International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 05 , May 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

MINIMIZING THE STORAGE COST BY MIGRATING THE AUTHORIZED DEDUPLICATED CONTENT INTO HYBRID CLOUD

M.Abinaya[1], Dr.G.S.AnandhaMala, Ph.D.,[2]

[1,2]Department of Computer Science and Engineering
Easwari Engineering College, Chennai, (India)

ABSTRACT

Cloud computing involves in deploying the groups of remote servers and software networks that allow data storage and also to access to computer services or resources in online. Nowadays increase in the data storage values there is more space needed.The data storage space are needed more and more for that data de-duplication is one of the important data compression techniques is usedfor eliminating the duplicate copies and it has been widely used in cloud storage to reduce the amount of storage space cost and save bandwidth. In this paper the main issues is to protect the confidentiality of sensitive datas while supporting de-duplication and also to utilize the cloud to serve the volatile requests with service response time guarantee all times, while incurring the minimum operational cost.

Keywords : Cloud Computing, Data De-Duplication, Confidentiality, Minimum Operational Cost.

I INTRODUCTION

Cloud computing technology is recently evolved computing terminology or metaphor based on utility and consumption of computing resources. Clouds can be classified as private, public or hybrid cloud. The criticisms about it are mainly focused on its social implication. It will take place when the owner of the remote server is a person or organization other than the user, as their interest may point in different direction, for example, the user may wish that his information is kept privatly, but the owner of the remote server may want to take advantage of their own business.

Fig.1. Architecture of Hybrid cloud computing

A condemnatory confrontation for cloud storage is to manage the aggregate volume of accumulating datas. In order to manipulates the data management, data compression techniques or data de-duplication techniques has been proposed. Since the amount of data storage is large, there may be large amount of duplicate copies. In order to avoid those unwanted datas and to save the storage space, a peculiar data compression techniqueshas been has been used to remove the redundant datas.Where the datadeduplication allow only the unique content to be stored in the cloud storage (Ref Fig 2). And also it increase the network bandwidth. Data deduplication is an approach to reduce the storage area that needs to store datas and the amount of data to be transfer over the network. The processes are partitioned into large data objects called chunks or blocks. For each blocks it generates the unique key by the cryptographic hash function called fingerprint. And then replace the duplicate chunks with their hash fingerprint value by the index lookup table in the cloud server. And finally transfers the unique chunks of data for the communication or to store datas in the cloud. Data deduplication that follows two approach for the implementation. They are finger prints and delta based deduplication approaches. The delta deduplication approach is the oldest method perform by chunking, but it is not searching similar, but necessarily identical data block. And in the fingerprint based data deduplication approaches, where the chunks are fingerprinted using the cryptographic hash functions. And then it can be search in the index lookup table in the cloud server for the identical file. If the file are identical file then it cannot be stored. If the file are not identical means then it can be stored in the index lookup table and then it can transfers over the network. The datas that are stored at the cloud backup can be secured. Because the personal data that cannot be visible to others and it cannot be easily downloaded. So the datas after deduplication that performs encryption and decryption process for secure backup.

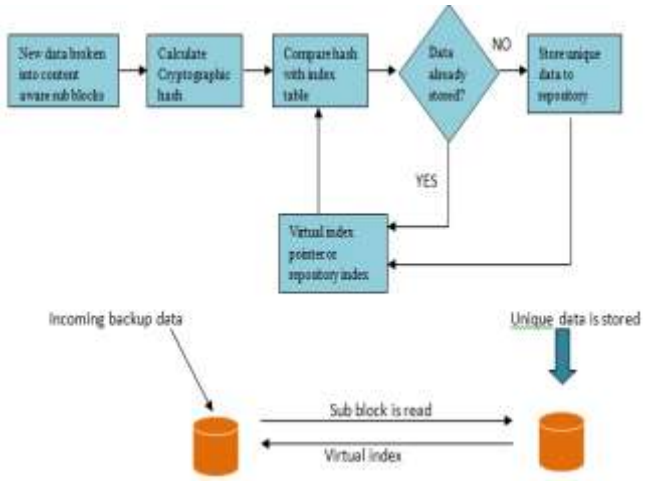

Fig 2: De-Duplication Process

II SYSTEM OVERVIEW

2.1 Authorized Duplication System

At a high level, our setting of interests is in Enterprise networks, consisting of group of affiliated clients who will use the S-CSP and store datas with deduplication technique. In this setting, Deduplication technique can be

frequently used in these setting for data backup and disaster recovery application while greatly reducing storage spaces in the cloud . Such system are widespread and are often moresuitable for the users to backup their files and the synchronizationapplication than richer storage abstraction.

There are three entities which define in our system as shown in figure 3, they are,

- Private cloud
- S-CSP in public cloud
- user

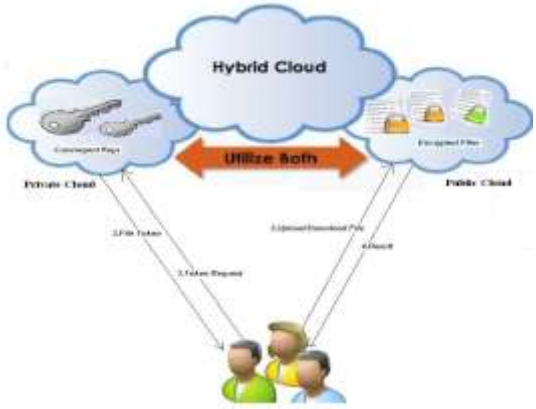

Fig.3. Architecture for Authorized Deduplication

De-duplication technique performed by S-CSP in the cloud, by checking if the contents of two files are same and it stores only one of them. Based on the set of privilege, the right to access a file is defined. The exact definition of the privilege varies across different application. Each privilege are represented in the form of a short message called token. Each file is associated with some file token, which denotes the tag with specified privilege. The user then compute and sends the duplicate-check tokento the public cloud for the authorized duplicate checks. If the file is duplicate, then all its block must be duplicates as well; otherwise, the user further performs the block-level duplicate checks and identifies the unique blocks that need to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

- *S-CSP*.The S-CSP in the cloud, provide the data outsourcing services andstore the datas on behalf of the user. To reduce the storagecost, the S-CSP eliminates the redundant datavia deduplication and keeps only unique data in the cloud storage.

  1. Start
  2. Get the unencrypted file tag for each file
  3. Accept the privileges based token from the user
  4. Validates the token and assert the privilege level
  5. Run deduplication checks only on the privileged files
  6. If the same tag is found along the privileged file, then

mark deduplication search as successful and grant
premission accessto the encrypted file.

7. Stop

- *Data Users.* A user is an entity who wants to outsource their data storage to the S-CSP and to access their data later. In a storage system that support deduplication, the user can only upload the unique data but does not upload any duplicate datas because to save the upload bandwidth, which may be owned by the same user or different user. In authorized de-duplication system, each user is issued with a set of privilege. Each file is protected with the convergent encryption keys and privilege keys to realize the authorized de-duplications with differential privilege.
- *Private Cloud.* Comparing with the traditional deduplication system in cloud computing, this is a new entity introduced for facilitating the users for secure usage of cloud service. Specifically, since the computing resources at datas user/owner side is restricted and the public cloud are not fully trusted in practice for storing datas, private cloud is able to provide data for the user/owner with an execution environments and infrastructure working as an interface between users and the public cloud. The private key for the privilege are managed by the private cloud, who answer the file token requests from the users.

1. Start
2. Authenticate the user based on their credential
3. Generate the unique privilege based token for each file
4. Return the token to the user for accessing the file
5. Stop

The public cloud doesn't have access to the user credentials. If compromised token generation key can be changed regardless of the user credential. Public cloud does not have an access to the decrypted files. Deduplication happens only on the privileged files.Hybrid cloud generally having twin clouds (i.e.,private cloud and public cloud). This architecture is used for data deduplication. For example, an enterprise might use a public cloud services, such as Amazon S3, for archived datas, but continue to maintain in-house storage for the operational customer data. Alternatively, the trusted private clouds could be a cluster of virtualized cryptographic co-processors, which are offered as a service by the third party and it provides the necessary hardware based security features to implement a remote execution environments trusted by the user.

2.2 Mathematical Model

Let S be the system that finds out duplicate copie of the file using the Authorized deduplication system in hybrid cloud. S={B,F,C,T,P,M,O}Where,

B={CBi,TBi,Pki}

F={F1,F2,F3,…Fn}

CBi= cipher block text

T = Token [16-Bit unique token]

P= Private Key (PKi)used for the encryption &decryption M=Metadata of files

O=Output consist reduce database sizes.

Following steps occur in the given proposed system architecture:

1. File F is divided into multiple block F= ΣBi, F = size (F) / 4096
2. KeyGen(1,λ)→k is key generating algorithm,generates secrete key using security parameter.

3. Enc (k,F)→C is encrypting algorithm, that takes the secrete keyk and then file and then F output the cipher text C.

4. Generate Token T for each blocks.

5. Dec(k,C)→F is Decrypting algorithm that take secrete key k and ciphertext C and then output the original file F.

6. Detect duplication.

Security Service generate TiBi Token on basic on Bi, If the same Bi comes in then it will generates the same TiBi. Then it will stores the TiBi to the Own Security Db. If file are found in database it generates response.

III PROPOSED METHODOLOGY AND DESIGN

3.1 Data De-duplication Mechanism

In storage server, De-duplication technique detects redundant data by creating cryptographic hash of the data to be stored. Hash is a fixed length representation used to detect duplication. Hashing reduces the complexity of comparing the two data chunks or data records because the size of the hash is much smaller when compared to the size of the datas. For each incoming records, server first calculates its hash signature value and search this hash signature is already present in the hash index which is present in the system. If the servers find that an entry is available for this signature in the hash index, servers only creates a reference for this redundant data, which points to the location of block which is already present on the disk. Otherwise servers store this record on the disk and add an entry for its hash signature in the hash index which is present in the cloud server , Refer Fig 4

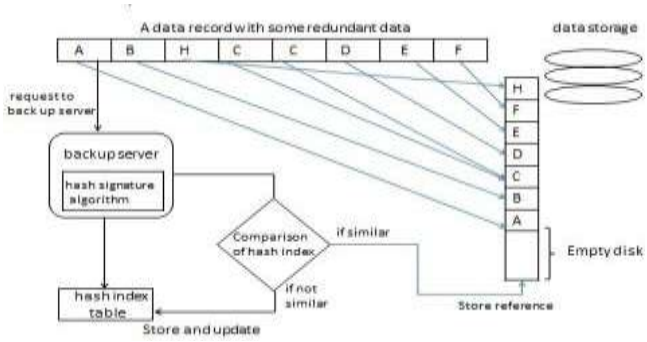

Fig 4. Data De-Duplication System

3.2 De-Duplication Types

3.2.1De-Duplication Based on Time

Data deduplication has been broadly classified into two types, they are

- In-line
- Post-process

These two classifications are described in following sections.

1) Post-process Data De-duplication: With this post-process de-duplication techniques, data de-duplication analysis are made after the datas are stored in storage device. Once the data are stored then only the process will be applicable. A benefit of using this post process is no one need to wait for the hash based calculation. The lookup is completed before storing the datas. The negative sides of this process isone may unnecessarily save the redundant data for a small time which could be an important issue if the system is near to full capacity[23] .

2) In-line De-duplication: With this inline de-duplication techniques,process are applied at the target devices. When the datas enter into the device in the real instance of time, the deduplication and hash calculation are performed. At this time if the device found a block that are already available in the system, in this case it will not store a new one rather then it will create a reference to the existing block. An advantage of using this inline data de-duplication is that it needs lesser amount of storage space as data is not redundant. On the other side because of the lookup and hash calculation takes a long time, it leads to slower data ingestion due to this decrement in throughput of backup of the device [23].

3.2.2 De-Duplication Based on Location

The location based de-duplication technique ,the de-duplication occur intwo locations. First process are applied where datas areproduced,known as “source de-duplication” and second process areapplied where the record are stored, referred as "target deduplication” [23].

1) Source versus Target De-duplication: When we describes the data de-duplication process for the backup systems and architecturethere are two kinds of de-duplication which can be applied.They are refer as the source based data deduplication and the target based data deduplication techniques.The data de-duplication process applied at the data source are knownas the source de-duplication. Source deduplication is one of the type of locationbased de-duplication. Source de-duplication process commonlyimplemented within the file systems directly. In such processa periodic scan are performed by file systems, in which deduplicationprocess will scan the new files and create hash value and then it compare these hashes with existing hash indexes ofrecords in the cloud server. If any the file or record found with the similar hash index, it will remove the copy of a file and this new file will pointto the old file with the reference. Duplicated copies are stored separately and incase of duplicate file , it is modified after sometimes , than by performing the copy to write, then another copy of modified file are produced. While applying de-duplication for backupthe files system causes the redundancy, the result with the biggerbackups rather than the source data. In target de-duplication redundant datas are removed at the secondary storage. Backup are store as the virtual tape library or the data storing repository are thegeneral type of backups which are provided in the target based data de-duplication. Choice of in-line or post process method dependson requirement at the target side storage .

3.3 De-duplication Level

3.3.1 Block Level De-duplication

In the block-level de-duplication technique is where the data block de-duplication method is applied. In this process the incoming data streams are divided into blocks or chunks, and then it is compared with the hash value

of the data block. After comparison, it determine whether it is same data which has been previously saved in the data blocks. If the hash value of the data block is unique , then it store this new block to the disk, and stores its identifier in the hash index, otherwise it store the reference to the same data blocks. It store a reference of a comparatively small size in place of the data blocks, rather than storing redundant data blocks again and again, hence a significant saving of disk storage spaces in the cloud server. Hash algorithm is used to examine and judge the duplicate data.(Refer fig 5).

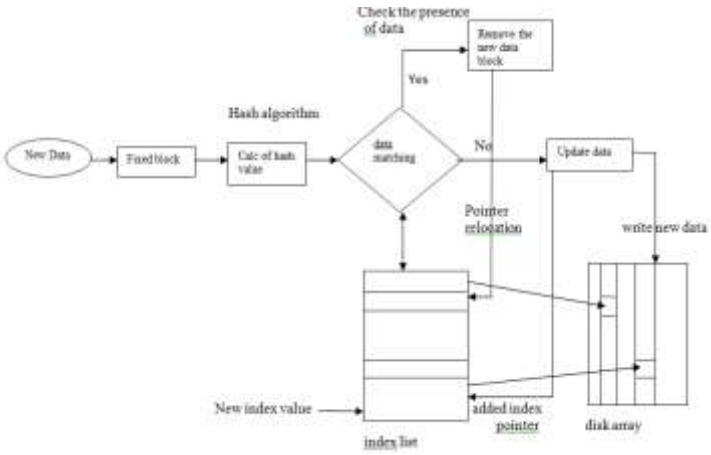

Fig 5. Block Level De-duplication

3.3.2 Types of Block Level De-Duplication

There are two types of Block Level de-duplication method , they are

A. Fixed Size Block De-Duplication

Fixed Block de-duplication method involve in determining a block sizes and segmenting the files/data into those block sizes which is mentioned. Then, those blocks are which are stored in the storage subsystem Suppose we take a fixed size 1 byte to divide an incoming file.

B. Variable Size Block De-Duplication

Variable Block de-duplication methominvolve in using the algorithms to determine the variable block size. Then the data are splitted, based on the algorithm's determination. Then, those blocks are stored in the subsystem.

3.3.3System Architecture

The main aim of designing this architecture is to design an improved techniques for the storage in Cloud computing. The overall architecture of the technique is shown in Figure 6.

Overall architecture is divided into four layers, they are:

Interface Layer - Interface layer provides the user interfaces to select the type/file for data deduplication and also to specify the split length.

Chunk Layer- Based on the split length mentioned the different Segments of the file are created. For these chunks hash value are computed by using the MD5 algorithm.

Deduplication Layer- It involves in the detection of the duplicate chunks by comparing hash value generated in chunk layer. Then the chunks are compressed after eliminating the duplicate value.

Storage layer - After eliminating the duplicate value, the compressed files are stored in the cloud server.

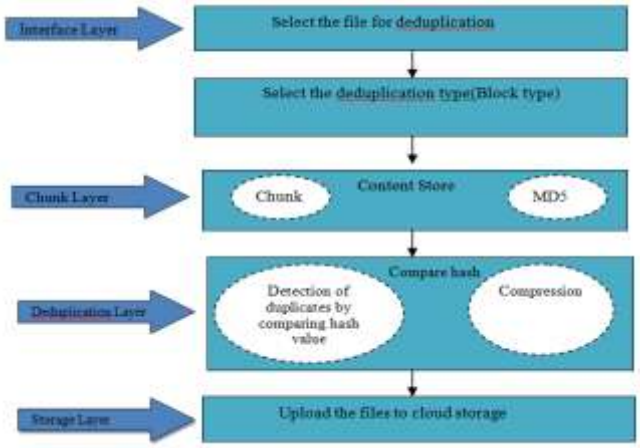

Fig 6: Overall Architecture of data de-duplicaion

IV RELATED WORK

Secure data Deduplication - With the advent of cloud computing, secure data deduplication has attracted much attention recently from research community. Bellare et al. [3] showed how to protect the data confidentiality by transforming the predicatable message intounpredicatable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek et al. [19] presented a novel encryption scheme that provides differential security for the popular data's and unpopular data's. For popular data that are not particularly sensitive, the traditional conventional encryption are performed. Another two-layered encryption scheme with stronger security while supporting deduplication is proposed for an unpopular data. In this way, they achieved better tradeoff between the efficiency and security of the outsourced data. Li et al. [11] addressed the keymanagement issue in block-level deduplication by distributing these key across multiple server after encrypting the files.Proof of ownership.Twin Clouds Architecture. Recently, Bugiel et al. [6] provided an architecture consisting of twin clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud.

V CONCLUSION

In this paper, the notion of the authorized data decompression technique was proposed to protect the data security by including the differential privileges of the users in the duplicate checks. We also presented several new de-duplication constructions supporting authorized duplicate check in the hybrid cloud architectures, in which the duplicate-checks token of files are generated by the private cloud server with private keys. The developed of these system must have a modest requirement, as only minimal or null changes are required for

implementing this system. The level of acceptance by the user solely depends on the methods that are employed to educate the users about the systems and to make the user familiar with it. Application of cloud computing theory is clear and carefully designed. Security model manipulate that our proposed system are secure in terms of outsider attack. With the advent of cloud computing technique, secure data de-duplication has attracted much more attention recently from the research community. We will try to implement the techniques of data compression, in order to avoid duplicates in videos and images.

REFERENCES

[1] A Hybrid Cloud Approach for Secure Authorized Deduplication Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou (2014)

[2] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010. [3] M.
Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart.Message-locked encryptionandsecurededuplication. InEUROCRYPT,pages296

[4] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009

[5] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.

[6] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[7] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In ICDCS, pages 617–624, 2002.

[8] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15th NIST-NCSC National Computer Security Conf., 1992.

[9] GNULibmicrohttpd.http://www.gnu.org/software/libmicrohttpd/.

[10] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491–500. ACM, 2011.

[11] J.Li,X.Chen,M.Li,J.Li,P.Lee,andW.Lou. Securededuplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[12] libcurl. http://curl.haxx.se/libcurl/.

[13] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[14] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–

446. ACM, 2012.

[15] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, ACM Symposium on Information, Computer and Communications Security, pages 81–82. ACM, 2012.

[16] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002.

[17] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui. A secure cloud backup system with assured deletion and version control. In 3rd International Workshop on Security in Cloud Computing, 2011.

[18] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-based access control models. IEEE Computer, 29:38–47, Feb 1996.

[19] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013.

[20] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008.

[21] Z. Wilcox-O'Hearn and B. Warner. Tahoe: the least-authority filesystem. In Proc. of ACM StorageSS, 2008.

[22] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.

[23] Q. He, Z. Li, and X. Zhang Data deduplication techniquesin Future Information Technology and Management Engineering (FITME), 2010 International Conference on, vol. 1, 2010, pp. 430–433.