Secure Data Deduplication on Hybrid Cloud Storage Architecture

Miss.Aparna Ajit Patil
Computer Department
Dr.D.Y.Patil College of Engineering
Pune, India
patilaparna.4@gmail.com

Prof.Dhanashree Kulkarni
Computer Department
Dr.D.Y.Patil College of Engineering
Pune,India
dskulkarni10@gmail.com

Abstract— Data deduplication is one of the most important Data compression techniques used for to removing the identical copies of repeating data and it is used in the cloud storage for the purpose of reduce the storage space as well as save bandwidth. To retain the confidentiality of sensitive data while supporting the deduplication, to encrypt the data before outsourcing convergent encryption technique has been proposed . This project makes the first attempt to formally address the problem of authorized data deduplication giving better protect data security, Different from the traditional deduplication system, distinctive benefits of the user are further considered the duplicate check besides the data itself. Hybrid cloud architecture having various new deduplication constructions supporting authorized duplicate check. The proposed security models contain the illustration of security analysis scheme. As a proof of concept, contains the implementation framework of proposed authorized duplicate check scheme and conduct experiments using these prototype. In proposed system contain authorized duplicate check scheme sustain minimal overhead compared to normal operations.

***Keywords**-Deduplication,Authorized deduplication check,Hrbrid cloud,Token generation,Block level deduplication.*

*****

I. Introduction

Cloud computing provides unlimited virtualized recourse to user as services across the whole internet while hiding the platform and implementing details. Cloud storage service is the management of evergreen increasing mass of data. To make data management scalable in cloud computing, deduplication has been a conventional technique. Data compression technique is used for eliminating the duplicate copies of repeated data in cloud storage to reduce 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 eliminates redundant data by keeping only one physical copy and refer other redundant data to that copy. Data deduplication occurs file level as well as 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 benefits, 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. The paper continue to exist as follows: Section II deals with an overview of the related work which is used in the secure deduplication method ,Section III is about system overview and Section IV represents System design, Section V present Result & analysis whereas the conclusions are drawn in Section VI.

II. Related Work

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] authorized deduplication technique is poven by Jin Li[4] which avoid the duplicate content in cloud storage system and incurs minimal overhead as compared to the normal operation by using convergent key encryption . It also provide the security to the given data. C.Ng[7] presented reverse deduplication technique for read to latest backup .W. K. Ng, Y. Wen, and H. Zhu[8] proposes private data deduplication Protocols in cloud storage for Enhance the efficiency of data S. Halevi, D. Harnik [9]proposes Proofs of Ownership in Remote Storage Systems which contain Performance measurements indicate that the scheme incurs only a small overhead compared to naive client-side deduplication. Bugiel[11]presented twin clouds: an architecture for secure cloud computing for Client uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries. Token generation technique and identity based signature for provide security to the give data in cloud storage.[14].

III. SYSTEM OVERVIEW

A. Problem Statement

To develop authorized data de-duplication for protecting the data security by including differential privileges of users in the duplicate check and conduct test bed experiments to evaluate the overhead of the prototype. De-duplication is one of important data compression techniques for eliminating duplicate copies of repeating data. For that purpose Authorized duplication check system is used. This paper addresses problem of privacy preserving de-duplication in cloud computing and propose a new de-duplication system supporting for Differential Authorization, Authorized Duplicate Check, Unforgeability of file token/duplicate-check token, Indistinguishability of file token/duplicate-check token, Data Confidentiality.

B. Authorized Deduplication System

There are three entities defined in our system, that is, users, private cloud and Secure cloud service provider(S-CSP) in public cloud as shown in following fig.1 The S-CSP accomplish 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 describe 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 realated 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. Users have access to the private cloud server and a semitrusted third party which will help in performing deduplicable encryption by generating file tokens for the requesting users. S-CSP. This is an entity that contribute a data storage service in public cloud. The S-CSP provides the data outsourcing service as well as stores data on behalf of the users. To diminish the storage cost, the S-CSP remove the storage of redundant data via deduplication and keeps 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 answers the file token requests from the users. The interface offered by the private cloud authorize user to submit files and queries.[1]

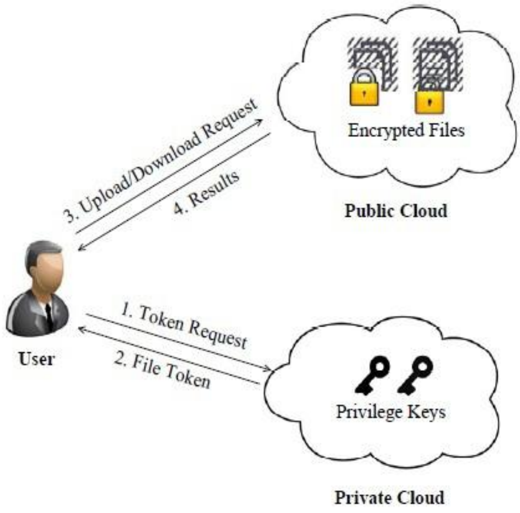

Fig 1: Architecture of Authorized Deduplication System.

C. Mathematical Model

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

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

Where,

$F=\{F1,F2,F3,\ldots Fn\}$

$F1=\{B1,B2,B3,\ldots Bn\}$

$B1=\{CBi,TBi,Pki\}$

CBi=Set of cipher text block

T = Token [16-Bit , unique token for Block]

P= Private Key (PKi)used for encryption &descrption mechanism

M=Metadata of file

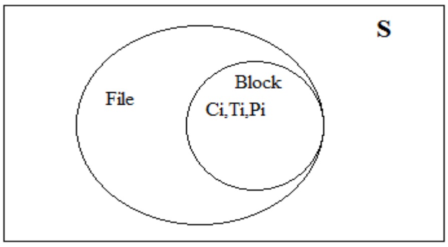

Fig.2 F∩(BU(C,T,P))

IV. System Design

A. Proposed System Architecture

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 i.e block size is 4KB by default. According to file size the block occurs. Each block contain there own cipher text , token for the unique identification and private key. After that deduplication dectection occurs which shown in following fig 3.

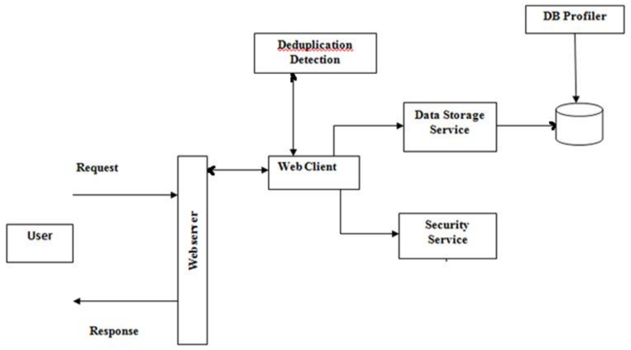

Fig 3: System Architecture

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.

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.

B. Workflow For File Upload /Download

Authorized user can access the file from cloud storage.Only privileged user have authority for upload and download the file from cloud. It contain the actual flow of the file with different operations. Following Fig 4 show the workflow of the upload and download file from cloud storage system.

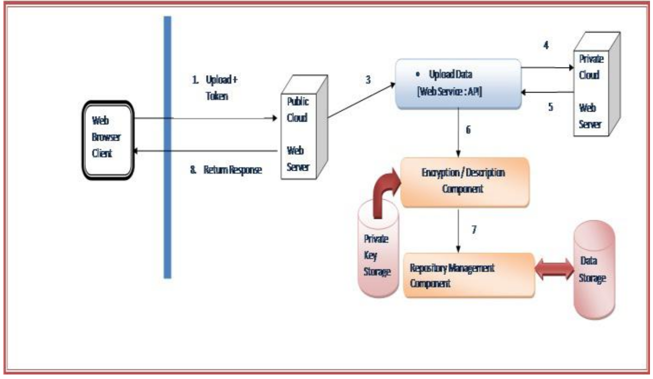

Fig 4: Workflow of Upload and Download File.

System Workflow:

1) Get the Token.

2) Upload / Download The File [after token generation].

3) Forward the request for upload to the web service API.

4) Web service API will connect to the Private cloud and validate request.

5) Response from the private cloud , for verification [Boolean response].

6) Data will given to the Security Component which will responsible for encryption / description. It will have its own private data base for storing keys and other information.

7) Actual Encrypted data will be store onto the repository.

C. Token Genration

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 5 show the steps of generating token for block.

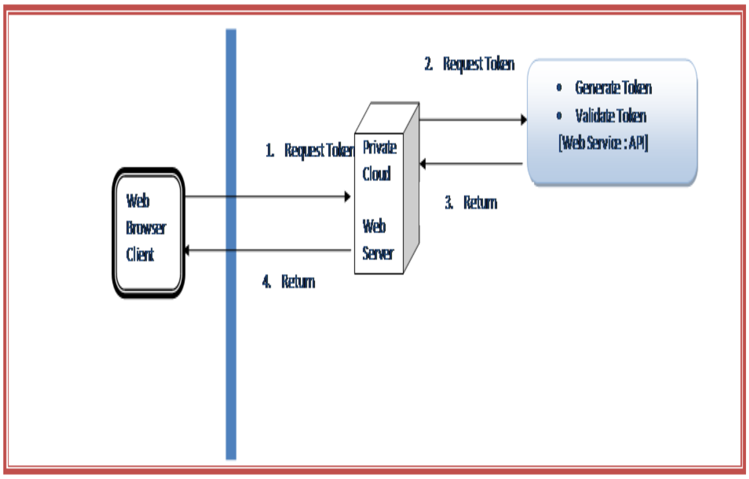

Fig.5 Token Generation

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

V. Result and Analysis

The authorizes 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 maintain security because each block contain their own token, cipher text and private key. The database size will be reduced by using this technique.

A. System Implementation

The prototype of proposed system is implemented by java. The server machine has configuration of Intel Core Duo CPU 2.4 GHz with 4GB of RAM. Heidisql used for the database storage. When user upload file,it divide into the blocks then another same file is uploaded at that time it compare with given database.The file is already store it stores only the metadata of file.When to files having same content at that time it didn't save the similar block it store only different blocks which was not in given database with the token for unique identification. Given uploaded file wiil be store in encrypted format. Block level duplication reduce the storage space as well as uploaded bandwidth as compare to file level duplication. The block size is inversely proportional to the time require to compare with database.The Advanced encryption algorithm is used for encrypt and decrypt tha data,providing high security performance.The following fig 6 shows the actual output of proposed system.

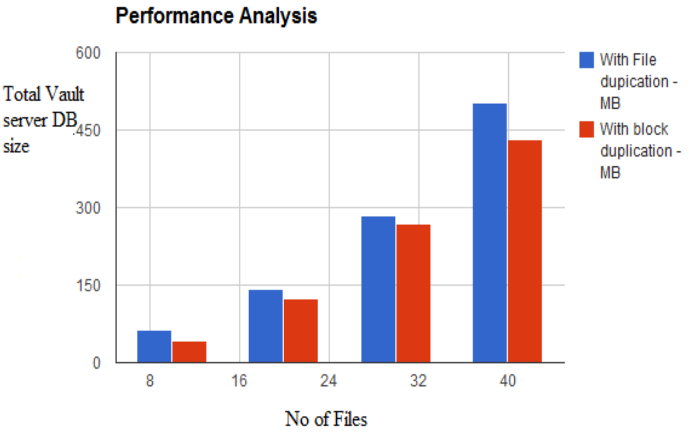

Fig.6 Comparison between file level deduplication and block level deduplication regarding storage space

The above graph is contain expected output from proposed system. X axis shows number of files and Y axis shows tatal 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.

VI. Conclusion

In this paper proposed the secure deduplication with the help of token generation and Secure upload download we can assure the user about high data security and also avoid data deduplication. Security analysis determine that our schemes are secure in terms of insider as well as outsider attacks specified in the proposed security model. As a proof of concept, we executed a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments on our prototype. In this paper we have to provide the different techniques to reduce the deduplication in cloud storage and maintain the security.

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.

ACKNOWLEDGMENT

The authors would like to express heartfelt gratitude towards the people whose help was very useful to complete this dissertation work on the topic of Secure Data deduplication On Hybrid Cloud Storage Architecture System. It is great privilege to express sincerest regards to P.G. Guide as well as H.O.D., Prof. Dhanashree Kulkarni for her valuable inputs, able guidance, encouragement, wholehearted cooperation and constructive criticism throughout the duration of this work.

The Work in this Paper is based on block level deduplication for avoid the duplicate copies of dada in hybrid cloud. Block level deduplication system executed by various techniques like authorized duplicate check. Author is very much thankful to them for such a precious work. Author would like to pay Sincere thanks to IJRITCC-2015 organizing committee.

References

.

[1] Jin Li and Yan Kit Li ˙A Hybrid cloud approach for secure authorized dedoplication, IEEE Transaction on parallel and distributed system,vol:pp:99 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.

[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 and Communications 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 Proceedings of 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