Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Block Level Data Duplication on Hybrid Cloud Storage System

Aparna Ajit Patil, Asst. Prof. Dhanashree Kulkarni
Dr.D.Y.Patil COE, Computer Department
Savitribai Phule Pune University
Maharashtra, India

Abstract—Data deduplication is one of the most important data compression technique,used for removing identical copies of repetitive data. For reduce duplication of data authorized duplication system is used. When a user uploads a file on the cloud, the file is split into a number of blocks, each block having a size of 4KB. Block is encrypted using a convergent key and subsequently a token is generated for it by using token generation algorithm. After encrypting the data using convergent key, users retain the key before sending the ciphertext to the cloud. Due to the deterministic nature of encryption, if identical data copies are uploaded the same convergent keys and the same cipher text will be produced thus preventing the deduplication of data. Each block is then compared with the database of cloud. After comparing,if a match is found in the cloud database then only metadata of the block is stored in DB profiler. This paper also prevents unauthorized access by using a secure proof of ownership protocol .The protocol uses authorizedduplicate check for hybrid cloud architecture. Thus, prevention is achieved by deduplication of data and protection of confidentiality of data while incurring minimal overhead compared to the normal operation.The proposed system has been compared with the existing system on the basis of database usage, security and bandwidth .The outcomes presented at the end conclusively prove that the proposed system gives better results as to the existing one.

Keywords—Block levelduplication, authorized duplicate check, confidentiality, hybrid cloud, Proof of ownership.

I. INTRODUCTION

Cloud computing provides extensive virtualized resource to user as services across the whole internet while hiding the platform and implementing details. Cloud storage service is management of ever increasing volume of data. To make data management scalable in cloud computing, deduplication has been standard technique. Data compression technique is used for eliminating the duplicate copies of repeated data in cloud storage to diminish the data duplication. This technique is used to improve storage utilization and also be applied to network data transfers to reduce the number of bytes that must be sent. Keeping multiple data copies with the similar content, deduplication remove redundant data by keeping only one physical copy and refer other redundant data to that copy data files. Although data deduplication takes a lot of benefits,security and privacy concerns arise as users' sensitivedata are capable to both insider as well as outsider attacks. In the traditional encryption providing data confidentiality,is contradictory deduplication occurs file level and block level. The duplicate copies of identical file eliminate by file level deduplication .For the block level duplication which eliminates duplicates blocks of data that occur in non-identical files. Although data deduplication takes a lot of advantages, security as well as privacy concerns arise as users' sensitive data are capable to both insider and outsider attacks. In the traditional encryption providing data confidentiality, is contradictory with data deduplication. Traditional encryption requires different users to encrypt their data with own keys.

For making the feasible deduplication and maintain the data confidentiality used convergent encryption technique. It encrypts decrypts a data copy with a convergent key, the content of the data copy obtained by computing the cryptographic hash value of. After the data encryption and key generation process users retain the keys and send the ciphertext to the cloud. Since the encryption operation is determinative and is derived from the data content, similar data copies will generate the same convergent key and hence the same ciphertext. A secure proof of ownership protocol is used to prevent the unauthorized access and also provide the proof to user regarding the duplicate is found of the same file.

II. RELATED WORK

Jin Li and Yan Kit Li presented hybrid cloud approach for secure authorized deduplication. It aims for solving the problem of the deduplication with different privileges in cloud computing [1]. M. Bellare, S. Keelveedhi, and T. Ristenpar proposes DupLess: Server aided encryption for deduplicated storage for cloud storage service provider like Mozy, Dropbox, and others perform deduplication to save space by only storing one copy of each file uploaded .Message lock encryption is used to resolve the problem of clients encrypt their file however the saving are lock. Dupless is used to provide secure deduplicated storage as well as storage resisting brute-force attacks [2].Jin Li proposed secure deduplication with efficient and reliable convergent key management contain the different techniques which is used in the secure deduplication and remove the duplicate copies of data for reduce the storage space in cloud system. For that

© 2015, IJARCSSE All Rights Reserved

purpose use the convergent encryption to providing the data confidentiality and encrypt /decrypt a data copy with a convergent key, which is given by computing the cryptographic hash value of the content of data copy itself. This technique is used for reduce the storage space and bandwidth also provide the confidentiality [4]

Twin clouds: Architecture for secure cloud computing proposed Client uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries. It store large amount of data and low latency [10]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg presented Proof of Ownership for check the ownership of user who having the authority for access the file or uploaded data from the given cloud storage system.Sometimes user not upload the file but it try to access the data from the given cloud to avoid this problem use proof of ownership algorithm used to provide authorized access to user [11].

III. SYSTEM OVERVIEW

A. Problem Statement

The file level duplication method have enormous storage overhead and less efficiency. To overcome this problem, Authorized Duplicate System is developed to avoid the duplicate copies of data, which reduces space used for storage as well as data overhead in cloud storage. It also protects the confidentiality of sensitive data.

B. Authorized Duplication System

There are three entities defined in this system, that is user, private cloud and secure cloud service provider(S-CSP) in public cloud. The S-CSP accomplishes deduplication by checking if the contents of two files are the identical and stores only one of them. The access right to a file is describing based on a set of privileges. The accurate definition of a privilege varies across applications.Token means each privilege is represented in the form of a short message. Each file is related with some file tokens, which denote the tag with specified privileges. A user computes as well as sends duplicate check tokens to the public cloud for authorized duplicate check.

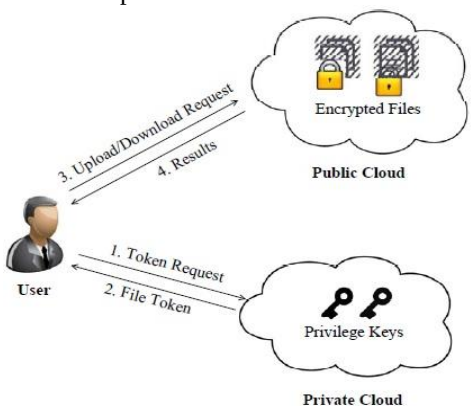

Fig.1 Authorized Duplicate system

- User: Users have access to the private cloud server and a semi trusted third party which will help in performingdeduplicable encryption by generating file tokens for the requesting users.
- S-CSP: This is entities that contribute a only unique data. In this paper, we assume that S-CSP is always online and has huge storage capacity and computation power. Data Users. A user is an entity that desire to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the identical user or different users. Each user is issued a set of privileges in the setup of the system for the authorized deduplication system. Each file is protected with the convergent encryption key as well as privilege keys to realize the authorized deduplication with differential privileges.
- Private Cloud:After Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating users secure usage of cloud service. Specifically, since the computing resources at data user/owner side are inadequate and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private cloud manages private keys for the privileges who answer the file token requests from the users. The interface offered by the private cloud authorize user to submit files and queries [1].

C. Mathematical Model

Let S be a system that find out duplicate copies of the file using Authorized deduplication system in hybrid cloud.

S={F,B,C,T,P,M,O}

Where,

F={F1,F2,F3,…Fn}

F1={B1,B2,B3,…Bn}

B1={CBi,TBi,Pki}

CBi=Set of cipher text block

© 2015, IJARCSSE All Rights Reserved

T = Token [16-Bit unique token for Block]
P= Private Key (PKi)used for encryption &description mechanism
M=Metadata of file
O=Output consist reduce database size
Following steps occurs in the given proposed system architecture:
1.File F is divided into multiple blocks F= ∑Bi, F = size (F) / 4096
2.KeyGen(1^{λ})→k is key generation algorithm,generate secrete key using security parameter 1^{λ} .Secret key stores in internal DB of Security Service (SS).
3.Enc (k,F)→C is encryption algorithm that takes secrete key k and then file and then F output is cipher text C.
4.Generate Token T for each block.
5.Dec(k,C)→F is Decryption algorithm that takes secrete key k and ciphertext C and then output is original file F.
F=∑PlainText (Bi)
PlainText (Bi) = SS (CipherText (Bi), TiBi)
6. Detect duplication.
Security Service generates TiBi Token on basic on Bi, If the same Bi comes in then it will generate the same TiBi.
i.e. TiBi = token generation (Bi);
Then it will store the TiBi to the Own Security Db.
If file is found in database it generates response.

IV. SYSTEM DESIGN

1. Proposed System

When the user wants to upload & download the file from cloud storage at that time first user request to the web server for uploading file means only authorized user can upload the file to web server for that purpose it use the proof of ownership algorithm. User to prove their ownership of data copies to the storage server. When file is uploaded it divides into blocks that are block size 4KB by default. The file having extension like .txt, .html, .js,css,.xml, .java, .c, .cppetc which open in notepad. According to file size the block occurs. Each block contain their own cipher text, token for the unique identification and private key. Given block compare with cloud storage if the block is already store in database it store only metadata of block.

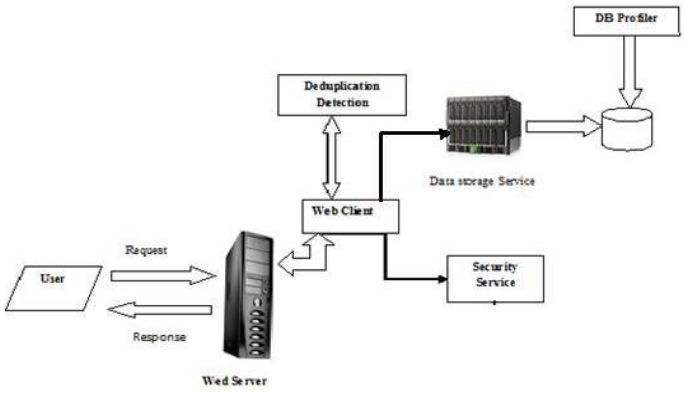

Fig.2 SystemArchitecture

- Web server: Web server is store, process and deliver the web pages to client. User want to upload file to cloud through web server. The communication between client and server takes place using the Hypertext Transfer Protocol.
- Web client: Web client connect to server and retrieve the web pages. Data owners want to divided file into multiple blocks. For each block it perform encryption operation and generate the response like cipher text, token and private key for each block.
- Security Services :After all the response has been generated PKi is store into internal db of SS[security service] . The main Idea behind to hide PKi is provide security to Cipher Text(Bi) , So no one else can used the key and try to decrypt the block.
- Duplication detection: Security Service generates TiBi Token on basic on Bi , If the same Bi comes in then it will generate the same TiBi. Token generation algorithm is used . then it will store the TiBi to the Own Security database. Now next time after generation of the code it will cross verify with the exiting token data and send back the notification accordingly.
- DB Profiler: It stores uploaded data, shared data, all list of users, and sequence of token of blocks. It also stores all encrypted data and metadata of file also store in the database.

The data storage server contain all the uploaded files and DB profiler store all the metadata of the file.
Case 1:When file F1 & F2 are different the all the data will be store in the database in different blocks.

© 2015, IJARCSSE All Rights Reserved

Case 2:If the file F1 =F2 it stores only one file in the database avoid duplication of the data.
Case 3:If F1€F2 then it compare the blocks with data storage and only different blocks of both file will be store in the database.
For execution of Authorized duplicate system ,first start different services which is used in cloud for deployment purpose. Following are the Pre-requisites for execution:

1. Tester for validation of authorized user.
2. Security service executer for checking duplicate copies over cloud server.
3. TTP Executer for validate email address and send one time password through email address.
4. Block level encryption which provide the registration and further details to user regarding duplication system.

2. Workflow of Encryption

When user uploaded file to cloud storage ,it divide into multiple blocks .Security service received data and start encryption using advanced encryption standards. After encryption, generate token for unique identification of block.

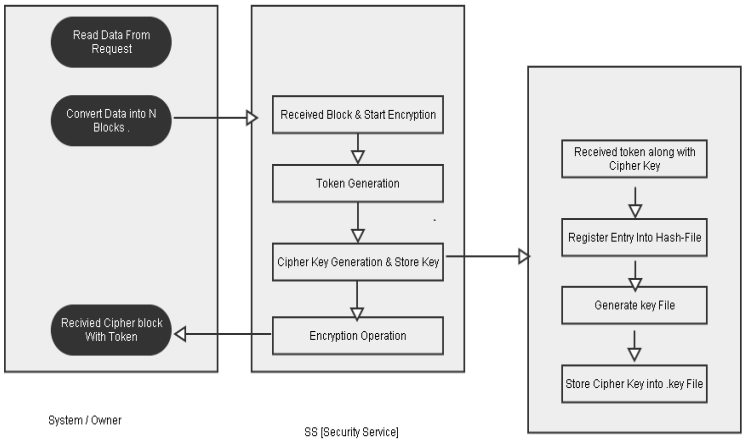

Fig. 3 Encryption Workflow

Security service generate cipher text as well as key which store in the database. It also register entry to hash file to maintain the sequence of block which is used to download file. It compresses the size of given file after encryption so it improves upload performance of the given data.Fig. 3 shows workflow of encryption.
The following steps are perform on encryption workflow:

1. Read the data form uploaded request that is uploaded file.
2. File is divided into multiple blocks.
3. Security Service receive the block and perform encryption.
4. Each block generate token, cipher text and private key.
5. In security services, hash table used to maintain the sequence of blocks and gives the original file.

3. Token Generation Algorithm

The token generation algorithm used for generate the token foe uniquely identification of blocks and maintain the proper sequence of the file block at the time of downloading the give file. The following Fig.4 show the steps of generating token for block.

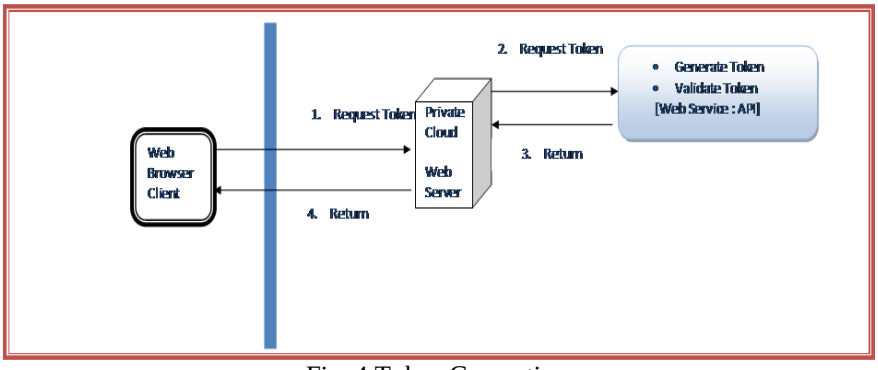

Fig. 4 Token Generation

Input: File as a input

1. Web browser client request token to private cloud
2. Web services validate token
3. Return token to web server
4. Web client got token
   Output: Generate token.

© 2015, IJARCSSE All Rights Reserved

V. RESULT AND ANALYSIS

The authorized deduplication system used to avoid duplicate copies of data in the given cloud. Proposed system implemented by using block level duplication which compare the given blocks with database, suppose the file is already stored in the database and that same file uploaded by another user at that time only metadata of file will be store not actually file so it reduce the storage space of data and proper utilization of space. The data will be store in encrypted format so it also maintains security because each block contains their own token, cipher text and private key. The database size will be reduced by using this technique.The proposed system has been compared with the existing system on the basis of database usage, and security using proof of ownership.

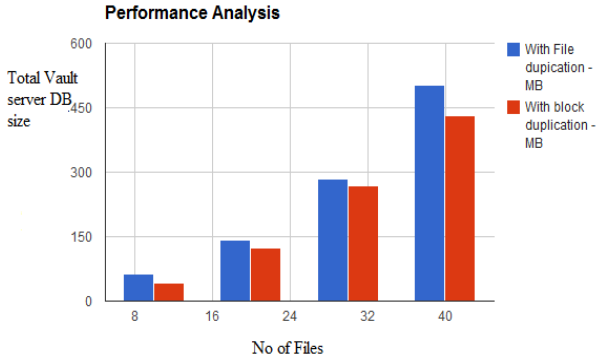

Fig.5 Comparison between file level duplication and block level duplication

The above Fig.5 show expected output from proposed system. X axis shows number of files and Y axis shows total database size. It shows actual storage space in database, with file level duplication having large storage space as compare to block level duplication. For that purpose, use the block level duplication for reduce storage space in database and reduce duplication.

Table1.Actual Result Comparison

Sr. No	Number of Files	File level duplication DB Size	Block level duplication DB Size
1	8	80	50
2	16	120	70
3	24	140	110
4	32	280	210
5	40	490	308

The above Table1.Shows the database usage for file level duplication and block level duplication. The file level duplication having extra storage space as compare to block level duplication. The block level duplication having less storage space and also provide extra security using proof of ownership concept.

VI. CONCLUSIONS

In this proposed article secure deduplication occurs with the help of token generation and secure upload/download of file. It assures the user about high data security and also avoids data duplication in cloud storage. This helps in eliminating duplicate copies of repeating data, reduces storage space used and saves bandwidth in cloud storage. Convergent Encryption protects the confidentiality of the sensitive data. The new deduplication construct supports Authorized Duplicate Check in hybrid cloud architecture. The proposed system aims to have minimal overhead in entire upload and download process and is negligible for moderate file size. The scheme is suitable to construct an authorized deduplication system for backup storage. It is more efficient and reliable than existing system. This system shows that Authorized Duplicate Check Scheme incurs minimal overhead compared to convergent and network transfer.

In future by using Cloud Service Provider (CSP) have significant resources to govern distributed cloud storage servers and to manage its database servers. It also provides virtual infrastructure to host application services. These services can be used by the client to manage his data stored in the cloud servers. The CSP provides a web interface for the client to store data into a set of cloud servers, which are running in a cooperated and distributed manner. In addition, the web interface is used by the users to retrieve, modify and restore data from the cloud, depending on their access rights. Moreover, the CSP relies on database servers to map client identities to their stored data identifiers and group identifiers

REFERENCES

[1] Jin Li and Yan Kit Li ˙A Hybrid cloud approach for secure authorized deduplication, IEEE Transaction on parallel and distributed system, 2014.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server aided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[3] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication .IACR Cryptology ePrint Archive, 2013.

© 2015, IJARCSSE All Rights Reserved

[4] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, pages 296– 312, 2013

[6] 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.

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

[8] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S Ossowski and P. 2012.

[9] 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 Security2012.

[10] 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.

[11] 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 andCommunications Security, pages 491–500. ACM, 2011

[12] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacy aware data intensive computing on hybrid clouds. In Proceedingsof the 18th ACM conference on Computer and communications security, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM

[13] 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

[14] M. Bellare, C. Namprempre , and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 2009.

[15] 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

© 2015, IJARCSSE All Rights Reserved