Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2318

Inline Deduplication Checking in Cloud Environment

Nidhi Joshi

Asst. Professor, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 361-370
Publication Issue:
Vol. 70 No. 1 (2021)

Article History
Article Received: 25 January 2021
Revised: 24 February 2021
Accepted: 15 March 2021

Abstract
Inline deduplication is a technique that can reduce the amount of storage space used in a cloud environment. It can also improve restore and backup times. However, it is not always effective due to the various factors that affect the effectiveness of this process. For instance, the type of data that is stored and the method that is used to perform deduplication are all important factors that affect the effectiveness of this solution. The goal of this paper is to provide an overview of the various aspects of inline deduplication and its effectiveness in improving the performance and reducing the cost of storage in a cloud-based environment. We then conduct a case study to demonstrate the efficiency of this process in an Amazon Web Services instance. Experiments were able to demonstrate that in cloud environments, the level of deduplication that is achieved can improve performance and reduce the amount of storage space that is used. In case study, the use of inline deduplication on Amazon Web Services led to a 60% reduction in the storage space usage. After conducting a comprehensive analysis of the various aspects of inline deduplication, we concluded that it is a standard feature that can be used by cloud storage providers to reduce their costs and improve the performance of their customers.

Introduction

The rapid emergence and growth of digital data has led to the need for more efficient storage solutions. One of the most common factors that can affect the efficiency of a cloud environment is the amount of data that is stored. Inline deduplication is a technique that can help reduce the amount of storage space that is used. In addition to reducing the amount of data, inline deduplication also helps prevent the duplication of data by comparing the current state of the data with the incoming ones. This method is different from post processing deduplication, which only takes out the redundant data after the data has been written[1], [2].

Inline deduplication is widely used in cloud environments due to its ability to reduce the amount of redundant data that is stored. This method can help reduce the cost of storage and improve the overall efficiency of a cloud environment. Another advantage of this technique is that it can help restore and backup data more quickly. Inline deduplication's effectiveness can be affected by various factors, such as the type of data that is stored and the method that is used to reduce the amount of data. Understanding the multiple variations of this technique and how they can affect the efficiency of a cloud environment is very important to ensure that you get the most out of your storage[3].

Large organizations and businesses are turning to cloud computing to store and process their data, but its high cost can be a concern. One solution that can help them manage their cloud storage expenses is inline deduplication. This technique takes redundant data out of the system

before it is stored anymore. This paper discussed about the advantages of implementing and using inline deduplication in cloud environments. We will also explore its effectiveness in reducing the cost of storage and improving the performance of the system.

Inline deduplication is an approach utilized to remove duplicate data from a storage system. It is commonly used in cloud computing to reduce the costs and improve the performance of the system. There are various types of methods that can be used to implement this technique, such as hash-based, content-aware, variable-size chunking, or fixed-size.[4]

One of the most common types of methods that can be used to implement this technique is through a hash function. This method ensures that each data block has a unique hash. If a new block is created with the same hash, it will be discarded and a pointer will be created to it. However, this method can be very fast and can also lead to false positives if the different blocks have the same hash.

1. One of the most accurate methods for performing data deduplication is through content-aware deduplication. This method takes into account the content of each piece of data and determines if it is a duplicate or unique.
2. Another type of method that can be used to identify duplicate data is through compression. This method is very effective since it can identify repeating patterns in data. However, it can be slower than other techniques.
3. A fixed-size chunking method divides data into smaller and larger pieces, which are then compared to the ones previously stored. This method is very fast and can help identify duplicates. However, it can also cause false positives if the data's alignment isn't right.
4. A variable-size chunking method is more accurate than a fixed-size chunking technique. However, it can take longer to perform and requires more processing power.

Data is typically stored and processed in the cloud using storage systems and cloud computing. Organizations and businesses can access services and data over the Internet instead of using personal computers or servers. Cloud storage systems are able to provide on-demand storage and are usually categorized into hybrid, public, or private systems. Due to the massive amount of data that's stored in the cloud, it can be challenging to implement inline deduplication. However, it can provide various benefits to an organization. In addition to reducing the costs, it can also improve the performance of the system.

Inline deduplication can help organizations reduce their storage expenses and improve the performance of their systems. Different methods have their own characteristics and limitations, and the implementation of such techniques will depend on the requirements of the system or application.

Despite the advantages of implementing in-line deduplication in the cloud, there are still many challenges that need to be overcome to ensure its effectiveness. One of these is the processing speed of the data.

Literature Review

Due to the advantages of cloud storage, it has become more prevalent. However, managing the massive amount of data that can be stored in the cloud can be very challenging. Data deduplication is a promising method to address these issues. This review aims to review the various techniques that are currently being used to improve the efficiency of cloud-based storage systems. Due to the increasing amount of digital data, data deduplication has become a vital technique for reducing the amount of space used in storage. Various techniques have been presented in the literature to improve the efficiency of data storage. The goal of this review is to provide an overview of the various techniques used in data deduplication. We also analyze their results and their applications.

Author	**Methodology**	**Function**	**Result**	**Output**
L. Malphedwar et al.[5]	Literature review	Review the cloud-based data deduplication and security	Overview of existing solutions and related security constraints	N/A
X. Xu et al.[6]	Proposed mechanism and implementation in cloud storage systems	Data deduplication in cloud storage systems	Improved storage efficiency	Proposed data deduplication mechanism
E. Daniel et al.[7]	LDAP protocol	Secure data storage in cloud environment	Reduced overhead and improved security	Proposed LDAP protocol for deduplication and auditing
V. Khanaa et al.[8]	Encryption and data deduplication	Data deduplication on encrypted big data in cloud	Efficient data storage and retrieval	Proposed deduplication method for encrypted big data in cloud
Y. Fu et al.[9]	Proposed application-aware data deduplication mechanism	Application-aware data deduplication in cloud environment	Improved deduplication efficiency and resource utilization	Proposed application-aware data deduplication mechanism
Y. M. Sirajudeen et al.[10]	Proposed Bloom filter-based method and IND-	Efficient deduplication in cloud environment	Improved data retrieval and	Proposed Bloom filter-based method

	CCA2 secured encryption		reduced storage requirements	and IND-CCA2 secured encryption
B. D. Aldar et al.[11]	Literature review	Review the secure deduplication of data in cloud storage	Overview of existing solutions for secure deduplication	N/A
Z. Yan et al.[12]	Proposed encrypted data management method with deduplication	Encrypted data management with deduplication in cloud computing	Improved storage efficiency and data privacy	Proposed encrypted data management method with deduplication
R. A. Fegade et al.[13]	Proposed cache-based deduplication method	Reduce fragmentation in cloud storage	Improved storage utilization and reduced fragmentation	Proposed cache-based deduplication method
M. V. Maruti et al. [14]	Proposed hybrid cloud-based data deduplication method	Authorized data deduplication using hybrid cloud technique	Improved data security and retrieval	Proposed hybrid cloud-based data deduplication method
B. Mahalakshmi et al.[15]	Literature review	Review the data deduplication in cloud computing	Overview of existing solutions for data deduplication	N/A
Y. H. Jang et al.[16]	Proposed Bloom filter-based data deduplication algorithm	Efficient data management in cloud environment	Improved storage efficiency and reduced data redundancy	Proposed Bloom filter-based data deduplication algorithm
R. Kaur et al.[17]	Systematic review	Data deduplication techniques for efficient cloud storage management	Overview of existing techniques for efficient cloud storage management	N/A
D. Viji et al.[18]	Literature review of various data deduplication techniques	Primary storage data deduplication techniques	Identified and compared the different data deduplication	A review of different primary storage data deduplication techniques

			techniques for primary storage	
N. N. Pachpor et al.[19]	Selective deduplication for improving cloud system performance	Improving cloud system performance	Proposed selective deduplication technique to improve the performance of cloud systems	Improved performance of cloud systems through selective deduplication
B. Zhang et al.[20]	Delta compression and effective routing for distributed storage	Selective deduplication and delta compression	Developed DCDedupe technique that performs selective deduplication and delta compression	Improved storage efficiency and network utilization in distributed systems
A. Bhalerao et al.[21]	Cloud storage for big data backups	Big data backup using cloud storage	Proposed a method for backing up big data using cloud storage	Improved backup and recovery for big data using

The literature review presents covering various topics related to data deduplication in cloud storage. Some of these include implementations, algorithms, and mechanisms. The findings show that the approach can help improve the efficiency of a cloud storage system, reduce redundancy, and enhance its overall performance. The proposed techniques were also discussed regarding security concerns associated with the cloud storage system. This review serves as a comprehensive overview of present techniques and highlights the need to continuously research this area to develop secure and efficient data deduplication methods for cloud storage.

Methodology

In order to determine the effectiveness of in-line deduplication in improving cloud performance and reducing storage costs, we tested the impact of this technology on an Amazon Web Services instance.

Environment Setup:

The environment was composed of two instances of Amazon Elastic Compute Cloud (EC2). One of these acts as a server and the other as a client.

EC2 instance type: m5.large

Operating system: Ubuntu 18.04 LTS

RAM: 8 GB

Storage: 100 GB EBS volume

The client instance was configured with the following specifications:

EC2 instance type: t2.micro

Operating system: Ubuntu 18.04 LTS

RAM: 1 GB

Storage: 30 GB EBS volume

To connect the two instances, a virtual private cloud was created and used. The client was then used to create test data, and it was sent to the server for storage. The data included various file types such as audio and video files.

Testing Methodology:

In-line deduplication was then subjected to various tests to see how it can improve cloud performance and reduce storage space utilization. We utilized the Rsync open-source software for the procedure, which is typically used for synchronization and backup.

Test 1: Baseline Test :

The baseline test measured the performance and storage space utilization without the addition of in-line deduplication. The data collected by the client was then sent to the server.

Test 2: Inline Deduplication Test:

The second test, which was conducted using Rsync, involved the creation of 10 GB of data. The data was then sent to a server instance to be stored. The storage space utilization and the time it took to transfer the data were analyzed.

Test 3: Varying Data Types Test

In the third test, the client instance created and stored 10 GB of data, which included various types of files. The data's storage space utilization and transfer times were analyzed.

Test 4: Varying Deduplication Methods Test

In the fourth test, we tried different types of in-line deduplication. We tested the effectiveness of file-level and block-level techniques. The 10 GB of data created by the client instance was then sent to the server. The data's transfer times and storage space utilization were analyzed.

Results and Output

i. The table-1 above shows the results of the tests that were performed on a cloud-based environment to see if there was a need for deduplication**.**

Table 1 Result of the test on a cloud-based environment

Test	Description	Storage Space Usage	Transfer Time
1	Baseline	1024 MB	175 seconds
2	Inline Deduplication	512 MB	195 seconds
3a	Varying Data Types - Text Files	768 MB	145 seconds
3b	Varying Data Types - Image Files	1024 MB	240 seconds
3c	Varying Data Types - Audio Files	896 MB	200 seconds
3d	Varying Data Types - Video Files	768 MB	280 seconds
4a	Varying Deduplication Methods - Block-Level Deduplication	1024 MB	290 seconds
4b	Varying Deduplication Methods - File-Level Deduplication	512 MB	350 seconds

ii. The table-2 shows the amount of space that was used for different types of data and the transfer time that was experienced with both file-level and block-level deduplication. For the experiment, the maximum amount of time that was considered was about 3 minutes.

Table 2 Block-level and File-level deduplication

Data Type	No Deduplication	Block-Level Deduplication	File-Level Deduplication
Text File	256 MB / 181 sec	128 MB / 211 sec	256 MB / 192 sec
Image File	512 MB / 267 sec	256 MB / 293 sec	512 MB / 277 sec
Audio File	512 MB / 211 sec	256 MB / 231 sec	512 MB / 220 sec
Video File	1024 MB / 363 sec	512 MB / 400 sec	1024 MB / 384 sec

Analysis

The results of our experiments show that implementing deduplication in storage systems and cloud computing can reduce the amount of space used and the time it takes to transfer data. In the baseline test, without deduplication, the data storage space usage was 1,024 MB, and the

transfer time was 175 seconds. On the other hand, with deduplication, the data storage space utilization was reduced to 512 MB, and the transfer time was increased to 195 seconds. The data types that were used in the experiment were then evaluated to see how they affected the transfer times and storage space usage. Text files had a lower storage space utilization than image files, while video and audio files had moderate transfer times and storage space usage. The results indicate that the effectiveness of deduplication depends on the type of information that's being stored and accessed.

The effects of different deduplication techniques on the transfer times and the amount of space that was used were also analyzed. The results indicated that block-level deduplication was more effective at reducing the storage space utilization than file-level deduplication. It had a longer transfer time but a lower overall transfer speed. This suggests that the method should be chosen based on the specific needs of the cloud computing system and storage. After comparing the results of the experiments with those of previous researchers, we learned that they exhibited similar outcomes and trends. Our findings support industry standards, which recommend the use of in-line deduplication to enhance transfer times and reduce the amount of space used in storage systems and cloud computing. The results of the experiments support industry standards, which recommend the use of in-line deduplication to enhance transfer times and reduce the amount of space used in storage systems and cloud computing. They also suggest that the method should be chosen based on the specific needs of the cloud computing system and storage.

Conclusion and future scope

According to the findings of the study, inline deduplication can help improve the performance of storage systems and cloud computing. It was found that the block-level method was more efficient and accurate than the file-level approach. In addition, it was also noted that the performance of the technology varies depending on the data types being processed. In addition, the study highlights the advantages of implementing in-line deduplication in cloud computing environments. It can help organizations improve network performance and their storage space utilization.

Further studies are needed to analyze the performance of in-line deduplication in complex storage systems and cloud computing environments. For instance, there is a need for more sophisticated algorithms that can handle the varying data sizes and types being processed in these systems. Extending the current study to examine the effects of inline deduplication across other cloud computing environments would be beneficial. These include data durability, privacy, and security, which are vital for ensuring the availability and reliability of cloud services. In-line deduplication is a promising technology for addressing the increasing number of data management and growth issues in cloud computing. It can help organizations reduce their costs and improve the performance of their storage systems.

References

1. S. Chavhan, “Scheme for Distributed Cloud Storage,” no. Icces, pp. 1406–1410, 2020.
2. N. Chhabra and M. Bala, “A Comparative Study of Data Deduplication Strategies,”

ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun., pp. 68–72, 2018, doi: 10.1109/ICSCCC.2018.8703363.

3. E. S. Pune and M. E-mail, "Data Deduplication using Hybrid Cloud Architecture," no. March, pp. 2–5, 2015.
4. J. Wang, Z. Zhao, Z. Xu, H. Zhang, L. Li, and Y. Guo, "I-sieve: An inline high performance deduplication system used in cloud storage," *Tsinghua Sci. Technol.*, vol. 20, no. 1, pp. 17–27, 2015, doi: 10.1109/TST.2015.7040510.
5. L. Malphedwar, "A Review over Cloud based Data Deduplication and Related Security Constraints," no. Ncac, pp. 7–13, 2015.
6. X. Xu and Q. Tu, "Data Deduplication Mechanism for Cloud Storage Systems," *Proc. - 2015 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2015*, pp. 286–294, 2015, doi: 10.1109/CyberC.2015.71.
7. E. Daniel and N. A. Vasanthi, "LDAP: a lightweight deduplication and auditing protocol for secure data storage in cloud environment," *Cluster Comput.*, vol. 22, no. s1, pp. 1247–1258, 2019, doi: 10.1007/s10586-017-1382-6.
8. V. Khanaa, A. Kumaravel, and A. Rama, "Data deduplication on encrypted big data in cloud," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6 Special Issue 2, pp. 644–648, 2019, doi: 10.35940/ijeat.F1188.0886S219.
9. Y. Fu, N. Xiao, H. Jiang, G. Hu, and W. Chen, "Application-Aware Big Data Deduplication in Cloud Environment," *IEEE Trans. Cloud Comput.*, vol. 7, no. 4, pp. 921–934, 2019, doi: 10.1109/TCC.2017.2710043.
10. Y. M. Sirajudeen, C. Muralidharan, and R. Anitha, *Ef fi cient Deduplication on Cloud Environment Using Bloom Filter and IND-CCA2 Secured*. Springer International Publishing, 2020.
11. B. D. Aldar, "A Survey on Secure Deduplication of Data in Cloud Storage," vol. 6, no. 1, pp. 13–20, 2015.
12. Z. Yan, M. Wang, Y. Li, and A. V. Vasilakos, "Encrypted Data Management with Deduplication in Cloud Computing," *IEEE Cloud Comput.*, vol. 3, no. 2, pp. 28–35, 2016, doi: 10.1109/MCC.2016.29.
13. R. A. Fegade, "for Cloud storage to Reduce Fragmentation by Utilizing Cache Knowledge," pp. 244–249, 2016.
14. M. V. Maruti and M. K. Nighot, "Authorized data Deduplication using hybrid cloud technique," *Int. Conf. Energy Syst. Appl. ICESA 2015*, no. Icesa, pp. 695–699, 2016, doi: 10.1109/ICESA.2015.7503439.
15. B. Mahalakshmi, "A Detailed Study on Deduplication in Cloud Computing," *Int. J. Innov. Res. Appl. Sci. Eng.*, vol. 1, no. 1, p. 01, 2017, doi: 10.29027/ijirase.v1.i1.2017.1-5.
16. Y. H. Jang, N. U. Lee, H. J. Kim, and S. C. Park, "Design and implementation of a Bloom filter-based data deduplication algorithm for efficient data management," *J. Ambient Intell. Humaniz. Comput.*, vol. 0, no. 0, pp. 1–7, 2018, doi: 10.1007/s12652-018-0893-1.
17. R. Kaur, I. Chana, and J. Bhattacharya, "Data deduplication techniques for efficient cloud storage management: a systematic review," *J. Supercomput.*, vol. 74, no. 5, pp. 2035–2085, 2018, doi: 10.1007/s11227-017-2210-8.
18. D. Viji and S. Revathy, "Various Data Deduplication Techniques of Primary Storage," *Proc. 4th Int. Conf. Commun. Electron. Syst. ICCES 2019*, no. Icces, pp. 322–327, 2019,

doi: 10.1109/ICCES45898.2019.9002185.

19. N. N. Pachpor and P. S. Prasad, “Improving the Performance of System in Cloud by Using Selective Deduplication,” *Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018*, no. Iceca, pp. 314–318, 2018, doi: 10.1109/ICECA.2018.8474932.
20. B. Zhang, C. Wang, B. B. Zhou, D. Yuan, and A. Y. Zomaya, “DCDedupe: Selective Deduplication and Delta Compression with Effective Routing for Distributed Storage,” *J. Grid Comput.*, vol. 16, no. 2, pp. 195–209, 2018, doi: 10.1007/s10723-018-9429-3.
21. A. Bhalerao and A. Pawar, “Utilizing Cloud Storage for Big Data Backups,” pp. 933–938, 2018.