Secure Deduplication Techniques: A Study

Vruti Satish Radia
ME – CE Student
PIET, Vadodara, Gujarat, India

Dheeraj Kumar Singh
Assistant Professor
PIET, Vadodara, Gujarat, India

ABSTRACT
Currently usage of cloud storage is increasing and to overcome increasing data issue, Data deduplication techniques are used. Moreover the Cloud storage service is provided by third party cloud providers thus security of data is needed. Data Deduplication techniques cannot be applied directly with security mechanisms. Thus here in this paper we would be discussing data deduplication techniques along with securing techniques thus forming secure deduplication.

Keywords
Data Deduplication, Cloud Storage, Data Security

1. INTRODUCTION
After the rapid development of cloud computing, users and enterprise would like to back up their data to cloud storage. According to prediction given by International Data Corporation the digital data will exceed 44 Zeta Bytes in 2020 [3] [6]. The development of cloud storage encourage the service provides to make the data storage service been outsourced to third-party cloud providers [2]. Management of ever increasing data over the cloud storage is a important issue to be looked upon. With the explosive growth of digital data, deduplication techniques are used widely to backup data and minimize storage and network overhead by detecting and eliminating redundancy among data [1].

2. DATA DEDUPLICATION
Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth [1]. When a data is uploaded its hash value is formed and then compared with the existing hash value, if duplicate value is found then that data is not uploaded and is replaced with pointer to the unique data else if no duplicate is found the data is uploaded to the server.

Various benefits of Data Deduplication technologies are

- Increases network efficiency
- Lower storage space requirements
- Storage cost is reduced
- Reduced upload bandwidth

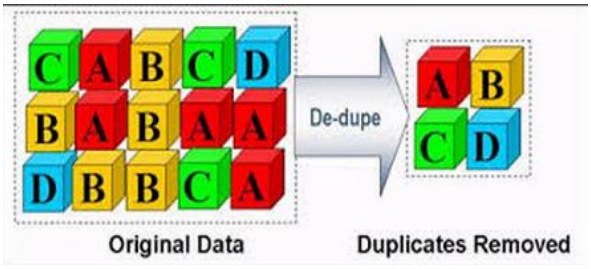

Fig 1.1 Data Deduplication Example

3. VARIOUS DATA DEDUPLICATION APPROACHES
Data Deduplication can be applied in various forms as follows:

- Based on Granularity
  - File Level Deduplication
  - Block Level Deduplication
- Based on Time of Application
  - Inline Deduplication
  - Post Process Deduplication
- Based on point of Application
  - Source Based Deduplication
  - Target Based Deduplication

A. File Level Deduplication Approach
File Level Deduplication is also referred to as single-instance storage (SIS). It compares a file to be backed up or archived with those already stored, by checking its attributes against an index. If the file found to be unique then it is stored and the index table is updated, but if the file is not unique then a pointer to the presented file is stored. The outcome is that only one occurrence of the file is saved and consecutive copies of the files are replaced with a pointer to the original file. [9]

B. Block Level Deduplication Approach
Block-level data deduplication operates on the sub-file level. As per the name, the file is normally broken down into segments, chunks or blocks that are checked for redundancy vs. previously stored information. [9]. Block Level Deduplication is further divided into Fixed Chunk Level Deduplication and Variable Chunk Level Deduplication. In Fixed chunk level deduplication the blocks are divided into fixed chunk size of say 4 KB, 8KB and so on. Then check for deduplication. And in Variable block size the blocks are divided into various size blocks and then checked for Deduplication.

C. Inline Deduplication Approach
Deduplicating the data before it is written to disk thus it reduces the storage requirement. The inline deduplication only checks the incoming raw blocks and it does not have any knowledge of the files. This forces it to use the fixed-length block approach. Extent of deduplication is less, and only fixed-length block deduplication approach can be used [8].

D. Post Process Deduplication Approach
In this approach data is first written to the storage device and then checked for deduplication. It can be applied on file-level or sub-file levels. Whole file data checksum can be easily compared with the existing checksums of previous backuped files and thus full file level duplicates can be eliminated easily [8].

E. Source Based Deduplication Approach

Deduplication is applied when data is on the source i.e. when data is created. [8] Then the non-duplicate data is backuped to the cloud. It helps in better and optimized utilization of resources. It is also helpful in incremental backup of new blocks in the user's instances.

F. Target Based Deduplication Approach

Deduplication occurs after data is been stored. Process of removing Deduplication occurs when data was not generated at that location [8]. User is unaware of the deduplication process occurrence. Thus this approach helps in storage utilization but does not helps in saving upload bandwidth [8].

Table 1 Comparison of various Data Deduplication Approaches

Deduplication Approach	Bandwidth Utilization	Storage Utilization	Throughput	Deduplication Ratio	Efficiency	Cost
File Level	Low	Medium	High	Low	Less	Low
Block Level	Medium	High	Low	High	High	Medium
Source Based	Low	Medium	Medium	Medium	Medium	Low
Target Based	High	High	Medium	Medium	Medium	High
Inline	Low	Low	Low	Low	Medium	Low
Post Process	High	Low	Medium	High	High	High

4. LITERATURE SURVEY

A Secure Deduplication with Efficient and Reliable Convergent Key Management [2]

It addresses the problem of achieving efficient and reliable key management in secure Deduplication. It implements Dekey using the Ramp secret sharing scheme that enables key management to adapt to different reliability and confidentiality levels [2]. It also uses key management and convergent encryption method to provide security to the data. Deduplication is done on file and block level deduplication.

- **Drawback**
  - Convergent keys are distributed across Multiple servers but the key servers are limited.
- Key space overhead needs to be taken care.

B SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management [3]

SecDep employs User Aware Convergent Encryption which helps in reducing computation overhead and it resists brute force attack [3] and it also uses Multi-Level Key Management which helps in reducing key space overheads (as file level key used for block level encryption and it splits file-level keys into share-level keys and distribute them to multiple servers to ensure security and reliability of file-level keys [3]

- **Drawback:**
  - Time overhead comes with multi-level key management can be reduced.

C Message-Locked Encryption and Secure Deduplication [4]

Message-Locked Encryption (MLE) is the cryptographic primitive, and the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure deduplication, an objective at present targeted by several cloud-storage providers. It provides definitions both for a form of integrity and privacy which is called tag consistency [4].

- **Drawback:**
  - Convergent encryption leads to significant number of convergent keys which are difficult to manage with increasing user.
  - Affected by brute-force attack.

D DupLESS: Server-Aided Encryption for Deduplicated Storage [5]

Dupless is an architecture that provides [5]

- secure deduplicated storage resisting brute-force attacks
- Clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It allows clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and however achieves strong confidentiality guarantees.
- Author shows that encryption for deduplicated storage can accomplish performance and space savings close to that of using the storage service with plaintext data.

- **Drawbacks**
  - Get and Put operations are time consuming.
  - Large computational overheads for chunk level

E Proofs of Ownership in Remote Storage Systems [7]

This author discusses solution based on Merkle Trees and specific encoding which identify attacks that exploit client side deduplication attempts to identify deduplication. Proofs-of-ownership (PoWs) concept in which Client proves to the server that it in fact holds the data of the file and not just some short information about it.

- **Drawback:**
  - Performance measurements indicate that this scheme incurs small overhead compared to naïve client side deduplication

F Secure Distributed Deduplication Systems with Improved Reliability [1]

Author has proposed a distributed Deduplication system with higher reliability (in storage over cloud) in addition to achieving confidentiality and integrity over data. Moreover distributed Deduplication system supports file-level Deduplication and block-level Deduplication [1]. For

reliability, Ramp Secret Sharing Scheme RSSS (2 algorithms Share and Recover are used) is used to provide better fault tolerance. For confidentiality, RSSS and Tag Generation Algorithm are used. For integrity – Message Authentication Code (MAC – use short cryptographic hash function) are used which also support process of secure deduplication system.

- **Drawbacks**
  - Only two types of attacks are considered. Type 1 Attack for Dishonest system and Type 2 attack for Collusion

5. CONCLUSIONS

In order to optimize upload bandwidth and storage space over cloud, Source Based Deduplication is one of the best options. Distributed deduplication systems helps to achieve security, confidentiality and reliability if data. Thus if both the approaches are combined then one can achieve better deduplication ratio along with reliability of data. Further deduplication algorithm can be modified to achieve better deduplication ratio.

6. REFERENCES

[1] Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Hassan and AbdulhameedAlelaiwi, "Secure Distributed Deduplication Systems with Improved Reliability", IEEE Transactions on Computers Volume: PP, Year – 2015

[2] Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management", IEEE Transactions on Parallel and Distributed Systems vol. 25(6), Year – 2014

[3] Yukun Zhou, Dan Feng, Wen Xia, Min Fu, Fangting Huang, Yucheng Zhang, Chunguang Li, "SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management", IEEE Mass Storage Systems and Technologies (MSST) 2015 31st Symposium, Year - 2013

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, ''Message-Locked Encryption and Secure Deduplication", Springer 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques at Athens Greece Proceedings, Year – 2013

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server-aided encryption for deduplicated storage", ACM SEC'13Proceedings of the 22nd USENIX conference on Security, Year - 2013

[6] "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm, April 2014, EMC Digital Universe with Research & Analysis by IDC.

[7] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems", ACM Conference on Computer and Communications Security, Year – 2011

[8] http://www.druva.com/blog/understanding -data-deduplication/

[9] http://searchdatabackup.techtarget.com/tip/The-pros-and-cons-of-file-level-vs-block-level - data-deduplication-technology

[10] http://www.cddatahouse.co.uk/solutions/dedupe-deduplication-of-data

[11] Neha Kaurav ," An Investigation on Data De-duplication Methods And it's Recent Advancements ", Proc. of the Intl. Conf. on Advances In Engineering And Technology - ICAET, Year – 2014

[12] Jyoti Malhotra, JagdishBakal, "A Survey and Comparative Study of Data Deduplication Techniques" Pervasive Computing (ICPC) 2015 International Conference IEEE, Year – 2015

[13] NagapramodMandagere, Pin Zhou, Mark A Smith, Sandeep Uttamchandani, "Demystifying data de-duplication", Proceedings of the ACM/IFIP/USENIX Middleware '08, Year – 2008

[14] IderLkhagvasuren, Jung Min So, Jeong Gun Lee, ChukYoo, Young WoongKo,"Byte-index Chunking algorithm for data deduplication system", International Journal of Security and its Applications, Year -2013

[15] Cai Bo, Zhang Feng Li, Wang Can, "Research on Chunking Algorithms of Data De-duplication", Proceedings of the 2012 International Conference on Communication, Electronics and Automation Engineering Springer, Year – 2012