International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Analysis on Deduplication Techniques for Storage of Data in Cloud

Ujjwal Rajput[1], Sanket Shinde[2], Pratap Thakur[3], Gaurav Patil[4], Prof. Poonam Deokar[5]

[1,2,3,4]*Student, Dept. of Information Technology, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra, India*
[5]*Professor, Dept. of Information Technology, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra, India*

Abstract - *Cloud storage service providers address the need for organizations and individuals by allowing them to store, transfer and back up their ever-increasing amount of data at low cost and provide access to other cloud services. In order to provide efficient data storage, cloud service providers use a widely used drag-and-drop method as it allows for single-data storage and removes duplicate copies, thereby reducing overhead storage and saving upload bandwidth. A customer who uploads his or her data to the cloud is very concerned about the security, integrity, privacy and confidentiality of his or her data. The extraction method is used to manage the duplication of data in the cloud. Although there are some defrosting methods used to avoid data overload, they still lack efficiency. The main purpose of this paper is to obtain sufficient information and a good idea of the extraction strategies by examining existing methods and this work can assist the researcher and the work in their future research in developing effective cloud storage management strategies.*

Key Words: **Big data, Cloud computing, Cloud storage, Data deduplication, Data management.**

1. INTRODUCTION

With the continuous development of the internet, growth, and usage of the internet of things and social networking environments, data size is also increasing exponentially which leads to the requirement of a huge amount of storage space. As per International Data Corporation report, Global Data sphere is the combination of data generated, captured or replicated through the digital content from all over the world. IDC predicts that the Global Data sphere will grow from 33 Zettabytes (ZB) (1 ZB = 10^{21} Bytes or 2^{70} Bytes) in 2018 to 175 ZB by 2025. Cloud computing offers many resources or services, especially the huge volume of storage to back up the big data [2]. Cloud computing is an optimal paradigm for providing storage, computing, and managing big data of the internet of things (IoT) or organization [3] [4]. Cloud service providers provide many services with the features of elasticity, scalability, and pay for usage [5]. To maintain data privacy and security, all data owners store only encrypted data on clouds. Many users are storing their own data, and it has the chance of data duplication in clouds such that different users may send the same data but with different encrypted technology. Even CSP provides a large amount of storage; data redundancy requires extra storage space and higher bandwidth, and also it is a tedious task for service providers to manage a large amount of storage space and duplicate copies. Deduplication is an optimal technique to manage data duplication [6]. It compresses the data by eliminating duplicate copies of data. Deduplication reduces the storage space up to 90 to 95 percent, bandwidth rate, and provides good storage management [7]. Most cloud service providers implement a deduplication mechanism to achieve the efficient storage management of big data [9] [10]. Data deduplication approaches overcome the issues in the storage management of increasing big data in clouds.

Despite of various choices available for data storage including cloud data storage, one of the major challenges faced by users & organizations is about data duplication. It has been observed that for a single user or single transaction, there is a lot of duplication of data resulting due to usage of different sources of information. Data deduplication is one of the effective techniques for data reduction. This technique ensures storage one single copy of each data. This is possible by comparison of data fingerprints with the existing stored data and thus identifies duplicate data. The replication factor is the minimum number of copies of the same data. The Cloud storage system maintains a replication factor for all data. If any data is greater than the replication factor, then the deduplication technique eliminates that data to reduce the storage requirement, cost, and bandwidth rate. All existing deduplication techniques are still lack of efficiency because of the demerits of data comparing and matching algorithms and security issues. All data are stored in the memory location, and each location is identified by pointer or address. Thus only one copy is maintained, and storing a pointer in the duplicate data place helps to free those locations in the storage space.

The main goal of this article is to study the efficiency and inefficiency of existing deduplication techniques. Hence the deduplication process is mandatory for the cloud service provider to reduce the huge amount of storage space requirement, cost, and higher network transfer rate. Another intention of this paper is to bring together researchers and practitioners to have a great idea in various deduplication approaches. This paper summarizes the merits, demerits, and

limitations of deduplication approaches. It may give many directions and future challenges to interested researchers. In paper the fundamental and background concepts in data deduplication are described and also discussed the various existing deduplication approaches. Finally, a conclusion is made based on the study of various data deduplication techniques.

1.1 Evolution of Data Deduplication

With the revolutionary development of the internet and communication technologies, the growth of data is increased rapidly. Hence the word big data is commonly used to denote the huge amount of data and datasets [14]. Big data are unstructured data and have three properties (i) volume, which represents the size of data; (ii) velocity denotes the frequency of data generation and (iii) variety such that data are in a different format and from various heterogeneous data sources. Handling these kinds of data is still a challenging task for researchers and developers. This part of the paper is utilized to describe the background and desirability of the data.

In the early years, Data redundant approach is used to detect and eliminate redundant data in the storage space. Two types of approaches loss-less and lossy are used in data redundant techniques [15] [16]. This scenario had gone into the next level; thus, delta intelligent data compression techniques are introduced to avoid duplicate files or chunks. In the 2000s, data deduplication approaches are proposed to remove redundant files or chunks. Various data deduplication techniques are Privacy-Preserving multi-domain big data deduplication, Leveraging Data deduplication, Cross-Domain Big Data Deduplication, and Attribute-Based Storage Supporting Secure Deduplication. The cloud service providers use any data deduplication technique to achieve efficient cloud storage management. Loss-less data redundant techniques propose both byte level and string level approaches to detect and eliminate all kinds of data. Huffman coding and dictionary coding are two main approaches to loss-less redundant techniques. In the delta compression technique, Xdelta and Zdelta are approaches used to remove the duplicate copies at the string level. Data deduplication techniques remove the redundant data at the file and chunk level by using the deduplication techniques. Fig. 1 shows the progress in the efficient storage management of big data.

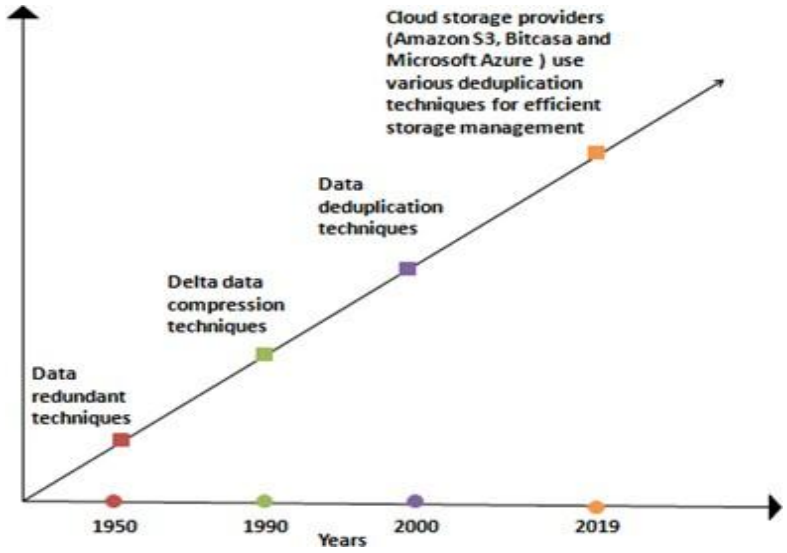

Fig. 1. Progress of efficient storage management.

1.2 Data Deduplication

With the rapid growth of information technology, the internet, social network environment, IoT devices, data are generated in high velocity, variety, and from different sources[18]. Big data is the right term to describe such kind of data. The task of storage management of big data is crucial for researchers and practitioners [19]. Data deduplication is a technique used for removing duplicate data chunks and improving cloud storage efficiency. Fig. 2 illustrates the general method of big data deduplication. Deduplication reduces storage space as it allows only single instance of data to be retained and removes all the duplicate copies of data files and all the redundant data files are replaced by pointer which points to the unique instance of data.

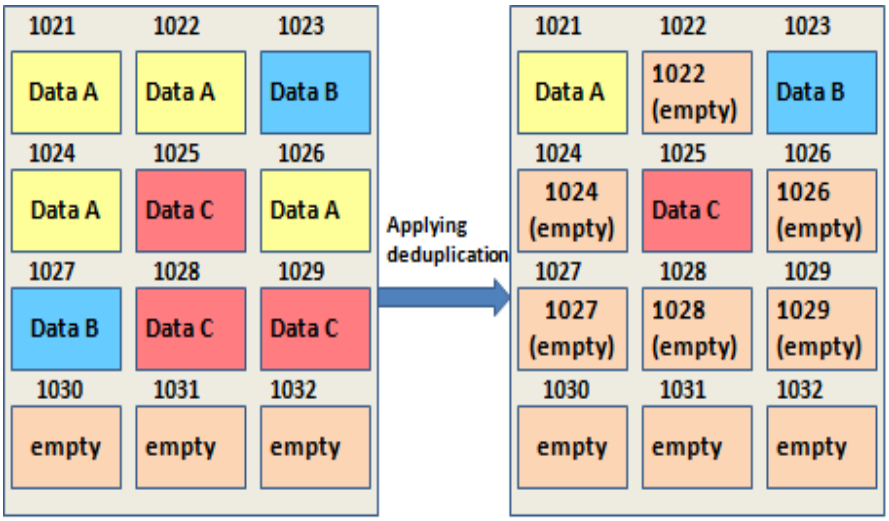

Fig. 2. General data deduplication approach.

1.3 Benefits and Limitations of Data Deduplication

This section gives a comparison of the benefits and limitations of the data deduplication.

Benefits of data deduplication:

- As duplicate data are removed, the amount of cloud storage is freed and available for any other new data.
- As redundant data are removed and only one copy is maintained, the rate of network bandwidth is reduced.
- Due to data deduplication, storage space, bandwidth rate, time, workforce, and cost are reduced.
- Deduplication provides efficient cloud storage management.

Limitations of data deduplication:

- Sometimes to increase the availability of data, multiple copies may be stored in the clouds. Due to data deduplication, the availability of data may be poor.
- Data are located using hash values for easy storage management. Different data may have the same hash index, which may be wrongly identified as duplicate copies (hash collision). It causes loss of data and integrity.
- An individual additional hardware system is required to perform deduplication. It increases the hardware cost. Due to having all rights to access and control over all cloud data, security and privacy is a big issue. Only careful and highly securable duplication techniques can address this issue.

2. DATA DEDUPLICATION PROCESS

The data deduplication is accomplished by some processes such as verifying the format or type of data chunks or files, determining the hash or fingerprinting value of new and existing chunks. Then this task extends by the process of matching the hash value of a new chunk with the hash values of existing chunks, and finally, this process ends at either removing the data chunk or storing it into the memory [21]. Fig. 3 shows the various processes involved in the data deduplication technique.

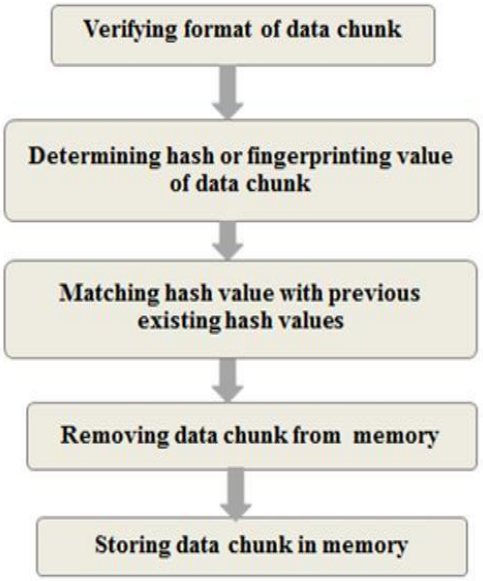

Fig. 3. Various processes in data deduplication approaches.

The steps involved in the general data deduplication technique are illustrated in Procedure-1.

Procedure: General data deduplication approach

1) Input: New data chunk of file DC.
2) Output: Redundant data chunk if it's appeared
3) Get the new data chunk or file DC is to be stored in the cloud
4) Identify the format of a data chunk DC and split it into objects (fixed or variable)
5) DC= {dc1, dc2,.., dcn} , dc1,dc2,..,dcn are objects.
6) Determine the hash value or fingerprinting value, hv(DC) of the new data chunk or file DC.
7) Match the hash value of a new data chunk hv(DC) with the previous or existing hash values {hv1,hv2,..,hvn} in the cloud storage or memory, {hv1,hv2,..,hvn} are hash values of data chunks already stored in the cloud.
8) If there is a match, then remove the data chunk (i.e. duplicate or redundant data), otherwise (not match) store a new data chunk in the memory and maintain the new index.

A new data chunk or file is stored in the cloud after accomplishing various tasks. The first task is identifying the format of the data chunk that the format of the file may be fixed or variable. Then the second task is finding the hash or fingerprint value of the new data chunk. The determined hash value is matched with the hash values of existing data chunks in the cloud. If the calculated hash value of a new data chunk is matched with any one of the existing hash values, then the new data chunk is declared as redundant data, and this redundant data is eliminated. Otherwise, the new data chunk is stored in the cloud.

2.1 Categorization of Data Deduplication Approaches

Data deduplication approaches are categorized or grouped based on (i) type of memory/storage, (ii) type of terminal sending or receiving data, (iii) time(before or after storing data in disk), and (iv) level how process accomplished. The storage-based group consists of Primary storage and secondary storage techniques, type-based consists of source and target approaches, time-based has inline and post-process techniques, and level-based consists of distributed and global techniques. Figure Fig. 4 lists the classification of data deduplication techniques.

1) *Primary storage technique:* In this approach, data deduplication is done at primary memory or main memory under the control central processing system. The performance of this approach may be decreased when a large amount of data is loaded into the primary memory [22]. It is a good approach for confidential data and efficiently performs primary activities. Example: primary workloads and mail servers.

2) *Secondary Storage technique:* The deduplication process is taken place at a backup server or secondary memory. It stores the important primary data in auxiliary memory to avoid the loss of data due to natural disasters. It is not controlled by the CPU. As data are stored in a remote device, this approach performs faster data deduplication. This eliminates all duplicated data and gives better performance than the primary data deduplication [2]. This technique dumps and loads old data efficiently. Example: archives and file backups.

3) *Source deduplication:* Data redundancy is done at the client or source terminal before data are sending to the backup server or target terminal. It requires less storage space and hardware comparing to the target deduplication. Thus the amount of memory space, network transfer rate, and time of storing data in the cloud is ultimately reduced. For small datasets, this approach is good for deduplication.

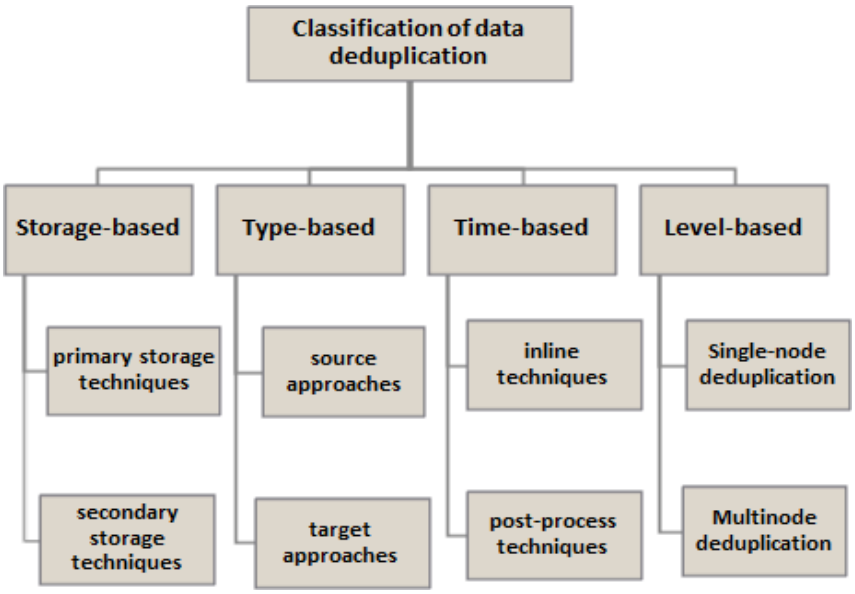

Fig. 4. Classification of data deduplication techniques.

4) *Target deduplication:* In contrast to the source deduplication approach, the removal of redundant data is done at the server after sending the data to the targeted terminal. Hence this approach needs more storage space and hardware than the previous approach. It needs a high communication channel or bandwidth. This approach is more beneficial to deduplicate large datasets.

5) *Inline deduplication approach:* Redundant data is removed at the source terminal or before storing data into the disk. It doesn't require extra storage spaces or a disk. This approach is a flexible and efficient technique thus it executes the data at one time only. The computation time is high.

6) *Offline or post deduplication approach:* In contrast to the inline approach, data deduplication is done after the data is written into the disk. As it requires less computation, this gives better performance than the inline approach.

7) *Local data or single node deduplication:* This approach eliminates the redundant data in the local area network. As the elimination of duplicate data is possible in a single node only, this technique fails to eliminate all redundant data.

8) *Multinode or global deduplication:* This approach performs data deduplication in multi nodes in the distributed storage system. It completely eliminates all redundant data in multi-servers in a distributed environment.

Data deduplication techniques can also be classified based on the type of data. This type-based classification has three categories text-based, image-based and Video-based deduplication. The data of the categories text, image, and video are duplicated in social networks continuously. The text deduplication approach verifies the text byte by byte to find redundancy of the data. The image deduplication implements image detection (exact or near) techniques to find the redundancy of the image. The video deduplication use frame- based techniques thus it converts the video into frames and performs the deduplication in all frames. We summarized the efficiencies and limitations of data deduplication techniques in Table -1.

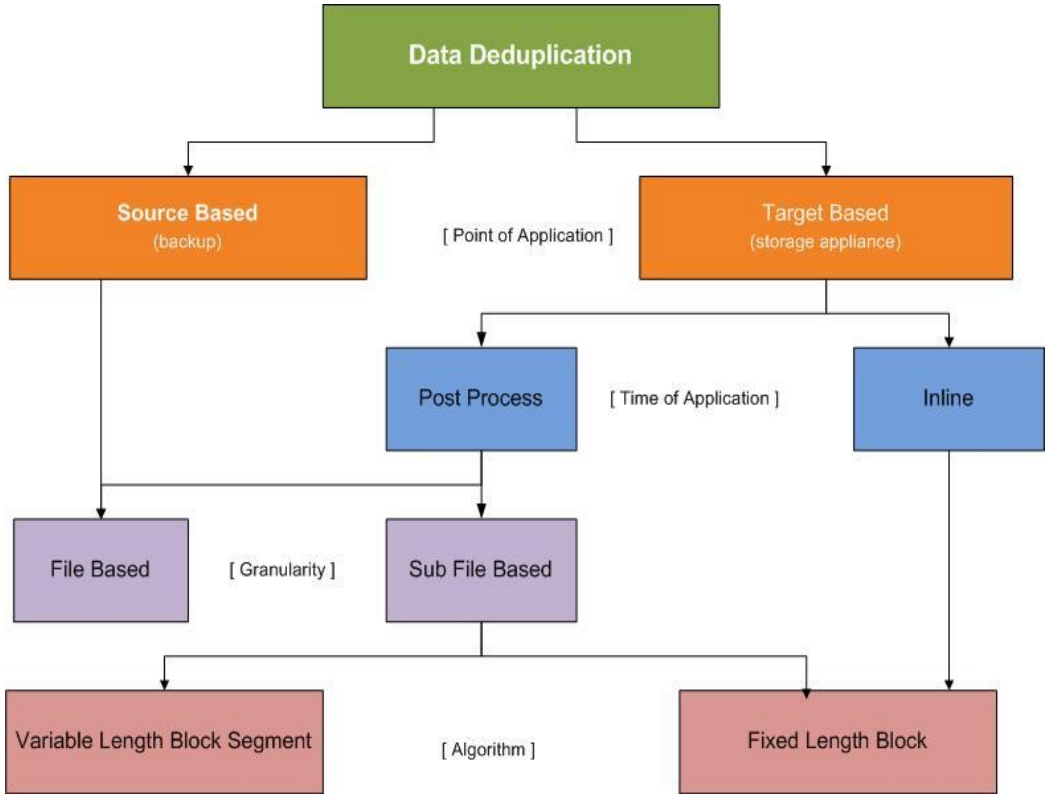

Fig. 5. Categorization of data deduplication approaches

2.2 Summarization of Data Deduplication Approaches

This section presents a comparative study of the classification data deduplication techniques. We used three features specific approaches (an actual mechanism), efficiency, and limitation to analyze the deduplication approaches based on our studies. Each approach is good in some environments however each restricts some features. The researchers or cloud service providers can use any one of them based on their requirements *Data deduplication* or *single instancing* essentially refers to the elimination of redundant data. In the deduplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. However, indexing of all data is still retained should that data ever be required. In the same way that the phrase “single instancing” turns the noun “single-instance storage” into a verb, the word “dedupe” becomes the verb for “deduplication”.

The exponential growth of digital data in cloud storage systems is a critical issue presently as a large amount of duplicate data in the storage systems exerts an extra load on it. Deduplication is an efficient technique that has gained attention in large-scale storage systems. Deduplication eliminates redundant data, improves storage utilization and reduces storage cost. Due to reduced data storage requirement, the data transfer also becomes fats and efficient. Unique data chunks are identified, and by using suitable approach the duplicate data is eliminated by replacing the redundant chunks. This chapter presents an overview of data deduplication methods, types, file chunking, metadata, and the applications of deduplication process.

2.3 The General Architecture of the Data Deduplication Approach

The general architecture of the data deduplication approach requires three components client-server, metadata- server, and data servers. The major task of the client-server is data preprocessing. Thus the client-server performs the process of splitting data chunks, fetching fingerprints or hash values of data chunks, and assigning super-chunks (new data file) to the data-server based on information Least Recently Used (LRU) cache. The client-server is capable of interacting with the data-servers[23]. The metadata-server maintains the necessary information of all data files, chunks, servers. This enables the cloud service provider to recover the loss of data files. The important role of metadata-server is to make two associations (i)file(id of a file)-super chuck and (ii) super chunk-data server(id of a data-server).

The data-server actually performs the deduplication process. It assigns super-chunk(new data chunk) to fingerprint (or hash value) and fingerprints to Existing data chunks. It compares the fingerprint or hash value of two chunks. If two chunks are the same based on the comparison of hash values or fingerprints, then one of them (redundant chunks) is removed, otherwise, the new chunk is stored in the cloud. All this information about storing the new chunk or removing the redundant data are stored in the metadata-servers by sending a request to the metadata-sever from the client-server. The general data deduplication algorithm with these three components is presented below.

Algorithm: Data Deduplication

1. Client-server split the new data file into fixed or variable size data chunks called super-chunks.
2. Client-server fetches the fingerprints or hash values of existing data chunks from the metadata-server.
3. Client-server assigns super-chunks to the data-server
4. Meta-data server gives a unique identifier (file-ID) to all the super-chunks and the identifier(Data-server-ID) to all data-servers.
5. Data-server determines the hash value of super-chunks.
6. Data-server compares the hash value of super-chunks with the hash values of existing chunks in the cloud. If there is any match, then that chunk (redundant chunk) will be removed, Otherwise, the new chunk is stored in the cloud.
7. Client-server send the request to the metadata-server to update the information of the data deduplication process.
8. Metadata-server stores the information of new data chunks or updates the information of the removed redundant chunk.

The hash value of the pair of two data-chunks is compared as follows.

$$SM(SD1,SD2)= \begin{cases} 1, & \text{if } \min(SD1)=\min(SD2) \\ 0, & \text{otherwise} \end{cases}$$

The data file is partitioned into data chunks. SD1 is the set of hash values of the data file D1 where D1 is divided into k number of data chunks $\{d_1,d_2,..,d_k\}$. Hence $SD1=\{hd1_1,hd1_2,hd1_3,..hd1_k\}$. Likely SD2 is the set of hash values of the data file D2, $SD2=\{hd1_1,hd1_2,hd1_3,..hd1_k\}$. The value min(SD1) is the k minimal hash value in the set SD1 and the value min(SD1) is the k minimal hash value in the set SD1. If min(SD1) is equal to min(SD2), then it is concluded that there is a duplicate chunk and the duplicated chunk is removed. Otherwise, the new chunk is stored in the cloud.

Table -1: Summarization of Data Deduplication

Technique	Specific Approach	Efficiency	Limitation
Primary storage	Executes on main memory or storage spaces which have direct control on CPU.	Confidential data and efficiently performs primary activities.	Performance may be decreased due to storing large amount of data.
Secondary storage	Executes on secondary memory that have no direct control on CPU.	A good approach for dumping and loading old data.	Consumes extra time to load data from secondary storage into the main memory.
Source-based	Before outsourcing the data into the cloud, data are deduplicated at the point where data are produced This point is known as a source point.	Memory space, network transfer rate, and time of storing data in the cloud is ultimately reduced.	Requires extra hardware (CPU, I/O devices) to do deduplication at the source point.
Target-based	Data deduplication is done after data are outsourced into the cloud.	Storage space capacity is increased.	Additional hardware is required to install and execute data deduplication functions.
Inline	Data deduplication is done before storing data in the disk.	Reduce the storage space because of storing only one copy of data.	High computation time.
Post-process	Data deduplication is done after storing data in the disk. This is known as offline deduplication.	Less compute time than inline deduplication	Additional storage space is required.

3. CONCLUSIONS

This paper gives the efficiency and inefficiency of existing deduplication techniques with the help of theoretical evidence. The deduplication process is mandatory for the cloud service provider to reduce the huge amount of storage space requirement, cost, and higher network transfer rate. This paper includes a summary of the merits, demerits, and limitations of existing approaches. This analysis may give many directions and future challenges. Although cloud computing offers a huge amount of storage space, data duplication decreases the efficiency and performance of cloud storage, and also it results in poor data management and the requirement of high bandwidth. The deduplication technique is used to manage data duplication in clouds. Data deduplication erases all redundant data and maintains only one copy of the data. Although there are some deduplication approaches used to avoid data redundancy, still they are in lack of efficiency. This paper may give sufficient knowledge and a good idea about deduplication techniques by surveying existing approaches, and this work may help the researcher and practitioner for their future research in developing efficient cloud storage management techniques.

In the future, efficient and reliable data deduplication approaches can be proposed and implemented to remove duplicate copies of data and also to maintain privacy, integrity, and confidentiality of the data. Better Data storage and management techniques in cloud computing can be developed in the future.

REFERENCES

[1] M. B. Waghmare and S. V. Padwekar, "Survey on techniques for authorized deduplication of encrypted data in cloud," 2020 Int. Conf. Comput. Commun. Informatics,ICCCI 2020, pp. 20–24, 2020, doi: 10.1109/ICCCI48352.2020.9104184.

[2] D. V. T. and V. R., "Retrieval of Complex Images Using Visual Saliency Guided Cognitive Classification," J. Innov. Image Process., vol. 2, no. 2, pp. 102–109, 2020, doi: 10.36548/jiip.2020.2.005.

[3] K. Vijayalakshmi and V. Jayalakshmi, "Big Data Security Challenges and Strategies in Cloud Computing : A Survey, International Conference on Soft Computing and Optimising Techniques,August 2019," no. 11824, pp. 11824–11833, 2020.

[4] V. Vijayalakshmi, K. and Jayalakshmi, "A Priority-based Approach for Detection of Anomalies in ABAC Policies using Clustering Technique," no. Iccmc, pp. 897–903, 2020, doi: 10.1109/iccmc48092.2020.iccmc-000166.

[5] P. Singh, N. Agarwal, and B. Raman, "Secure data deduplication using secret sharing schemes over cloud," Futur. Gener. Comput. Syst., vol. 88, pp. 156–167, 2018, doi: 10.1016/j.future.2018.04.097.

[6] C. B. Tan, M. H. A. Hijazi, Y. Lim, and A. Gani, "A survey on Proof of Retrievability for cloud data integrity and availability: Cloud storage state-of-the-art, issues, solutions and future trends," J. Netw. Comput. Appl., vol. 110, no. December 2017, pp. 75–86, 2018, doi: 10.1016/j.jnca.2018.03.017.

[7] Z. Yan, L. Zhang, W. Ding, and Q. Zheng, "Heterogeneous data storage management with deduplication in cloud computing," IEEE Trans. Big Data, vol. 5, no. 3, pp. 393– 407, 2019, doi: 10.1109/TBDATA.2017.2701352.

[8] K. Vijayalakshmi and V. Jayalakshmi, "Shared Access Control Models for Big data : A Perspective Study and Analysis," I Pandian A.P., Palanisamy R., Ntalianis K. Proc. Int. Conf. Intell. Comput. Inf. Control Syst. Adv. Intell. Syst. Comput. vol 1272. Springer, Singapore. https//doi.org/10.1007/, 2021.

[9] Y. Zhou et al., "A similarity-aware encrypted deduplication scheme with flexible access control in the cloud," Futur. Gener. Comput. Syst., vol. 84, pp. 177–189, 2018, doi: 10.1016/j.future.2017.10.014.

[10] L. Meegahapola, N. Athaide, K. Jayarajah, S. Xiang, and A. Misra, "Inferring Accurate Bus Trajectories from Noisy Estimated Arrival Time Records," 2019 IEEE Intell. Transp. Syst. Conf. ITSC 2019, pp. 4517–4524, 2019, doi: 10.1109/ITSC.2019.8916939.

[11] A. Bhalerao and A. Pawar, "Utilizing Cloud Storage for Big Data Backups," pp. 933–938, 2018.

[12] L. Araujo, "Genetic programming for natural language processing," Genet. Program. Evolvable Mach., vol. 21, no. 1–2, pp. 11–32, 2020, doi: 10.1007/s10710-019-09361-5.

[13] H. Hou, J. Yu, and R. Hao, "Cloud storage auditing with deduplication supporting different security levels according to data popularity," J. Netw. Comput. Appl., vol. 134, pp. 26–39, 2019, doi10.1016/j.jnca.2019.02.015.

[14] R. Kaur, I. Chana, and J. Bhattacharya, "Data deduplication techniques for efficient cloud storage management: a systematic review," J. Supercomput., vol. 74, no. 5, pp. 2035–2085, 2018,doi:10.1007/s11227-017-2210-8.

[15] K. P, "Deep Learning Approach to DGA Classification for Effective Cyber Security," J. Ubiquitous Comput. Commun.Technol., vol. 2, no. 4, pp.203–213,2021, doi:10.36548/jucct.2020.4.003.

[16] S. Li, C. Xu, and Y. Zhang, "CSED: Client-Side encrypted deduplication scheme based on proofs of ownership for cloud storage," J. Inf. Secur. Appl., vol. 46, pp. 250–258, 2019, doi: 10.1016/j.jisa.2019.03.015.

[17] D. Viji and S. Revathy, "Various Data Deduplication Techniques of Primary Storage," Proc. 4th Int. Conf.Commun.Electron.Syst.ICCES2019,no.Icces,pp.322-327,2019,doi:10.1109/ICCES45898.2019.9002185.