The Secure Integrity Verification in Cloud Storage Auditing with Deduplication

[1]Pyla. Naresh, [2]K. Ravindra, [3]Dr. A. Chandra Sekhar
[1,2,3]Dept. of CSE, Avanthi Institute of Engineering and Technology, Vizianagaram, A.P, India

Abstract
The cloud computing innovation appeared amid the21st century; outsourcing data to cloud benefit for capacity turns into a helpful yet proficient pattern, which benefits in saving endeavors on data support and administration. By the by, since the outsourced cloud stockpiling administration is not completely reliable, it raises security worries on the most proficient method to acknowledge data de-duplication in cloud while accomplishing integrity auditing. In this work, we concentrate the issue of secure de-duplication on cloud data, likewise guaranteeing integrity. In particular, going for accomplishing both data integrity and de-duplication in cloud, we propose a framework, specifically D-cloud. D-cloud presents an auditing element with support of the cloud, which creates hash esteem before transferring and audit the integrity of data having been put away in cloud. Contrasted and past work, the calculation by client in D-Cloud is extraordinarily lessened amid the document transferring and auditing stages. D-cloud is planned understanding the way that clients dependably need to encode their data being transferred, and empowers integrity auditing and secure de-duplication on scrambled data. The primary danger for this cloud data stockpiling is data security as far as keeps up data integrity and data deduplication on cloud. Taking care of both issue normal time is the troublesome assignment. SecCloud and SecCloud+ are two new cloud auditing frameworks which help in keeping up cloud data integrity with productive data deduplication, In SecCloud framework, client can ready to create data labels before putting away data on cloud which encourages amid performing audit to check integrity of data, opposite side SecCloud+ framework give encryption of data before transferring it, which empowers integrity check and secure deduplication of encoded data.

Keywords
Reliability, Data Sharing, Deduplication, Distributed Storage System, Auditing

I. Introduction

Cloud computing comprises a gathering of PCs and servers that are flexible through the Internet. Client get to the data's and will pay according to client premise. Cloud computing has four fundamental qualities: versatility and the capacity to scale here and there, self-benefit provisioning and programmed de-provisioning, application programming interfaces (APIs), charging and metering of administration utilization in a compensation as-you-go show. While Cloud Computing makes these points of interest more engaging than any other time in recent memory, it likewise brings new and testing security dangers towards users‟ outsourced data. Since cloud specialist organizations (CSP) are separate regulatory substances, data outsourcing is really giving up clients extreme control over the destiny of their data. The accuracy of the data in the cloud is being put at hazard because of the accompanying reasons. As a matter of first importance, despite the fact that the foundations under the cloud are significantly more intense and dependable than individualized computing gadgets, they are as yet confronting the wide scope of both inner and outside dangers for data integrity and capacity administration. The cloud stockpiling administration (CSS) soothes the weight of capacity administration and support. Be that as it may, if such an essential administration is helpless against assaults or disappointments, it would convey lost misfortunes to clients since their data or documents are put away into an unverifiable stockpiling pool outside the ventures. The ability gave to the purchaser is to utilize the supplier's applications running on a cloud framework. The applications are open from various customer gadgets through a thin customer interface, for example, a web program (e.g., online email). The shopper does not oversee or control the fundamental cloud foundation including system, servers, working frameworks, stockpiling data, or even individual application abilities, with the conceivable exemption of constrained client particular application arrangement settings. Despite the fact that data deduplication brings a great deal of advantages, security and isolation nerves emerge as clients' delicate data are helpless to both inside and outside assaults. Accordingly, indistinguishable data duplicates of various clients will prompt to various figure writings, making deduplication unimaginable. Jenkins encryption has been proposed to authorize data classification while making deduplication feasible. It scrambles/decodes a data duplicate with a Jenkins key, which is acquired by computing the cryptographic hash estimation of the substance of the data copy. After key era and data encryption, clients safeguard the keys and send the figure content to the cloud. Since the encryption operation is dismay monistic and is gotten from the data fulfilled, indistinguishable data duplicates will cause the same focalized key and thus the equivalent figure content.

II. Related Work

Late years have seen the pattern of utilizing cloud-based administrations for expansive scale satisfied capacity, handling, and dissemination. Security and protection are among top attentiveness toward the general population cloud situations. That is, each customer processes according to data key to scramble the data that he expects to store in the cloud. In that capacity, the data get to is fared by the data proprietor. Second, by absorbing access benefits in metadata record, an endorsed client can decode a scrambled document just with his private key. In spite of the critical points of interest in sparing assets, customer data deduplication brings numerous security issues, significantly due to the multi-proprietor data ownership challenges. For example, a few assaults target either the transmission capacity utilization or the secrecy and the security of honest to goodness cloud clients. For instance, a client may check whether another client has as of now transferred a record, by attempting to outsource a similar document to the cloud. This paper presents another cryptographic technique for secure Proof of Ownership (PoW), in view of the joint utilization of Jenkins encryption and the Merkle-based Tree, for enhancing data security in cloud stockpiling frameworks, giving element sharing amongst clients and guaranteeing productive data deduplication. Our thought comprises in utilizing the Merkle-based Tree over scrambled data, so as to start an unmistakable identifier of subcontracted data. On one hand, this identifier serves

to check the accessibility of similar data in remote cloud servers. Then again, it is utilized to guarantee proficient get to control in element sharing situations. From the point of view of cloud stockpiling security, there have been two prominent thoughts: Proof of Data Possession (PDP) It permits a cloud customer to confirm the integrity of its data subcontracted to the cloud in an extremely effective manner. This is conceivable in light of the fact that it could be extremely asset expending to stack an expansive data record from optional stockpiling to memory. Confirmation of Retrievability (POR) This idea was presented by Juels and Kaliski.. This clarifies the expression "deduplication". This issue was initially acquainted with the examination group. Since direct deduplication is powerless against assaults Halevi proposed the thought called Proof of Ownership (POW) and additionally solid developments.

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.

VI. The System Model

The system model consist three different entities: the cloud user, the cloud server (CS) and the third- party auditor (TPA). As shown in fig. 1. The cloud user is the one who has large amount of data files that are stored in the cloud; the cloud server is the one who provides the data storage service like resources, software to the user. The cloud server is managed by cloud service provider; the third- party auditor is the one who has belief to access the cloud storage service for the benefit of user whenever user request for data access. The TPA has capabilities and competence that the user does not have. They can also interact with cloud server to access the stored data for different purpose in different style. Every time it is not possible for user to check the data which is stored on cloud server that arrives online burden to the user .so that"s why to reduce online burden and maintain that integrity cloud

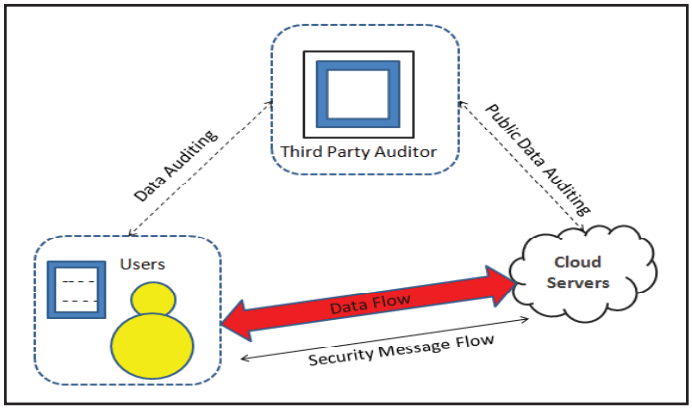

Fig. 1: The Architecture of Cloud Data Storage

User may resort to TPA. The data stored on cloud server is come from internal and external attacks, which is having data integrity threads like hardware failure, software bug, hackers, and management errors. The Cloud Server can maintain reputation for its self-serving. The CS might even decide to hide these data correction incidents to user. So that's why here we are giving third-party auditing service for users to gain belief on cloud.

In this, we address the problem of privacy preserving de-duplication in cloud computing and propose a new deduplication system supporting for, the ϖ

- **Differential Authorization:** To perform duplicate check based on privilege of user is able to get his/her individual token. Without aid from the private cloud server and for the duplicate check outs token cannot generate by the user.
- **Authorized duplicate check:** 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 duplicate check directly and tells the user if there is any duplicate. The security requirements considered in this paper lie in two folds, including the security of file token and security of data files. For the security of file token, two aspects are defined as un-forge ability and in-distinguish ability of file token. The details are given below.
- **Unforgeability of file token/duplicate-check token:** Unauthorized users without appropriate privileges or file should be prevented from getting or generating the file tokens for duplicate check of any file stored at the S-CSP. The users are not allowed to collude with the public cloud server to break the unforgeability of file tokens. In our system, the S-CSP is honest but curious and will honestly perform the duplicate check upon receiving the duplicate request from users. The duplicate check token of users should be issued from the private cloud server in our scheme.
- **Indistinguishability of file token/duplicate check token.** It requires that any user without querying the private cloud server for some file token, he cannot get any useful information from the token, which includes the file information or the privilege information.
- **Data Confidentiality.** Unauthorized users without appropriate privileges or files, including the S-CSP and the private cloud server, should be prevented from access to the underlying plaintext stored at S-CSP. In another word, the goal of the adversary is to retrieve and recover the files that do not belong to them. In our system, compared to the previous definition of data confidentiality based on convergent encryption, a higher level confidentiality is defined and achieved.

VII. Proposed System

All the existing applications discussed are kind of more commercial and money making, but this web application is different. Mainly this web application deals with the important factor like De-duplication, Security, Integrity and Availability. The sharing of data is easy but the one thing we should take care of is the security because we don't want anybody should see our data in the cloud without permission of the primary user. Here using Cloud storage[1], it will help the users to store their data on a network and can retrieve it easily from their when it is needed and don't need to store it on hard-drive. Also the reason for making this web application is that, the data in cloud is not fully trustworthy and raise security concerns. The high cost of data storage devices and the use of data rapidly make us to use cloud storage

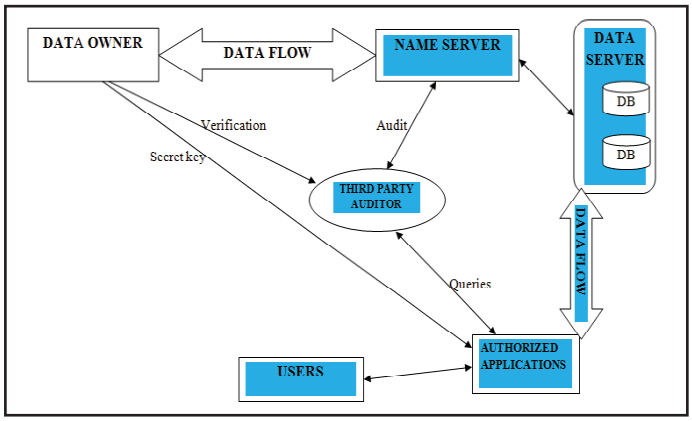

Fig. 2: Proposed System Architecture

VIII. Conclusion

Data de-duplication is important technique used in cloud computing. But data deduplication technique can't be used alone in cloud, because there is often need of data security. So data de--duplication and convergent encryption work in collaboration such that, data deduplication is possible with security of data. But convergent encryption does not provide much security, as it can be susceptible to guessing and brute force attacks. Also current data deduplication technique does not provide support for differential privilege level deduplication. This system is useful in currently changing industry where it is necessary to consider privilege levels of employees in data deduplication, so that, it will enhance data deduplication process and security. This paper provides an abstract view of different schemes proposed in recent past for cloud data security using auditor. Most of the authors have proposed schemes which rely on encrypting the data using some encryption algorithm and make auditor store a message digest or encrypted copy of the same data that is stored with the service provider. The Auditor is used to resolve any kind of conflicts between service provider and client.

References

[1] P. Anderson, L. Zhang,"Fast and secure laptop backups with encrypted de-duplication", USENIX LISA, 2010.

[2] M. Bellare, S. Keelveedhi, T. Ristenpart,"Dupless: Serveraided encryption for deduplicated storage", In USENIX Security Symposium, 2013.

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

[4] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon,"In ICDCS, pp. 617 -624, 2002. Reclaiming space from duplicate files in a serverless distributed file system.

[5] J. Li, X. Chen, M. Li, J. Li, P. Lee, W. Lou,"Secure deduplication with efficient and reliable convergent key management", In IEEE Transactions on Parallel and Distributed Systems 2013.

[6] S. Quinlan, S. Dorward. Venti,"A new approach to archival storage", USENIX FAST, Jan 2002.

[7] A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, J. C. S. Lui.,"A secure cloud backup system with assured deletion and version control", In International Workshop on Security in Cloud Computing 2011.

[8] Z. Wilcox-O'Hearn, B. Warner. Tahoe,"The leastauthority filesystem", ACM StorageSS, 2008.

[9] J. Xu, E.-C. Chang, J. Zhou,"Weak leakage resilient client side deduplication of encrypted data in cloud storage", In ASIACCS, pp. 195–206, 2013.

[10] J. Yuan, S. Yu.,"Secure and constant cost public cloud storage auditing with deduplication", 2013.

Pyla.naresh Pursuing M.Tech (CSE) From Avanthi Institute Of Engineering And Technology, Vizianagaram, A.P. His area of interest includes Cloud Computing and Network Security.

K.RAVINDRA, M.Tech(Ph.D) From KL University ,Presently working as Associate Professor, Department of CSE Avanthi Institute of Engineering &Technology, Vizianagaram, AP, India. He published several International Papers which shows his zeal towards research.

He believes in the wordings of **"A.P.J Abdul Kalam"**:
"All of us do not have equal talent. But , all of us have an equal opportunity to develop our talents".

Dr. A. Chandra Sekhar, Professor & HOD of Department of Computer Science And Engineering Avanthi Institute of Engineering &Technology, Vizianagaram, AP, Affiliated to JNTU Kakinada. He attended several seminars and workshops. He published several International Papers which shows his zeal towards research.

He believes in the wordings of **"Albert Einstein":**
"The True Sign of Intelligence is Not Knowledge But Imagination"