ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Secure Deduplication on Encrypted Big Data in Cloud Computing Environment

Sukanya Gunjal[1], Prof. B. S. Kurhe[2]

M.E. Student, Department of Computer Engineering, SPCOE, Otur, India[1]

Assistant Professor, Department of Computer Engineering, SPCOE, Otur, India[2]

ABSTRACT: Cloud computing has arrived as one of the fastest-growing segments of the Information technology industry which provides variant services such as software, platform and infrastructure for internet users. The greatest test for enormous information from a security perspective is the assurance of client's protection. Be that as it may, encoded information present new difficulties for cloud information deduplication, which gets to be significant for huge information stockpiling and preparing in cloud. Customary deduplication plans can't take a shot at encoded information. Existing arrangements of scrambled information deduplication experience the ill effects of security shortcoming. They can't adaptably bolster information get to control and renouncement. Hence, few of them can be promptly sent by and by. In this paper, we propose a plan to deduplicate scrambled information put away in cloud in light of proprietorship test and intermediary re-encryption. It incorporates cloud information deduplication with get to control. We address redundancy issues in Cloud Computing environments, we propose a plan to deduplicate encrypted data put away in cloud in view of possession test and intermediary re-encryption. It incorporates cloud data deduplication with access control.

KEYWORDS: Access control, Key generation, Cloud computing, Data-deduplication.

I. INTRODUCTION

A technique which has been adopted to manage large redundant data is Deduplication which plays a key role in Cloud Computing services. Our meant to minimize repetitive information and augment space funds. A strategy which has been generally embraced is cross-client deduplication. The basic thought behind deduplication is to store copy information (either documents or pieces) just once. Accordingly, if a client needs to transfer a record (piece) which is now put away, the cloud supplier will add the client to the proprietor rundown of that document (square). Deduplication has demonstrated to accomplish high space and cost reserve funds and numerous Huge Information stockpiling suppliers are as of now receiving it. Deduplication can diminish capacity needs by up to 90-95% for reinforcement applications and up to 68% in standard document frameworks. Distributed computing gives apparently boundless "virtualized" assets to clients as administrations over the entire Web, while concealing stage and usage subtle elements. Today's cloud benefit suppliers offer both exceedingly accessible capacity and enormously parallel figuring assets at moderately low expenses. As distributed computing gets to be predominant, an expanding measure of information is being put away in the cloud and imparted by clients to determined benefits, which characterize the get to privileges of the put away information. One basic test of distributed storage administrations is the administration of the regularly expanding volume of information. To make information administration versatile in distributed computing, de-duplication has been an outstanding strategy and has pulled in more consideration as of late. Information de-duplication is a specific information pressure system for wiping out copy duplicates of rehashing information away. The strategy is utilized to enhance stockpiling use and can likewise be connected to network information exchanges to diminish the quantity of bytes that must be sent. Rather than keeping numerous information duplicates with similar substance, de-duplication disposes of repetitive information by keeping stand out physical duplicate and alluding other excess information to that duplicate. De-duplication can occur at either the document level or the piece level. For record level de-duplication, it disposes of copy duplicates of similar document. De-duplication can likewise happen at the piece level, which takes out copy squares of information that happen in non-indistinguishable documents. Distributed computing is a rising administration display that gives calculation and capacity assets on the Web. One appealing

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

usefulness that distributed computing can offer is distributed storage. People and undertakings are regularly required to remotely file their information to stay away from any data misfortune in the event that there are any equipment/programming disappointments or unanticipated fiascos. Rather than buying the required stockpiling media to keep information reinforcements, people and ventures can basically outsource their information reinforcement administrations to the cloud benefit suppliers, which give the fundamental stockpiling assets to have the information reinforcements. While distributed storage is appealing, how to give security certifications to outsourced information turns into a rising concern. One noteworthy security test is to give the property of guaranteed cancellation, i.e., information records are for all time blocked heaps of erasure. Keeping information reinforcements for all time is undesirable, as delicate data might be uncovered later on in view of information break or wrong administration of cloud administrators. Subsequently, to dodge liabilities, endeavors and government organizations normally keep their reinforcements for a limited number of years and demand to erase (or crush) the reinforcements a short time later. For instance, the US Congress is figuring the Web Information Maintenance enactment in approaching ISPs to hold information for a long time, while in Joined Kingdom, organizations are required to hold wages and compensation records for a long time.

II. RELATED WORK

In this section, we present some important observations drawn from previous and our analysis of the vendor problem in cloud storage, showing how to make an efficient deduplication that allows very appealing reductions in the usage of storage resources.

Pasquale et al. present the inherent security exposures of convergent encryption and propose Cloud Deduplication, which preserves the combined advantages of deduplication and convergent encryption. Zheng et al. propose a scheme based on attribute-based encryption (ABE) to deduplication encrypted data stored in the cloud and support secure, effective and efficient data access control.

Fatema et al. extended there work for enterprise data deduplication framework is composed of three steps: In the first step, the data and its metadata is indexed in such a way as to ensure complete data privacy against a semi-honest cloud service provider. The second step consists in performing the multi-user private keyword searchable encryption on the encrypted data within a particular enterprise, keeping the searches and the resulting files secret from the cloud service provider. Step 3 makes use of a strategy to support data sharing between users, by utilizing the existing metadata, the indexing structures, and the searchable encryption scheme.

Xiaolong et al. triednew approach combining client-side deduplication and target-side deduplication for cloud storage system. The contribution of this paper is twofold. Firstly, file-level and chunk-level duplication are detected and eliminated locally by Client andglobally by Metadata Server (MS) to improve the deduplication ratio. Secondly, we put forward the Delay Dedupe strategy, a delayed target-deduplication scheme based the chunk level deduplication and the access frequency of chunks in the Snodes.

N. Lakshmi et al. proposed system uses ALG datadedupulication technique, which avoids data redundancy saving space ultimately increasing the efficiency of the data storage. This system also uses the standard AES algorithm for data encryption thus adding data security. Finally this also uses the RSS key method, which is generated dynamically in a unique way, which is far secured than the privilege keys.
N. Jayapandian et al. takes a solution called Convergent Encryption (CE). CE is a deterministic symmetric encryption scheme in which the key K is derived from the message M itself by computing K = H(M) and then encrypting the message as C = E(K;M) = E(H(M);M) where H is a cryptographic hash function and E is a file cipher.Using CE, any user holding the same message will produce the same key and cipher text, enabling deduplication. And suggested MLE for small files.

Jan etal.found a secure deduplication scheme for encrypted data that has dynamic ownership management capability. There proposed scheme is constructed based partially on a randomized convergent encryption scheme in order to

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

randomize the encrypted data, which renders the proposed scheme secure against the chosen-plaintext attack while still allowing deduplication over the data. There proposed scheme is further integrated into the re-encryption protocol for owner revocation. The owner revocation is executed by re-encrypting the outsourced cipher text and selectively distributing the re-encryption key to valid owners by the cloud server.

III. MOTIVATION

Deduplication may be a storage saving technique that has been adopted by several cloud storage suppliers like Dropbox. In cloud storage services, deduplication technology are commonly accustomed cut back the world and (knowledge and data) live necessities of services by eliminating redundant knowledge and storing entirely one copy of them. Deduplication is best once multiple users supply an identical data to the cloud storage, but it raises issue concerning security and possession. Issues over information security still forestall several users from migrating information to remote storage. The standard resolution is to write in code the info before it leaves the owner's premises. Client-side information deduplication specifically ensures that multiple transfers of constant contentsolely consume network information measure and space for storing of one upload.

Traditional deduplication schemes cannot work on encrypted data, same or different users may upload duplicated data in encrypted form to CSP, especially for scenarios where data are shared among many users. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few of them can be readily deployed in practice. So the development of numerous services further makes it urgent to deploy efficient resource management mechanism. System present a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control.

IV. EXISTING SYSTEM

From the above literature survey we have concluded that an existing data de-duplication system, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges. Such architecture is practical and has attracted much attention from researchers. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

V. SYSTEM ARCHITECTURE

In the proposed research work to design and implement a system which will provide the parallel processing to detect the data de-duplication problem in cloud storage. We enhance our system in security. Specifically, we present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

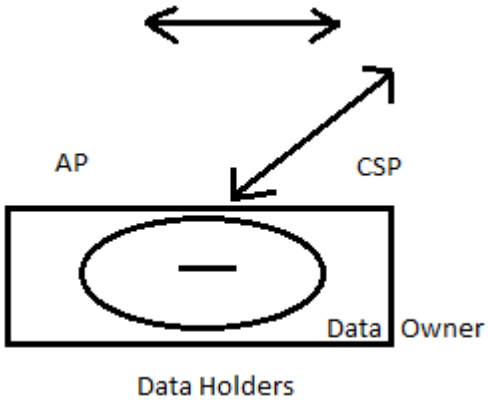

Fig. 1. System Model

As shown in Fig. 1, the system contains three types of entities: 1) CSP that offers storage services and cannot be fully trusted since it is curious about the contents of stored data, but should perform honestly on data storage in order to gain commercial profits; 2) data holder (ui) that uploads and saves its data at CSP. In the system, it is possible to have a number of eligible data holders ðui ; i ¼ 1; . . . ; nÞ that could save the same encrypted raw data in CSP. The data holder that produces or creates the file is regarded as data owner. It has higher priority than other normal data holders, which will be presented in Section 4; 3) an authorized party (AP) that does not collude with CSP and is fully trusted by the data holders to verify data ownership and handle data deduplication. In this case, AP cannot know the data stored in CSP and CSP should not know the plain user data in its storage. Proposed scheme contain following main aspects.

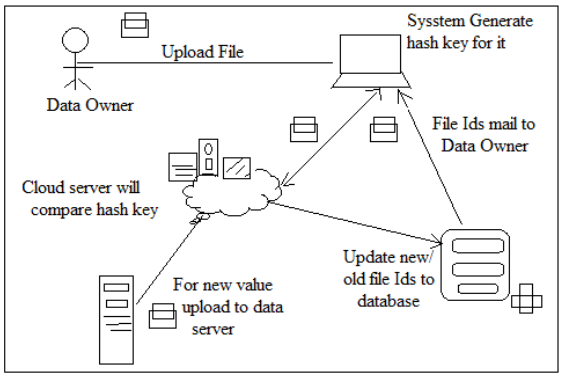

Fig. 2. Proposed System

Encrypted Data Upload:
If data duplication check is negative, the data holder encrypts its data utilizing an arbitrarily culled symmetric key DEK in order to ascertain the security and privacy of data, and stores the encrypted data at CSP together with the token utilized for data duplication check. The data holder encrypts DEK with pkAP and passes the encrypted key to CSP.

Data Deduplication:
Data duplication occurs at the time when data holder u endeavors to store the same data that has been stored already at CSP. This is checked by CSP through token comparison. If the comparison is positive, CSP contacts AP for deduplication by providing the token and the data holder's PRE public key. The AP challenges data ownership, checks the eligibility of the data holder, and then issues a re-encryption key that can convert the encrypted DEK to a form that can only be decrypted by the eligible data holder.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Data Deletion:
When the data holder effaces data from CSP, CSP firstly manages the records of duplicated data holders by abstracting the duplication record of this utilizer. If the rest records are not vacuous, the CSP will not efface the stored encrypted data, but block data access from the holder that requests data effacement. If the rest records are vacuous, the encrypted data should be abstracted at CSP.

Data Owner Management:
In case that an authentic data owner uploads the data later than the data holder, the CSP can manage to preserve the data encrypted by the authentic data owner at the cloud with the owner engendered DEK and later on, AP fortifies re-encryption ofDEK at CSP for eligible data holders.

Encrypted Data Update:
In case that DEK is updated by a data owner with DEK0 and the incipient encrypted raw data is provided to CSP to supersede old storage for the reason of achieving better security, CSP issues the incipient re-encrypted DEK0 to all data holders with the fortification of AP.

VI. PROPOSED ALGORITHM

Algorithm: AES: Key Generation, Encryption, Decryption Algorithm:
AES encrypts messages through the following algorithm, which is divided into 3 steps:

1. Key Generation:

I. Choose two distinct prime numbers p and q.
II. Find n such that n = pq. n will be used as the modulus for both the public and private keys.
III. Find the totient of n, (n)
(n)=(p-1)(q-1).
IV. Choose an e such that 1 ¡ e ¡ (n), and such that e and (n) share no divisors other than 1 (e and
(n) are relatively prime). e is kept as the public key exponent.
V. Determine d (using modular arithmetic) which satisfies the congruence relation de = 1 (mod (n)).

In other words, pick d such that de - 1 can be evenly divided by (p-1)(q-1), the totient, or (n). This is often computed using the Extended Euclidean Algorithm, since e and (n) are relatively prime and d is to be the modular multiplicative inverse of e. d is kept as the private key exponent.The public key has modulus n and the public (or encryption) exponent e. The private key has modulus n and the private (or decryption) exponent d, which is kept secret.

2. Encryption:

I. Person A transmits his/her public key (modulus n and exponent e) to Person B, keeping his/her private key secret.
II. When Person B wishes to send the message ”M” to Person A, he first converts M to an integer such that 0 ¡ m ¡ n by using agreed upon reversible protocol known as a padding scheme. III. Person B computes, with Person A’s public key information, the ciphertext c correspondingto,
c = me (mod n).
IV. Person B now sends message”M” in ciphertext, or c, to Person A.

3. Decryption:

I. Person A recovers m from c by using his/her private key exponent, d, by the computation m = cd (mod n).
II. Given m, Person A can recover the original message ”M” by reversing the padding scheme. This procedure works since,
c = me (mod n), cd =(me)d (mod n), cd = mde (mod n).
By the symmetry property of mods we have that
mde = mde (mod n).

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Since de = 1 + k(n), we can write mde = m1 + k(n) (mod n), mde = m(mk)(n) (mod n), mde = m (mod n).

VII. PERFORMANCE ANALYSIS

Convergent encryption suffers from some weaknesses which have been widely discussed in the literature, As the encryption key depends on the value of the plaintext, an attacker who has gained access to the storage can perpetrate the so called "dictionary attacks" by comparing the ciphertexts resulting from the encryption of well-known plaintext values from a dictionary with the stored cipher texts. Indeed, even if encryption keys are encrypted with users' private keys and stored somewhere else, the potentially malicious cloud provider, who has no access to the encryption key but has access to the encrypted chunks (blocks), can easily perform offline dictionary attacks and discover predictable files. This issue arises in where chunks are stored at the storage provider after being encrypted with convergent encryption. Since this has direct access to server data are less secure. In attribute based encryption the complexity of implementation is high and it is not scalable and flexible for large volume of data but this supports large number of duplicate copies so it is much suitable for small data that has much copies. Since it is implemented in symmetric key policy the key distribution and computation is the bigger problem and could damage when compressed. The data that needs to be stored is first preprocessed if necessary. In certain cases the preprocessing takes up a lot of time based on the type of processing that is implemented. Data cleaning, Normalization, Data hiding, Structuring of data, etc are some of the preprocessing steps available. The next step is the chunking. The process of splitting the given data into any blocks or chunks of data is called as chunking. This is the crucial step in the deduplication process. This is because, based on the size of each chunk the number of duplicate data changes. Based on the initial data, the chunk size should be fixed in such a way that the obtained chunks have large number of duplicates and thus the storage size will be reduced as much as possible.

VIII. CONCLUSION

In this paper, we proposed a practical scheme to manage the encrypted big data in cloud with deduplication based on ownership challenge and PRE. Our scheme can flexibly support data update and sharing with deduplication even when the data holders are offline. Encrypted data can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption. Extensive performance analysis and test showed that our scheme is secure and efficient under the described security model and very suitable for big data deduplication. The results of our computer simulations further showed the practicability of our scheme. Future work includes optimizing our design and implementation for practical deployment and studying verifiable computation to ensure that CSP behaves as expected in deduplication management.

REFERENCES

[1]. Jin Li, Yan Kit Li, XiaofengChen,Patrick P. C. Lee and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", *IEEE Transaction On Parallel And Distributed System*,Vol.PP,No.99, 2014.
[2]. Maneesha Sharma, Himani Bansal and Amit Kumar Sharma, " Cloud Computing: Different Approach & Security Challenge", *IJSCE, Volume-2, Issue-1,March* 2012.
[3]. Kangchan Lee, "Security Threats in Cloud Computing Environments", *International Journal of Security and Its Applications*, Vol. 6, No. 4, October, 2012.
[4]. Sashank Dara, "Cryptography Challenges for Computational Privacy in Public Clouds", *International Journal of Security and Its Applications, Volume 4*, 2002.
[5]. David Pointcheval, "Asymmetric Cryptography and Practical Security", *International Journal of Security and Its Applications, Volume* 4,2002.
[6]. Yogesh Kumar, Rajiv Munjal and Harsh Sharma, "Comparison of Symmetric and Asymmetric Cryptography with Existing Vulnerabilities and Counter-measures", *International Journal of Computer Science and Management Studies*, Vol. 11, Issue 03, Oct 2011.
[7]. Jan Stanek, Alessandro Sorniottiy, Elli Androulakiy, and Lukas Kencl, "A Secure Data De-duplication Scheme for Cloud Storage", *IBM Research, Zurich, May* 1994.
[8]. Fatema Rashid, Ali Miri, Isaac Woungang., "Secure Enterprise Data Deduplication in the Cloud" *IEEE Sixth International Conference on Cloud Computing*, 367-374, 2013.
[9]. Pasquale Puzio, RefikMolva, Melek O nen, Sergio Loureiro., "ClouDedup: Secure Deduplication with Encrypted Datafor Cloud Storage" *IEEE CloudCom*, 2013.
[12] MihirBellareSriramKeelveedhi, Thomas Ristenpart., "Message-Locked Encryption and Secure Deduplication" *A preliminary version of this paper appears in the proceedings of Eurocrypt*, 1-29, 2013.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

[13] R.Manjusha, R.Ramachandran., "Comparative Study of Attribute Based Encryption Techniques in Cloud Computing," *International Conference on Embedded Systems (ICES)*, 116-120, 2014.

[14]. 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.

[15]. C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: A survey" *in J. Big Data, vol.* 2, no. 1, pp. 1-32, 2015, doi:10.1186/s40537-015-0030-3.