An Approach for Secure Distributed Deduplication System

Anita A. Kundgir
Dept. of CSE, (CNIS)
SGGS IE & T, Nanded, India

S. S. Hatkar
Dept. of CSE,(CNIS)
SGGS IE & T Nanded, India

ABSTRACT

In this paper we introduce a deduplication system with improved reliability. As we all know the cloud computing performs a number of operations on data. All commercial cloud service providers know that the users demand for huge amount of storage space because number of online operations. Deduplication is good thing to implement but it will not work on encrypted data because of conflict in encryption. As we know convergent encryption, where key is derived from the hash of data which is recovered by the same encryption key.to overcome from all such interrupts we need a secure system which will perform the deduplication. The system checks at block level as well as file level with assignment of tags. Digital libraries contains huge amounts of data. Most of the times while data storage the number of copies of the same data are stored, again and again, so to remove such duplicate data copies we need a Deduplication technique. The deduplication removes unnecessary data, but at the same time, it is beneficial if it retrieves it with more reliability. In this paper, we have studied various approaches to improve the reliability of data after removing duplicate copies in data storage. The use of such deduplication schemes also reduces the data transfer rate to upload and download data .The time consumption should be less which will effect on its accessibility.

Keywords

Cloud storage, Deduplication, Integrity; Secret Sharing Scheme, Message locked encryption.

1. INTRODUCTION

By the unusual improvement of advanced information, deduplication systems are extensively drawn in to backup information and Reducing network and storage Transparency by notice and annihilate excess amount of information. As an option of keeping up numerous information duplicates with a similar substance, deduplication reduces repetitive information by keeping up just single duplicate and allowing other excess information to that duplicate. Data Deduplication Scheme is used for removing all the unnecessary data which is replica of other.

There are basically two types of data deduplication on basis of implementations as follows.

1.1 File level Deduplication

In this type of compares a file that has to be archived or backup that has already been stored by checking all its necessary attributes against the index. The index is the record of the files. The index is updated and stored only if the file is unique if not then only a pointer to existing file store references. The only single instance is saved and the other are replaced by the stub to the original file. In this sort of thinks about a record that must be filed or reinforcement that has as of now been put away by checking all its vital properties against the list. The file is the record of the documents. The list is refreshed and put away just if the record is exceptional if not then just a pointer to existing document store references. The main single occurrence is spared and the other are supplanted by the stub to the first record.

1.2 Block level deduplication

This type of duplication operates on the basis of the sub file level. The time is broken into segments. The blocks of chunks will be examined for previously stored information. The popular approach to determine redundant data is assigning identities to the chunk of data by using a just algorithm, which generates the unique ID of that particular block, the particular id is compared to the central index. Data Deduplication scheme allows compression of data for removing the replicas of same data type. To improve the storage performance the scheme is applied to the online data to reduce the size of data. As in block level deduplication the blocks and chunks are compared to storage data size .the chunks are used to store the index values. The index maintains the records of uploaded data to system. As Compress which is a program based on LZW algorithm, which performs fast in less storage space. Another is DEFLATE a lossless data compression algorithm which is a combination of LZ77 and Huffman coding.

2. TYPES OF DEDUPLICATION

Data deduplication type is classified on the basis of deduplication timing. As the deduplication has to perform data deduplication, it creates a record for new files or data items with respect to the time [13].

2.1 Offline Deduplication

In this case the data deduplication algorithm is performed on the all data which is to be stored in storage system. The advantage of this scheme is that it is performed on all static data which has been already stored is storage system, and it will improve the efficiency .only the disadvantage is that it effect on performance and system will be little slower.

2.2 Online Deduplication

As comparing to offline data deduplication, it is performed when the data is uploaded at the time. The advantage of this scheme is that it allows space reallocation but the problem is that it increases the waiting time since the write operation of file is stopped until the duplicate files are removed.

2.3 Whole File Hashing

In this method the whole file is sent to hashing function. The MD5 and SHA1 hashing functions are used basically hash function is used to map the data to fixed size .the cryptographic hash allows data to map the hash values .hash functions are used for building caches of large data sets. Advantage is that due to full data backup the efficiency of system increased. Only drawback is that increasing granularity of duplicate data, it prevents the duplicate data it prevents the duplicate copies which differ by data byte.

2.4 Sub File Hashing

In this method before mapping the hash value the file is divided into several chunks i.e. called subfiles .The subfiles are divided with fixed length chunking and variable length chunking. Various fingerprinting schemes are used to determine chunk size, the broken files are transferred to cryptographic hash function [6].

3. LITERATURE SURVEY

In 2013 Mihir Bellare. Sriram Keelveedhi,T. Ristenpart , ”Dupless Server aided encryption for deduplicated storage “.they introduced the concept of security and confidentiality by the scheme of symmetric encryption they explained symmetric encryption as well as convergent encryption. The system enables client to store encrypted data with an service which checks it for deduplicated storage ,can reach the goal of performance with more reliability[2]

The data Deduplication scheme is used for neglecting the replicas of similar data types. This scheme provides homogeneous results which reduces the unnecessary space occupied by replicas of the same data type.

Data deduplication used for removing replication copies of data. Data deduplication techniques are very interesting techniques. The reliability produces stable and consistent results. They only focused on files without encryption, without considering the reliable deduplication over cipher text. Cipher text is also known as encrypted or encoded information. In1997 Mihir Bellare introduced the idea of security and scheme for symmetric encryption in concentrate security framework.

Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which will unnecessarily use more bandwidth and network. So here some common technique is being defined which finds the hash for the particular file and with that the process of duplication can be simplified, David Geer.

The concept of proof of ownership in cloud storage system is explained by D. Harnik. they identified the security issues related to the dropping out of date and time consumption by using the client side encryption[1]. In 1997Author Adi Shamir proved the Secure Sharing Scheme they shown that how to divide the data D into the n pieces such that they can be easily recovered by key K.Where every k-1 piece does not have the information about D.They proved the scheme for robust key management for system which can be secure and reliable for working.

In 1989 Author Michel O.R developed an information dispersal algorithm which breaks the file into blocks or small pieces it has numerous application to secure the storage information of computer networks for fault tolerance and efficient transmission of network[11].

In 2002 J. R. Douceur,Willium J.B,D.Simon ,M. Theimer ”Reclaiming space from Duplicate Files in a Serverless Distributed File System”, provides a Farsite distributed file system for the purpose of reclaiming storage spaces from the storage. The system enables the identification and collect together all data files when they are provided with encryption with different users. They provided two schemes convergent encryption and other is SALAD Self-Arranging, Losssy,Associative Database for aggregating file contents and location information scalable and fault tolerant manner. The system removes the copies of files with ideal content to storage systems[12].

In 2013 at EUROCRYPT ,Author formalized a new cryptographic encryption called Message Locked Encryption for deduplication[3].

4. METHODOLOGY

4.1 Convergent Encryption

It is used to provide a data confidentiality in secure deduplication. It performs encryption and decryption operations by using the convergent key which is obtained from a cryptographic hash value. In data encryption process the user regains the keys and forwards the cipher text. Since the copies of similar, identical data will generate the similar convergent key and same cipher text. Gather for the decryption the cipher text and convergent key are used to get the original data. The convergent Encryption is defined with four functions.[3]

- ➢ *KeyGen (M) →K*

In the key generation algorithm, it takes original data M as input. And maps it into a convergent key K.

- ➢ *Encrypt (K, M) →C*

In Encryption algorithm like symmetric encryption takes both key K and Data M as input and then outputs ciphertext c.

- ➢ *Decrypt (K, C) → M*

In decryption algorithm, it takes input as ciphertext C and the same convergent key K which is used for encryption and outputs the original data M.

- ➢ *TagGen (M) →T(M)*

The Target generation algorithm target the original data and outputs a tag to the data. By using this method two users having ideal data copies can get two ciphettexts with same encryption key hence the cloud service provider will easily able to perform deduplication scheme on this data.

As the encryption keys are randomly generated from the data provided .so no need to communicate between data owner and users.

The Disadvantage of the scheme is that as encryption key is derived from the data itself. so if the intruder get access to data storage can break security such attacks are called as Dictionary Attacks. In such attack shared secret is compared with the data.

4.2 Symmetric Encryption

This encryption uses the similar secret key for the purpose of encryption and decryption. They can share the secret between two or more parties, which maintains private resource Information. This encryption defined with the three functions.[10]

- ➢ *Key Generation (KeyGen())→k*

This algorithm generates the key for encryption. The random key ‘k' is said to be distinctive and nonspecific.

- ➢ *Encryption (Enc(m,k)) →C*

This algorithm takes an input of identical data ‘ M ' and random key ‘ K' which produce the output cipher text.due to the unique key k the output C is also identical every time.

- ➢ *Decryption (Dec(C,K)) → M*

This algorithm generates the output as identical data plaintext M by taking input as a Key and Cipher text. The same key must be used which is used at the time of encryption.

4.3 MLE Scheme

MLE is a symmetric encryption scheme in which the encryption and decryption keys are same and are derived from the message shared. As it enables duplication of cipher text, it lets key to the message. The encryption algorithm allow the cipher text and key to recover the message M. The tag generation algorithm maps the cipher text to a tag used for the server to detect deduplication in files .this scheme accepts the tag generation to cpphertext.it is resistant to fake attacks as practically they provide ROM security Analysis .they make connections with deterministic encryption hash function secure on correlated input for different message sources[9].

5. IMPLEMENTATIONS

5.1 Existing System

Several suggestions made for securing remote data in the cloud using various security aspects which may be usefull to consume the security of user data . As we know that deduplication reduces the storage space at cloud server side.Data reliability is very critical issue in storage systems ,because only a single instance of file has been accessed by multiple user.

To ensure private information the secret sharing scheme is utilized which is relating to disseminated capacity systems. In this paper the secret sharing method is utilized for security of private data. In detail a document is isolate and encode into parts by utilizing secret sharing scheme. These parts will be conveyed over numerous independent stockpiling servers. A cryptanalysis hash estimation of the substance will likewise be ascertained and send to storage server as the characteristic of the section put away at every server.

5.2 Proposed System

Deduplication term is defined as removing the same data copy or replicas of data. Data deduplication can be measured by following two function as Deduplication ratio and Throughput. [12]

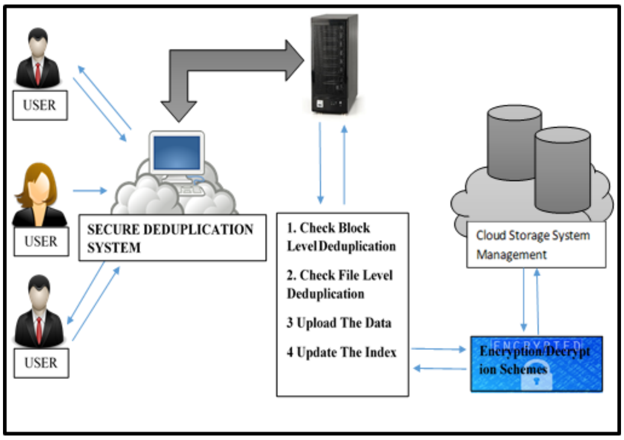

USER
USER
USER
SECURE DEDUPLICATION SYSTEM
1. Check Block Level Deduplication
2. Check File Level Deduplication
3 Upload The Data
4 Update The Index
Cloud Storage System Management
ENCRYPTED
Encryption/Decrypt ion Schemes

Fig. 1 System Architechture

Several suggestions made for securing remote data in the cloud using various security aspects which may be usefull to consume the security of user data. As we know that deduplication reduces the storage space at cloud server side.Data reliability is very critical issue in storage systems, because only a single instance of file has been accessed by multiple user.

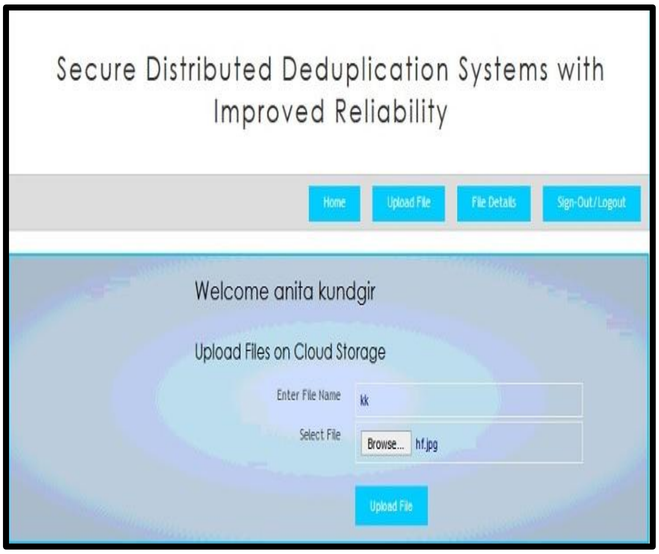
To ensure private information the secret sharing scheme is utilized which is relating to disseminated capacity systems. In this paper the secret sharing method is utilized for security of private data. In detail a document is isolate and encode into parts by utilizing secret sharing scheme. These parts will be conveyed over numerous independent stockpiling servers. A cryptanalysis hash estimation of the substance will likewise be ascertained and send to storage server as the characteristic of the section put away at every server.

Components of our System Includes Following

5.2.1 Data Owner

The data owner who first upload the data is required to calculate and distribute such secret shares and following users own same data copy do not need to calculate and stores these shares. Retrieve data copies owner must access a minimum number of storage server by a validation and obtain the secret shares to alter the data.in different way, the authorized uses will access the secret shares data copy. Methodology:

When the Data owner wants to upload the file on cloud server then it will register and after that continue towards login. After authenticating all things Data owner will be able to upload the Data. The key generation is done with the SHA-1 algorithm.



Secure Distributed Deduplication Systems with Improved Reliability

Welcome anita kundgir

Upload Files on Cloud Storage

Enter File Name: kk

Select File: hf.jpg

Fig.2 Data Upload

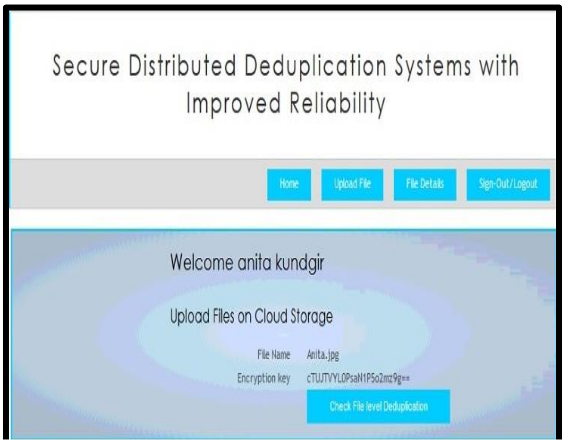
5.2.2 System Checks Deduplication

After uploading the file it turns towards the deduplication method

a. *File Level Deduplication*

In this the AES algorithm is used for File Encryption in System, And Sha-1is used for sharing the files.

- If the duplicate file is found in deduplication then it will send its information to storage cloud service provider to check whether the file is already present in the index ,and will assign the reference of that stored file and discard the upload of new file.
- If the duplicate file is not found then system will send the File to storage cloud service provider and upload the file with new entry in index.
- We use Secret sharing scheme which allow only secure and reliable sharing.



Secure Distributed Deduplication Systems with Improved Reliability

Welcome anita kundgir

Upload Files on Cloud Storage

File Name Anita.jpg

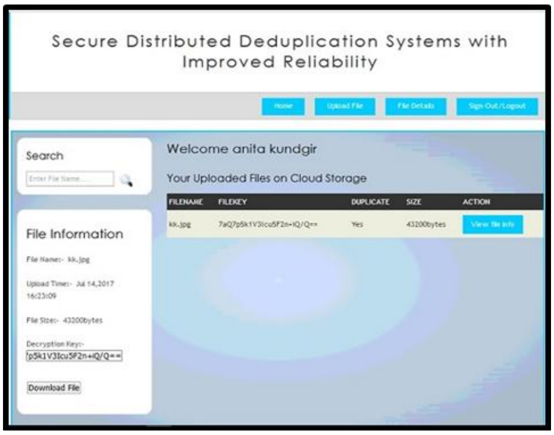
Encryption key cTUJTVYL0PsaN1P5o2mz9g==

Fig. 3 Checks Deduplication

b. Block Level Deduplication

In this after checking the file level Deduplication if no duplicate found then it will check it for blocks.it divides the file in number of blocks and perform the deduplication.

- If the duplicate is found in any part of the blocks by using SHA-1 algorithm for computing hash value of the block, it checks the index saved at cloud service provider. Then it will point towards the previously stored files.
- If the duplicate block is nor found in checking then it will assign it into the index with new entry at cloud service provider side and upload the file.



Secure Distributed Deduplication Systems with Improved Reliability

Welcome anita kundgir

Your Uploaded Files on Cloud Storage

FILENAME	FILEKEY	DUPLICATE	SIZE	ACTION
kk.jpg	7aQ7p5k1V3Icu5F2n+IQ/Q==	Yes	43200bytes	View file info

File Information

File Name:- kk.jpg

Upload Time:- Jul 14,2017 16:23:09

File Size:- 43200bytes

Decryption Key:- 'p5k1V3Icu5F2n+iQ/Q==

Fig. 4 File Uploading Details

5.2.3 Secret Sharing Scheme

The scheme contains two algorithms Share and Recover. In this two algorithms present share and recover. The secret is divided and shared by using Share algorithm. RSSS use to secretly splitting of secret into shards. Specifically (n,k,r)-RSSS (where n>k.r>=0)generates n shares from secret. Finally user uploads set of values {Uid, Fid, Fname, Encrypted file, si, H(F)}By using Share the user shares the key to access the file and by using Recover they can access any file on a cloud. In RSSS no information about the file reduced by any no. of shares.[7]

5.2.4 System with Tag Consistency

As we are preventing cipher text to be getting duplicate, because the main aim is to maintain only single copies of identical data. So tag provides security guarantees against the fake duplication attacks. The tag generation is purely performed by the data owner that's why chances of to be suffering from malicious attacks are minimized [2].

5.2.5 Download File

In this section the Files Stored at Storage cloud Service provider .If the other user wants to download the file then it have to register on the system after that it will have to take permission from the Data owner to refer the file. Because the Share of the file is with the data owner until it allow .After having valid decryption key from the data Owner the user can access the file.

6. CONCLUSION AND FUTURE SCOPE

In this we study the several approaches to maintain data with identical copies, deduplication schemes. And try to implement a system with improved functions. Deduplication Systems are very useful in today's life. Another discernable element of our proposition is that information culmination includes tag consistency, can be inferred. To clarify further if a similar esteem is put away in different distributed storage then deduplication check by methods. It can't restrict the crash assault set up by numerous servers. As far as anyone is concerned no related work on secure deduplication can appropriately address, the unwavering quality and tag consistency issue. As most of IT industries now turning towards data security and integrity various technologies are implemented. By stating all this scheme we expect the better enhancement.

7. REFERENCES

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

[2] Mihir. Bellare, Sriram. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage," in USENIX Security Symposium, 2013.

[3] "Secure deduplication with efficient and reliable convergent key management" by J. li,Chen, jingwei,parteic and W.Lou ,in IEEE 2013.

[4] "A Survey on Deduplication Scheme in Cloud Storage" of author Deepa D.,Revathi M.

[5]" An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing"in IEEE transactions in 2012 ,Author K. yang and X.jia.

[6] "Fingerprinting by random polynomials",by M. Rabin in 1981.

[7] "How to share secret" by Adi Shamir in ACM commun.,1979.

[8] "Multiple Ramp schemes",by A. santis and B.Masuci in IEEE Transactions july 1999.

[9] "Message Locked Encryption in secure deduplication"in Eurocrypt,2013.

[10] "A Secure data deduplication scheme for cloud storage" in technical Report, 2013 by J. Stanek, A. Sorniotti, Androulaki, kenel L.

[11]"Efficient dispersal of information for security ,load balancing,fault tolerance" in ACM journal 1989,M.Rabin.

[12] "Reclaiming Space from Duplicate Files in a Serverless Distributed File System". John R.Douceur, Atul Adya, William J. Bolosky, D.Simon, M. Theimer Microsoft .

[13] "A Survey on Data Deduplication in Cloud Storage Environment", in IJSRET 2015.by Mani U.V., G. Mohan