ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Deduplication of Distributed Cloud Storage by Improving Confidentiality, Integrity and Reliability

Prerna Lahane, Prof. Sarika Bodake

ME Student, Dept. of Computer Engineering, P.V.P.I.T Bavdhan, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, P.V.P.I.T Bavdhan, Pune, Maharashtra, India

ABSTRACT: Deduplication is a method of eliminating or removing duplicate files or data from the storage or database, and is used in cloud computing environment to minimize storage space as well as network bandwidth. In data deduplication single copy of file is stored in cloud environment, which is owned by a large number of data users which may leads to reduced reliability, availability and security. Here we are proposing new distributed data deduplication system, which achieves more confidentiality, integrity and reliability, as compare to the traditional deduplication system. In the proposed system single data chunk or file is divided and distributed on the multiple cloud storage servers, in such a way that the single part of file is unpredictable. This is done by introducing a Shamir secret sharing scheme and hashing algorithms in distributed cloud storage system, without using traditional ways of encryption-decryption scheme. To achieve principles of storage system like reliability, integrity, security, availability and confidentiality, proposed system focusing on various kinds of attacks which affects these principals. Also proposed system will be focusing on the recovery and reconstruction of corrupted data or failed storage site without using traditional backup or recovery methods as like RAID array method.

KEYWORDS*:* Deduplication, distributed storage system, reliability, availability, integrity, Shamir secret sharing, Hashing algorithms.

I. Introduction

Data deduplication removes duplicate data or files from the storage, and is widely used in cloud computing to save storage space and network bandwidth. Data security, integrity and reliability are critical issues in a deduplication storage system. In the deduplication system only one copy of file is shared among multiple users, which leads to improved storage utilization and network bandwidth with less reliability. If such a file or data block was lost, then large amount of data becomes inaccessible due to unavailability of these files or data chunk. Also to provide data security and achieve integrity to single file or data chunk should be taken into consideration. Thus, how to achieve high data security, integrity and reliability in deduplicated system is a major problem. The distributed deduplication system is introduced to improve the reliability, confidentiality and integrity of the users outsourced data without using traditional encryption technique. To secure data on the cloud is the main challenge in the cloud computing environment. Here we attempt to propose new distributed deduplication system with higher confidentiality, integrity and reliability in which the secret or main data is distributed on the multiple cloud storage servers in secret share form. This security requirements of data is achieved by introducing Shamir secret sharing scheme in distributed storage system [1]. A number of de-duplication strategies have been proposed such as client-side and server-side de-duplication, file-level and block-level De-duplication [2].

A. Design goals :

- Confidentiality: To achieve confidentiality we are focused on the secret sharing scheme, which divides original data into multiple parts and distributed across multiple storage servers. Due to this scheme we are able to protect data from chosen distribution attack, collusion attack and collision attack, brute force attack kind of attacks.

- Integrity: To achieve integrity of data we are focused on the hash calculation and authentication check to avoid insider attack and duplicate replacement attack.
- Reliability: To achieve reliability we have used distributed system to provide fault tolerance means even if any server fails or data is corrupted, then recovery option is available to detect failed server or corrupt and repair the user’s data.

II. RELATED WORK

Following work has been performed in previous deduplication schemes to achieve data deduplication and confidentiality on the data.

Jin Li et al [1] provided secure deduplication system by introducing distributed deduplication server system and tag consistency. They focused on achieving the data reliability. Here they proposed distributed deduplication system along with Ramp secret sharing scheme instead of using convergent encryption scheme. Tag consistency used for integrity purpose and tag generation is done by the end user. This scheme gives kind of workload to the end user.

Li [3] addressed convergent encryption, provides data security in deduplication. The convergent encryption (CE) scheme uses the Dekey scheme, where Dekey constructs secret share on plain text and distributes across multiple cloud service providers. Dekey can be used by multiple users instead of different key for different users. Here multiple user shares the same block, due to which storage space is minimized. Key-management is the issue in this work. They focused only on the confidentiality of data. In this work cipher text can be easily duplicated.

M Bellare et al [4] In DupLESS(Duplicate Encryption for Simple Storage), clients encrypt data under message-based keys, which is obtained from the key-server. Group of clients encrypt data with the help of key server which is separate from storage server. Clients have to authenticate themselves to the key server. But the point of failure is key server, unless key server is secure whole system is secure. DupLESS technique achieves strong confidentiality. They have focused on the Server side deduplication. But key server is the single point of failure.
Mihir Bellare et al [5] Message-Locked Encryption (MLE), performs encryption and decryption of the text by the key, which is derived from the message (encryption primitive). Message is mapped to the key for encryption and decryption technique. Here symmetric encryption scheme has been used by clients for secure deduplication. This scheme provides both privacy and integrity of data. But they worked fine deduplication with single client not for multiple clients.
Stanek et al [6] explained encryption scheme which differentiates security of popular data (easily available e.g. audio, video etc.) and unpopular data (e.g. Password, private photos etc.). They provides secure data deduplication scheme for cloud storage. Popular data are normally not sensitive hence the traditional encryption mechanism is performed. For unpopular data they have provided semantic security and multilayered cryptosystem for supporting deduplication. So they provides security to the outsourced data in the cloud deduplication system.

III. PROPOSED WORK

A. Problem Definition

Proposed system supports client, server, file and block level data deduplication over the cloud to minimize storage space and upload bandwidth with improved security, reliability, Integrity and availability. Here we introduced distributed storage system, where data chunks or secrets are distributed across multiple storage servers with the help of secret sharing scheme. Instead of using traditional cryptography we have focused on the secret sharing scheme. Here we proposed a solution for recovery and reconstruction of corrupted data or failure site without use of traditional recovery plans (RAID method).

B. System Architecture

In the proposed work we are focused on the file level and block level, client side and server side deduplications. Fig 1 describes, when user wants to upload the file on the cloud Storage at that time user requests to the web browser for uploading the file. Only approved user can upload or download the file through web server. To check the authorized user we are calculating hash value of the data at file and block level using two different hashing algorithms(Rabin fingerprint[8] and MD5[9]), those hash values are shared with users and servers as a proof of ownership. By matching these hash values we are easily detect the original user and attacker. Process to be followed in upload and download operation is shown in Fig 2. When file is uploaded, the original file splits into fixed size blocks. According to the file size the number of blocks are created. After this process deduplication check occurs. Data storage server contains all the uploaded files in the form of secret share of that files along with hash value. With the help of share and recover, algorithms of Shamir secret share encryption and decryption takes place. Distributed storage structure in deduplication system provides better fault tolerance and high availability. Further to protect data confidentiality, Shamir secret sharing algorithm [7] is proposed, which having greater compatibility with the distributed storage system

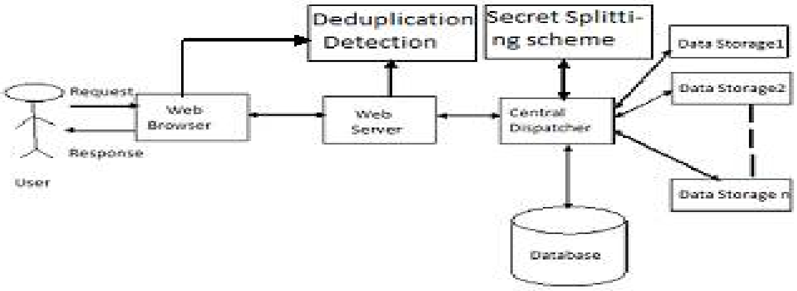

Fig 1: Architecture of proposed system

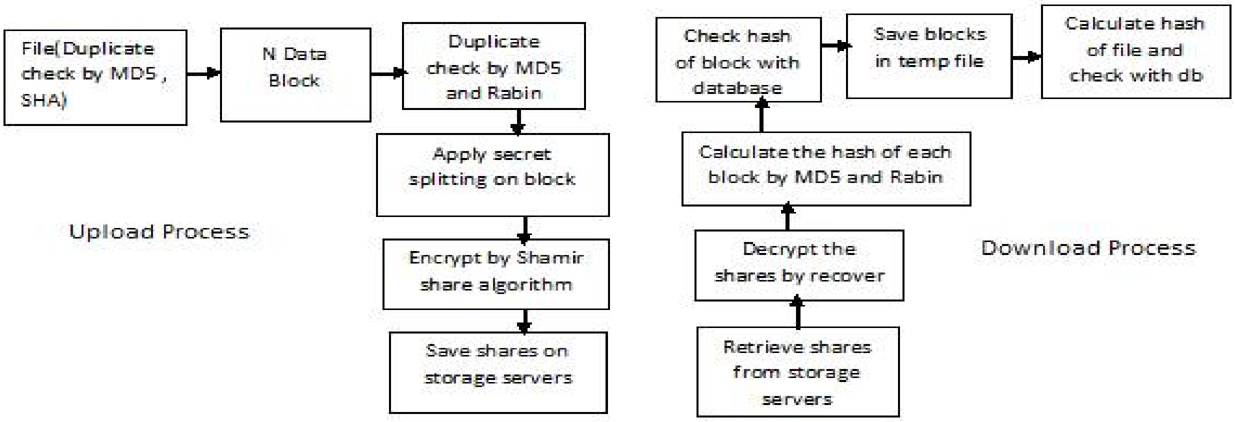
File(Duplicate check by MD5 , SHA)
N Data Block
Duplicate check by MD5 and Rabin
Apply secret splitting on block
Encrypt by Shamir share algorithm
Save shares on storage servers
Upload Process
Check hash of block with database
Save blocks in temp file
Calculate hash of file and check with db
Calculate the hash of each block by MD5 and Rabin
Decrypt the shares by recover
Retrieve shares from storage servers
Download Process

Fig 2: Detailed process of proposed system

IV. IMPLEMENTATION PLAN

Following modules are implemented in the proposed scheme:

A. *Distributed deduplication system:* The distributed Deduplication system stores data on the distributed cloud storage by achieving confidentiality, integrity, reliability and integrity. It distributes the data across multiple storage servers more reliably. Our new construction uses the technique of secret splitting to split data item (secret) into multiple secret shares, and distributes these shares across multiple distributed storage servers.On the distributed sites central dispatcher stores secret share along with their hash values.

B. *File level and block level deduplication*: In this Module file and fine grained block level deduplication check occurs. After which hash value calculated twice by two different algorithms to avoid collusion attack. Such a block then dispatched by central dispatcher over distributed storage site with the help of secret splitting technique.

C. *Data Construction module:* In the proposed system we can recover the corrupt data or failure site, with the help of Shamir algorithm. Also one of the more important feature of the proposed system is we can reconstruct the failure site, with the help of database and Shamir secret share algorithm without using traditional recovery ways like RAID etc.

D. *Data Modification module:* Data modification Operations like create, read, write and update takes place in a secure way on the shared data blocks such that no user will be affected by the other user's modification on the stored data, if both users pointed to the same file. Modified data will be shown to the respective users only. Collusion attacks handled properly.

E. *Mathematical Model:*

- Let D be the Data Deduplication System.
- **Upload Process**
- Input = {File F}
- F = {Binary, Text, Audio, video, Doc, image}
- H is the set of calculated hash values.
- H = h1, h2,h3,h4
- B = b1, b2,b3...
- where, h1 ← md5(F)
- h2 ← SHA(F)
- h3 ← md5(B)
- h4 ← Rabin (B)
- F = File, B= Block of file
- File deduplication ← Compare (h1, h2)
- Block deduplication ← Compare (h3, h4)
- File F = B1, B2,B3...
- B1 is set of various divided blocks.
- B1 =B11,B12,B13...
- Use (K,n)threshold scheme to generate share secrets
- where K < n , n = number of shares and k = minimum shares or threshold
- Determine n points by,

$$f(x) = a0 + a1x + a2x^2 + \ldots + a_{k-1}^{k-1}$$

- Put x=1, 2,3...
- Generate Secret share(SS)
- Let CS = CS1, CS2, CS3...
- CS be the distributed cloud storage servers.
- CS1 ← SS1
- Store all secret shares over distributed Storage server.
- Output : Secret shares of whole file

- **Download Process**
- Input: Secret shares(SS)
- Determine Secret S
- To reconstruct the secret we are using following formula:
- For convenience, let yi denote P(xi). We can generate the coefficients of P using Lagrange Interpolation. Define,

$$P(x) = \sum_{i=0}^{t-1} y_i \prod_{\substack{0 \le j \le t-1 \\ j \ne i}} \frac{x - x_j}{x_i - x_j}.$$

- Block B SS1, SS2, SS3...
- Generate complete file F
- F ← B1,B2,B3...

Output: Complete File F

F. Algorithms:

1) Shamir Secret share:

- Secret Sharing Scheme is given by two algorithms: sharing (Share) algorithm and recovery (Recover) algorithm. Secret sharing technique, divides secret into multiple parts, each participant has given unique part of secret. To reconstruct the secret all or some parts of the secret must be present. The property of the Shamir Secret Share is: Recover (Share (M)) = M.

2) Rabin Fingerprint

- It is a procedure that maps an arbitrarily large data item to a shorter bit string.
- Its fingerprint that uniquely identifies the original data.
- This fingerprint can be used for data deduplication purpose.
- Easily detect modification of file or data.
- Works as high-performance hash functions.
- Used to check data collision attack.

3) MD5 and SHA-256

- Widely used cryptographic hash function
- Used to verify data integrity
- In the proposed system MD5 and SHA-256 used to calculate hash value of data chunks.
- Used to avoid file level or block level collision attack
- Used for Sub file hashing or whole file hashing.

V. Results

We are proposed the system which provides the storage functionality across the cloud or distributed topology. This system achieves file or block level deduplication without affecting the principles of storage system like reliability, integrity, security, availability and confidentiality. The system is strong enough to tackle various attacks like insider attack, collision attack, chosen distribution attack etc. System can be easily recoverable in case of failure. The system will have negligible operational overhead as compare to other ways to achieve above principles and provides user friendly interface for operations. CRUD (create, read, update, delete) operations will be supported by system on user data. Table 1 describes how we can save storage and network bandwidth in duplicate and half duplicate files. As shown in table 1 different files are taken as an input to store on the distributed cloud storage. Here we mentioned how much percentage we can save network bandwidth and storage size when we are uploading any type of file. Result varies according to the status of file like duplicate file, half duplicate file (some of the contents in the file are same with other stored file. Also we are proposing the system which recover data even if any distributed storage site fails with the help of Shamir scheme and database.

TABLE 1
Data Table

File	Duplicate Check	Required storage space	Required Bandwidth	Saved space and bandwidth
Abc.txt	Unique	100%	100%	0%
Xyz.txt	Duplicate	0%	0%	100%
Ping.jpg	Duplicate	0%	0%	100%
Info.doc	Half duplicate	50%	50%	50%
Ring.avi	Duplicate	0%	0%	100 %

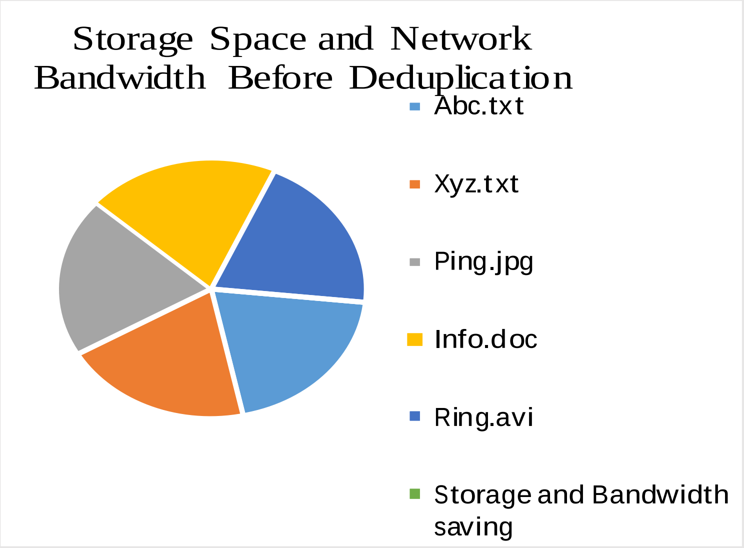

Fig 3: Storage Space and Network Bandwidth Saving before deduplication

Fig 3 pie chart has been drawn by considering values in Table 1. As shown in Fig 3, storage space and network bandwidth saving is not done here, because each file is taken as unique file and stored on the cloud. Here deduplication concept is not considered before saving any type of file. Due to which wastage of storage space and network bandwidth has been done. To overcome this issue proposed system given one of the best solution with maximum security. Fig 4 shown that with the help of proposed system one can save maximum amount of storage space and network bandwidth. Here proposed system considered duplicate and half duplicate concept to store a file. As shown in fig 4, Xyz.txt, Ping.jpg and Ring.avi is not stored actually on the cloud because those files are stored before on the cloud. Those files are duplicate files. Info.doc is half duplicate means some part of the file is already present, so unique part only stored on the cloud. Such a system is useful not only for cloud service provider but also normal user. Because it saves storage space of cloud service provider and bandwidth of user.

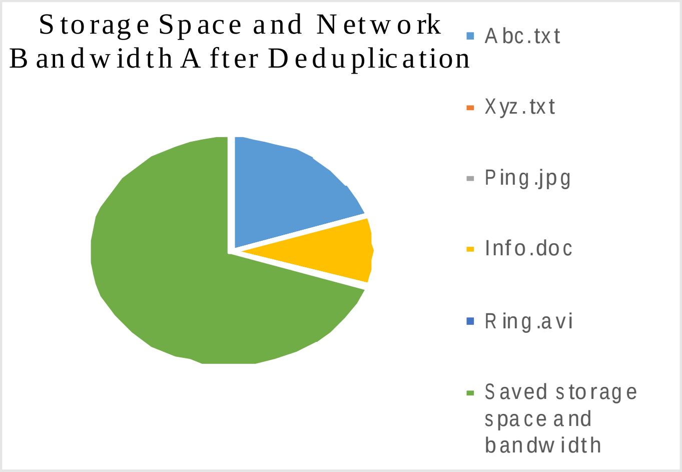

Fig 4: Storage Space and Network Bandwidth Saving after deduplication

VI. Conclusion and Future Work

We propose the distributed deduplication system to improve the reliability, confidentiality and integrity of the user data without using traditional encryption mechanism. Our construction is proposed to work on the file-level and fine-grained block level data deduplication to minimize the storage space and bandwidth. We implements our deduplication system using the Shamir secret sharing scheme for storing the user data. In secret sharing, Secret is divided into multiple parts or shares, to reconstruct the secret some parts of secret must be available. Single share of secret is difficult to predict the original secret. Proposed system achieves confidentiality, Integrity and reliability by removing various attacks and provides secure distributed deduplication system in the cloud environment.

This mechanism achieves strong security against insider and outsider attack, minimum storage space as well as saves network bandwidth. This mechanism itself is a cloud infrastructure with great performance as compared to the traditional deduplication and can be applied to the current cloud service provider to achieve the various feature implemented in the proposed mechanism. The proposed mechanism gives minimum overhead as compare to the traditional deduplication system.

Currently we are targeting to use fixed size block level deduplication. In future we can work on to use variable sized blocks deduplication which has better disk space utilization but with more complex computation. We can also focus on improvement in Shamir's secrete algorithm or look for fitment of other secrete sharing algorithm. Also we can move for the Recovery of failure storage site with the help of database and Shamir secret sharing scheme instead of traditional recovery or backup methods (RAID or any other).

References

1. Jin Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang Senior Member, Mohammad Mehedi Hassan Member, and Abdulhameed Alelaiwi Member, "Secure Distributed Deduplication Systems with Improved Reliability." In IEEE Transactions on Computers, Volume: PP , 2015.
2. Deepak Mishra, and Dr. Sanjeev Sharma, "Comprehensive study of data deduplication." in International Conference on Cloud, Big Data and Trust . Nov 13-15, 2013.
3. Mihir Bellare, Sriram Keelveedhi and Ristenpart, "Message-locked encryption and secure deduplication." in EUROCRYPT, pp. 296312,2013.
4. J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou, "Secure deduplication with efficient and reliable convergent key . management" in IEEE Transactions on Parallel and Distributed Systems, vol.25(6), pp. 16151625,2014.
5. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage" in Technical Report, 2013.
6. M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage." in USENIX Security Symposium, 2013
7. Shamir, Adi, "How to share a secret." in Communications of the ACM 22 (11): 612613.
8. Xie Tao, Fanbao Liu, and Dengguo Feng , "Fast Collision Attack on MD5." in 2013.

9. Calvin Chan, and Hahua Lu , "Rabin's Method." Dec 2001.
10. Yevgeniy Dodis, Marisa Debowsky, "Exposure-Resilient Cryptography " in January 2007.
11. Aparna Ajit Patil, Dhanashree Kulkarni, "A Survey on: Secure Data Deduplication on Hybrid Cloud Storage Architecture " International Journal of Computer Applications (0975 8887) Volume 110 No. 3, January 2015.
12. https://en.wikipedia.org/wiki/Rabin-Karp algorithm.
13. Jean-Sebastien Coron, Emmanuel Prouff and Thomas Roche, "On the Use of Shamir's Secret Sharing Against Side-Channel Analysis ".
14. www.jpinfotech.org

BIOGRAPHY

Prerna Lahane student of ME Computer Engineering second year from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune.

Prof. Sarika Bodake is a faculty in the Computer Engineering from the college TSSM's Padmabhushan Vasantdada Patil Institute of Technology, Bavdhan, Pune, India..