A Study on Authorized Deduplication Techniques in Cloud Computing

Bhushan Choudhary, Amit Dravid

***Abstract*— Data deduplication is the technique which compresses the data by removing the duplicate copies of identical data and it is extensively used in cloud storage to save bandwidth and minimize the storage space. To secure the confidentiality of sensitive data during deduplication, the convergent encryption technique is used to encrypt the data before outsourcing. For better data protection, this paper talks about the issue of data deduplication authorization. There are several new deduplication implementations providing authorized deduplication verification in a hybrid cloud approach.**

***Index Terms*— Authorized Deduplication, Secured, duplicate check, confidentiality, Hybrid Cloud computing.**

I. INTRODUCTION

Cloud computing is the new emerging trends in the new generation technology. Every user has huge amount of data to share to store in a quickly available secured place. The concept of deduplication is arrived here to efficiently utilize the bandwidth and disk usage on cloud computing. To avoid the duplication copies of the same data on cloud may cause lose of time, bandwidth utilization and space.

Cloud computing is internet-based, a network of remote servers connected over the Internet to store, share, manipulate, retrieve and processing of data, instead of a local server or personal computer. The benefit of cloud computing are enormous. It enables us to work from anywhere. The most important thing is that customer doesn't need to buy the resource for data storage. When it comes to Security, there is a possibility where a malicious user can penetrate the cloud by impersonating a legalize user, there by affecting the entire cloud thus infecting many customers who are sharing the infected cloud. There is also big problem, where the duplicate copies may upload to the cloud, which will lead to waste of band width and disk usage. To improve this problem there should be a good degree of encryption provided, that only the customer should be able to access the data and not the legitimate User. Yan Kit Li et al.[6] shown To formally solve the problem of authorized data deduplication. Data deduplication is a data compression techniques for removing duplicate copies of identical data, and it is used in cloud storage to save bandwidth and to reduce the amount storage space. The technique is utilized to enhance the storage use and can likewise be applied to network data exchange to reduce the amount of bytes that must be sent. Keeping multiple data copies with the identical content, de-duplication removes redundant data by keeping only one copy and referring other identical data to that copy. De-duplication occurs either at block level or at file level. In file level de-duplication, it removed duplicate copies of the identical file. Deduplication can also take place in the block level that eliminates duplicate blocks of data that is occurred in non identical files. Data deduplication having huge amount of advantages like providing security as well as privacy concerns arise as users sensitive or delicate data are at risk to both insider and outsider attacks. The traditional encryption requires many different customers for encrypting the data files with their own private keys. Thus, the same data copies of different customers will lead to different cipher texts, making de-duplication impossible. To secure the privacy of sensitive information while supporting deduplication, the convergent encryption strategy has been proposed to encode the information before outsourcing.

This paper will work to dissolve the security issue and to evaluate the efficient utilization of cloud band width and disk usage.

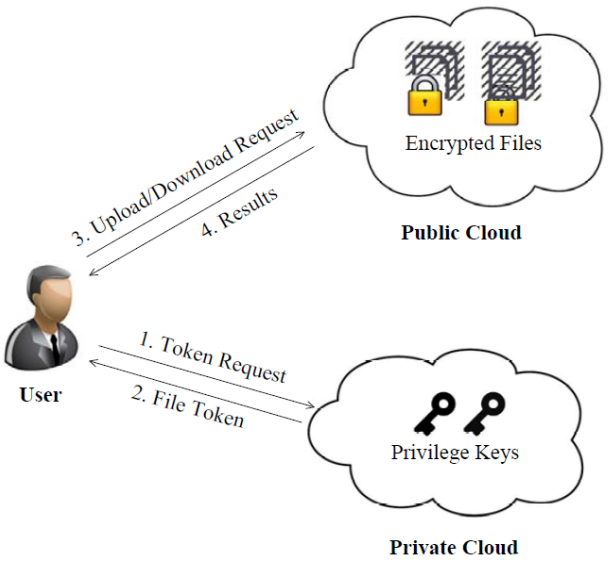

Fig. 1 Architecture of Authorized deduplication
(Source: Li et al. [6])

II. PRELIMINARIES

In this portion, the definition of notation used in this paper, survey some protected primitives utilized as a part of our safe deduplication. The notations used in this paper are given in TABLE 1.

***Bhushan Choudhary**, Dept. of M.E. (ComputerEngineering) Savitribai Phule Pune University, Pune, Maharashtra(INDIA).*

***Amit Dravid**, Dept. of M.E. (ComputerEngineering) Savitribai Phule Pune University, Pune, Maharashtra(INDIA).*

ISSN: 2278 – 1323

Acronym	Description
S-CSP	Storage-cloud service provider
PoW	Proof of Ownership
(pk_U, sk_U)	User's public and secret key pair
k_F	Convergent encryption key for file F
P_U	Privilege set of a user U
P_F	Specified privilege set of a file F
$\phi'_{F,p}$	Token of file F with privilege p

Table 1

Symmetric encryption: Symmetric encryption utilizes a regular secret key κ to encode the decoded data. A symmetric encryption plan comprises of three basic function:

• Keygense(1^{λ}) $\rightarrow \kappa$ -key generation algorithm generates κ utilizing security parameter 1^{λ}.
• Encse(κ ,*M*) $\rightarrow$*C* -symmetric encryption alogorithm that receives secret key κ and message*M* and gives ciphertext *C*.
• Decse(κ ,*C*) $\rightarrow$*M* -symmetric decryption algorithm that receives the secret key κ and ciphertext *C* and gives the original message *M*.

Convergent Encryption gives information secrecy in deduplication. Customers get a convergent key from each and every unique data copy and encrypt the unique data copy with the convergent key. And also, the customer determines a tag for the unique data copy, which will utilize the tag to recognize duplicate copies. The consideration of the tag accuracy holds[4] that means if both the data copies are the same, then the tags of the data copies are same. To discover the duplicate copies, the customer first sends the tag to the server to verify if the duplicate copy has been already available. The convergent key and tags are individually evaluated, and tags cannot understand the convergent key to distract the data security. The encrypted data copy and the respective tag will store on the server. The convergent encryption system can be defined by four basic functions:
• $KeyGen_{ce}$ (*M*) $\rightarrow$*K* -key generation algorithm which maps an information data copy *M* to convergent key *K*.
• Enc_{ce}(*K,M*) $\rightarrow$*C* -symmetric encryption algorithm that receives the input of both data copy *M* and convergent key *K,* then gives output cipher text *C.*
• Dec_{ce}(*K,C*) $\rightarrow$M –decrypting algorithm which receives the input of the convergent key *K* and cipher text *C,* then gives the output of the original data copy *M*.
• TagGen(*M*) $\rightarrow$ *T(M)* –tags generating algorithm which maps original data copy *M* and gives output tag *T(M).*

Proof of Ownership: The idea of proof of ownership (Pow) [8] allows customers to verify the ownership of the information data copies to storage server. Particularly, PoW is developed as an communicative algorithm (indicated by PoW) run by a verifier (i.e. customer) and a prover (i.e. storage server). The storage server derives a short term $\phi(M)$ from an information data copy *M*. To demonstrate the ownership of information data copy *M*, the customer needs to send ϕ' to the storage sever such that $\phi' = \phi(M)$. The security definition for PoW follows threat system in content distributed network, where the attacker doesn't knows the whole document, yet has accessories who have the record. The accessories follows "bound retrieval system", that it can help the attacker to get the document, subject to restrict or give limitation that they must send some few bits than the starting min-entropy of the document to the attacker [8].

Idetification Protocol: This protocol can be depicted with two stages: Proof and Verify. In the phase of Proof, a prover/client *U* (User) can explain his identity to a verifier by demonstration or presenting some recognizable proof of indentity. The information of the prover/client is his private key sk_u that is delicate data for example private key of a public key in its debit card number or certificate etc. that the client doesn't wants to share others. The verifier performs the confirmation process with input of public data pk_u correlated with sk_u. At the final inference of the protocol, the verifier give output of accepts or rejects to specify that the proof is correct or not. There are numerous effective identification proof protocol, with identity based and certificate based identification.

III. Related Work

The new start of cloud computing, secure information deduplication has pulled in much consideration and attention from research group. A deduplication system in the cloud storage Yuan et al [10] proposed to reduce the storage size of the tags for integrity check. To upgrade the security of deduplication and secure the information secrecy, Bellare et al [3] demonstrated to secure the information by transforming the predictable message into unpredictable message.

Mihir Bellare et al [3] given the security verifications or assaults for an expansive number of identity-based recognizable proof and signature schemes characterized either explicitly or implicitly in present information. Fundamental this are a system that on the one hand benefits clarify how these schemes are determined, and then again empowers integrated security investigations, consequently serving to understand, streamline and bind together past work. [3] Given in the paper that how to secure the data confidentiality by translating the predictable message into unpredictable. The use of third party (key server) is implemented to produce the file tag for the duplicate copy check. Stanek et al. [11] The innovative encryption scheme which provides many different security of known and unknown data. For known information that are not especially delicate or sensitive, the traditional or classic ordinary encryption is performed. An alternate two-layered encryption plan with higher security while giving support to deduplication is proposed for unknown information. Along these lines, they accomplished better tradeoff between the proficiency and security of the outsourced information. Li et al. [12] tended to the key management problem in block level deduplication by circulating these keys crosswise over numerous servers after scrambling the records.

Convergent Encryption: [8] This guarantees information protection in deduplication. Bellare et al. [4] formalized a primary message-locked encryption, and analyzed its application in efficient space secure outsourced capacity storage. Xu et al. [13] additionally tended to the problem and demonstrated a protected convergent encryption for effective encryption, without considering problems of the block level deduplication and key-management. There are likewise

different implementations of convergent encryption for secure deduplication. It is realized that some business cloude storage suppliers, for example, Bitcasa, likewise send converent encryption.

Proof of Ownership: The thought of "Proof of ownership"(PoW) Halevi et al. [8] for deduplication frameworks, such that a customer can effectively prove to the cloud storage server that he owns a record without transferring the record itself. A few PoW developments established on the[8] Merkle-Hash Tree is proposed to allow customer side deduplication, which include the delimited leakage setting. Pietro and Sorniotti [9] proposed an alternate PoW plan by selecting the projection of a record onto some randomly chosen bit-positions as the record verification. Note that all the above plans don't consider information security. Newly, Ng et al. [14] enhanced PoW for encryption documents, yet they don't show how to reduce the key management overhead.

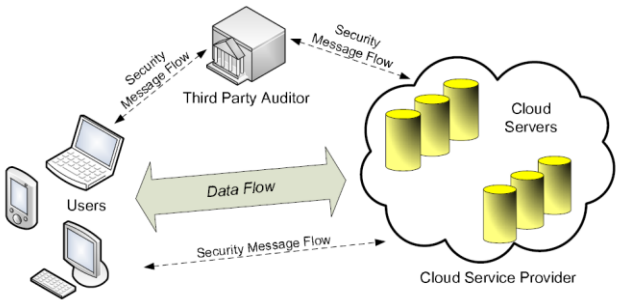

Fig 2. Architecture of Cloud Data storage.
(Source: Wang et al. [16]

Twin Clouds Architecture: Bugiel et al. [7] given a framework comprising of twin cloud for protected outsourcing of information and subjective processing to an untrusted service cloud. Zhang et al. [15] also introduced the hybrid cloud methods to support security conscious data intensive computing. The work considers pointing the authorized deduplication issue over information in public cloud. The security model of the frameworks is same as related work, in which the private cloud is expect to be completely trustworthy and remarkable.

IV. Discussion

The idea of Authorized Data deduplication was proposed to secure the information security by counting differential benefits of clients in the copy check. Yan Kit Li et al [6] additionally exhibited a few new deduplication developments supporting approved copy check in hybrid cloud construction modeling, in which the copy check tokens of documents are created by the private cloud server having private keys. Security examination shows that our plans are secure as far as insider and outsider attacks determined in the proposed security model. As an issue verification of idea, they actualized a model of the proposed approved copy check plan and behavior test bed investigates their model. They indicated that their authorized copy check plan brings about insignificant overhead comparing convergent encryption and system exchange.

The issue of giving secure outsourced capacity that both supports deduplication and defend brute-force attacks. The framework [3], Dupless, that consolidates a CE-type baseMLE plan with the capacity to get message-derived keys with the assistance of a key server (KS) imparted among a gathering of clients. The customers connect with the KS by a protocol for absent pseudorandom functions(PRF), guaranteeing that the KS can cryptographically blend in mystery material to the every message keys while adapting nothing about documents put away by clients. These instruments guarantee that Dupless gives solid security against outside attacks which compromise the SS (Storage Service) and interacting channels, also that the security of Dupless rapidly corrupts despite contained frameworks. Allowing a client be compromise, taking in the plaintext fundamental an alternate client's cipher text requires mounting an online bruteforce attack (which can be abated by a rate-restricted KS). Allowing the KS be compromised, the aggressor must in any case endeavor offline brute-force attack, matching the sureties of MLE plans. The generous increment in security takes a stab at a humble cost as far as execution, and a little increment in capacity prerequisites with respect to the base framework. The low execution overhead brings about part from enhancing the client to-KS oblivious pseudorandom function convention, furthermore from guaranteeing Dupless utilizes a low number of associations with the SS. Demonstrated that Dupless is not difficult to convey: it can work straightforwardly on top of any SS executing a basic capacity interface, as demonstrated by the model for Dropbox furthermore Google Drive.

V. Advantages of Authorised Deduplication System

1) The client is permitted to perform the duplicate copy check for records selected with the particular subject.
2) The complex subject to help stronger security by encoding the record with distinct privilege keys.
3) Decrease the storage space of the tags for reliability check. To strengthen the security of deduplication and ensure the data privacy.

VI. Conclusion

The thought of authorized information deduplication was proposed to ensure the information security by counting differential benefits of clients in the duplicate copy check. The presentation of a few new deduplication developments supporting authorized duplicate copy in hybrid cloud architecture, in that the duplicate check tokens of documents are produced by the private cloud server having private keys. Security check exhibits that the methods are secure regarding insider and outsider assaults detailed in the proposed security model. As an issue verification of idea, the developed model of the proposed authorized duplicate copy check method and tested the model. That showed the authorized duplicate copy check method experience minimum overhead comparing convergent encryption and data transfer.

ISSN: 2278 – 1323

REFERENCES

[1] Bugiel, Sven, et al. "Twin clouds: Secure cloud computing with low latency." Communications and Multimedia Security. Springer Berlin Heidelberg, 2011.
[2] Anderson, Paul, and Le Zhang. "Fast and Secure Laptop Backups with Encrypted De-duplication." LISA. 2010.
[3] Bellare, Mihir, Sriram Keelveedhi, and Thomas Ristenpart. "DupLESS: server-aided encryption for deduplicated storage." Proceedings of the 22nd USENIX conference on Security. USENIX Association, 2013.
[4] Bellare, Mihir, Sriram Keelveedhi, and Thomas Ristenpart. "Message-locked encryption and secure deduplication." Advances in Cryptology–EUROCRYPT 2013. Springer Berlin Heidelberg, 2013. 296-312.
[5] Bellare, Mihir, Chanathip Namprempre, and Gregory Neven. "Security proofs for identity-based identification and signature schemes." Journal of Cryptology 22.1 (2009): 1-61.
[6] Li, Jin, et al. "A Hybrid Cloud Approach for Secure Authorized Deduplication."
[7] Bugiel, Sven, et al. "Twin clouds: An architecture for secure cloud computing." Proceedings of the Workshop on Cryptography and Security in Clouds Zurich. 2011.
[8] Halevi, Shai, et al. "Proofs of ownership in remote storage systems." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.
[9] Di Pietro, Roberto, and Alessandro Sorniotti. "Boosting efficiency and security in proof of ownership for deduplication." Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM, 2012.
[10] Yuan, Jiawei, and Shucheng Yu. "Secure and constant cost public cloud storage auditing with deduplication." Communications and Network Security (CNS), 2013 IEEE Conference on. IEEE, 2013.
[11] Stanek, Jan, et al. A secure data deduplication scheme for cloud storage. Technical Report, 2013.
[12] Li, Jin, et al. "Secure deduplication with efficient and reliable convergent key management." (2013): 1-1.
[13] Douceur, John R., et al. "Reclaiming space from duplicate files in a serverless distributed file system." Distributed Computing Systems, 2002. Proceedings. 22nd International Conference on. IEEE, 2002.
[14] Ng, Wee Keong, Yonggang Wen, and Huafei Zhu. "Private data deduplication protocols in cloud storage." Proceedings of the 27th Annual ACM Symposium on Applied Computing. ACM, 2012.
[15] Zhang, Kehuan, et al. "Sedic: privacy-aware data intensive computing on hybrid clouds." Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011.
[16] Wang, Cong, et al. "Privacy-preserving public auditing for data storage security in cloud computing." INFOCOM, 2010 Proceedings IEEE. Ieee, 2010.

Mr. Bhushan P. Choudhary received the B.E. degree in Information Technology in 2012 from University of Pune. He is Master of Computer Engg. Student at GHRIET, Pune. His interest is in Cloud Computing and wants to know new innovative things in I.T. field.

Mr. Amit Dravid did his M.S. in 2004 from Virginia Tech, USA, and his B.E. in 1999. He is a Ph.D scholar at DA-IICT, Gandhinagar, and his current interests are in Image Processing and Image Retrieval. He likes to learn new things and explore new areas.

