International Research Journal of Computer Science (IRJCS)
Issue 04, Volume 3 (April 2016)
ISSN: 2393-9842
www.irjcs.com

A MECHANISM FOR PREVENTING DUPLICATE FILES IN CLOUD

Komathi.R *
Department of CSE,
Kingston Engineering college,
Vellore, Tamil Nadu

Deepapriya.V*
Department of CSE,
Kingston Engineering college,
Vellore, Tamil Nadu.

Mohana priya.M*,
Department of CSE,
Kingston Engineering college
Vellore, Tamil Nadu.

Natteshan N.V.S,
Assistant Professor/ CSE,
Kingston Engineering college
Vellore, Tamil Nadu

Abstract— cloud computing provides a way of storing a voluminous data and can be easily accessed anywhere. This work deals about the prevention of a duplicate file storage in cloud. Here there are three important components in our system namely the owner of the data who generated and will store it in the cloud, the user can be a valid person who will download the file after providing the suitable credentials namely the user will provide the encrypted key he obtained through his mail to download a file from the cloud, the third important component is the cloud. Here the file is provided a unique value which can be used to identify it SHA is used for this purpose. Then for the secure encryption and decryption RSA algorithm is used. The user needs to provide this key to decrypt and obtain file ensure confidentiality of data. A owner can upload a file and when he uploads a duplicate copy of the file the cloud server will notify a error. This duplicate prevention in a cloud ensures that the memory space is effectively utilized thereby reducing the processing overhead.

Key words- cloud computing, Secure Hash Algorithm (SHA), encryption, decryption, RSA, confidentiality.

I. INTRODUCTION

Cloud computing can be considered as the use of servers by the user for storing which has high availability. Cloud as a service is a emerging technology and this particular work focuses on the utilization of the cloud service for storing file effectively in the way that there is a elimination of duplicate files in the process of storage. This intern has an advantage of effective memory utilization. Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent.

Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Duplication removal can take place at the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can be done at the block level, which eliminates duplicate blocks of data that occur in non-identical files. Here in this research work both the deduplication at the file level and block level is done. Here in this work a encryption technique combined with duplication prevention is done. Here as the encryption and decryption key will not be generated for the duplicate files. The reason behind not generating a different key for the duplicate files is due to the generation of a identity value for the file by the SHA algorithm.

A. PURPOSE OF THE WORK

To provide convergent encryption technique to enforce data deduplication.
To provide security and privacy concerns for both inside and outside attacks.
To provide a secured proof of ownership protocol is to find duplicate data for same convergent key in plain and cipher text

B. OBJECTIVE

To increase storage of data shared by users
To eliminate duplicate copies of data
To provide confidentiality of data secured in cloud.

The overall organization of this paper is as follows, section II gives the detail about the works related to this deduplication prevention, section III gives the design of the system, section IV describes the experimental setup and section V gives the evaluation of the experiments in a java environment, section VI concludes the work.

II. RELATED WORK

This section gives the details about the related works done in this duplication prevention. S.Keelveedhi et al. in their work Dupless: Server aided encryption for deduplicated storage propose an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees.

T.Ristenpart et al. in their work Message-locked encryption and secure deduplication formalize a new cryptographic primitive that we call Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure deduplication (space-efficient secure outsourced storage), a goal currently targeted by numerous cloud storage providers.

G.Neven et al in their work, Twin clouds: architecture for secure cloud computing promises a more cost effective enabling technology to outsource storage and computations

H.Zhu et al. in their work, Private data deduplication protocols in cloud storage in their research proposes a new notion which we call private data deduplication protocol, a deduplication technique for private data storage is introduced and formalized. Intuitively, a private data deduplication protocol allows a client who holds a private data proves to a server who holds a summary string of the data that he/she is the owner of that data without revealing further information to the server.

III. A MECHANISM FOR DUPLICATION PREVENTION IN CLOUD

In this section we describe the overall architecture of the system.

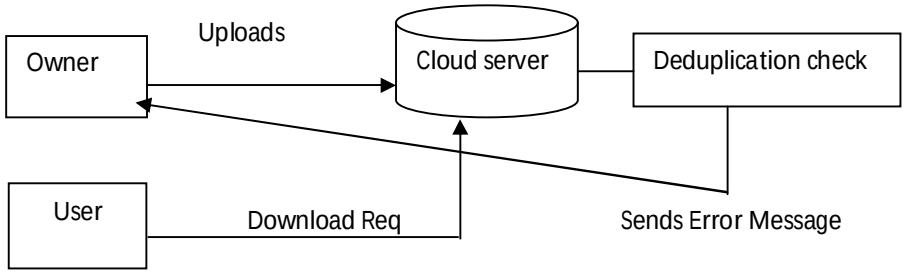

Fig. 1 Overall Architecture of the Deduplication Mechanism

A. DESCRIPTION OF THE ARCHITECTURE

Here the owner performs the uploading of file to the cloud server. Then a unique value is generated by the use of SHA algorithm. Then a Token is generated by using ID. Then a RSA algorithm is used to encrypt the file and will be checked for duplication in the naming and also the block level duplication check is done. Then if its not a duplicate file then the file to be uploaded by the owner will be uploaded. If there is any duplication found then a error will be send by the cloud. Then the process of uploading is a success. Then the user can perform a download request. Then the server asks for a valid mail-id and then sends to that mail-id the key for decrypting the file. Then the user uses it and downloads the file.

IV. EXPERIMENTAL SET UP

The experiments were conducted with large amount of input files and for each file the Tag value is computed and a Encryption and decryption key is generated by RSA Algorithm. The duplication check mechanism is designed which prevents duplicate file storage. The implementation of the cloud server and the duplication check mechanism is done by using Java Net beans IDE. The number of owners who did the upload process and the number of duplicate files detected and the number of files downloaded and the running time is computed for each tasks and is tabulated in the experimental evaluation.

V. EXPERIMENTAL EVALUATION AND RESULTS

The experiments were conducted and the results obtained are tabulated below. First the number of owners and users and the files are tabulated below

TABLE I
DETAILS OF THE COMPONENTS IN THE SYSTEM

NO OF CLOUD SERVER	NO OF OWNERS	NO OF USERS	NO OF FILES
1	10	50	100

Here the numerical count of the cloud server and number of owners, users and the details about their files are tabulated. Here the owners perform the process of uploading and users downloads the files

TABLE II
THE NUMBER OF FILES AND ITS TAG AND TOKEN VALUE

FILE ID	FILE TYPE	TAG VALUE	TOKEN VALUE
FID_001	PDF	4	45
FID_002	PDF	6	47
FID_003	DOC	3	52
FID_004	DOC	5	51
FID_005	PDF	5	47
FID_006	TXT	4	41

The above table describes the type of file used and the TAG value computed by using SHA and token value generated by TAG, Id of the file.

TABLE III
DETAILS OF FILE UPLOAD, DOWNLOAD AND DUPLICATE FILES

TOTAL NO OF FILES	FILES UPLOADED	FILES DOWNLOADED	DUPLICATE FILES
100	80	35	20

The above table III describes the number of files uploaded and downloaded and the number of duplicate files detected.

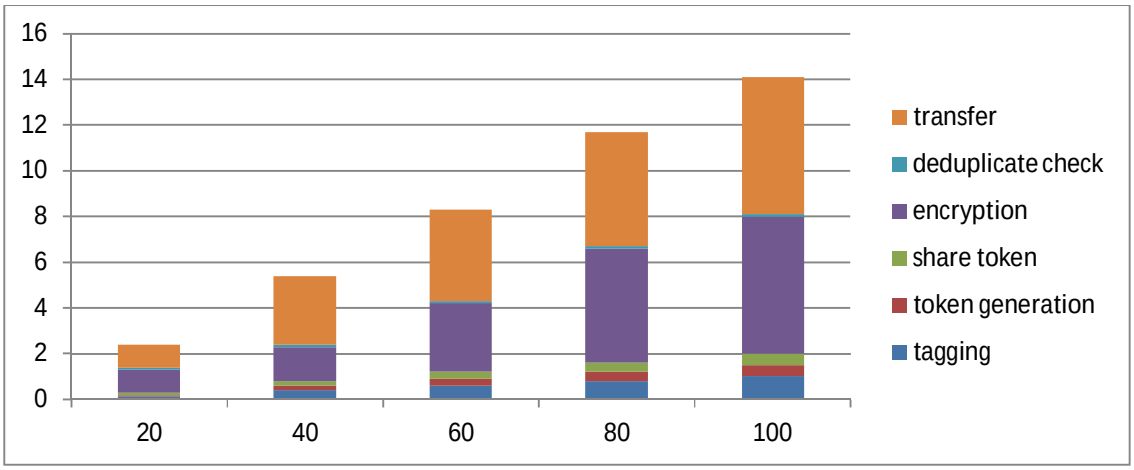

Fig. 2 Bar graph showing the time for performing various tasks

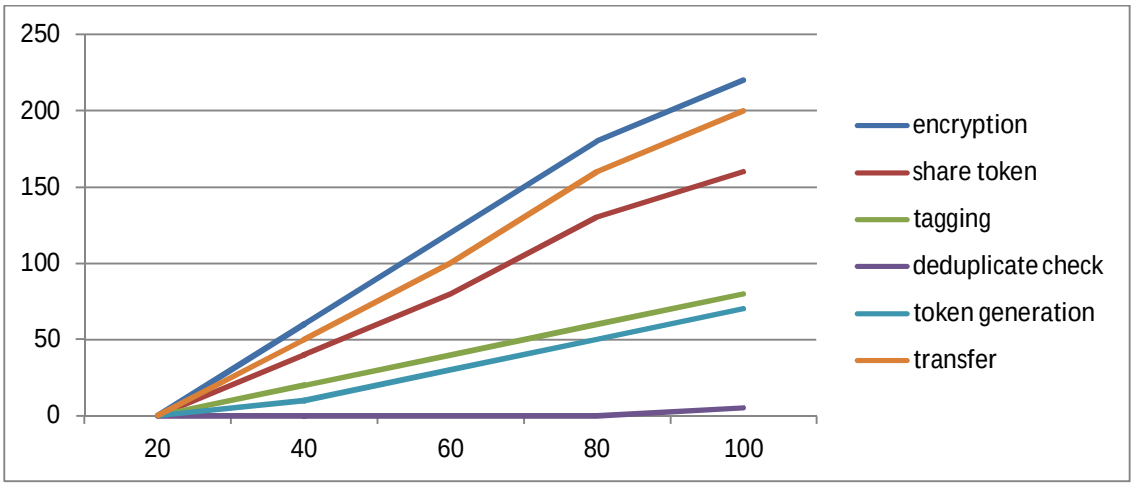

Fig. 3 A performance plot for the various tasks.

VI. CONCLUSION AND FUTURE WORK

A mechanism for the duplicate file upload is implemented in cloud environment and from the performance evaluation the tasks are performed quickly and there is an effective computation of the duplicate files by the detection module and there is a good and effective utilization of memory. The future research can be focused on effective utilization of memory still to further level and the encryption algorithm can be changed to include some other algorithm and some other technique to generate a value for file may be proposed.

REFERENCES

[1]. Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26, NO. 5, MAY 2015.
[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]. M. Bellare and A. Palacio, "Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks," in Proc. 22nd Annu. Int. Cryptol. Conf. Adv. Cryptol., 2002, pp. 162–177.
[4]. P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, pp. 29–40.
[5]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proc. ACM Conf. Comput. Commun. Security, 2011, pp. 491–500.
[6]. W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proc. 27th Annu.ACM Symp. Appl. Comput., 2012, pp. 441–446.
[7]. 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.
[8]. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," IEEE Comput., vol. 29, no. 2, pp. 38–47, Feb. 1996.
[9]. 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
[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.