Deduplication in Cloud Storage

Pronika, S.S.Tyagi

Abstract: ***Cloud Computing is well known today on account of enormous measure of data storage and quick access of information over the system. It gives an individual client boundless extra space, accessibility and openness of information whenever at anyplace. Cloud service provider can boost information storage by incorporating data deduplication into cloud storage, despite the fact that information deduplication removes excess information and reproduced information happens in cloud environment. This paper presents a literature survey alongside different deduplication procedures that have been based on cloud information storage. To all the more likely guarantee secure deduplication in cloud, this paper examines file level data deduplication and block level data deduplication.***

Keywords: Cloud Computing, Data Deduplication, Data Storage, Security.

I. INTRODUCTION

With the help of Cloud the performance of storing data in computer will be increase day by day and access of data will be fast. When user stored the large amount of data in cloud it faces some challenges like data loss, infected application, data theft, privacy issue, data location, security at user level and data duplication. When data was in reading or writing mode and transfer from one network to another network on cloud then these issues damage the routine of system.

So as to upgrade the consistency and accessibility and give disaster retrieval, information is by and large copied on various storage areas. The majority of this copied information applies an additional load on the capacity framework as far as extra space and transmission capacity to move the copied information on the system. Efficient information storage is serious and information deduplication system is considered as an empowering innovation for active storage of huge information. Deduplication method is a unique information compression strategy to remove the excess information and decrease transmission rate and loading space in the distributed storage frameworks [1]. This method find out the copy information, spare just one duplicate of the information and deliberately utilize consistent pointers for copied information. Deduplication tends to the developing interest for storage size. Many distributed storage suppliers like Microsoft Azure, Bitcasa, Amazon S3 and backup administrations, for example, Memopal and Dropbox are utilizing information deduplication methods to improve capacity efficiency.

We study data or information deduplication is primary significant issue in huge capacity on cloud. The deduplication methods are information type specific, and various systems are utilized on various kinds of information, for example video, picture and text information. All three sorts of information have distinctive capacity groups and qualities. In view of sort of information, deduplication strategies have various procedures to find and reduce copy data [2]. In this way, kind of information is significant for the advancement of deduplication methods. The arrangement of data is serious for finding, reading and coordinating the data. The strategies to checked duplication in video, picture and content have various procedures because of different pattern of information.

The lowest number of information repetition called repetition element or kept up in a big dispersed loading framework to accomplish high information accessibility. Any copy information above repetition element is expelled to shrink storage prerequisite, cost of storage, calculation and energy [2]. Because of this significant profit to industry, deduplication procedures for a huge circulated storage framework picked up energy in the industry and academia. In any case, these strategies are facing difficulty in efficacy and efficiency of information coordinating methods. Many company providers provides cloud founded storage like Google Drive, Dropbox and Mozy that can protect currency on capacity costs with the help of deduplication. For example, when some users send the same type of data or record then supervisor knows it and stores only a unique data or information. This research determines some latest deduplication improvements on approved duplication check in a hybrid, private and public cloud storage. The calculation of rate of data deduplication can be found from the portion of data deduplication expression [3].

The data duplication level as:

$$\frac{\text{Planned Storage Size-Actual Storage Size}}{\text{Planned Storage Size}} \quad \text{X} \quad 100$$

Where planned storage size is the addition of all data or information to standby and the storage size which is actually taken because of deduplication. For example, if user want to save a combined data from different machines X and Y is 100MB and 60MB is required for actual storage, in that case data duplication rate for X and Y is 40%.

Revised Manuscript Received on December 15, 2019.

***** Correspondence Author

Pronika*, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad(Haryana), India. Email: pronika.fet@mriu.edu.in

S. S. Tyagi, Department of Computer Science & Engineering, Manav Rachna International Institute of Research & Studies, Faridabad(Haryana), India. Email: shyam.fet@mriu.edu.in

Data 1	Data 2	Deduplication	Data 1	Empty	
Data 3	Data1		Data 2	Empty	
Data 1	Data 3		Data 3	Empty	

Figure 1 Data Deduplication Process

Figure 1 shows the data deduplication of unique section of data that reduce the duplicate section of data or information [2]. The complete data is divided into different segments or sections like variable and fixed size segments. Due to the process of deduplication the unique copy of every segment is saved and indicators are utilized for identical data of every segment.

II. EVOLUTION OF DEDUPLICATION

In start of 1950, data reduction methods were introduced and it can be divided into two types such as lossy and lossless data reduction methods. After that in 1990, compression techniques like delta compression came; its objective is to compress the similar chunks or similar files. In early 2000s, data deduplication techniques came to help the huge storage infrastructure. It removes both intra-file and inter-file level repetition over huge datasets across many dispersed storage servers. As in case of traditional technique it reduces repetition over little set of files this was based only on intra-file repetition. With the help of cryptographic hash of every file or its chunk find out the duplication of the data [2]. In 2008, these methods applied on multimedia matters and find out the duplicate content in multimedia by evaluating the same type of frames or images using hashing and extraction methods.

These information deduplication methods appeared to untie the issue identified with expanding information size in the capacity frameworks. Some coding methods like Dictionary and Huffman are the conventional compression systems that effort at string or byte level, while deduplication strategies remove the repetition at block or file level. Repetitive information reduction strategies appeared in 1950s, which were to a great extent lossless and lossy information compression procedures, trailed by delta pressure in 1990s.

III.DATA DEDUPLICATION TECHNIQUES

In 2000, data deduplication methods was first introduced to sustain overall compression at coarse granularity. The past methodologies are very inefficient to identity relative pieces and are not versatile, while information deduplication strategies can be functional at file level or sub file like block level. It wraps information by utilizing variable or fixed size pieces. The hash estimations of these pieces are produced utilizing cryptographic hash measurements, and copies are identified by coordinating hash standards. Data deduplication techniques classified into sub file and file level.

A. File-level deduplication:

This method tests at file level, and complete file is reflected as a solo segment. According to this method, it authorizes the backup file record to match the components saved in the data. On the off chance that the equivalent file occurs, it enhances a pointer to the current file; else it updates and save the new content of the file. In this way, just one occurrence of the file is stored, and it is additionally called as unique storage. Here, entire file hashing method is easy to apply. Since, file hash numbers are anything but difficult to create, and it needs a smaller amount of preparing control. A couple or one byte change in file triggers an age of an alternate or new hash amount that requires dissimilar space. It is the main issue of file-level deduplication prompts the overview of sub file or block level deduplication methods.

It is also known as Record level deduplication [3], as the name suggests, is always implemented over a private article. The Familiar confirmation of same hash assessment of two records checks that the reports are equivalent. It doesn't break the records into little pieces but it utilizes whole document as piece.

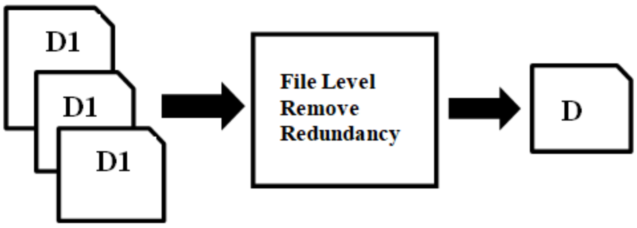

Figure 2 This method only eliminates duplicate files and keep on single instance of file.

i. Disadvantage of File level deduplication:
(a) File level deduplication cannot be useful for the big files with shifting or moving data.
i i. Advantages of File level deduplication:
(a) Little CPU utilization
(b) Indexing performance

B. Block level deduplication

Block level deduplication is called as square level deduplication and it is executed over pieces. Firstly, it segregates the data into pieces and saves only a private copy of each block or square. In these strategies, when a file is broken into different little blocks, it may be variable-size or fixed size blocks or squares. SHA-I, MD5, Rabin fingerprinting and comparative hash calculations are utilized to distinguish related squares.[4] In this way, the unique square is kept in touch with disk and its file is refreshed. Something else, a pointer is added to similar information block's root division. It requires all the more preparing control in light of the fact that the quantity of identifiers increments significantly that should be handled. It is additionally classified as fixed-length or variable-length deduplication [3].

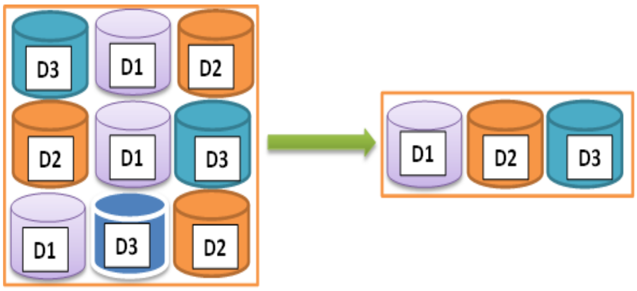

Figure 3 shows block level deduplication where deduplication is based on piece of file

Fixed length block deduplication:

Fixed-length methodologies observe squares of information with a prearranged length. It split the data into fixed-size squares. The framework doesn't reclaim up of same square of information twice. In this the complete file not considered as a smallest one so it divide it into equal size chunks or pieces, if a big file is changed then at that time changed pieces must be re indexed and move it to backup.

Advantages of Fixed length block deduplication: The main benefit of using this method is its simplicity. When two characters are added in the data then it require a moving of data or information by two bytes. The subsequent all information chunks are backed up once more.

Disadvantage of Fixed length block deduplication: It neglects to notice excess information if a few bytes are documents into equal sized pieces be re-filed and moved to the backup area erased or inserted from the record since piece boundaries are controlled by counterbalance as opposed to by content.

The drawback of this method clues to the invention of variable-length method.

Variable length block deduplication

It uses different methods for deciding the length of the block. In variable length deduplication, it first partitions the file into variable-length data or information squares. This helps the limits of information squares to "float" with in information stream so changes in a single piece of square have no effect on the limits in different areas of squares. The file is divided into subject subordinate way, and the section is of any bytes long inside a series. It gives more prominent granularity control and provides flexibility to add information in square. It characterizes breakpoints. This is typically done by fixed size coinciding sliding. At each offset of a document, the matter of the sliding turns out to be genuine window is considered and a unique mark f is determined. On the off chance that f fulfills the break condition, another break point has been found and new piece is made.

IV. LITERATURE SURVEY

Ng, W.K., Wen, Y. and Zhu, H. (2012) Authors discussed the private data deduplication in which a user or customer has a private data and it shows to server with the string of the data in brief form. It’s basically a customer duty that he should not disclose to server regarding their personal data. Authors analyzed private deduplication method which include cryptography standard. It shows that simulation based system which include hash function are more secure that is collision resistant but discrete algorithm is complex.

Sun, Zhang and Zhu (2014) Authors analyzed when normally encrypt the data it faced many key management issues and can’t help complex necessity for example fine grained approval, parallel alternation and query. Authors suggested fully homomorphic encryption strategy for data confidentiality, which can work for any activity that can do it in clear content and not taking the procedure of decrypting the data. The homomorphic encoding calculation is inefficient, a light weight component purposed for database encryption known as transposition, substitution, folding and shifting (TSFS) calculation. In this paper authors purposed various procedures for information assurance and accomplish most abnormal amount of information security in the cloud but these methods become more effective when the gaps in the study can be filled.

Mahalle & Shahade (2014) Authors purposed a hybrid encryption system utilizing AES and RSA algorithm in this paper. High Security of content records and make the entrance of unique document by intruders close to impossible is given by utilizing hybrid encryption calculation. Be that as it may, just a single drawback here is the execution time. The time taken to execute the entire procedure is higher than other strategies.

Hussain & Ashraf (2014) Authors discuss various kinds of security issues and present a trust model to upgrade the security and interoperability of cloud computing condition. Authentication issue can be solved by utilizing digital signature and he recommended that information is encoded utilizing cryptography strategies in every cell of a table in cloud. At whatever point a client needs to make a question, the inquiry parameters are assessed against the information stored. The principle behind it is to part the information over various hosts that are non-communicable. To determine the issue of multi Tenancy, the authors recommended the cloud suppliers should utilize intrusion detection framework to protect their clients in cloud environment.

Usman Jan and He (2016) Authors proposed a protected, lightweight, strong and effective plan for information interchange among the media clouds and mobile clients. The aim of this paper is to help real time formulating with power saving requirement. The purposed methodology is basically a mixture of PUK, SK and PRK and no requirement of consistent association between the clients. In this paper, high proficiency video coding related plan of encryption is discussed which is time effective, less unpredictable as far as calculations. The proposed plan encrypt the secret information in compressed space, not in encoding area and it comprises of different stages, which are information encryption, video encoding and decoding with or without interpreting. The analyzed plan attempts to maintain the visual nature of that video and keep the size of video stream same as before encryption process.

Aloraini & Hammoudeh (2017) Authors present a survey of the three fundamental information security characteristics with regards to cloud computing specifically accessibility, integrity and privacy. Authors give a technique for securing the data in cloud conditions utilizing RSA algorithm and steganography method yet in this including an additional layer of security includes more expense regarding time. Authors proposed an arrangement which uses the HMAC algorithm to guarantee information reliability and random number secret key to give greater security. By far most of the information secrecy arrangements in the cloud give static information storage approaches and don't examine the modification of these strategies

Singh, Nafis & Sethi (2017) Authors proposing a hybrid approach of RSA and SHA1 for improving the security. RSA is for encryption and decoding of the information and SHA1 is for creating hash value. It gives security which just the approved client can get to it. According to them before putting away the information into the cloud, it initially encodes the information. After receiving the request from the customer, CSP verifies the client and decrypts the message and conveys it to the client.

Authors take a string and produce public and private key and encode the string utilizing RSA algorithm lastly creating the hash value of a similar message utilizing SHA1. In this, we selected the tool which requires highly skilled professionals and it is nearly costly.

Xiong & Shi (2018) Authors suggested the protection preservative contract out plan of flexible information hiding over encoded picture information in distributed computing. It gives guarantees that multimedia information security without depending on the dependability of cloud servers. Authors' purposed two reversible information hiding schemes like reversible data hiding and reversible information covering with homomorphic encryption in encoded space. Authors analyzed the flexible information hiding can be worked above the encoded picture at the various phases. This paper demonstrates the better execution in terms of capacity and security as contrast with existing arrangements.

Abdulhamid, Sadiq, Abdullahi, Rana and Chiroma (2019) Authors bring the possibility of advancement of Blowfish encryption scheme that empowers them to encode their information before storage in the cloud. Authors considered and tested the created encoded application with ordinary measurements and contrast it with other scheme. The purposed plan is utilized for secure information storage in the public and business cloud computing condition. As per authors Blowfish acknowledged as toughest, highest speed and permit free algorithm and it produce an extraordinary key for encoding the message and similar key is utilized to recover the information from the cloud.

Pachpor, N.N. and Prasad, P.S. (2018) Authors discussed some issues related to primary storage system in which the performance as well as data security of large size file can be improved but when the size of the file from 4KB to 8KB then deduplication cannot improve the performance and security of the system. Author purposed some method that reduce the traffic on the network. These methods are not capacity oriented but it improves the primary storage system.

Kaur, R., Chana, I. and Bhattacharya, J. (2018) Authors discussed different types of deduplication like file level deduplication, block level deduplication and it applied on text, image and video data. Author suggested that deduplication reduce overall storage cost, reduce storage space and improve the network bandwidth but this paper discuss some challenges related to deduplication like performance issue, small size files, optimization techniques.

Tan, C.B., Hijazi, M.H.A., Lim, Y. and Gani, A. (2018) Cloud Storage has modern development for storing the data over the traditional storage system. Authors suggested some accessibility methods for storing the data in cloud like proof of retrievability and provable data possession for ensuring the customer that their data will be in secure mode in the cloud like traditional storage system. Providing accessibility with the help of these methods but they analyzed some challenges like finding the actual location of server where the information is stored. Geo location of the stored data, assured deletion and deduplication which are related to data storage still a challenge for cloud storage.

V. RESULTS AND DISCUSSIONS

A. Advantages of data deduplication

(1) Reduce overall storage cost: It helps in the investments of space, budget, human resources, network bandwidth and time. It principals to improved efficacy and efficiency of storage network [5].

(2) Reduce storage space: It supports in decreasing the storage space needed for file, backups, or some other applications [6]. It provides a single copy of data or information which is saved and additional copies are detached. So, it makes extra space to save more data [7]

(3) Minimize energy usages: Deduplication is a capacity enhancement method that decreases energy and storage necessities. At the point when the smaller space requires less power and coolants. So, it means saves energy and storage supplies and decrease burden on network resources.

(4) Efficiently increase network bandwidth: The exclusive data are saved on memory and indicators are made for matching information [8]. Basically, it gives benefits in dropping network bandwidth necessities.

B. Disadvantages of data deduplication

(1) Downfall of information integrity: The information chunks are listed through hash standard for better query. The hashes of same type can be produced for various information chunks because of hash impact or collision that can root loss of information integrity [9] [10]. In this way, impact of hashes must be deliberately routed to maintain a strategic distance from any loss of information and its honesty.

(2) Security and privacy: These strategies have complete control on whole storage. It tends to be abused to get total fetch of capacity. The security of these strategies should to be deliberately intended to monitor framework from such security breaks and damage of personal information [11].

(3) Effect on capacity execution: In essential storage framework, fixed-size methodology prompts many pieces put away at various memory areas. It prompts breakup issue, which unfavorably impacts the presentation. It requires extra assets like memory, CPU and data transfer capacity for its implementation [12].

(4) Issues of Backup Machine: It may need a different equipment scheme to move and process information. Such reinforcement machine may prompt extra amount and effect storage execution.

VI. CONCLUSION

Data deduplication is a specialized data compression technique for eliminating duplicate copies of data in storage. In this paper, different deduplication methods like block level deduplication and file level deduplication was studied. In view of the literature study of deduplication methods, it has been seen that information deduplication procedure on cloud storage framework is a potential research area. In future, to make information deduplication financially savvy and energy efficient as far as space, there is a requirement to build up a capable deduplication method with ideal utilization of memory, CPU and system assets will remain in core interest.

REFERENCES

1. Inbaraj, X.A.,Rao, A.S.:Modified Secure Data Deduplication Computing in Cloud based Environment 2017.
2. Kaur, R., Chana, I.,Bhattacharya, J.:Data deduplication techniques for efficient cloud storage management: a systematic review. The Journal of Supercomputing, Vol.74, 2018, pp. 2035-2085.
3. More, S.,Gaikwad, S.:Secure Cloud Using Secure Data Deduplication Scheme 2018.
4. Rawal, B.S., Vijayakumar, V., Manogaran, G., Varatharajan, R.,Chilamkurti, N.: Secure disintegration protocol for privacy preserving cloud storage. Wireless Personal Communications Vol.103, 2018, pp. 1161-1177.
5. Jose, G.S.S.,Christopher, C.S.: Secure cloud data storage approach in e-learning systems. Cluster computing 2018, pp. 1-6.
6. Velapure, S.S., Barde, S.S.:A Hybrid Cloud Approach for Secure Authorized Deduplication 2014.
7. Wu, S., Li, K.C., Mao, B., Liao, M.: DAC: improving storage availability with deduplication-assisted cloud-of-clouds. Future Generation Computer Systems, Vol.74, 2017, pp. 190-198.
8. Pachpor, N.N.,Prasad, P.S.:Improving the Performance of System in Cloud by Using Selective Deduplication. In Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) 2018, pp. 314-318.
9. Zafar, F., Khan, A., Malik, S.U.R., Ahmed, M., Anjum, A., Khan, M.I., Javed, N., Alam, M., Jamil, F.: A survey of cloud computing data
10. integrity schemes: Design challenges, taxonomy and future trends. Computers & Security, Vol.65, 2017, pp. 29-49.
11. Tan, C.B., Hijazi, M.H.A., Lim, Y., Gani, A.: A survey on Proof of Retrievability for cloud data integrity and availability: Cloud storage state-of-the-art, issues, solutions and future trends. Journal of Network and Computer Applications Vol. 110, 2018, pp. 75-86.
12. Waghmare, V., Kapse, S.: Authorized Deduplication: An Approach for Secure Cloud Environment. Procedia Computer Science, Vol.78, 2016, pp. 815-823.
13. Singh, P., Agarwal, N.,Raman, B.: Secure data deduplication using secret sharing schemes over cloud. Future Generation Computer Systems, Vol.88, 2018, pp. 156-167.
14. Ng, W.K., Wen, Y.,Zhu, H.: Private data deduplication protocols in cloud storage. In Proceedings of the 27th Annual ACM Symposium on Applied Computing 2012, pp. 441-446
15. Singh, S., Scholar, M.T., Nafis, T., Sethi, A.: Cloud Computing: Security Issues & Solution Vol.13, 2017, pp. 1419-1429.
16. Xiong, L., Shi, Y.: On the privacy-preserving outsourcing scheme of reversible data hiding over encrypted image data in cloud computing. Computers, Materials & Continua Vol. 55, 2018, pp. 523-539
17. Aloraini, A., Hammoudeh, M.: A survey on data confidentiality and privacy in cloud computing. In Proceedings of the International Conference on Future Networks and Distributed Systems 2017.
18. Usman, M., Jan, M.A.,He, X.:Cryptography-based secure data storage and sharing using HEVC and public clouds. Information Sciences Vol. 387, 2017, pp. 90-102.
19. Sun, Y., Zhang, J., Xiong, Y., Zhu, G.: Data security and privacy in cloud computing. International Journal of Distributed Sensor Networks Vol. 10, 2014, pp.1-9.
20. Mahalle, V.S., Shahade, A.K.: Enhancing the data security in Cloud by implementing hybrid (Rsa & Aes) encryption algorithm. In International Conference on Power, Automation and Communication (INPAC) 2014, pp. 146-149.
21. Hussain, I., Ashraf, I.: Security issues in cloud computing-a review. International Journal of Advanced Networking and Applications Vol. 6, 2014, pp. 2240-2243.
22. Chhabra, N., Bala, M.:A Comparative Study of Data Deduplication Strategies. In First International Conference on Secure Cyber Computing and Communication (ICSCCC) 2018, pp. 68-72.
23. Yan, Z., Ding, W., Zhu, H.: A scheme to manage encrypted data storage with deduplication in cloud. In International Conference on Algorithms and Architectures for Parallel Processing, 2015, pp. 547-561.
24. Fan, Y., Lin, X., Liang, W., Tan, G., Nanda, P.: A secure privacy preserving deduplication scheme for cloud computing. Future Generation Computer Systems 2019.
25. Reddy, C.R.K., Phaneendra, K.N.:Compreive Anasysis On Secure Data Deduplication With Dynamic Ownership Management In Cloud Data Storage 2018.
26. Tulasi, K.,Karanam, S.: Heterogeneous Data Storage Management with Deduplication Aware in Cloud Computing. In IADS International Conference on Computing, Communications & Data Engineering (CCODE) 2018.

AUTHORS PROFILE

Ms. Pronika, pursed Bachelor of Technology from KUK, University, India in 2007 and Master of Technology from Banasthali Vidyapith, Jaipur in 2009.She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies. Her area of research is cloud computing, database, computer network and security. She has 11 years of teaching experience.

Dr. S. S. Tyagi, is presently working as a Professor, Computer Engineering and Dean at Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad. He completed his Ph.D in Computer Science and Engineering from Kurukshetra University, Kurukshetra. He did his M.E from BITS Pilani and B.Tech in Computer Technology from Nagpur University.

He is having an experience of more than 27 years in academics/teaching and research. He is a senior member of various professional organizations like IEEE, ACM, CSI, QCI, ASQ etc. He is past chair, IEEE Computer Society, Delhi Section. There are more than 70 publications to his credit in National and International Journals. He is associated as an editor/reviewer of various journals. He has guided 06 Ph.Ds and several M.Tech Thesis and guiding Ph.D scholars in the field of Software Defined Networking, Cloud Computing, Adhoc Networks, Wireless Security etc.