RESEARCH ARTICLE OPEN ACCESS

A Proposed Approach for Cloud Storage Optimization with Authorized Byte level De-duplication

Ruchi Agrawal [1], Prof. D. R. Naidu [2]
Research Scholar [1], Assistant Professor [2]
Department of Computer Science and Engineering
SRCOEM, Nagpur
India

ABSTRACT

Now days, with increasing the population of whole world, the cloud based storage space popularity is also increasing rapidly. The main advantage of using cloud storage is, customer can reduce their expenditure for purchasing and maintain the storage area, but they have to use the cloud in optimized manner. In cloud, with data de-duplication strategy, the storage optimization can be achieved. It will be reliable to support dynamic datasets in the cloud. In this paper, Storage Optimization Algorithm is proposed for authorized de-duplication. In the SOA algorithm byte level de-duplication is performed with CPABE encryption technique for authorized de-duplication. To upload any file in storage, check its uniqueness and encrypt that. For uniqueness, file is divided into chunks as character by character. Then apply unique hash value to each chunk. By matching with hash, chunks are divided into unique or duplicated category.

For encryption, first public key is generated by KGC (Key Generation Center) and private key is generated by the user. Each unique chunk is now encrypted with public & private keys. The encrypted file is stored in public cloud and it's both keys are stored at private cloud. The duplicated chunks are referred to previously store unique chunk.

To download the file, concept of locality locates the related chunks via the reference pointer and decrypts all chunks then combines them into a single file. All of these aiming to improve storage efficiency and maintaining the reliability of cloud.

Keywords: - Byte level De-duplication, CPABE, Chunking, Key Generation Center (KGC).

I. INTRODUCTION

Cloud computing has recently as a popular business model for utility the computing system. The concept of cloud is to provide computing resources as a storage service on demand to customers over the Internet. The concept of cloud computing is similar as grid computing, which aims to achieve resource virtualization. In cloud computing SAAS (Storage-as-a-service) is one of the services that has been increasing popularly. It is an important for business storage at low cost. Cloud computing provides a large storage in various areas as government, enterprises, and use for storing personal data on cloud. Here user can access and share different resources over cloud. The large amount of storage space and its security issues are important concern in cloud computing. The major critical challenge of cloud storage is, to management of ever-increasing volume of data and its security. Data de-duplication is most important technique to improve performance with reliability, scalability & storage problem and has attracted more attention recently. Data de-duplication and other methods of reducing the storage consumption plays an important role in managing today's explosive growth of data. This paper will explore the significance of de-duplication ratios related to specific capacity optimization techniques with storage management. Optimizing storage capacity is beneficial for process improvement, cost savings and risk reduction. Increasing the efficiency and effectiveness of their storage environments helps companies remove constraints on data growth, improve their service levels, and better leverage the increasing quantity and quality of data to improve their competitiveness. It is an important technique for data compression, it simply avoid the redundant copies of data and store only single copy of data. It is a space-efficient method for storage, has gained increasing attention and popularity. It splits files into multiple chunks that are each uniquely identified by a SHA-512 hash signature, also called a fingerprint. It removes duplicate segments by checking their fingerprints. Data de-duplication not only reduces the storage space overheads, but also minimizes the network transmission of redundant data in the network system. One of the main challenges for centralized backup services based on de-duplication is the scalability of fingerprint-index search.

According to the data granularity, de-duplication technique can be categorized into two main categories: file-level and

block-level, which is nowadays the most common strategy. In block-based de-duplication, de- duplicate blocks of data that occur in non-identical files, the block size can either be fixed or variable. In file level approach duplicate files are eliminate. Again the block is also granular to finer level named as Byte Level De-duplication. In this the block are generated with each character (byte) and then check for the duplicity.

II. PROPOSED SYSTEM

Capacity optimization technologies can complement each other in two senses. First, they may be used to optimize different infrastructure elements based on their applicability. For example, software with source de-duplication capabilities may be used for remote office data protection, storage systems with de-duplication capabilities may be used as a backup target for enterprise data center data protection, and compression may be used to reduce the storage requirements for active data. Second, some of the techniques can be used together to achieve additive benefits. For example, compression can be applied to data that has been capacity optimized by other space reduction techniques to gain additional space savings.

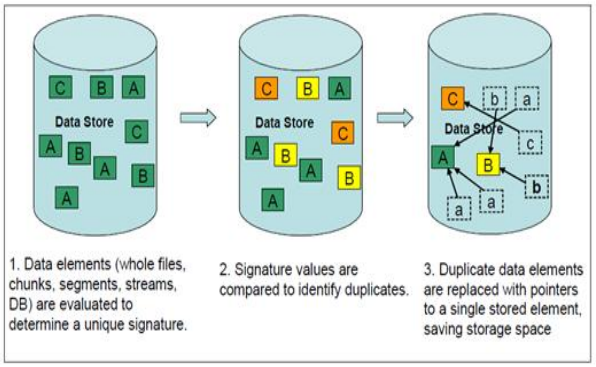


Figure: Byte level Data de-duplication

In our project, we implement a system that includes the public cloud and the private cloud and also the hybrid cloud which is a combination of the both public cloud and private cloud. In general, if we use the public cloud we can't provide the security to our private data and hence our private data will be loss. So that we have to provide the security to our data for that we make a use of private cloud also. When we use private clouds the greater security can be provided. In this system we also provide the data de-duplication, which is used to avoid the duplicate copies of data. User can upload and download the files from public cloud but private cloud provides the security for that data. That means only the authorized person can upload and download the files from the public cloud. For that user generates the key and stored that key onto the private cloud. at the time of downloading user request to the private cloud for key and then access that Particular file.

The System Model: If the user wants to upload the files on the public cloud then user first encrypt that file with the keys and then at the same time also generate the keys from the key generation center for that file and sends that key to the private cloud for the purpose of security. In the public cloud we use one algorithm for de-duplication, which is used to avoid the duplicate copies of files that is uploaded in the public cloud. Hence it also minimizes the bandwidth which means we require the less storage space for storing the files on the cloud. In the public cloud any person that means the unauthorized person can also access or store the data so we can conclude that in the public cloud the security is not provided. In general for providing more security user can use the private cloud instead of using the public cloud. User generates the key at the time of uploading file and stores it to the private cloud.

When user wants to download the uploaded file, then he/she sends a request to the public cloud. Public cloud provides the list of files that are uploads the many user of the public cloud because there is no security is provided in the public cloud. When user selects one of the file from the list of files then cloud sends a message like enter the key. User has to enter the key for that file. When user enter the key the private cloud checks the key for that file and if the key is correct that means user is valid then cloud provides access to that user for downloading that file successfully. Then file from the public cloud decrypt, by using the key which is used at the time of encrypt and the file will be downloaded. In this way user can make a use of the architecture.

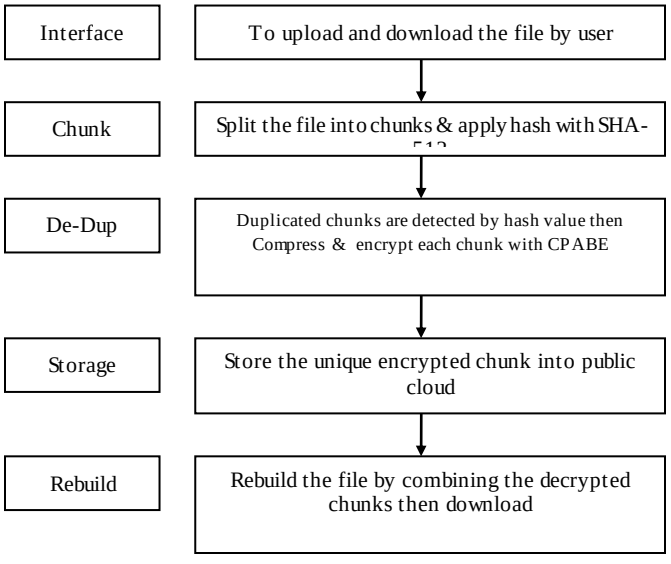


FIG: PROPOSED SYSTEM MODULES

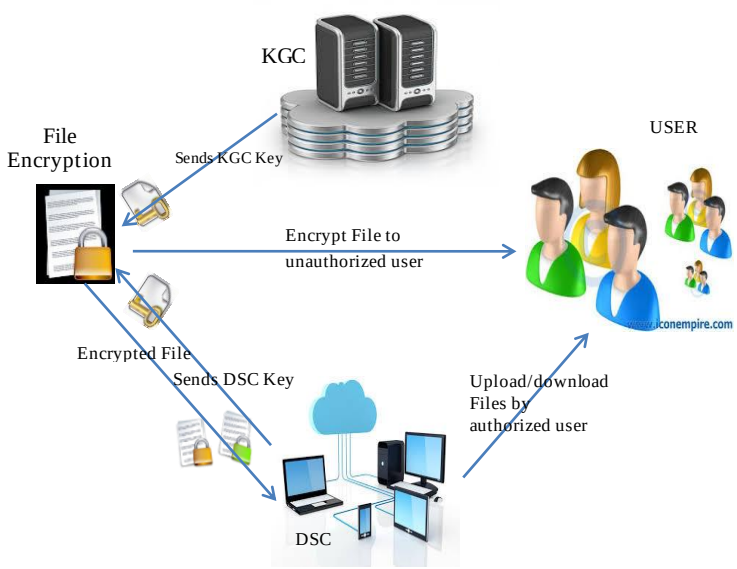
III. IMPLEMENTATION

STEPS:

1. Chunking In Bytes - Every file that user uploads will be of a different size. The uploaded whole file can be hashed at once but there is a disadvantage because if its sub-file is duplicated then it can't be determined. Therefore files are chunked into blocks. The unit chosen for chunking is a single character. The user input file is scanned character by character, and whenever scanned a character, the block is chunked.

2. Comparison and Storage - The application contains two databases for this purpose. In the first table, the hash codes are mapped against contents of original file. Here the hash code is set as a unique key constraint. If another file has same contents as of the original file, then the hash codes generated will be identical. But due to the unique key constraint, only unique hash codes are stored in the table.
In second table, the hash codes are mapped against the usernames and filenames. In case of identical hash codes, the duplicate hash codes are also be stored in the table because they are mapped against username and filename which are identical. This is done because it is easier to retrieve a file when a user requests to download a file.

3. **SECURITY** – CPABE IMPLEMETATION



ALGORITHM:

The proposed algorithm for De-duplication: STORAGE OPTIMIZATION ALGORITHM,

A. Client Side
- Get the keys from KGC
- Select the file to upload and convert into bit-stream.
- Apply hash to each chunk (byte).
- Encrypt the chunk with public key.
- Send the file to store.

B. Server Side
- Hash Indexing.
- Pass the references.

C. Client Side
- Retrieve the chunks and decrypt.
- Rebuild the file.

In summarized manner, we can take any file as input. Then, we create the chunks and generate fingerprint or hash for each chunk by using SHA512 to check the efficiency of hashing algorithm for best result to our system. Then, we will implement the Elasticity techniques like if the interdependent chunks are present then we put the single chunks in memory and locate the related chunks i.e. the concept of locality of index.
We will maintain the indexes of each chunk with the relevant user and file and if any duplicate chunks found then we just keep the hash in database but not the chunks. Finally, we recreate the file by mapping all related chunks in database.
With the recent adoption and diffusion of the data sharing paradigm in distributed systems such as cloud computing; there have been increasing demands and concerns for distributed data security. One of the most challenging issues in data sharing systems is the enforcement of access policies and the support of policies updates. Cipher text policy attribute-based encryption (CP-ABE) is becoming a promising cryptographic solution to this issue.
For example, in figure there are two files named as morning.txt (having content GOOD MORNING) and evening.txt (having content GOOD EVENING) then before apply byte level de-duplication it having 24 chunks of each character, but after that we only have to store 10 unique chunks.

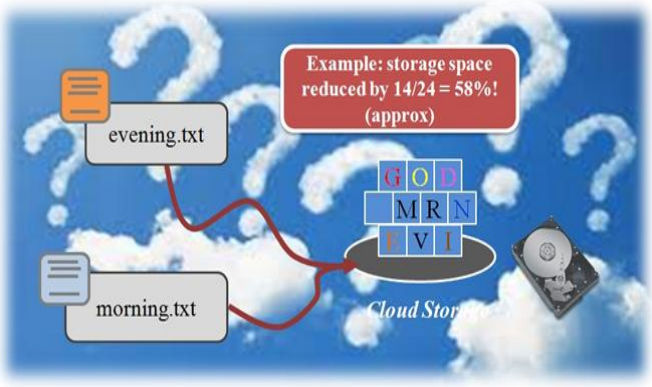


Fig. After de-duplication storage at Cloud

EXPECTED OUTCOME

From the literature survey, the result of various project will be find as,

- Data De-duplication ratio is, for de-duplication process over a particular time period is

Number of Bytes (Input) (Units In bytes)
Number of bytes (Output)

- Space reduction ratios as, typically depicted as "ratio: 1" or "ratio X,
  Storage Capacity of system (Units In bytes)
  Usable storage capacity.

File Name	Total No. of Chunks	User Storage	Cloud Storage
morning.txt	12	12	8
Evening.txt	12	12	8

For both files,

Total number of chunks: 24
Unique no of chunks: 10
For both files,
Total no of chunks: 24
At Cloud Storage: 10

IV. CONCLUSIONS

De-duplication aids in saving the storage space as well as bandwidth. This application helps in easy maintenance of data. The efficiency of the application is dependent on the amount of duplicated data present in the system. Therefore, the data de-duplication application is found to be efficient in eliminating redundant data.
All file formats are accepted and de duplicated successfully. Application willTest for text, docx, ppt, jpeg, png, sql, ppt, pptx, mp3, mp4, wmv file formats. Since the application will implement on cloud, it can accessed from anywhere, anytime. It gives us a flexibility on how to operate on our data. From future perspective, implementation on cloud gives us the edge to handle Big-Data. Since, the application is finding duplicate data, large amounts of data have to be processed.

REFERENCES

1. Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang, "Secure Distributed Deduplication Systems with Improved Reliability" *0018-9340 (c) 2015 IEEE.*
2. Jingwei Li, Jin Li, Dongqing Xie and Zhang Cai, "Secure Auditing and Deduplicating Data in Cloud 10.1109/TC.2015. 2389960, *IEEE 2015 Transactions* on Computers.
3. Prof. N.B. Kadu, Mr. Amit Tickoo, Mr.Saurabh I. Patil, Mr. Ganesh B. Divte, "A Hybrid Cloud Approach for Secure Authorized Deduplication" International Journal of Scientific and Research Publications, April 2015
4. "A SURVEY: DEDUPLICATION ONTOLOGIES, *International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 1, January 2015.*
5. Yufeng Wang, Chiu C Tan, Ningfang Mi "Using Elasticity to Improve Inline Data Deduplication Storage Systems" 2014
6. Waraporn Leesakul, Paul Townend, Jie Xu, "Dynamic Data Deduplication in Cloud Storage" *2014 IEEE 8th International Symposium.*
7. Jin Li, Xiaofeng Chen, Mingqiang Li, and Wenjing Lou. *"Secure De-duplication with Efficient and Reliable Convergent Key Management". In IEEE Transactions On Parallel And Distributed Systems*, Vol. 25, No. 6, June 2014.
8. Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, and Lei Xu," Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage" *IEEE ,May 2014*
9. Quanlu Zhang, Shenglong Li, Zhenhua Liy, Yuanjian Xingz, Zhi Yang, and Yafei Dai, "CHARM: A Cost-efficient Multi-cloud Data Hosting Scheme with High Availability" 10.1109, *IEEE 2014 Transactions* on Cloud Computing
10. Nesrine Kaaniche, Maryline Laurent, 'A Secure Client Side Deduplication Scheme in Cloud Storage Environments" *©2014 IEEE.*
11. "OpenfMRI Datasets," Accessed in 05/2013, https://openfmri.org/data-sets.
12. "FILE DEDUPLICATION WITH CLOUD STORAGE FILE SYSTEM, 2013 *IEEE 16th International Conference on Computational Science and Engineering.*
13. "COMPREHENSIVE STUDY OF DATA DE-DUPLICATION, *International Conference on Cloud, Nov 13-15.*
14. D. Harnik, O. Margalit, D. Naor, D. Sotnikov, and G. Vernik, "Estimation of deduplication ratios in large data sets," in *Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on*, 2012.
15. J. Min, D. Yoon, and Y. Won, "Efficient deduplication techniques for modern backup operation," *IEEE Transactions on Computers*, 2011.
16. Amrita Upadhyay, Chanchal Dhaker, Pratibha BR, Sarika Hablani, Shashibhushan Ivaturi, *Application of data deduplication and compression techniques in cloud design*, IIITB, April 2011

17. HiTech Whitepaper "Effective Data Deduplication Implementation 05 2011
18. M. Dutch, "Understanding data deduplication ratios," in *SNIA Data Management Forum*, 2008.
19. D. Geer, "Reducing burden via data deduplication," *Computer*, 2008.
20. J. H. Burrows, "Secure hash standard," DTIC Document, Tech. Rep., 1995.