ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

Authorized De-duplication of Data over Cloud

[1] Shaikh Basirat Tazin, [2] S.D Pingle
[1] M.E. Student, [2] Ass.Professor
CSE Department, PES College of Engineering, Aurangabad, India

Abstract: **With a large number of users migrating towards cloud storage for storing their important data, data compression technique managing data stored over cloud had gained much importance. This technique removes duplicate copies of same data on a server and keeps only one copy. Again to protect data from attackers, data is encrypted using the linear congruential algorithm which employs convergent encryption concept. An authorized duplicate check is performed by users after encryption of data with help of generated tags. The file is uploaded if it is not already present otherwise POW is implemented. Hybrid cloud architecture is used with the public and private cloud, where public cloud stores all unique copies of data and private cloud manages authenticated access by maintaining privileges associated with file and users.**

Keywords: **- Data de-duplication, convergent encryption, authorized duplicate check, hybrid cloud.**

I. INTRODUCTION

Cloud computing provides you access to a networked storage and computing resources in a virtualized environment. Peoples are becoming more interested in big data for utilizing them in their own various application fields. Data tends to be in enormous amount when dealing with big data, where cloud computing can prove to be a useful solution for management of data. Main functions performed by cloud computing are efficient management of distributed data, offering services to users which they are requesting for, and solving complex problems. Cloud computing manages data residing over it by compressing amount of data. And for compressing amount of data, technique used is data de-duplication which deals with removing redundant copies of data. It stores only single copy of data, and for other copies of data reference is passed to that single copy. It is a simple storage optimization technique which is used by various cloud storage providers such as Dropbox, Bitcasa [7], JustCloud, Mozy [8], Amazon S3 (Simple Storage Service) and Google Drive, etc. With data deduplication technique bandwidth needed to transfer same content multiple times is reduced. When data is outsourced to cloud server, security and privacy issues arises since it is susceptible to vulnerable attacks. To overcome these issues before outsourcing data to cloud, it is encrypted. However, when different users will encrypt data by using their own keys, cipher text produced will be different thereby making deduplication impossible. Therefore, data is encrypted with help of convergent key which is derived from data itself and this technique of encryption is referred as convergent encryption. Data deduplication with convergent encryption provides secure and optimized storage.

II. PRIMITIVES

Data de-duplication

To efficiently manage data, well- known data compression technique used is data de-duplication which deals with removing redundant data. Using this technique, only single copy of data is stored on cloud server and other copies of similar data are passed reference to that unique copy. Data de-duplication can be performed at file level where redundant copies of similar files are removed or block level where redundant blocks of data that occur in dissimilar files are removed. De-duplication can be performed at client side. When de-duplication is performed at client side, bandwidth for transmitting data to server is reduced.

Convergent Encryption

Convergent encryption is a technique that produces similar cipher text from similar plain text. It performs this operation with help of convergent key which derived by calculating cryptographic hash value from data copy. After encryption, user stores key with him/her and sends cipher text to cloud. By using this technique, users will have same encryption keys and cipher text for similar data.

Proof of Ownership

De-duplication of data is performed by calculating corresponding hash values with help of which similarity between data copies is determined. If file is already present at cloud, the instead of uploading again that file pointer to ownership of file is updated and user can download that corresponding file. When an encryption is done at client side, an attacker may somehow retrieve hash values and access corresponding file. To defend from such an attack PoW is used to assure whether user owns a file or not.

Hybrid Cloud Architecture

This architecture consists of three components, namely, public cloud, private cloud and user. Architecture is illustrated in figure below.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

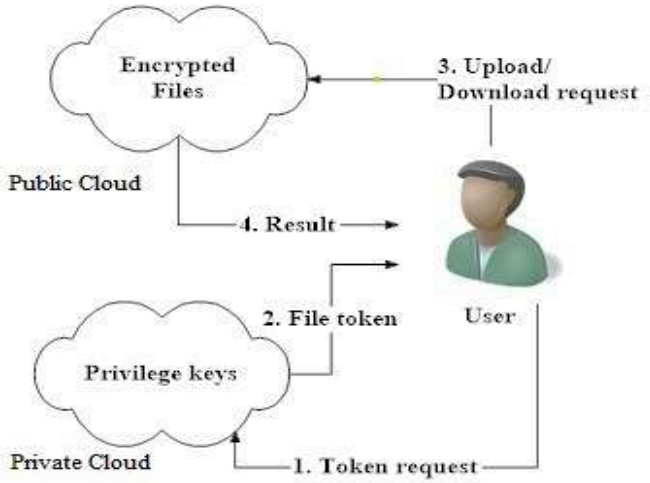

Fig: Hybrid cloud architecture.

- ***Public cloud:*** Usually, public cloud deals with storage of data. Users store data on public cloud which they want to access later. Public cloud uses S-CSP (Storage-cloud service provider) for storing data and stores only single copy of a particular data.
- ***Private cloud:*** Private cloud acts as an interface between public cloud and user. User who wants to upload data on public cloud first need to request for token from private cloud where private keys for users are stored. Only authenticated users are issued token for accessing public cloud.
- ***Users:*** User is an entity who wants to upload data to public cloud. User should upload only unique data and if data is already present then it should not uploaded to save bandwidth. In storage systems, performing deduplication of data, users are assigned privileges to predefine type of access they can perform.

III. RELATED WORK

Due to increase in use of cloud computing, data deduplication had gained much importance. Data deduplication is mostly performed with convergent encryption and together they are used in used in various systems such as Bitcasa, Backup system patent filed by Stac Electronics, Farsite, [8]etc.

Dropbox:

Dropbox [6] offers cloud storage for users to store storage, data amount will increase. To manage this data, de-duplication is needed. Dropbox performs de-duplication by considering fixed-size blocks. Each fixed-size block is of 4MB size. Hashing is done using SHA 256 technique. Calculated hash values of files and mapping of file and its corresponding block is managed by control server. And blocks representing unique blocks of data are stored in storage server in Amazon S3. Control server receives all calculated hash values sent by client and returns back only unique values to client. As different users are accessing this cloud Clients retrieves this values and sends corresponding blocks to storage server.

Bitcasa:

In his interview [7], Bitcasa CEO Tony Gauda had explained the use of convergent encryption in Bitcasa cloud storage provider for performing deduplication. In his interview he mentioned use of AES-256 hash, SHA 256 hashing for data to be stored. Encryption is performed on client side.

Encrypted de-duplication with fast and secure laptop backups:

In this paper [1], a community of laptop users is considered, which are using backup scheme to back up their data. This scheme emphasizes on increasing speed of performing backup and storage required for storing backup data. Algorithm this scheme uses is based on convergent encryption technique, in which data de-duplication is performed by using data common between different users. File to be backed up is first searched in list of backup storage. If file is not present, then file is backed up otherwise index is returned indicating location of file already backed up. If data to be backed up is confidential, then personnel can perform encryption on client-end also. Main disadvantage of this scheme is that direct backing up data to cloud can be very costly.

Dupless: Server aided encryption for de-duplicated storage:

In this paper [2], message locked encryption scheme is used which is based on convergent encryption technique. It consists of keys derived from messages which can be obtained from key server and is shared amongst different users. Users need to perform authentication to key server and do not supply any information about data to it. Data is stored on storage server where unique copies of data are maintained. Users interact with storage server very less number of times.

A secure data de-duplication scheme for cloud storage:

This scheme [6] considers that data are of two types popular and unpopular depending upon number of users accessing them. Popular data is considered as less sensitive and unpopular data is considered as sensitive data. Multi- layered cryptosystem is used in this scheme, with two cryptosystems, namely, convergent and threshold cryptosystem. The unpopular data is sensitive and is used by less number of users. Therefore, it needs security and is protected by two layers. On the other hand, popular data is less sensitive and used by many users. Therefore, it needs weaker security and is protected by only one layer. Security to be applied depends on layers that have been deployed depending upon how sensitive data is.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 2, February 2018

Secure data de-duplication:
In this paper [5], existing data is divided into small chunks from which keys for encrypting these chunks are generated. Two models are implemented in this paper, one is authenticated model which has similar design to convergent encryption technique that has been deployed in Farsite system and other one is anonymous model which mainly focuses on hiding details of authors and readers. In both models, client first divides a file to be uploaded into set of chunks using content based chunking. After chunks of data are formed, data is encrypted using convergent encryption technique at client side.

Enhanced De-key Approach to Reduce Data DeDuplication in Cloud Storage:
In this paper [3], main problem dealt is management of convergent keys. Dekey scheme is applied in this paper in which key management is not needed. In this scheme, convergent keys derived from messages are distributed over different server and client side deduplication is implemented with POW. Ramp Secret Sharing scheme is implemented in this paper.

IV. PROPOSED SYSTEM DESCRIPTION

Problem Statement:
To design an authorized deduplication scheme that will perform data de-duplication in cloud computing with twin clouds: public and private cloud and to use linear congruentiality algorithm for encryption of data and generate tags associated to it and to associate right of ownership of uploaded file along with access control rights in list.

Proposed Scheme:
An authorized deduplication scheme is proposed in this paper, where hybrid cloud architecture with public and private cloud is used. Public cloud deals with storing data. Private cloud deals with maintaining privileges for files and users and issuing private keys. Users first need to interact with private cloud by providing files and his privileges as input. Private cloud receives request, matches privileges and generates tag which is used for performing duplicate check. After performing duplicate check at public cloud, if that file is not already present at public cloud then token is generated for that corresponding user with help of which he can upload that file at public cloud. If file is already present on public cloud, then ownership list is updated and pointer is returned to user with help of which he can download corresponding file. The architectural diagram for authorized deduplication system is shown in figure.

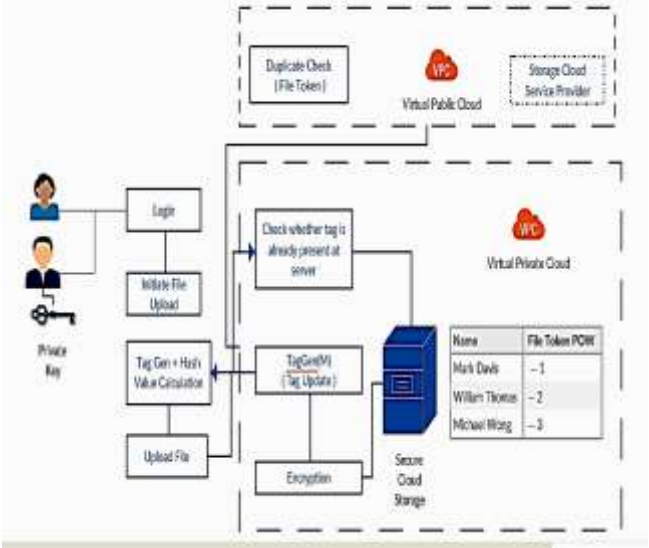

Fig: Architecture of authorized deduplication system.

The processes involved are described as follows:
• Access control list: Different users will have different privileges and these privileges will be stored using access control list. This access control list will have few attributes that will identify user by his credentials and it will also store access right of that user.
• Hash value calculation and tag generation: Different users will have different privileges and these privileges will be stored using access control list. This access control list will have few attributes that will identify user by his credentials and it will also store access right of that technique.
• Token Generation: Users will be issued token to upload file at public cloud after performing authorized duplicate check.
• Encryption and Decryption: To perform encryption and decryption linear congruentiality algorithm is used which uses layered approach for performing encryption and decryption. This algorithm uses symmetric encryption technique in which key is embedded within message itself and output is transmitted as bitmap file. Encryption and decryption process consists of three layers for performing its operation. These processes layer wise are described below:
O Encryption:
Layer-1: This layer is considered as mapping layer since it maps each of its character by another character which is also present in same set. It confuses attacker by jumbling characters. It considers two sets: one is repeated character set and other is non-repeated. Repeated characters are those whose probability of occurrence is maximum while non-repeated characters are those which occur very often. Each character is replaced by another character in similar set and thus performing first layer of encryption. No key is used in

this layer. Number character will also be replaced, to cause a mismatch.

Layer-2: This layer is considered as core-encoding layer since it performs encoding of characters at this stage by using bitwise logic and ASCII format. In this layer, each character obtained in first-layer is converted to an ASCII character. Starting from first character of message obtained in layer-1 each character is XORed with negated ASCII character of key starting from first character onwards. Each and every character is encoded using this process. Since key is of a small length, it is repeatedly applied to the message. This can be formulated as follows:

$$char\_new = (char\_old) \wedge (\sim key[i])$$

o Decryption:

Layer-1-This layer is referred character-restructuring layer since it restructures characters in and it forms groups of bits from bitmap fields to form ASCII characters. Each 8-bit data is considered and its ASCII value is found. Then character representing that ASCII value is identified.

Layer-2: This layer is referred as core-decoding layer since it decodes each character. While performing encryption we applied XOR logic, by using same logic twice we can retrace original character. Hence, by using same algorithm we can perform decryption and bitwise logic is also used here. So, with help of this algorithm we performed symmetric encryption since we used same key for encryption and decryption. Furthermore, in encryption and decryption process layered approach except one layer other layers are dependent on keys.

V. CONCLUSION

In this paper, an authorized data deduplication scheme is proposed in which deduplication is performed in hybrid cloud architecture with twin clouds, that is, private cloud and public cloud by using linear congruentiality algorithm. Data is secured by assigning differential privileges of users for performing duplicate check maintained with help of access control list. User performs authorized duplicate check where private cloud server issues token with help of which user can upload or download file from public cloud depending on whether file exists on server or not and accordingly owners of file are updated.

REFERENCES

[1] P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de-duplication”, In Proc. of USENIX LISA, 2010.

[2] M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Server-aided encryption for deduplicated storage”, In USENIX Security Symposium, 2013.

[3] Mr. Antony Xavier, Dr. V. Sai, Dr. S. P. Rajagopalan, “Enhanced De-key Approach to Reduce Data De-Duplication in Cloud Storage”, p. 5316-5320, 2016.

[4] M.W. Storer, K. Greenan, D.D.E. Long, and E.L. Miller, “Secure Data Deduplication”, p. 1-10, 2008.

[5] D. Kim, S. Song, B. Choi, “Data deduplication for data optimization for storage and network systems.” p. 42-44, 2017.

[6] 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.

[7] https://techcrunch.com /2011 /09/18 /bitcasa-explains-encryption

[8] https://en. wikipedia.org /wiki/ Convergent_ encryption