An Advanced Data Management Scheme for Secure De-duplication

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

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

ABSTRACT
Data De-duplication is one of the effective ways now a days employed in the cloud computing. Energy consumption ratio of IT companies is increasing day by day and reason for it is storing redundant copy of data. Data De-duplication is a cutting edge technology that is used to remove redundant copies of data and store only unique copy of data. It reduces network bandwidth and storage space as well. There are many Data De-duplication methods used however in proposed work main focus is on block level Data De-duplication along with security. In Block level Data De-duplication redundancy would be checked based on blocks of data.

Keywords
Cloud Storage; Security; Block-level Data De-duplication

1. INTRODUCTION
Now a day all the things are being computerized so why the storage of digital data is be raised. Due to large data storage persons are moving to cloud storage as it provides large amount of data storage and is easy to use with proper credentials and via internet servers. The main advantage of using cloud storage is that user can save cost by only having for the storage been used. But the issue arises when the repeated copy of stored data arrived. Progressive data may initiate problem for cloud storage which is finite rather than infinite. To make data management scalable in cloud computing, de-duplication has been a well-known technique and has attracted more and more attention recently [7]. But it is also very important to have secured data with benefits of data de-duplication.

2. DATA DE-DUPLICATION
Data De-duplication is a kind of method in which redundant data will be eliminated and solely distinctive data will be saved. Data De-duplication is the foremost topic in cloud storage due to which reduction in cost on storage is achieved. Cloud data storage providers like Google Drive[9], Mozy[10], Dropbox[11], and others use data deduplication. Sudden growing digital contents have demanded for new network and storage capacities along with its further economical use, so it desires to be solved. For the cloud provider it is essential to achieve deduplication on the data being stored to increase the utilization of current space for storage.

For cloud provider it is terribly useful as a result of user can de-duplicate what is being stored and can higher utilize the current space for storing. If storage is less than required hardware is additionally less, which suggests that cost on hardware and network prices will be saved.

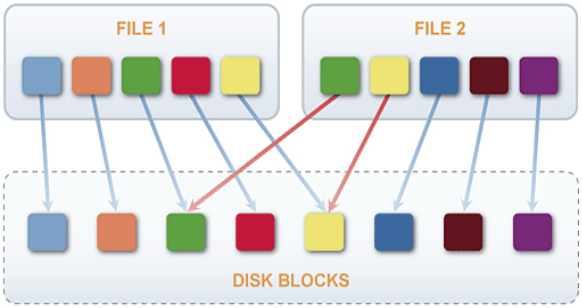

Fig 1: Data De-Duplication [8]

As shown in above Figure 1 Data De-duplication is the method in which only unique copy is being stored and for repeated copy pointer would be provided to previously arrived copy. To achieve Data De-duplication there are basic two approaches: File-level de-duplication and Block-level de-duplication. In File-level de-duplication only duplicate copies of the file are rejected based on hash value of the file. In block-level de-duplication file is further additionally divided into the blocks and then per block hash value is calculated and then based on that de-duplication will be performed. In block-level degree of de-duplication will be more than file-level de-duplication.

Although Data de-duplication brings many benefits, confidentiality and security alarms arise. Traditional encryption does not support Data de-duplication. Because in tradition encryption different users encrypt data with their own keys and it may leads to different cipher text. So it is possible that same data of different users will lead to different cipher text. Convergent encryption supports Data de-duplication and data confidentiality as well. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy[7]. So identical data copies will produce same convergent key and hence same cipher text will be produced.

3. RELATED WORK
Data De-duplication is the field of interest for many researchers. As Cloud storage needs to be efficiently used, Data De-duplication is one of the best ways. There are many data De-duplication systems proposed by researchers. M. Bellare et. al. [1] discusses method for de-duplicated storage at file-level with encryption which shows desired space saving close to storage service with plaintext data and improved message level encryption to deliver security against Brute force attack.

Secure client side De-duplication scheme in cloud storage is addressed by Nesrine Kaaniche et. al. [2] where data access is managed by providing two level of access control so only

authorized user can decrypt the encrypted data and higher confidentiality can be ensured towards unauthorized users.

Chan-I et. al. [3] introduces hybrid data de-duplication with mixture of encrypted and un-encrypted data on cloud storage having three blocks: Check block, enabling block and cipher block. The check block would be used for checking unique data. The session key is used to encrypt data is stored in enabling block and then in cipher block the encrypted data would be stored.

Dynamic data de-duplication has been presented by Waraporn Leesakul et. al. [4] where cloud storage performance is being improved by using load balancer and redundancy manager and client side de-duplication is used. Zhaocong Wen et. al. [5] proposed a verifiable Data De-duplication scheme at file-level for image in which Storage server will store image data and Verifiable Server will verify de-duplication process.

S. Bugiel et. al. [6] presented twin cloud used for secure outsourcing of data and it provides secure execution and computation environment. Jin Li et. al. [7] introduced authorized de-duplication using hybrid cloud and proposed Differential duplicate check so user can check duplicate based on privilege sets. Public cloud is used for data storage and private cloud is used as secure execution environment and provides Unforgeability and Indistinguishability of file token. Vidya Maruti et. at. [12] addressed authorized De-duplication where proof of ownership is set at the time of file upload to decide the access privilege to the file.

4. PROPOSED WORK

In proposed work block-level de-duplication is going to be used. When user wants to upload or download the file from cloud storage at that time first user request to cloudserver and only authenticated user can access cloudserver. When user uploads data, encryption is performed on data and then checked for file-level de-duplication. If no duplicate is found then further block-level de-duplication would be performed. Because at final goal attempt to reduce the storage, which is very massive potential. If only few bytes of documents are going to be changed then only changed blocks would be stored that means no need to store entire file again. This manner makes block-level de-duplication more efficient though it takes more processing power. Security is also being considered as encrypted data will be stored on cloudserver. The goal of proposed work is to reduce storage utilization and increase data de-duplication with security.

A key will be used for encryption and decryption. For encryption and decryption main three functions are there:

- KeyGen(M) -> K : To generate key K from Data copy M
- Enc(K, M) -> C : It takes key K and Data copy M as input and generates Ciphertext C as output
- Dec(K, C) -> M: It takes key K and Ciphertext C as input and generates Data copy M as output

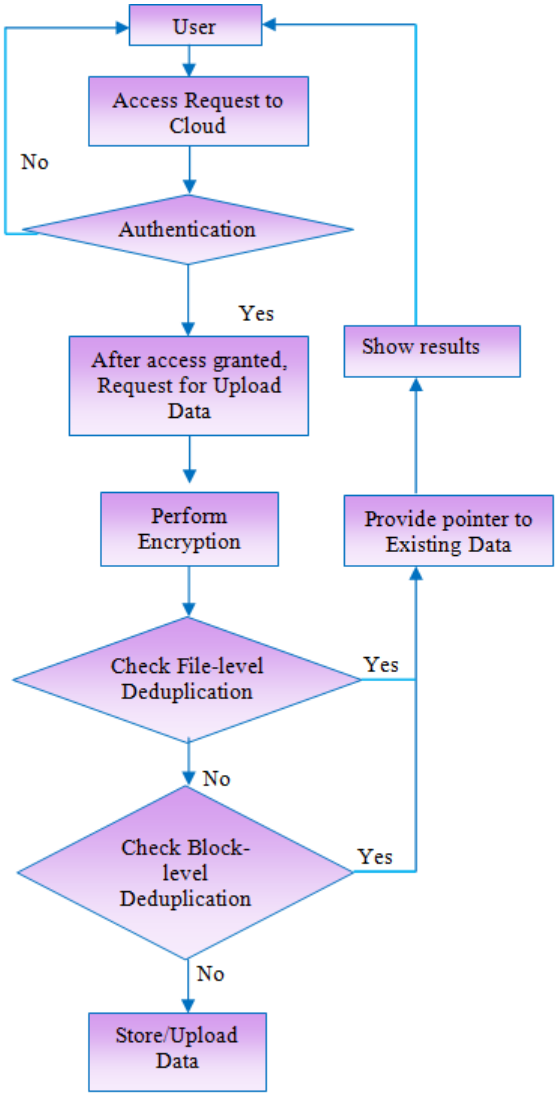

Fig 2: Proposed work flow

How proposed system is going to work is shown below:

Step –1 Start

Step –2 User request access to cloud

Step –3 Cloud authenticate user and grant access to cloud

Step –4 If true then

Request for upload file

Perform encryption on data

Else

Go to step 2

Step –5 Check file level Deduplication

If duplicate file then

Provide pointer to existing data

Show details

Else

Check Block level Deduplication

If duplicate block then

Provide pointer to existing data

Show details

Else

Store/Upload data to cloud

Step –6 End

Data De-duplication is the best way to handle this much data as it contains repeated copy of data. This method is very much convenient for the systems where backup is going to be taken which requires storing unique and repeated copies of data as well.

5. RESULT

A prototype is conducted using Java and Rackspace cloud. Various inputs are taken. Figure 3 shows duplication detection in File-level and Block-level when various files are uploaded on cloudserver.

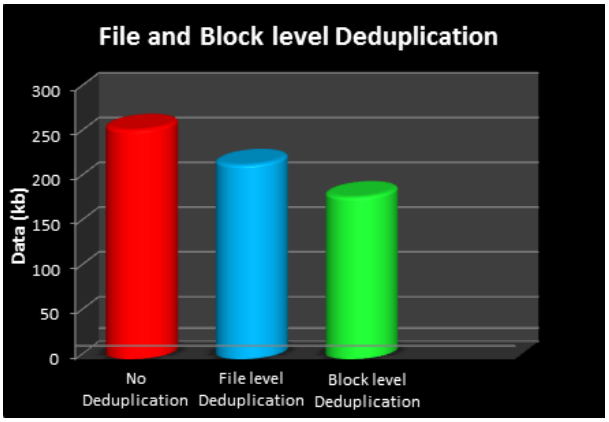

Fig 3: File and Block-level De-duplication detection

6. CONCLUSION

Data De-duplication is a useful technique for eliminating duplicate copies of data in the cloud. In this paper an approach is presented with the block-level De-duplication. For security concern regarding data encryption technique is used and regarding user cloud based authentication is used. Though in future byte-level De-duplication can be explored, which can further reduce storage space.

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] https://www.starwindsoftware.com/data-deduplication

[9] GoogleDrive : https://www.drive.google.com

[10] Mozy : https://www.mozy.com

[11] Dropbox : https://www.dropbox.com

[12] Mane Vidya Maruti, Mininath K.Nighot," Authorized Data Deduplication Using Hybrid Cloud Technique" 978-1-4673-6817-9/15/$31.00 ©2015 IEEE