International Journal of Science Engineering and Advance Technology

The cloud Methodology for Reliable approved deduplication

Ravi Raj Kuna[1] R.Chandra sekhar[2]
M.Tech(Research scholar in Software Engineering)[1],Assistant professor[2]
in Dept. of Computer science and engineering
kakinada institute of engineering & technology,korangi,AP,India

Abstract
Data deduplication is one of imperative information pressure methods for wiping out copy duplicates of rehashing information, and has been broadly utilized as a part of cloud storage to decrease the measure of storage room and spare transmission capacity. To secure the secrecy of touchy information while supporting deduplication, the merged encryption strategy has been proposed to scramble the information before outsourcing. Not the same as customary deduplication frameworks, the differential benefits of clients are further considered in copy check other than the information itself. We likewise show a few new deduplication developments supporting approved copy check in a half and half cloud structural planning. In this paper, we introduce a plan that allows an all the more fine-grained exchange off. The instinct is that outsourced information may require distinctive levels of security, contingent upon how well known it is: substance shared by numerous clients, As more corporate and private clients outsource their information to cloud storage suppliers, late information break occurrences make end-to-end encryption an undeniably unmistakable necessity. Lamentably, semantically secure encryption plans render different practical stockpiling enhancement procedures, for example, information deduplication, incapable. We exhibit a clever thought that separates information as indicated by their prevalence. In light of this thought, we plan an encryption plan that ensures semantic security for disliked information and gives weaker security and better stockpiling and transmission capacity advantages for famous information. Along these lines, information deduplication can be compelling for well-known information, whilst semantically secure encryption ensures disagreeable substance. We demonstrate that our plan is secure under the Symmetric External Decisional Diffe-Hellman Assumption in the irregular prophet model.

Keywords: Data deduplication, Confidentiality, Hybrid cloud, Authorized Duplicate check, Authorization.

I. Introduction

Regardless of cloud computing huge fame numerous organizations are being disappointed as they have not discovered what they require in the single cloud environment. • Private mists, which organizations run inside, are secure and available inside of the neighborhood. • Public mists are worked and keep running over web and are less expensive, adaptable and simple to utilize. Up to this point, organizations couldn't monetarily and effectively incorporate and work with both the sorts of architectures as a one cloud framework. So these days' organizations are combining both the general population and private methodologies, making a solitary half breed cloud that will profit from each expressed Panos Tsirigotis, prime supporter of mixture cloud seller cloud speed and boss programming planner. Receiving half breed cloud is simple for some organizations as they will be having in-house cloud and will require just influence the current open cloud abilities noted Professor Kantarcioglu. Business association information volumes are expanding as organizations store and gather tremendous measure of information for their own particular use in cloud. As indicated by Business association technique bunch, by industry investigation more associations want to store their information on to cloud. Accordingly obliges association to have more stockpiling and expend more power and vitality for overseeing and taking care of the information, more system assets are used for transmitting the information and additional time is spend on capacities, for example, replication and information reinforcement. A large portion of the data that is put away is copy information, diverse sources in the same associations for the most part make comparative documents or copy records that as of now exist by which they can work freely. On the off chance that it was conceivable, IT associations would just shield the interesting information from their reinforcements. Rather than sparing everything more than once, the perfect situation is one where just the new or remarkable substance is spared. Information de-duplication gives this essential ability. It offers the capacity to find and expel excess information from inside of a dataset. A dataset can compass a solitary application or compass a whole association. Repetitive information components can be whole records or sub-document information portions inside of a record. In all cases, the target of

the deduplication procedure is to store one of a kind information components just once, yet have the capacity to reconstitute all substance in its unique structure on interest, with 100 percent unwavering quality at plate speeds. Information de-duplication is key to enhancing data insurance, streamlining reinforcement operations, diminishing reinforcement framework, shortening reinforcement windows, and expelling weight from data systems.

To make information administration adaptable in cloud computing, deduplication [5] has been a surely understood procedure which is being utilized by the greater part of the clients. Information Deduplication is one of the particular information pressure methods which is utilized to wipe out copy duplicates of information. Deduplication can happen at record level or either square level. For record level deduplication, it dispenses with copy duplicates of the same document. Deduplication can likewise happen at the square level, which dispenses with copy pieces of information that happen in non-indistinguishable records. In spite of the fact that there are a few preferences of information deduplication security and protection concerns emerge as clients' delicate information are powerless to both insider and outcast assaults. Encryption strategies which were utilized generally were not perfect with information deduplication while giving information secrecy. Conventional encryption requires distinctive clients to encode their information with their own keys by which indistinguishable information duplicates of diverse clients will prompt distinctive figure writings, making deduplication incomprehensible. United encryption [4] has been proposed to uphold information secrecy while making deduplication attainable. It scrambles/unscrambles an information duplicate with a merged key, which is acquired by processing the cryptographic hash estimation of the substance of the information duplicate. At whatever point the key is produced clients hold the keys and send the figure content to the cloud. Keeping in mind the end goal to avert unapproved access, a protected evidence of proprietorship convention [2] is likewise expected to give the verification that the client to be sure possesses the same record when a copy is found. Subsequently joined encryption permits the cloud to perform deduplication on the ciphertexts and the verification of possession keeps the unapproved client to get to the document. Conventional deduplication frameworks in light of focalized encryption, in spite of the fact that giving secrecy to some degree; don't bolster the copy check with differential benefits. Inconsistency happens when we attempt to acknowledge both deduplication and differential approval copy check in the meantime.

II. Related Work

"A secure cloud backup system with assured deletion and version control. A. Rahumed", H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui [1],has presented Cloud storage is an emerging service model that enables individuals and enterprises to outsource the storage of data backups to remote cloud providers at a low cost. Hence results shows that FadeVersion only adds minimal performance overhead over a traditional cloud backup service that does not support assured deletion."A reverse deduplication storage system optimized for reads to latest backups", C. Ng and P. Lee. Revdedup [2] had present RevDedup, a de-duplication system designed for VM disk image backup in virtualization environments. RevDedup has several design goals: high storage efficiency, low memory usage, high backup performance, and high restore performance for latest backups. They extensively evaluate our RevDedup prototype using different workloads and validate our design goals. "Role-based access controls", D. Ferraiolo and R. Kuhn [3],has described the Mandatory Access Controls (MAC) are appropriate for multilevel secure military applications, Discretionary Access Controls (DAC) are often perceived as meeting the security processing needs of industry and civilian government. "Secure deduplication with efficient and reliable convergent key management", J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou [4], had proposed Dekey, an efficient and reliable convergent key management scheme for secure de-duplication. They implement Dekey using the Ramp secret sharing scheme and demonstrate that it incurs small encoding/decoding overhead compared to the network transmission overhead in the regular upload/download operations." Reclaiming space from duplicate files in a server less distributed file system", J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. [5],has presented the Farsite distributed file system provides availability by replicating each file onto multiple desktop computers. Measurement of over 500 desktop file systems shows that nearly half of all consumed space is occupied by duplicate files. The mechanism includes 1) convergent encryption, which enables duplicate files to coalesced into the space of a single file, even if the files are encrypted with different users' keys, and 2) SALAD, a Self Arranging, Lossy, Associative Database for aggregating file content and location information in a decentralized, scalable, faulttolerant manner.."A secure data deduplication scheme for cloud storage", J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl [6],has provided the private users outsource their data to cloud storage providers, recent data breach incidents make end-toend encryption an

increasingly prominent requirement data deduplication can be effective for popular data, whilst semantically secure encryption protects unpopular content. "Weak leakage-resilient clientside deduplication of encrypted data in cloud storage", J. Xu, E.-C. Chang, and J. Zhou [7],has described the secure client-side deduplication scheme, with the following advantages: our scheme protects data confidentiality (and some partial information) against both outside adversaries and honest-but-curious cloud storage server, while Halevi et al. trusts cloud storage server in data confidentiality. "Secure and constant cost public cloud storage auditing with deduplication", J. Yuan and S. Yu[8] has proposed, Data integrity and storage efficiency are two important requirements for cloud storage. The author proposed scheme is also characterized by constant realtime communication and computational cost on the user side. "Privacy aware data intensive computing on hybrid clouds", K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan [9] has proposed, the emergence of cost-effective cloud services offers organizations great opportunity to reduce their cost and increase productivity.The system, called Sedic, leverages the special features of Map Reduce to automatically partition a computing job according to the security levels of the data it works. "Gq and schnorr identification schemes Proofs of security against impersonation under active and concurrent attacks", M. Bellare and A. Palacio[10] has provided, the proof for GQ based on the assumed security of RSA under one more inversion, an extension of the usual onewayness assumption that was introduced. Both results extend to establish security against impersonation under concurrent attack.

III. Data duplication problem in cloud

Storage efficiency functions such as deduplication afford storage providers better utilization of their storage back ends and the ability to serve more customers with the same infrastructure. It is the process by which a storage provider only stores a single copy of a file owned by several of its users and there are four different deduplication strategies, depending on whether deduplication happens at the client side (i.e. before the upload) or at the server side, and whether deduplication happens at a file level or at a block level. Deduplication is most rewarding when it is triggered at the client side, as it also saves upload bandwidth but For these reasons, deduplication is a critical enabler for a number of popular and successful storage services which offers a cheap, remote storage to the broad public by performing client-side deduplication, thus it will saving both the network bandwidth and storage costs. Indeed, data deduplication is arguably one of the main reasons why the prices for cloud storage and cloud backup services have dropped so sharply. As the world moves to digital storage for archival purposes, there is an increasing demand for systems that can provide a secure data storage in a cost-effective manner. By identifying the common chunks of data both within and between files and storing them only once, by this deduplication can yield cost savings by increasing the utility of a given amount of storage but Unfortunately, deduplication exploits identical content, while encryption attempts to make all content appear random, when the same content encrypted with two different keys results in very different ciphertext. Thus, in encryption combining the space efficiency of deduplication with the secrecy aspects is problematic. Although data deduplication brings a lot of benefits to cloud user, security and privacy concerns arise as users sensitive data are susceptible to both insider and outsider attacks. While Traditional encryption, providing data confidentiality, is incompatible with data deduplication. Specifically, traditional encryption requires different users to encrypt their data with their own keys. Thus, identical data copies of different users will lead to a different ciphertexts, which makes deduplication impossible. Thus Convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible.

IV. Security issues in cloud

The security will be analyzed in terms of two aspects, that is, the confidentiality of data and the authorization of duplicate check. We suppose that all the files are sensitive and needed to be fully protected against both public cloud and private cloud. Under this assumption, two kinds of adversaries are considered, that is, adversaries which aim to extract secret information as much as possible from both public cloud and private cloud, and internal adversaries who aim to obtain more information on the file from the public cloud and duplicate-check token information from the private cloud outside of their scopes. The data will be encrypted in our deduplication system before outsourcing to the storage cloud to maintain the confidentiality of data. The data is encrypted with the traditional encryption scheme and the data encrypted with such encryption method which guarantees the security of data. System address the problem of privacy preserving deduplication in cloud computing and propose a new deduplication system supporting for Differential Authorization and Authorized Duplicate Check. Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges. Under this assumption, any unauthorized user cannot generate a token for duplicate check out of his privileges or without the aid from the private

cloud server. Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs the duplicate check directly and tells the user if there is any duplicate. The security requirements considered in two folds, including the security of data files and security of file token. For the security of file token. Unauthorized users without appropriate privileges or file prevented from getting or generating the file tokens for duplicate check of any file stored at the Storage cloud. The users are not allowed to collude with the public cloud server. It requires that any user without querying the private cloud server for some file token, he cannot able to get any useful information from the token, which includes the privilege or the file information and to maintain the data confidentiality unauthorized users without appropriate privileges or files, prevented from access to the underlying plaintext stored at Storage cloud.

V. A Detailed Look at Data De-Duplication

Data de-duplication has many forms. Typically, there is no one best way to implement data de-duplication across an entire an organization. Instead, to maximize the benefits, organizations may deploy more than one de-duplication strategy. It is very essential to understand the backup and backup challenges, when selecting de-duplication as a solution. Data de-duplication has mainly three forms. Although definitions vary, some forms of data de-duplication, such as compression, have been around for decades. Lately, single-instance storage has enabled the removal of redundant files from storage environments such as archives. Most recently, we have seen the introduction of sub-file de-duplication. These three types of data de-duplication are described below

A. Data Compression

Data compression is a method of reducing the size of files. Data compression works within a file to identify and remove empty space that appears as repetitive patterns. This form of data de-duplication is local to the file and does not take into consideration other files and data segments within those files. Data compression has been available for many years, but being isolated to each particular file, the benefits are limited when comparing data compression to other forms of de-duplication. For example, data compression will not be effective in recognizing and eliminating duplicate files, but will independently compress each of the files.

B. Single-Instance Storage

Removing multiple copies of any file is one form of the de-duplication. Single-instance storage (SIS) environments are able to detect and remove redundant copies of identical files. After a file is stored in a single-instance storage system than, all the other references to same file, will refer to the original, single copy. Single-instance storage systems compare the content of files to determine if the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically equipped with single-instance storage functionality. While file-level de-duplication avoids storing files that are a duplicate of another file, many files that are considered unique by single-instance storage measurement may have a tremendous amount of redundancy within the files or between files. For example, it would only take one small element (e.g., a new date inserted into the title slide of a presentation) for single-instance storage to regard two large files as being different and requiring them to be stored without further de-duplication.

C. Sub-file De-Duplication

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated—even after the duplicated data exist, within separate files. This form of de-duplication discovers the unique data elements within an organization and detects when these elements are used within other files. As a result, sub-file de-duplication eliminates the storage of duplicate data across an organization. Sub-file data de-duplication has tremendous benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization. Sub-file de-duplication implementation has two forms. Fixed-length sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files. Although simple in design, fixed-length segments miss many opportunities to discover redundant sub-file data. (Consider the case where an addition of a person's name is added to a document's title page—the whole content of the document will shift, causing the failure of the de-duplication tool to detect equivalencies). Variable-length implementations are usually not locked to any of arbitrary segment length. Variable-length implementations match data segment sizes to the naturally occurring duplication within files, vastly increasing the overall de-duplication ratio (In the example above, variable-length de-duplication will catch all duplicate segments in the document, no matter where the changes occur). So most of the organizations widely use data duplication technology, which is also called as, single-instance storage, intelligent compression, and capacity optimized storage and data reduction.

Fig. Example of De-duplication

Fig. shows, de-duplication finds the redundant data and eliminates all but keeps one copy which creates logical pointers to the file so that users could access the file as and when needed. [4] Pointers and References. De-duplication systems removes redundant data that they find then creates a reference or logical pointer to single instance of data that host keeps. There are pointers at places where multiple users store the same single piece of information.

VI. Propose authorized duplication checker for clouds

There are three entities defined in system, that is, users, private cloud and storage cloud servise provider in public cloud as shown in Fig. 1. The Storage cloud performs deduplication by checking if the contents of two files are the same and stores only one of them and The access right to a file is defined based on a set of privileges. Each privilege is represented in the form of a short message called token. Each file is associated with some file tokens, which denotes the tag with specified privileges. A user computes and sends duplicate-check tokens to the public cloud for authorized duplicate check. While Users have access to the private cloud server, a semitrusted third party which perform duplicable encryption by generating file tokens for the requesting users.

Fig. Proposed Architecture for Authorized Deduplication

Storage Cloud

This is an entity that provides a data storage service in public cloud. The storage cloud service provider provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the storage cloud eliminates the storage of redundant data via deduplication and keeps only unique data.

Data User

A user is an entity that wants to outsource data storage to the S-CSP and access the data later when needed. In a storage system supporting deduplication, to save the upload bandwidth the user can only uploads unique data but does not upload any duplicate data, which may be owned by the same user or the different users. In authorized deduplication system, each user is issued a set of privileges in the setup of the system and each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

Private Cloud

The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users and this interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

In deduplication system, a hybrid cloud architecture is introduced to solve the problem of unauthorized deduplication of file. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server. The user needs to send a request to the private cloud server to get a file token. The user needs to get the file token from the private cloud server to perform the duplicate check for some file. The private cloud server also check the user's identity before issuing the corresponding file token to the user. The user perform the authorized duplicate check for this file with the public cloud before uploading this file. The user either uploads this file or prove their ownership based on the results of duplicate check. If a file duplicate is found, the user needs to run the Proof of ownership protocol with the cloud storage service provider to prove the file ownership. Otherwise, if no duplicate is found then the data owner performs an identification to prove its identity with private key. If it is passed, the private cloud server will find the corresponding privileges of the user from its stored table list and send to the user then user can upload his files. The same way user can download his file from storage cloud.

VII. Proposed Algorithm

A convergent encryption scheme can be defined with four primitive functions:

• KeyGenCE(M) !K is the key generation algorithm that maps a data copy M to a convergent key K;

• EncCE(K, M) !C is the symmetric encryption algorithm that takes both the convergent key K and the data copy M as inputs and then outputs a ciphertextC;
• DecCE(K, C) !M is the decryption algorithm that takes both the ciphertextC and the convergent key K as inputs and then outputs the original data copy M; and
• TagGen(M) !T (M) is the tag generation algorithm that maps the original data copy M and outputs a tag T (M).

The notion of proof of ownership(PoW) [11] enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm (denoted by PoW) The verifier derives a short value $\phi(M)$ from a data copy M. To prove the ownership of the data copy M, the prover needs to send ϕ' to the verifier such that $\phi' = \phi(M)$

VIII. Conclusion And Future Work

Several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct tested experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

Future Work

We plan to investigate the secure deduplication issue in cloud backup services of the personal computing environment. We can further explore and exploit index lookup parallelism availed by the application-aware index structure of Deduplication in multi core environment.

References

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou" A Hybrid Cloud Approach for Secure Authorized De-duplication" in vol: pp no-99, IEEE, 2014
[2] OpenSSL Project. http://www.openssl.org/.
[3] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.
[4] M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.
[5] M. Bellare, S. Keelveedhi, and T. Ristenpart. Messagelocked encryption and secure eduplication. In EUROCRYPT, pages 296– 312, 2013.
[6] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.
[7] M. Bellare and A. Palacio. Gq and schnorr dentification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.
[8] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. chneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.
[9] 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.
[10] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15th NIST-NCSC National Computer Security Conf., 1992.

Authors:

Ravi Raj Kuna pursuing Master of Computer Technology in Software Engineering in KIET Kakinda, East Godavari, A.P, India.

R.Chandra sekhar working as an assistant Professor in KIET college of engg, Kakinada. Presently he is pursuing Ph.D in Big Data. His area of interests are cloud computing and Big data.