SECURE DATA DEDUPLICATION IN CLOUD STORAGE BY USING HYBRID CLOUD APPROACH

[1] M.ANIL KUMAR, [2] P.NAMRATHA REDDY

[1]M.Tech Student, Department of CSE, INTELL ENGINEERING COLLEGE ,
ANANTHAPURAMU, A.P, India.
[2] Assistant Professor, Department of CSE, INTELL ENGINEERING COLLEGE ,
ANANTHAPURAMU, A.P, India.

Abstract*— *Data deduplication could be a technique for reducing the number of space for storing a company needs to save its knowledge. In most organizations, the storage systems contain copies of many items of information. for instance, identical file is also saved in many totally different places by different users, additional files that are not identical should still embody abundant of identical data. Deduplication eliminates these further duplicates by saving only one original copy of the info and replacing the opposite copies with pointers that lead back to the first copy. Companies frequently use deduplication in backup and disaster recovery applications, however it is used to unlock house in primary storage in addition. To avoid this duplication of information and to keep up the confidentiality within the cloud we tend to victimization the conception of Hybrid cloud. to safeguard the confidentiality of sensitive knowledge whereas supporting deduplication, the focused secret writing technique has been planned to write in code the info before outsourcing. to higher shield knowledge security, this paper makes the primary arrange to formally address the matter of approved knowledge deduplication.

I. INTRODUCTION

In computing, information deduplication may be a specialised information compression technique for eliminating duplicate copies of continuation information. connected and somewhat similar terms area unit intelligent (data) compression and single-instance (data) storage. this method is employed to improve storage utilization and might even be applied to network information transfers to cut back the number of bytes that has got to be sent. within the deduplication method, distinctive chunks of information, or byte patterns, area unit known and keep throughout a method of research. because the analysis continues, other chunks area unit compared to the keep copy and whenever a match happens, the redundant chunk is replaced with atiny low reference that points to the keep chunk. on condition that identical computer memory unit pattern may occur dozens, hundreds, or maybe thousands of times (the match frequency depends on the chunk size), the number of information that has got to be keep or transferred is greatly reduced.

A Hybrid Cloud could be a combined type of non-public clouds and public clouds during which some critical information resides within the enterprise's non-public cloud whereas alternative information is keep in and accessible from a public cloud.

To make information management scalable in cloud computing, deduplication has been a widely known technique and has attracted more and a lot of attention recently. information deduplication may be a specialised information compression technique for eliminating duplicate copies of continuance information in storage. The technique is employed to boost storage utilization and might even be applied to network information transfers to cut back the amount of bytes that has to be sent. rather than keeping multiple information copies with c1onstant content, deduplication eliminates redundant information by keeping only 1 physical copy and referring alternative redundant information to it copy.

Deduplication will occur at either the file level or the block level. For file level deduplication, it eliminates duplicate

copies of constant file. Deduplication also can occur at the block level, that eliminates duplicate blocks of knowledge that occur in non-identical files.

Cloud computing is associate degree rising service model that has computation and storage resources on the net. One attractive practicality that cloud computing offers is cloud storage. people and enterprises ar usually needed to remotely archive their knowledge to avoid any data loss just in case there are any hardware/software failures or unforeseen disasters. rather than getting the required storage media to stay knowledge backups, people and enterprises will merely source their knowledge backup services to the cloud service suppliers, which offer the mandatory storage resources to host the data backups. whereas cloud storage is enticing, the way to offer security guarantees for outsourced knowledge becomes a rising concern. One major security challenge is to produce the property of assured deletion, i.e., knowledge files ar for good inaccessible upon requests of deletion. Keeping knowledge backups for good is undesirable, as sensitive data is also exposed within the future due to knowledge breach or inaccurate management of cloud operators. Thus, to avoid liabilities, enterprises and government agencies typically keep their backups for a finite variety of years and request to delete (or destroy) the backups afterwards. as an example, the U.S. Congress is formulating the net knowledge Retention legislation in asking ISPs to retain knowledge for two years, where as in uk, firms ar needed to retain wages and remuneration records for 6 years.

Although information deduplication brings plenty of advantages, security and privacy issues arise as users' sensitive information square measure susceptible to each business executive and outsider attacks. ancient cryptography, whereas providing information confidentiality, is incompatible with information deduplication. Specifically, ancient cryptography needs totally different users to cypher their information with their own keys. Thus, identical information copies of various users can cause different ciphertexts, creating deduplication not possible. Convergent Encryption has been projected to enforce information confidentiality whereas creating deduplication possible. It encrypts/ decrypts a data copy with a focused key, that is obtained by computing the scientific discipline hash worth of the content of the information copy. After key generation and encryption, users retain the keys and send the ciphertext to the cloud. Since the cryptography operation is settled and comes from the information content, identical information copies can generate an equivalent focused key and hence an equivalent ciphertext. to stop unauthorized access, a secure proof of possession protocol is additionally required to produce the proof that the user so owns an equivalent file once a replica is found. when the proof, ulterior users with an equivalent file will be provided a pointer from the server with no need to transfer an equivalent file. A user will transfer the encrypted file with the pointer from the server, which may solely be decrypted by the corresponding information house owners with their focused keys. Thus, convergent cryptography permits the cloud to perform deduplication on the ciphertexts and therefore the proof of possession prevents the unauthorized user to access the file.

II. RELATED WORK

However, previous deduplication systems cannot support differential authorization duplicate check, that is vital in many applications. In such a licensed deduplication system, every user is issued a group of privileges throughout system initialization. every file uploaded to the cloud is additionally delimited by a group of privileges to specify which type of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for a few file, the user has to take this file and his own privileges as inputs. The user is in a position to seek out a reproduction for this file if and provided that there's a duplicate of this file and a matched privilege keep in cloud. for instance, in an exceedingly company, many various privileges are going to be allotted to employees. so as to avoid wasting value and with efficiency management, the information are going to be affected to the storage server supplier (S-CSP) in the public cloud with nominal privileges and therefore the deduplication technique are going to be applied to store only 1 copy of constant file. attributable to privacy thought, some files are going to be encrypted and allowed the duplicate

check by staff with specified privileges to understand the access management. ancient deduplication systems supported focused cryptography, although providing confidentiality to some extent, don't support the duplicate consult with differential privileges. In different words, no differential privileges are thought of within the deduplication supported focused cryptography technique. It appears to be contradicted if we would like to understand each deduplication and differential authorization duplicate check at constant time.

A. Symmetric Encryption

Symmetric encryption uses a common secret key κ to encrypt and decrypt information. A symmetric encryption scheme consists of three primitive functions:

KeyGenSE(1 λ)= κ is the key generation algorithm that generates κ using security parameter 1 λ.

EncSE(κ,M)= C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C.

DecSE(κ,C)= M is the symmetric decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

B. Convergent Encryption

Convergent encoding ,provides information confidentiality in deduplication. A user (or information owner) derives a confluent key from every original information copy and encrypts the info copy with the confluent key. additionally, the user conjointly derives a tag for the data copy, specified the tag are accustomed find duplicates. Here, we have a tendency to assume that the tag correctness property holds, i.e., if 2 information copies square measure a similar, then their tags square measure a similar. To find duplicates, the user initial sends the tag to the server side to envision if the identical copy has been already keep. Note that each the confluent key and also the tag square measure severally derived, and also the tag can\'t be accustomed deduce the confluent key and compromise information confidentiality. each the encrypted information copy and its corresponding tag are keep on the server facet.

C. Proof of Ownership

The notion of proof of possession (PoW) permits users to prove their possession of knowledge copies to the storage server. Specifically, prisoner is enforced as AN interactive algorithmic program (denoted by PoW) go by a prover (i.e., user) and a admirer (i.e.,storage server). The admirer derives a brief worth Φ(M) from a knowledge copy M. To prove the possession of the info copy M, the prover must send Φ' to the admirer specified $\Phi' = \Phi$ (M). The formal security definition for prisoner roughly follows the threat model in a very content distribution network, wherever AN assailant doesn't grasp the complete file, however has accomplices World Health Organization have the file. The accomplices follow the “bounded retrieval model”, specified they'll facilitate the assailant get the file, subject to theconstraint that they need to send fewer bits than the initial min-entropy of the file to the assailant .

D. Identification Protocol

An identification protocol $\prod$ will be delineated with 2 phases: Proof and Verify. within the stage of Proof, a prover/user U will demonstrate his identity to a admirer by acting some identification proof associated with his identity. The input of the prover/user is his non-public key skU that's sensitive data like non-public key of a public key in his certificate or mastercard range etc. that he wouldn't wish to share with the opposite users. The admirer performs the verification with input of public data phenylketonuria related to skU. At the conclusion of the protocol, the admirer outputs either settle for or reject to denote whether or not the proof is passed or not. There area unit several economical identification protocols in literature, together with certificate-based, identity-based identification etc.

III. FRAME WORK

In the projected system we have a tendency to square measure achieving the info deduplication by providing the proof of ata by the info owner. This proof is employed at the time of loading of the file. Each file uploaded to the cloud is additionally finite by a group of privileges to specify which type of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for a few file, the user has to take this file and his own privileges as inputs. The user is in a position to seek out a replica for this file if and on condition that there\'s a replica of this file and a matched privilege hold on in cloud.

A. Encryption Of *Files*

A. Here we are using the common secret key k to encrypt as well as decrypt data. This will use to convert the plain text to cipher text and again cipher text to plain text. Here we have used three basic functions,

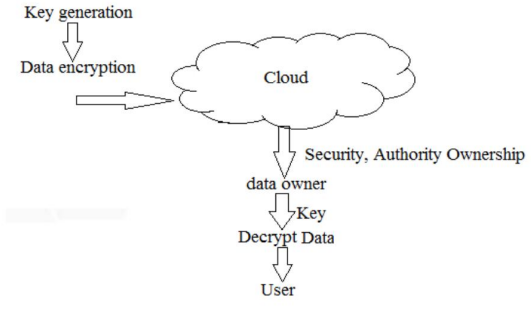
B. ***KeyGenSE****: k is the key generation algorithm that generates κ using security parameter 1.*

C. ***EncSE (k, M)****: C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C.*

DecSE (k, *C*): *M* is the symmetric decryption algorithm that takes the secret *κ* and ciphertext *C* and then outputs the original message *M*.

B. Confidential Encryption

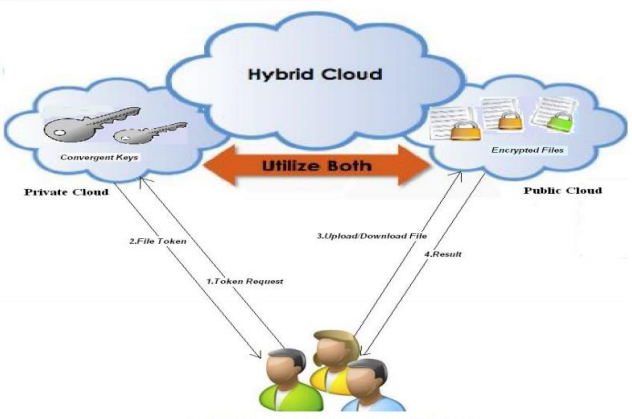
We u It provides knowledge confidentiality in deduplication. A user derives a focused key from each original knowledge copy and encrypts the info copy with the focused key. additionally, the user additionally derives a tag for the info copy, such the tag are going to be wont to sight duplicates.



confidential data encryption

C. Proof Of Data

The user have to be compelled to prove that the info that he need to transfer or transfer is its own data. which means he have to be compelled to give the convergent key and supportive knowledge to prove his ownership at server.

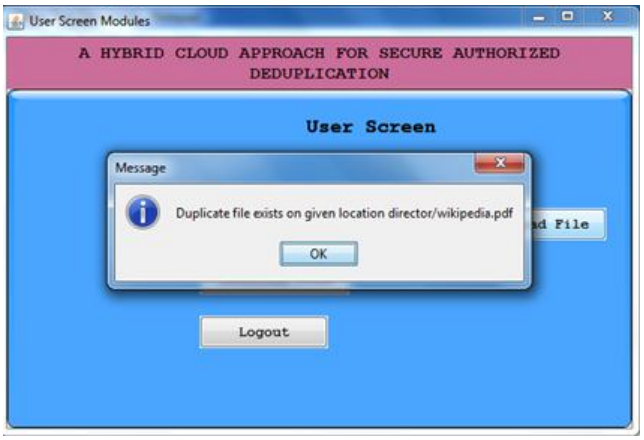


Architecture of Authorized Deduplication

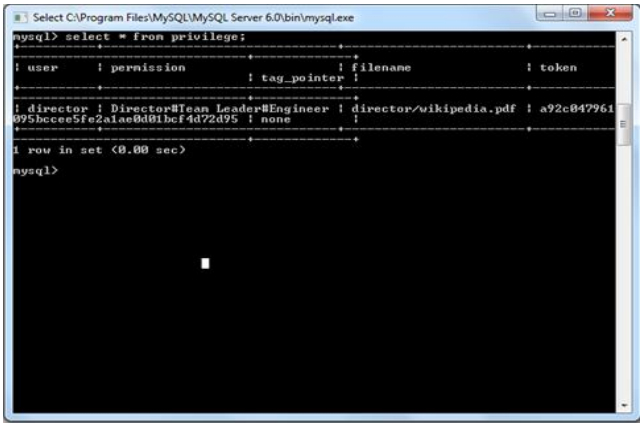
this paper we proposed the system consist of hybrid cloud After registration the users can uploa ther data in to cloud.

data.If you are trying to upoload any similar data(In our proposed system cloud servers not allows the duplicate already uploaded data) it will not allow and shows message like as shown below.

IV. EXPECTED RESULT



When you are uploaded any data the data will be stores in the encrypted format as shown.



V. CONCLUSION

Cloud computing has reached a maturity that leads it into a productive part. This means that most of the most problems with cloud computing are self-addressed to a degree that clouds became attention-grabbing for full business exploitation. This but doesn't mean that all the issues listed on top of have really been resolved, solely that the according risks is tolerated to an exact degree. Cloud computing is so still the maximum amount a groundwork topic, as it is a market giving. For higher confidentiality and security in cloud computing we've planned new deduplication constructions supporting approved duplicate sign on hybrid cloud architecture, within which the duplicate-check tokens of files square measure generated by the personal cloud server with personal keys. planned system includes proof of information owner thus it'll facilitate to implement higher security problems in cloud computing.

REFERENCES

[1] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In *Workshop on Cryptography and Security in Clouds (WCSC 2011)*, 2011.

[2] R. D. Pietro and A. Sorniotti. Boosting efficiency and security in proof of ownership for deduplication. In H. Y. Youm and Y. Won, editors, *ACM Symposium on Information, Computer and Communications Security*, pages 81–82. ACM, 2012.

[3] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, pages 296– 312, 2013.

[4] OpenSSL Project. http://www.openssl.org/.

[5] GNU Libmicrohttpd . http://www.gnu.org/software/libmicrohttpd/.

[6] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In *ICDCS*, pages 617–624, 2002.

[7] D. Ferraiolo and R. Kuhn. Role-based access controls. In *15th NIST-NCSC National Computer Security Conf.*, 1992.

[8] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg. Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, *ACM Conference on Computer and Communications Security*, pages 491–500. ACM, 2011.

[9] libcurl. http://curl.haxx.se/libcurl/.

[10] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In *Proc. of APSYS*, Apr 2013.

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

[12] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan. Sedic: privacyaware data intensive computing on hybrid clouds. In *Proceedings of the 18th ACM conference on Computer and communications security*, CCS'11, pages 515–526, New York, NY, USA, 2011. ACM.