ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

EFFICIENT DEDUPLICATION SCHEME FOR HYBRID IaaS CLOUD ARCHITECTURE

[1]Joshla L Soni, [2]Dr.Deepa A.J.
[12]Ponjesly College of Engineering, Nagercoil, TN, India
Email: joshlasoni@yahoo.com

Abstract---- **Data deduplication is the most significant data compression method for removal of replica copies , and is used in storage of cloud to reduce the storage space and also save bandwidth. To save the secrecy of sensitive data in deduplication, the convergent encryption technique is proposed to encrypt the data before sending to the cloud. To protect data security, in this paper we address the problem of certified data deduplication. From traditional way of deduplication systems, the disparity rights of users are always considered in duplicate check besides the data itself. We have a new deduplication constructions supporting certified duplicate check in a hybrid cloud architecture. Security analysis says that our scheme is most secure.**

Index Terms— **Deduplication, certified duplicate check, secrecy, hybrid cloud**

I. INTRODUCTION

The cloud helps to get various information from any part of the world at any time. Data management is measurable in cloud computing, deduplication [5] has been a significant technique recently. Data deduplication is a expert data compression method for eliminating replica copies in cloud. The method helps to improve storage utilization and can also can be used in network data transfers to decrease the number of bytes that can be sent. Instead of having multiple data copies with the same content, deduplication eliminates repeated data by keeping only one physical copy and referring other repeated data to that copy. Deduplication is mainly two types namely the file level and the block level. For file-level deduplication, it makes the elimination of replica copies of the same file. For the block level, it makes elimination of replica blocks of data that happens in non-identical files.

In Traditional encryption, data secrecy is unsuited with data deduplication. It needs various users to encrypt their data with their own keys. So, the same data copies of various users will direct to dissimilar cipher texts, making deduplication impossible. Convergent encryption [4] is proposed to implement data confidentiality while making deduplication possible. It encrypts/decrypts a data copy with a convergent key, which is got by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption are done, users maintain the keys and send the cipher text to the cloud.

As the encryption operation is done in deterministic manner and is got from the data content, matching data copies will be generating the same convergent key and also the same cipher text. The user will find out a duplicate of the file only if there is a copy of the file and a matched privilege which is stored in cloud. To save cost and efficiently manage, the data will be moved to the storage server provider in the public cloud with particular privileges and the deduplication method will be applied to store only one copy of the same file.

II. RELATED WORK

A. Secure deduplication.

With the advancement of cloud computing, secure data deduplication has more significance in the recent days. Yuan and Yu [11] proposed a deduplication system in the cloud storage to reduce the storage size of the tags for integrity check. Bellare et al. [1] showed how to protect the data confidentiality by transforming the predictable message into unpredictable message. In their system, another third party called key server is introduced to generate the file tag for duplicate check. Stanek et al. [7] presented a novel encryption scheme that provides differential security for popular data and unpopular data. For popular data that are not particularly sensitive, the traditional conventional encryption is performed. Another two layered encryption scheme with stronger security while supporting deduplication is proposed for unpopular data. In this way, they achieved better tradeoff between the efficiency and security of the outsourced data.

B.Convergent encryption.

Convergent encryption [4] ensures data privacy in deduplication. Bellare et al. [2] formalized this primitive as

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

message-locked encryption, and explored its application in space-efficient secure outsourced storage. Xu et al. [10] also addressed the problem and showed a secure convergent encryption for efficient encryption, without considering issues of the key-management and block level deduplication. There are also several implementations
of convergent implementations of different convergent encryption variants for secure deduplication (e.g. [6], [8], [9]).

C.Twin clouds architecture.

Bugiel et al. [3] provided architecture of twin clouds for secure outsourcing of data to an entrusted commodity cloud. Zhang et al. [25] also presented the hybrid cloud techniques to support privacy-aware data intensive computing. The private cloud is always said as to be honest but curious.

III. CONVERGENT ENCRYPTION METHOD

Convergent Encryption is used to enforce data confidentiality while making deduplication feasible. Since the encryption operation is deterministic and is derived from data content, identical data copies will generate same convergent key and hence same cipher text. It encrypts or decrypts data copy with convergent key which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain keys and send the cipher text to the cloud. To prevent unauthorized access, a secure Proof of Ownership (POW) protocol is also needed to provide the proof that the user indeed owns the same file while a duplicate is found. After the proof, subsequent users with the same file will be provided a pointer from the server without need to upload the same file. User can download the encrypted file with the pointer from the server which can only be decrypted by the corresponding data owners with their convergent keys.

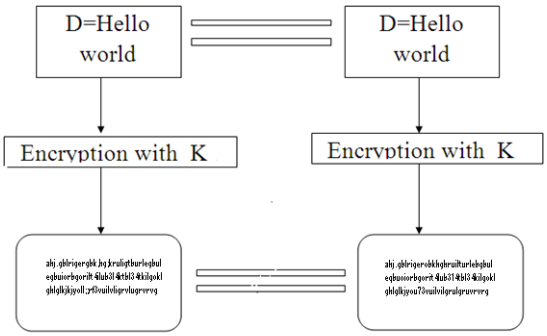

Figure 1 Convergent Encryption Method

In Figure 1 Convergent Encryption method, the user gives a data D that is Hello World which is encrypted with a convergent key K. For both the users, the cipher text that is got is always the same value in the encrypted form. So it's very easy to **find** out the duplicates with the help of the cipher texts.

The various advantages of convergent encryption method are,

- Data confidentiality is maintained.
- It makes overhead minimal.
- Reduce storage space

The system is mainly planned to explain the differential privilege problem in secure deduplication. The security is always calculated in two aspects, they are, the approval of duplicate check and the secrecy of data. Some tools are used to create the secure deduplication, which are implicit to be secure. The tools are mainly the convergent encryption scheme, symmetric encryption scheme.

Symmetric encryption.

In Symmetric encryption user has a common secret key for both encryption and decryption.

Convergent encryption.

Convergent encryption [2], [4] has data privacy in deduplication. A user gets a convergent key from each and every actual data copy and encrypts the data copy with the convergent key. At this time, the user also gets a tag for the data copy, where the tag will be helpful to find out the duplicates. To find out the duplicates, the user will first send the tag to the server to check if the actual copy is already stored. Both the encrypted data copy and its related tag will be stored on the server side for future references.

Data secrecy.

Illegal users without proper rights or files must be prohibited from access. The main goal of the attacker is to get

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

back and recover the files that are not belonging to them. In convergent encryption, a superior level secrecy is clear and achieved.

IV. FRAME WORK FOR SECRECY DEDUPLICATION

In this framework for secrecy deduplication, we can clearly see that, a cloud user first splits the data, encrypts and uploads to the cloud server. The cloud server then authenticates it and sends to the Metadata Manager. From the metadata manager, it is stored to the cloud storage provider which has both Private cloud and also Public cloud. Only Authorized users can access from the private cloud with their keys and other users can access from the public cloud.

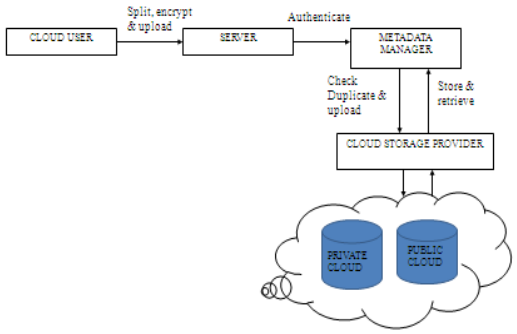

Figure 2 Framework of Secrecy Deduplication

Figure 2, shows the framework of secrecy deduplication which consists of both Public cloud and Private cloud in it. Hybrid cloud helps to make more secure storage in it. Users can access the files in the Public cloud and view it easily. But accessing the Private cloud needs authorization from the users. The key can be got only for the private cloud users. The cloud user will split the file, encrypt and upload into the Server. Then the server will authenticate whether it can be uploaded or not and send to the metadata manager. The metadata manager checks the file again and sends to the Cloud Storage Provider (CSP) which has a hybrid cloud. Hybrid cloud has both public and private clouds in it. While saving the file it asks whether to save in public or in private cloud. For storing in private cloud the authorized users should have the privileges and needs keys to download it. For storing in public cloud the registered users can do. Files in public cloud can be viewed by both public cloud users and also by private cloud user. Once the files are stored in the hybrid cloud, we can download or retrieve it with the help of the keys.

If the user requests for a file download, then the metadata manager checks whether the user is authorized user or registered user. If it's an authorized user, both the private cloud and public cloud files can be downloaded by that user. If it's a registered user then only public cloud files can be viewed and downloaded. One the authentication is over, the files which is stored in the hybrid cloud will be allowed to download by the user. It is then send to the cloud storage provider and then to the metadata manager for authentication and then to the server and finally reaches the user.

Administrator has the rights to monitor all the files that are uploaded and also to delete the files from both private and public cloud at any time. Administrator has a login address to monitor the servers which the files are uploaded and the time taken for each download. Administrator also monitors any failure time that is made during the file upload. Administrator is the main person to monitor all the files that are uploaded, viewed and also downloaded from the hybrid cloud. Hybrid cloud is always secure to use. Only authorized users can upload, view and also download the files from the Hybrid cloud.

Private cloud.

Private cloud users can have the authorized access. The private keys for the privileges and authorized users are maintained by the private cloud. It provides keys to the users who make the requests. The private cloud always allows user to submit files which can be stored and easily computed.

V. PSEUDO CODE

The pseudo code for finding out the duplicate file is given below.

```
begin
    Create a cloud environment
    Encrypt file & Upload file
 Create a Hybrid cloud
    Insert either in private or public cloud
if(a[m].equals(b[m])
        Duplicate File
    else
      Upload File
   Download & decrypt
end
```

In this pseudo code, first a cloud environment should be created for splitting and encrypting and also uploading the files in the cloud. Next a Hybrid cloud is created. In a hybrid cloud, both the private and also public cloud is present. Private cloud is only for the authorized users. Public cloud is common for all the people. The person who uses private cloud

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

can also view the files from the public cloud. But the person using public cloud can never view the files of the private cloud. A file is uploaded in the public cloud by a person and the same file is tried to be uploaded by another person, the it says that 'File already exists". So it is not possible to upload the file. If the file which is about to be uploaded is not present in any clouds then it is easy to upload. This pseudo code says the duplicate checking of files in the cloud.

VI. Performance Analysis

Performance Analysis involves gathering formal and informal data to help customers and sponsors define and achieve their goals. Performance analysis uncovers several or barriers to successful performance and proposing a solution system based on what is discovered. The accomplishment of a given task measured against pretest. Known standards of accuracy, completeness, cost and speed. In a contrast, performance is deemed to be the fulfillment of an obligation, in a manner that releases the performer from all liabilities under the contrast.

The definition for performance analysis given is : A specific performance based needs assessment technique that precedes any design or development activities by analyzing the performance problems of a work organization. It has three basic steps in the performance analysis process: data collection, data transformation and data visualization. Data collection is the process by which data about program performance and obtained from an executing program. Data are normally collected in a file, either during or after execution, although in some situation it may be presented to the user in real time. Data transformation is applied, often with the goal of reducing total data volume. Transformation can be used to determine mean values or other higher order statistics or to extract profile and counter data from traces.

To evaluate the deduplication to be made, experiments are conducted in five machines with different configurations and in different operating system. The upload can be done in steps such as Token generation, Duplicate check and Encryption. Both start time and end time are recorded. The average time is taken for each data set. The two parameters used are namely, File size, Number of stored Files.

1. File Size

The file sizes are evaluated by uploading 100 unique files of same file size and record the time. The worst case scenario is evaluated by uploading all data files. The average time is plotted in the Figure 3. The file metadata is only used for duplicate check and the time remains constant. The file size increases and the overhead decreases.

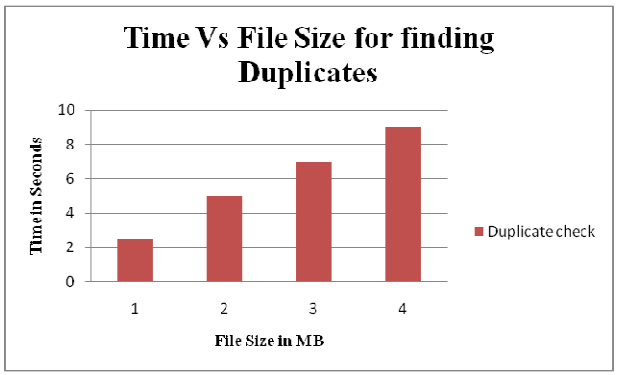

Figure 3 Time Vs File Size for finding Duplicates

In the Figure 3, the File sizes are given in X axis and Time is given in Y axis. When the files are uploaded and checked for duplicates, as the file size increases the duplicate checking time also gets increased in parallel. The max time taken for checking the duplicate is 9 seconds for a 4 MB file.

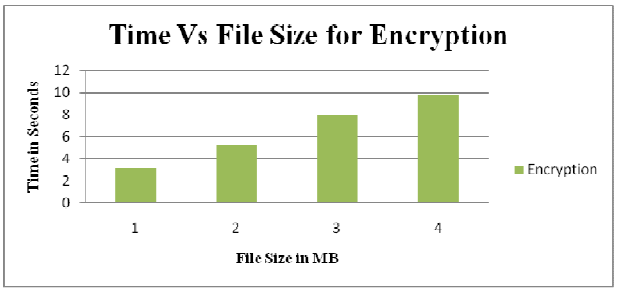

Figure 4 Time Vs File Size for Encryption

In the Figure 4, the File sizes are given in X axis and Time is given in Y axis. When the files are uploaded and checked for encryption, as the file size increases the encryption time also gets increased in parallel. The max time taken for encrypting is 9.7 seconds for a 4 MB file.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

2. Number of stored files

The number of stored files is evaluated by uploading 10MB unique files in 100 numbers. The time for each file upload is recorded. The steps remain constant. Duplicate check is stable as the collision is low.

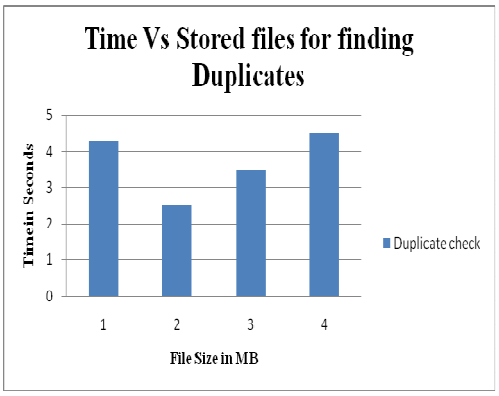

FIGURE 5 TIME VS STORED FILES FOR FINDING DUPLICATES

In Figure 5, the File sizes are given in X axis and Time is given in Y axis. When the files are uploaded and checked for duplicate, as the file size increases the duplication time will remain stable as there is low collision in it.

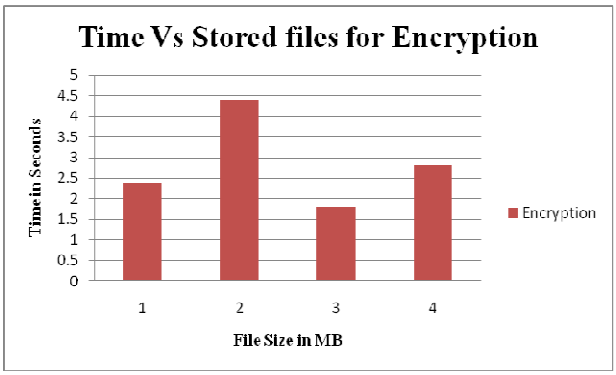

FIGURE 6 TIME VS STORED FILES FOR ENCRYPTION

In Figure 6, the File sizes are given in X axis and Time is given in Y axis. When the files are uploaded and checked for encryption, as the file size increases the encryption time also increases according to the number of files given.

VII. CONCLUSION

The notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

The data in cloud will be exactly one copy and if any failures or data loss occurs, it's a very huge damage. So in future it should be placed in the distributed server which should be in more than one place. If any cloud damage or loss happens then we can upload the contents from the distributed server. Instead of having more number of duplicates in the cloud, it's better to be stored in four servers and can be got when needed. Each server is in distributed manner that it will be placed in some remote servers. Only the administrator can upload the contents in distributed servers and also be aware about any damage to the cloud. Periodical checkups for the data in cloud are to be made.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–194.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

[3] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, "Twin clouds: An architecture for secure cloud computing," in Proc. Workshop Cryptography Security Clouds, 2011, pp. 32–44.

[4] 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 Proc. Int. Conf. Distrib. Comput. Syst., 2002, pp. 617–624.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 1, January 2016

[5] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," in Proc. 1st USENIX Conf. File Storage Technol., Jan. 2002, pp. 7.

[6] 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 Proc. 3rd Int. Workshop Secutiry Cloud Comput., 2011, pp. 160–167.

[7] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," Tech. Rep. IBM Research, Zurich, ZUR 1308-022, 2013.

[8] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. 4th ACM Int. Workshop Storage Security Survivability, 2008, pp. 1–10.

[9] Z. Wilcox-O Hearn and B. Warner, "Tahoe: The least-authority filesystem," in Proc. ACM 4th ACM Int. Workshop Storage Security Survivability, 2008, pp. 21–26.

[10] J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient clientside deduplication of encrypted data in cloud storage," in Proc. 8th ACM SIGSAC Symp. Inform. Comput. Commun. Security, 2013, pp. 195–206.

[11] J. Yuan and S. Yu, "Secure and constant cost public cloud storage auditing with deduplication," IACR Cryptology ePrint Archive, 2013:149, 2013. IEEE Transactions on, Vol. 15, Issue 1.

AUTHORS BIOGRAPHY

JOSHLA L. SONI was born on 14th January 1991. Received her B.Sc degree in Computer Science from Women's Christian College, Nagercoil which is under M.S University in 2011, M.C.A degree in Sun College of Engineering and Technology which is under Anna University, Chennai in 2014 and now presuming her M.E degree in Ponjesly College of Engineering which is under Anna University, Chennai. Her area of interests includes Mobile Computing, Security in Cloud Computing, Distributed Storage and Deduplicaton.