Implementation Of Integrated Data De-duplication For Stress Reduction On Cloud

Shaikh Masira Maruf
Department Of Computer Engineering
Vishwabharti Academy's College Of Engineering
Ahmednagar(MH), India

Mr. Rathod V.U
Department Of Computer Engineering
Vishwabharti Academy's College Of Engineering
Ahmednagar(MH), India

***Abstract*—One of the most important issue in today's cloud computing is duplication for any organization, therefore this needs to be analyze to avoid the reparative files on cloud storage. Avoidance of the file is advantages the cloud size issue. To protect the confidentiality of sensitive data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. To better protect data security, on cloud storage. In this system we check the duplicate file on cloud storage also security apply using encryption. We use the AES encryption algorithm for encrypt the file simultaneously we check the duplicate file using the hashing algorithm. Also enhanced this system using recover option, cloud provide the deleted file backup on requesting. This paper study on the plain text as a input the system for checking the duplicate file, we next stage we analysis the encrypted file as a input to the system and find the duplicate file on the cloud storage.The data is store in a splitted form on the cloud.**

Keywords— Data Duplication, Authorized User Check, Splitting Data,Discretion, Cloud computing.

I. INTRODUCTION

Today's work is all dependant on the online and number of customer who are connected to the online cloud network for use the storage, access any resources, where data is stored in pools of storage which are generally hosted by third parties. Therefore in this case the cloud storage is increasing rapidly. The online storage provides users with multiple benefits, ranging from cost saving and simplified convenience, to mobility opportunities and scalable service. Above all properties used for customers to use and storage their personal data to the cloud storage. As per our analysis the volume of data menace size of the cloud storage capacity is expected to achieve 50 trillion gigabytes in 2020. The cloud storage system has been widely used in various organization, institution and corporate offices in the world; it fails to accommodate some main emerging needs such as the abilities of auditing integrity of uploaded data cloud files by clients and we detecting duplicated files by cloud servers. We generate the system and analysis both problems below. The first step to solve the problem is integrity auditing in the cloud computing. The local cloud is able to remove unwanted action clients from the heavy burden of storage management and maintenance. The online cloud data storage used for further access and the data is transferred via Internet and stored in an uncertain domain, not under control of the clients at all. All of these concerns originate from the fact that the cloud storage is highly susceptible to security threats from both outside and from the clients may be hidden by the uncontrolled cloud servers to maintain the reputation.

The most important thing is that for ordinary clients the data which is rarely accessed is deliberately deleted by the servers to maintain the cost and space. We considering the large size of the outsourced data files uploaded by user and the clients' constrained resource capabilities, the first problem is as how can the client efficiently perform periodical in verifications even without the local copy of data files. So we solve this problem using the detecting is secure de-duplication file on cloud storage. The remove increased volumes of data stored at remote cloud servers accompany the rapid adoption of cloud services. According to the last survey of EMC the most of the remotely stored files are deduplicated. Recently the 75% of the digital data is deduplicated. Due to this the term came that is deduplication in which the cloud servers just keep only one file and keeps the link of that file for the user's who wants the same file to store. Due to this it leads to a number of threats affecting the storage system, for example, a server telling the client that it does not need to send or store the file which is same as other user and it can be dangerous sometimes.. These attacks originate from the proof that client owns a file that totally uses static or we can say a hash code. Thus, the second problem is generalized as how can the cloud servers efficiently confirm that the client owns the uploaded file before creating a link to this file for him/her and security of the data on cloud.

II. LITERATURE SURVEY

A. Convergent Key

In this paper [1] tries to explain the concept of convergent encryption; the data uploaded is encrypted under a key derived by hashing the data itself. This convergent key is mainly used for encrypting and decrypting of a data uploaded. After key generation and data encryption. The used the encryption process where identical data copies will generate the same convergent key and the same cipher text. Thus convergent encryption allows to detect the data deduplication on cipher texts. The cipher texts can be decrypted by the corresponding data owners only with their convergent keys. Duplicate check authority is given only for a authorized user .a set of privileges is given based on their authorization during system initialization. This set of privileges specifies that which kind of users is allowed to perform duplicate check and access the files.

B. Policy Based Deduplication Scheme:

Policy based deduplication scheme in order to limit the ability and the knowledge needed for deduplication and to have a certain degree of security assurance they proposed a policy-based de-duplication proxy scheme to enable different trust relations among cloud storage components. The de-duplication proxy scheme will de-duplicate data of its registered users based on the capabilities it has received from the key center. The user can decide which data will be de-duplicated by submitting the tags of the data chunks.

C. Application –Aware Source Deduplication

In this paper [3] they proposed An Application -Aware source deduplication[AA] as the process of cloud back up system is increasing , deduplication can achieve high space efficiency and it reduces the cloud storage cost. The process of cloud efficiency becomes critical for cloud clients in the personal computing environment due to its limited system resources. To achieve high space efficiency EMC Avamar applies CDC based chunk level deduplication with high computational overhead and lookup overhead .In this AA-deduplication improves deduplication efficiency significantly by intelligent data chunking methods with application awareness. But exploits application awareness by limiting the search for redundant data to the chunks within the same kind of applications specified by the file format information. The direction for future work, they planned to investigate the secure deduplication issues in cloud backup services of personal computing environment and further explore and exploit index lookup parallelism by application aware index structure of AA deduple in a multi-core.

D. Lock Dependent Message

In this paper [4] they have explained how to provide security even for lock dependent messages. First approach is to avoid using tags that are derived deterministically from the message .to this end ,we design a fully randomized scheme that supports an equality testing algorithm defined on the cipher texts. They design an algorithm that encrypt message under a key that is highly correlated with message and still remain secure. Secondly the part of ciphertext that allows the equality test must not leak any information about message from an adversarial chosen min-entropy distribution even given the public parameters.

E. Deduplication Indexing

In this paper[5] they explained how to achieve storage efficiency, limited memory usage for deduplication indexing, and to achieve high throughput of multiple backup streams using RevDedup .Revdedup which removes duplicates of old backups and mitigates fragmentation of latest backups. RevDedup applies global deduplication over large size data units. To maintain high deduplication efficiency, it maintains the data placement as sequential as possible for the latest version, and removing any redundant data of old versions and referring it to the identical data of the latest version.

F. Security Issues

In this paper [6] They have explained about the number of security issues in cloud computing as it encompasses many technologies including network, databases, operating system, resource scheduling, transaction management, load balancing.

For example, the network that interconnecting the systems in a cloud has to be secured and mapping the virtual machine to the physical machines has to be carried out securely. Cloud service providers need to inform their customers the level of security that they provide on their cloud.

G. One Time Password

In this paper [7] they have explained about the how to design and implement a fast and secure data backup process. They used the technique of one time password which is easy for a admin to recognize the user if the OTP generated and received will be same then the user is same. The OTP is send on the registered account either on mobile phone or the email address,

III. PROPOSED METHODOLOGY

A. Architecture

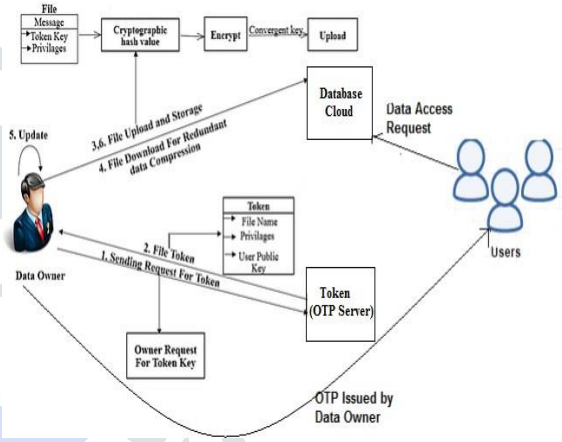

Fig 1. System architecture

- *User*:

  User is the person who is going to upload or download the data from the cloud. The user will be the registered one and his identity will b recognized to the admin. This is done to make a cloud secure.

- *Data Owner:*

  Data owner will create a account on the cloud the account will be approved by the admin .Once the data owner is registered he can upload or download a file from the cloud. While accessing the account the one time password (OTP) will be send to the data owner to verify if he is the right user. If it matches then further process is allowed or rejected. If user uploads a file the hash value will be generated and the hash value will be checked by the files on the cloud.

- *Data deduplication with secured manner:*

  While data uploading by user or data owner into public cloud the hash value will be identified of duplicate data will be notified by showing the warning pop message to users if the user wan to upload existing file again, still user wan to upload file the new file need to update with existing file. While user uploading data into cloud user can distinguish responsive and non-responsive data and can provide encryption for only sensitive data.While downloading data the splited data is merge and send to user.

B. Flowchart Of The System

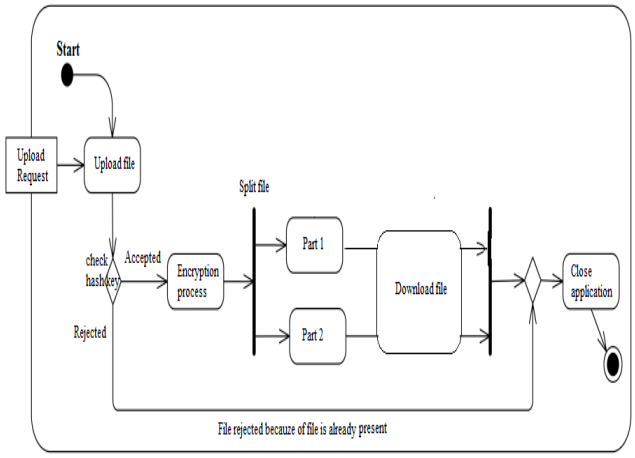

Fig 2: Flowchart Of the system

The above figure will explain the flow of the system the first of all the user need to register and upload the file on the cloud the hash key will be generated and when next the user will upload the file the hash value will be generated After the generation of hash key the file will be encrypted and splited the file in two parts. If the user need to download the file they need to login and download the file easily.

C. Algorithms

Input: Request to upload or download the file.

Output: Duplicated file will be rejected.

START
Step --1 Login Request
Step –2 OTP to the user
Step –3 Details will be checked by the server
Step –4.1 Requests to upload file
1 Read file
2 Calculate hash value
3 Checks the Hash value
If same Hashvalue
Reject (File Exist)
Else
Upload
4 File is splitted into blocks
5 Block consist of ID and data
Step – 4.2 Request to download
1. Blocks are merged and decrypted
2. Display the contain of the file
END

D. Preliminary Ideas

In this section we will define the AES algorithm use in our paper. The three main steps of generation of key encryption of data which is to be uploaded and the last is decryption of data when the use request to download the file.

- KeyGenSE(1) ! κ is the key generation algorithm that generates κ using security parameter 1 ;
- EncSE(κ, M) ! C is the AES encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C; and
- DecSE(κ, C) ! M is the AES decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

Convergent encryption helps in providing the data deduplication while uploading the data on the cloud .The original data is encrypted and the convergent key is generated .This convergent key provides a hash value and due to which we can avoid the deduplication.

The main role will be played by the hash tag it checks the hash tag value of the uploaded file if the hash value matches with any of the file on the cloud it means that the uploaded file is duplicated and the uploading of the file will be rejected. If the hash value is different then the file will be uploaded on the cloud.

While uploading the file on the cloud the file is splitted into small fragments so that the hacker may not be able to collect the whole file at a same time only a small fragment will be hacked by the hackers .Due to this the security issue may to resolve. The file is divided into the small blocks with the same size and each block will have its id and the data of the file contents. If any user needs to download the files then the blocks are merge and the file is decrypted and send the user.

Bloom Filter:

A Bloom filter is a simple space-efficient randomized data structure for representing a set S in order to support membership queries . Bloom filter is initialized by InitBF(em) algorithm, which means a bit array with size em is allocated. It has two Operations: AddBF(x) and QueryBF(y), where x and y are two elements.

1.The AddBF operation consists of hashing an element with ek hash functions h1, · · ·, hn.

2.QueryBF operation repeats the same hashing procedure, and then checks if the appropriate bits are set 1

IV. RESULT AND DISCUSSIONS

TABLE I Analysis of existing and proposed algorithm parameter

Algorithm	Constraints				
	Storage Space	Key Usage	Time Overhead	Cost	Security
Proposed AES	High	High	High	Medium	Low
SHA-1	Medium	High	Low	High	Low
MD5	Medium	Low	Medium	Low	Medium

According to existing paper, the analysis of various research issues is described in the TABLE I. It is classified as three types. High indicates the work has been completed in that area. There is an algorithm solving these types of problem. Medium- it shows which achieved half the successes in that constraints. Low- it depicts that there is need to explore optimized algorithm for the particular domain focus on different aspects such as storage, key usage and mainly on security. An advanced encryption algorithm is

faster than DES. It is a popular symmetric encryption algorithm. While using AES in deduplication it covers more storage space, key usage and time overhead with low security. SHA is secure hash algorithm. It is considered has stronger encryption and most preferred algorithm used by government. But usage of this algorithm causes high cost in deduplication. MD5 is secured hashing algorithm. The message authentication protocol verifies content of the message.

TABLE II Analysis of existing and proposed algorithm in security

	Constraints				
	Encrypted Deduplication	Pow	Side Channel Attack	SQL Injection Attack	DOS Attack
Existing System	Yes	No	Yes	No	No
Our Scheme	Yes	Yes	Yes	Yes	Yes

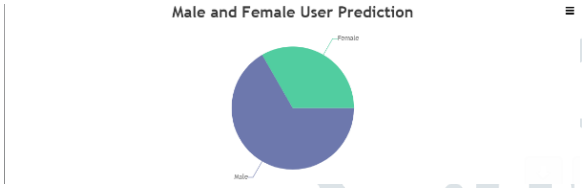

Fig 3: Male and Female User Prediction

On the basis of the registration we can identify the people according to its criteria.

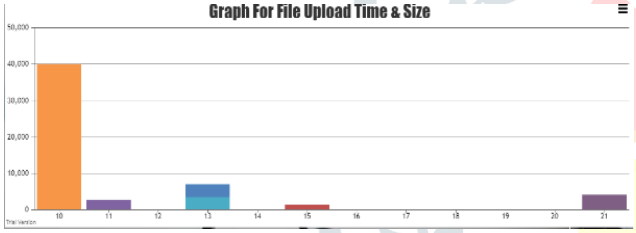

Fig 4: Graph For File Upload Time And Size

The above graph will show the time required to upload the file of a particular file.

V. CONCLUSIONS

In this paper we reviewed the deduplication techniques for better discretion and secure in cloud computing. The detection of redundant data and removal of this duplicated data is an important task for keeping the cloud storage clean and scalable. Also the security issue is consider and tried to solve it at the max level. This redundant data elimination has a great advantage for cloud storage. We have surveyed various techniques for deduplication and maintaining the security on the cloud.

REFERENCES

[1] J. Douceur, A. Adya, W. Bolosky, S. Dan, and M. Theimer, "Reclaimingspacefrom duplicate files in a serverless distributed filesystem," in *in Proceedings of the 22nd International Conferencevon Distributed Computing Systems, IEEE*, 2002, pp. 617–624.

[2] Chuanyi LIU 1, Yancheng WANG2, Jie LIN. (2013)A Policy-based De- duplication Mechanism for Encrypted Cloud Storage. Trustworthy Computing and Services ISCTCSV 2012 .Communication in computer and information Science ,vol 320,Springer,Berlin.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication,"in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2013, pp. 296–312.

[5] S.Keelveedhi, M.Bellare, and T. Ristenpart, "Dupless: server aide dencryption for deduplicated storage," in *Presented as partof the 22nd USE NIX Security Symposium (USENIX Security 13)*,2013, pp. 179–194.

[6] Cloud Computing: Security Issues and Research Challenges. Rabi Prasad Padhy, Manas Ranjan , PatraSuresh, Chandra Satapathy.

[7] Proceedings of the 5th Symposium on Operating Systems Design and Implementation. Boston, Massachusetts, USA December 9–11, 2002.

[8] "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage," in *Proceedings of the 8th ACMSIGSAC symposium on Information, computer and communications security*., 2013, pp. 195–206.

[9] J. Hur, D. Koo, Y. Shin, and K. Kang, "Secure data deduplication with dynamic ownership management in cloud storage,"2016.

[10] Y. Zhou, D. Feng, W. Xia, M. Fu, F. Huang, Y. Zhang, and C. Li, "Secdep: A user-aware efficient fine-grained secure deduplication scheme with multi-level key management," in *31stSymposium on Mass Storage Systems and Technologies (MSST),IEEE*. IEEE, 2015, pp. 1–14.

[11] L. Li, X. Chen, X. Huang, S. Tang, Y. Xiang, M. M. Hassanmad, and Alelaiwi, "Secure distributed deduplication systems with improved reliability,"*IEEE Transactions on Computers*, vol. 64,no. 12, pp. 3569– 3579, 2015.

[12] R. Chen, Y. Mu, G. Yang, and F. Guo, "Bl-mle: Block-level message- locked encryption for secure large file deduplication,"*IEEE Transactions on Information Forensics and Security*,vol. 10, no. 12, pp. 2643–2652, 2015.

[13] J. Xu, E. C. Chang, and J. Zhou, "Leakage-resilient client- side deduplication of encrypted data in cloud storage," *IACR ePrintArchive, 15pages*, 2011.

[14] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," *IEEE Security & Privacy*, vol. 8, no. 6, pp. 40–47, 2010.

[15] B. H. Bloom, "Spacetime trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp.422–426, 1970.

[16] https://www.cryptopp.com/.

[17] OpenSSL Project. http://www.openssl.org/.