IPHV8I1063X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .I
ISSN Online: 2319 – 9253
Print: 2319 – 9245

SECURE DATA DEDUPLICATION WITH EFFICIENT KEY MANAGEMENT IN CLOUD DATABASE

[#1]Dr.K. RAMESHWARAIAH, Professor and H.O.D, Dept of CSE,
[#2]SATURI RAJESH, Associate Professor, Dept of CSE,
[#3]GOVARDHAN REDDY KANDAKAT, M.Tech Student, Dept of CSE,
NALLA NARSIMHA REDDY GROUP OF EDUCATIONAL SOCIETY, HYD, T.S., INDIA.

ABSTRACT— Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data and has been widely used in cloud storage in order to minimize the amount of storage space and save bandwidth. For protection of data security, this paper makes an attempt to primarily address the problem of authorized data deduplication. To protect the confidentiality of important data while supporting deduplication, the convergent encryption technique has been proposed to encrypt the data before outsourcing. Along with the data the privilege level of the user is also checked in order to assure whether he is an authorized user or not. Security analysis demonstrates that our scheme is secure in terms of the definitions specified in the proposed security model. We show that our proposed authorized duplicate check scheme has minimal overhead compared to normal operations. As a proof of concept, we implement a prototype of our proposed authorized duplicate check scheme and conduct tested experiments using our prototype. This paper tries to minimize the data duplication that occurs in hybrid cloud storage by using various techniques.

KEYWORDS— *Deduplication, authorized duplicate check, confidentiality, hybrid cloud.*

I.INTRODUCTION

Cloud computing technique which is most widely used today. In that, computing is done over the large communication network like Internet. It is an important solution for business storage in low cost. Cloud computing provide vast storage in all sector like government, enterprise, also for storing our personal data on cloud. Without background implementation details, platform user can access and share different resources on cloud. The most important problem in cloud computing is that large amount of storage space and security issues. One critical challenge of cloud storage to management of ever-increasing volume of data. To improve scalability, storage problem data reduplication is most important technique and has attracted more attention recently. It is an important technique for data compression, It simply avoid the duplicate copies of data and store single copy of data. Data deduplication take place in either block level or file level. In file level approach duplicate files are eliminate, and in block level approach duplicate blocks of data that occur in non-identical files. Deduplication reduce the storage needs by up to 90-95% for backup application,68% in standard file system. Important issues in data deduplication that security and privacy to protect the data from insider or outsider attack. For data confidentiality, encryption is used by different user for encrypt there files or data, using a secrete key user perform encryption and decryption operation. For uploading file to cloud user first generate convergent key, encryption of file then load file to the cloud. To prevent unauthorized access proof of ownership protocol is used to provide proof that the user indeed owns the same file when deduplication found. After the proof, server provide a pointer to subsequent user for accessing same file without needing to upload same file. When user want to download file he simply download encrypted file from cloud and decrypt this file using convergent key.

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.

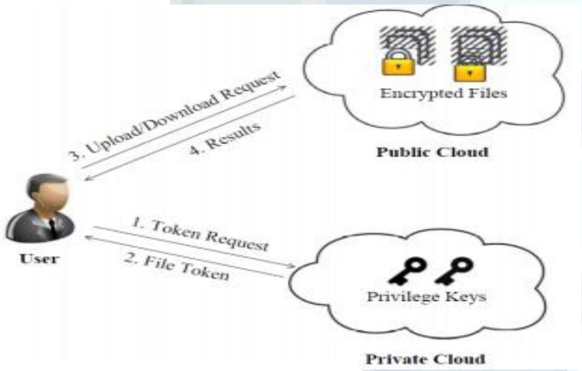
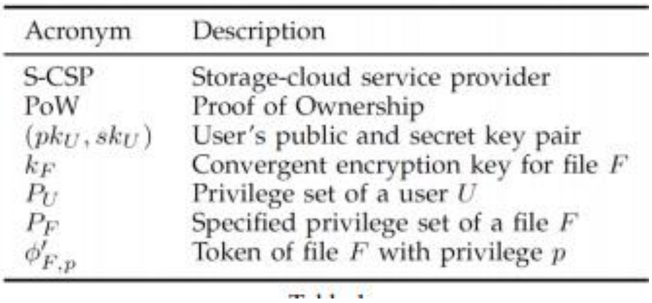


Fig. 1 Architecture of Authorized deduplication

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.



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 functions:

- Keyence (1**λ**) →κ -key generation algorithm generates utilizing security parameter 1**λ**.
- Encse (κ,M) →C -symmetric encryption alogorithm that receives secret keyκ and messageM and gives ciphertext C.
- Decse(κ,C) →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:

- KeyGence (M) →K -key generation algorithm which maps an information data copy M to convergent key K.
- Encce(K,M) →C -symmetric encryption algorithm that receives the input of both data copy M and convergent key K, then gives output cipher text C.
- Decce(K,C) →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) → 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 ϕ(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 ϕ′ = ϕ(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].

Identification 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 sku 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 pku correlated with sku. 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. Bellaire 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 cloud storage suppliers, for example, Bitcasa, likewise send convergent 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.

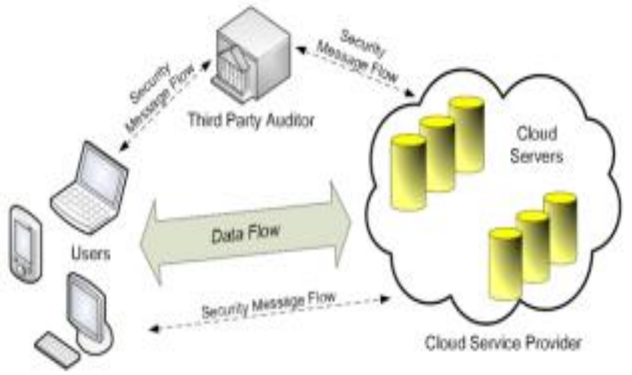


Fig 2. Architecture of Cloud Data storage

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. PROPOSED SYSTEM

In our system we implement a project that includes the public cloud and the private cloud and also the hybrid cloud which is a combination of the both public cloud and private cloud. In general by if we used the public cloud we can't provide the security to our private data and hence our private data will be loss. So that we have to provide the security to our data for that we make a use of private cloud also. When we use private clouds the greater security can be provided. In this system we also provides the data deduplication. Which is used to avoid the duplicate copies of data? User can upload and download the files from public cloud but private cloud provides the security for that data. that means only the authorized person can upload and download the files from the public cloud. for that user generates the key and stored that key onto the private cloud. at the time of downloading user request to the private cloud for key and then access that Particular file.

System Model: Now we see the architecture of our system. in our architecture there are three modules . [1] user [2] public cloud [3] private cloud.etc First if the user want to upload the files on the public cloud then user first encrypt that file with the convergent key and then sends it to the public cloud at the same time user also generates the key for that file and sends that key to the private cloud for the purpose of security. In the public cloud we use one algorithm for deduplication. Which is used to avoid the duplicate copies of files which is entered in the public cloud. Hence it also minimizes the bandwidth. that means we requires the less storage space for storing the files on the public cloud. In the public cloud any person that means the unauthorized person can also access or store the data so we can conclude that in the public cloud the security is not provided. In general for providing more security user can use the private cloud instead of using the public cloud. User generates the key at the time of uploading file and stores it to the private cloud. When user wants to downloads the file that he/she upload,. He/she sends the request to the public cloud. Public cloud provides the list of files that are uploads the many user of the public cloud because there is no security is provided in the public cloud. When user selects one of the file from the list of files then private cloud sends a message like enter the key!. User has to enter the key that he generated for that file. When user enter the key the private cloud checks the key for that file and if the key is correct that means user is valid then private cloud give access to that user to download that file successfully. then user downloads the file from the public cloud and decrypt that file by using the same convergent key which is used at the time of encrypt that file. in this way user can make a use of the architecture.

V. ROLES OF ENTITIES

S-CSP: The purpose of this entity to work as a data storage service in public cloud. On the half of the user S-CSP store the data. The S-CSP eliminate the duplicate data using deduplication and keep the unique data as it is. SSCP entity is used to reduce the storage cost. S-CSP hand abundant storage capacity and computational power. When user send respective token for accessing his file from public cloud S-CSP matches this token with internally if it matched then an then only he send the file or cipher text Cf with token, otherwise he send abort signal to user. After receiving file user use convergent key KF to decrypt the file. Data User: A user is an entity that wants to access the data or files from S-SCP. User generates the key and stores that key in private cloud. In storage system supporting deduplication, The user only upload unique data but do not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. Each file is protected by convergent encryption key and can access by only authorized person. In our system user must need to register in private cloud for storing token with respective file which are store on public cloud. When he wants to access that file he access respective token from private cloud and then access his files from public cloud. Token consist of file content F and convergent key KF. Private Cloud: In general for providing more security user can use the private cloud instead of public cloud. User stores the generated key in private cloud. At the time of downloading system ask the key to download the file. User can not store the secrete key internally. for providing proper protection to key we use private cloud. Private cloud only stores the convergent key with respective file. When user wants to access the key he first check authority of user then an then provide key. Public Cloud: Public cloud entity is used for the storage purpose. User uploads the files in public cloud. Public cloud is similar as S-CSP. When the user wants to download the files from public cloud, it will be ask the key which is generated or stored in private cloud. When the users key is match with files key at that time user can download the file, without key user can not access the file. Only authorized user can access the file. In public cloud all files are stored in encrypted format. If any chance unauthorized person hack our file, but without the secrete or convergent key he doesn't access original file. On public cloud there are lots of files are store each user access its respective file if its token matches with S-CSP server token.

Operations performed on Hybrid Cloud

File Uploading: When user want to upload the file to the public cloud then user first encrypt the file which is to be upload by make a use of the symmetric key, and send it to the Public cloud. At the same time user generates the key for that file and sends it to the private cloud. in this way user can upload the file in to the public cloud.

File Downloading: When user wants to download the file that he/she has uploaded on the public cloud. he/she make a request to the public cloud. Then public cloud provides a list of files that many users are upload on it. Among that user select one of the file form the list of files and enter the download option. at that time private cloud sends a message that enter the key for the file generated by the user. then user enters the key for the file that he/she is generated. then private cloud checks the key for that file and if the key is correct that means the user is valid. only then and then the user can download the file from the public cloud otherwise user can't download the file. When user download the file from the public cloud it is in the encrypted format then user decrypt that file by using the same symmetric key.

VI. 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.

VII. 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.

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.