A Survey of Secure Data Deduplication

Riddhi Movaliya
Department of Computer Engineering
PIET, Vadodara,
Gujarat, India

Harshal Shah
Department of Computer
Science and Engineering
PIET, Vadodara,
Gujarat, India

ABSTRACT
Now days due to advancement of storage technology and computer technology, larger fraction of data is being maintained in digitized form. The same data is stored over and over again, consuming unnecessary storage space on the disc. Deduplication is ideal for highly redundant operations like backup, which requires repeatedly coping and storing the same data. Data deduplication is one of the most alive topics in storage because it enables companies to save a lot of money on storage costs. For cloud provider it is very helpful because you can deduplicate what you store. Due to reduction in cost it is being more popular. This paper will briefly describe Data Deduplication and give a comprehensive survey.

Keywords
Data Deduplication, Cloud Computing, Cloud Storage, Data Security

1. INTRODUCTION
Cloud computing allows access to resources from anywhere and at any time through the internet. The main advantage of using cloud storage from the customer's point of view is that customers can reduce their expenditure in purchasing and maintaining storage infrastructure while only paying for the amount of storage requested, which can be scaled-up and down upon demand [4]. But it is also very true that cloud Storage is not infinite. Data deduplication is the best way to handle these data.

2. DATA DEDUPLICATION
Data Deduplication is rapidly growing technique now days especially in backup storage due to reduction in cost of storage. Data deduplication is very important in management of data because it will store only unique data among duplicate data copies. Data Deduplication is efficient technique to handle these large duplicate data. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy [7]. Unique Id of data copy would be generated using hash algorithm, and then would be used for comparison. Data deduplication can be target based and source based. In the target based Data Deduplication user will upload their data and deduplication will take place at target side. So target based approach can improve storage utilization but cannot save bandwidth as whole data needs to be transferred at target side. In Source based deduplication client will check at storage side whether the data copy already exists or not, that means deduplication will be perform at client side and then after only unique copy will be stored. So Source based approach can improve bandwidth as well as storage. There is also granularity based deduplication: 1) File level deduplication 2) Block level deduplication. In File level only unique copy of file will be stored and duplicate copy will be discarded. In Block level each file is divided in the blocks and then only unique block will be stored. Length of divided block can be fixed or variable. Level of deduplication in block level is more than file level deduplication that means deduplication ratio is high in block level deduplication. There are disadvantages and advantages to each approach. Deduplication can also be applied at byte level. The differences lie in the amount of reduction each produces and the time each approach takes to determine what's unique.

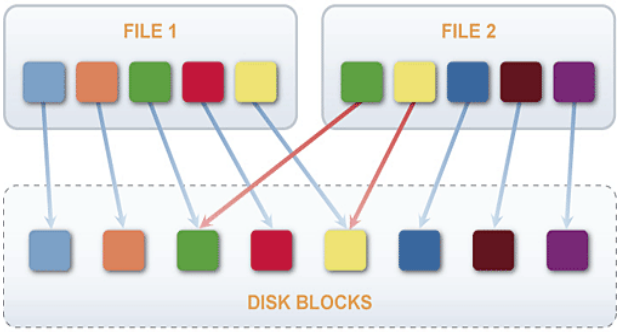

Fig 1: Data Deduplication

Data deduplication gives noteworthy benefits but security and data confidentiality is still sensitive issues. So, usual way to provide security is encryption. But there is confliction between data deduplication and encryption. Because it is possible that same plaintexts may lead to different ciphertexts. If one can realize data deduplication on ciphertexts, the cloud server must be able to identify all of the ciphertexts of the same plaintext [3]. One more thing needs to be there in data deduplication is authorized deduplication in which users would have set of privileges because in many of applications differential authorized duplication is needed. User can not check duplicate out of his privilege set. For example any role based application may require authorized deduplication.

3. COMPARING VARIOUS DEDUPLICATION APPROACHES

Table 1.Comparing various Deduplication approaches

Approach	Cost	Throughput	Used Bandwidth	Deduplication ratio	Required Storage
File Level Deduplication	Low	High	Low	Low	Medium
Block level Deduplication	High	Low	Low	High	Less
Source based Deduplication	Relatively Low	Medium	Low	Medium	Medium
Target based Deduplication	High	Medium	High	Medium	Medium

4. METHODS USED IN DATA DEDUPLICATION

4.1 Symmetric Encryption [7]

In Symmetric encryption common key will be used to encrypt or decrypt the information.

- $KeyGen_{SE}(1^{\lambda})$ -> k is the key generation algorithm that generates k using security parameter 1^{λ};
- $Enc_{SE}(k, M)$ -> C is the symmetric encryption algorithm that takes the secret k and message M and then outputs the Ciphertext C; and
- $Dec_{SE}(k, C)$ -> M is the symmetric decryption algorithm that takes the secret k and Ciphertext C and then outputs the original message M.

4.2 Convergent Encryption [7]

In convergent encryption secret key will be derived as hash value of plaintext. So the same plaintext will lead to the same cipher text. In addition tag is also derived to detect the duplicate.

4.3 Proof of Ownership

This will be used to prove ownership of data by user to the storage server. It is a kind of interactive algorithm.

4.4 Identification Protocol

There will be two phase: 1) Proof 2) Verify. In Proof user can give his identity by some secret credentials and in Verify stage verifier will verify with public information of input credentials.

5. LITERATURE SURVEY

5.1 DupLESS: Server-Aided Encryption for Deduplicated Storage [1]

DupLess Server-Aided Encryption for Deduplicated Storage provide simple storage interface. Cloud Storage provider like Dropbox, Mozy, and other providers can use deduplication technology to save space by storing single copy of data. Message lock encryption is used to resolve the problem of clients encrypt their file however the saving are lock. Dupless is used to provide secure Deduplicated storage as well as storage resisting brute-force attacks. Clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol in dupless server. It allows clients to store encrypted data.

Characteristics:

- More Security
- High Performance
- User friendly

Drawback:

- Operations are time consuming

5.2 Secure Client Side Deduplication Scheme in Cloud Storage Environment [2]

Ensure better confidentiality towards unauthorized users by cryptographic usage of symmetric encryption used for enciphering the data file and asymmetric encryption for meta data files. Data access is managed by the data owner by providing two level of access control Only authorized user can decipher the encrypted file.

5.3 Hybrid Data Deduplication in Cloud Environment [3]

A message in the proposed hybrid data deduplication mechanism consists of a triple of blocks, i.e., (check block, enabling block, and cipher block). The check block can be used to check the repetition of encrypted files. A session key is used to encrypt data is stored in the enabling block. The cipher block contains the encrypted data that are encrypted with the session key.

Characteristics:

- Easy to implement
- Provides transparency

5.4 Dynamic Data Deduplication in Cloud Storage [4]

Day by day popularity of cloud storage services has been increasing. User can store their data on the cloud storage. A dynamic data deduplication scheme is used to achieve a balance between changing storage efficiency and fault tolerance requirements. This will also improve the performance in cloud storage systems. We dynamically change the number of copies of files according to the changing level of Qos.

5.5 Verifiable Data Deduplication Scheme in Cloud Computing [5]

Two servers will be there S1 and S2. S1 will store data and S2 will be used to verify the deduplication process by S2. In this model client cannot be cheated by the wrong response by server. So, verifiable deduplication is ensured.

5.6 Twin Clouds: An Architecture for Secure Cloud Computing (Extended Abstract) [6]

The client communicates with the Trusted Cloud over a low bandwidth, secure channel. The two clouds are connected with an insecure, high bandwidth channel. The Commodity Cloud further provides untrusted storage. a Trusted Cloud is used as proxy between the client and the Commodity Cloud. Trusted Cloud provides an interface for secure storage and

computations to the client while abstracting from the service provider's cloud infrastructure.

5.7 Hybrid Cloud Approach for secure Authorized Deduplication [7]

Characteristics:

- Differential authorization
- Authorized duplicate check
- Unforgeability of file token/duplicate-check token
- Indistinguishability of file token/duplicate-check token
- Data confidentiality

Table 2. Comparison of various methods

Paper	Feature	Result
DupLESS: Server-Aided Encryption for Deduplicated Storage [1]	Space saving High performance	Simple Storage
A Secure Client Side Deduplication Scheme in Cloud Storage Environments	Access control Privacy	Ensure better confidentiality
Hybrid Data Deduplication in Cloud Environment [3]	Easy to implement Provide transparency	Data can be store as per requirement either in encrypted or un-encrypted area
Dynamic Data Deduplication in Cloud Storage [4]	Improve Storage efficiency	Maintaining redundancy for fault tolerance
A Verifiable Data Deduplication Scheme in Cloud Computing [5]	Storage saving	Verifiable data deduplication
Twin Clouds: An Architecture for Secure Cloud Computing (Extended Abstract) [6]	Secure computation Store large amount of data Secure execution environment	Client uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries.
A Hybrid Cloud approach for secure Authorized Deduplication	Differential authorization Authorized duplicate check Unforgeability of token	Reduce storage space and save network bandwidth

6. CONCLUSION

Cloud Computing is warm topic in IT industries for research. As in the starting of paper issues of ever increasing data is familiarized and it is very clear that data deduplication is the one of efficient ways to handle repeated data. But again data security issue is known. To better protect data, authorized deduplication is efficient way that provides various privileges to users by private cloud. In future block level approach can also be used. This paper would be helpful to new researcher who wants to study on secure deduplication.

7. REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.

[2] Nesrine Kaaniche, Maryline Laurent "A Secure Client Side Deduplication Scheme in Cloud Storage Environments" IEEE Transactions on Mobility and Security (NTMS) in Cloud Computing, Issue Date:April.2.2014

[3] Chun-I Fan, Shi-Yuan Huang, and Wen-Che Hsuz" Hybrid Data Deduplication in Cloud Environment" 978-1-4673-2588-2/12/$31.00 ©2012 IEEE

[4] Waraporn Leesakul, Paul Townend, Jie Xu" Dynamic Data Deduplication in Cloud Storage" 2014 IEEE 8th International Symposium on Service Oriented System Engineering

[5] Zhaocong Wen, Jinman Luo, Huajun Chen, Jiaxiao Meng, Xuan Liand Jin Li "A Verifiable Data Deduplication Scheme in Cloud Computing" 2014 International conference on Intelligent networking and collaborative System. 978-1-4799-6387-4/14 $31.00 © 2014 IEEE

[6] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, "Twin clouds: An architecture for secure cloud computing," in Proc. Workshop Cryptography Security Clouds, 2011, pp. 32–44.

[7] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou "A Hybrid Cloud Approach for Secure Authorized Deduplication" IEEE Transactions On Parallel And Distributed Systems, Vol. 26, No. 5, May 2015

[8] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, pp. 29–40.

[9] D. Ferraiolo and R. Kuhn, "Role - based access controls," in Proc. 15th NIST - NCSC Nat. Comput. Security Conf., 1992, pp. 554–563.

[10] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

[11] J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," IACR Cryptology ePrint Archive, 2013:149, 2013.