International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

A Survey of Various Authorization & Deduplications in Cloud

Shweta Biswas[#1], Prof. Amit Saxena[*2], Dr. Manish Manoria[#3]

[#]*Department of CSE, Truba Institute of Engineering & Information technology*
Bhopal, India
[1]shwetabiswascs@rediffmail.com
[2]amitsaxena@trubainstitute.ac.in
[3]manishmanoria@trubainstitute.ac.in

***Abstract*— As the rapid data expansion of data most current years, cloud computing has turn out to be one of the furthermost accepted area in IT industry with considers to cost, proficient use of computing possessions for example storage space and system bandwidth in the cloud computing communications. Data deduplication refers to a method that reduces the unnecessary data on the storage and transmitting on the network, and is think about to be one of the most-enabling storage technologies that presents well-organized resource exploitation in the cloud computing. On the additional hand, be appropriating data deduplication acquires protection vulnerabilities in the cloud storage method so that untrusted entities including a cloud server or unauthorized users may break data confidentiality, privacy and integrity on the outsourced data. It is difficult to resolve the difficulty of data defense and confidentiality regarding data de-duplication, but positively essential for presenting amature and established cloud storage service.**

***Index Terms*— Cloud Storage, Data deduplication, Cloud Computing**

I. INTRODUCTION

Cloud Computing (CC) is an important Information Technology (IT) move and an innovative representation of computing over distributed computing resources, for example bandwidth, storage space, processing power, servers, services and functions [1]. Nowadays, this innovative example has become well-liked and is getting several considerations from researchers in the educational and business societies. Cloud computing is a capable tool that cost-effectively allows data outsourcing as an examination using Internet tools with elastic provisioning and usage-based pricing [2]. Many cloud service vendors provide remote data outsourcing and support services by developing storage space and network resources on cloud storage communications. Since they promise to offer users a convenient and efficient storage service that is available anyplace and anytime over the Internet application on cloud storage services such as Drop-Box, Mozy and Memopal are increasing recognition.

As the rapid expansion of data volumes enhances require for data outsourcing on cloud storage services, pay-as-you-use cloud model takes the require for cost-competent storage space, specifically for reducing storage space and network bandwidth overhead, which is directly related to the financial cost savings. With the aim of decrease the operating costs on storage space and network, business cloud storage service suppliers make use of their resources efficiently through data deduplication, which refers to a technique that finds redundant data units transversely customers, removes duplicate copies of them, and make available associations to the continuing data as an alternative of storing the replica. By storing and shifting only a single copy of unnecessary data, deduplication provides savings of both storage space and network bandwidth. Data deduplication technique is considered to be one of the most-impactful storage technologies, and it is estimated that the ratio of applying deduplication will increase steadily among the storage service providers [3]. However, these capable storage space examines bring several demanding plan issues; considerably due to the loss of data manage. These confronts, explicitly data privacy and data reliability, have important control on the safety measures and presentations of the cloud system. Some risk representations take for granted that the cloud facility earner cannot be dependences, and for that reason security exclusives offer a high level security promise for instance storing encrypted data in cloud servers.

Using cloud servers the distant data reviewing service encompasses a set of protocols calculated to verify the intactness of the isolated data exist in cloud storage additional

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

consistently and proficiently, devoid of downloading the complete data. Additionally, the outsourced data is also issue to manage by unpredictable third party cloud providers [4]. The RDA frameworks use spot checking method to authenticate the outsourced data in which a small section of entire data only is necessitated to be right to use by the assessor. This method presents either a probabilistic or deterministic declaration for the data intactness [5]. This idea of fortifying data in cloud storage amenities has attracted many researchers to work in this field with the aim of constructing a trusted control model of cloud storage. In this study, we are interested in providing a secure framework for data security in the cloud storage services such as dropbox, that offer more functionalities such as file sharing and organization in addition cloud storage essential functionalities. Deduplication can be applied to data which are in primary storage space, cloud storage space; backup storage space for replication relocates [6]. The methods from which deduplication technique implement, many kinds of there are but essentially only three kinds are in concern which are as best deduplication procedure category as block level, file level and byte level by the names itself deduplicate procedure effort in that order on that substance.

Even if data deduplication has lot of fore deal, but user with sensitive data are worried about both outsider/insider attacks. So deduplication of data to be required to hold protection and confidentiality. But with conventional encryption unusual clients encrypt data with their own key, which put together similar to data with unusual consumer key makes unusual ciphertext for that data which is incapable for deduplication. The convergent encryption allocates encrypt/decrypt data with convergent key on the data consequently formulates feasible to be appropriate to ensure duplicates. Thus with uploading user's data as ciphertext to cloud resolved security concerns. To check form unauthorized right of entry, proofs of possession protocol can be utilized as confidentiality restriction [7]. Proof of possession; consumer can download the decrypted and acquire exacting data with convergent keys by identifying its rights. Consequently by means of convergent encryption and proof of rights both protection and confidentiality concerns determine. At rest the method can't effort on advantage level field, indicates customer upload file with various set of privileges on its and on the beginning on convergent encryption doesn't provide any deduplication on it [6]. Thus not provision duplicate checkered with dissimilar rights set providing by the data owner.

II. Theoretical Background

Cloud storage service providers take advantage of efficient resource utilization by applying data deduplication techniques over data which are outsourced from customers (i.e., clients). In order to keep security of outsourced data from an untrusted cloud server, several previous approaches [8] planned some ideas to permit data deduplication over the encrypted data. Their explanations are generally supported on alleged convergent encryption, which acquires a hashed value of a file as an encryption key. Using this method, identical ciphertexts are always generated from identical plaintexts in a deterministic manner. Hence, the cloud server can achieve deduplication over encrypted data, without knowing the exact content of the file. From the observation of cryptography, conversely, it is implicit that deterministic encryption, including convergent encryption naturally, is not as secure as randomized one.

Keeping confidentiality of the outsourced data from the untrusted cloud server is not the only protection trouble for the structure use data deduplication. It has been shown that most of storage services by means of client-side deduplication frequently acquire a side channel all the way through which an opposition may possibly learn information about the contents of files of other users [9]. This is since these amenities portion two intrinsic properties; 1) data transmission in a network can be visible to an adversary, and 2) deterministic and small-size data, such as a hashed value of a file, is queried to the cloud server before file uploading. A user who has not access to but curious about some file might mount an online-guessing attack exploiting the side channel. The attack is as follows: an adversary first constructs a search domain comprising possible versions of the target file, and then queries a hash of each version to the server until a deduplication event occurs.

III. Cloud Remote Storages

Remote storage is the method to extend users' disk space without addition more firm disks on local computers. The remote storages usually are services running on other machines (servers) which automatically copies files from user's local disks. In some cases, after copying, the files physical size on user disks are reduced to minimum (a few KBs) while the logical sizes displayed remain the same.

Similarly, cloud remote storages are remote storages which use cloud services. Cloud services copy files from user's machine and store them in logical pools. The physical storages can be distributed, replicated among multiple servers or locations. Cloud services providers are hosting companies who own these physical structures. They also provide

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

services, applications to interact with user's devices and logical structures. Cloud remote storages allow users accessing their stored data from other supported devices as long as they have internet connection. They also enable users to share data with others. Among the most popular cloud remote storages are OneDrive such as SkyDrive, Dropbox and Google Drive. Since the encryption process is deterministic and is found from the data gratified, matching statistics copies will produce the same convergent key and consequently the identical ciphertext. To avoid unauthorized right to use, a protected proof of rights protocol [10] is also required to provide the proof that the client certainly be in possession of the identical file when a duplicate is originated. After the proof, following customers with the identical file will be making available a pointer from the server exclusive of requiring uploading the similar file. A operator can transfer the encrypted file with the baton from the server, which can only be decrypted by the identical data proprietors with their convergent keys. Consequently, convergent encryption permits the cloud to execute deduplication on the ciphertexts and the proof of rights avoids the unauthorized consumer to right of entry the file.

However, since files are stored and distributed over the internet, security problems emerge. One major problem is data exfiltration. By uploading sensitive information to cloud storages, users have to put their trust on the providers to protect those data. Nonetheless, there are many situations that can affect this trust. Cloud providers can peek into data or transfer data to other parties in some circumstance. Cloud servers can be hacked, resulting in exposing user's sensitive data. It also can be the consequence of poor security design decision coming from the provider architects.

IV. Data Deduplication

For the most part in cloud remote storage context, deduplication fulfills one of the most mutual requirement as reducing the overall costs of providing the same service that client could do themselves in their own data centers. Not only the process utilizes storage hardware costs, it also reduces backup and recovery costs while improving network efficiency. Cloud services like Dropbox employs deduplication across multiple user client, as the user base grows, service cost per user decreases subsequently.

From end-users' point of view, one of the most notable advantage of deduplication is the storage service cost. Basically, they pay less for the same amount of storage. For example, Dropbox offers 1TB storage for only $9.99 while an external hard drive would cost more than a few hundred dollars for the same capacity. Another noteworthy gain is the upload time/bandwidth for duplicated files. Large files such as .iso or movies could be uploaded almost instantly if they are duplicated, providing significant boost to user experience. In general, deduplication process involves 4 steps:

1. Split data into small units, (ex: files, blocks).
2. Calculate unique hash value for each unit.
3. Detect if there is any file on the storage having the same hash value.
4. Put reference to the duplicated file.

Among deduplication schemes, the first and the second steps vary the most. Data can be compared at file level where the scheme generates hashes of files and compares. Meanwhile, block/bye level units allow deduplication in a finer grained manner with the cost of more operations. Hashing mechanisms are also varying among schemes but the general idea is to generate a unique, mathematical representation of unit that can be located and compared. Data Deduplication is a method to prevent duplication of repeating data. During deduplication processes, unique chunks of data such as files or block of bytes are analyzed followed by discarding the duplication. The result is only one instance of repeat data are stored/transferred hence greatly improve storage utilization and transfer time/bandwidth. Initially, deduplication is heavily used on duplication prone situations such as backups or virtual desktops [11].

V. Security Issues on Cloud Storage with Data Deduplication

Even though the data deduplication method is reflect to be efficient and valuable in storage systems, there are several challenging issues of statistics security and confidentiality in the cloud storing services where the data deduplication technique is applied. These issues of security and privacy originate from the following facts:

- In the cloud computing environments, cloud servers are usually outside of the trust domain of the data owners (i.e., users). In fact, a wide range of the users are more than willing to put their data outsourcing task to a cloud storage provider.
- Cloud storage services are typically based on multi-tenant architecture, where there is no trust relationship among users. Chasing efficiency in terms of utilizing resources such as the storage space and the network bandwidth leads to applying client-side data deduplication across multiple (untrusted) users.

In the cloud stowage scheme with statistics deduplication, untrusted entities including a cloud server and users may

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

cause security threats to the storage system. By exploiting some vulnerabilities in data deduplication, both an inside adversary, who act as a cloud server, and an outside adversary, who act as a user, will attempt to break data confidentiality, privacy and integrity on the outsourced data. More concretely, for cloud storage system with deduplication, we are concerned with several security issues that are raised by the adversaries: 1) sacrificing data security for deduplication, 2) evidence leakage finished side channel, and 3) unauthorized arbitrary data access.

Sacrificing Data Confidentiality for Deduplication: Cloud storage service providers that are using deduplication, however, are typically reluctant to apply encryption on the stowed statistics, since encrypting on statistics obstructs executing statistics deduplication [9][12]. They may be unlikely to stop using deduplication due to the high cost savings offered by the technique. This eventually incurs the loss of confidentiality for the data stored on the cloud storage.

Information Leakage through Side Channel: In a storage system using client-side data deduplication, losing confidentiality of the outsourced data is not the only security problem. As shown in [9], client-side deduplication commonly incurs a side channel through which a malicious user (i.e., an adversary) may get sensitive information about the other user's data. The side channel is caused by two inherent properties; 1) data transmission over the network is visible to an adversary, and 2) a small-sized hashed value of a file is used to determine the existence of the same file on the cloud server.

Using the side channel, an opponent can easily classify the existence of a file by uploading the file and monitoring its network activity. If the whole file is not transmitted over the network, the adversary learns that the file already exists on the server. Furthermore, the adversary is even able to learn the content of the file by mounting an online-guessing attack. That is, an adversary trying to figure out the content of the file will build a dictionary that consists of guessed versions of that file and repeat uploading each guess to the server until finding out that a deduplication event occurs.

Unauthorized Arbitrary Data Access: Besides an information leakage through the side channel, a cloud storage system that applies the client-side deduplication technique is also vulnerable to another type of attack [10]. This new security threat creates from the datum that client-side deduplication systems typically adopt a hashing based strategy, in which a small-sized hash value is calculated for each file (or block) and is used to find duplicate copies of the same file. In such a storage system, by accepting the hash value for the file, the cloud server allows anyone who possesses the hash value to download the entire file.

An adversary, who does not have a whole file except its hash value, will exploit this vulnerability to get the ownership of that file. The adversary can convince the cloud storage server that it owns that file by presenting just the hash value, hence can download the entire file from the server.

VI. Literature Survey

In this paper author, Harnik et al. [9] aimed at weakening the correlation between deduplication and the existence of files in the storage, and proposed a randomized approach that obfuscates the occurrence of deduplication. Given a security parameter d, a threshold value is randomly generated and assigned for each file in the cloud storage. When a client uploads a new copy of the file, the cloud server checks whether the accumulated number of file uploads exceeds the threshold. If the number is at slightest as high as the verge, the server performs data deduplication at client side and the copy of that file is not sent over the network. Otherwise, deduplication is performed at server side, in which the actual file content is sent to the server.

Li et al. [13] observed that secure deduplication schemes based on convergent encryption suffer from managing an enormous number of convergent keys, since both encrypted data chunks and their convergent keys should be handled together. To address this key management problem, they proposed Dekey, which is an efficient and dependable key organization scheme in secure deduplication. This construction makes cloud storage users to distribute their convergent key stocks crossways multiple mist storing waitpersons instead of managing the keys on their own.

DuPLESS [14] is a real deduplication system, which is built on commercial cloud storage services, that provides security against brute-force attacks launched by malicious clients or an untrusted server. In order to achieve the desired goals while satisfying the required security, an Oblivious Pseudo Random Function (OPRF) protocol and message-locked encryption (MLE) [15] were utilized for their construction. OPRF is a randomized protocol between clients and the key server, which ensures that the key server learns nothing about the inputs and the resulting outputs, and the clients learn nothing about the key. MLE is a generalized version of convergent encryption designed.

In this paper author Jin Li,et.al [16] give explanation for deduplication to save from harm the privacy of susceptible data while sustaining deduplication, the convergent encryption method has been offered to encrypt the data earlier than

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

outsourcing. To improved defend data protection; this paper creates the initial effort to properly concentrate on the difficulty of authorized data deduplication. Different from conventional deduplication methods, the discrepancy benefits of customers are additional well thought-out in duplicate check alongside the data itself. They also present quite a lot of novel deduplication buildings secondary official matching checkered in hybrid cloud construction. As a resistant of notion, they implement a prototype of our proposed authorized duplicate verify method and accomplish test bed experiments. They show that their proposed authorized duplicate confirm method acquires smallest operating cost evaluated to standard

S. No.	Paper	Author	Advantages	Issues
1.	Secure De-duplication And Data Security With Efficient And Reliable CEKM[17]	N.O.AGRAWAL, S.S.KULKARNI	Add security features insider attacker on De-duplication and outsider attacker by using the detection of masquerade activity by risk-averse attackers.	The problem of injury of pinched data if we diminution the value of that pinched evidence to the aggressor to achieving effectual and consistent key organization in protected de-duplication.
2.	A Study on Authorized De-duplication Techniques in Cloud Computing.[18]	Bhushan Choudhary, Amit Dravid	Security by counting differential benefits of clients in the duplicate copy check. Some endangered primitives applied as a part of our harmless de-duplication i.e. Symmetric encryption, Convergent Encryption, Proof of Ownership, Identification Protocol.	The security issue is to appraise the effectual operation of cloud band width and disk usage.
3.	Private Data De-duplication Protocols in Cloud Storage.[19]	Wee Keong Ng, Yonggang Wen, Huafei Zhu	It formalized in the context of two-party computations using private data de-duplication protocol. Algorithm is best for de-duplication protocol for private data storage.	how to define the security of private data de-duplication protocols how to formalize the functionality of private data de-duplication protocols, and how to construct private data de-duplication protocols if exist
4.	A Tunable Proof of Ownership Scheme for Deduplication Using Bloom Filters.[20]	Jorge Blasco, Agustin Orla	Their computational competence in expressions of bandwidth and I/O, for both legal clients and the server. In addition, PoW methods should not entail the server to load the large portion from its back-end storage at each execution of PoW.	The main issue is cause root of these risks lies in the precision that proof of ownership only relies on the knowledge of a static, small portion of information.
5.	A Secured and Authorized Data Deduplication in Hybrid Cloud with Public Auditing.[21]	Sharma Bharat, Mandre B.R.	Specific clients are only allowable to achieve the matching check and storage provider for distinct files with the equivalent privileges to access.	Issue is that duplicate check do not support differential privileges from convergent encryption even though provide confidentiality.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 1, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-109 PAGE NO: 630-635.

process.

VII. Conclusion

In this paper we can review common uses of cloud computing have brought massive simplicity for data sharing and collection. One significant challenge of cloud storage space services is to deal with the rising volume of data content accumulated. Even though data deduplication gives numerous advantages, beside with protection and confidentiality concerns happen as customer's susceptible data are vulnerable to both inside and outside attacks remove matching duplicates of storage space statistics and boundary the harm of stolen data if we reduce the importance of that stolen information to the attacker.

References

[1] Peter Mell and Timothy Grance. The NIST definition of cloud computing (draft). NIST special publication 800 (2011), 145.

[2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al., "Above the clouds: A berkeley view of cloud computing," Technical Report UCB/EECS-2009-28, Dept. EECS, UC Berkerely, 2009.

[3] D. Russell, "Data deduplication will be even bigger in 2010," Gartner, 2010.

[4] Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, Osama Khan, Lea Kissner, Zachary Peterson, and Dawn Song. 2011. Remote data checking using provable data possession. ACM Trans. Inf. Syst. Secur. 14, 1 (2011), 1–34.

[5] Bo Chen, Reza Curtmola, Giuseppe Ateniese, and Randal Burns. 2010. Remote data checking for network coding-based distributed storage systems, 2010.

[6] Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou Jin Li, "A Hybrid Cloud Approach for Secure Authorized," IEEE Transactions on Parallel and Distributed Systems, vol. pp, pp. 1-12, 2014.

[7] D. Harnik, B. Pinkas, and A. Shulman-Peleg. S. Halevi, "Proofs of ownership in remote storage systems." ACM Conference on Computer and Communications Security, pp. 491-500, 2011.

[8] L. Marques and C. J. Costa, "Secure deduplication on mobile devices," in Proc. Workshop on Open Source and Design of Communication (OSDOS'11), pp. 19-26, 2011.

[9] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," IEEE Security and Privacy Magazine, vol. 8, pp. 40-47, 2010.

[10] 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.

[11] Meyer, D. T., and Bolosky, W. J. A study of practical deduplication. ACM Transactions on Storage (TOS), 2012.

[12] M. Mulazzani, S. Schrittwieser, M. Leithner, M. Huber, and E. Weippl, "Dark clouds on the horizon: using cloud storage as attack vector and online slack space," in Proc. USENIX Security Symposium (SEC'11), 2011.

[13] L. Jin, C. Xiaofeng, L. Mingqiang, L. Jingwei, L. Patrick, and L. Wenjing, "Secure deduplication with efficient and reliable convergent key management," IEEE Transactions on Parallel and Distributed Systems, vol. PP, 2013.

[14] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server-aided encryption for deduplicated storage," in Proc. USENIX Security Symposium (SEC'13), pp. 179-194, and 2013.

[15] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," Advances in Cryptology - EUROCRYPT'13, LNCS 7881, pp. 296-312, 2013.

[16] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou" A Hybrid Cloud Approach for Secure Authorized Deduplication" IEEE Transactions On Parallel And Distributed System VOL: PP NO:99 YEAR 2013.

[17] N.O.AGRAWAL1, S.S.KULKARNI, "Secure Deduplication and Data Security with Efficient and Reliable CEKM" International Journal of Application or Innovation in Engineering & Management (IJAIEM), November 2014.

[18] Bhushan Choudhary, Amit Dravid , "A Study on Authorized Deduplication Techniques in Cloud Computing" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3, Issue 12, April 2014.

[19] Wee Keong Ng, Yonggang Wen, Huafei Zhu, "Private Data De-duplication Protocols in Cloud Storage" ACM 978-1-4503-0857-1/12/03, 2011.

[20] Jorge Blasco, Agustin Orfila, "A Tunable Proof of Ownership Scheme for Deduplication Using Bloom Filters" June 18, 2014.

[21] Sharma Bharat, Mandre B.R. "A Secured and Authorized Data Deduplication in Hybrid Cloud with Public Auditing" International Journal of Computer Applications (0975 – 8887) Volume 120 – No.16, June 2015

[22] John R. Douceur, Atul Adya, William J. Bolosky, Dan Simon, Marvin Theimer, "Reclaiming Space from Duplicate Files in a Serverless Distributed File System" July 2002.