IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Data Deduplication, Revision and Sharing in Hybrid Cloud Supporting Differential Authorized Privilege

Priyanka Bhopale[1], Dr. A.M. Kanthe[2]

Student, Computer Engineering, Sinhgad Institute of Technology, Lonavala, Pune, India[1]

Associate Professor, Computer Engineering, Sinhgad Institute of Technology, Lonavala, Pune, India[2]

Abstract: Cloud storage units are becoming more and more popular these days and hence they require the mechanism to down their utilization cost and provide more efficiency and security to client's data. Deduplication is the technology which stores only a single copy of echoing data and so it reduces the required storage space. In order to provide confidentiality to the client's data and at the same time acquiring data deduplication, so we provide a technique called as convergent encryption. In this data is encrypted before outsourcing. To provide better data security this first talks the problem of authorized data deduplication, in this we further enhance this with sharing the files among the users on the basis of privileges. Secondly, this gives right to the user to revise the content of its data and share it with other users with keeping the integrity of data intact. Also we have presented several security schemes with data deduplication construction in a hybrid cloud model. The proposed method shows that the users are able to share the data and revise the data, in cloud and with efficiency and security. Our performance measures show that the proposed system gives more security in terms of insider and outsider attacks.

Keywords: authorized data deduplication, encryption, data revision, hybrid cloud, user privileges

I. INTRODUCTION

As the request for computing supremacy amplified with more users coming on-line and hence the data being published on-line became increasing so some strong architectural modifications had to be introduced to bring compute resources extra efficiently. So the cloud computing becomes current technology, in which increasing amount of data is being kept within the cloud and shared by users with some defined privileges, that plan the access rights of the data present in the cloud. One important challenge of cloud storage facilities is the organization of the increasing capacity of information. Cloud computing is associate rising service model that has computation and storage properties on the web. One attractive practicality that cloud computing can produce is cloud storage. The concern of individual user or large enterprises is that they should retrieve their data without any data loss. Instead of buying the required storage media to keep info backups, user and enterprises will simply source their info backup to the cloud service suppliers.

Data deduplication is a very important technique to compress the data which will provide better storage utilization and that is the reason it has attracted more and more people nowadays. This technique eliminates the copies of the repeating data and stores a single copy of the data. Instead of storing the multiple copies of the echoing data. Eliminating redundant data can knowingly shrink storage requirements and improve bandwidth competence. This technology is used to increase storage consumption and also reduce the number of bytes sent in a network. Deduplication can work at two places either at block level or file level. In block level deduplication it removes duplicate block of the same file and in file level deduplication it removes duplicate copies of the same file.

The data which is stored on the cloud shared among the users under some predefined privileges, which outlines the access rights to the stored data. While doing the registration process, privileges are assigned to each user which provide the security to data. The registered users can revise the data as per need. Such as if a user wants can add some new contents to the original file and save it also, he/she can delete the file if required. These operations can be performed at the run time and hence gives the more flexibility to user and also it reduces the time of the system as compared to the traditional models. At the same time the security, confidentiality and integrity of data also have to be maintained.

Though data deduplication gives us many benefits but it has privacy concerns also which leads system insecure to insider and outsider attack. Traditional encryption technique provides data confidentiality but it is impossible to achieve data deduplication using the traditional method. In traditional method different users produce different cipher text for the same data copies, which makes deduplication impossible. We want a method which produces identical cipher text for same data copies of the same/different users which makes deduplication possible. Convergent encryption [1] method is proposed to make data deduplication possible while making data confidentiality practical. It generates the key with the cryptographic hash values of the data and the generated key is useful to encrypt or decrypts the block information.

Hence the same data file create the same hash values and hence same key and same encrypted/decrypted data. This system also provide support to differential authorized deduplication check, in this he/she is bounded by the set of privilege to signify which user is provided with which authority.

The rest of this paper is arranged as follows, section II talks about Literature survey. Section III gives problem statement with design of the system, Section IV shows the details of proposed system. Section V shows the mathematical model of system, Section VI shows the result of the system, Section VII shows the conclusion and the references are given.

II. LITERATURE SURVEY

Jin Li, Yan Kit Li, Xiaofeng Chen, P. P. C. Lee, Wenjing Lou [1] proposed authorized data deduplication system which is used to guard the data safety by giving differential privileges to users in the duplicate check method. To achieve data deduplication different techniques are utilized in this hybrid cloud structure which finds duplicate files. Private cloud server generate the tokens of file with the help of private keys. This paper make use of convergent encryption technique to achieve deduplication and the trial shows that authorized duplicate check acquires negligible overhead if compared with convergent encryption time required and network transfer data rate.

Jin Li, Xiaofeng Chen, M. Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou [2] proposed a system to find Data deduplication. Convergent encryption technique is used instead of the normal encryption which allows to achieve the deduplication of files but at the same time the management of these keys is not given in previous research. So the convergent encryption method has been adopted to secure deduplication. A significant issue of creating convergent encryption practical is to expertly and consistently handle a massive number of convergent keys. This paper creates the first effort to address the matter of achieving effective and reliable key administration in secure deduplication. They first propose a style in which each user holds an independent master key which encrypt the convergent keys later outsourcing them to the cloud. Conversely, the system finds it difficult to manage the newly generated enormous number of keys. As the number of users increases the number of keys also which makes it difficult to handle. To solve this problem, they propose a fresh method called as Dekey, in this users do not require to manage any of the keys on their own but in its place securely assign the convergent key is shared across multiple servers. Security study shows that Dekey is safe in terms of the definitions given in the planned security system. As a proof of concept, they implementedDekey using the Ramp secret allotment method and demonstrates that Dekey experiences limited overhead in truthful environments.

Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, and Mohammad Mehedi Hassan [3] this paper creates the first try to support the idea of distributed consistent deduplication system, authors offer unusual distributed deduplication systems with advanced reliability in which the blocks of data are spread across various cloud servers. The security requirements of data confidentiality and tag trustworthiness are also attained by proclaiming a deterministic secret sharing technique in distributed storage structures, instead with convergent encryption system as in previous deduplication systems. Security analysis demonstrates that the proposed model is secure in terms of the classifications specified to achieve security in the system. As a proof of concept, authors created the system and prove that the proposed method experienced less overhead.

S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg [4] they proposed data deduplication as one of acute info packing plans to remove the copy of duplicates, and has been typically used as a part of Cloud storage to reduce the amount of storage area and improves data transfer ability. To secure the data while supporting the data deduplication convergent encryption is produced which encodes the data before outsourcing. They propose a method which efficiently prove the server that the file is belong to the specific user. The proof of owner ship is applied where the tag and token of the file are also save in the server and when the user wants to access the file later it has to match the token of the file this achieves high level security to the data.

Chun-Ho Ng and Patrick P. C. Lee [5] offer RevDedup, which removes the duplicate from old data instead of new data file, in contrast with conventional deduplication system. This system increases reads using a method called as reverse deduplication and addresses the problem of fragmentation. This proposed system increases efficiency of saving data, throughput and provides high backup

III. PROBLEM STATEMENT

A. Problem Definition

The proposed paper is aimed to efficiently solve the problem of data deduplication with access rights to the user i.e. differential privilege in hybrid cloud computing environment. We provide our users a facility that they can share their copies of data to other users in the cloud under differential authorized privilege and produce a facility that they could even add some data or delete some copies of data from cloud. We propose an architecture which has public cloud and private cloud. Private cloud is providing the authentication to the user in order to achieve the confidentiality of the data and public cloud will store the actual data in encrypted format. That means the actual data is maintained by the public cloud while the differential authorized check is performed by private cloud. Then we improve our system in terms of security. The user is not able to do the duplicate check without the privilege and such users are also not allowed to perform any other operations.

B. System design

The architecture of the network is illustrated here.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

The units used in the scheme are described as follows

1. Private cloud

In general we can say, all data which proves the identity of the users are stored with the private cloud. Any user wants to access the data in the public cloud has to prove the authentication to the system and is proved with the help of file tokens, privileges and convergent keys. Every user has to prove the identity.

2. Public cloud

The data in encrypted format is stored in public cloud. Every user can get access to the data copies in the public cloud after proving the authentication to the private cloud. Hash values of the data copies are generated from the original data files and convergent encryption technique is applied for duplicate check. Here the single print of a file is kept, if it is deduplicate file a reference is added to the original file.

3. User

User is assigned privilege at the time of registration in the system, who can store, retrieve, alter the data or share data copy with other users in the system.

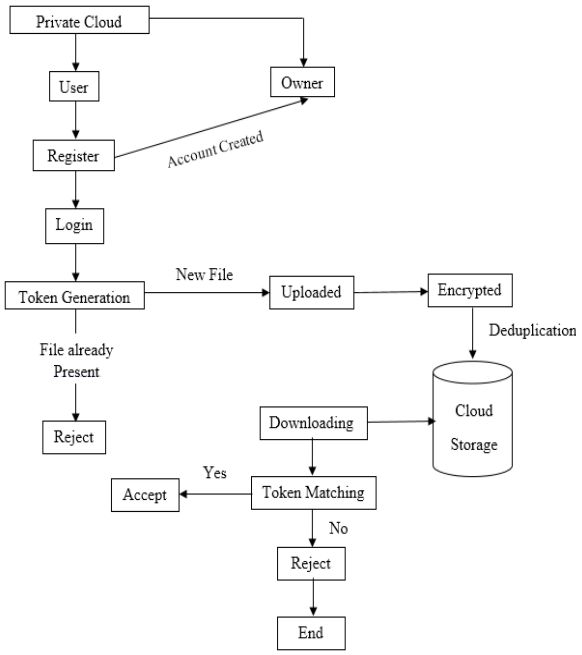

Fig 1Flow of Data Deduplication

C. Design Goals

1. Authorized duplicate check

Authorized user is capable to use his/her distinct private keys to produce query for certain file and the privileges he/she maintained with the help of private cloud, whereas the public cloud performs duplicate check directly and states the user if there is any duplicate. Role based privilege [9][10] is provided in the system.

2. Differential authorized privilege

Each user is assigned a privilege and on that basis the entry to the user is provided to the cloud and can retrieve their individual file token to perform duplicate check. None of the user is not able to generate a token for a file which is not of his/her privilege.

3. Data confidentiality

The data stored at the cloud can access by the user on the basis of the suitable privilege. Further the files are stored in cloud in an encrypted format. In this construction, advance level of confidentiality is achieved in cloud on the basis of convergent encryption.

IV. PROPOSED METHODOLOGY

To provide more flexibility to the user while working in the cloud we proposed a method called as data revision using which we provide options to the client in order to make the required changes in the block of data. Once a user has uploaded the data in a cloud and is now want to make any changes to the previously stored data as per the requirement so that can possible through this technique. But while performing this action the user has to prove its own identity to the private cloud that the data belong to the same user and carries the required privilege. That means only the authorized users are able to perform the operations. Once the authorized user has done any changes in the data then the deduplicate check method, checks the duplicate copies done for the revised data. Further we have extended with method, through which the users are competent to share their file with other users in the system on the basis of privilege provided to them. This adds more strength in the system.

Algorithm for data alteration

Step 1: Start (request to alter data)
Step 2: Select the appropriate option
If add a block of data go to step 3
If remove a block of data go to step 4
Step 3: Insert block b at specified i^{th} location
Update file f to f'
Remove the reference r
Check for duplicate of f'
If duplicate exist
Then add reference to existing copy
else
create a new reference r'
Step 4: Remove a file f
Check the ownership of f
Generate the token t
If t=t'
Then remove file f
Remove the reference r

The given algorithm works for data revision. When the user has uploaded files and are being saved in a system in an encrypted format. With the changes in any file the reference for that respective file will be removed and a new reference is created after deduplicate check.

The given diagram shows the operations performed in the cloud where the user requests a token to cloud which is used to discover the deduplicate check and accordingly that the file is stored or downloaded to or from cloud.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

After this user can change the data in future according to the need.

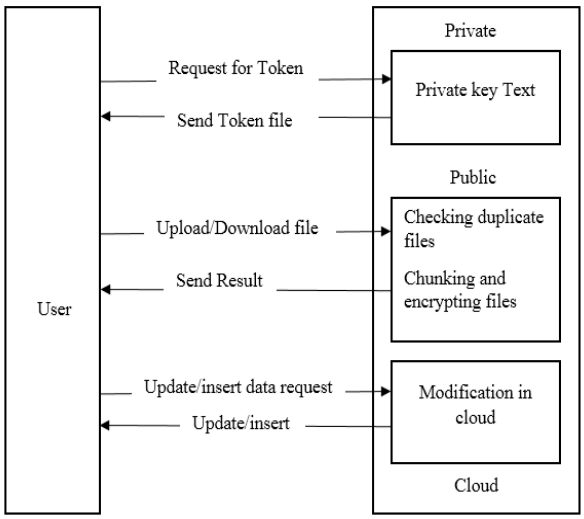

Fig 2 Data alteration in cloud system

V. MATHEMATICAL MODEL

The number of files stored in the system shows working of the architecture. The number of unique copies stored in the system will give us a worst case scenario because the system have to upload each and to store every file will automatically increases the time period of system and uses maximum bandwidth. If we check the duplicates while uploading the file then system will give us the better performance.

Input = {f}
Output = {enc}
Where,
S is a system which represented flow of project
Sign is signature for each file
f = Data file
enc = Encrypted File
p = privilege
kpi = symmetric key
FileTag(F) – Tag ϕF= TagGen(F) – this is on private cloud
Return {ϕ′F, pT= TagGen(ϕF, kpT)}
TokenReq{ϕ′ F,pT } - this is on public cloud
DupCheckReq(token {ϕ′ F,pT }) - Send File to public cloud
ShareTokenReq(Tag, {P}) – It requests the Private Server to create the Share Token of file with the FileTag(f) and Privilege Set of user;
FileEncrypt(EncCE(kF, F)) - convergent key kF= KeyGenCE(F)

It encrypts the File with a method called as Convergent Encryption which uses 256-bit AES algorithm in cipher block chaining (CBC), where the convergent key is obtained from SHA-256 Hashing of the file; and FileUploadReq(CF , { ϕ′ F,pT }) – Update (File) – It updated the file data and recheck De duplication of file
Insert (File, data) – Is append some data to the file which is previously present and re check de duplication.
Delete (f, kF) – It deletes file which we are select.
Share_file (f) – It Share file to other privileges which is already stored

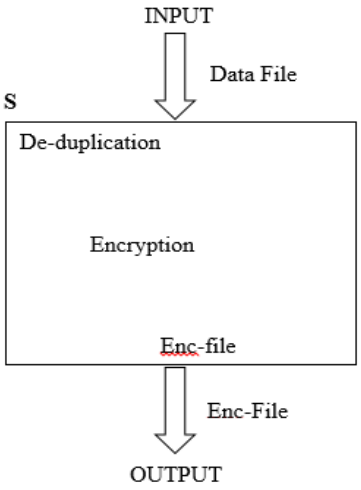

Fig 3 Input and Output in System

VI. RESULT ANALYSIS

In this system we give file as input to the system. Our proposed system is divided in following parts. First we have focused on creating a GUI for the proposed application. This GUI serves as a platform to users for browsing the data. Next for any data encryption or decryption is done by generating hash values and keys, which also checks that the file is belonged to the respective user. After this upload/download part will focus on the changes in data part where the user is able to modify the data, share the data and delete the data.

We have created the proposed system on java platform with the help of MySQL database to catch the users and files stored. This structure uses the Windows7 operating system. Then we have uploaded different sizes of files to note down different execution time, we have used JfreeChart and received the output in graph format. We have find the execution time of the system and the following parameters are checked. Results are calculated on local host

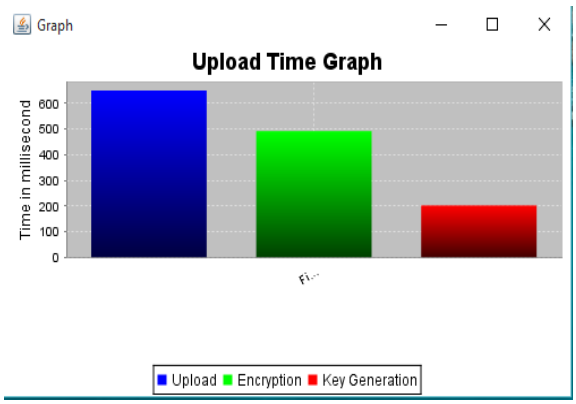

Fig 4 Time required for Upload, Encryption and Key Generation

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Figure 4 shows the total time required to upload the file this includes time for deduplication check. When user modifies the data and the revised file is again saved in the cloud is also included in this time. It again shows the time required for encryption and key generation for upload of file.

Figure 5 shows the total time required to download a file. X axis shows the size of file and Y axis shows the time required to download. Actually when file is uploaded that time the syst0em generates the key in order to encrypt the data and using that data it checks the duplicate copies if any, then uploads the file.

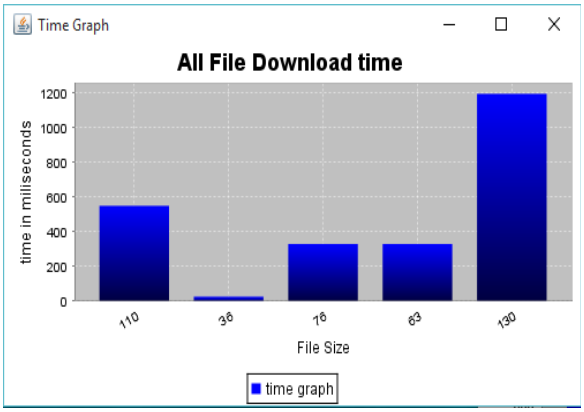

Fig 5 File size Vs Time for Download

Again when user wants to download the file then the system has to decrypt the data using the same key which is used for encryption. So the download time also includes the time required for decryption. When user requests to download the file then stored file must have to be decrypted with the key stored in dataset.

Figure 6 shows the time required for download and decryption

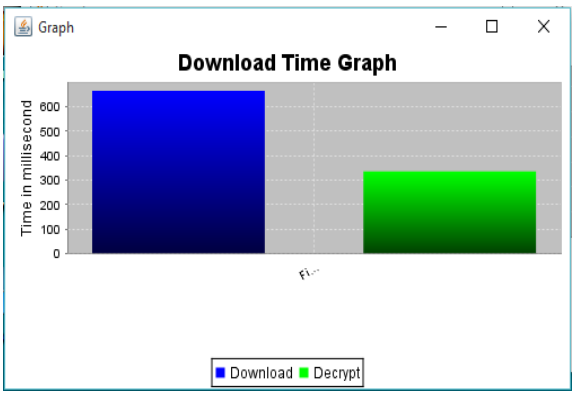

Fig 6 Time for download and decryption

Decryption time may depend upon the size of the file also.

Figure 7 shows the storage required for file when duplicate copies present in system and when the duplicate copies are removed and only single instance is stored. To find the result of this, we upload the files of same or varying file size and calculated result without checking the duplicates and again for the same files we have checked the duplicates and removed it and stored only one instance of duplicate copies. The result shows deduplication saves the space.

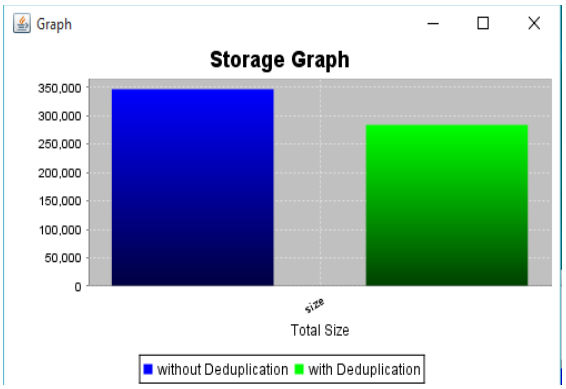

Fig 7 without Deduplication Vs with Deduplication

Figure 8 shows time required to upload the file after making the changes in the file already present in the database. Here we want to show that the proposed system will require less time than the existing system to do the procedure. Where x axis shows the existing and proposed approach and y axis show the total time required to upload the data.

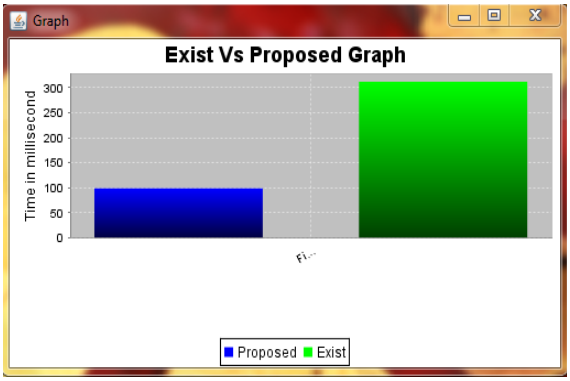

Fig 8 Existing Vs Proposed

VII. CONCLUSION

This method requires less time and provides a more flexibility to the user as compared to previous versions and at the same time it achieves high level confidentiality to the data which is stored in the system. After doing changes in the file the user is able to perform the deduplicate check in order to achieve our proposed aim. All users are able to provide the authentication with the help of privilege, token of file and can retrieve or modify the data with the help of convergent key. Also the list of users are allowed to share their files with other users with some defined rules. So this paper is achieving the confidentiality to the data, supporting authorized data deduplication and flexibility to the user with more options such as to add new contents to the file and check deduplication, delete the file from the system and remove the reference of the respective file. The result shows it requires less time.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, Hybrid Cloud Approach for Secure Authorized Deduplication, Parallel and Distributed Systems, IEEE Transactions on Volume: 26, Issue: 5, pages 1206-1216, 2014

[2] Jin Li, Xiaofeng Chen, M Li, Jingwei Li, Patrick P.C. Lee, and Wenjing Lou, Secure Deduplication with Efficient and Reliable Convergent Key Management, IEEE transactions on parallel and distributed systems, vol. 25, no. 6, pages 1615-1625,2014

[3] Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang and Yang Xiang, Mohammad Mehedi Hassan, Secure Distributed Deduplication Systems with Improved Reliability, Computers, and IEEE Transactions on Computers (accepted) pages 1-11, 2015

[4] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs ofownership 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.

[5] Chun-Ho Ng, Patrick P. C. Lee, RevDedup: A Reverse Deduplication Storage System Optimized for Reads to Latest Backups, Proc. of APSYS, 2013

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

[7] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-lockedencryption and secure deduplication. In EUROCRYPT, pages 296–312, 2013.

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

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

[10] 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.