SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I11P106

Volume 12 Issue 11, 57-85, November 2025

Original Article

An Extensive Analysis on Examining Several Data Deduplication Techniques in Cloud Computing

Bharti Duhan[1], Anju Sangwan[2], Anupma Sangwan[3]

[1,2,3]*Department of CSE, Guru Jambheshwar University of Science & Technology, Hisar, India.*

[1]*Corresponding Author : bhartiduhan12@gmail.com*

Received: 06 September 2025 Revised: 08 October 2025 Accepted: 07 November 2025 Published: 29 November 2025

Abstract - *As computer technologies and internet applications are developing at a fast rate, the volume of data is also increasing dramatically. It becomes necessary to store this huge amount of data in the cloud. Following the outbreak of COVID-19 (Coronavirus Disease 2019), it has been observed that offices started working from home and educational institutes began offering online education. As everything becomes online, the demand for storing online data grows. Cloud computing technology has existed long before COVID-19, but it has grown in popularity as a result of the pandemic. Sometimes the same data is being stored multiple times on the cloud by different users, which consumes more storage space, but the storage memory is limited. As a consequence, some storage optimization technique is required, which gives birth to a technique named Deduplication. It is a technique in which duplicated or redundant data is removed to save storage space. This paper presents an extensive analysis of several data deduplication techniques used in cloud computing. The goal is to study the existing techniques of deduplication and then to determine the tradeoff in terms of performance metrices. The graphical and tabular comparison between various existing deduplication techniques is done using parameters like efficiency, throughput, memory consumption, deduplication rate, and computation time. This paper aims to identify the appropriate technique to be used based on the user's requirements.*

Keywords - *Cloud Computing, Chunking, Data Deduplication, Hashing, Indexing.*

1. Introduction

Cloud Computing means storing and retrieving data from the internet rather than accessing it from the hard drive of a computer. Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user [1]. In basic terms, the cloud is analogous to the Internet [2]. Generally, the internet is shown by a cloud, as shown in Figure 1. A huge amount of data is generated by ever-increasing technologies and applications developed every single day.

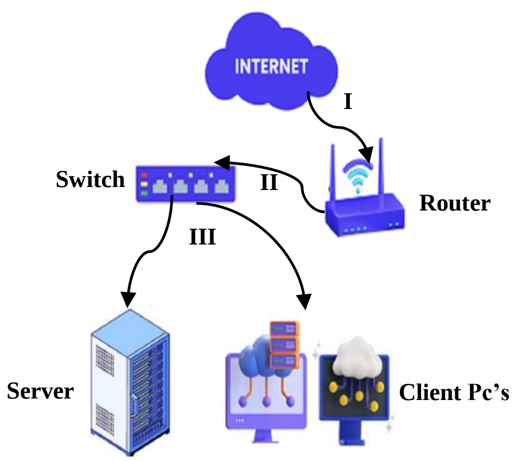

Fig. 1 Internet is represented as a cloud

Many individuals and organizations want to store their data in the cloud so as to get rid of the burden of storage and so that they can share the data with others effectively [3]. Cloud Computing is used when there is a need for a service to be delivered over a network. Customers can access files by using any device over the cloud, but the condition is that their device must be connected to the internet. It is a model that gives access to computing resources on demand [4]. These resources could be applications, services, networks, storage, etc. [5]. There are many dealers who provide services of cloud computing to the customers whenever there is a demand, and these dealers are known as Cloud Service Providers (CSPs). Some popular cloud service providers are Google Cloud Platform (GCP), Apple iCloud, Amazon Web Services (AWS), and Microsoft Azure [6]. Google has cloud apps like Google Docs, Google Slides, and Google Sheets, which provide online storage for its users. Some other services of Google can also work as cloud computing, like Gmail, Calendar, Google Maps, Picasa, Google Analytics, and many more. Apple iCloud provides services like online storage, backup, mail synchronization, calendar, contacts, and many more. In this, generally, data is present on Mac OS, iOS, etc. Amazon Cloud Drive stores data, such as images and music that can be purchased from it. It provides a service like Amazon Prime through which its user gets unlimited storage. Microsoft started providing cloud services in 2010 under the name Azure. All the Microsoft applications run on this cloud. In today's era, Azure is the most reliable and demanded CSP. There are many sectors that use these services, such as business, government, and educational organizations. These organizations access data from cloud servers presented at the data centers by using the internet. Cloud Computing is evolving at a very fast rate in the IT (Information Technology) sector, as well. It is playing a very significant role in handling the growing demands of users for storage and infrastructure [7]. Cloud is different because of its unique property of providing resources, such as hardware and software, through a network. Users can hire resources on the cloud according to their needs by paying only for the required resource.

According to the type of user, clouds are being divided mainly into 4 types. These types are Private cloud, Public cloud, Community cloud, and Hybrid cloud [8-10]. A private cloud is explicitly used by a particular organization. A private cloud achieves the highest security as it is used by only one organization, but at the same time, the organization must pay a high cost. Public cloud works for the general public. Usually, public cloud services can be used free of cost, but security is often very poor, as anyone can access the public cloud. Community cloud is for organizations that have similar interests, such as educational institutes, hospitals, etc. Cost is reduced as the total cost is being divided among organizations mutually. No one outside these mutual organizations can access the data, but sometimes the security can be breached by an insider organization. A hybrid cloud is constructed by combining public and private clouds. This cloud provides the cost and scaling features of public cloud and high security like private cloud [11]. There is a survey report generated every year, named the Flexera State of the Cloud report. This report reveals the necessary information surveyed about the clouds in a particular year. Flexera's State of the Cloud Report for the year 2025 is based on a survey of 759 cloud decision makers and users from all over the world. The division of the types of cloud used by the respondents is shown in Figure 2. Among the 759 respondents, 12% used only public cloud, 2% used private cloud only, while 86% used a multi-cloud approach. 86% of multi-cloud has 2% multiple private clouds, 14% multiple public clouds, and 70% hybrid cloud [12].

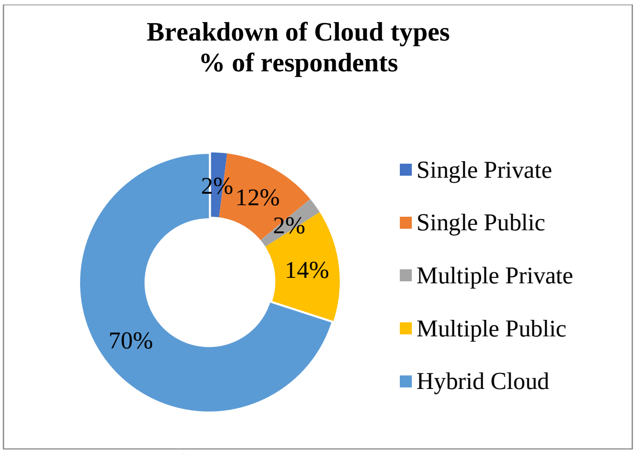

Fig. 2 Breakdown of cloud types used by the respondents according to the Flexera state of the cloud report 2025 [12]

1.1. Motivation and Contributions

After COVID-19 emergence, cloud computing gained popularity. There was a rapid increase in cloud usage due to COVID-19 in 2020. After that, the dependence on the cloud is increasing day by day. As users started using cloud storage more, it became very important to study cloud computing, and as the storage space is limited, it becomes necessary to use the storage resources efficiently.

Data deduplication cannot be ignored in the context of cloud computing, as it emerges as a vital solution to address the problem of storage space utilization in cloud computing. The motivation behind this review paper is to comprehensively explore the various dimensions of deduplication in cloud computing. Study the various existing techniques of deduplication and then evaluate these techniques by comparing them in terms of some specified parameters.

The major contributions of this paper are listed as follows:

- This paper presents a comprehensive study of cloud computing, deduplication, classification of deduplication, and comparison between existing deduplication techniques, making it easy for a layman to understand the concept of cloud computing and deduplication in one place, which is seldom done in any existing review paper.
- This paper uses a Flexera state of the cloud report generated in 2021 and 2025 for a better understanding of cloud computing usage by connecting it with practical implications in today's world.
- A clear and concise tabular comparison is presented between various types of deduplication techniques, based on different performance parameters.
- This paper has a lot of justified diagrammatic representations, tabular comparisons, flow charts, and comparison graphs, which make this paper quite understandable for the scientific and academic community.
- The future research direction of some deduplication techniques has been highlighted in the tabulation summary of deduplication techniques, providing insights for researchers to explore emerging areas of research.

1.2. Characteristics of Cloud Computing

Cloud computing offers very interesting and unique characteristics to its users [13,14]. Some of these characteristics are shown in Figure 3.

1.3. Cloud Computing Service Models

Cloud computing is a general way for technical companies to access technological resources such as hardware or software. It is not a single piece, such as a cell phone or any chip. Instead, it is a whole system which consists of 3 services. These services are IaaS (Infrastructure as a Service), SaaS (Software as a Service), PaaS (Platform as a Service), and an additional RaaS (Recovery as a Service) [15-18]. The description of these models is shown in Table 1.

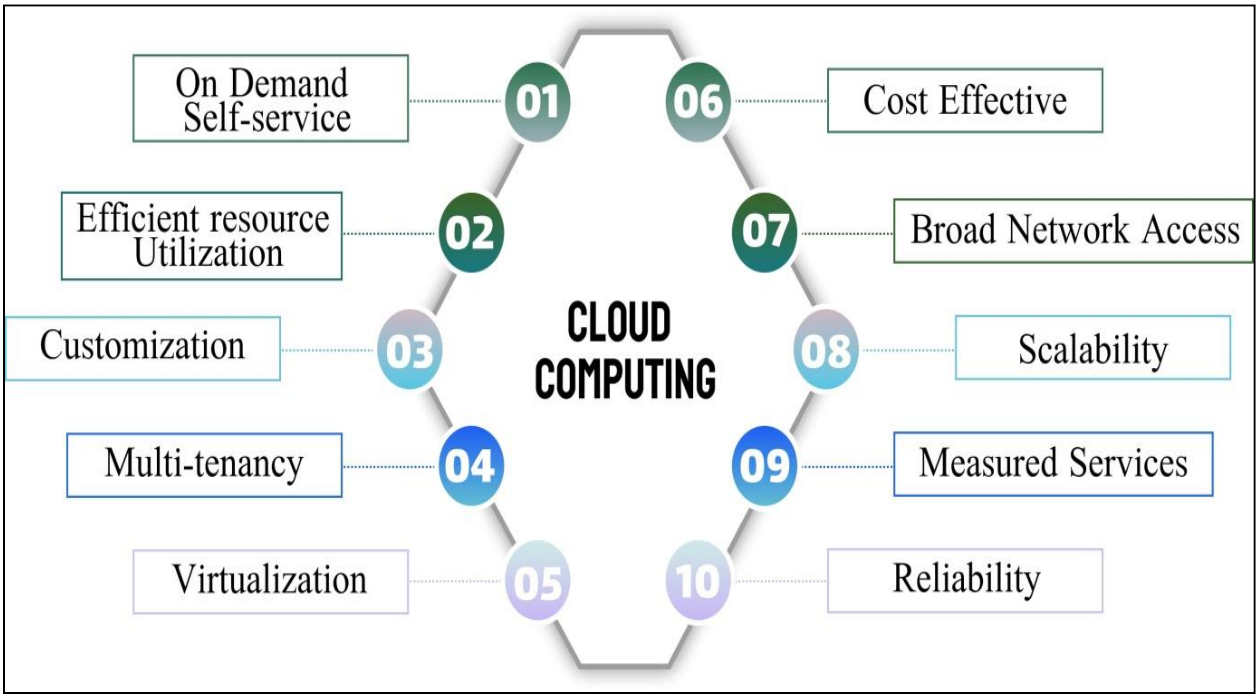

Fig. 3 Characteristics of cloud computing

Table 1. Cloud computing service models

Type	Description	Benefits	Examples
IaaS	It provides online services like storage, a database, and computer capabilities.	• Cost effective • Access to enterprise-grade IT resources and infrastructure • Pay according to usage • Scale up or down resources anytime	Google Docs, Salesforce.com, Acrobat.com
PaaS	It provides a platform for designing, developing, building, and testing applications.	• Simplified deployment • Cost effective • Companies need not worry about upgrades or updates	Win Azure, Google App Engine, Web 2.0
SaaS	It provides online service of desktop applications, which are hosted on cloud infrastructure.	• Rapid scalability • Mobility access • Removal of infrastructure maintenance burden • Customized services available	Acrobat.com, Salesforce.com, Google Docs
RaaS	It provides recovery of data such as the operating system, database files, and applications.	• No data loss • Recovery is cost-effective • Faster recovery with accuracy • Great flexibility on the type of backup needed	Geminare, WindStream Business

1.4. Applications of Cloud Computing

Cloud Computing is one of the most dominant fields of online computing. Resource sharing and management become easy using the cloud [19]. Due to the COVID-19 pandemic, all educational institutions, including colleges, have shut down. Providing education to students becomes possible through online classes, in which cloud computing makes a major contribution. Apart from this, cloud computing plays an important role in many fields [20]. Some of them are shown in Figure 4. The Flexera state of the cloud report for the year 2021 indicates that the emergence of the COVID-19 pandemic has had a great impact on planned cloud usage by the organizations [21]. The report explores the thinking of 750 global users. This percentage change in cloud usage is shown in Figure 5.

1.5. Cloud Computing Challenges

Regardless of its developing influence, problems concerning cloud computing are still an issue [22-24]. Some challenges faced by cloud computing are:

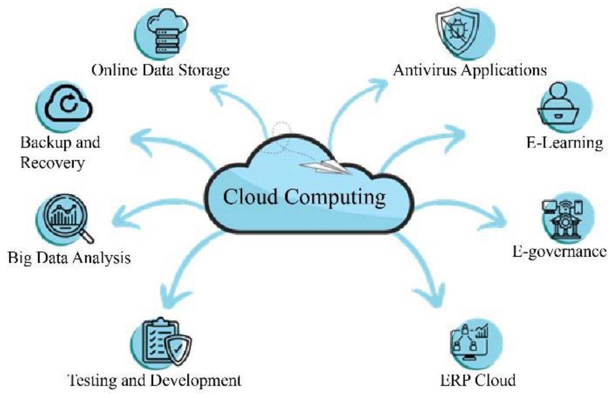

Fig. 4 Applications of cloud computing

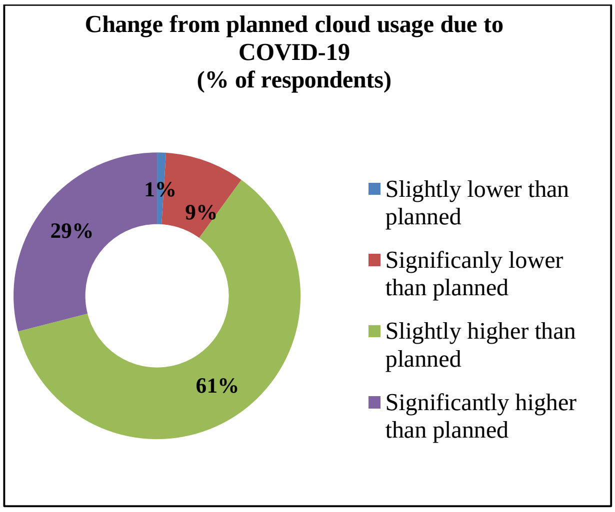

Fig. 5 COVID-19 impact on planned cloud usage for organizations

1.5.1. Security

Data security has always been a major concern for users, as sensitive data stored in a shared environment is exposed to breaches. Everyone wants their data to be secure, whether it is a single user or an organization. Organizations must obtain assurance of their data protection from their vendors. Users have a fear of losing data to competitors or attackers. Organizations aim to keep the data of their users confidential [25-27].

Most organizations keep the actual storage location confidential so that no one can steal their data. Firewalls at data centres are used to protect confidential data. In this model, CSPs have the responsibility of providing data security, and organizations become dependent on CSPs [28].

1.5.2. Data Availability and Service Reliability

CSPs rely heavily on uninterrupted access to data and applications. Cloud services rely on internet connectivity, and any disruption will stop the service, resulting in severe loss of productivity. Sometimes, top CSPs like AWS, Microsoft Azure, and Google Cloud also experience downtime, which results in heavy losses. Ensuring high availability is the responsibility of CSP by managing system checking, disaster recovery, capacity, performance management, and maintenance (Runtime Control), etc.

1.5.3. Governance

There are some countries where governments do not allow the location of personal information or any confidential data outside the country. To fulfill these requirements, CSPs have to set up their data centres within the same country only. Sometimes it becomes a big challenge for cloud providers, as it seems infeasible to them. Data migration from one CSP to another is also not possible.

1.5.4. Software Licence and Data Management Abilities

The management of the platform and infrastructure is still an issue, regardless of the many cloud providers available. There is a high demand for dynamic scaling and dynamic resource allocation by many organizations. Till now, there is also a high possibility of improvement in the scalability and load balancing provided by organizations.

1.5.5. Lack of Resources/ Expertise

There is a shortage of resources required for cloud management, and to date, cloud providers do not have much expertise in handling data in the cloud. There is a high demand for experts in this field, and there is also a need to provide expertise skills to the already present cloud providers.

1.5.6. Managing Cloud Spend

Managing the cost of cloud services is the biggest challenge, as cloud usage and the cost associated with this are continuously rising. There is a need to manage or execute cloud cost optimization strategies. It is no surprise that cloud spend is increasing, and therefore, controlling this fast-growing cost is a top challenge.

According to the Flexera 2025 state of the cloud report, managing cloud spend has overtaken the security challenge for the first time in a decade. Figure 6 indicates the top 5 cloud challenges faced by the respondents according to the Flexera 2025 state of the cloud report [12].

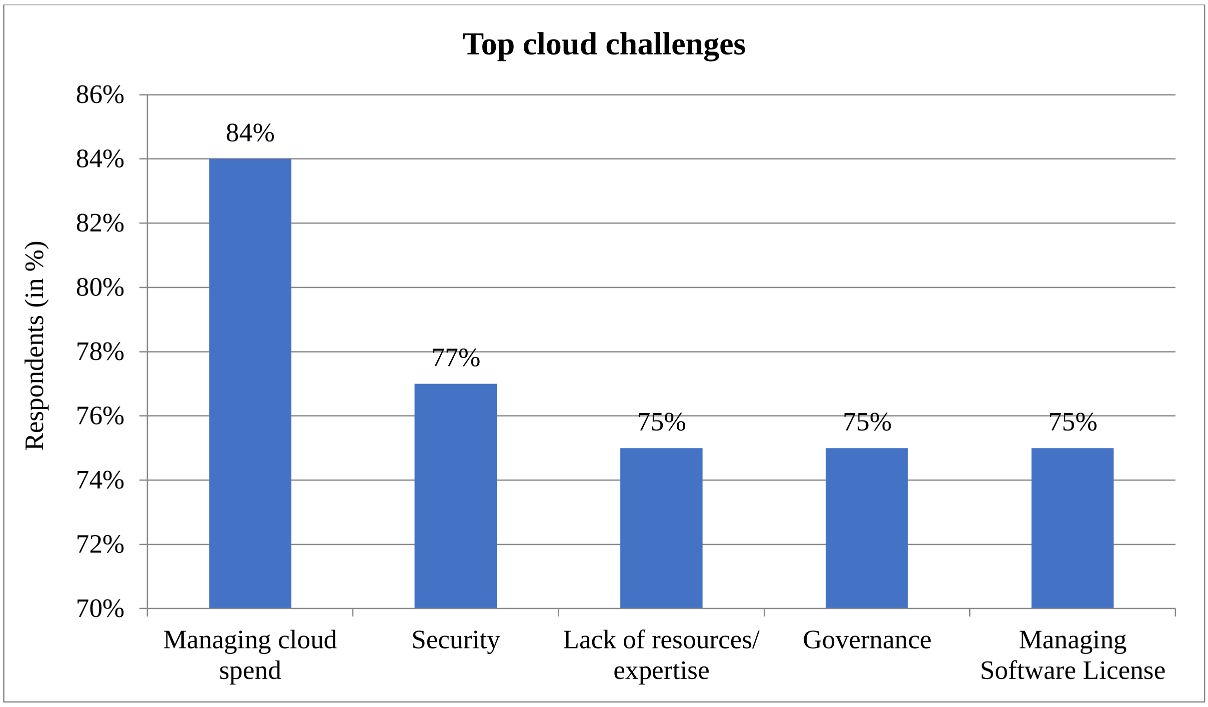

Fig. 6 Top 5 cloud challenges faced by the respondents according to the RightScale 2025 state of the cloud report from Flexera

2. Deduplication in Cloud Computing

In the current era, users are adopting cloud computing, due to which the volume of data on the cloud storage system is increasing rapidly. The reason behind the increasing volume is the widespread use of the internet and the data generated by the various social networking sites. The exponential growth of digital data in cloud systems is a big problem in the current era. Sometimes, the storage of the same data at different sites leads to a huge amount of data duplication, and because of that, memory is blocked with the same copy of data. Due to this, the traffic on the internet and bandwidth for transferring the same data on the network increase [29]. The duplication of data exerts additional load on the storage system and results in wastage of very useful storage capacity. Full storage leads to a shortage of space for upcoming data. There is a need for better storage management and optimization techniques to provide the storage space in cloud systems for the upcoming data. A technique known as Deduplication is introduced for better storage management. The deduplication technique removes redundant data and duplicated copies of the same data. As duplicated copies are removed, only the unique data consumes storage space. The technique is one of the top techniques to handle similar copies of data and provides better storage utilization. The technique has been proven to be an effective technique that eliminates duplicated data and reduces the unwanted use of bandwidth, unwanted storage consumed, and cost. Apart from all this, it is beneficial for cloud service providers as now they can store more data in their existing storage capacity [30-32]. Figure 7 depicts that the deduplication process removes all the duplicated copies of data, and after deduplication, only the unique set of data is saved on the disk storage. If required, then only the logical pointer to the duplicated segments of data is stored [33,34]. There are several cloud service providers (Amazon S3, Microsoft Azure, Dropbox, Google Drive, Bitcasa, etc.) that use data deduplication techniques for better storage utilization.

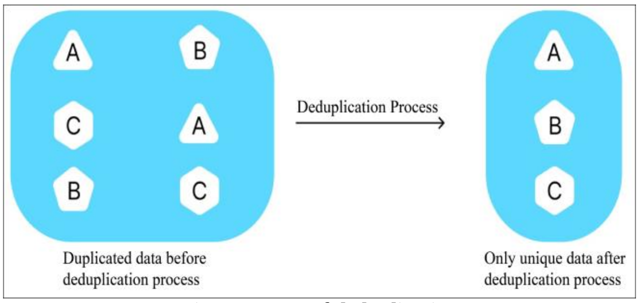

Fig. 7 Process of deduplication

The term data reduction was introduced in the early 1950s, which was further classified into two types, i.e., lossy and lossless data reduction techniques. After this, in the 1990s, the data encoding technique was introduced, in which the compression of similar files is done to minimize the space required for storing these files. Later on, in 2000, the term deduplication came into existence for removing the redundant data from large files both at the inter and intra level. The deduplication technique is very different from data compression techniques (LZ77 and LZ78). In compression, extra data from the particular file is identified and symbolized efficiently, but in the deduplication technique, duplicate data from the same file or from different files is identified and removed for storage optimization. From 2011 onwards, the deduplication technique was also applied to multimedia data successfully, which finds the similarity between multimedia files like images using feature extraction and hash-based techniques [35-38]. The year-wise evolution of deduplication is shown in Figure 8.

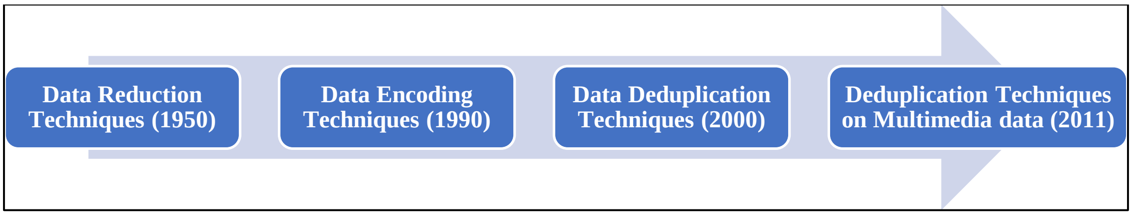

Fig. 8 Evolution of deduplication

The type of data file determines how the deduplication technique has to be applied. The process of deduplication varies from data to data type. Different data types like text, image, audio, and video use different types of deduplication techniques because these have different storage formats and implied characteristics. Basically, the process of deduplication includes four steps, and these are data chunking, calculating a cryptographic hash, index lookup, and storage of a new chunk [34]. First of all, the file on which processing has to be done is divided into fixed or variable-sized chunks. The type of chunking method to be used depends on the file format. After chunking, a hash value is calculated for each chunk by using hashing techniques like MD5, SHA-1, etc. A hash function is used for assigning a unique hash value to every chunk. After calculating the hash, the hash value is compared with the existing hash values from the index lookup table [39]. If the value matches the existing hash value, then only the pointer to existing chunks is saved; otherwise, the new chunk is added into the memory, and the index lookup table is updated. Figure 8 depicts the steps required in the data deduplication process.

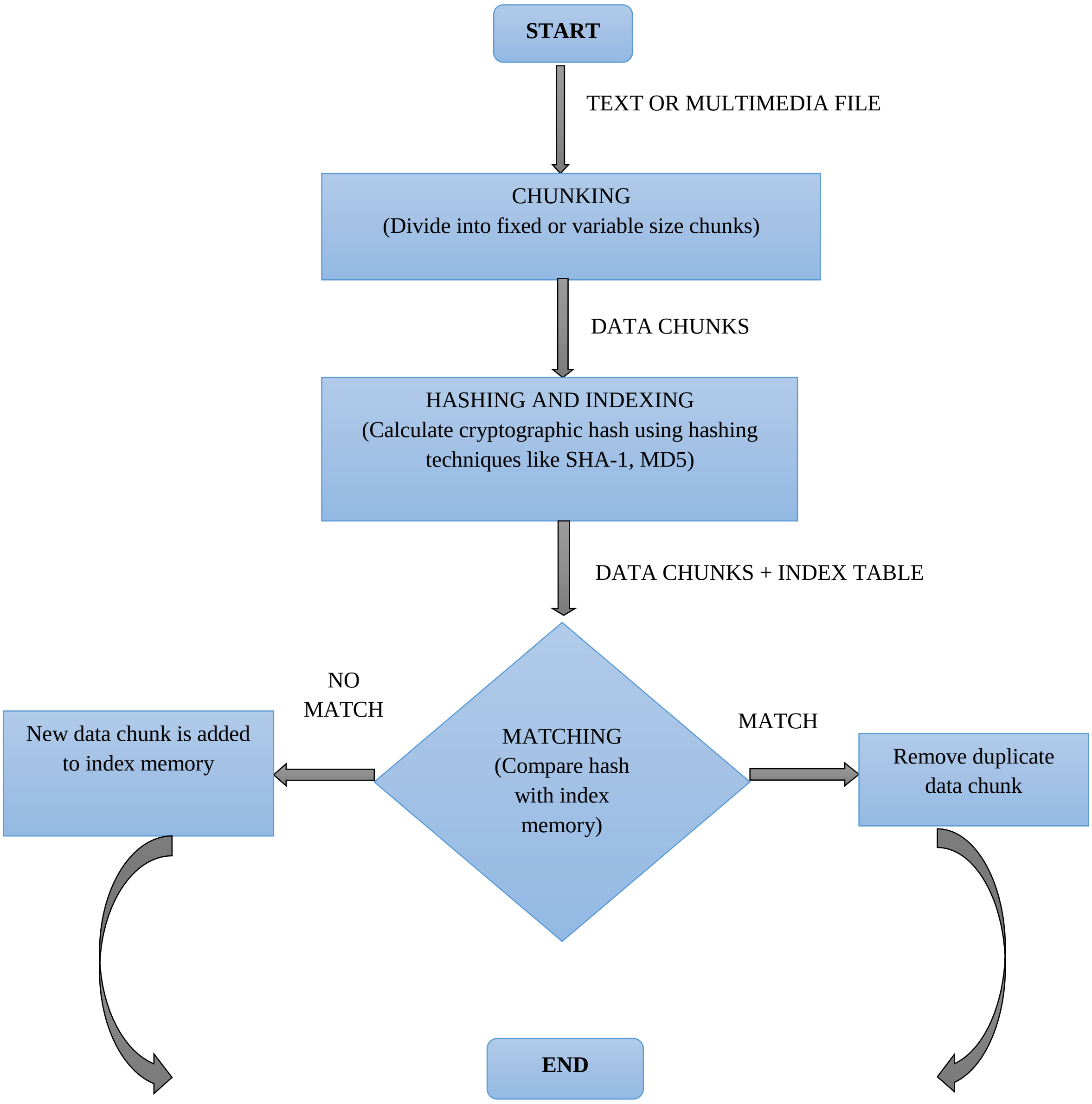

Fig. 9 Steps of deduplication process

2.1. Merits and Security Threats Related to Data Deduplication

Some merits of data deduplication are [40]:

- Storage Efficiency: Deduplication of data will increase the storage space, as the redundant copies are removed.
- Profitability of Cloud Service Providers: As duplicated copies are being deleted, the storage space has increased, so that CSP can assign that space to more users and will get more profit.
- Network Bandwidth Utilization: The server will store only the original copy, and for duplicated data, only links are provided. By using deduplication, less network bandwidth is required, and hence, increases utilization.
- Energy Consumption Efficiency: Energy consumed by the server is decreased as the data stored by the server has also decreased due to the removal of duplicate copies.
- Green Computing: As the data stored at different data centers is reduced. Cooling systems required for maintaining these data centers will generate less carbon, and hence, environmental pollution will be decreased.

Broadly, there are 3 types of threats that can be faced by data deduplication [41,42]:

- Insider Threat: The adversary can be present in the cloud environment, like CSP, and can harm or steal the data. In this type, threats include data breaches like data confidentiality, privacy of data, and integrity of data.
- Outsider Threat: The adversary is from outside the cloud environment. It includes attacks like DDoS (Distributed Denial of Service) attacks, Access Control, Masquerading, etc.
- Network Threat: This is vulnerable when deduplication takes place at the client side. Attacks like Index Tempering, in which an attacker can perform an attack on the index information that has to be sent by the CSP to the client through the network.

2.2. Classification of Data Deduplication

Broadly, data deduplication is divided into six categories on the basis of location, time, storage location, data type, and implementation [43-45].

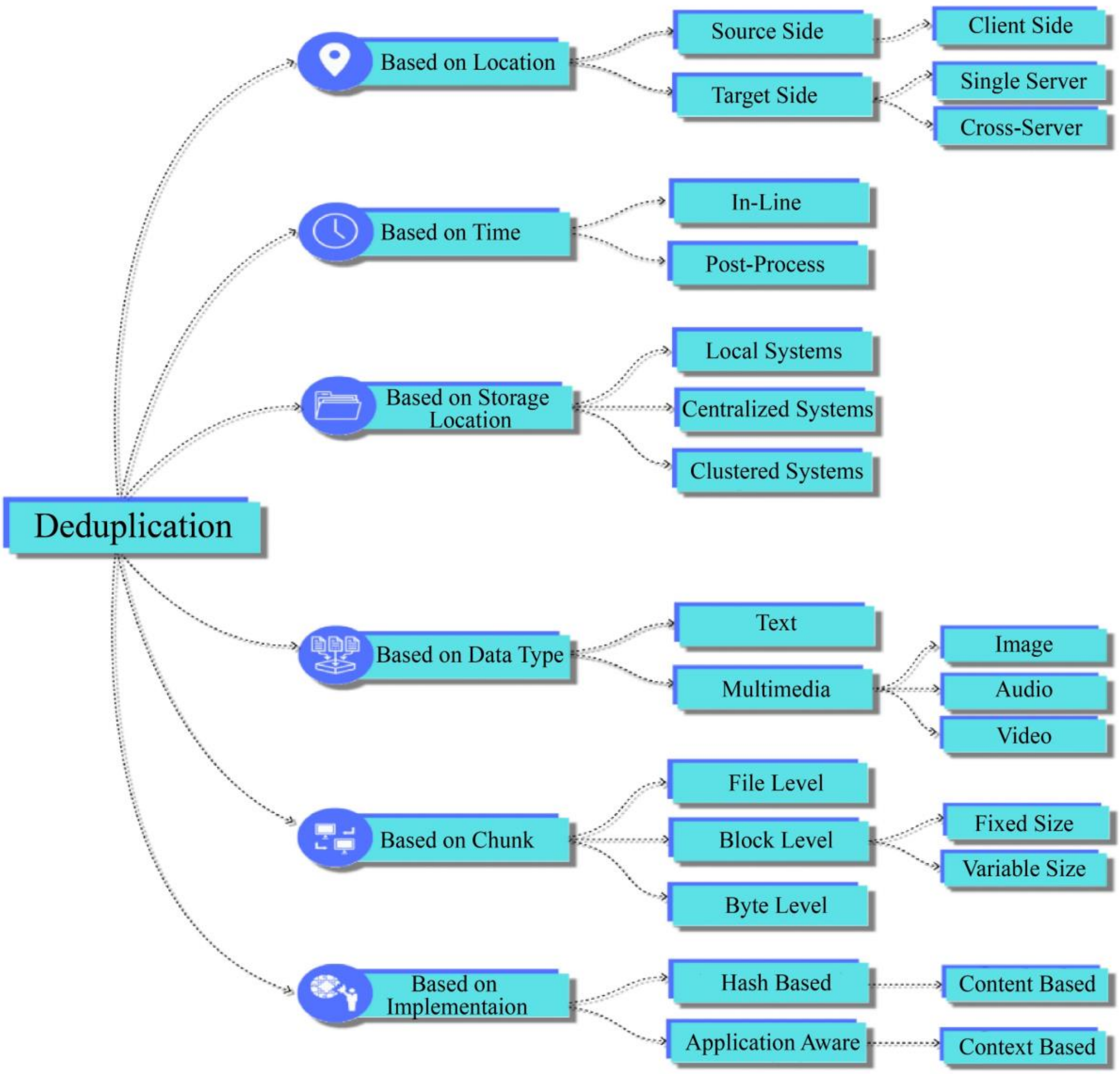

Fig. 10 Classification of data deduplication

These categories are further divided into subdivisions, which are shown in Figure 10. Based on the location, it refers to which side (Client side/ Server side) the deduplication has to be performed. By time means when deduplication has to be performed, whether during the transfer of data from client to server or after the data has been transferred to the server completely. Storage location tells about the location of the deduplication process, like at the local site or at some other sites. Data type refers to the format of the file on which the deduplication has to be applied. The file could be text, image, audio, video, etc. Implementation refers to the process of executing deduplication in reference to the content or context of data. The detailed description of these categories is given further. Tabular comparison is done between the classified categories on the basis of some appropriate performance metrices like bandwidth, storage, throughput, deduplication ratio, efficiency, cost, risk of data loss, index overhead, granularity, processing time, and metadata overhead.

2.2.1. Deduplication with Respect to Location

On the basis of location, deduplication is divided into two types. These are deduplication at the source side and the target side. In source-side deduplication, identical copies are removed at the client side, i.e., before broadcasting data to the target machine. Here, bandwidth is reduced as only unique data is transferred. The hardware requirement is decreased, but processing of resources is increased at the client side [46-51]. In Deduplication at the target side, identical copies are removed at the server, i.e., after the broadcasting of data at the destination. Bandwidth has increased, but performance is also enhanced when compared to deduplication at the source side [52-55]. Deduplication at the target side can further take place at a single server or between cross-server, depending on the type of file. The diagrammatic representation of client and server-side deduplication is shown in Figures 11 and 12, respectively. Table 2 presents a comparison between them in terms of bandwidth, storage, throughput, deduplication ratio, efficiency, and cost [49, 56].

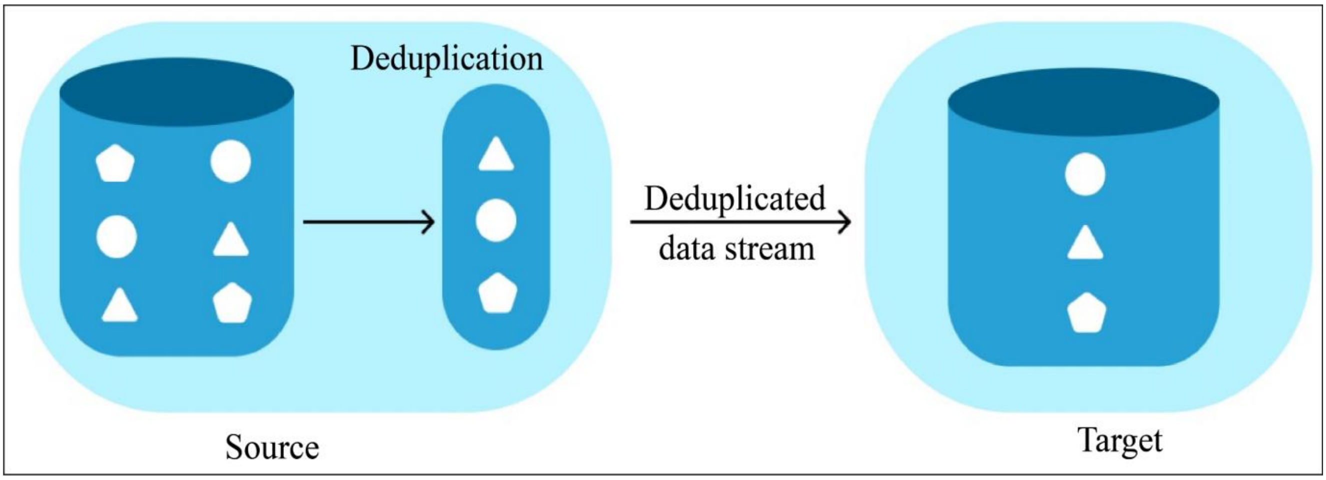

Fig. 11 Source / Client Side deduplication

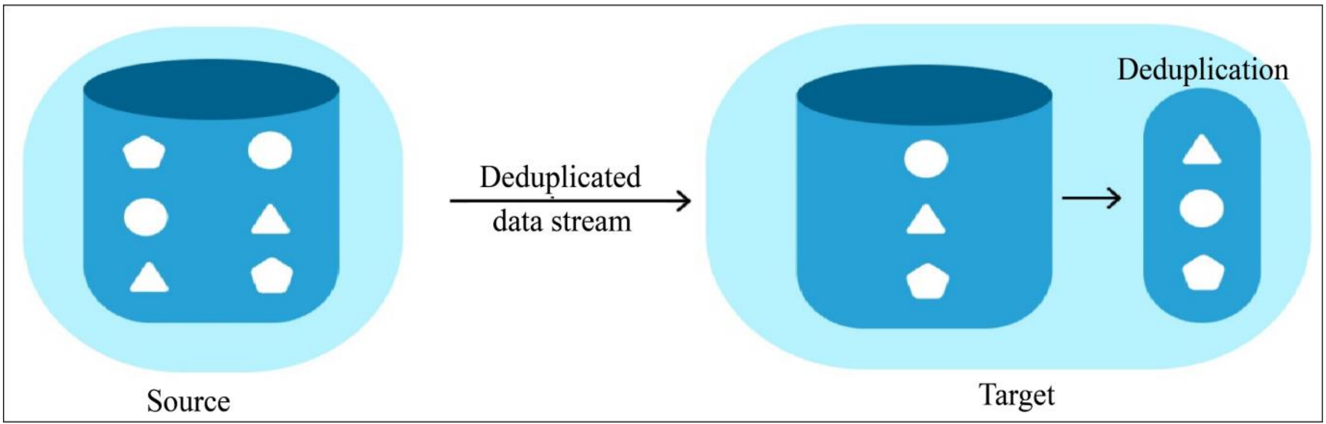

Fig. 12 Target/Server-side deduplication

Table 2. Comparison of performance metrics of deduplication in reference to location

Parameters	Bandwidth	Storage	Throughput	Deduplication Ratio	Efficiency	Cost
Source Side	Low	Nominal	Average	Average	Nominal	Less
Target Side	High	High	Average	Average	Nominal	More

2.2.2. Deduplication with Respect to Time

On the basis of time, deduplication is divided into two types. These are inline deduplication and post-process deduplication. In inline deduplication, it takes place before being written to the disk at the client side or while transferring data from the client to the server. Deduplicated data is being transferred to the server, as deduplication has already taken place at the client side, so the network overhead is reduced [57-59]. In post-process deduplication, deduplication takes place after writing to the disk at the server side. Here, the whole data is stored on the server, which contains duplicated copies as well, and then the duplicated copies are removed. In this, more disk space is required as compared to inline deduplication [52, 54]. The process of inline and post-process deduplication is shown in Figures 13 and 14, respectively. The comparison of performance metrices between them is shown in Table 3.

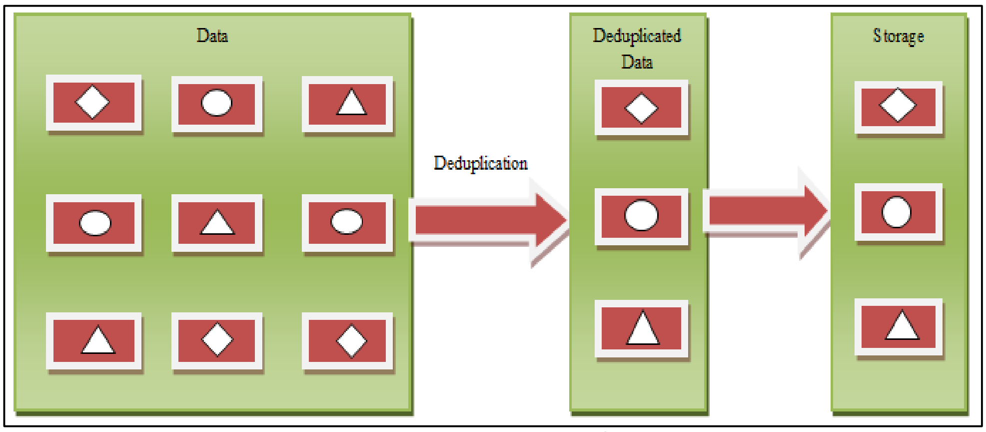

Fig. 13 Inline deduplication of data

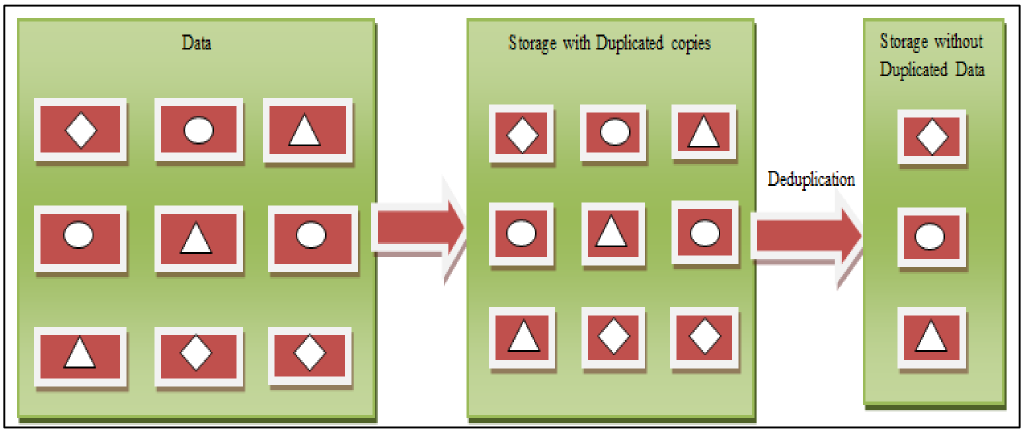

Fig. 14 Post-Process deduplication of data

Table 3. Comparison of performance metrics of deduplication in reference to time

Parameters	Bandwidth	Storage	Throughput	Deduplication Ratio	Efficiency	Cost
Inline	Low	Less	Low	Low	Nominal	Less
Post-Process	High	More	Nominal	High	High	More

2.2.3. Deduplication with respect to Storage Location

On the basis of storage location, deduplication is divided into three types. These include storage at local systems, storage at centralized systems, and storage at clustered systems. In a local system-based type of deduplication, the deduplication process, including chunking, hashing, indexing, and storage of data, took place at a single system, as shown in Figure 15. The main motto behind this is to maintain a tradeoff between efficiency and computation overhead [60-63]. In centralized storage deduplication systems, unique data is stored at one central server for achieving high efficiency and for managing resources, as shown in Figure 16. As data is stored on a single server, there are chances of data loss, which is a main drawback of this type [64-66]. For overcoming the data loss situation due to a server crash, the data is stored at distinct locations by distributing the data among them, as shown in Figure 17. This type of deduplication is known as cluster-based deduplication. But this system faces the problem of load balancing and consistency among nodes [67-69]. The comparison is done on the performance metrices like risk of data loss, cost, efficiency, and throughput, and is shown in Table 4.

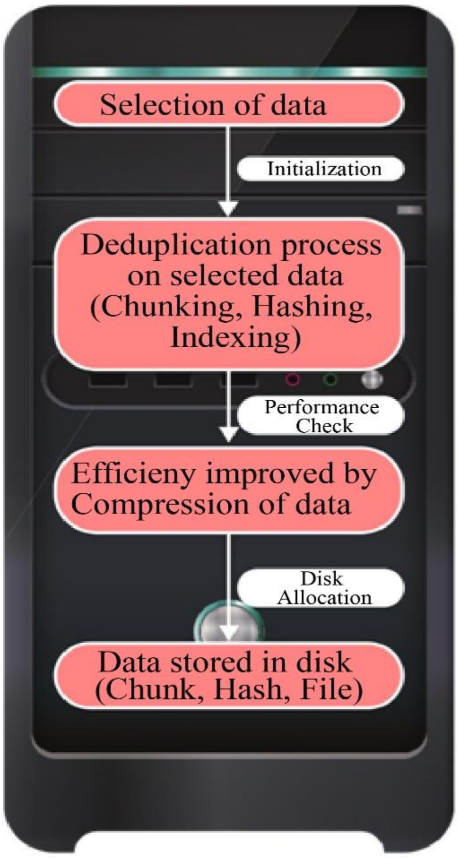

Fig. 15 Local system-based deduplication

2.2.4. Deduplication with Respect to Data Type

On the basis of the type of data on which deduplication has to be performed, deduplication is divided into two types. These are Text and Multimedia. Multimedia includes Images, audio, and videos. Different deduplication techniques are required for each data type because every type has a different file format. The file format plays a crucial role in reading, writing, and executing files. The similarity index of these data types is checked to ensure the quality, such as luminance, contrast, and structure, and then deduplication is done. Encryption methods are used for storing this type of deduplicated big data with efficiency [70-77]. The comparison of performance metrices like bandwidth, storage, deduplication ratio, and cost is shown in Table 5.

2.2.5. Deduplication with respect to Chunks

On the basis of chunks being divided, deduplication is divided into three levels. These are single instance storage, i.e., file-level chunking, block-level, and byte-level chunking. Block level is further divided into fixed-size chunking and variable-size chunking. At the file level, processing takes place on the complete file at a single time. The index value is calculated for the complete file, and this index value is compared with the existing index table to find duplicates. As there is a single value for each file, the entries in the index table are very few, due to which the storage space required is also less, as shown in Figure 18. If there is any modification in the file, then the unique index is again generated for the complete modified file, which should be generated for modified data only, rather than the complete file, due to which efficiency is decreased [47,55,59,78,79]. To overcome this problem, the file is being divided into chunks known as block-level chunking. In fixed-size chunking, the whole file is divided into various equal-sized chunks, as shown in Figure 19. An index value is generated for the individual block and saved in the index table. So, when there is any modification in the file, then only the index value for that particular block is calculated again, rather than for the complete file. Here, the entries in the index table are more, and hence the size of the table is enlarged, due to which memory consumption is increased [53,78,80]. In variable-size chunking, the whole file is not divided into equal-sized chunks, but chunks could be of any different sizes, and this size depends on the content of the data, as shown in Figure 20. Data boundaries shift when there is a change in data according to the need, which has not been possible earlier in fixed-size chunking [81-83]. In byte-level chunking, chunks are divided on the basis of their byte value, as shown in Figure 21. The index value of each byte is generated and compared with the existing values in the index lookup table. If there is any modification in any chunk, then only the byte value of that chunk requires change [84-87]. The comparison done in terms of deduplication ratio, index overhead, processing time, efficiency, memory requirement, and throughput is shown in Table 6.

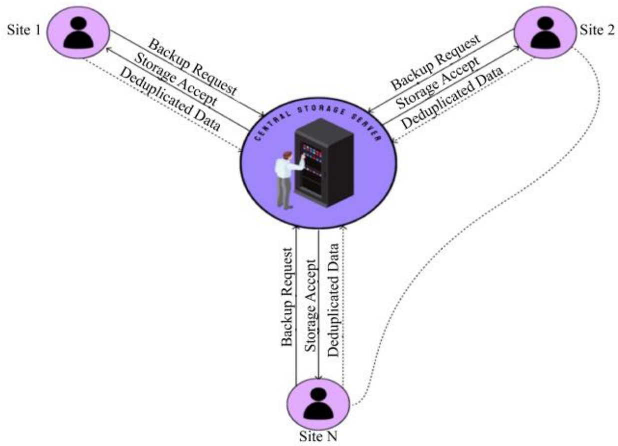

Fig. 16 Central system-based deduplication

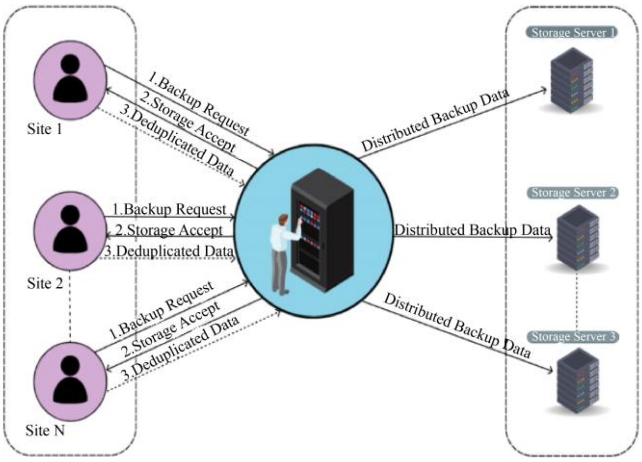

Fig. 17 Clustered system-based deduplication

Table 4. Comparison of performance metrics of deduplication according to storage location

Parameters	Data Loss's Risk	Cost	Efficiency	Throughput
Local Systems	Most	Less	Low	Less
Centralized Systems	More	More	High	More
Clustered Systems	Less	Most	Highest	Most

Table 5. Comparison of performance metrics of deduplication according to data type

Parameters	Bandwidth	Storage	Deduplication Ratio	Cost
Text	Least	Lower	Most	Least
Image	Less	Low	More	Less
Audio	More	High	Less	More
Video	Most	Highest	Least	Most

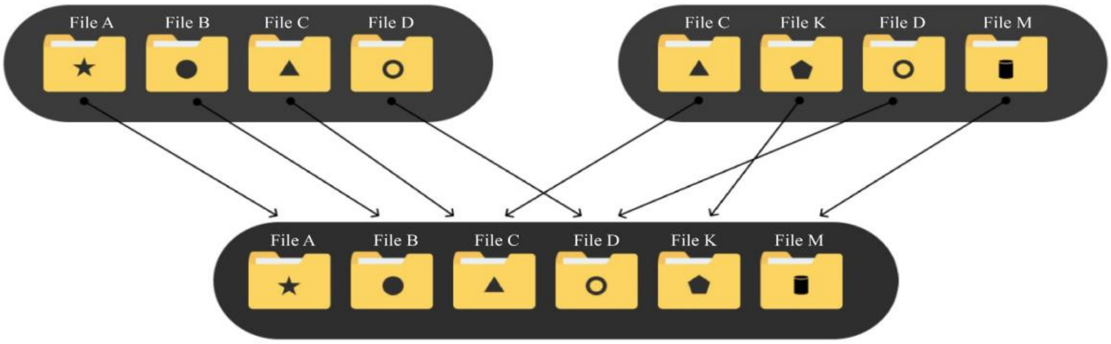

Fig. 18 File-level deduplication

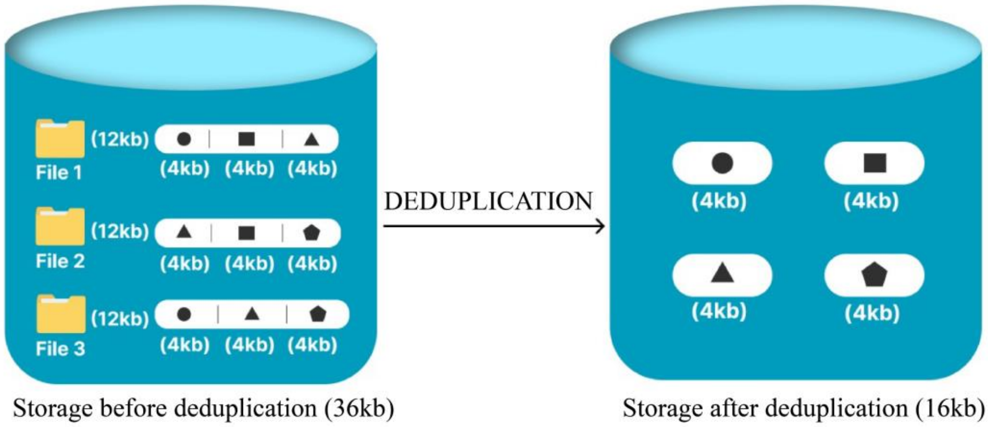

Fig. 19 Fixed block-level chunking-based deduplication

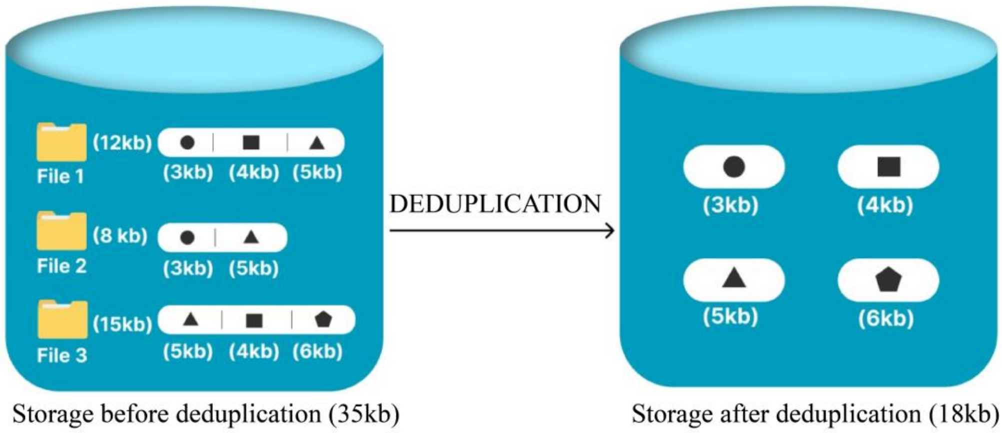

Fig. 20 Variable block level chunking based deduplication

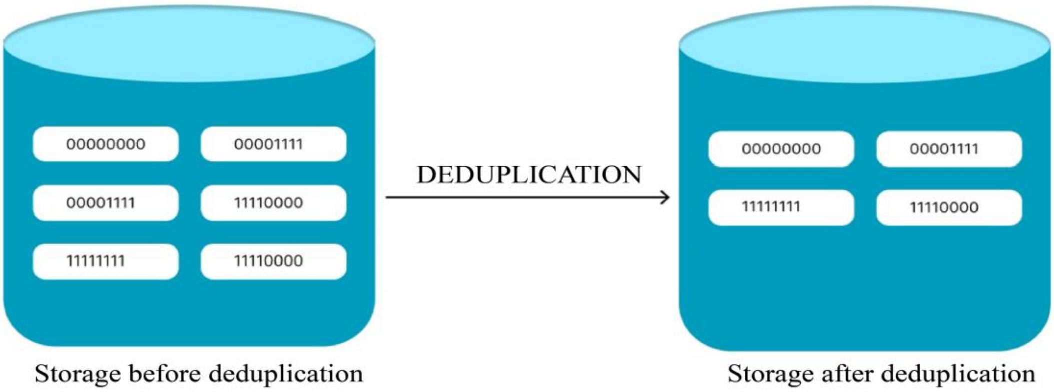

Fig. 21 Byte-level chunking-based deduplication

Table 6. Comparison of performance metrics in reference to chunking methods

Parameters	Deduplication Ratio	Index Overhead	Processing Time	Efficiency	Memory Requirement	Throughput
File Level Chunking	Less	Best	Medium	Less	More	Least
Fixed Block Level Chunking	Medium	Worse	Less	Moderate	Moderate	Less
Variable Block Level Chunking	High	Worse	High	More	Less	More
Byte Level Chunking	Highest	Worst	Highest	Most	Least	Most

2.2.6. Deduplication with respect to Implementation

On the basis of implementation, deduplication is divided into two types. These are hash-based (content-based) and application-aware (context-based). In content-based based, a hash value is calculated and compared for the data content. Hashing techniques like MD5, SHA-256, and SHA-512 are used for calculating hash values [88-92]. This process is shown in Figure 22. In application-aware deduplication, data is considered as an object, and during the deduplication process, only similar types of objects are being compared, like a Word file given as input is being compared with the existing Word files only. Here, deduplication takes place at the byte level. Only the unique bytes of the object are saved into the disk [32,93,94]. This deduplication process is shown in Figure 23. The comparison of performance metrices like granularity, processing time, efficiency, metadata overhead, and cost is done in Table 7.

3. Discussion and Analysis

In this section, the deduplication techniques are discussed. The type of deduplication used in various papers is shown in Table 8, which clearly depicts that a deduplication technique can be a part of more than one type of classification. For example, the technique employed in paper [95] uses target side, clustered, block level, and hash-based deduplication on the basis of location, storage location, chunking used, and implementation, respectively. This section will answer questions like the rate at which the duplicated data is removed by the technique. How much memory space is required to store the data? How much time does the technique require for computing the necessary computations to remove redundancy? How much data is found to be duplicated? So, for analysis, comparison of some deduplication techniques is done on the basis of parameters like computation time, efficiency, deduplication rate, throughput, and memory consumption. In reference to deduplication, computation time refers to the total time required by the technique to perform all the necessary computations needed in the deduplication process. It is measured in seconds. Efficiency refers to the ability of the deduplication technique to remove redundant data. It is calculated as [(original data-removed data)/original data] * 100. Deduplication rate refers to the rate at which the amount of data is deduplicated with respect to the total given data. Throughput refers to the amount of duplicated data removed in a particular amount of time (Amount of data/ given time). It is calculated as MB/Sec. Memory consumption refers to the storage space required by the technique for processing and saving data. It is measured in MB. These techniques include DupLESS [46], Boafft [53], REBL [61], H.P. Dedup [62], P-Dedup [64], MMSD [66], ∑-Dedupe [68], S.L. [69], App-Dedup [96], MECC [97], SDD [98], BDKM [99], MUUE [100], ISFDA [101], SS [102], SSIMI [103], SLDF [104], R-Dedup [105], PAKE [106], RCE [107], Sim-Dedup [108], NIDF [109], Rev-Dedup [110], DIODE [111], DEDIS [112] and RMD [113]. The performance comparison graphs of these techniques are shown in Figures 24-28.

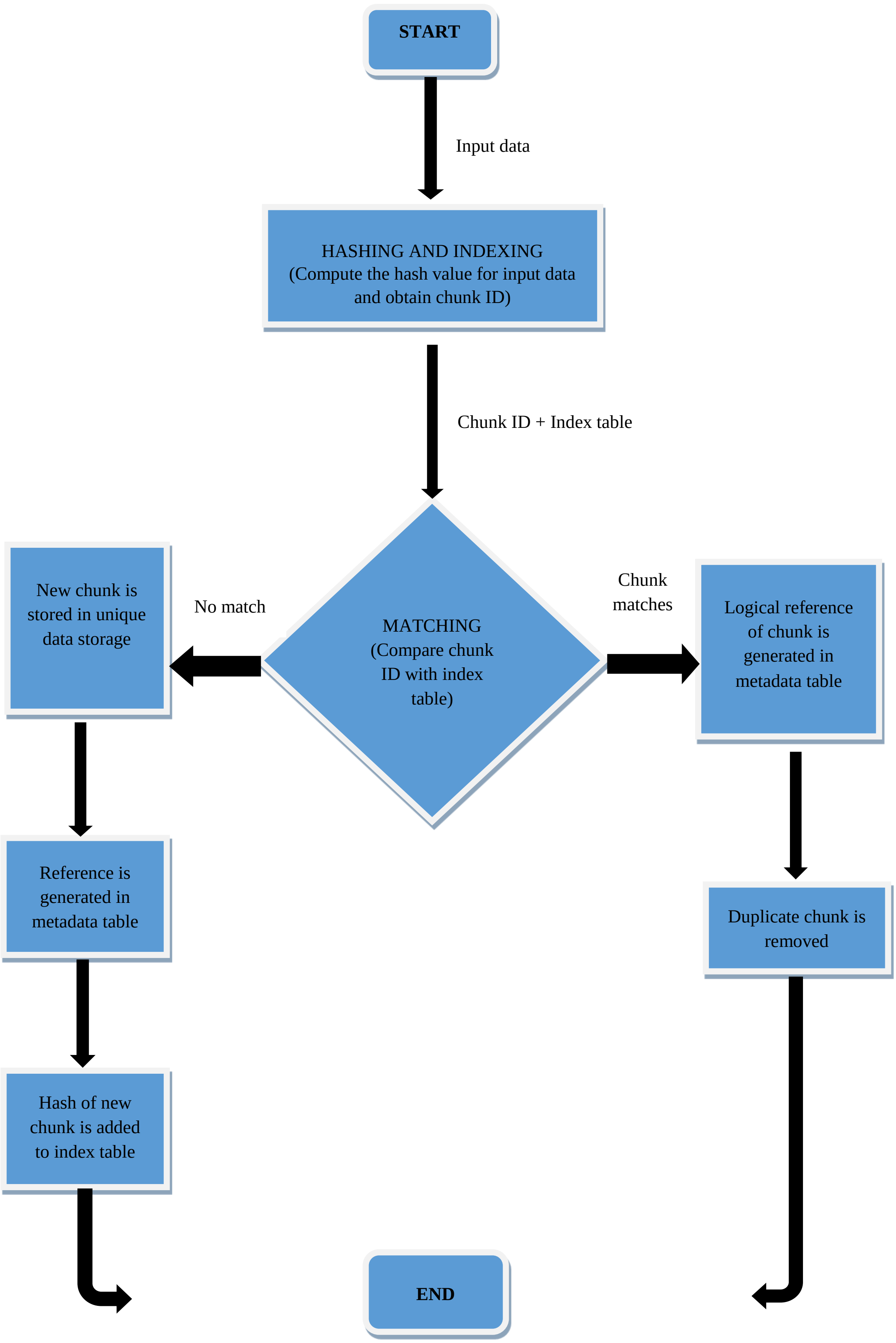

Fig. 22 Hash/Content-based deduplication

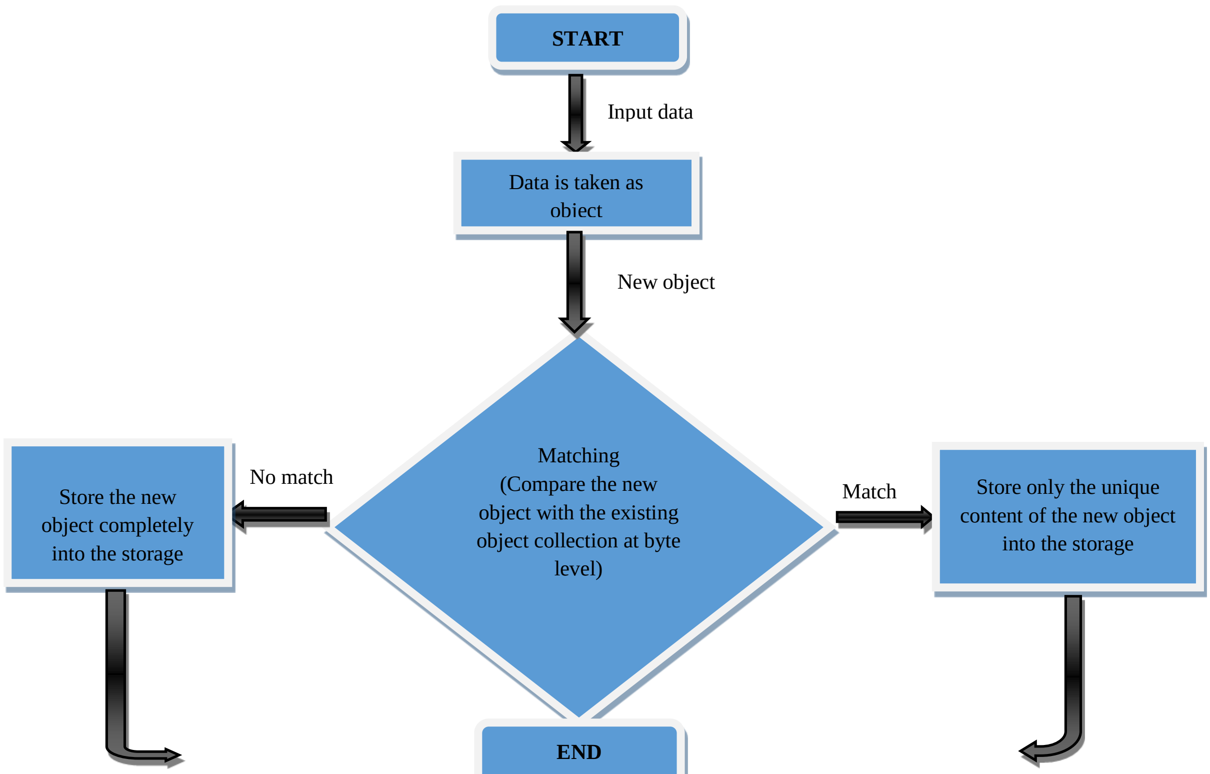

Fig. 23 Application/ Content-aware based deduplication

Table 7. Comparison of performance metrics of deduplication in reference to implementation

Parameters	**Granularity**	**Processing Time**	**Efficiency**	**Metadata overhead**	**Cost**
Hash Based	Chunk/ Block level	Slow	Less	Only basic details required	Less
Application-aware	Byte level	Fast	More	Additional metadata required	More

Table 8. Type of deduplication used in various papers (Y-> Yes and N-> No)

			[50]	[95]	[114]	[115]	[116]	[65]	[117]		[118]	[119]	[120]	[121]	[112]	[123]	[96]
Deduplication	Based on Location	Source Side	Y	N		N									N	Y	
		Target Side	N	Y		Y		Y	Y				Y	Y	Y	N	
	Based on Time	Inline			Y							Y			Y	Y	Y
		Post-process			N	Y										N	N
	Based on Storage Location	Local Systems		N			Y										
		Centralized Systems		N	N			Y	N		N	N					
		Clustered Systems		Y	Y				Y		Y	Y	Y				Y
	Based on the Data type	Text									Y			Y			
		Multimedia	Y				Y	Y				Y	Y	Y			Y
	Based on Chunk	File Level	Y	N		Y		Y					N	Y	N		
		Block Level	N	Y	Y	Y		Y	Y		Y		N	Y	N	Y	Y
		Byte Level	N	N								Y		Y	Y		
	Based on Implementation	Hash Based		Y	Y	Y					Y					Y	N
		Application-aware		N												N	Y

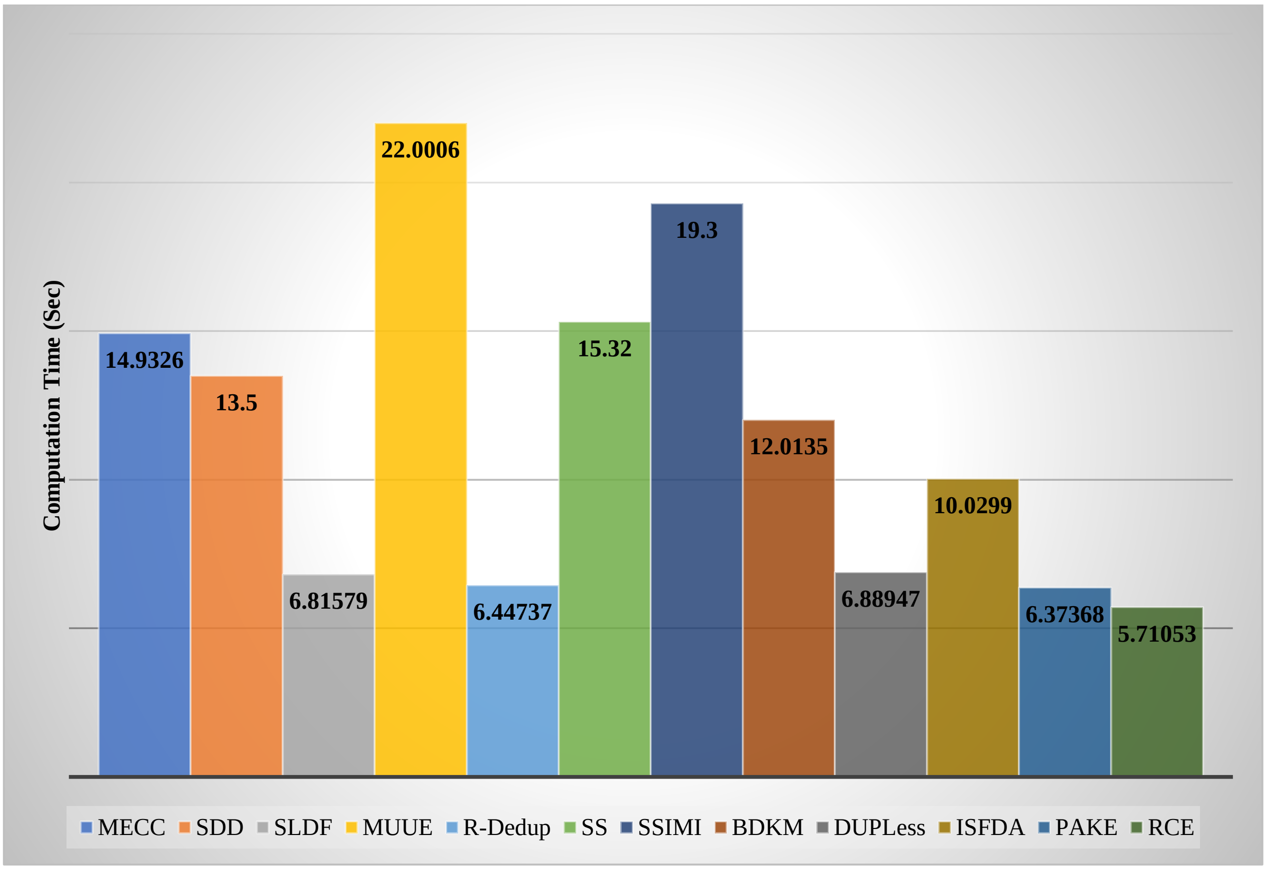

Fig. 24 Comparison graph of deduplication techniques on the basis of computation time

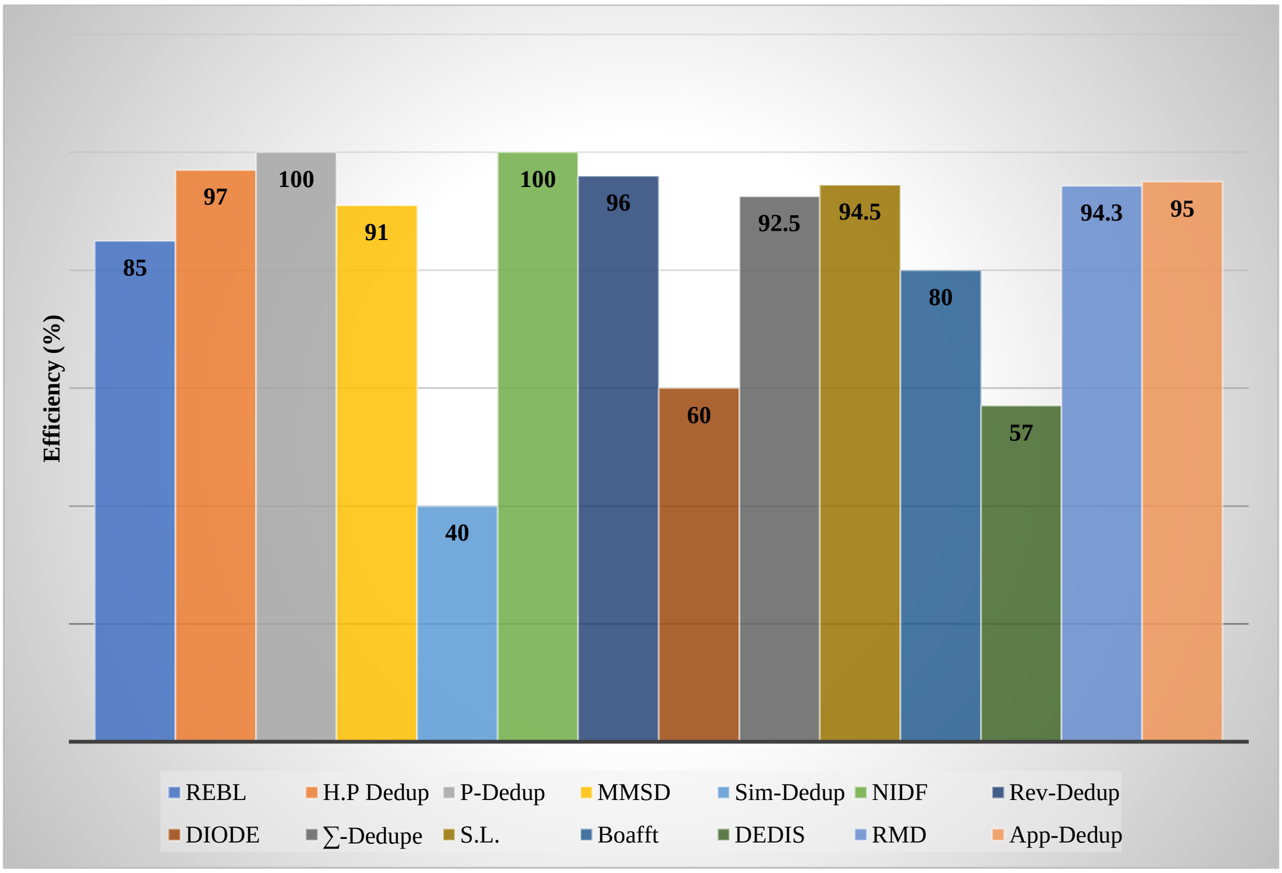

Fig. 25 Comparison graph of deduplication techniques on the basis of efficiency

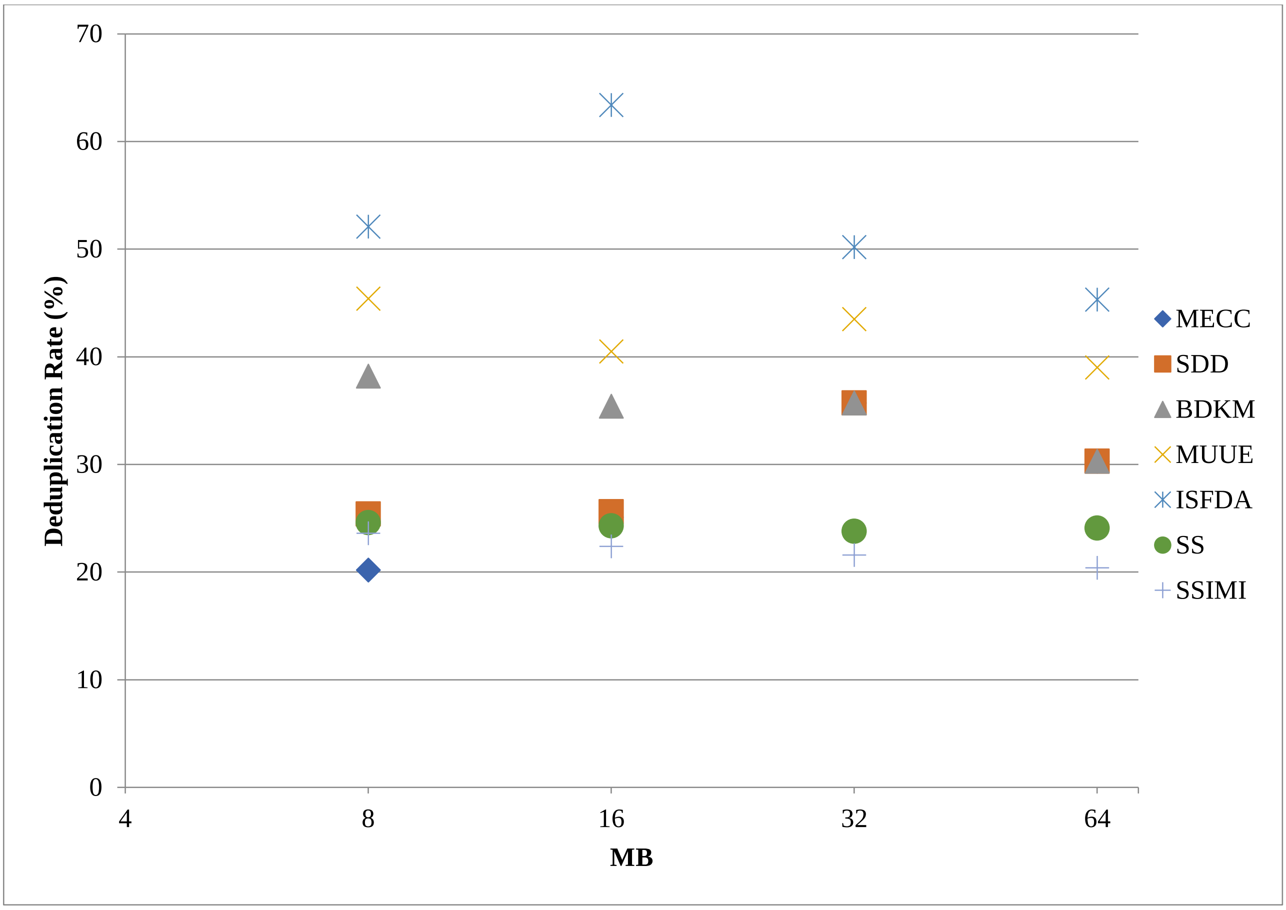

Fig. 26 Comparison graph of deduplication techniques on the basis of deduplication rate

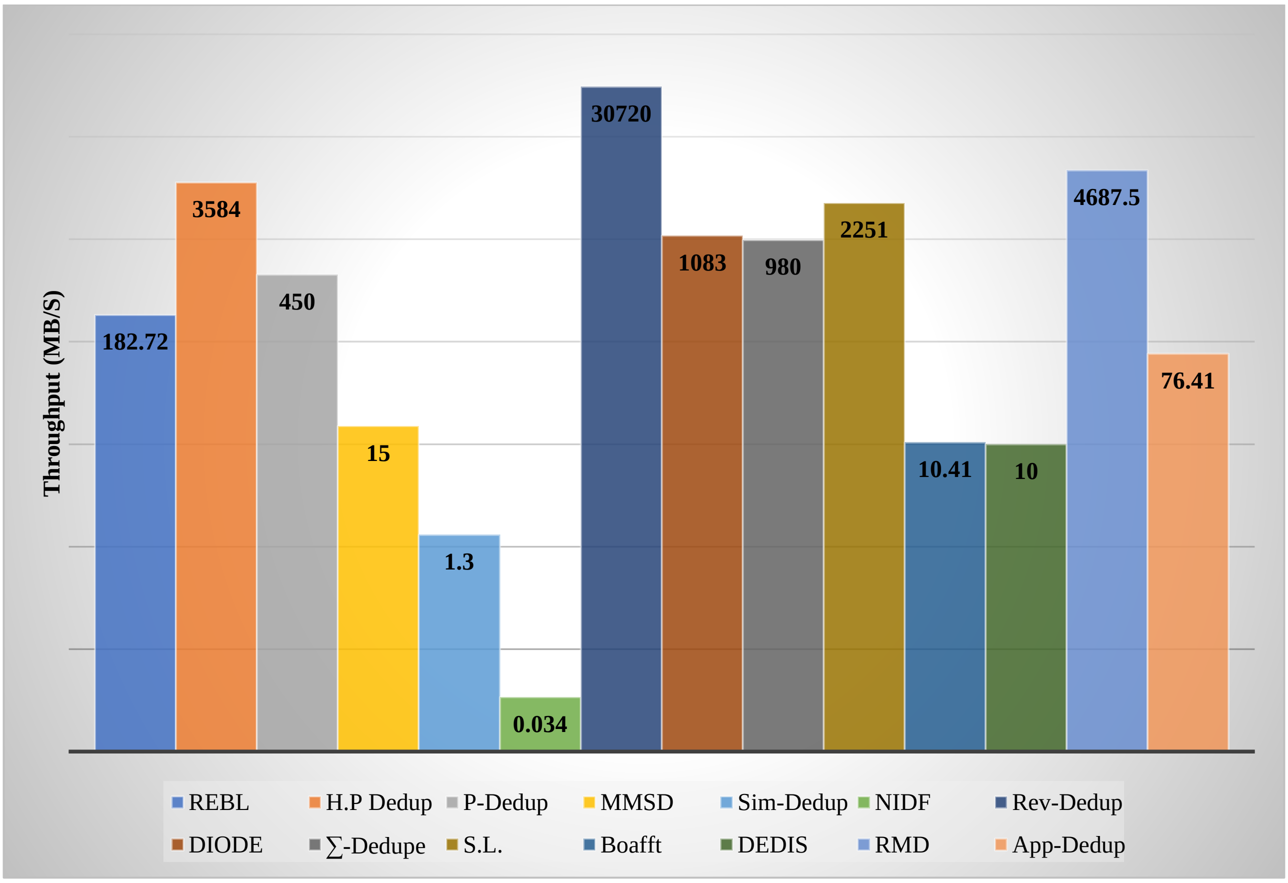

Fig. 27 Comparison graph of deduplication techniques on the basis of throughput

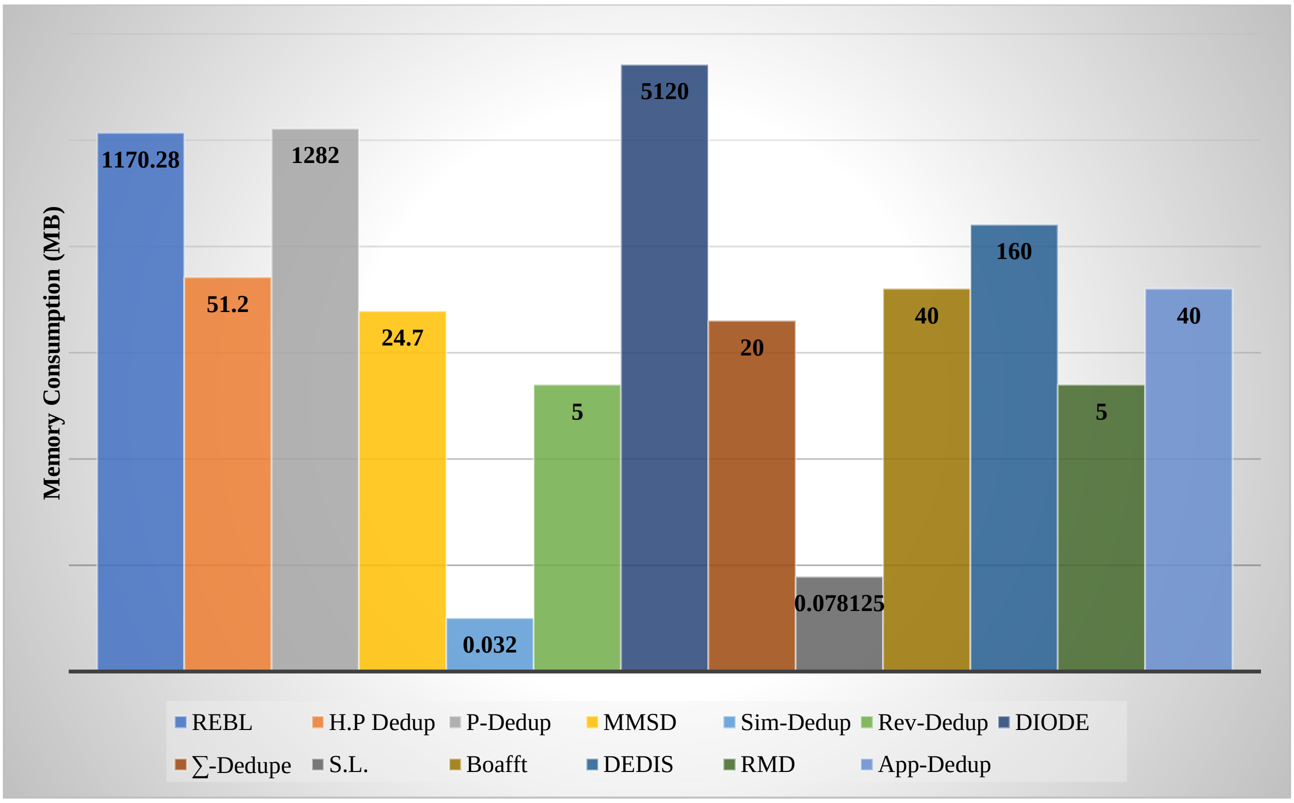

Fig. 28 Comparison graph of deduplication techniques on the basis of memory consumption

It is evident from Figure 24 that the technique MUUE [100] has taken the most computation time, while the technique RCE [107] has taken the least computation time for the 64 MB data file. Figure 25 shows that the techniques P-Dedup [64] and NIDF [109] achieve the best efficiency, i.e, 100% while the worst efficiency is achieved by the technique Sim-dedup [108]. It is clear from Figure 26 that the technique ISFDA [101] has the highest deduplication rate for all the 8MB, 16MB, 32MB, and 64 MB datasets. The technique MECC [97] has the lowest deduplication rate for the 8MB dataset, while for the remaining 16MB, 32 MB, and 64 MB datasets, technique SSIMI [103] has the lowest deduplication rate. It is illustrated in Figure 27 that the technique Rev-Dedup [110] achieved the highest throughput, and the lowest throughput is achieved by the technique NIDF [109]. The highest memory is consumed by the technique DIODE [111] while the technique Sim-Dedup [108] consumes the lowest memory, as shown in Figure 28. So, it is clear from the analysis that none of the techniques is best or worst in terms of each performance matrix. Every technique has some merits and demerits. For selecting the appropriate technique, it depends on the decision maker which parameters have to be given preference. The summarized description of these deduplication techniques is given in Table 9.

Table 9. Tabulation summary of various deduplication techniques

Year	Authors	Proposed Technique	Description	Findings	Performance
2020	Menon et al. [97]	MECC (Modified Elliptic Curve Cryptography)	• The MECC system is designed for integrated cloud-edge networks • Uses SHA for hashtag generation • File encrypted using CE (Convergent Encryption), like RSA, and re-encrypted using MECC.	Enhances security by 96% and performance of the system in a fog environment. In the future, this model can be implemented for IoT (Internet of Things) applications and cyber-physical systems.	Computation Time= 14.9326sec Deduplication Rate 20.2% (8MB), 25.1%(16 MB), 35.7% (32 MB), 30.3% (64MB) (Best)

2021	Ebinazer et al. [98]	Secure Data Deduplication (SDD)	• Secure data deduplication approach • Uses bloom filter and radix tree • Uses authorized deduplication, proof of ownership, role key update, and tag consistency preservation	SDD is a client-side deduplication model with a high deduplication rate. In the future, performance can be enhanced by working on queuing techniques and lightweight cryptographic algorithms.	Computation Time= 13.5 sec Deduplication Rate 25.4% (8MB), 25.6%(16 MB), 35.7% (32 MB), 30.3% (64MB)
2002	Douceur et al. [104]	Serverless Distributed Filesystem (SLDF)	• SLDF is a distributed file system that reclaims space from duplicate files • Uses convergent encryption • Uses Self Arranging, Lossy and Associative Database (SALAD) for collecting file data, like piggybacking in DHCP (Dynamic Host Configuration Protocol)	SLDF enhances the reliability and security by storing duplicated files at multiple (585) sites. Results show that this system of coalescing files is scalable, effective, and also fault-tolerant. This system is outdated, but it becomes necessary to study it for a better understanding of other systems.	Computation Time= 6.81579 sec
2021	Wang et al. [100]	Multi-User Updatable Encryption Scheme (MUUE)	Secure deduplication scheme that supports revocation for unauthorized users Uses multi-user updatable encryption and a binary tree for management of group key	MUUE achieves high efficiency and security by reducing memory space on the cloud storage, as communication and computing cost is reduced.	Computation Time= 22.0006 sec Deduplicaion Rate= 45.4% (8MB), 40.5% (16 MB), 43.5% (32 MB), 39% (64MB)
2020	Guo et al. [105]	Randomized Deduplication (R-Dedup)	R-Dedup is randomized, cross-user, and secure client-side deduplication without any additional cloud server. Uses ELGamal encryption, SHA-256, and bilinear mapping.	R-Dedup is a lightweight deduplication system that is not dependent on any third party. Achieves high security and data integrity. Low computation overhead at the client side. Resistant to brute force attacks from both the cloud server and users.	Computation Time= 6.44737 sec
2018	Singh et al. [102]	Secret Sharing scheme (SS)	• SS is secure data deduplication, which uses a secret sharing scheme over the cloud.	This secure data deduplication scheme resolves the problem of fault tolerance. Manages keys reliably and	Computation Time= 15.32 sec Deduplicaion Rate= 24.6% (8MB),

			• Uses Permutation Ordered Binary (POB) for data distribution • Uses Proof of Ownership (PoW) and Chinese Remainder Theorem (CRT) for overhead minimization	efficiently and achieves data confidentiality.	24.3%(16 MB), 23.8% (32 MB), 24.1% (64MB)
2017	Liu et al. [103]	SSIMI	A based data deduplication scheme that uses an integration of both MLE and BL-MLE techniques. A new data-defined algorithm is given.	This similarity-based data deduplication scheme achieves the desired security against PVD$-CDA (Privacy chosen distributed attacks) by maintaining a tradeoff between security and computing overhead.	Computation Time= 19.3 sec Deduplicaion Rate= 23.6% (8MB), 22.4%(16 MB), 21.6% (32 MB), 20.4% (64MB)
2022	Zhang et al. [99]	BDKM	Secure distributed deduplication approach that uses blockchain and reliable key management Uses Message Locked Encryption (MLE) and Ramp Secret Sharing Scheme (RSSS) Uses Merkle Hash Tree (MHT) for PoW and SHA-256 (Secure Hash Algorithm)	BDKM achieves data confidentiality at the file level and block level deduplication. Reliability is increased by using a secret share scheme. Limited computation overhead. Resists brute force attacks and collisions. In the future, a study on implementing integrity verification on data deduplication without getting any information about the data can be done by using blockchain.	Computation Time= 12.0135 sec Deduplicaion Rate= 38.2% (8MB), 35.4%(16 MB), 35.7% (32 MB), 30.3% (64MB)
2013	Keelveedhi et al. [46]	Duplicate less (DupLESS) Encryption for Simple Storage	• Server-aided encryption for deduplication storage • Provides secure deduplication for storage, avoiding brute-force attacks • Uses message-level encryption	DupLESS enhances the performance, along with saving space, which is nearly the same as using the storage service only with plain text data. Easy to deploy on any other storage interface. Provides confidentiality, but it faces problems in controlling the data access of other users' data in a better manner.	Computation Time= 6.88947sec
2022	Mangeshkumar et al. [101]	Improved Secure File Deduplication	• An improved secure file deduplication avoidance, which	An ISFDA is implemented for both block-level and file-	Computation Time= 10.0299 sec

		Avoidance (ISFDA)	uses the Chaotic Krill Herd Optimization (CKHO) algorithm for generating a secret key • Uses a deep learning classifier • Uses dynamic perfect hashing and the Advanced Encryption Standard (AES) algorithm	level deduplication. This model removes the duplicates without compromising integrity, and attacks are reduced by 12% for a 50 MB dataset. In the future, this model can be implemented by using blockchain, and that too without using key servers.	Deduplication Rate= 52.1% (8MB), 63.4%(16 MB), 50.2% (32 MB), 45.3% (64MB)
2015	Liu et al. [106]	Password Authenticated Key Exchange (PAKE)	Pake is a deduplication system without additional independent servers Additionally, homomorphic encryption is done at the client-side Single server and cross-user deduplication Uses password-authenticated key exchange	PAKE is resistant to online brute force attacks. Hence, it provides better security without additional independent servers. Failed to find some duplicates. So, a little negative effect is seen with minimum overhead, which uses a proof-of-concept prototype.	Computation Time= 6.37368 sec
2013	Bellare et al. [107]	RCE (Randomized Convergent Encryption)	• Message Locked Encryption (MLE) is introduced, in which the key is derived from the message itself • Uses the Randomized Convergent Encryption (RCE) scheme • Uses correlated secure hash functions and Deterministic Public Key Encryption (D-PKE)	MLE is based on a symmetric encryption scheme, which is designed by keeping in mind the theoretical as well as practical domains. For the practical domain, a ROM (Random Oracle Model) security analysis is done, and for the theoretical domain, connections with deterministic encryption and hash functions are made.	Computation Time= 5.71053 sec (Best)
2004	Kulkarni et al. [61]	Redundancy Elimination at the Block Level (REBL)	Redundancy elimination within a large collection of files at the block level Uses compression, duplicate block suppression, delta encoding, and super fingerprints	REBL enhances the performance by reducing data sizes effectively and efficiently as it focuses on exploiting the relationship among similar blocks. For a future purpose, the effectiveness of REBL can be checked for a new environment, such as	Efficiency= 85% Memory consumption= 1170.28 MB Throughput= 182.72 MB/s

				the Google Gmail system.	
2011	Guo et al. [62]	H.P. Dedup System	• High-performance deduplication system • Uses progressive sampling indexing • Group marked-sweep mechanism • Uses hashing (MD5, SHA) and fingerprinting techniques	H.P. Dedup enhances performance by increasing single-node performance. Scalability and throughput increased by using a multi-threaded environment. In the future, work on the boundary shifting problem can be done.	Efficiency= 97% Memory consumption= 51.2 MB Throughput= 3584 MB/s
2012	Xia et al. [64]	P-Dedup	Exploits pipelining and parallelism in the data deduplication system Uses FSC (Fix Size Chunking) and CDC (Content Defined Chunking) based parallel chunking algorithms Uses parallel fingerprinting algorithms	P-Dedup is a fast and scalable deduplication system. Achieves high deduplication write throughput by a factor of 2~4. For the future, with increasing processor cores, the thread-level parallelism can be exploited.	Efficiency= 100% (Best) Memory consumption= 24.7 MB Throughput= 1282 MB/s
2013	Meng et al. [66]	Metadata-Aware Multi-Tiered Source Deduplication (MMSD)	Cloud system designed for a personal computing environment Shorter backup window Uses WFC (Whole File Chunking) policy of file size< 1 MB	MMSD achieves an optimum tradeoff between efficiency and storage overhead of just 33.8%. For the future, semantic-based multi-tiered source deduplication can be designed within a linux environment.	Efficiency= 91% Memory consumption= 24.7 MB Throughput= 15.11 MB/s
2013	Yao et al. [108]	Sim-Dedup	A deduplication scheme based on Simhash Exploits file similarity and chunk locality Uses SHA-1 hash function, CDC (Content Defined Chunking) algorithm on 4 KB, 250 MB segment, and in-memory cache	Simdedup tried to maintain high deduplication throughput and low system computation overhead, but this results in very little throughput. The plus point with this scheme is that the memory required is very little.	Efficiency= 40% Memory consumption= 24.7 MB (Best) Throughput= 0.032 MB/s
2014	Madhubala et al. [109]	Nature-Inspired Data Deduplication Framework (NIDF)	• Nature-inspired enhanced data deduplication framework, which uses text matching algorithms like SM (Sequence	The framework provides an efficient and reliable system for identifying duplicates. 100% efficiency is achieved, but the	Efficiency= 100% (Best) Throughput= 0.034 MB/s

			matching), LA (Levenshtein Algorithm) for text comparison • Uses Genetic Programming for the closest matching	throughput is very low. In the future, this work can be extended to other file formats and can be used to find out the dual possession of Ration Cards and employment cards, etc.	
2014	Li et al. [110]	Reverse Deduplication (Rev Dedup)	A hybrid of inline and outline deduplication for backup storage Uses the CDC algorithm Uses 2-level reference management for tracking chunks, multi-threading, and prefetching	Aims for high performance in backup, restore, and low deletion overhead for expired backups. Efficiency and Throughput achieved are very high, but need extra Input/outputs for identification and removal of chunks.	Efficiency= 96% Memory consumption= 5 MB Throughput= 30720 MB/s (Best)
2016	Tang et al. [111]	Dynamic Inline-Offline Deduplication (DIODE)	DIODE works at 2 deduplication phases, i.e, the inline deduplication phase and offline deduplication phase Uses CTA (Context-aware Threshold Adjustment) for inline and DPE (Deferred Priority-based Enforcement) for offline deduplication Uses SHA-1 and CDC algorithm (4 KB)	DIODE achieves better read/write performance and space saving for primary storage systems as compared to other conventional schemes. In the future, a compatible index structure with high lookup efficiency can be designed, and machine learning techniques can be implemented for adjusting parameters.	Efficiency= 60% Memory consumption= 5120 MB Throughput= 1083 MB/s
2012	Jiang et al. [68]	∑-Dedupe	A scalable inline cluster deduplication framework for big data protection Uses a similarity-based data routing algorithm Chunk fingerprint caching Parallel container management	∑-Dedupe achieves an optimal tradeoff between parallel cluster deduplication efficiency and scalability with low overhead and low RAM usage. Enhances the significance of stateless and extreme binning in this framework.	Efficiency= 92.5% Memory consumption= 20 MB Throughput= 980 MB/s
2013	Zhang et al. [69]	Similarity Locality (S.L.)	Similarity locality approach based on cluster data deduplication Similarity in data and Locality of	S.L. removes duplicacy in nodes by using bloom filters, which exchange necessary	Efficiency= 94.5% Memory consumption= 0.078125 MB

			data are used for finding deduplication between nodes Uses the bloom filter algorithm, which stores the fingerprints of data	information between nodes. In the future, the performance of S.L. can be enhanced by removing the false positive rate of bloom filters.	Throughput= 2251 MB/s
2015	Luo et al. [53]	Boafft	Distributed deduplication for big data storage in the cloud Modification in HDFS(Hadoop distributed file system) Uses the MinHash function and the similarity routing algorithm for finding two similar blocks	The Boafft uses multiple data servers in parallel for deduplication of data, achieving scalable throughput and efficiency with an insignificant loss of deduplication proportion. Also results in a good load balance.	Efficiency= 80% Memory consumption= 40 MB Throughput= 10.41 MB/s
2016	Paulo et al. [112]	Dependable and Decentralized System (DEDIS)	A dependable and decentralized system that performs offline deduplication on clusters Uses DF (Duplicate Finder) and GC (Garbage Collector) algorithm (4KB) Uses the SHA-1 hashing function and in-memory cache	DEDIS is free and open source, and it does not depend upon local data assumptions. Minimizes deduplication overhead and requires acceptable memory consumption, but very little throughput in the primary storage cloud infrastructure.	Efficiency= 57% Memory consumption= 160 MB Throughput= 10 MB/s
2017	Zhang et al. [113]	Resemblance and Mergence-based Indexing (RMD)	• Resemblance and mergence-based scheme • Uses Dynamic Bloom Filter Array (DBA), resemblance algorithm, Bin Address (BA) tables • Uses frequency-based fingerprinting and RAM hit table	RMD is a fast deduplication scheme that speeds up the performance of the fingerprinting index and minimizes the need for RAM during this process, but it wastes some memory on storage of redundant data.	Efficiency= 94.3% Memory consumption= 5 MB Throughput= 4687.5 MB/s
2017	Fu et al. [96]	AppDedupe	Application-aware big data deduplication in cloud environment Uses an inline distributed deduplication architecture Uses Two Threshold Two Divisor (TTTD) chunking algorithm and two-tiered data routing scheme	AppDedupe provides a scalable, inline, distributed deduplication architecture. Enhances the efficiency by using application awareness, data similarity, and locality. Provides very low overhead in cluster-based deduplication.	Efficiency= 95% Memory consumption= 40 MB Throughput= 76.41 MB/s

4. Conclusion and Future Scope

Cloud computing and data deduplication are the hottest topics in today's world. The unlimited fast generation of digital data leads to the increased demand for efficient storage systems in cloud computing, which further increases the demand for data deduplication. After the evolution of data deduplication, cloud computing is booming like a fire in a forest. In this paper, an extensive review of cloud computing is done, and the various associated deduplication techniques are discussed. The thorough comparison of deduplication techniques under each classification category on the basis of performance metrices like computation time, efficiency, throughput, deduplication rate, and memory consumption is done. By using deduplication, duplicated or redundant data is removed, which results in increased storage capacity, bandwidth, and cost decreases, resulting in improved efficiency of cloud storage systems. But as everything has pros and cons, deduplication techniques also have some issues. The study revealed that if a technique has high efficiency, then it could have taken high computation time, and vice versa. Likewise, if the time is decreased, then the memory required by the technique could be high. Based on the study of existing deduplication techniques, it has been observed that deduplication technique on cloud storage systems has a great potential for research in the future. There is still plenty of scope for improvement in existing deduplication techniques. It has been concluded that a deduplication technique for the cloud storage system should be introduced, which maintains a perfect tradeoff between the discussed performance metrices. The future work in this field can certainly focus on developing a deduplicated cloud storage system that is reliable, secure, scalable, cost-effective, and energy efficient. The future scope associated with some deduplication techniques is depicted individually in the findings of the techniques mentioned in Table 9.

References

[1] Cloud Computing, Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Cloud_computing

[2] Ali Sunyaev, *Cloud Computing*, Internet Computing, Springer, pp. 1-413, 2020. [Google Scholar] [Publisher Link]

[3] Safana Alzide, "Cloud Computing: Evolution, Challenges, and Future Prospects," *Journal of Information Technology, Cybersecurity, and Artificial Intelligence*, vol. 1, no. 1, pp. 52-63, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Aisha Hassan Abdalla Hashim et al., *Cloud Computing's Transformative Power in Computing Environments*, IGI Global Scientific Publishing, pp. 1-538, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Gurmeher Singh Puri, Ravi Tiwary, and Shipra Shukla, "A Review on Cloud Computing," *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, pp. 63-68, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Praveen Borra, "An Overview of Cloud Computing and Leading Cloud Service Providers," *International Journal of Computer Engineering and Technology (IJCET)*, vol. 15, no. 3, pp. 122-133, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Nathaniel Brooks, Corinna Vance, and Dorian Ames, "Cloud Computing: A Review of Evolution, Challenges, and Emerging Trends," *Journal of Computer Science and Software Applications*, vol. 5, no. 4, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Priyanshu Srivastava, and Rizwan Khan, "A Review Paper on Cloud Computing," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 6, pp. 17-20, 2018. [Google Scholar]

[9] Kyle Chard et al., "Social Cloud: Cloud Computing in Social Networks," *2010 IEEE 3rd International Conference on Cloud Computing*, Miami, FL, USA, pp. 99-106, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[10] Sumit Goyal, "Public vs Private vs Hybrid vs Community - Cloud Computing: A Critical Review," *International Journal of Computer Network and Information Security*, vol. 6, no. 3, pp. 20-28, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ali Ryadh Abdulhafidh, and Hebah H. O. Nasereddin, "Building of Private Cloud Computing Architecture to Support E-Learning," *High Technology Letters*, vol. 26, no. 12, pp. 853-860, 2020. [Google Scholar] [Publisher Link]

[12] State of the Cloud Report, Flexera, 2025. [Online]. Available: https://info.flexera.com/CM-REPORT-State-of-the-Cloud

[13] Chunye Gong et al., "The Characteristics of Cloud Computing," *2010 39th International Conference on Parallel Processing Workshops*, San Diego, CA, USA, pp. 275-279, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mark Stieninger, and Dietmar Nedbal, "Characteristics of Cloud Computing in the Business Context: A Systematic Literature Review," *Global Journal of Flexible Systems Management*, vol. 15, pp. 59-68, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[15] Aaqib Rashid, and Amit Chaturvedi, "Cloud Computing Characteristics and Services: A Brief Review," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 2, pp. 421-426, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Andrew Joint, and Edwin Baker, "Knowing the Past to Understand the Present1 – Issues in the Contracting for Cloud Based Services," *Computer Law & Security Review*, vol. 27, no. 4, pp. 407-415, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[17] Rehmana Younis et al., "A Comprehensive Analysis of Cloud Service Models: IaaS, PaaS, and SaaS in the Context of Emerging Technologies and Trend," *2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, Kuala Lumpur, Malaysia, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] C.N. Höfer, and G. Karagiannis, "Cloud Computing Services: Taxonomy and Comparison," *Journal of Internet Services and Applications*, vol. 2, pp. 81-94, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[19] Somya Agrawal, "A Survey on Recent Applications of Cloud Computing in Education: COVID-19 Perspective," *Journal of Physics: Conference Series: International Symposium on Automation, Information and Computing*, Beijing, China, vol. 1828, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Rashid Nazir et al., "Cloud Computing Applications: A Review," *EAI Endorsed Transactions on Cloud Systems*, vol. 6, no. 17, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Flexera Releases 2021 State of the Cloud Report, Flexera, 2021. [Online]. Available: https://www.flexera.com/about-us/press-center/flexera-releases-2021-state-of-the-cloud-report

[22] Deepika Saxena et al., "Secure Resource Management in Cloud Computing: Challenges, Strategies and Meta-Analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 55, no. 4, pp. 2897-2912, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Fan Yunlong, and Luo Jie, "Incentive Approaches for Cloud Computing: Challenges and Solutions," *Journal of Engineering and Applied Science*, vol. 71, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Muhammed Golec et al., "Quantum Cloud Computing: Trends and Challenges," *Journal of Economy and Technology*, vol. 2, pp. 190-199, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Oluwafemi Clement Adeusi et al., "IT Standardization in Cloud Computing: Security Challenges, Benefits, and Future Directions," *World Journal of Advanced Research and Reviews*, vol. 22, no. 5, pp. 2050-2057, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Vinayak Raja, "Exploring Challenges and Solutions in Cloud Computing: A Review of Data Security and Privacy Concerns," *Journal of Artificial Intelligence General science (JAIGS)*, vol. 4, no. 1, pp. 121-144, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Bader Alouffi et al., "A Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies," *IEEE Access*, vol. 9, pp. 57792-57807, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Doaa M. Bamasoud et al., "Privacy and Security Issues in Cloud Computing: A Survey Paper," *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan, pp. 387-392, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Isaac Odun-Ayo et al., "An Overview of Data Storage in Cloud Computing," *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)*, Jammu, India, pp. 29-34, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[30] E. Manogar, and S. Abirami, "A Study on Data Deduplication Techniques for Optimized Storage," *2014 Sixth International Conference on Advanced Computing (ICoAC)*, Chennai, India, pp. 161-166, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[31] Urs Niesen, "An Information-Theoretic Analysis of Deduplication," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5688-5704, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Zheng Yan et al., "Encrypted Data Management with Deduplication in Cloud Computing," *IEEE Cloud Computing*, vol. 3, no. 2, pp. 28-35, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[33] Ravneet Kaur, Inderveer Chana, and Jhilik Bhattacharya, "Data Deduplication Techniques for Efficient Cloud Storage Management: A Systematic Review," *The Journal of Supercomputing*, vol. 74, pp. 2035-2085, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[34] A. Venish, and K. Siva Sankar, "Framework of Data Deduplication: A Survey," *Indian Journal of Science and Technology*, vol. 8, no. 26, pp. 1-7, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[35] Wen Xia et al., "A Comprehensive Study of the Past, Present, and Future of Data Deduplication," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1681-1710, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[36] Anmol Jyot Maan, "Analysis and Comparison of Algorithms for Lossless Data Compression," *International Journal of Information and Computation Technology*, vol. 3, no. 3, pp. 139-146, 2013. [Google Scholar]

[37] Wen Xia et al., "Ddelta: A Deduplication-Inspired Fast Delta Compression Approach," *Performance Evaluation*, vol. 79, pp. 258-272, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[38] Min Gu, Xiangping Li, and Yaoyu Cao, "Optical Storage Arrays: A Perspective for Future Big Data Storage," *Light: Science & Applications*, vol. 3, pp. 1-11, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[39] Jiwei Xu et al., "Clustering-based Acceleration for Virtual Machine Image Deduplication in the Cloud Environment," *Journal of Systems and Software*, vol. 121, pp. 144-156, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[40] Basappa B. Kodada, and Demian Antony D'Mello, "Secure Data Deduplication (*SD^2 eDup*) in Cloud Computing: Threats, Techniques and Challenges," *Advances in Communication and Computational Technology*, pp. 1239-1251, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[41] Anjuli Goel et al., "Security Concerns and Data Breaches for Data Deduplication Techniques in Cloud Storage: A Brief Meta-Analysis," *International Journal of Safety & Security Engineering*, vol. 14, no. 2, pp. 435-446, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[42] Philip Shilane, Ravi Chitloor, and Uday Kiran Jonnala, "99 Deduplication Problems," *Proceedings of the 8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage'16)*, Denver, Colorado, pp. 1-5, 2016. [Google Scholar] [Publisher Link]

[43] Nipun Chhabra, and Manju Bala, "A Comparative Study of Data Deduplication Strategies," *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, pp. 68-72, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[44] A. Venish, and K. Siva Sankar, "Study of Chunking Algorithm in Data Deduplication," *Proceedings of the International Conference on Soft Computing Systems*, pp. 13-20, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[45] João Paulo, and José Pereira, "A Survey and Classification of Storage Deduplication Systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1-30, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[46] Mihir Bellare, and Sriram Keelveedhi, "DupLESS: Server-Aided Encryption for Deduplicated Storage," *Proceedings of the 22nd USENIX Security Symposium*, Washington, D.C., USA, pp. 179-194, 2013. [Google Scholar] [Publisher Link]

[47] Himshai Kamboj, and Bharati Sinha, "DEDUP: Deduplication System for Encrypted Data in Cloud," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, pp. 795-800, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[48] Jiwei Xu et al., "A Lightweight Virtual Machine Image Deduplication Backup Approach in Cloud Environment," *2014 IEEE 38th Annual Computer Software and Applications Conference*, Vasteras, Sweden, pp. 503-508, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[49] Hardik Gajera, and Manik Lal Das, "DeDOP: Deduplication with Cross-Server Ownership Over Encrypted Data," *2020 Third ISEA Conference on Security and Privacy (ISEA-ISAP)*, Guwahati, India, pp. 36-40, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[50] Taek-Young Youn et al., "Efficient Client-Side Deduplication of Encrypted Data with Public Auditing in Cloud Storage," *IEEE Access*, vol. 6, pp. 26578-26587, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[51] Jiaojiao Wu et al., "CPDA: A Confidentiality-Preserving Deduplication Cloud Storage with Public Cloud Auditing," *IEEE Access*, vol. 7, pp. 160482-160497, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[52] Ankush R. Deshmukh, R.V. Mante, and P.N. Chatur, "Cloud Based Deduplication and Self Data Destruction," *2017 International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT)*, Warangal, India, pp. 155-158, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[53] Shengmei Luo et al., "Boafft: Distributed Deduplication for Big Data Storage in the Cloud," *IEEE Transactions on Cloud Computing*, vol. 8, no. 4, pp. 1199-1211, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[54] Jinbo Xiong et al., "Secure Encrypted Data with Authorized Deduplication in Cloud," *IEEE Access*, vol. 7, pp. 75090-75104, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[55] Xueyan Liu et al., "Verifiable Attribute-Based Keyword Search Over Encrypted Cloud Data Supporting Data Deduplication," *IEEE Access*, vol. 8, pp. 52062-52074, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[56] Priteshkumar Prajapati, and Parth Shah, "A Review on Secure Data Deduplication: Cloud Storage Security Issue," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 3996-4007, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[57] Wande Chen et al., "Low-Overhead Inline Deduplication for Persistent Memory," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[58] Samiksha Chavhan, Pragati Patil, and Gajanan Patle, "Implementation of Improved Inline Deduplication Scheme for Distributed Cloud Storage," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, pp. 1406-1410, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[59] Xinyu Tang et al., "A Secure and Lightweight Cloud Data Deduplication Scheme with Efficient Access Control and Key Management," *Computer Communications*, vol. 222, pp. 209-219, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[60] Neha Kaurav, "An Investigation on Data De-duplication Methods And it's Recent Advancements," *International Conference on Advances in Engineering and Technology (ICAET)*, 2014. [Google Scholar]

[61] Purushottam Kulkarni et al., "Redundancy Elimination within Large Collections of Files," *Proceedings of the General Track: USENIX Annual Technical Conference*, Boston, MA, USA, pp. 1-15, 2004. [Google Scholar] [Publisher Link]

[62] Fanglu Guo, and Petros Efstathopoulos, "Building a High-Performance Deduplication System," *2011 USENIX Annual Technical Conference (USENIX ATC 11)*, Portland, OR, 2011. [Google Scholar] [Publisher Link]

[63] Dirk Meister et al., "A Study on Data Deduplication in HPC Storage Systems," *SC '12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Salt Lake City, UT, USA, pp. 1-11, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[64] Wen Xia et al., "P-Dedupe: Exploiting Parallelism in Data Deduplication System," *2012 IEEE Seventh International Conference on Networking, Architecture, and Storage*, Xiamen, China, pp. 338-347, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[65] Hongyang Yan et al., "Centralized Duplicate Removal Video Storage System with Privacy Preservation in IoT," *Sensors*, vol. 18, no. 6, pp. 1-15, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[66] Haiyan Meng et al., "MMSD: A Metadata-Aware Multi-Tiered Source Deduplication Cloud Backup System in the Personal Computing Environment," *International Review on Computers and Software*, vol. 8, no. 2, pp. 427-679, 2013. [Google Scholar] [Publisher Link]

[67] Yu-Xuan Xing et al., "AR-dedupe: An Efficient Deduplication Approach for Cluster Deduplication System," *Journal of Shanghai Jiaotong University (Science)*, vol. 20, pp. 76-81, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[68] Yinjin Fu, Hong Jiang, and Nong Xiao, "A Scalable Inline Cluster Deduplication Framework for Big Data Protection," *ACM/IFIP/USENIX 13th International Middleware Conference*, Montreal, Canada, pp. 354-373, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[69] Xingyu Zhang, and Jian Zhang, "Data Deduplication Cluster Based on Similarity-Locality Approach," *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, Beijing, China, pp. 2168-2172, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[70] Walzade Arti, and Zine Datta, "Survey on Data Deduplication of Text File Over Cloud," *International Journal of Science and Research (IJSR)*, vol. 6, no. 1, pp. 402-405, 2017. [Google Scholar] [Publisher Link]
[71] Ahmed El-Shimi et al., "Primary Data Deduplication—Large Scale Study and System Design," *2012 USENIX Annual Technical Conference (USENIX ATC 12)*, Boston, MA, pp. 285-296, 2012. [Google Scholar] [Publisher Link]
[72] S. Uthayashangar et al., "Image and Text Encrypted Data with Authorized Deduplication in Cloud," *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[73] Zhou Lei et al., "An Improved Image File Storage Method Using Data Deduplication," *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, Beijing, China, pp. 638-643, 2014. [CrossRef] [Google Scholar] [Publisher Link]
[74] Anuja A. Sawant, and Pravin S. Game, "Deduplication of Audio Files to Remove Redundancy in Cloud Storage," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[75] Suganthi Dewakar et al., "Storage Efficiency Opportunities and Analysis for Video Repositories," *7th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 15)*, Santa Clara, CA, pp. 1-5, 2015. [Google Scholar] [Publisher Link]
[76] Fatema Rashid, Ali Miri, and Isaac Woungang, "Proof of Storage for Video Deduplication in the Cloud," *2015 IEEE International Congress on Big Data*, New York, NY, USA, pp. 499-505, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[77] Weijing You et al., "Deduplication-Friendly Watermarking for Multimedia Data in Public Clouds," *25th European Symposium on Research in Computer Security, ESORICS 2020*, Guildford, UK, pp. 67-87, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[78] Shunrong Jiang, Tao Jiang, and Liangmin Wang, "Secure and Efficient Cloud Data Deduplication with Ownership Management," *IEEE Transactions on Services Computing*, vol. 13, no. 6, pp. 1152-1165, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[79] Vivek Waghmare, and Smita Kapse, "Authorized Deduplication: An Approach for Secure Cloud Environment," *Procedia Computer Science*, vol. 78, pp. 815-823, 2016. [CrossRef] [Google Scholar] [Publisher Link]
[80] M. Chandra Sekar, and H.J. Shanthi, "Secure Data Deduplication for Efficient Cloud Storage Using Blockchain Technologies," *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India, pp. 1229-1235, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[81] Qing Liu et al., "Hadoop Based Scalable Cluster Deduplication for Big Data," *2016 IEEE 36th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, Nara, Japan, pp. 98-105, 2016. [CrossRef] [Google Scholar] [Publisher Link]
[82] Rongmao Chen et al., "BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2643-2652, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[83] Shuguang Zhan et al., "SecDedup: Secure Encrypted Data Deduplication with Dynamic Ownership Updating," *IEEE Access*, vol. 8, pp. 186323-186334, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[84] Ider Lkhagvasuren et al., "Byte-index Chunking Algorithm for Data Deduplication System," *International Journal of Security and its Applications*, vol. 7, no. 5, pp. 415-424, 2013. [CrossRef] [Google Scholar] [Publisher Link]
[85] D. Viji, and Dr.S. Revathy, "Comparative Analysis for Content Defined Chunking Algorithms in Data Deduplication," *Webology*, vol. 18, pp. 255-268, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[86] Wen Xia et al., "The Design of Fast Content-Defined Chunking for Data Deduplication Based Storage Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 9, pp. 2017-2031, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[87] Lauren Whitehouse, Data Deduplication Methods: Block-level Versus Byte-level Dedupe, TechTarget Park, 2016. [Online] Available: https://www.techtarget.com/searchdatabackup/tip/Data-deduplication-methods-Block-level-versus-byte-level-dedupe
[88] Jyoti Malhotra, and Jagdish Bakal, "A Survey and Comparative Study of Data Deduplication Techniques," *2015 International Conference on Pervasive Computing (ICPC)*, Pune, India, pp. 1-5, 2015. [CrossRef] [Google Scholar] [Publisher Link]
[89] Hala AbdulSalam Jasim, and Assmaa A. Fahad, "New Techniques to Enhance Data Deduplication using Content based-TTTD Chunking Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 116-121, 2018. [CrossRef] [Google Scholar] [Publisher Link]
[90] Fei Xie, Michael Condict, and Sandip, "Estimating Duplication by Content-based Sampling," *2013 USENIX Annual Technical Conference (USENIX ATC 13)*, San Jose, CA, pp. 181-186, 2013. [Google Scholar] [Publisher Link]
[91] G. Sujatha, and Jeberson Retna Raj, "A Comprehensive Study of Different Types of Deduplication Technique in Various Dimensions," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 316-323, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[92] S. Hema, and A. Kangaiammal, "Distributed Storage Hash Algorithm (DSHA) for File-Based Deduplication in Cloud Computing," *Second International Conference on Computer Networks and Communication Technologies*, pp. 572-581, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[93] Yinjin Fu et al., "Application-Aware Local-Global Source Deduplication for Cloud Backup Services of Personal Storage," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 5, pp. 1155-1165, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[94] Yinjin Fu et al., "AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment," *2011 IEEE International Conference on Cluster Computing*, Austin, TX, USA, pp. 112-120, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[95] Jonathan Takeshita, Ryan Karl, and Taeho Jung, "Secure Single-Server Nearly-Identical Image Deduplication," *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, Honolulu, HI, USA, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[96] Yinjin Fu et al., "Application-Aware Big Data Deduplication in Cloud Environment," *IEEE Transactions on Cloud Computing*, vol. 7, no. 4, pp. 921-934, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[97] P.G. Shynu et al., "A Secure Data Deduplication System for Integrated Cloud-Edge Networks," *Journal of Cloud Computing*, vol. 9, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[98] Silambarasan Elkana Ebinazer, Nickolas Savarimuthu, and S. Mary Saira Bhanu, "An Efficient Secure Data Deduplication Method using Radix Trie with Bloom Filter (SDD-RT-BF) in Cloud Environment," *Peer-to-Peer Networking and Applications*, vol. 14, pp. 2443-2451, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[99] Guipeng Zhang et al., "BDKM: A Blockchain-Based Secure Deduplication Scheme with Reliable Key Management," *Neural Processing Letters*, vol. 54, pp. 2657-2674, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[100] Yunling Wang et al., "Secure Deduplication with Efficient user Revocation in Cloud Storage," *Computer Standards & Interfaces*, vol. 78, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[101] N. Mageshkumar, and L. Lakshmanan, "RETRACTED ARTICLE: An Improved Secure File Deduplication Avoidance using CKHO based Deep Learning Model in a Cloud Environment," *The Journal of Supercomputing*, vol. 78, pp. 14892-14918, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[102] Priyanka Singh, Nishant Agarwal, and Balasubramanian Raman, "Secure Data Deduplication using Secret Sharing Schemes Over Cloud," *Future Generation Computer Systems*, vol. 88, pp. 156-167, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[103] Jinfeng Liu et al., "Secure Similarity-Based Cloud Data Deduplication in Ubiquitous City," *Pervasive and Mobile Computing*, vol. 41, pp. 231-242, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[104] J.R. Douceur et al., "Reclaiming Space from Duplicate files in a Serverless Distributed File System," *Proceedings 22nd International Conference on Distributed Computing Systems*, Vienna, Austria, pp. 617-624, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[105] Cheng Guo et al., "R-Dedup: Secure Client-Side Deduplication for Encrypted Data without Involving a Third-Party Entity," *Journal of Network and Computer Applications*, vol. 162, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[106] Jian Liu, N. Asokan, and Benny Pinkas, "Secure Deduplication of Encrypted Data without Additional Independent Servers," *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver Colorado USA, pp. 874-885, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[107] Mihir Bellare, Sriram Keelveedhi, and Thomas Ristenpart, "Message-Locked Encryption and Secure Deduplication," *Advances in Cryptology – EUROCRYPT 2013: 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques*, Athens, Greece, pp. 296-312, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[108] Wenbin Yao, and Pengdi Ye, "Simdedup: A New Deduplication Scheme Based on Simhash," *Web-Age Information Management: WAIM 2013 International Workshops: HardBD, MDSP, BigEM, TMSN, LQPM, BDMS*, Beidaihe, China, pp. 79-88, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[109] G. Madhubala et al., "Nature - Inspired Enhanced Data Deduplication for Efficient Cloud Storage," *2014 International Conference on Recent Trends in Information Technology*, Chennai, India, pp. 1-6, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[110] Yan-Kit Li et al., "Efficient Hybrid Inline and Out-of-Line Deduplication for Backup Storage," *ACM Transactions on Storage*, vol. 11, no. 1, 1-21, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[111] Yan Tang et al., "DIODE: Dynamic Inline-Offline DE Duplication Providing Efficient Space-Saving and Read/Write Performance for Primary Storage Systems," *2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, London, UK, pp. 481-486, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[112] João Paulo, and José Pereira, "Efficient Deduplication in a Distributed Primary Storage Infrastructure," *ACM Transactions on Storage (TOS)*, vol. 12, no. 4, pp. 1-35, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[113] Panfeng Zhang et al., "Resemblance and Mergence Based Indexing for High Performance Data Deduplication," *Journal of Systems and Software*, vol. 128, pp. 11-24, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[114] Awais Khan, Prince Hamandawana, and Youngjae Kim, "A Content Fingerprint-Based Cluster-Wide Inline Deduplication for Shared-Nothing Storage Systems," *IEEE Access*, vol. 8, pp. 209163-209180, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[115] Wen Xia et al., "Similarity and Locality Based Indexing for High Performance Data Deduplication," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 1162-1176, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[116] Peizhen Guo, and Wenjun Hu, "Potluck: Cross-Application Approximate Deduplication for Computation-Intensive Mobile Applications," *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*, Williamsburg VA USA, pp. 271-284, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[117] Qinlu He et al., “Research on Routing Strategy in Cluster Deduplication System,” *IEEE Access*, vol. 9, pp. 135485-135495, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[118] Niteesha Sharma, A. V. Krishna Prasad, and V. Kakulapati, “File-level Deduplication by using Text Files - Hive Integration,” *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-6, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[119] Fatema Rashid, and Ali Miri, *Deduplication Practices for Multimedia Data in the Cloud*, Guide to Big Data Applications, Springer, pp. 245-271, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[120] P.M. Ashok Kumar, E. Pugazhendhi, and Rudra Kalyan Nayak, “Cloud Storage Performance Improvement Using Deduplication and Compression Techniques,” *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp. 443-449, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[121] Manogar Ellappan, and Abirami Murugappan, “A Smart Hybrid Content-Defined Chunking Algorithm for Data Deduplication in Cloud Storage,” *Soft Computing*, vol. 28, pp. 9037-9052, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[122] Lianghong Xu et al., “Online Deduplication for Databases,” *Proceedings of the 2017 ACM International Conference on Management of Data*, Chicago Illinois USA, pp. 1355-1368, 2017. [CrossRef] [Google Scholar] [Publisher Link]
[123] Shweta Pal, Kiran More, and Priya Pise, “Content-Based Deduplication of Data Using Erasure Technique for RTO Cloud,” *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, Sangamner, India, pp. 109-113, 2018. [CrossRef] [Google Scholar] [Publisher Link]