOGY

S. Seethalakshmi
Research Scholar,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: seethasri.lakshmi@gmail.com

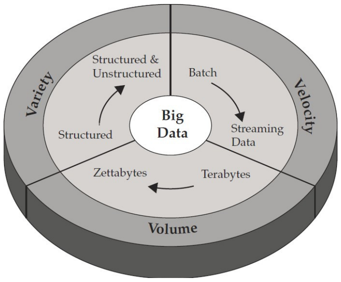
Dr. B. Balakumar
Assistant Professor,
Manonmaniam Sundaranar University, Tirunelveli.
Tamilnadu, India
Email: balakumarcite@msuniv.ac.in

Abstract — A complicated collection of massive data sets is referred to as data. Using traditional and existing database management tools or data processing programmes, processing such a large volume, diversity, and velocity of data is difficult. Data storage is the most significant and widely used cloud service. Data is frequently kept in the cloud in encrypted form to protect data holders' privacy. Encrypted data presents additional issues for cloud data deduplication, which is becoming increasingly important for huge data storage and processing in the cloud. Flexible Deduplication Scheme leverages both the data owner and CSPs for deduplication control based on encryption behaviour in the present system. It also improved a heterogeneous data storage management strategy that allows for customizable data deduplication and access control in the cloud. Our scheme is adaptable to a wide range of application scenarios and demands, and it provides cost-effective big data storage management across numerous cloud service providers. Content hash keying, also known as convergent encryptions, is a cryptosystem that generates identical ciphertext from identical plaintext files. The fundamental shortcoming of the previous method was that the key required to verify deduplication was distributed to users in a decentralised manner. For the verification process, this will result in significant CPU consumption and high bandwidth connection. Dekey presented a new architecture in the project, in which customers don't have to manage any keys themselves, but instead safely distribute convergent key shares among numerous servers. Dekey is secure according to the definitions stated in the proposed security model, according to security analysis. The fundamental benefit of the proposed method is that no user keys are required, and the DeKey Generation calculation is done only on servers. In addition, the system employs both Full File Analysis and Block Analysis, which contribute in the system's ability to consume less memory and CPU. The DeKeys will be stored on the storage server and delivered to the verification system in a centralised way, preventing key loss or update.

Keywords — Cloud Computing; DeDuplication; Cloud Storage

I. INTRODUCTION

A complicated collection of data files is referred to as cloud computing. Using traditional and existing database management tools or data processing programmes, processing such a large volume, diversity, and velocity of data is difficult.



This applies to capture, creation, storage, search, sharing, transfer, analysis, and visualisation, among other things. To correlate and recognise business trends, determine research quality, prevent diseases, link legal citations, battle crime, and establish real-time roadway traffic conditions, a single huge group of relevant data must be analysed as a whole.

Handling and effectively utilising Big Data is largely dependent on the organization's ability to manage such huge sets, as well as its applications' ability to process and analyse the data sets in its domain. When confronted with hundreds of gigabytes of data for the first time, some organisations may need to reassess their data management strategies. Others may need tens or hundreds of terabytes of data before the scale of the data becomes a relevant factor in Big Data adoption.

By arranging multiple resources across the Internet, cloud computing provides a new way of providing services. Data storage is the most significant and widely used cloud service. Data is frequently kept in the cloud in encrypted form to protect data holders' privacy. Encrypted data presents new obstacles for cloud data deduplication; it becomes critical for cloud big data storage and processing. On encrypted data, traditional deduplication algorithms are ineffective.

II. LITERATURE SURVEY

In [1], This paper author proposes a provable data possession protocol. Recently the demand of the cloud computing increased due to its ability, flexibility and large storage capacity. Normally in public clouds the user keeps the data in the server but they can’t access the data remotely. The information security is a significant problem in public cloud storage, such as data confidentiality, integrity, and availability. Some cases the client can’t check the possession of the data. This paper talks about a proxy provable data possession (PPDP). In public clouds, PPDP is a matter of crucial importance when the client cannot perform the remote data possession checking. Here it focuses on the PPDP system model, the security model, and the design method. Based on the bilinear pairing technique, it verifies whether the server possess correct data.

Technology: Proxy Provable Data Possession (PPDP)

Advantages

- The overhead at the server is low.
- Performance of PDP is bounded by disk I/O and not by cryptographic computation

Drawbacks

- Performance decreases as the number of users increases

In [2], This paper author describes a broadcast encryption. Some set S of the users will be listening to the broadcast channel. A broadcaster will encrypt the message. Inorder to decrypt it the user can use the private key. We can say that the system is collusion resistant in the sense that the users outside the broadcast can’t collude the information. The broadcaster can encrypt to any subset S of his choice. In the file system the access control can be provided by this broadcast encryption.

Advantages

- Each user only has to keep a single secret key

Drawbacks

- It needs more storage and time for to implement.

In [3], This paper author describes Cloud Security Using Service Level Agreements. It makes use of some of the service level agreements. Users can specify the needs. And with the help of the security level agreement, they can be provided. Te security level agreement consists of different security parameters in different point of view.

Technology: Service Level Agreements

Advantages

- Clients outwardly analyze CSPs in view of their offered secSLAs.

Drawbacks

- Nothing mentioned about advanced security metrics/Cloud secSLA notions
- Uncertainty, end-to-end security evaluation

In [4], This paper author says about a Proxy reencryption (PRE). A significant benefit brought by PRE is that each user only has to keep a single secret key and does not suffer from the key escrow problem. In a PRE system, a data owner can generate a reencryption key to help Bob transform a ciphertext under her own public key to ciphertext of the same message under Bob’s public key, such that Bob can decrypt it by using his own secret key to obtain the original message. However, the access control provided by a traditional PRE (including ID-based PRE) is in an “all or nothing” manner. Namely, a user with a corresponding reencryption key can read all the data of the data owner, but a user without the re-encryption key cannot read any private data of the data owner. This is not applicable to the case that the data owner only wants to share part of his data with others. In order to realize flexible access control, a new variant of PRE can be used.

Technology: Proxy Re-Encryption (PRE)

Advantages

- Implementation is simple

Drawbacks

- Some area is error prone

In [5], This paper creator says in regards to a strategy for deduplicate information put away in cloud in light of the possession test and intermediary re-encryption. It integrates cloud data deduplication with access control. We evaluate its performance based on extensive analysis and computer simulations. The outcome demonstrates the proficiency and adequacy of the plan, particularly for huge information deduplication in distributed storage.

Advantages

- Low Cost of Storage
- Can efficiently perform big data deduplication

Drawbacks

- The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

OBJECTIVE

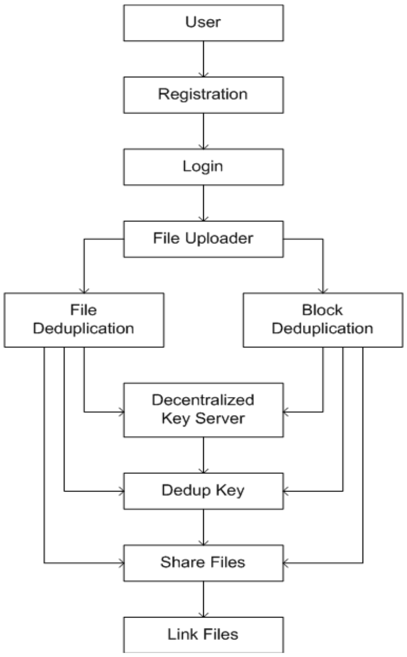
- To save cloud storage and preserve the privacy of data holders by proposing a scheme to manage encrypted data storage with deduplication.
- To support flexibly for data sharing with deduplication even when the data holder is offline, and it does not intrude the privacy of data holders.
- To propose an effective approach to verify data ownership and check duplicate storage with secure challenge and big data support.
- To prove the security and performance of the scheme through analysis and simulation.

MOTIVATION

For Cloud storage, integrity was checked with storage deduplication. It overcomes this open-source dilemma by employing novel methods such as polynomial-based authentication tags and holomorphic linear authenticators. This architecture allows for the deduplication of both files and authentication tags. At the same time, data integrity auditing and storage deduplication are accomplished. On the user side, the proposed approach is also classified by constant real-time communication and computational cost. Both public auditing and batch auditing are available. The suggested technique outperforms prior schemes while also including deduplication functionality.

III. SYSTEM MODEL

We improve the security of our system in this project. We offer a better method for increasing file security by encrypting it with differential privilege keys. As a result, people without the required permissions are unable to perform the duplicate check. Even if such unauthorised people cooperate with the DeKey, they will be unable to decrypt the encrypted text. According to security analysis, our system is secure according to the definitions provided in the suggested security model.



File Upload Process

Data storage will be provided via the Cloud Server. Cloud customers upload personal or confidential data to a Cloud Service Provider's data centre and authorise it to store it. Because intrusions and attacks on sensitive data at Cloud Service Providers are unavoidable, cloud users should assume that CSP cannot be entirely trusted. Data security concerns, particularly data privacy leaks, are increased when people lose control of their own personal data. In order to protect data security and user privacy as the privacy issue grows more serious, it is a good practise to only outsource encrypted data to the Cloud.

Key Generation

Cloud consumers may not be able to fully trust CSP. Data is frequently kept on the Cloud in encrypted form to protect the privacy of data owners. Encrypted data, on the other hand, presents significant obstacles for Cloud data deduplication. However, the same or different users may upload copied data to a Cloud service provider in encrypted form, where the data is shared among numerous users. Despite the large amount of storage space available in the cloud, data duplication wastes a lot of network resources, consumes a lot of energy, and makes data management more difficult. To circumvent this, it presents a novel ownership verification approach to improve prior work and to support Big Data data integrity in an effective manner.

Data Redundancy Process

The same encrypted raw data could be saved in the Cloud service provider by a number of qualified data holders. Data owner refers to the person who produces or creates the file. It has a higher priority than other data holders since it is an authorised party that does not work with CSP and is fully trusted by data holders to verify data ownership and conduct data deduplication.

DATA RETRIEVE AND KEY VERIFICATION

The data owner encrypts his or her private key and sends it to CSP along with the encrypted data. After a challenge, the genuine data owner can be validated. For example, the data owner should produce a specific certificate to prove ownership. Simply verify the key and retrieve the encrypted data from CSP, after which the data will be decrypted using the generated random key.

Algorithm For Secure Access

The selective restriction of access to a location or other resource is known as secure access. Consuming, entering, or using are all words that can be used to describe the process of accessing. Locks and login credentials are two similar access control techniques. Once the key has been generated, the encryption and decryption processes are relatively simple and computationally simple. Where the private key is used to decrypt a communication transmitted using a public key and vice versa, each person's private and public keys must be scientifically related. The cryptosystem is used in a number of well-known asymmetric encryption methods. Although the public key can be published in a public directory, such as with a certifying authority, the owner must keep the private key completely secret.

```
Begin
  User Registration
    Generate User Profile Key
    Generate User Encryption Key
  File Storage
    Upload Files
    Generate De-Dup Key for File
    Slice File Segments
    for Each Slice in Segments
      Generate Block De-Dup Key for each
    end for
  Encryption
    File Encryption
    Block Encryption
End
```

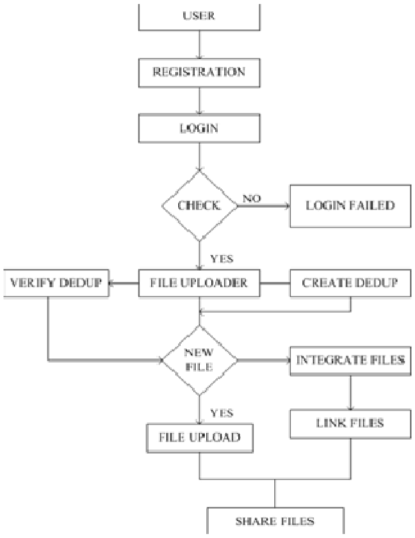
Block Level Deduplication Algorithm

The block-level distributed deduplication approach is known as Block-Level Deduplication. Before uploading a file in a block-level deduplication system, the user must first execute file-level deduplication. If no duplicates are identified, the user breaks the file into blocks and deduplicates them at the block level. The system will be set up similarly to the file-level deduplication system, with the exception that the block size parameter will be defined separately. Following that, we go over the specifics of the File Upload and File Download algorithms.

```
Begin
  User File Upload
    Generate DEDUP Key
    Generate File Blocks
  Deduplication Processing
    Generate Hash Keys
    Index Keys to File Blocks
    Encrypt File using File Hash
    Encrypt File Blocks using Block Hash
  Deduplication Verification
    File Upload
    If (New File)
      Generate Index File
      Generate Block Hash Code
    Else
      Link Index File and Block Hash Code
    End If
End
```

IV. PROJECT DESCRIPTION

Deduplication can be done at either the chunk or file level. Because it identifies and eliminates redundancy at a finer granularity, chunk-level deduplication is preferable over file-level deduplication.



Chunking is one of the most difficult aspects of the deduplication process. One of the most important factors determining overall deduplication performance is chunking efficiency. In addition, the chunking step has a big impact on the deduplication ratio and performance. CPU heavy phases include chunking and chunk fingerprinting, whereas memory and disc intensive steps include fingerprint indexing and querying, as well as data storage and administration.

V. FUTURE SCOPE

Managing encrypted data with deduplication is significant in the way for achieving a successful Cloud storage service, particularly for Big Data storage. To proposed a practical scheme to manage the encrypted Big Data in Cloud with deduplication based on ownership challenge. This scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only the authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that this scheme is secure and efficient under the described security model and very suitable for Big Data deduplication. The results of our computer simulations further showed the practicability of our scheme

V. CONCLUSION

The appropriated deduplication frameworks with higher unwavering quality in which the information pieces are conveyed over various cloud servers. The system uses traditional De Key based deduplication detection detection and avoids unwanted traffic and storage in servers using Big Data processing. The protection analysis demonstrates that proposed deduplication systems are secure. As a proof of concept, the system implemented incurred overhead is very limited in realistic environments. The proposed system provides a solution for preserving the data in cloud with the avail of encryption protocol.

REFERENCES

[1] Huaqun Wang, “Proxy Provable Data Possession in Public Clouds”, 2013.

[2] Dan Boneh, Brent Waters, Craig Genry, “Collusion Resistant Broadcast Encryption with Short Ciphertexts and Private Keys”, 2005.

[3] Y. Xiang, W. Zhou, and M. Guo,"Cloud Security Using Service Level Agreements,"2009.

[4] Giuseppe Ateniese, Kevin Fu, Matthew Green, and Susan Hohenberger,"Improved proxy re- encryption schemes with applications to secure distributed storage,”2016.

[5] Sabale Nikita C, Prof. N. G. Pardeshi, “A Survey Paper on Deduplication on Encrypted Big Data Using HDFS Framework,”2017.

[6] Rongmao Chen, Yi Mu, Guomin Yang, Fuchun Guo,"BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication,"2015.

[7] Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart,"DupLESS: Server-Aided Encryption for Deduplicated Storage,"2013

[8] Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart,"Message-Locked Encryption and Secure Deduplication,"2013.

[9] Fatema Rashid, Ali Miri, Isaac Woungang, “Secure Enterprise Data Deduplication in the Cloud”, 2013.