DOI: http://dx.doi.org/10.26483/ijarcs.v9i3.6006

ISSN No. 0976-5697

Volume 9, No. 3, May-June 2018

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

DIFFERENT SECURE DATA DEDUPLICATION APPROACHES FOR CLOUD STORAGE: A REVIEW

Akanksha Upadhyay
PG Scholar,
Computer Science & Engineering,
RITS, India

Abha Sharma
Asst. Prof., Computer Science & Engineering,
Computer Science & Engineering,
RITS, India

Chetan Agrawal
HOD, Computer Science & Engineering,
RITS, India

Abstract–With the fast growth of the current era, world are moving to digital storage for accomplishment purposes, there is a rising demand for systems that can deliver safe data storage in effective manner. As more corporate and private users outsource their data to cloud storage providers, new data breach incidents make end-to-end encryption progressively prominent obligation. With the growing awareness of data privacy, more and more cloud users choose to encrypt their sensitive data before outsourcing them to the cloud storage. Data deduplication is single instance data storage widely used in cloud storage system to reduce space and upload bandwidth. Duplication-less storage system which de-duplicates the data using file level deduplication and block level deduplication. In this paper we gives an overview of the state of data Deduplicationapproachesand describes the various application to produce specification according need of current generation.

Keywords: Data Deduplication, Security, Cloud Storage, Data Encryption, Convergent Encryption;

I. INTRODUCTION

The data is growing every day the volume and varsity of data have raised a critical and increasing requirement for data storage space and confidentiality of private data in computing environment. Cloud computing is getting increasingly popular because it can provide low-cost and on-demand use of vast storage and processing resources. With the rapidly increasing amounts of data produced worldwide, networked and multi-user storage systems are becoming very popular. However, concerns over data security still prevent many users from migrating data to remote storage. The conventional solution is to encrypt the data before it leaves the owner's premises. Cloud storage providers constantly look for techniques aimed to minimize redundant data and maximize space savings. We focus on deduplication, which is one of the most popular techniques and has been adopted by many major providers [1] [2].

From a user's perspective, data outsourcing raises security and privacy concerns. We must trust third-party cloud providers to appropriately impose confidentiality, integrity checking, and access control mechanisms against any insider and outsider attacks. Authorized data deduplication objectives at information security to keep the information secured and maintain a strategic distance from unapproved get to. Deduplication at encryption level saves a lot of memory and memory can be used proficiently. In particular, conventional encryption requires diverse clients to scramble their information with their own keys. In this way, indistinguishable information duplicates of various clients will prompt diverse cipher texts, making deduplicationimpossible [3].

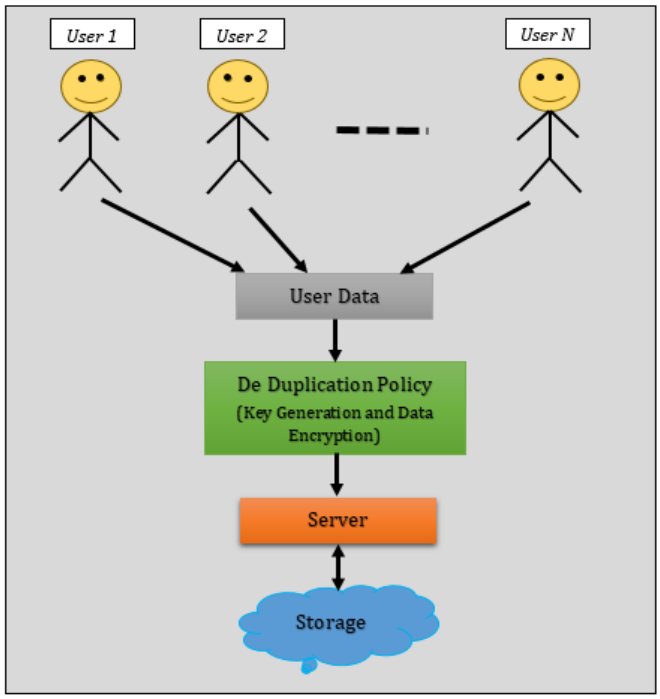

Figure 1: General Demonstration of Data Deduplication

The above diagram depicts the general process of de duplication scenario for a single or multiple users' data security over the cloud storage.

For making the feasible deduplication and keep up the information privacy utilized focalized encryption method. It scrambles unscrambles an information duplicate with a focalized key, the substance of the information duplicate got by figuring the cryptographic hash estimation of. After the

information encryption and key age process clients hold the keys and send the cipher text to the cloud. Since the encryption activity is determinative and is gotten from the information content, comparable information duplicates will produce the same united key and thus the same cipher text [4]. Presently a day's cloud turns into an appealing pattern toward the general population. Cloud gives the space to client to store information in virtualized pool stockpiling which might be available wherever you need with the assistance of web proficient gadget. This transportability and also the versatile administrations gave by the cloud influence that people groups liked to store their own information on distributed storage. Cloud additionally advantageous for cost sparing .It gives the asset sharing element.

The main objective of this paper is: To present various aspect of data Deduplication for securely store data on cloud storage their applicability; to review some late and existing procedures of data deduplication using convergent key encryption combined cryptography techniques; to give generalize problem formulation in the aspects of available de duplication issue.

Rest of the paper is organized as follow: in section II background study of data Deduplication system is done, in section III previous researches done in data Deduplication is listed, Comparison of prior work on Secure Data Deduplication is described in section IV, section V describe problem statement section VI is the solution domain is short. Lastly in section VII summarize the brief summary of the paper i.e. conclusion.

II. BACKGROUND

The background of a study is an important part of our survey paper. It provides the context and purpose of the study. Hence there is need for background study that contributes to prepare different aspects of the data Deduplicationsystem.

A. Cloud Data Storage

In the most recent decade, the request of outsourcing data is significantly expanded. Data storage and superior performance are the primary needs which must be satisfied. These services are given by numerous distributed computing specialist organizations like Drop box, Google App Engine, Amazon Simple Storage Service (AmazonS3), and so on. The upside of putting away information in cloud servers is that the information proprietors can diminish the overhead of purchasing additional solid servers and furthermore abstain from employing of server administration engineers. Distributed storage is these days extremely well known stockpiling framework. Distributed storage is putting away of information off-site to the physical stockpiling which is kept up by outsider. By utilizing distributed storage we can get to data from any PC over web which lost confinement of getting to data from indistinguishable PC where it is stored [5].

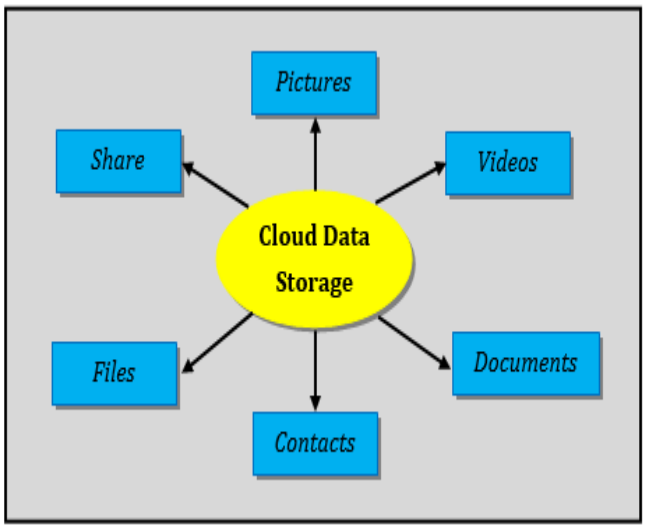

Figure 2: Cloud Storage

B. Advantage and Disadvantage of Cloud Storage

Most of the services are free up to certain number of gigabytes, and storage also. All the cloud provider provides all the features to the end user like drag and drop, syncing files and folder in your desktop, mobile device and soon [6] [7].

Advantages:

- ✓ *Usability* – the cloud provider are always usable at any time.
- ✓ *Bandwidth* – avoid of sending the files to individual instead of send a web link to the end user through email it.
- ✓ *Accessibility* – stored files and folder are accessible from any were in cloud platform

Disadvantages:

- ✓ *Usability* – Be carefully of using the drag and drop option which is used in cloud storage that will permanently remove your file. While using drag and drop option used the copy and paste option to it which will save your files from the permanent delete.
- ✓ *Bandwidth* – limited bandwidth allowed in the cloud storage. If you want more bandwidth then it should be payable. However, some providers have unlimited bandwidth on it.
- ✓ *Accessibility* – to access the data and files you need an internet connection. With-out it not possible.

C. Data Deduplication

Data Deduplicationrefers to a technique for taking out repetitive data in an informational collection. In the strategy for Deduplication, additionally duplicates of comparable information are erased; deed just a single duplicate to be kept. Information is analyzed to perceive reproduction byte examples to ensure the single occasion is extremely the single document. At that point, copies are traded with a reference that focuses to the keep chunks.

Data Deduplicationis a technique to decrease storage space. By recognizing redundant data by means of hash values to

relate data chunks, keeping only one copy, and making logical pointers to other copies in its place of storing other actual copies of the redundant data. Deduplicationdecreases data volume so disk space and network bandwidth can be reduced which decrease costs and energy depletion for successively storage systems [8] [9]. Figure 3 shows the example of de duplication.

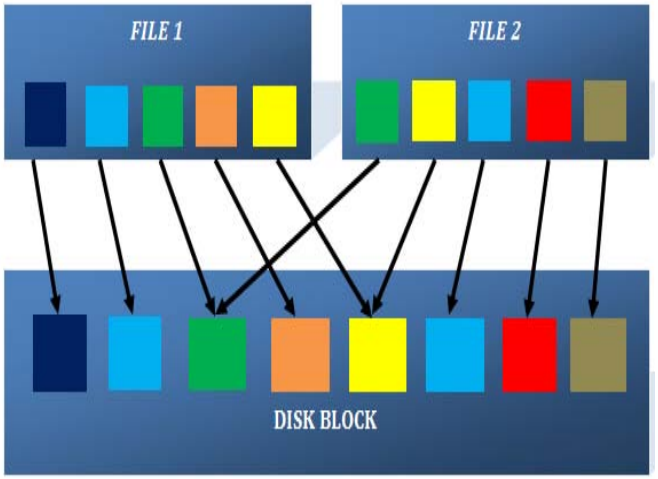

Figure 3: Example of data Deduplication

D. Applications of Data Deduplication

Data Deduplicationprovides practical ways to achieve these goals, including [10]:

- *Capacity optimization.* Data Deduplicationstoresextra data in less physical space. It attains greater storage effectiveness than was possible by using features such as Single Instance Storage (SIS) or NTFS compression.
- *Scale and performance*. Data Deduplicationisextremely scalable, resource well-organized, and nonintrusive.
- *Reliability and data integrity*. When data Deduplicationis applied, the integrity of the data is preserved. Data Deduplicationuses checksum, steadiness, and uniqueness validation to confirm data integrity.
- *Bandwidth efficiency with Branch Cache.* Through integration with Branch Cache, the same optimization techniques are applied to data transferred over the WAN to a branch office.
- *Optimization management with familiar tools.* Data Deduplicationhas optimization functionality made into Server Manager and Windows Power Shell.

III. LITERATURE SURVEY

This section provides the recently made contribution and the research work performed for improving software quality by fixing bug by prediction technique. Thus different research articles and papers are included in this section.

Jin Li et al. [11] makes the first effort to formally address the issue of accomplishing effective and reliable key administration in secure deduplication. Creators initially present a standard technique in which every client holds a free ace key for scrambling the joined keys and outsourcing them to the cloud. By and by, such a key administration framework produces tremendous number of keys with the developing number of clients and necessities clients to relentlessly shield the ace keys. To this end, creators propose Dekey, another working in which clients don't have to finish any keys individually however in its place safely allot the merged key offers transversely various servers. Security examination exhibits that Dekey is secure as far as the definitions evaluated in the proposed security demonstrate. As a proof of idea, creators actualize Dekey utilizing the Ramp mystery sharing plan and show that Dekey endures limited overhead in precise conditions.

To better protect data security, **Jin Li et al. [12]** makes the primary attempts to formally address the issue of endorsed data deduplication. Not exactly the same as standard deduplication structures, the differential advantages of customers are moreover considered in duplicate check other than the data itself. We furthermore demonstrate a couple of new deduplication advancements supporting affirmed duplicate check in a hybrid cloud designing. Security examination demonstrates that our arrangement is secure the extent that the definitions showed in the proposed security show. As a proof of thought, makers execute a model of this proposed endorsed duplicate check design and lead testbed tests utilizing our model. Creators demonstrate that our proposed approved copy check conspire causes insignificant overhead compared with typical operations.

Pasquale Puzio et al. [13] proposed ClouDedup, a safe and effective storage services which guarantees block level deduplication and information security in the meantime. Based on convergent encryption, ClouDedup remains secure on account of the meaning of a part that actualizes an extra encryption task and an entrance control instrument. Besides, as the prerequisite for deduplication at square level raises an issue as for key administration, we recommend incorporating another segment keeping in mind the end goal to execute the key administration for each piece together with the real deduplication task. We demonstrate that the overhead presented by these new segments is insignificant and does not affect the general stockpiling and computational expenses.

Encrypting data on client-side before uploading it to cloud storage is essential for protecting users' privacy. However client-side encryption is at odds with the standard practice of deduplication. Reconciling client-side encryption with cross-user deduplication is an active research topic. **Jian Liu et al. [14]** present the first secure cross-user deduplication scheme that supports client-side encryption without requiring any additional independent servers. Interestingly, the scheme is based on using a PAKE (password authenticated key exchange) protocol. We demonstrate that our scheme provides better security guarantees than previous efforts. Authors show both the effectiveness and the efficiency of our scheme, via simulations using realistic datasets and an implementation.

Expecting to address the security challenges, **Jin Li et al. [15]** makes the principal endeavors to formalize circulated tried and true deduplication structure. Creators propose new scattered deduplication structures with higher reliability in which the data pieces are coursed over different cloud servers. The security necessities of data arrangement and mark consistency are similarly proficient by exhibiting a deterministic secret sharing arrangement in scattered limit structures, as opposed to

using combined encryption as in past deduplication systems. Security examination displays that our deduplication structures are secure the extent that the definitions demonstrated in the proposed security appear. As a proof of thought, we complete the proposed systems and show that the achieved overhead is extremely constrained in practical situations.

There are many techniques which is used for eliminating duplicate copies of rehashing data .From that one of the essential procedure is information deduplication. Information Deduplicationis particular information pressure procedures for expelling copy duplicates of rehashing information and has been generally utilized as a part of distributed storage to diminish the measure of storage room and spare data transmission. To ensure the secrecy of touchy information on cloud, the united encryption strategy is utilized to scramble the information before outsourcing. **VaishaliKeshavDhokne et al. [16]** proposed framework in which emit sharing plan is utilized. In that records are separated into number of squares called as offers and this offers store on the diverse number of hubs. By utilizing recoup procedure number of offers isjoining into single record. That implies utilizing this plan give data security, confidentiality, and data reliability.

IV. COMPARATIVE STUDY

In order to optimize upload bandwidth and storage space over server, Deduplication is one of the best options. Depending upon what type of deduplication approach we used for secure data and redundancy. Therefore, following table will possess the prior work on secure data Deduplicationwith different techniques.

Table 1: Comparison of prior work on Secure Data Deduplication

First author et. al. [ref no]	Method/ scheme	Advantage	Limitation
Jin Li [11]	Dekey	Dekey incurs limited overhead in realistic environments. Users do not need to manage any keys on their own	The first problem is the enormous storage overhead in key management. Master key presents the single-point-of failure and needs to be securely and reliably maintained by the user.
Jin Li [12]	Authorized Duplicate Check	The token generation introduces only minimal overhead in the entire upload process. secure in terms of insider and outsider attacks	Maintain differential authorization on every key generation process which is incurred long time consuming to process convergent key encryption.
PasqualePuzio [13]	ClouDedup	No need of additional independent server for client side encryption.	Limited to file level deduplication and is not scalable in the case of block level deduplication, which achieves higher space savings.
JianLiu [14]	Secure Cross-User Deduplication	Uses a per-file rate limiting strategy to prevent online brute-force attacks	Block-leveldeduplication and data confidentiality is dependent on only convergent key encryption.
Jin Li [15]	Distributed Deduplication Systems	provide higher reliability as well as data chunks are distributed across multiple cloud servers	The work is limited for tag consistency i.e. tag consistency can be used for existing problem. Tag generation taking higher time for tag generation process.

V. PROBLEM IDENTIFICATION

Storage efficiency functions like compression and deduplication give stockpiling providers higher usage of their stockpiling back-closes and the ability to serve a considerable measure of clients with consistent framework. Information deduplication is the technique by that a capacity provider exclusively stores one duplicate of a record close by a large number of its clients. There are four totally unique deduplication routes, contingent upon whether deduplication occurs at the purchaser side (i.e. before the transfer) or at the server side, and whether deduplication occurs at a square level or at a document level. Deduplication is most beneficial once it is activated at the purchaser side, since it additionally spares transfer data transmission. Consequently, deduplication is a fundamental empowering influence for assortment of prominent and successful stockpiling administrations (e.g. Dropbox, Memopal) that offer ease, remote stockpiling to the expansive open by action customer side deduplication, subsequently sparing both the system data transmission and capacity costs. Without a doubt, information deduplication is ostensibly one among the most reasons why the expenses for

distributed storage and cloud storage and cloud backup services have dropped so sharp.

Data deduplication is the process by which a storage provider just stores a solitary duplicate of a record possessed by a few of its clients. There are four distinctive deduplication procedures, contingent upon whether deduplication occurs at the customer side (i.e. before the transfer) or at the server side, and whether deduplication occurs at a square level or at a record level. Deduplication is most remunerating when it is activated at the customer side, as it additionally spares transfer data transmission. Hence, deduplication is a genuine empowering agent for various across the board and productive stockpiling administrations that offer low-estimated, inaccessible capacity to the wide open by acting customer side deduplication, in this way sparing both the system data transmission and capacity costs. For sure, information deduplication is seemingly one of the fundamental reasons why the costs for distributed storage and cloud reinforcement administrations have dropped so forcefully.

As an outcome, storage systems are relied upon to experience major rebuilding to keep up the present circle/client proportion within the sight of end-to-end encryption. The outline of capacity proficiency works when all is said in done and of deduplication works specifically that don't lose their viability in nearness of end-to-end security is in this way still an open issue.

VI. SOLUTION DOMAIN

The proposed system is designed to provide a solution for the drawbacks of general de duplication systems. This proposed system reduces the redundant data to a maximum extent and also provides facility to effectively handle sensitive data which is shared among users. It also provides a facility to check the existence of duplicated data before uploading to the cloud storage server the proposed system aims to:

- ✓ Replace manual Deduplicationsystem with an automated one.
- ✓ Enhance the storage space utilization
- ✓ Reduce the chances of attacks associated with shared sensitive information
- ✓ Reduce the workload involved in processing
- ✓ Prior duplicate checks before uploading the data
- ✓ Access through different privileges

VII.CONCLUSION

With the information and network technology, rapid development, rapid increase in the size of the data center energy consumption in IT spending in the increasing proportion of data deduplication to optimize storage system can greatly reduce the amount of data, thereby reducing energy consumption and reduce heat emissions. Data security has consistently been a major issue in information technology. As digital data is growing tremendously, cloud storage services are gaining popularity since they promise to provide convenient and efficient storage services that can be accessed anytime, from anywhere. At the same time, with the advent of cloud computing and its digital storage services, the growth of digital content has become irrepressible at both the enterprise and individual levels. This paper deal with different terminology of data deduplication paradigm. Additionally, in this paper we described various data deduplication techniques and applications which is solely used in real time environment.

We are proposing data security approach for cloud data storage which is ensures to secure data Deduplication using different privilege of the end user. For the prospect of the data security along with the improve efficiency of storage space we are implementing a system by means of user can prevent their sensitive data over the storage. Our Implementation should be fully based on cryptographic approach. After implementation, in near future the proposed work considering following directions are:

- ✓ This scheme can be further enhanced by uploading very large data and can do compression on those data. Also, enhance to measure the Quality of Service (QoS
- ✓ We also plan to implement a privacy preserving public auditing of the uploaded data in the cloud such that the third party auditor (TPA) should not learn about the content of the data.

REFERENCES

[1] DiptiBansode and Amar Buchade, “Study On Secure Data Deduplication System with Application Awareness Over Cloud Storage Systems”, International Journal of Advanced Computational Engineering and Networking, Volume-3, Issue-1, Jan.-2015.

[2] AshwiniShete and B. M. Patil, “Deduplication in Hybrid Cloud with Secure Data”, International Journal of Computer Applications (IJCA), Volume 148 – No.8, August 2016.

[3] MadhuriKavade and A.C. Lomte, “Secure Deduplication using Convergent Keys (Convergent Cryptography) for Cloud Storage”, International Journal of Computer Applications (IJCA), Volume 126 – No.10, September 2015.

[4] AparnaAjitPatil and DhanashreeKulkarni, “Secure Data Deduplication on Hybrid Cloud Storage Architecture”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 5, May 2015, pp. 2906-2909.

[5] Moritz Borgmann and Tobias Hahn, “On the Security of Cloud Storage Services”, SIT Technical Reports, March 2012.

[6] Zeng, Wenying, et al. "Research on cloud storage architecture and key technologies." Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. ACM, 2009.

[7] Leesakul, Waraporn, Paul Townend, and JieXu. "Dynamic data deduplication in cloud storage." Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on. IEEE, 2014.

[8] RiddhiMovaliya and Harshal Shah, “A Survey of Secure Data Deduplication”, International Journal of Computer Applications (IJCA), Volume 138 – No.11, March 2016

[9] Deepa .D and Revathi .M, “A Survey on Deduplication Scheme in Cloud Storage”, International Journal of Science and Research (IJSR), Volume 3 Issue 12, December 2014.

[10] Rashid, Fatema. "Secure Data Deduplication in Cloud Environments." PhD dissertation. PhD thesis, Ryerson University, Toronto, Ontario, Canada, 2015.

[11] Li, Jin, Xiaofeng Chen, Mingqiang Li, Jingwei Li, Patrick PC Lee, and Wenjing Lou. "Secure deduplication with efficient and reliable convergent key management." IEEE transactions on parallel and distributed systems 25, no. 6 (2014): 1615-1625.

[12] Li, Jin, Yan Kit Li, Xiaofeng Chen, Patrick PC Lee, and Wenjing Lou. "A hybrid cloud approach for secure authorized deduplication." IEEE Transactions on Parallel and Distributed Systems 26, no. 5 (2015): 1206-1216.

[13] Puzio, Pasquale, RefikMolva, MelekOnen, and Sergio Loureiro. "ClouDedup: secure deduplication with encrypted data for cloud storage." In Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, vol. 1, pp. 363-370. IEEE, 2013.

[14] Liu, Jian, N. Asokan, and Benny Pinkas. "Secure deduplication of encrypted data without additional independent servers." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 874-885. ACM, 2015.

[15] Li, Jin, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Mehedi Hassan, and AbdulhameedAlelaiwi. "Secure distributed deduplication systems with improved reliability." IEEE Transactions on Computers 64, no. 12 (2015): 3569-3579.

[16] Dhokne, MsVaishaliKeshav, and VarshaPatil. "Secure Data Deduplication System with Tag Consistency." IJETT 1, no. 2 (2017).

