IJSEAT
International Journal of
Science Engineering and Advance Technology

The cloud Approach for Consistent Appropriate deduplication

A.Dheeraj[1], P.Krishna Sai[2]

#1. M.Tech Graduate with Specialization in Information Technology, Department of Computer Science & System Engineering Andhra University College of Engineering

#2. M.Tech Graduate with Specialization in Software Engineering, Department of Computer Science & System Engineering Andhra University College of Engineering, Visakhapatnam, India.

Abstract

In Cloud processing includes conveying gatherings of remote servers and programming systems that permit brought together information stockpiling and online access to PC administrations or assets. Mists can be named open, private or mixture. Cloud administration suppliers offer both exceedingly accessible capacity and hugely parallel registering assets at moderately low expenses. As Cloud computing gets to be pervasive, an expanding measure of information is being put away in the cloud and imparted by clients to indicated benefits, which characterize the entrance privileges of the put away information. One basic test of Cloud storage administrations is the administration of the constantly expanding volume of information. Datadeduplication is a particular information pressure strategy for wiping out copy duplicates of rehashing information. This method is utilized to enhance stockpiling usage and can likewise be connected to network information exchanges to diminish the quantity of bytes that must be sent and spare transmission capacity. To ensure the secrecy of delicate information while supporting deduplication, the merged encryption system is utilized to scramble the information before outsourcing. It encodes/unscrambles an information duplicate with a joined key, which is acquired by figuring the cryptographic hash estimation of the substance of the information duplicate. Focalized encryption permits the cloud to perform deduplication on the ciphertexts and the confirmation of possession keeps the unapproved client to get to the document. To improve the framework in security OAuth is utilized. OAuth (Open Authorization) is an open convention for token-construct confirmation and approval with respect to the Internet utilized as a part of half and half cloud to improve the security. OAuth empowers the framework to guarantee that the client is a verified individual or not. Just such confirmed client got the token for transferring and downloading in pubic cloud.

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

I. Introduction

Cloud computing is the new developing patterns in the new era innovation. Each client has immense measure of information to share to store in a rapidly accessible secured place. The idea of deduplication is touched base here to proficiently use the data transfer capacity and plate utilization on Cloud computing. To keep away from the duplication duplicates of the same information on cloud may bring about loss of time, data transfer capacity use and space. Cloud computing is web based, a system of remote servers associated over the Internet to store, offer, control, recover and handling of information, rather than a neighborhood server or PC. The advantage of Cloud computing are gigantic. It empowers us to work from anyplace. The most imperative thing is that client doesn't have to purchase the asset for information stockpiling. With regards to Security, there is a plausibility where a vindictive client can infiltrate the cloud by mimicking an authorize client, there by influencing the whole cloud in this manner tainting numerous clients who are sharing the contaminated cloud. There is additionally huge issue, where the copy duplicates may transfer to the cloud, which will prompt misuse of band width and circle use. To enhance this issue there ought to be a decent level of encryption gave, that just the client ought to have the capacity to get to the information and not the honest to goodness User. Yan Kit Li et al.[6] appeared To formally tackle the issue of approved information deduplication. Information deduplication is an information pressure methods for evacuating copy duplicates of indistinguishable information, and it is utilized as a part of Cloud storage to spare transfer speed and to diminish the sum storage room. The procedure is used to improve the capacity utilize and can similarly be connected to network information trade to decrease the measure of bytes that must be sent. Keeping different information duplicates with the indistinguishable substance, de-duplication evacuates repetitive information by keeping stand out duplicate and alluding other indistinguishable information to that duplicate. De-duplication happens either at piece level or at record level. In document level de-duplication, it evacuated copy duplicates of the indistinguishable record. Deduplication can likewise happen in the square level that kills copy pieces of information that is happened in non-indistinguishable records. Information deduplication having immense measure of focal points like giving security and additionally protection concerns emerge as client's touchy or fragile information are at danger to both insider and untouchable assaults. The customary encryption requires a wide range of clients for encoding the information records with their own private keys. Along these lines, the same information duplicates of distinctive clients will prompt diverse figure writings, making de-duplication inconceivable. To secure the protection of delicate data while supporting deduplication, the united encryption procedure has been proposed to encode the data before outsourcing. This paper will work to break down the security issue and to assess the proficient use of cloud band width and plate utilization.

II. Related Work

"A protected cloud reinforcement framework with guaranteed cancellation and variant control. A. Rahumed", H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S.Lui [1],has introduced Cloud stockpiling is a developing administration demonstrate that empowers people and ventures to outsource the capacity of information reinforcements to remote cloud suppliers requiring little to no effort. Henceforth results demonstrates that FadeVersion just includes negligible execution overhead over a conventional cloud reinforcement benefit that does not backing guaranteed deletion."A reverse deduplication stockpiling framework advanced for peruses to most recent reinforcements", C. Ng and P. Lee. Revdedup [2] had present RevDedup, a de-duplication framework intended for VM circle picture reinforcement in virtualization situations. RevDedup has a few outline objectives: high stockpiling effectiveness, low memory use, high reinforcement execution, and high restore execution for most recent reinforcements. They widely assess our RevDedup model utilizing distinctive workloads and accept our outline objectives. "Part based access controls", D. Ferraiolo and R. Kuhn [3],has depicted the Mandatory Access Controls (MAC) are fitting for multilevel secure military applications, Discretionary Access Controls (DAC) are regularly seen as meeting the security handling needs of industry and non military personnel government. "Secure deduplication with productive and solid concurrent key administration", J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou [4], had proposed Dekey, an effective and solid concurrent key administration plan for secure de-duplication. They actualize Dekey utilizing the Ramp mystery sharing plan and show that it brings about little encoding/unraveling overhead contrasted with the system transmission overhead in the consistent transfer/download operations." Reclaiming space from copy documents in a server less dispersed record framework", J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. [5],has displayed the Farsite conveyed document framework gives accessibility by recreating every record onto different desktop PCs. Estimation of more than 500 desktop record frameworks demonstrates that almost 50% of all devoured space is involved by copy documents. The component incorporates 1) merged encryption, which empowers copy documents to combine into the space of a solitary record, regardless of the fact that the documents are encoded with diverse clients' keys, and 2) SALAD, a Self Arranging, Lossy, Associative Database for totaling document substance and area data in a decentralized, adaptable, faulttolerant manner.."A secure information deduplication plan for distributed storage", J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl [6],has gave the private clients outsource their information to distributed storage suppliers, late information rupture occurrences make end-toend encryption an undeniably conspicuous necessity information deduplication can be compelling for prevalent information, whilst semantically secure encryption ensures disagreeable substance. "Frail spillage versatile clientside deduplication of encoded information in distributed storage", J. Xu, E.- C. Chang, and J. Zhou [7],has portrayed the safe customer side deduplication plan, with the accompanying points of interest: our plan ensures information privacy (and some incomplete data) against both outside enemies and fair however inquisitive distributed storage server, while Halevi et al. trusts distributed storage server in information secrecy. "Secure and consistent expense open distributed storage reviewing with deduplication", J. Yuan and S. Yu[8] has proposed, Data uprightness and capacity proficiency are two imperative necessities for distributed storage. The creator proposed plan is likewise described by steady realtime correspondence and computational expense on the client side. "Security mindful information concentrated figuring on half breed mists", K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan [9] has proposed, the rise of financially savvy cloud administrations offers associations extraordinary chance to lessen their expense and expand productivity.The framework, called Sedic, influences the exceptional elements of Map Reduce to naturally segment a figuring occupation as per the security levels of the information it works. "Gq and schnorr distinguishing proof plans Proofs of security against mimic under dynamic and simultaneous assaults", M. Bellare and A. Palacio[10] has given, the confirmation for GQ in light of the expected security of RSA under one more reversal, an expansion of the standard onewayness supposition that was presented. Both results stretch out to set up security against mimic under simultaneous assault.

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. 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.

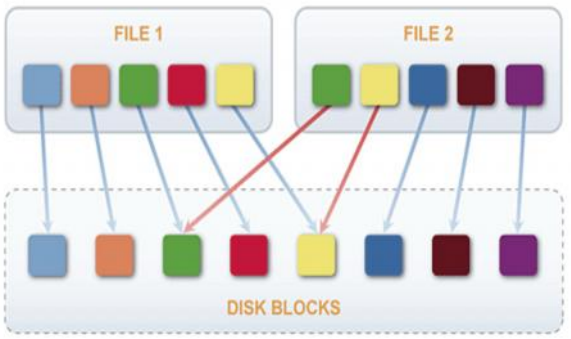

Fig. Example of De-duplication

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.

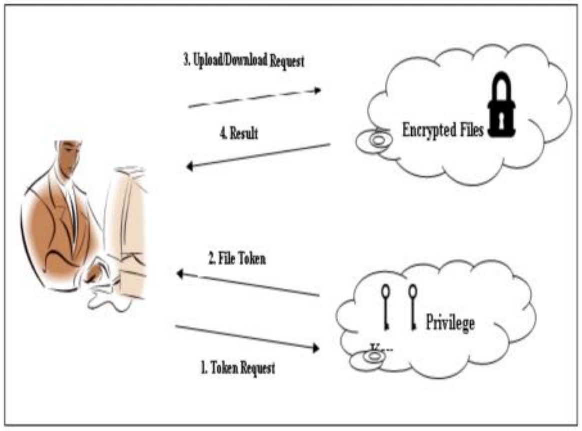

Fig. Proposed Architecture for Authorized Deduplication

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

Dheeraj Alamuri is an M.Tech graduate with Information Technology as specialization in Department of Computer Science & System Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India. His area of interest includes Web technologies, Network Security & Cryptography and DBMS.

Krishna Sai A.R. Patnana is an M.Tech graduate with Software Engineering as specialization in Department of Computer Science & System Engineering Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India. His area of interest includes Image Processing, Database Management Systems and Network Security & Cryptography.