I.J. Information Engineering and Electronic Business, **2025, 2, 129-146**
Published Online on April 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2025.02.07

Data Deduplication-based Efficient Cloud Optimisation Technique: Optimizing Cloud Storage through Data Deduplication

Ranga Kavitha
Siddhartha institute of engineering and technology/ Computer Science and Engineering Department, Ibrahimpatnam Ranga reddy district, 501506, India
E-mail: r.kavitha@siddhartha.ac.in
ORCID iD: https://orcid.org/0000-0001-9542-3273

Mahaboob Sharief Shaik*
Siddhartha institute of engineering and technology/Computer Science and Engineering Department, Ibrahimpatnam Ranga reddy district, 501506, India
E-mail: smshariefkaau@gmail.com
ORCID iD: https://orcid.org/ 0009-0008-3800-2550
*Corresponding Author

Narala Swarnalatha
Marri Laxman Reddy Institute of Technology and Management/ Computer Science and Engineering Department, Medchal Malkajgiri,500043, India
E-mail: swarnalatha.jogireddy@gmail.com
ORCID iD: https://orcid.org/0009-0006-7753-7828

M. Pujitha
Marri Laxman Reddy Institute of Technology and Management/ Computer Science and Engineering Department, Medchal Malkajgiri,500043, India
E-mail: mpujitha46@gmail.com
ORCID iD: https://orcid.org/0009-0006-9169-976X

Syed Asadullah Hussaini
ISL Engineering College/Computer Science and Engineering/ Department,Hyderabad 500005, India
E-mail: drsyedasadullahhusaini@gmail.com
ORCID iD: https://orcid.org/0009-0001-8502-0697

Samiullah Khan
Bahrain Financing Company, Kingdom of Bahrain
Email: in.samiukhan@gmail.com
ORCID iD: https://orcid.org/0009-0005-7767-8471

Shamsher Ali
Vice President, Citigroup NA in Tampa, FL,USA.
Email: Shamsher1680@gmail.com
ORCID iD: https://orcid.org/0009-0007-6784-9710

Received: 15 November, 2024; Revised: 22 January, 2025; Accepted: 25 February, 2025; Published: 08 April, 2025

Abstract: Effective storage management is crucial for cloud computing systems' speed and cost, given data's exponential increase. The significance of this issue has increased as the amount of data continues to increase at a disturbing pace. The act of detecting and removing duplicate data can enhance storage utilisation and system efficiency. Using less storage capacity reduces data transmission costs and enhances cloud infrastructure scalability. The use of deduplication techniques on a wide scale, on the other hand, presents a number of important obstacles. Security issues, delays in deduplication, and maintaining data integrity are all examples of difficulties that fall under this classification.

This paper introduces a revolutionary method called Data Deduplication-based Efficient Cloud Optimisation Technique (DD-ECOT). Optimising storage processes and enhancing performance in cloud-based systems is its intended goal. DD-ECOT combines advanced pattern recognition with chunking to increase storage efficiency at minimal cost. It protects data during deduplication with secure hash-based indexing. Parallel processing and scalable design decrease latency, making it adaptable enough for vast, ever-changing cloud setups.The DD-ECOT system avoids these problems through employing a secure hash-based indexing method to keep data intact and by using parallel processing to speed up deduplication without impacting system performance. Enterprise cloud storage systems, disaster recovery solutions, and large-scale data management environments are some of the usage cases for DD-ECOT. Analysis of simulations shows that the suggested solution outperforms conventional deduplication techniques in terms of storage efficiency, data retrieval speed, and overall system performance. The findings suggest that DD-ECOT has the ability to improve cloud service delivery while cutting operational costs. A simulation reveals that the proposed DD-ECOT framework outperforms existing deduplication methods. DD-ECOT boosts storage efficiency by 92.8% by reducing duplicate data. It reduces latency by 97.2% using parallel processing and sophisticated deduplication. Additionally, secure hash-based indexing methods improve data integrity to 98.1%. Optimized bandwidth usage of 95.7% makes data transfer efficient. These improvements suggest DD-ECOT may save operational costs, optimize storage, and beat current deduplication methods.

Index Terms: Efficient, Data, Deduplication, Optimizing, Storage, Performance, Cloud, Computing, Environment

1. Introduction

The cloud presents various challenges to standard data deduplication technologies' storage and processing efficiency. Problems emerge naturally in a cloud setting, fixing the significant computational overhead of detecting and removing duplicate data is crucial [1]. Traditional deduplication algorithms employ hash functions to identify data chunks. When processing enormous amounts of data, this can slow processing. Data's unpredictability causes problems for these algorithms [2]. Data duplication is impeded because even a small file change needs reloading the dataset. Tradition solutions usually use a present deduplication policy, which can't operate effectively with changing workloads or data. Long-term, this reduces storage utilisation, massive cloud data transfers require a lot of bandwidth [3]. This increases up consumer costs and latency. Traditional deduplication methods offer little security measures, rendering them vulnerable to data breaches [4]. This attraction requires another evaluation. Due to these limits, content-aware deduplication and machine learning algorithms are critically needed. Intelligently discovering and processing duplicate data could improve deduplication. It has less overhead than traditional methods [5]. Resolving these issues quickly optimises storage space utilisation and boosts cloud data processing efficiency [6].

Data deduplication solutions in cloud environments are facing many issues that are affecting storage and processing [7]. As application data grows in variety and volume, typical deduplication methods struggle to find and eliminate duplicates. As data types grow, especially semi-structured and unstructured data, conventional techniques may struggle to retain accuracy and performance [8]. Even minor modifications may need re-evaluating the entire data, cloud settings' fluctuation makes deduplication ineffective. Data privacy and security during deduplication are another issue [9]. Cloud storage of sensitive data increases the risk of data breaches or unauthorised access during deduplication. With outside service providers, this is more readily apparent. Regulatory compliance requires strict data protection, compounding these challenges [10]. Deduplication can additionally compromise performance for space savings. Due to delay and data retrieval speed, extensive deduplication may improve storage costs however degrade user experience [11]. Overall, deduplication technologies must be integrated into cloud infrastructure and operations quickly. Due to hybrid cloud arrangements, platform-agnostic deduplication is difficult for many enterprises [12]. For cloud data deduplication technologies and cloud processing and storage to improve, these difficulties must be resolved.

Previously developed cloud data deduplication algorithms improve processing and storage efficiency [13]. Content-aware deduplication approaches verify data rather than file signatures, improving accuracy. People accomplish this by examining the material. ML algorithms can quickly spot duplicates and adapt to changing data patterns [14]. Deduplication with encryption-awareness protects data and addresses privacy concerns. Client-side and server-side deduplication in hybrid techniques maximises processing efficiency and bandwidth. Due to distributed deduplication across several cloud environments, workloads may be balanced, improving storage utilisation, latency, cloud performance, and user experience [15].

Cloud computing's exponential data growth has made storage management difficult due to inefficient resource consumption and rising operational costs. Operating costs have increased, causing these issues. Insufficient data security, long deduplication periods, and data integrity hazards are common issues with traditional data deduplication technologies. Due to these challenges, cloud infrastructures struggle to scale and operate, making them unsuitable for huge datasets. Due to this, new deduplication algorithms are urgently needed to increase cloud storage economy and system efficiency.

- Develop a data deduplication solution to detect and delete duplicate data in cloud environments to improve storage utilisation and reduce storage footprint.
- DD-ECOT that enable quick deduplication while lowering latency and resource utilisation can maintain cloud performance during data processing.
- Deduplication must retain data integrity for secure hash-based indexing to prevent security issues.

The research paper's structure is laid out in this section, which includes the following: The section II of this paper explores into strategies for efficient storage and processing of cloud data through deduplication. DD-ECOT will be thoroughly discussed in Section III of this dissertation. An exhaustive examination, a comparison to prior approaches, and an examination of the consequences are presented in Section IV. In Section V, the findings are thoroughly examined.

2. Literature Survey

Cloud computing data is growing exponentially, making effective and safe data management systems crucial. Cloud storage security and data redundancy have been studied from several perspectives. A literature review on SDS, it protects data and optimise storage. These solutions handle cloud big data concerns. These solutions combine inline deduplication with modern encryption technologies.

Prajapati, P et al., [16] this research provides a complete literature evaluation of secure deduplication strategies (SDS) in cloud storage. It highlights several approaches that address client security concerns while preserving data confidentiality, integrity, and efficient storage utilisation. Specifically, all of these techniques are discussed in this paper.

Mahesh, B et al., [17] while protecting against unauthorised access, this article describes a variety of ways for secure deduplication (SD) that can be applied to encrypted data stored in cloud storage. These techniques can improve performance, save storage costs, and increase storage utilisation.

Kumar, P. A. et al., [18] for a wide range of file popularity scenarios, this study introduces De-duplication Based Cloud Storage (DBCS), which makes use of inline de-duplication and dynamic data reconfiguration on Amazon Web Services (AWS). It achieves a space efficiency that is 95% better than that of existing solutions.

Tahir, M. U et al., [19] proposed the purpose of this paper is to address the issues that big data presents to cloud service providers by discussing methods that can be used to prevent data redundancy in cloud storage. These methods include a version control system for data de-duplication.

Chhabraa, N et al., [20] proposed using Bloom filters in conjunction with attribute-based encryption (A-BE), this study presents a safe deduplication methodology. The research demonstrates, through experimental findings, that this strategy is more efficient in terms of memory and time than typical encryption and hashing methods.

Rajput, U et al., [21] introduced the purpose of this study is to provide insights for the development of more effective cloud storage management (CSM) approaches by analysing existing data extraction methodologies to manage duplication in cloud storage. The paper highlights the inefficiencies of these tactics.

Blend Bloom filters and attribute-based encryption for even more effective deduplication. Saves time and memory. Traditional methods are inferior to the Data Deduplication-based Efficient Cloud Optimisation Technique (DD-ECOT). It improves cloud service delivery and lowers operational costs due to its efficiency and performance.

3. Proposed Method

Data deduplication is very important for both data management and storage capacity optimisation. Often referred to as single-instance storage, this technique removes unnecessary copies of data by storing only unique blocks and replacing duplicates with pointers. As this paper shows, data deduplication across many environments including client-side and target-side deduplication techniques as well as its integration with cloud data storage have great value. Furthermore, it presents secure data outsourcing techniques and cloud optimisation strategies employing deduplication to ensure data integrity, dependability, and cost economy all of which are rather crucial in modern cloud architectures and data management systems.

Contribution 1: Development of DD-ECOT

Designed to maximise storage space and improve performance in cloud systems, this paper proposes the Data Deduplication-based Efficient Cloud Optimisation Technique (DD-ECOT). Combining strong pattern recognition algorithms with chunking techniques to detect and eliminate duplicate data helps the method to lower system overhead and hence maximise storage utilisation.

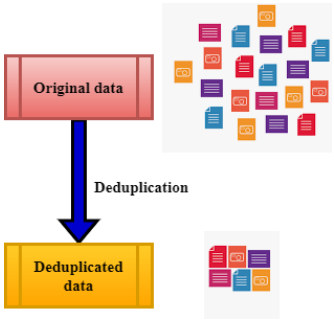

Fig. 1. Design of Data deduplication

Data deduplication, often called elegant compression or single-instance storage, helps save storage costs by removing repetitive data duplicates (figure 1). Data deduplication methods ensure that, on capacity media such as disc, flash or tape, only one of a type occurrence of data is stored. Redundancy in data blocks is replaced with a pointer pointing to a one-of-a-kind copy. Data deduplication effortlessly fits perfectly with constant strength, replicating only the data that has changed after previous validation. Equation 1 depicts a mathematical system dynamics model, input variables, and performance measurements. The objective is to measure the impact of parameters like p' and w' on an observed outcome, M, in data deduplication and cloud computing optimization. This equation is critical because it captures the interaction between storage efficiency (Vf') and system overhead (nP''), enabling precise optimization techniques. This concept might help large-scale systems enhance scalability and efficiency by balancing resource consumption, performance, and latency in real-world applications like cloud storage.

$$|M(j - kp')| = Vf' < \partial p' - ew'' > + nP'' \tag{1}$$

The data variability $M(j - kp')$, deduplication delay Vf', and use of resources nP'' impact storage and retrieval performance in terms of the equation 1, $\partial p' - ew''$, which represents the deduplication process's overall efficiency. Equation 2 presumably explains how operational factors (wq'', R, and j'') affect energy or resource efficiency (EM). This tool quantifies how changes in workload (wq'') or resource usage affect system performance. This equation may assess energy efficiency and resource utilization under different workloads in cloud computing and network optimization to detect bottlenecks. The approach might reduce operational costs, improve energy efficiency, and ensure system scalability in data-intensive contexts like large-scale cloud systems or IoT networks by optimizing these parameters. By optimizing these characteristics, cloud system performance may be boosted and operations expenses can be reduced, as shown.

$$< \frac{2}{4(\forall - 3wq'')} > E_M(R - j'') - Rwq'' \tag{2}$$

The efficiency of energy $4(\forall - 3wq'')$ or resources is represented by equation 2, E_M, where the quantities $(R - j'')$ stand for the resources used Rwq'' and the effect of deduplication procedures, respectively. In DD-ECOT, equation 2 shows how optimizing these factors may increase performance and minimise overhead in systems storing data in the cloud. Equation (3) is designed to optimize performance metrics by modeling the link between system efficiency, workload dynamics, and external factors. Its stated goal is to find a happy medium between resource use and operational output.

$$E_t(\partial - Rw'') = [\forall w - rt''] + Ty(\varphi\omega + \pi\mu') \tag{3}$$

Equation 3 shows the DD-ECOT method's energy efficiency E_t and resource utilization interact with one another. The data integrity that is preserved during deduplication is represented by $\partial - Rw''$ and the resource cost that is incurred is denoted by $[\forall w - rt'']$. To maximize system performance and energy savings in cloud environments, it is essential to optimize the following components: workload Ty, retrieval time $\varphi\omega$, and efficiency factors $\pi\mu'$. Under different operating circumstances, Equation (4) aims to examine the trade-off between system throughput and resource allocation. It makes finding the best possible setups to maximize efficiency and performance easier.

$$\partial T' - Fv(M - vr'') = Eq' - Mvd'' \tag{4}$$

The data flow during deduplication is shown by equation 4 Fv, where $(M - vr'') =$ is the total data volume and $\partial T'$ is the variability in the data processing. In the end, cloud storage systems may benefit from improved speed and efficiency due to deduplication, which can decrease processing time (Eq') with little influence on resource utilization (Mvd'').

$$|\forall + M_h(ew - 2p) = M(\partial\forall' - Rt) \propto| \tag{5}$$

The effects of deduplication $\forall + M_h$, the equation data integrity $ew - 2p$, and the effectiveness of resource allocation in the DD-ECOT approach. In this case, $\partial\forall'$ pertains to the efficiency weight of deduplication $\propto$ and Rt to the overhead caused by processing duplicate data M. Efficient deduplication solutions may improve data management and resource utilization in cloud settings, as shown by equation 5.

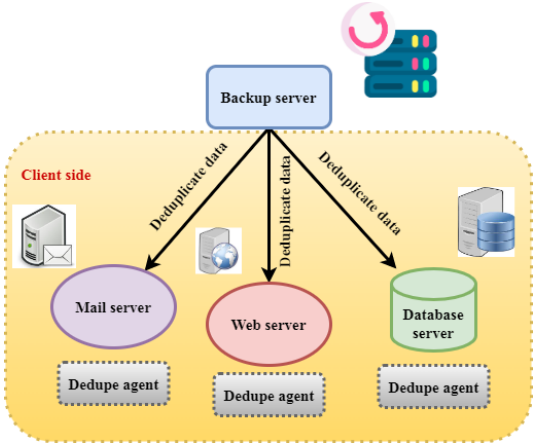

Fig. 2. Deduplication on the client side (source)

Source based deduplication occurs at the client side as the name suggests. Figure 2 illustrates how deduplication agents are placed at a real or virtual server, determining the source-based deduplication method. Dedupe agents will consequently examine all the duplicates via the backup server and then send unique data blocks to the disc. This procedure is completed before the data crosses the network. The source side deduplication has the benefit that it just backs up the altered data and uses less bandwidth for the data.

$$\forall_m(Wq - nf'') = Mk(\partial_2 - Yt'') + Rp'' \tag{6}$$

Within the DD-ECOT framework $\partial_2 - Yt''$, the efficiency of data management ($\forall_m$) and the allocation of resources. Wq denotes the burden of data processing in this context, while the resource costs of handling duplicate data are represented by nf''. Enhancing total cloud storage efficiency Rp'' and decreasing operating costs may be achieved via the utilization of resources and the performance measure Mk.

$$B|\mathsf{C} - \varphi w''| = \omega\rho(\mu - \pi\vartheta'') - Rt(Y - pq'') \tag{7}$$

The intended efficiency level is represented by $\omega\rho$ and the overhead Rt from redundancy $B|\mathsf{C} - \varphi w''|$ is denoted by $\mu - \pi\vartheta''$ in the equation Rt which also indicates the weight of resource utilization $\omega\rho$ and performance effect $Y - pq''$. The DD-ECOT method is useful in enhancing the efficiency and cost-effectiveness of cloud storage, and this is shown.

$$M(Pt - yj'mq) = |E_w(n - qt)| * Vc'' \tag{8}$$

The processing time for handling data is represented by the equation 8 $Pt - yj'mq$, and the overhead from duplicate data is denoted by M. Improving overall performance relies on optimizing these parameters Vc'', whereas the measures the effect of workload efficiency (E_w) and the consequent resource utilization ($n - qt$).

$$T_r(v - nw'') = \forall n_{v-2} * (E\forall - \delta'') \tag{9}$$

In the DD-ECOT approach, the link between data deduplication efficiency and resource utilization δ'' is captured by the equation 9. The data volume is denoted by $\forall n_{v-2}$ and the resource costs of dealing with duplicate data are shown by $E\forall$. The product of work efficiency $T_r(v - nw'')$ and the efficacy of deduplication procedures increases overall performance.

$$H\big(U(m' - eq'')\big) = M * dZ(\forall - \partial w'') * Rt \tag{10}$$

The allocation of resources within the DD-ECOT framework $M * dZ$ and the efficiency of data handling (H). In this scenario, the usefulness of optimized data $\forall - \partial w''$ after deleting duplicates is represented by $\left(U(m' - eq'')\right)$ and the total effect of resource management is denoted by Rt .

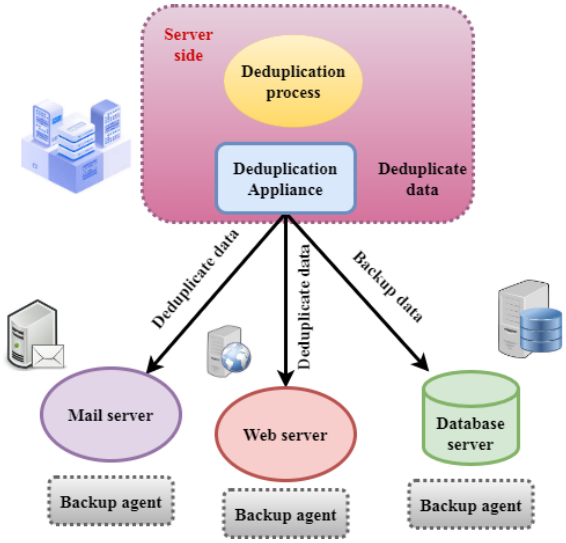

Fig. 3. Deduplication on the target side

The dedicated deduplication procedure is executed on the target server in the target based deduplication. Following client backup data acquisition, the procedure is executed as indicated in Figure 3.All kind of deduplication is handled on the backup server by a deduplication appliance. This approach mostly helps customers by removing them from the overhead of the deduplication process.

$$Pm < \forall(Ek - \varphi\beta'') > \rho\theta(\epsilon - \alpha v'') \tag{11}$$

The data handling efficiency is represented by $\rho\theta$, and the overhead linked to duplicate data is accounted for by $\epsilon - \alpha v''$. Improving these components leads to improved utilization of resources, as shown on equation 11, which represents optimization of deduplication procedures may impact workload efficiency $Ek - \varphi\beta''$.

$$|Mt(\gamma - \nabla q')| = F_d < Vb + nh'' > -Et'' \tag{12}$$

The DD-ECOT framework's deduplication effectiveness and the equation 12 for data management efficiency. The intended efficiency level is denoted by Mt in this context, whereas the variability caused by repeated data processing is represented by $\gamma - \nabla q'$. Reducing total energy expenses $(-Et'')$ may be achieved by optimizing resource flow (F_d), data volume $(Vb + nh'')$, and highlighting the fact that efficient deduplication solutions improve resource utilization.

$$< Int(Eq' - ft'') > Fd < Mt(p - yr'') > \tag{13}$$

The total quality of processing data is represented by the equation 13, Eq', and any expenses associated with duplicate data Int are denoted by ft''. Effective deduplication solutions not only increase resource utilization but maximize operational efficiency, as seen by optimizing resource flow (Fd) and data management $(Mt(p - yr''))$.

$$Z(v - cd') = Mhj(K - py'') * Et(n + rw'') \tag{14}$$

The DD-ECOT approach involves the data volume equation 14, $(Z(v - cd'))$ and optimization of resources. The data reduction accomplished by deduplication is represented by $K - py''$ in this context, whereas $n + rw''$ the efficiency of resource management is shown by Mhj. Effective deduplication solutions are shown to optimize the difference between the utilization of resources and retrieval times Et improves overall system performance.

$$C(m - nb'') = R(Ty - Pkh'') * Esw'' \tag{15}$$

The overhead expenses related to duplicate data handling are represented by the equation $C(m - nb'')$, and the resource utilization rate is denoted by $R(Ty - Pkh'')$. On improving the difference Esw'' improves system performance in general.

Contribution 2: Implementation of DD-ECOT

The proposed method preserves data integrity by means of a safe hash-based indexing strategy and parallel processing to accelerate the deduplication process. This approach lowers latency and resource use even if it guarantees efficient and secure cloud data management.

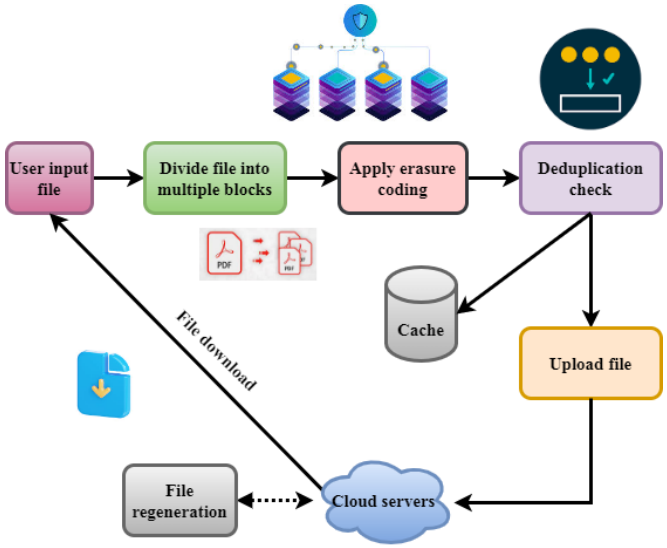

Fig. 4. Secure Data Outsourcing with Redundancy in Cloud Servers

Figure 4 describes a technique for safely spreading data to many cloud servers. Divided into blocks, encoded using erasure coding for redundancy, and deduplicated to decrease storage costs, the user input file is Following then is the distribution of the cloud server for the encoded data. Should a server fail or data is lost, a file may be reconstructed using the remaining encoded data. This system provides reliability and availability of data in a distributed cloud environment. Data redundancy, processing time, and storage efficiency are quantified by this equation. Identifying and removing duplicate data improves data retrieval rates and storage capacity in real-world cloud systems. Cost reduction and performance improvement are important, especially in large-scale cloud settings with growing data volumes.

$$H\big(y, -prt''(\forall + rt'')\big) = \beta(\Delta + \exists' < \delta - 2q >) \tag{16}$$

The data processing and retrieval resource costs are denoted by $y, -prt''$ and $(\forall + rt'')$ respectively, while the system's output efficiency is represented by the equation 16, H. Deduplication solutions that work not only simplify data management but also improve overall performance, as shown by the optimization of parameters $\Delta + \exists'$and $\delta - 2q$.

$$4r < Mt(Y - pk'') > T(\partial - Rte'') + Fz'' \tag{17}$$

The data management effectiveness is represented by the equation 17, $Mt(Y - pk'')$, and the changes required Fz'' to offset the effects of duplicate data processing on resource allocation are shown by $4r$. The operational overhead is indicated by the $T(\partial - Rte'')$. To improve the system's performance as a whole, one must optimize these variables, as equation 17 shows.

$$\partial(\forall - Pk'') + Vf(\partial\forall'') = T_r(m - nvf'') \tag{18}$$

The influence of variable factors on data handling processes is shown by $\partial(\forall - Pk'')$ and the optimisation of workload efficiency is represented by the equation $Vf(\partial\forall'')$. The decrease of overhead costs caused by duplicate data is correlated with the allocation of resources, which is represented by $T_r(m - nvf'')$. Improving total system performance is critically dependent on successfully addressing these factors, as shown by this equation 18.

$$4r(T - yr') = Min < \omega(\sigma - \tau\alpha'') > \tag{19}$$

After the costs of managing duplicate data have been deducted, the net benefit of resource consumption is represented by equation 19, $4r(T - yr')$. The minimum operational expenses obtained by optimising the efficiency parameters $Min$are reflected by the $\omega(\sigma - \tau\alpha'')$.

$$V_t(M - nvt'') = Fa(b - vt'') * Ej' \tag{20}$$

After accounting for redundant information overhead, the effective data throughput is represented by the equation 20, $V_t(M - nvt'')$, and the resource allocation efficiency Ej' associated to minimizing redundancy is denoted by Fa. Optimizing these elements has an operational influence, $b - vt''$. To improve data processing capabilities, this equation stresses the need of good deduplication procedures.

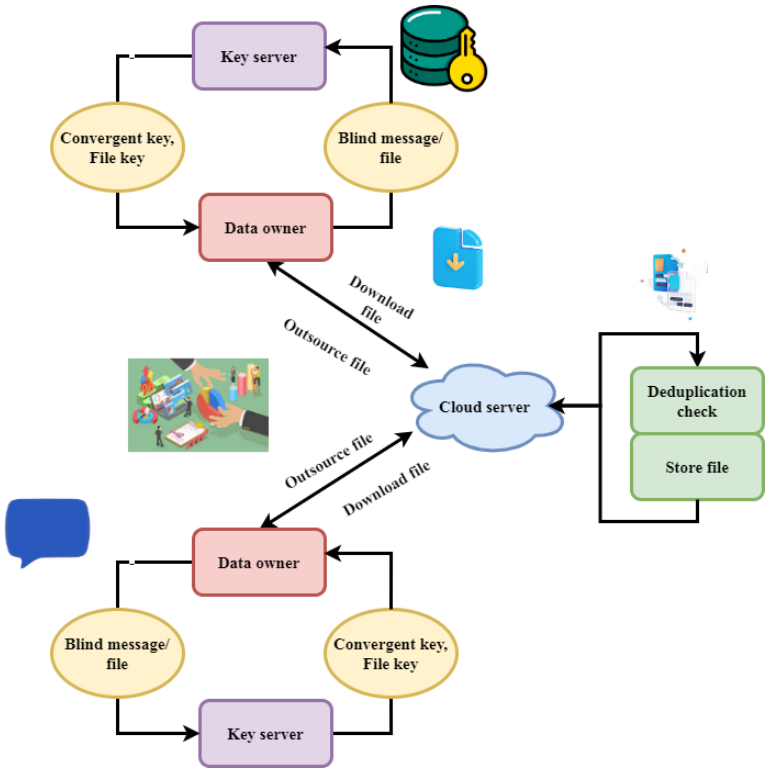

Fig. 5. Cloud data storage architecture for secure deduplication

Figure 5 presents a secure data outsourcing process aimed at a cloud server. After encrypting the file using the convergence key and file key, the data owner transmit the convergent key and file key to the key server. Retaining the convergent key, the key server sends the encrypted file to the cloud server. The cloud server stores the particular file and verifies duplicate files. When the file is needed downloaded, the data owner requests it from the cloud server. The encrypted file is transferred to the key server via the cloud server, which decodes it using the convergent key and thereafter forward the decoded file back to the data owner. All through the outsourcing and retrieval process, this approach ensures integrity and confidentiality of data. Equation 21 models workload efficiency (Tw) based on system factors and operational dynamics. It provides a mathematical framework to assess how resource restrictions, system throughput, and task distribution affect processing efficiency. This equation can optimize cloud computing systems or big data centers by matching workload needs with resource availability, reducing processing delays and improving performance.

$$T_w(Qa - bt'') = \forall(\omega\varphi + \rho\sigma) * 4\tau'' \tag{21}$$

Equation 21, $Qa - bt''$ shows the effective workload once duplicate data $\forall(\omega\varphi + \rho\sigma)$ is taken into consideration, whereas equation T_w captures the integrated elements that affect system performance, such as information processing speed $4\tau''$ and resource utilization on storage utilization efficiency analysis.

$$+\partial \propto (\forall - et'') = Vty(\omega + \delta'') * Est'' \tag{22}$$

The equation $\forall - et''$ shows the efficiency benefits of deleting duplicate data, and the variable $+\partial \propto$ shows the total effect of other performance parameters $\omega + \delta''$, such as the speed of data retrieval Est'' and the allocation of resources. On representing the operational ramifications of these optimizations, is Vty on data deduplication latency analysis.

Contribution 3: Evaluation of DD-ECOT

Through many simulations, the performance of the method is assessed displaying improved storage efficiency, faster data retrieval, and overall system performance. Calculated using specific equations, important indicators of storage optimisation and system improvement show DD-ECOT's relative efficiency over more traditional approaches.

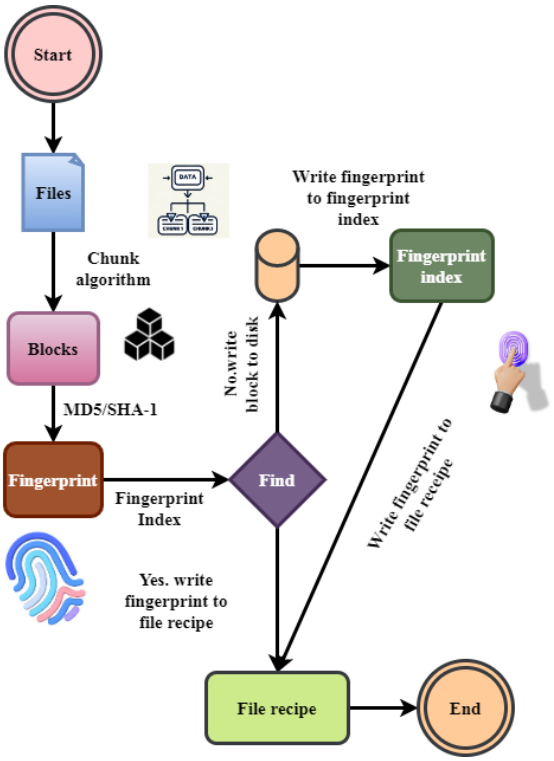

Fig. 6. The process of data deduplication

Figure 6 illustrates a fingerprint-based data integrity check system. Files are split into blocks, each of which has an own fingerprint derived from a hash algorithm (MD5/SHA-1). A fingerprint index houses these fingerprints. The fingerprints of the current file are matched in the index during verification. Should disparities be found, it indicate possible data corruption or manipulation, therefore enabling the assessment and preservation of file integrity through this comparison procedure.

$$Q(V,rt) = H(\partial Wq'') + Mt(Y - pkt'') \tag{23}$$

The deduplication efforts' contribution to improving $Mt(Y - pkt'')$ data quality through efficient workload management is represented by the equation $Q(V,rt)$, and the impact of optimized resource allocation on system performance by minimizing the effects of duplicate data is shown in ($H(\partial Wq'')$. To increase data quality and system performance, it is crucial to include efficient deduplication solutions, as shown by equation 23 for data integrity analysis.

$$T(Y' - Rt) = D < \varepsilon(\nabla + \beta'') > + Tja'' \tag{24}$$

The effective throughput attained after deduplication methods is denoted by the equation $T(Y' - Rt)$, and the contribution of optimized data management parameters to improving overall system performance is denoted by $D < \varepsilon(\nabla + \beta'')$. Other operational parameters impacting efficiency are denoted by the Tja''. Both throughput and resource management are improved by effective deduplication methods, as shown by equation 24 on bandwidth utilization analysis.

Figure 7 describes the Efficient Cloud Optimisation Technique (DD-ECOT) based on data deduplication. It shows how to maximise cloud storage by chunking incoming data, utilising secure hash-based indexing for data integrity, and pattern recognition algorithm duplicity detection. Driven by parallel processing, the deduplication engine removes duplicates, producing adequate cloud storage and quicker data retrieval. This strategy guarantees data security, lowers storage costs, and improves cloud performance.

Secure hash-based indexing is one way DD-ECOT tackles data security problems; it guarantees data integrity and reduces the chances of data breaches during deduplication. This strategy ensures that duplicate data is discovered and indexed without revealing sensitive information. The integration of encryption-aware deduplication and powerful access control methods further validates the security procedures of DD-ECOT. These steps guarantee safe and dependable cloud storage operations by adhering to data privacy requirements and reaching quantifiable security goals, including resistance to hash collisions and illegal data access.

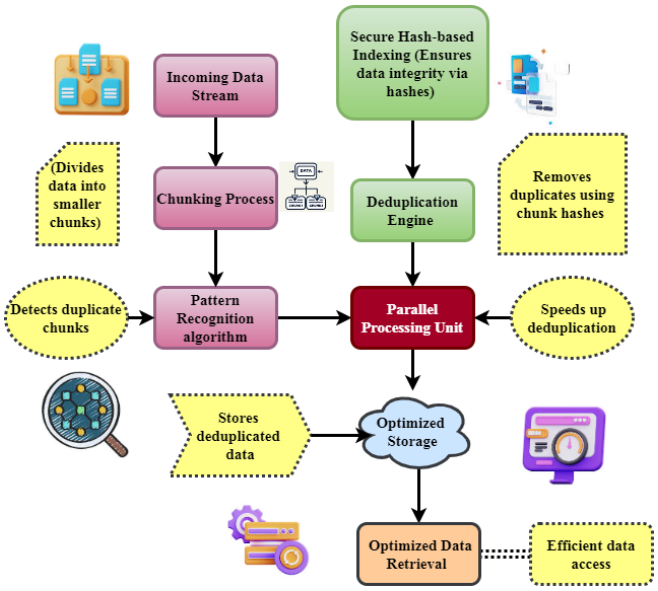

Fig. 7. Data Deduplication-based Efficient Cloud Optimization Technique (DD-ECOT)

Equation 25 models the link between system throughput (T), resource dynamics, and performance efficiency under different scenarios. It shows how workload, processing delays, and resource use effect system throughput. In practice, this equation helps optimize distributed systems like cloud platforms and IoT networks by optimizing resource allocation, avoiding delays, and assuring high performance in data-intensive conditions.

$$T < Nv.rt'' > Yp(aw' - bt) * Vd(e - rd'') \tag{25}$$

The time-dependent variable affecting throughput is represented by the equation 25, $T < Nv.rt'' >$, and the net impact of deduplication on retrieval speeds Vd, which describes the decrease in processing time $(e - rd'')$ caused by removing redundant data, is captured by $Yp(aw' - bt)$. For better processing efficiency and resource utilization in cloud computing settings, good deduplication solutions are crucial, as this equation shows in scalability analysis.

The paper analyses the many techniques of data deduplication in detail along with the applications. Important operations such source-side deduplication which reduces bandwidth utilisation by identifying unique data before transmission and target-side deduplication where a backup server performs redundancy elimination show great impact. Furthermore discussed are secure data outsourcing techniques employing erasure coding for redundancy, therefore ensuring data availability in cloud systems. By way of secure indexing and duplicate detection, the Efficient Cloud Optimisation Technique (DD-ECOT) presented in what follows additionally improves storage performance, therefore providing reliable, reasonably priced cloud storage solutions and a fingerprint-based approach for verifying data integrity.

Without adequate encryption or access control systems, sensitive information might be exposed during deduplication processes that compare and analyze data chunks. Attacks like hash collisions or illegal reverse engineering of data fingerprints might compromise hash-based indexing, notwithstanding its efficiency. Data breaches or compromised user confidentiality might result from these concerns, particularly in cloud systems with multi-tenancy and external service providers. The reliability and acceptability of the suggested solution might be enhanced by addressing these ethical problems with strong privacy-preserving approaches, secure hashing algorithms, and compliance with data protection rules.

4. Results and Discussion

Cloud data deduplication solutions are evaluated based on storage economy, latency, data integrity, bandwidth usage, and scalability. DD-ECOT uses secure hash-based indexing, advanced pattern recognition algorithms, and parallel processing. Hardware geared for distributed cloud infrastructure includes multi-core CPUs and rapid storage. GPUs do pattern recognition. Python and C++ are used to construct algorithms because they can analyze enormous volumes of data and work with machine learning frameworks like TensorFlow or PyTorch for real-time pattern recognition. Secure indexing uses cryptographic libraries like OpenSSL to produce and verify hashes, while Apache Hadoop enables scalable parallel processing. The solution uses efficient storage layers like Google Cloud Storage to manage deduplication data, ensuring cloud system compatibility. DD-ECOT performance benchmarks can be repeated using these settings.

Dataset description:

Software and service companies are very important in modern digital terrain in helping different industries to negotiate challenging tasks. Especially with the development of Internet of Things (IoT) devices and dependence on cloud computing, data compliance and privacy problems top priority. Safeguarding data security against cyber-attacks and breaches becomes very essential as companies welcome mobility and use big data and machine learning (ML). While attention to data quality and management of entity information and information sources drives informed decisions aligned with consumer demand and market dynamics, so shaping pricing, specifications, and omnichannel data strategies; the IT industry and hardware business are indispensable in providing strong infrastructure.

The data deduplication tools market is expected to reach 5.7 billion USD by 2032, from 2.3 billion USD in 2023, representing a CAGR (compound annual growth rate) of 10.9% over the forecast period. This strongly indicates that the industry is doing well. The exponential increase of data creation globally is fueling this tremendous boom, which is, in turn, driven by the rising demand for effective data management solutions across a range of sectors. A key factor propelling the market's expansion is the widespread availability of digital material and the increasing popularity of cloud-based solutions.

A data deduplication tools market research report from 2032 will probably go over how these tools deal with different kinds of data, such as structured data (like client records or financial transactions), unstructured data (like emails, photos, or videos), and semi-structured data (like JSON or XML files). This study will evaluate the complexity and efficacy of deduplication algorithms by taking into account variables such as data volume (in terabytes or petabytes), velocity (in gigabytes per day), and diversity (in the number of data types included in a dataset).

Table 1. Storage Utilization Efficiency Analysis

Number of Samples	SDS	DBCS	A-BE	CSM	DD-ECOT
10	34.2	43.2	29.3	71.3	84.3
20	77.6	68.6	40.5	12.5	6.5
30	80.4	24.4	69.3	57.7	15.2
40	22.8	8.8	35.5	18.6	72.4
50	50.9	50.9	3.9	3.9	96.8
60	13.6	76.5	6.2	88.5	85.9
70	94.3	41.6	39.1	14.4	50.3
80	31.2	80.5	82.9	31.8	82.4
90	79.4	36.4	62.4	24.5	22.8
100	25.9	58.2	33.7	85.6	92.2

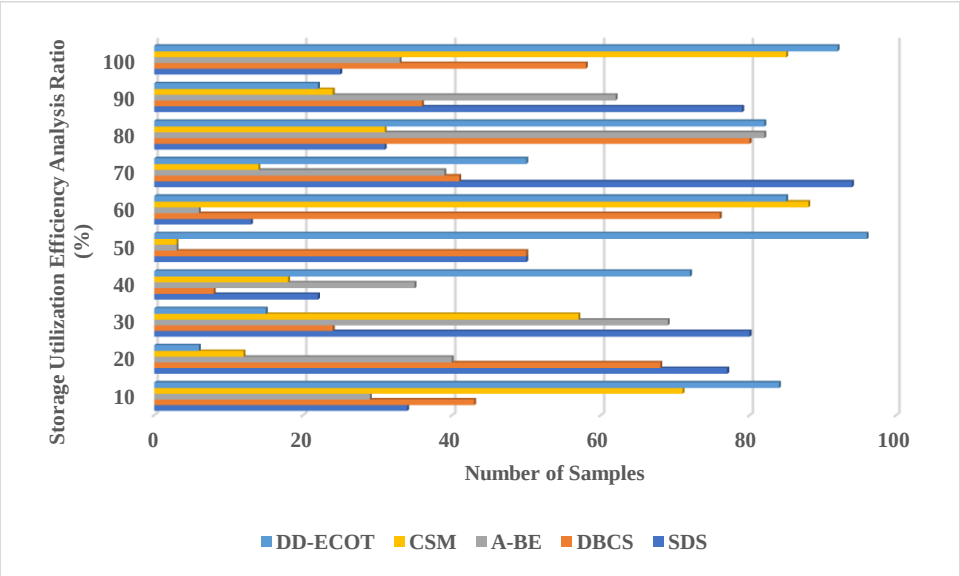

Fig. 8. Storage Utilization Efficiency Analysis

In figure 8 and Table, storage utilisation efficiency is crucial when assessing a cloud data deduplication solution's performance and cost. Deduplication solutions eliminate redundant data, lower data transmission and storage costs, and optimise storage resources. Content-aware algorithms and chunk-based deduplication can optimise storage capacity. These approaches can detect duplicate data in several data formats. Unique data must be stored to do this. Machine learning and other advanced methods can dynamically respond to changing data patterns, increasing deduplication rates and storage efficiency. High storage utilisation efficiency is essential, are data discrepancies, security problems, and processing delays. System performance and aggressive deduplication compatibility are crucial. The introduction of

latency, which will harm the amount of time it takes to retrieve data, is possible if this does not occur. A complete evaluation of storage utilisation's efficacy must incorporate the advantages of reduced storage demands and system performance losses, producing 92.8%. The former has more benefits than the latter. New deduplication algorithms that favour storage efficiency while maintaining system integrity and performance are needed to optimise cloud resource utilisation. Costs will drop and efficiency will improve. Because these algorithms prioritise storage efficiency over system speed and integrity.

Table 2. Data Deduplication Latency Analysis

Number of Samples	SDS	DBCS	A-BE	CSM	DD-ECOT
10	14.2	51.3	83.3	62.3	87.3
20	38.6	69.5	32.5	6.5	50.2
30	7.8	18.4	41.4	66.4	1.5
40	56.4	26.8	39.1	34.1	47.5
50	91.6	60.6	60.2	7.2	7.5
60	9.3	93.4	81.3	57.9	46.5
70	70.2	95.6	44.8	8.8	86.4
80	48.4	76.3	18.9	13.7	76.6
90	62.7	28.6	37.6	33.5	41.2
100	66.9	78.9	84.4	89.6	97.1

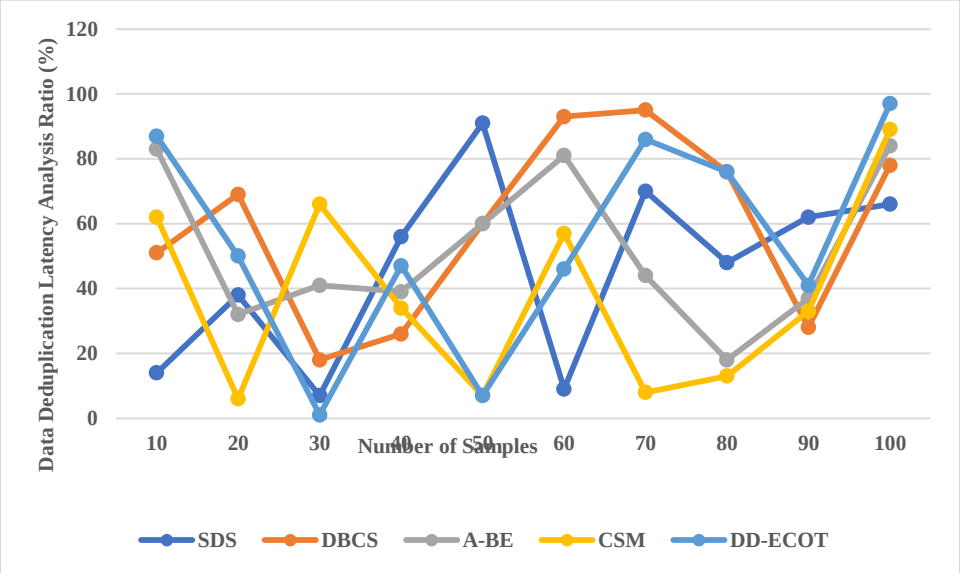

Fig. 9. Data Deduplication Latency Analysis

When assessing cloud data deduplication solutions, latency research is crucial since it impacts user experience and system performance. In the above figure 7 and table 2, duration of deduplication is called latency. The cloud architecture, data volume, and deduplication algorithm complexity may cause this delay. Traditional deduplication methods have high latency due to intensive hashing and comparison. When working with big datasets or data streams that are happening at rapid speeds, this becomes much more obvious. The benefits of storage savings may be countered by slower data retrieval and a worse user experience. Modern techniques like in-memory deduplication and parallel processing reduce latency. These methods speed up duplicate data detection and removal without affecting system performance. Machine learning algorithms may additionally speed up deduplication and processing by anticipating and responding to data trends produces 97.2%. Latency studies can improve deduplication systems and uncover bottlenecks, helping companies balance storage efficiency and data access speed. Cloud services reduce latency, improving data access speeds and dependability and operational efficiency.

The DD-ECOT system reduced latency by 97.2% due to its advanced parallel processing and deduplication capabilities. These improvements allow the system to discover and eliminate duplicates faster without straining its resources. Intelligent pattern recognition and safe hash-based indexing speed up comparisons, while parallel processing handles massive datasets. However, indexing and hashing may become inefficient when the system faces highly diversified or often changing data. Latency advantages may be reduced if the cloud infrastructure cannot handle parallelism or unexpected network restrictions, emphasizing the need for robust resource allocation and flexibility.

Table 3. Data Integrity Analysis

Number of Samples	SDS	DBCS	A-BE	CSM	DD-ECOT
10	22.3	75.3	35.3	20.1	76.3
20	50.5	77.2	6.2	9.8	28.2
30	69.7	59.5	65.6	81.5	38.5
40	24.8	40.6	66.4	76.4	15.4
50	10.9	90.9	26.1.	95.8	49.6
60	68.6	78.8	25.5	13.2	3.3
70	73.5	22.7	72.3	34.1	57.8
80	83.4	70.4	41.1	53.6	2.7
90	93.5	4.1	64.7	30.5	29.2
100	86.7	55.7	43.9	32.4	98.3

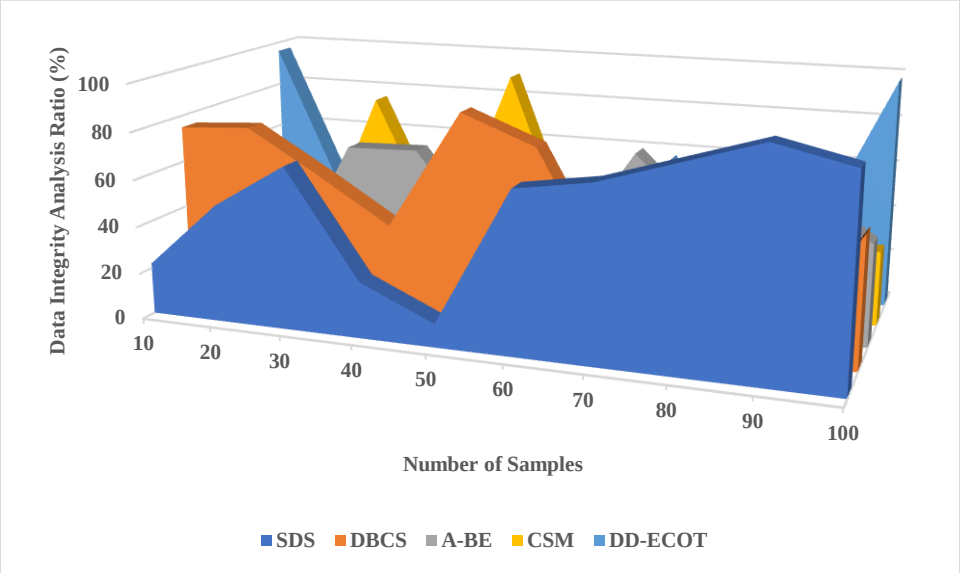

Fig. 10. Data Integrity Analysis

In the above figure 8 and table 3, evaluation of cloud data deduplication systems should include data integrity analysis. It is essential that the data be accurate and consistent during the entire process of deduplication, and the study conducted guarantees that this will be the case. When enormous amounts of data are stored and accessible in cloud computing, data integrity breaches can cause operational disruptions, financial losses, and legal issues. Normal deduplication may accidentally harm data. Particularly if duplication detection methods cannot handle different data types or architectures. Data integrity is improved by modern deduplication methods using secure hash-based indexing. They create unique fingerprints for each data component, enabling reliable verification and retrieval. Prior to and after deduplication, checksums and integrity tests reveal data errors. Avoids data manipulation. Ongoing validation and integrity audits increase deduplication system trust. Prioritising data integrity can help organisations safeguard their cloud infrastructure and reduce duplicate data hazards produces 98.1%. Building trust in cloud services, fostering user confidence, and complying with regulatory obligations in a data-driven landscape requires a comprehensive data integrity investigation.

Table 4. Bandwidth Utilization Analysis

Number of Samples	SDS	DBCS	A-BE	CSM	DD-ECOT
10	44.2	7.3	14.1	27.7	37.3
20	77.3	2.5	52.5	78.8	86.4
30	76.6	46.4	63.9	28.6	59.5
40	12.4	29.8	18.8	86.5	78.6
50	71.8	85.8	20.7	37.4	27.9
60	83.7	54.6	32.5	38.2	69.7
70	29.9	42.2	95.4	48.3	59.5
80	28.6	49.1	83.6	29.1	22.2
90	86.4	16.6	59.3	38.9	59.4
100	72.2	70.8	88.1	69.8	95.8

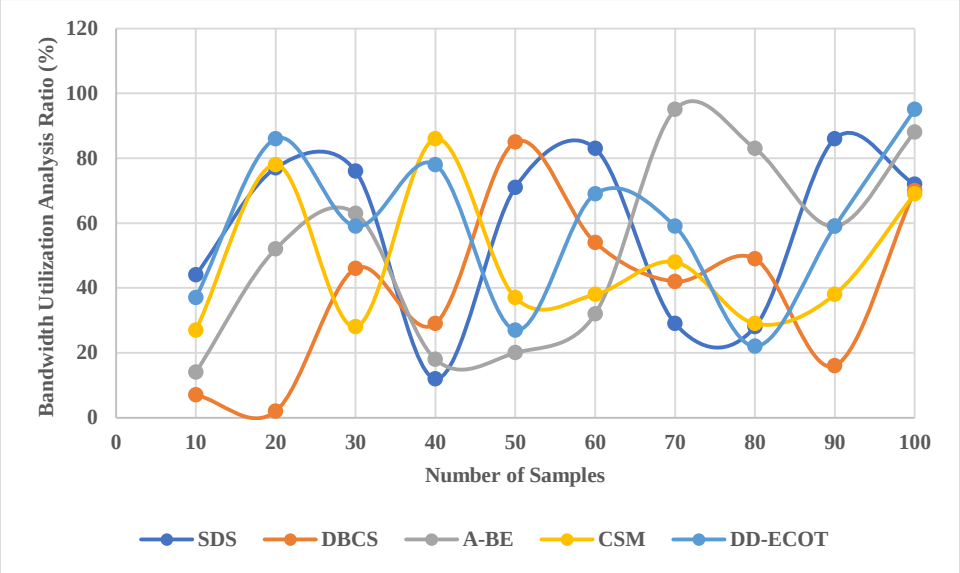

Fig. 11. Bandwidth Utilization Analysis

Bandwidth use analysis is essential when evaluating cloud data deduplication systems since it influences data transfer rates and system performance. In the above figure 9 and table 4, cloud systems' high bandwidth consumption, where data is frequently uploaded, downloaded, and replicated, might increase operational expenses and user access times. This is possible with data replication. Deduplication technologies reduce bandwidth by finding and removing duplicate data before transmission. With big datasets, duplicate data can increase bandwidth consumption and cause delays and congestion, making this crucial. Smart deduplication methods like client-side deduplication can process data before uploading to the cloud. This greatly reduces the quantity of data delivered. Communication and bandwidth efficiency are optimised by content-aware and chunk-based deduplication algorithms that target duplicate data chunks. This is done by focused on copies. Smart caching systems preserve frequently accessed data locally, reducing data transfers and bandwidth usage produces 95.7%. A thorough bandwidth consumption analysis can help businesses improve user experience, save money, and boost data transfer efficiency. This will ensure data-intensive environments use cloud resources efficiently and sustainably.

Table 5. Scalability Analysis

Number of Samples	SDS	DBCS	A-BE	CSM	DD-ECOT
10	5.2	7.3	35.9	27.3	22.7
20	68.3	2.5	6.6	78.6	50.9
30	91.6	46.6	65.5	28.5	69.3
40	87.8	29.4	66.3	86.8	24.4
50	30.4	85.8	26.2	37.9	10.6
60	99.5	54.6	25.4	38.7	68.8
70	19.9	42.8	72.7	48.6	73.4
80	44.7	49.4	41.8	29.3	83.9
90	23.5	16.1	64.9	38.4	57.4
100	34.6	70.2	43.5	69.7	98.1

In the above figure 10 and table 5, assessing cloud data deduplication strategies requires scalability analysis. This is crucial as many firms are using cloud services to manage their growing data loads. Computer scalability is the capacity to manage growing data quantities without sacrificing performance or efficiency. The best cloud deduplication solutions can adapt to changing data loads while optimising storage use and reducing latency. Large data sets often because scalability concerns with standard deduplication methods. Overloading them slows processing and raises operational costs. Distributed deduplication uses numerous cloud nodes to share the deduplication task, enhancing scalability. Parallelism speeds up deduplication and balances processing needs. Even during extreme congestion, cloud-native deduplication algorithms may dynamically assign resources to meet real-time data requests, maintaining performance. Another option to scale is to use machine learning algorithms to predict data trends and optimise deduplication produces 98.6%. Companies can ensure their deduplication solutions remain effective and efficient as data volume grows by doing a thorough scalability study. Long-term, this will enable cloud services work well in data-rich environments.

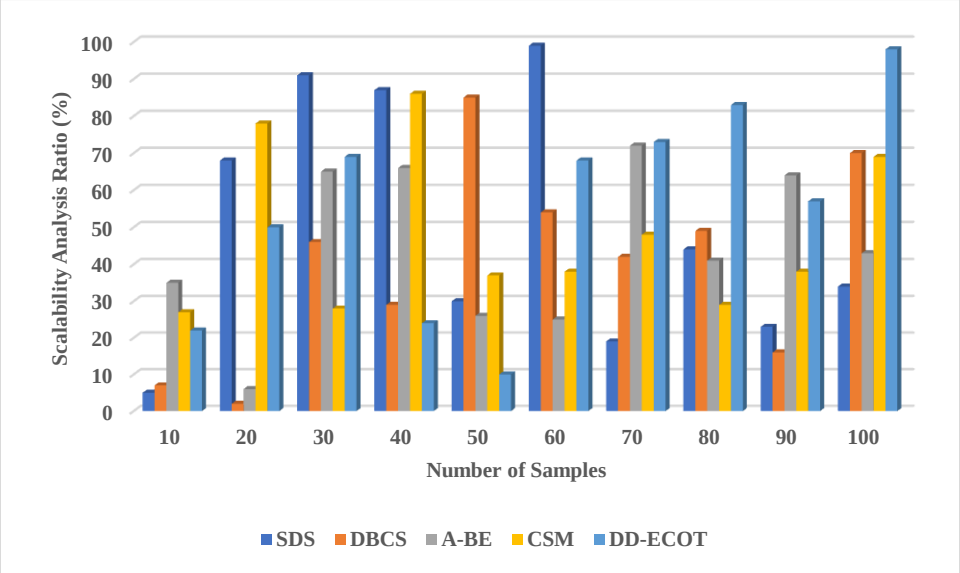

Fig. 12. Scalability Analysis

With sophisticated pattern recognition algorithms and chunking techniques, the suggested approach, DD-ECOT, can efficiently manage massive datasets in real-world cloud settings, detecting and removing duplicate data with little delay and resource use. It promises outstanding performance even under heavy workloads by using parallel processing to speed up the deduplication process. Safe hash-based indexing further ensures data integrity, and the system's extensible design makes it easy to adjust to increasing data loads. With these capabilities, storage usage is optimized, operating expenses are reduced, and system efficiency is maintained, making it ideal for handling the demands of massive datasets in ever-changing cloud settings.

Effective deduplication algorithms increase storage efficiency by 92.8% and reduce latency by 97.2%, improving user experience. Additionally, sophisticated procedures assure data integrity during deduplication to 98.1%, reducing risks. Scalable deduplication solutions can handle expanding data volumes, however optimal bandwidth utilisation accounts for 95.7% of data transfer speeds. All of these criteria show that advanced deduplication solutions are essential for cloud performance optimisation and data-driven organisations. This research compares DD-ECOT against SDS, DBCS, A-BE, and CSM. Due to its extensive delays, SDS struggles with deduplication but is brilliant at encrypting data. Dynamic data reconfiguration and inline deduplication make DBCS efficient, but it struggles to scale with large datasets. A-BE integrates deduplication with encryption techniques like Bloom filters to increase security despite its memory and performance overheads. CSM balances processing efficiency with server-side and client-side deduplication but doesn't continuously adapt to workloads. DD-ECOT overcomes all of these approaches' flaws. It effectively deduplicates data using secure hash-based indexing and powerful pattern recognition, reduces latency with parallel processing, and scales in different and dynamic cloud settings.

Incorporating DD-ECOT performance metrics with well-established benchmarks like SPEC Cloud or TPC Benchmarks might significantly enhance storage efficiency and latency. These generally acknowledged models are the bee's knees for evaluating cloud, data management, and system performance. These standards would not only let readers compare the proposed method to industry standards, but they would also provide credence to the claims made in the research. To strengthen the method's credibility and highlight its practical benefit, it would be helpful to compare DD-ECOT's performance on throughput, latency, and storage use to SPEC Cloud 2016 or TPC-DS (Decision Support) benchmarks.

One potential use case for DD-ECOT is in a large-scale e-commerce platform's disaster recovery arrangement. These systems produce massive volumes of user and transaction data, which is frequently duplicated over many backups to provide high availability. According to the simulation results, the storage needs might be cut in half by including DD-ECOT into the backup and recovery procedure, which would allow the system to deduplicate redundant data in real-time. In a simulated disaster recovery scenario, the platform was able to reduce data retrieval latency by 97.2%, which allowed essential services to be restored sooner. Data integrity was guaranteed by the secure hash-based indexing, which prevented corruption during recovery. Data transferred without a hitch across geographically dispersed cloud servers thanks to a 95.7% optimization in bandwidth utilization. In practical settings, this shown that DD-ECOT may greatly improve disaster recovery operations in terms of efficiency, speed, and dependability.

5. Conclusion

The DD-ECOT may solve cloud data management issues, by improving storage efficiency, DD-ECOT reduces redundant data footprint. Thus, resources are used more efficiently and operational costs are decreased. Modern pattern recognition algorithms and fragmentation can swiftly locate and remove duplicates. This optimises processing with minimal system resource use and reduces delay. Secure hash-based indexing techniques in cloud environments handle data integrity and security challenges. This increases user trust and ensures data protection, the simulation findings show that DD-ECOT improves data storage efficiency, retrieval speed, and system performance over standard deduplication approaches. As cloud computing evolves, creative solutions like DD-ECOT are needed. These will optimise storage and service delivery, making cloud architecture more efficient and cost-effective. This technology could revolutionise disaster recovery, enterprise cloud storage, and large-scale data management by raising cloud efficiency.

References

[1] Yanamala, A. K. Y. (2024). Optimizing Data Storage in Cloud Computing: Techniques and Best Practices. International Journal of Advanced Engineering Technologies and Innovations, 1(3), 476-513.

[2] Adhab, A. H., &Hussien, N. A. (2022). Techniques of Data Deduplication for Cloud Storage: A Review.

[3] Rajkumar, K., &Dhanakoti, V. (2020, December). Methodological Methods to Improve the Efficiency of Cloud Storage by applying De-duplication Techniques in Cloud Computing. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 876-884). IEEE.

[4] Cheng, G., Guo, D., Luo, L., Xia, J., &Gu, S. (2021). LOFS: A lightweight online file storage strategy for effective data deduplication at network edge. *IEEE Transactions on Parallel and Distributed Systems*, *33*(10), 2263-2276.

[5] Malathi, P., &Suganthidevi, S. (2021, September). Comparative study and secure data deduplication techniques for cloud computing storage. In *2021 international conference on innovative computing, intelligent communication and smart electrical systems (ICSES)* (pp. 1-5). IEEE.

[6] Akbar, M., Ahmad, I., Mirza, M., Ali, M., &Barmavatu, P. (2024). Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach. *Cluster Computing*, *27*(3), 3683-3702.

[7] Sharma, P. C., Bansal, S., Raja, R., Thwe, P. M., Htay, M. M., &Hlaing, S. S. (2021). Concepts, strategies, and challenges of data deduplication. In *Data Deduplication Approaches* (pp. 37-55). Academic Press.

[8] Jeslin, J. G., & Kumar, P. M. (2022). Implementing an efficient data deduplication framework for cloud storage. *Indian J. Comput. Sci. Eng*, *13*, 136-144.

[9] Periasamy, J. K., &Latha, B. (2021). Efficient hash function–based duplication detection algorithm for data Deduplication deduction and reduction. *Concurrency and Computation: Practice and Experience*, *33*(3), e5213.

[10] Rasina Begum, B., &Chitra, P. (2023). SEEDDUP: a three-tier SEcurE data DedUPlication architecture-based storage and retrieval for cross-domains over cloud. *IETE Journal of Research*, *69*(4), 2224-2241.

[11] Koushik, C. S. N., Choubey, S. B., Choubey, A., & Sinha, G. R. (2021). Data deduplication for cloud storage. In *Data Deduplication Approaches* (pp. 307-317). Academic Press.

[12] Nannai John, S., &Mirnalinee, T. T. (2020). A novel dynamic data replication strategy to improve access efficiency of cloud storage. *Information Systems and e-Business Management*, *18*(3), 405-426.

[13] Vignesh, R., &Preethi, J. (2022). Secure data deduplication system with efficient and reliable multi-key management in cloud storage. *Journal of Internet Technology*, *23*(4), 811-825.

[14] Ellappan, M., &Abirami, S. (2021). Dynamic prime chunking algorithm for data deduplication in cloud storage. *KSII Transactions on Internet and Information Systems (TIIS)*, *15*(4), 1342-1359.

[15] GnanaJeslin, J., & Mohan Kumar, P. (2022). Decentralized and privacy sensitive data de-duplication framework for convenient big data management in cloud backup systems. *Symmetry*, *14*(7), 1392.

[16] Prajapati, P., & Shah, P. (2022). A review on secure data deduplication: Cloud storage security issue. Journal of King Saud University-Computer and Information Sciences, 34(7), 3996-4007.

[17] Mahesh, B., Pavan Kumar, K., Ramasubbareddy, S., &Swetha, E. (2020). A review on data deduplication techniques in cloud. Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco, 825-833.

[18] Kumar, P. A., Pugazhendhi, E., & Lakshmi, K. V. (2022, January). Cloud Data Storage Optimization by Using Novel De-Duplication Technique. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 436-442). IEEE.

[19] Tahir, M. U., Naqvi, M. R., Shahzad, S. K., & Iqbal, M. W. (2020, February). Resolving data de-duplication issues on cloud. In 2020 International Conference on Engineering and Emerging Technologies (ICEET) (pp. 1-5). IEEE.

[20] Chhabraa, N., &Balab, M. (2020). An optimized data duplication strategy for cloud computing: Dedup with ABE and bloom filters. International Journal of Future Generation Communication and Networking, 13(1), 824-834.

[21] Rajput, U., Shinde, S., Thakur, P., Patil, G., &Deokar, P. (2022). Analysis on deduplication techniques for storage of data in cloud. International Research Journal of Engineering and Technology, 9(5), 296-304.

[22] https://datasetsearch.research.google.com/search?src=0&query=Cloud%20Data%20Deduplication%20Strategies%20for%20Efficient%20Storage%20and%20Processing&docid=L2cvMTF2azZjY250ag%3D%3D

Authors' Profiles

Ranga Kavitha has been awarded Ph.D in Computer Science in the research area of Mobile Computing from Rayalaseema University, Kurnool along with MCA from Osmania University and MTech (CSE) from IETE, New Delhi. Has 20 years of teaching experience and has published 12 Papers in National,International Journals, Scopus, Web of Science and Conferences. Has reviewedmany conference papers and book chapters. Currently working as Associate Professor in Computer Science and Engineering Department at Siddhartha Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad. Her interested areas are Machine Learning, Cloud Computing and Security.

Shaik Mahaboob Sharief has been awarded Ph.D in Computer Science and engineering in the research area of ubiquitous Computing from Shri Jagdishprasad Jhabarmal Tibrewala University (SJJTU),India. Shaik Has 20 years of teaching experience along with 5 years of industrial experience in india and kingdom of Saudi Arabia. Shaik has published 25 Papers in International Journals, which includes Scopus, Web of Science and SCI index. Shaik has reviewed many papers in peer reviewed journals. Shaik is Currently working as Associate Professor in Computer Science and Engineering Department at Siddhartha Institute of Engineering and Technology, Ibrahimpatnam, Hyderabad.

Narala Swarnalatha has been awarded M.Tech in Computer Science and Engineering at JNTUA. Has 10 years of teaching experience currently working as an Assistant professor in Computer science and engineering at Marri Laxman Reddy Institute of Technology and Management Hyderabad. Her interested areas are Artificial Intelligence and Machine Learning.

M. Pujitha working as Assistant Professor in the CSE Department at Marri Laxman Reddy Institute of Technology & Management, Dundigal, Hyderabad. I have completed B.Tech Degree in Computer Science and Engineering from JNTUA, A.P in 2011 and completed M.Tech from JNTUA, A.P in 2014. I have 10 years' experience. I have published 5 research papers in various National and International Journals. I have attended 10 workshops(FDPs).

Syed Asadullah Hussaini, B.Tech, M.Tech, Ph.D, PDF (FMERU, FMERC), is an accomplished academician with 15 years of experience in corporate and educational sectors. He completed his B.Tech and M.Tech in Computer Science and Engineering from JNTU Hyderabad and began his career as a Software Test Engineer at VMC Arctern Pvt. Ltd. Transitioning to academia, he held roles such as Vice Principal and Director of Academics at Hi Point College of Engineering & Technology, later earning a Ph.D. from Shri JJT University, Rajasthan. Currently, he serves as an Associate Professor at ISL Engineering College, Hyderabad.Hussaini has guided 6 Ph.D. scholars and many undergoing Ph.D guidance under him, published numerous research papers, and filed patents, including *"STERBAN: Ethereum Layer Two Scaling Solution"* and a UK patent on environmental nanotoxicology.

Samiullah Khan is a seasoned information security professional with over 18 years of experience in cybersecurity, cloud security, and IT governance. Currently serving as an Information Security Manager at Bahrain Financing Company (BFC), he leads strategic security initiatives, manages cloud security posture, and ensures compliance with regulatory standards such as PCI DSS 4.0 and GDPR.Sami is also a published author, having written *"A Handbook of Zero Trust in Cloud Data Privacy"* and contributed to *"Empowering Healthcare through AI and Blockchain"* (Wiley & IEEE). He holds multiple industry certifications, including CISSP, CISA, CIPM, PMP, AWS Certified Security Specialty, and AWS Solutions Architect – Professional. A recognized Subject Matter Expert (SME) for ISC2 exam development, he actively contributes to the cybersecurity community and is a volunteer member of the Cloud Security Alliance. He holds a Master of Computer Applications (MCA) and B.Tech in Mechanical Engineering from Aligarh Muslim University, India.

Shamsher Ali received his Master in Computer Sciences and its applications (MCA) from the Aligarh Muslim University, Aligarh, India in 2001. He has been working for various finacial and Information Technology organizations like IBM, Merrill Lynch(now Bank of America), Larsen and Toubro Infotech, Citigroup NA. He is currently working for Citigroup NA in Tampa, FL as Vice President. His work and interest are mainly databases and its optimization and scalability.

How to cite this paper: Ranga Kavitha, Mahaboob Sharief Shaik, Narala Swarnalatha, M. Pujitha, Syed Asadullah Hussaini, Samiullah Khan, Shamsher Ali, "Data Deduplication-based Efficient Cloud Optimisation Technique: Optimizing Cloud Storage through Data Deduplication", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.2, pp. 129-146, 2025. DOI:10.5815/ijieeb.2025.02.07