Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.130 – 135

A Hybrid Cloud Model for Detection and Prevention of Duplicate files with Improved Security Using Dual Encryption

Diyab A K[1], Ambarish A[2], Unnikrishnan S Kumar[3]

[1]Department of Computer Science and Engineering, Malabar CET, India
[2]Department of Computer Science and Engineering, Malabar CET, India
[3]Department of Computer Science and Engineering, Malabar CET, India
[1] diyabak@gmail.com; [2] ambruappat@gmail.com; [3] uksknair@gmail.com

Abstract— The ever increasing volume of data in the cloud storage is the crucial challenge faced by the cloud computing environment today. Data deduplication can be used as the solution for this problem. Deduplication eliminates the redundant copies of file from the cloud by keeping only single copy of the file using the convergent encryption scheme. The authorized deduplication can be carried out during the duplicate checking by considering the user privileges. This is done by generating a file token which is bound with the user privilege and the file tag. For reducing the possibility of brute force attacks and the confirmation of file attacks, Dual encryption approach is used by generating the encryption keys from the data itself.

Keywords—deduplication, file tag, convergent encryption, hybrid cloud, dual encryption, file token.

I. Introduction

Cloud computing is the most popular any time available resources which provides unlimited virtual resources for data storage and their processing. The most important feature of cloud computing is that it can provide us the on demand and low cost online resources for storage and processing. Cloud platforms hide the platform and implementation details. As cloud computing becomes popular large amount of data is being stored in the cloud. Each data will be bounded with some access right so that the specified user with the privilege can only access the data. Nowadays the cloud computing platforms are prevalent and used by most of the users so that the amount of data stored in the cloud also increased so that the critical challenge in the cloud computing environment is the exponential growth on amount of data stored in the cloud. As per the analysis report by IDC the amount of data in the world will reach 40 trillion gigabytes in 2020. As the volume of data in the cloud increases, the cost overhead for the management and the processing of data also increases. For example storage infrastructure cost, human administration cost, data management cost etc., so that the management of the ever increasing volume of data in cloud is a critical challenge.

Data deduplication is a well known data compression technique for reducing the amount of data stored in the storage space. Data deduplication keep only single physical copy of the redundant data and provide a link to that copy for all the repeating copies of data rather than storing the duplicate copies. Data deduplication is a popular technique used in the backup process by storing only the single copy of the redundant data. The deduplication process can be applied to both file level and block level. In file level deduplication we keep only a single copy of a file and all other repeating copies will be provided with a link to the single copy. In block level deduplication it eliminates the duplicate blocks inside the files. The overhead in the network can be eliminated by applying deduplication scheme by which the amount of data transferring through the network can be reduced and we can improve the network bandwidth.

Before outsourcing any data to the cloud, the data must be properly encrypted for the better security of the data. The traditional encryption scheme such as symmetric encryption cannot make the deduplication possible. In symmetric encryption each user uses their own secret key to encrypt their data. So that the same data or file will have different ciphertext, this makes the deduplication technique impossible. Data deduplication on data can be making possible by using the convergent encryption scheme. In convergent encryption the key is generated from the data itself using some cryptographic hash function so that the same data will have similar ciphertext and this makes deduplication possible.

This system is different from the traditional deduplication system as here we are considering the privileges of the user during the duplicate check. The duplicate check tokens are generated using a file tag which is generated from the file and the privileges of the user. Unlike conventional cloud architecture we use hybrid cloud architecture. The hybrid system consists of three entities such as a user, private cloud and the public cloud service provider. The complete security of the data which are uploading to the public cloud is attained by dual encryption scheme. In dual encryption the keys are generated from the data itself using two cryptographic hash functions. After creating the keys the data will be encrypted twice using the keys under convergent encryption.

II. BACKGROUND

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

A. Cloud computing

Unlimited virtualized storage resources and computational environment can be provided by cloud computing platforms as services through the internet. Today’s cloud service providers guarantee highly available and parallel computing resources with low costs. But the most important problem faced by the cloud computing platform is the increasing amount of data in the cloud. Data deduplication is a well known compression technique for solving this problem.

B. Deduplication

Based on the data granularity, deduplication technique can be classified into two major categories: file level and block level deduplication. When the deduplication is carried out in client side, then it is called as source-based deduplication. If it is carried out isn the server-side it is called as target based deduplication. In source based deduplication the client derive a hash value from the data and send it to server for duplicate check. Source based deduplication approach reduces the channel overhead where as in target level deduplication scheme does not decrease the communication overhead.

C. Encryption scheme

The traditional encryption scheme such as symmetric encryption will not provide support for the deduplication procedure. Deduplication can be achieved by the convergent encryption scheme. Here the encryption key is generated from the hash of the plaintext. The simplest form of convergent encryption can be explained as follows: Alice derives the hash value from the message (M) to be encrypted as the convergent key. Such that K=H (M) where, H is the cryptographic hash function applied to the message. Hence, C=E(K,M)=E(H(M),M) where, E is the symmetric encryption algorithm applied to M using K. Using this technique, two users having two identical plain text will obtain two identical ciphertext. Furthermore, the encryption keys are retained and protected by the users in the private cloud.

III.RELATED WORK

So much of systems have been developed for the secure storage of data in the cloud. But the traditional encryption schemes are not suitable for deduplication over data in the cloud. The backup solution [1] uses an algorithm to backup the data on laptops or home computers using the deduplication procedure. The system uses convergent encryption for encrypting the data in the client side for the security of data before backup. The convergent encryption supports for deduplication so that it increases the speed of backups and reduces the storage space requirements.

Message locked encryption scheme [2][3] supports for deduplication where the encryption key is generated from the known set. However it is inherently subject to brute force attack that can reconstruct the original file from the known set. In order to overcome this, the keys will be maintained in a separate key server [2]. Proof of security [5] against concurrent attacks uses an identification scheme in which each user is provided with a secret key for authentication. The prover/user sends the key to the verifier/server and the verifier checks whether the prover is authorized user or not so that the user without having the privilege cannot access any data in the server.

The twin-cloud architecture [6] provides an secure outsourcing of data to the cloud provider. Before uploading or downloading the data in the untrusted cloud, the user needs to interact with the trusted cloud. The trusted cloud gives access to the public cloud for the user who has the authorization privileges. Role based access control [7][10] mechanism used by commercial firms and corporations provides security for the data. In this control system each user is provided with a particular role so that the user without the corresponding privilege (role) cannot access the information.

IV. PRELIMINARIES

This section gives are view on the secure primitives used in the secure deduplication system.

A. Symmetric encryption

This is a traditional encryption scheme in which a common secret key 'K' is used to encrypt and decrypt information. Symmetric encryption procedure involves three primitive functional operations such as a key generation function, encryption function and a decryption function.

- KeyGen(1λ) ⟹ K, is the symmetric key generation algorithm that generate the key K using the security parameter 1λ.
- EncCE(K,M) ⟹ C, is the symmetric encryption algorithm that generate the ciphertext C using the key k and message M.
- Dec(K,C) ⟹ M, is the symmetric decryption algorithm that reconstruct the original message M using the secret key K.

B. Convergent encryption

The deduplication in cloud platform can be done by using the convergent encryption [3] scheme. The convergent encryption provides data confidentiality in deduplication. The data owner derives the convergent key from the original data and encrypts the data with the convergent key. The convergent encryption procedure can be defined with four primitive functions.

- KeyGen(M) ⟹ k is the algorithm which generate the convergent key k from the data copy M.
- EncCE(K,M) ⟹C is the encryption algorithm that generate the ciphertext C by taking the convergent key K and message M as input.
- DecCE(K,C) ⟹ M is the decryption algorithm that reconstruct the original message M by taking the convergent key k and ciphertext C as input.
- TagGen(M) ⟹ T(M) is the tag generation algorithm that generate tag from the message M.

The tag generation algorithm mentioned above, generate a tag from each message and we use the tag for duplicate check in the cloud server. If two data copies are identical then the tag generated from the data will also the same. In the duplicate check process, the user send the file tag to the cloud server to check whether the identical copy of the file tag is already there in the cloud server or not. If the same copy of the file tag is already exist in the cloud, then the user need not to upload the same file and furthermore the user get an access link to the file in the cloud server. There is a clear distinction between the file tag and the convergent key. Both are independently derived from the data and the tag cannot be used to get the convergent key and compromise the confidentiality of the data. The encrypted data and the file tag will be stored in the cloud server.

C. User identification

The identification of the user can be done in two steps: proof and verify. In the first step the user/prover can prove their identity to the verifier by sending some identification information. The user/prover input the private key associated with each user to the verifier. The verifier verifies the user by testing the public information related to the private key. In the final stage the verifier send accept or reject signal to the prover.

V. SYSTEM COMPONENTS

The secure cloud deduplication system uses hybrid cloud architecture which involves a user, a private cloud and public cloud.

A. User

The user uses the public cloud server for efficiently store their data in the cloud and access the data later. In a storage platform which supports deduplication, the user need to upload only a single copy of data and does not need to upload any duplicate copy of the same data to save the upload bandwidth and cloud storage. The data may be owned by the same user or different users. In this authorized secure deduplication system each user is provided with a set of privileges so that the user with the corresponding privilege can only access the file.

B. Private cloud

Private cloud is a new entity compared with the traditional deduplication architecture. This entity facilitates the users with a secure usage of cloud service. Specifically the computational procedures at the user/owner side are restricted. We cannot fully trust the public cloud. So the private cloud provides a fully trusted computational environment and work as an interface between the user and the public cloud. So the user first authenticate with the private cloud for uploading or retrieving data to or from the public cloud

C. Public cloud

Public cloud is the cloud service provider which provides us the online data storage facility. To reduce the cost of storage used for storing the data, the public cloud uses the deduplication strategy for eliminating the redundant data by keeping only unique copy of data. The public cloud will be always online and provide a large capacity of storage and computational power.

VI. SYSTEM ARCHITECTURE

The hybrid cloud deduplication system is focusing on the enterprise network consisting of group of authorized clients who uses the public cloud storage for storing their data by applying deduplication technique over the data. The hybrid cloud architecture of this system involves the components such as user, private cloud and the public cloud who provide us the storage and computational capabilities.

The privileges of the users will be maintained in the private cloud and the file token will be generated here for the duplicate check. The privileges of the users will be represented as privilege keys. So one user can have a privilege represented as a privilege key.

To support authorized deduplication the file token will be generated using the file tag which is sent by the user to the private cloud and the privilege key of the particular user. The system is setuped such a way that only the owner of the file can delete the file and all other users who is having the access privilege can only access the file.

Consider there are N number of users in the system and the privileges in the universe is defined as P={p1,p2,……pm}. for each privilege p in P, a private key kp will be selected. A user U with a privilege pu will be assigned a keys kpu. The privileges of the users will be stored and maintained in the private cloud in the form of privilege keys.

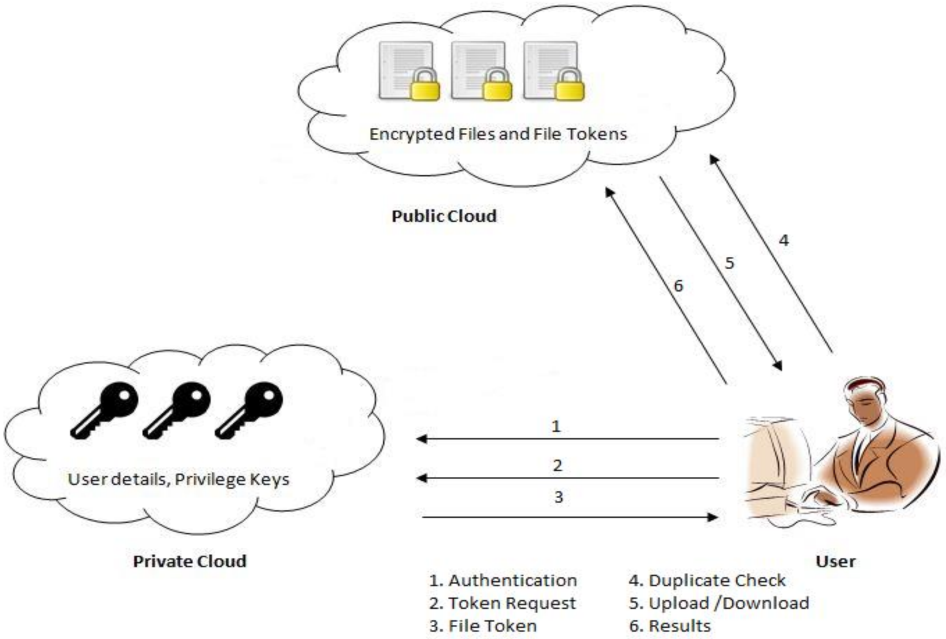

Fig 1: Architecture

A. Identification

The user sends the authentication parameters to the private cloud. The private cloud verifies the details and identify whether the user is authorized or not. If the user is an authorized one then the private cloud checks the privilege keys of the user which identifies the privilege of the particular user.

B. Uploading

Consider a user U with privilege set Pu wants to upload and share a file F with users having the privilege set PF.

- After identification the user derives a file tag ϕF = TagGen(F) to the private cloud server.
- The private cloud server accept the file tag and return file tokens ϕF,pu=TagGen(ϕF,kpu) which are generated by using the file tag (ϕF) and the privilege PU.
- The user then interact with the public cloud by sending the file token ϕF,pu.
- If a duplicate token is found, the response signal from the public cloud will be return to the user. The user then send the privilege set PF to the private cloud and the private cloud creates the file tokens for all the privilege keys in PF. This file tokens will be send to the public cloud. Then the file tokens of the file will be the union of the token for the privileges in the set PF and Pu.
- If no duplicate is found, a response signal from the public cloud will be returned. Then the user request for the file token for privileges in the set PF and the current user privilege Pu. The private cloud returns the file tokens to the user. The user module then generate the encryption keys from the file such as k1=TagGen(M) and k2=TagGen(M). Using the keys the file will be dual encrypted using convergent encryption scheme as follows: C1=EncCE(k1,M), C2=EncCE(k2, C1). The keys of the user will be then stored and maintained in the private cloud for security. The user then uploads the ciphertext along with the file tokens to the cloud server.

C. Retrieving

Consider a user who wants to download a file F, a request will be send to the public cloud by the user along with the file name. Upon receiving the request with the file name, the public cloud check whether the user has the privilege to download F. If the user has the privilege then the user will get the link to the file for access it.

VII. EVALUATION

In traditional scheme for every upload of a file, the file needs to be encrypted and transferred it to the public cloud. It increases the data usage and the processing time and also decreases the resource utilization. In this proposed approach if a duplicate is found while uploading a file, the file need not be encrypted and uploaded again.

The following graph (fig 2), explains the advantage of using this proposed approach in the cloud computing environment.

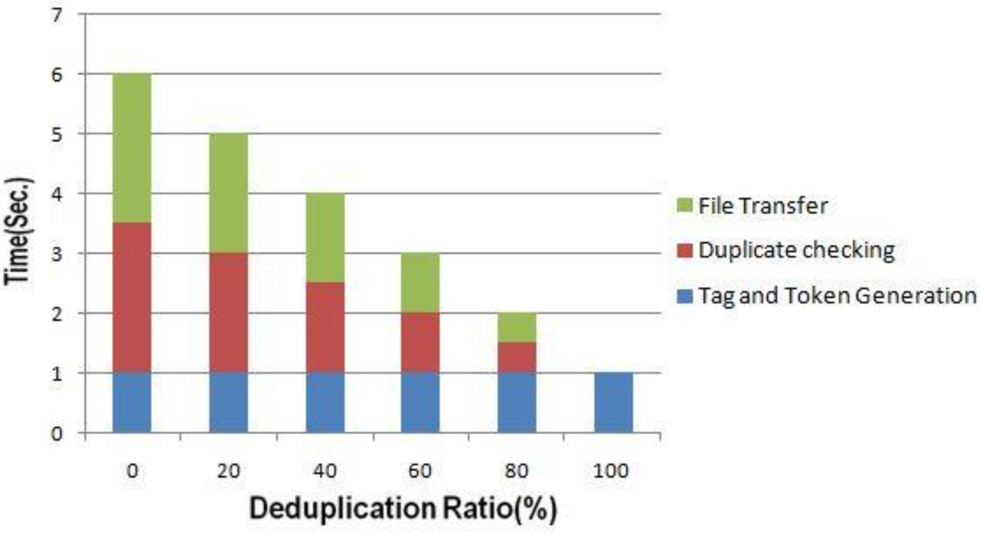

Fig 2: Time for deduplication ratio

VIII. FUTURE WORKS

For improving the security of the convergent keys, the keys can be stored and managed in a separate key server. Whenever the user needs to access the key from the server, the user can authenticate with the key server and retrieve the keys. The user can then use the keys for decrypting the files.

For improving the security of the ciphertext and avoiding the chance for brute force attack, ciphertext can be segmented and the segments can be distributed to multiple servers. In this procedure we use two algorithms to segment and rejoin the ciphertext before uploading and downloading the cypher to and from the public cloud.

IX. CONCLUSIONS

The authorized deduplication system eliminates the redundant copies of data in the cloud and protects the data by considering differential privileges of users in the duplicate check. The deduplication is achieved through authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with privilege keys and the file tag. The data security is achieved by applying convergent encryption scheme. The privilege keys of the users and the encryption keys for each file are stored and maintained in the private cloud so that the security can be guaranteed. The better security of the data is achieved by using the dual encryption procedure.

REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou, “A Hybrid Cloud Approach for Secure Authorized Deduplication”, in IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26, NO. 5, MAY 2015.

[2] 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.

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

[4] 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.

[5] M. Bellare, C. Namprempre, and G. Neven, “Security proofs for identity-based identification and signature schemes,” J. Cryptol., vol. 22, no. 1, pp. 1–61, 2009.

[6] 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.

[7] 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.

[8] D. Ferraiolo and R. Kuhn, “Role-based access controls, ” in Proc. 15th NIST-NCSC Nat. Comput. Security Conf., 1992, pp. 554–563.

[9] 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.

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

[11] 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.

[12] Ms. Madhuri. A. Kavade, Prof A. C. Lomte,”A Literature Survey on Secure De-Duplication Using Convergent Encryption Key Management”, International journal of Engineering and Computer Science ISSN:2319-7242 Volume 3 Issue 11 November,2014.