IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Secure Distributed Deduplication in Cloud with Improved Reliability

Reshma D. Kapadi[1], Prof. Pramod Patil[2], Prof. Prashant V. Raut[3]

Student, Computer Networks, NMIET Talegaon Dabhade, Pune, India[1]

Asst. Professor, Computer Engineering, NMIET Talegaon Dabhade, Pune, India[2]

Asst. Professor, Computer Engineering, Sinhgad Institute of Technology, Pune, India[3]

Abstract: Data deduplication is best way for eliminating identical copies of data. This technique has been greatly used in cloud. It results in reduced storage space and network bandwidth for upload. Only one copy of file stored in cloud even though file has number of owners. This gives improved of storage space but reduces reliability. Furthermore users also stores their sensitive data on cloud. Security of this sensitive data becomes a great challenge. This research proposes new deduplication system with higher reliability. This technique is used to save storage space and bandwidth under cloud. In this research data chunks are stored on multiple cloud servers. Data deduplication has two types one is file level deduplication and other is block level deduplication. The new deduplication system also maintains direct communication between deduplication System's users and owner of the file. Users can directly communicate with owner and request for file. Owner of the file send the respective file to the user in secure way. The deduplication system is mainly used in educational institutes and industry. The deterministic secret sharing scheme is used to achieve security requirements of data privacy and tag consistency. This research incurred small overhead in realistic environments.

Keywords: Deduplication, cryptography, distributed storage system, reliability, security.

I. INTRODUCTION

Now a day's every user have data which knowingly or unknowingly to him or her gets stored on the cloud for the future use of that user. But, sometimes it may happen that number of users also try to save the same copy of the data on cloud. It results in duplication of the data, and there will be wastage of storage space over the cloud. Secondly, if multiple users are trying to store same copy of data over the cloud, then it will reduce the network bandwidth also. Storage cost is very important factor while dealing with storage space. Today's world is around the cloud, and need of storage space is increasing day by day. If storage space is directly proportional to the storage cost then there will be wastage of money also. So, there is a need of system that prevents to store copy of same data. Deduplication of data keeps only one copy of data on cloud and if some other user is trying to store same data he/ she will get notification like the file or data already exist with owner's identity and you can use this copy of data. The user who uploaded the file first will be the owner of that data.

Deduplication of data can be done in two ways i.e. file level and block level as:

1. **File Level:** In this type of deduplication, file gets divided into number of shares viz. S1, S2, S3… etc. These shares are stored on number of cloud service provider. When user wants download that file then Kout of n shares will recollect and after reconstruction forms original file.
2. **Block Level:** In block level deduplication each file is divided into number of blocks as B1, B2, B3,… Bn. These blocks gets encrypted and stored over the cloud. In this case encryption is used to prevent data from attackers.

II. LITERATURE SURVEY

Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang [1] proposed data deduplication to eliminate duplicate copies of data. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

M. Bellare, S. Keelveedhi, and T. Ristenpart, conclude Distributed storage administration suppliers, for example, Dropbox, Mozy, and others perform deduplication to spare space by just putting away one duplicate of each file transferred. Should customers expectedly scramble their files, on the other hand, investment funds are lost. Message-bolted encryption (the most noticeable indication of which is focalized encryption) re-understands this strain. On the other hand it is characteristically subject to animal power assaults that can recuperate _les falling into a known set. The system propose a construction modeling that ace vides secure deduplicated stockpiling opposing savage power assaults, and acknowledge it in a framework called DupLESS. In DupLESS, customers encode under message-based keys obtained from a key-server by means

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

of a careless PRF convention. It empowers customers to store encoded information with an exist-ing administration, have the administration perform deduplication for their sake, but accomplishes solid confidentiality guarantees. The system demonstrate that encryption for deduplicated stockpiling can accomplish execution and space reserve funds near that of utilizing the stockpiling administration with plaintext info.

J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer proposed framework gives accessibility by recreating every record onto numerous desktop PCs. Since this replication expends critical storage room, it is essential to recover utilized space where conceivable. Estimation of more than 500 desktop document frameworks demonstrates that about portion of all devoured space is possessed by copy records. The system introduce an instrument to recover space from this accidental duplication to make it accessible for controlled document replication. Deduplication System incorporates 1) focalized encryption, which empowers copy records it combine into the space of a solitary document, regardless of the fact that the documents are scrambled with diverse clients' keys, and 2) SALAD, a Self-Arranging, Lossy, Associative Database for conglomerating document substance and area data in a decentralized, adaptable, issue tolerant way. Substantial scale recreation analyses demonstrate that the copy document mixing framework is versatile, exceedingly successful, and de_ciency tolerant.

A. D. Santis and B. Masucci gives plan to disseminate a mystery s picked in S among a set P of n members in a manner that: (1) sets of members of cardinality more noteworthy than or equivalent to k can reproduce the mystery s; (2) sets of members of cardinality not as much as or equivalent to t have no data on s, though (3) sets of members of cardinality more noteworthy than t and not as much as k may have some data on s. In this correspondence the system examine various incline plans, which are conventions to share numerous privileged insights among a set P of members, utilizing distinctive slope plans. Specifically, the system demonstrate a tight lower bound on the offers' extent held by every member and on the merchant's arbitrariness in numerous slope plans. A. Shamir demonstrates that industry standards to gap information D into n pieces in a manner that D is effortlessly reconstructable from any k pieces, however even finish learning of k - 1 pieces uncovers truly no data about D. This method empowers the development of hearty key administration plans for cryptographic frameworks that can work safely and dependably notwithstanding when adversities wreck a large portion of the pieces and security ruptures uncover everything except one of the remaining pieces.

III.DATA DEDUPLICATION

A. System Flow Diagram

The system flow diagram shows two main activities first is upload process and second is download process. In upload process user has data block, he generates hash key for that particular data block. User forwards the data block and hash key to the metadata manager where deduplication is going to check. In download process user request for data at this point metadata manager checks availability of the data If data is available then he decrypt that data and make it available for user.

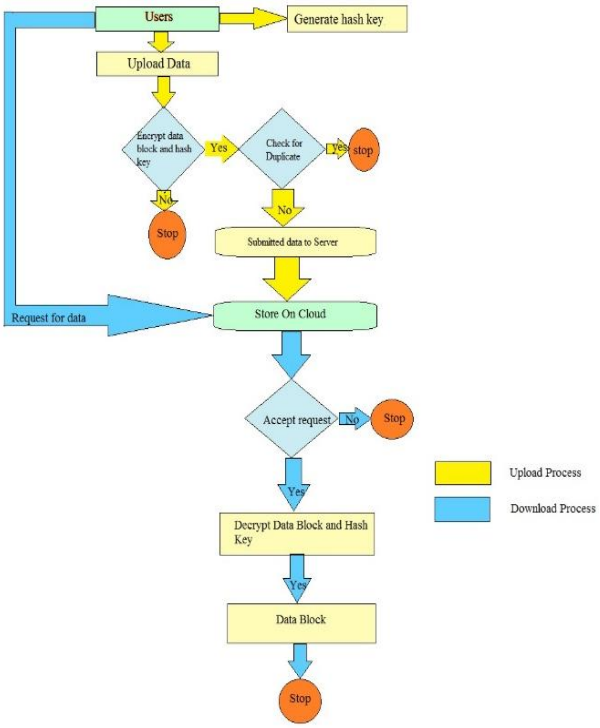

Fig.1 System Flow Diagram

B. Operations Performed

The main operations performed are:

1. Registration
2. Upload
3. Download
4. Edit Details
5. Upload Details
6. Request File
7. Check File Request
8. Send File

1. Registration:

Every user must be register with deduplication system. In Registration process user must have to fill his personal details like first name, last name, address, contact details etc. User can also choose his username and password for login in to the system.

2. Upload:

To upload a file, hash key must be generated. The data block and hash key should get encrypted. Then file should get checked that whether it is duplicate or not. If file is duplicate then user must get notification that file already exist. And if file not already present then file must get upload successfully.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

3. Download:
To download a file user must send request. Request is forwarded to metadata manager. Metadata manager checks for a file decrypt it and forwards to the user.

4. Edit Details:
User has facility to change his username and password.

5. Upload details:
Every details of the user gets uploaded to the cloud like his first name, last name, username, contact details etc.

6. Request File:
With this operation all registered users can send request for file to the owner of that particular file.

7. Check File Request:
With this operation owner can check all file requests sent by the registered user.

8. Send File:
Owner of the file send the particular file to the user who sent request for a file.

C. Mathematical Model
Let S be a system that find out duplicate copies of the _le using deduplication system in cloud. S=F,B,C,T,P,M,O
Where,

F= {F1, F2, F3, Fn}
F1={B1, B2, B3, B}
B1={CBi, TBi, PKi}
CBi=Set of cipher text block
T = Token [16-Bit unique token for Block]
P= Private Key (PKi) used for encryption and description mechanism
M=Metadata of file
O=Output consist reduce database size

Following steps occurs in the given proposed system architecture:

1. File F is divided into multiple blocks
F=∑ Bi,
F = size (F) / 4096,
2. KeyGen(l)→ k is key generation algorithm, generate secrete key using security parameter l.
Secret key stores internal DB of Security Service (SS).
3. Enc (k,F) →C is encryption algorithm that takes secrete key k and file and then output is cipher text C.
4. Generate Token T for each block.
5. Dec(k,C) →F is Decryption algorithm that takes secrete key k and ciphertext C and then output is original file F.

F=∑ PlainText (Bi)
PlainText (Bi) = SS (CipherText (Bi), TiBi).

6. Detect duplication.

Security Service generates TiBi Token on basic on Bi, If the same Bi comes in then it will generate the same TiBi. i.e. TiBi = token generation (Bi); Then it will store the TiBi to the Own Security Db. If file is found in database it generates response.

D. Algorithm Used
The following algorithms are used in this deduplication system.
1. Ramp's Secret Sharing Scheme.
2. Tag Generation Algorithm.
3. Message Authentication Code.

These three algorithms are used in this deduplication system they are described as follows:

1. Ramp's Secret Sharing Scheme:
There are two algorithms in a secret sharing scheme, which are Share and Recover. The secret is divided and shared by using Share. With enough shares, the secret can be extracted and recovered with the algorithm of Recover. In deduplication System implementation, The system will use the Ramp secret sharing scheme (RSSS) [7], [8] to secretly split a secret into shards. Speci_cally, the (n, k, r)-RSSS (where n _ k _ r = 0) generates n shares from a secret so that (i) the secret can be recovered from any k or more shares, and (ii) no information about the secret can be deduced from any r or less shares. Two algorithms, Share and Recover, are defined in the (n, k, r)-RSSS.

- **Share** divides a secret S into (k -r) pieces of equal size, generates r random pieces of the same size, and encodes the k pieces using a non-systematic k-of-n erasure code into n shares of the same size;
- **Recover** takes any k out of n shares as inputs and then outputs the original secret S.

2. Tag Generation Algorithm:
In deduplication System constructions below, two kinds of tag generation algorithms are defined, that is, TagGen and TagGen'. TagGen is the tag generation algorithm that maps the original data copy F and outputs a tag T (F). This tag will be generated by the user and applied to perform the duplicate check with the server. Another tag generation algorithm TagGen' takes as input a file F and an index j and outputs a tag. This tag, generated by users, is used for the proof of ownership for F.

3. Message Authentication Code:
A message authentication code (MAC) is a short piece of information used to authenticate a message and to provide integrity and authenticity assurances on the message. In this construction, the message authentication code is applied to achieve the integrity of the outsourced stored files. It can be easily constructed with a keyed (cryptographic) hash function, which takes input as a secret key and an arbitrary-length file that needs to be authenticated, and outputs a MAC. Only users with the same key generating the MAC can verify the correctness of the MAC value and detect whether the file has been changed or not.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

E. Results

The encoding and decoding times of deduplication systems for each block (per 4KB data block) are always in the order of microseconds. To check the efficiency of the system the RSSS algorithm taken into consideration. The number of S-CSPs n = 6 and the reliability level n - k = 2 are fixed. From the figure, it can be easily found that the encoding/decoding time increases with r.

Case 1:

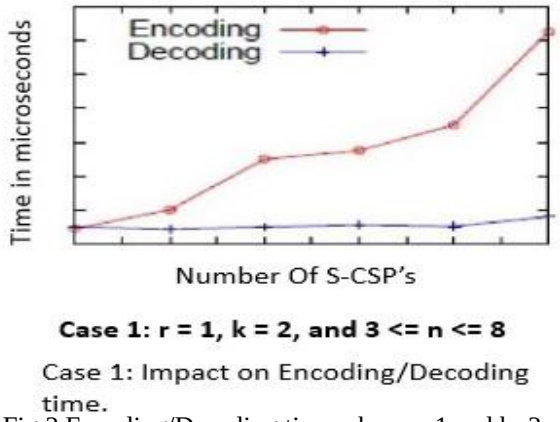

Fig 2 Encoding/Decoding time when r =1 and k=2

Case 2:

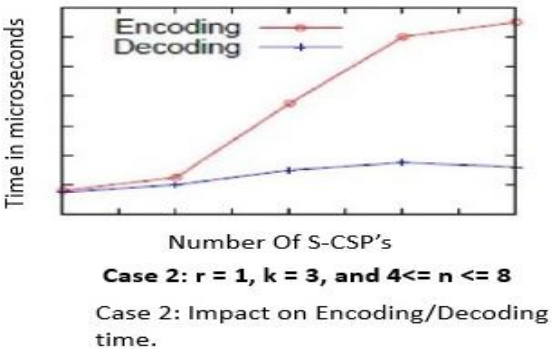

Fig 3 Encoding/Decoding time when r =1 and k=3

Case 3:

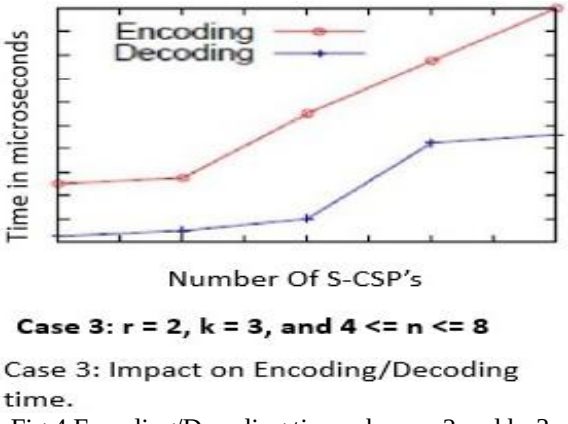

Fig 4 Encoding/Decoding time when r =2 and k=3

Case 4:

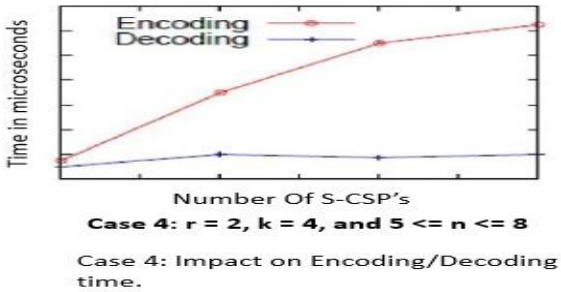

Fig 5. Encoding/Decoding time when r =2 and k=4

F. Deduplication System

The following figures show the various operation regarding deduplication system.

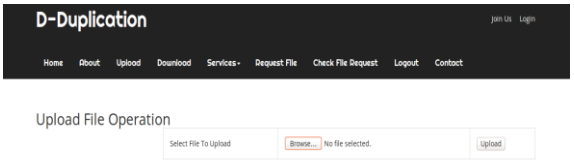

Fig. 6. Upload Operation of Deduplication System.

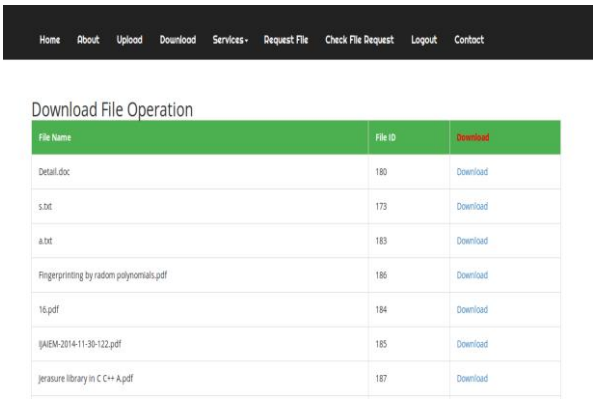

Fig. 7. Download Operation of Deduplication System

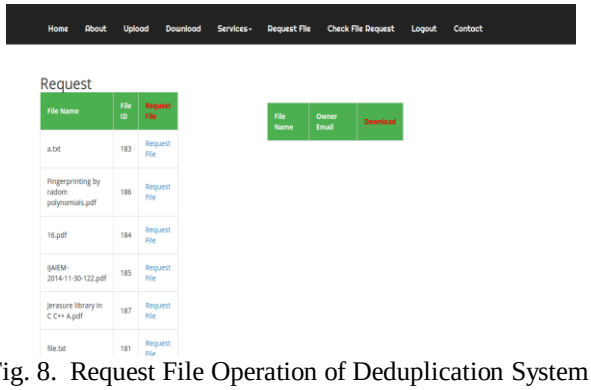

Fig. 8. Request File Operation of Deduplication System

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

The figure 6 shows the upload operation of the deduplication system. User can upload the file from here. The figure 7 shows the download operation and the figure 8 shows Request File operation.

IV. CONCLUSION

The fundamental thought is that the deduplication system set that protected deduplication administrations can be executed given extra security highlights insider aggressor on Deduplication and pariah assailant by utilizing the discovery of masquerade action. The assailant's perplexity and the extra expenses caused to recognize genuine from false data, and the discouragement impact which, albeit difficult to quantify, assumes a critical part in forestalling masquerade movement by danger unwilling aggressors. Deduplication System set that the mix of these security elements will give uncommon levels of security to the deduplication.

ACKNOWLEDGMENT

I would like to thank **Prof. Pramod Patil** and **Prof. Lomesh Ahire** and anonymous reviewers for their valuable suggestions that significantly improved the quality of this paper.

REFERENCES

[1] Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang Senior Member, IEEE and Mohammad Mehedi Hassan Member, IEEE and Abdulhameed Alelaiwi Member, Secure Distributed Deduplication Systems with Improved Reliability, 2015.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, Dupless: Server-aided encryption for deduplicated storage, in USENIX Security Symposium, 2013.

[3] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, Reclaiming space from duplicate les in a serverless distributed le system. in ICDCS, 2002, pp. 617624.

[4] A. D. Santis and B. Masucci, Multiple ramp schemes, IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 17201728, Jul. 1999.

[5] Message-locked encryption and secure deduplication, in EUROCRYPT, 2013, pp. 296312.

[6] A. Shamir, How to share a secret, Commun. ACM, vol. 22, no. 11, pp. 612613, 1979.

[7] 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. 491500.

[8] M. Li, C. Qin, P. P. C. Lee, and J. Li, Convergent dispersal: Toward storage efficient security in a cloud-of-clouds, in The 6th USENIX Workshop on Hot Topics in Storage and File Systems, 2014.

[9] J. S. Plank, S. Simmerman, and C. D. Schuman, Jerasure: A library in C/C++ facilitating erasure coding for storage applica- tions - Version 1.2, University of Tennessee, Tech. Rep. CS-08-627, August 2008.

[10] P. Anderson and L. Zhang, Fast and secure laptop backups with encrypted deduplication, in Proc. of USENIX LISA, 2010.[18] Z. Wilcox-OHearn and B. Warner, Tahoe: the least-authority.

[11] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, A secure data deduplication.

BIOGRAPHIES

Miss. Reshma D Kapadi P.G. Student NMVPM's Nutan Maharashtra Institute of Engineering and Technology Talegaon Dabhade, Pune, India. She has received B.Tech degree in Computer Engineering from VIT Pune.

Prof. Pramod Patil Assistant Professor at NMVPM's Nutan Maharashtra Institute of Engineering And Technology Talegaon Dabhade, Pune, India.

Prof. Prashant V. Raut Assistant Professor, Department of Computer Engineering, Sinhgad Institute of Technology, Pune, Maharashtra, India.