A Comprehensive Study of Different Types of Deduplication Technique in Various Dimensions

G.Sujatha[1], Dr.Jeberson Retna Raj[2]
Department of Computer Science and Engineering
School of Computing, Sathyabama Institute of Science and Technology, Chennai, India[1, 2]
Department of Networking and Communications, School of Computing
SRM Institute of Science and Technology, Chennai, India[1]

Abstract—In the current digital era, the growth of digital data is highly exceptional. There are various sources available for these digital data. The quantity of digital data being produced rose exponentially with time because of organizations and even by individuals, finally end up in the need of huge storage space. Cloud storage provides the storage space for such requirement. Since the storage space is utilized by many different users, having the duplicate data cannot be avoided. So it is necessary to make use of some storage optimization technique to handle such duplicate contents. Deduplication is a technique which is used to evade redundant data get stored. Among the various digital data, the possibility of having duplicate copies is high for data. In this research work, we review the benefits of having deduplication in optimizing the usage of storage space and study about the various types of deduplication techniques in different dimensions which can be used for data. It helps to select the appropriate data deduplication technique to increase their effective storage utilization and reduce the wastage of memory space because of duplicate data.

Keywords—Digital data; deduplication; storage optimization; cloud storage service; duplicate copies; bandwidth utilization

I. Introduction

The growth of information in the current period is very massive as shown in the Table I and Fig. 1. Many organizations and even individuals face lot of issues in storing and securing their huge volume of data.

The best way out for these issues is Cloud storage. Cloud storage is the service offered by the cloud to their users on demand [1]. It is nothing but a collection of data storage servers which can be located in different geographical locations. It can support any type of digital data like text, audio, video and image. The cloud storage providers takes the complete responsibility of data protection and also provide reliable data access. Apart from this, there are many other benefits in using cloud storage service [2]. There are many cloud storage providers existing like Microsoft Onedrive, Google drive, Dropbox, Box, Amazon Drive, and Apple icloud. Eventhough the cloud provides the storage space for their user as a service, it is not efficiently utilized because of duplicate copies of data. Such duplicate copies will occupy storage space unwantedly. This will decrease the efficient utilization of storage space. There are many storage optimization techniques which are employed to increase the better utilization of storage space. They are data compression, thin provisioning, snapshots, clones and Data deduplication. Among these techniques data deduplication [3,4] is widely used for efficient utilization of cloud storage by preventing the wastage of storage space due to duplicate copies.

The result of deduplication is shown in the Fig. 2. With the help of deduplication technique, nearly 60% of storage space wastage can be controlled.

TABLE I. DATA GROWTH RATE

Year	Data volume (ZB)
2006	0.16
2007	0.28
2008	0.48
2009	0.8
2010	1
2011	1.8
2015	8
2020	40

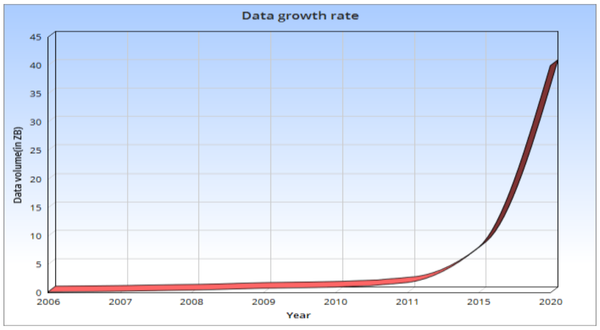

Fig. 1. Growth Rate of Data.

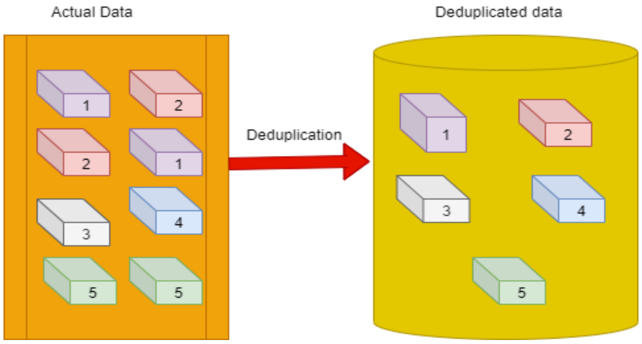

Fig. 2. Process of Deduplication.

The size of storage space prevented from redundant copies will depend upon the dataset or volume of the data. When the size of the data is very huge, then the possibility of having duplicate copies can also be very high. And it also depends upon the type of data. A sample for the same is represented in the below Table II.

TABLE II. COMPARISON OF DIFFERENT TYPES OF APPLICATION WITH RESPECT TO THEIR STORAGE SPACE PREVENTION FROM DUPLICATE COPIES

Type of Application	**Type of content**	**Approximate space saving**
User documents	Official documents, Images, Entertainment data	30-50%
Deployment shares	Software binaries, Cab files, symbols	70-80%
Virtualization libraries	ISOs, virtual hard disk files, etc.	80-95%
General File share	All types of content	50-60%

The major benefits of data deduplication are as follows:

1) Cost effective.
2) Clear storage space.
3) Clever replication.

Among the multimedia data types, audio, image and video occupy huge storage space when compared to text. Audio deduplication [5,6] is a process of finding and removing duplicate copies of audio data. This will reduce the wastage of memory due to duplicate copies of audio data. In cloud storage, storing cinema songs and favourite dialogues from movies are very common. And the possibility of having duplicate copies in such cases is also very high. If we apply deduplication techniques to avoid such duplication, it is possible to improve the utilization of storage space effectively. Data deduplication can be done in various ways. And all of the techniques we are going to discuss in the further sections is not with respect to the usage of any analytics algorithm or any of the classifier to identity the duplicate copies. In this study, we discuss about the process of deduplication and perform a comprehensive study on different digital data deduplication techniques in various dimensions that can be applied to data.

The rest of the paper flows as literature survey in Section II, Section III explain the architecture of the data deduplication process, Section IV discuss the various data deduplication techniques, Section V describes the comparison and analysis of various data deduplication techniques, and Sections V and VI briefs the conclusion and future work.

II. RELATED WORK

The various types of storage optimization techniques are categorized as Location based deduplication, Time based deduplication and Chunk based deduplication [7]. The characteristics and performance of various deduplication techniques under these categories are analysed and the author concluded that variable sized deduplication technique is comparatively better than other techniques. They also suggested to carry out the future work in optimizing the processing time of variable sized deduplication techniques.

The various chunking algorithms like Rabin fingerprint, Two Divisors, MAXP, Bimodal, MCDC, Leap-Based, AE algorithms [8] are compared and analysed their advantages and disadvantages. The authors concluded that AE algorithm is more efficient and mentioned that there is a space on chunking size variance issue for future work.

The deduplication in primary storage can be done by the following ways. They are inline, offline, Post processing and cache based algorithm [9]. The author analyzed the performance of inline, offline and cache based algorithm with respect to various criteria. They mentioned the future work as concentrating on deduplication techniques in backup storage system.

III. ARCHITECTURE OF DATA DEDUPLICATION

Image, Video and Audio data are most frequently handled by many users and even many users store their digital data in cloud storage space. And the possibility of having duplicate files from same user or from different user is also very high. For example, duplicate copies of songs or favorite dialogues of movie. This will actually decrease the effectual utilization of cloud storage space. To overcome this problem, we have the concept called data deduplication.

Fig. 3 represents the architecture of data deduplication. The following are the steps which are followed in the process of Data deduplication:

1) When a input data is to be uploaded in the storage space, the user initially raises the request for the storage provider to do the same.

2) The cloud storage provider has to decide the deduplication technique.

3) Then the respective deduplication technique will be carried out to find the duplicate copy.

4) When it finds a match, then it is considered as a duplicate data and it should not be stored again in the storage space.

5) But reference to access the data will be shared with the user.

6) The user can get the data whenever it is required.

7) When there is no match, the input data will be uploaded in the storage space.

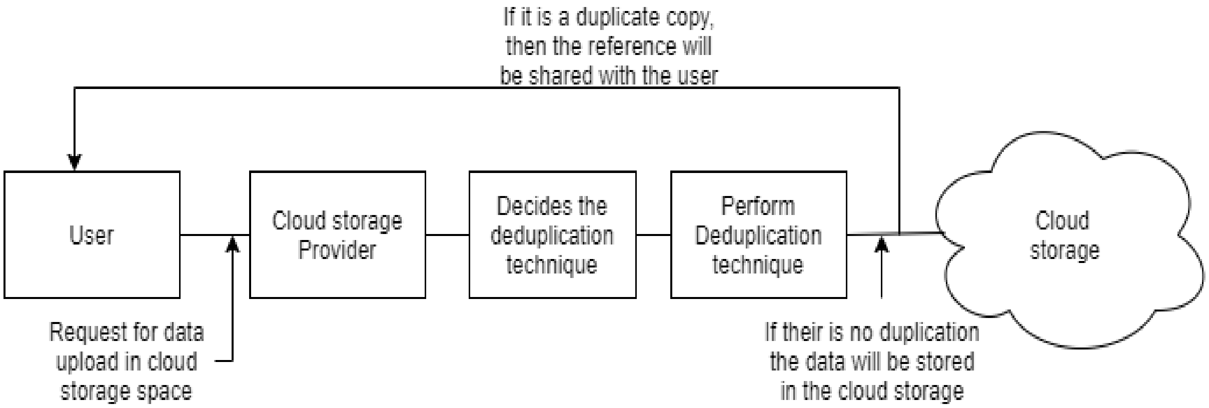
If it is a duplicate copy, then the reference will be shared with the user
User
Cloud storage Provider
Decides the deduplication technique
Perform Deduplication technique
Cloud storage
Request for data upload in cloud storage space
If their is no duplication the data will be stored in the cloud storage

Fig. 3. Architecture of Data Deduplication.

IV. Types of Digital Data Deduplication

Data deduplication will reduce the cost of storage space to be spent by the organisation. And it is comparatively better than compression technique. There are various types of deduplication exists as shown in the Fig. 4 and we can also view this process in different dimensions. They are where, when and how. The Different types of Data Deduplication in various dimensions are:

1) Source deduplication.
   a) Local source deduplication.
   b) Global source deduplication.
2) Target deduplication.
3) Inline deduplication.
4) Post-process deduplication.
5) Content-based/hash based deduplication.
6) Content-aware deduplication.
7) Chunk-level deduplication.
   a) Fixed sized deduplication.
   b) variable sized deduplication.
8) File-level deduplication.

A. Source Deduplication

Source deduplication [10] is the method of identification and removal of duplicate copy before the data is getting transmitted to the backup storage system. It does the data deduplicationin the client side as shown in the Fig. 5. This will work with the help of client software which is used to communicate with the storage device to check whether the data to be stored is already present in the backup storage space. The advantage of using source deduplication is utilization of low bandwith for data transmission but the disadvantage is, it uses client resources for the entire deduplication process. This includes generation of fingerprint for the comparison and the process of identifying the duplicate copy. This source deduplication can be done by two ways. They are local deduplication and global deduplication. In local-level deduplication, the duplicate copy will be initially identified locally, then proceed with the backup process which may depends upon the presence of that particular data in the backup storage. Here the deduplication will be done at that device only. In case of global deduplication, the fingerprint of the data is calculated in the client side and that is getting transferred to the target storage system to compare with the existing data fingerprint. And the data will be stored only when it does not already exist in the storage space. So the deduplication in done across all the users and their devices.

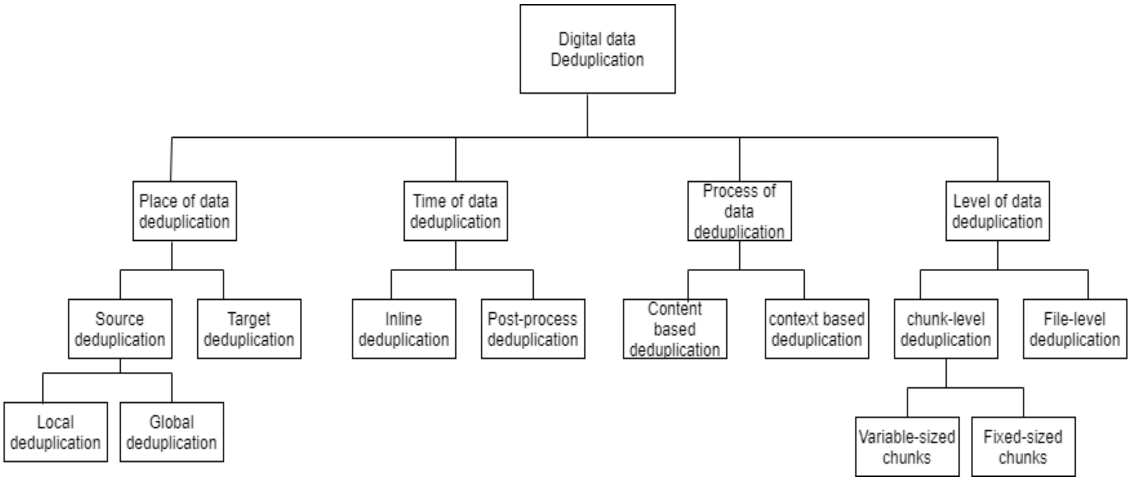
Digital data Deduplication
Place of data deduplication
Time of data deduplication
Process of data deduplication
Level of data deduplication
Source deduplication
Target deduplication
Inline deduplication
Post-process deduplication
Content based deduplication
context based deduplication
chunk-level deduplication
File-level deduplication
Local deduplication
Global deduplication
Variable-sized chunks
Fixed-sized chunks

Fig. 4. Different Types of Data Deduplication in Various Dimensions.

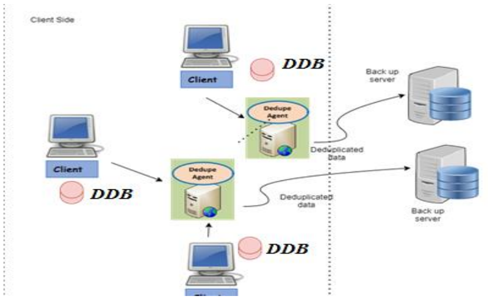

Fig. 5. Client- Side Data Deduplication.

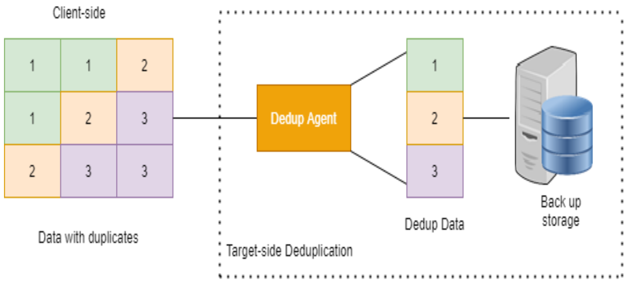

Fig. 6. Target-Side Data Deduplication.

B. Target Data Deduplication

The method of target data deduplication is shown in the Fig. 6. In Target data deduplication [11], the deduplication process is done in the backup server storage system. In the backup side, a dedicated device is employed to carry out the deduplication process. This will reduce the overhead in client side. But the disadvantage is, it requires more bandwidth for data transmission since it may contains duplicate data too.

C. Inline Data Deduplication

The inline data deduplication [12,13,14] can be otherwise called as synchronous deduplication as it allows the data can be stored only when it is unique and not already present in the storage. Not all the data is getting written in the storage space, but only the unique. It is shown in the Fig. 7.

D. Post-process Data Deduplication

In Post-process data deduplication [15], the data is initally written in the storage space as it comes as shown in the Fig. 8. Then the deduplication process will optimize the storage space with unique data. The time of performing this deduplication process is varied with various systems. It can be in seconds, minutes or even hours after the data got stored in the storage space.

E. Content-based or Hash-based Deduplication

The deduplication is performed by considering the content of the digital data [16]. A hash value is generated for the content and that hash value is used to check for the duplication. Any hashing algorithm like MD5, SHA-5256, SHA-512 can be used to calculate the hash value for the digital data and it is shown in the Fig. 9.

F. Content-aware Deduplication

Content-aware deduplication [17] considers the data as an object for the deduplication process. It does the process by comparing object with the other objects. For example if the input file is word document, then it restricts its comparison only with all other word documents existing in the storage. It achieves byte-level deduplication. This content-aware technique tries to find the similar segments or bytes and those bytes which are really changed or unique will alone be stored in the storage space. The steps in the content-aware deduplication are shown in the Fig. 10.

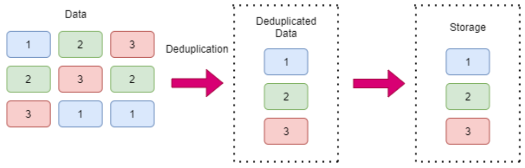

Fig. 7. Inline Data Deduplication.

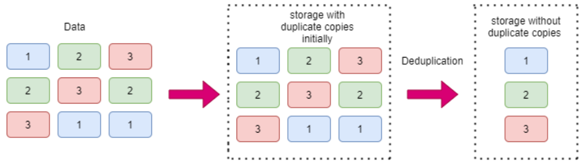
Data
1
2
3
2
3
2
3
1
1
storage with duplicate copies initially
1
2
3
2
3
2
3
1
1
Deduplication
storage without duplicate copies
1
2
3

Fig. 8. Post-process Data Deduplication.

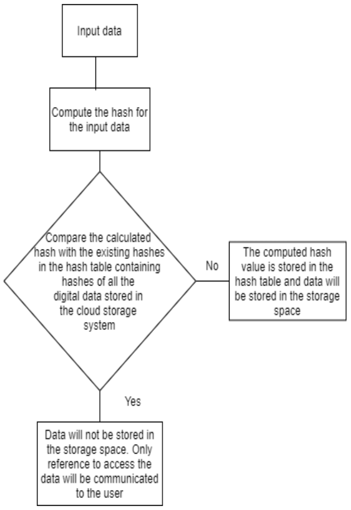
Input data
Compute the hash for the input data
Compare the calculated hash with the existing hashes in the hash table containing hashes of all the digital data stored in the cloud storage system
No
The computed hash value is stored in the hash table and data will be stored in the storage space
Yes
Data will not be stored in the storage space. Only reference to access the data will be communicated to the user

Fig. 9. Content-based Data Deduplication.

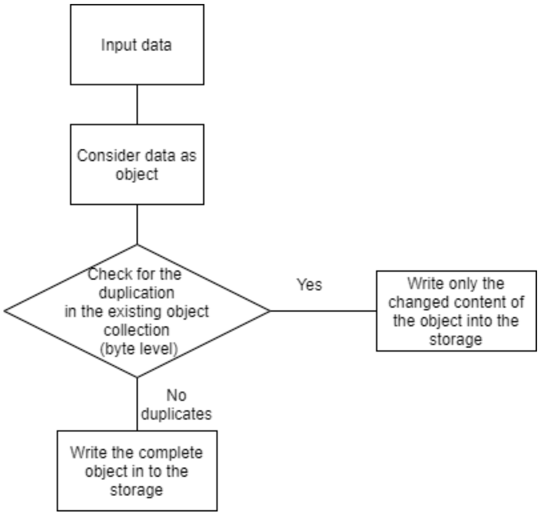
Input data
Consider data as object
Check for the duplication in the existing object collection (byte level)
Yes
Write only the changed content of the object into the storage
No duplicates
Write the complete object in to the storage

Fig. 10. Content-aware Data Deduplication.

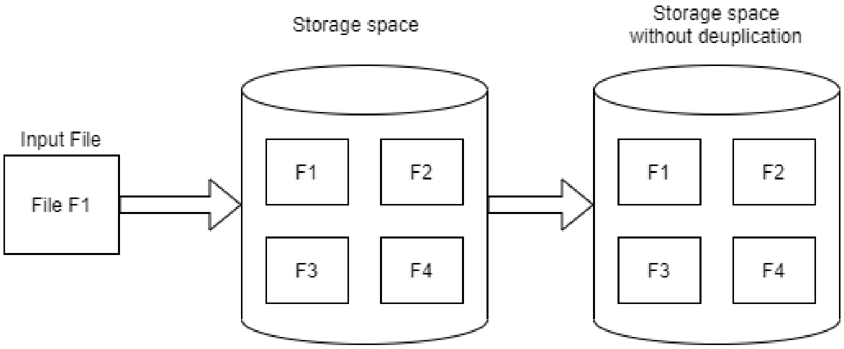

Fig. 11. File-level Deduplication.

G. File-level Deduplication

In file-level deduplication [18] as shown in the Fig. 11, the complete file is taken for comparison. According to the process of deduplication, for example, if hash-based deduplication is decided, then the hash value is calculated for the complete file and the calculated hash value is used to compare the hash values of the existing files stored in the hash table.

H. Chunk Level Deduplication

In chunk level deduplication technique [19,20] the entire content of the digital data is splitted into various chunks. The data upload process involves the following steps. Initially the entire content is divided into chunks. There are many algorithms existing for chunk creation. Then the hash value will be calculated for each and every chunk. Using those hash values duplicity of particular chunk can be identified. This will increase the granularity or degree of deduplication. Which means the wastage of memory is even reduced due to this chunk level deduplication. The chunk level deduplication can be done in two ways. They are Fixed-size chunking [21] and variable-size chunking. In fixed-size chunking, the entire content is divided into equal size chunks. But in variable-size chunking, it is not manditory that all the chunks will be in the same size. But it requires additional computation for each byte [22]. Fig. 12 and Fig. 13 show the process of chunk-level-Fixed size and Variable size deduplication process.

1) Chunk level - fixed size deduplication: In Fixed-size chunk-level deduplication technique, the entire content is divided into equal size chunks. The duplicate copy of chunk-level data can be identified with the help of this technique. For example, if a user wants to upload their data in the cloud storage space, then the data must be divided into chunks and the duplicity of each chunk will be checked before storing the chunk in the storage space. When any of the chunks is already present in the storage space, then that particular chunk alone not gets stored again. This will improve the utilization of storage space.

2) Chunk level- variable sized deduplication: In variable-size chunk-level deduplication technique, the entire content is divided into variable size chunks. Which means, each chunk of any size and not compulsorily same. The system will decide the boundary of the each chunks. It also divides the data in content-based strategy. Actually because of this variable-size chunks, the process of deduplication yield very good results in the finding of duplicate copies.

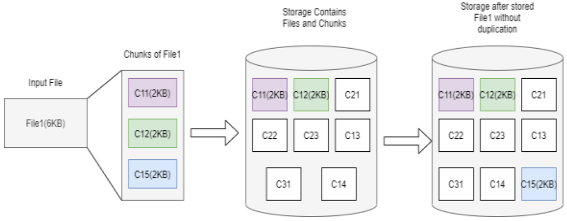

Fig. 12. Chunk-level –Fixed Size Deduplication.

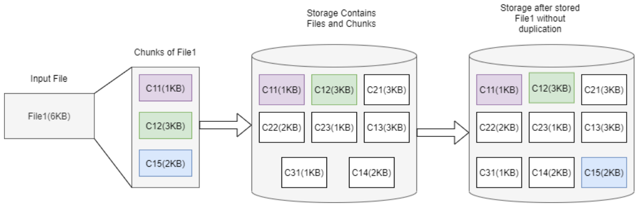

Fig. 13. Chunk-level –Variable Size Deduplication.

V. Comparison and Analysis of Various Deduplication Techniques

A. Source vs Target based Deduplication Techniques

Table III shows the comparison of Source and Target deduplication techniques in various metrics. The source based deduplication requires less LAN bandwidth as it has the deduplication process in the client-side itself. But it requires more client resources for handling deduplication. In target based deduplication the process overhead in the client-side is less and comparatively fast in process.

TABLE III. Source vs Target based Deduplication Techniques

Metrics	**Source based Deduplication**	**Target based Deduplication**
Place of deduplication	Deduplication occurs at client side	Deduplication occurs at backup medium.
Utilization of Bandwidth	Requires Less LAN bandwidth	Required more LAN bandwidth
Client resources	Requires more client resources	Required less client resources
Process overhead at client	More	Less
Speed	Comparitively slow	Fast

B. Inline vs Post-process Deduplication Techniques

Table IV shows the comparison of Inline and Post-process deduplication techniques. Inline deduplication has better storage throughput when compared to post-process deduplication. The Post-process deduplication is faster than in-line process in terms of storage performance.

TABLE IV. INLINE VS POST-PROCESS DEDUPLICATION TECHNIQUES

Metrics	InlineDeduplication	Post-process Deduplication
Time of deduplicatin	Deduplication occur at the time of data flow	Deduplication occur after it has been written
Storage performance	Slow	Fast. Since the hash calculation is deferred.
Storage requirement and network traffic	Less	Comparatively more
Storage throughput	Reduced storage throughput	Better comparatively
Temporary storage space	Not required	Required

C. Content-aware vs Content-based Deduplication Techniques

Table V shows the comparison of Content-aware and Content-based deduplication techniques. It shows that Content-aware deduplication process has more efficiency.

TABLE V. CONTENT-AWARE VS CONTENT-BASED DEDUPLICATION TECHNIQUES

Metrics	Content-aware Deduplication	Content-based Deduplication
Granularity of deduplication	Even Byte level is possible	Chunk/Block level is possible
Efficiency	More when compared to content-based deduplication	Better
Execution time	Fast. Since it handles data in the form of object. The comparison will be done with the same type of objects only and not with all.	Slow. Since it has to check for all the chunks individually
Metadata overhead	Additional metadata is required to store the type of the object.	Usual details required to handle chunk-based informations.

D. File Level vs Chunk Level Deduplication Techniques

Table VI shows the comparison of File-level and chunk-level deduplication techniques. The chunk-level deduplication process has good efficiency since it does the deduplication at chunk level. And it requires more resource and also it has high computational complexity.

Consider in the storage space, it has audio file F1already and the new request comes to store another audio file F2. The contents of the file F2 is a part of the File F1.

The MD5 hash value calculated for File F1 is 34195925cbe68a7dc78859b93e13e33e and the hash value calculated for File F2 is 09fd1195ba1c83e51966feb33faa4a80.

In File-level deduplication technique, the files will be considered as different files since their hash values are different. But actually the content of File F2 is the part of the File F1 content. But both are considered as different files in File-level deduplication technique. In chunk-level deduplication technique, the files will be divided into chunks and hash value will be calculated for each chunks. Then the chunk-level duplication can be checked.

Using this chunk-level deduplication technique, the content of F2 can be identified as duplicate copy. So the content of File F2 will not get stored again in the storage space. This will improve the utilization of storage space. Assume, if the size of the File F1 is 9.85KB and size of file F2 is 9.78KB (without compression). The memory consumption for these two files in storage space is as shown in the Table VII and Fig. 14.

From the observation, using chunk-level deduplication technique, the storage space utilization can be improved without storing duplicate copies of data.

E. Fixed-size vs Variable-size Chunk Level Deduplication Techniques

From the Table VIII, it is observed that Variable-size chunk-level deduplication is a better choice since it does the chunking meaningfully and it also yields good level of efficiency in terms of storage space utilization.

TABLE VI. FILE LEVEL VS CHUNK LEVEL DEDUPLICATION TECHNIQUES

Metrics	File level Deduplication	Chunk level Deduplication
Granularity of deduplication	File level	Chunk level
Resource utilization	Less	More
Execution time	Fast. Since comparison is done for entire file only once.	Slow. Since it has to check for all the chunks individually
Metadata overhead	Little metadata overhead	More metadata overhead
Computational complexity	Low	High
Level of efficiency in terms of storage space utilization	Less	High. Since the deduplication is chunk level. Avoid the duplicate copy of even part of the data.

TABLE VII. MEMORY REQUIREMENT IN STORAGE SPACE

Deduplication Technique	Memory required (in KB)
File-level	19.63
Chunk-level	9.85

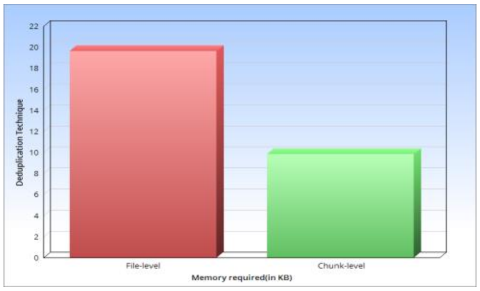

Fig. 14. Memory Requirement in Storage Space using File-level and Chunk-Level Deduplication Techniques.

TABLE VIII. FIXED-SIZE VS VARIABLE-SIZE CHUNK LEVEL DEDUPLICATION TECHNIQUES

Metrics	**Fixed-size Chunk level Deduplication**	**Variable-size Chunk level deduplication**
Granularity of deduplication	Chunk level	Chunk level
Size of the chunk	Fixed	Varibale
Execution time	Fast. Since the content is to be divided into fixed size without any constraints.	Slow. Since it has to divide the content into variable size in a content-based manner.
Boundary shift problem	It has boundary shift problem even if one byte is either added or deleted in their boundary.	It does not have any boundary shift problem.
Generating indexes	More	Few
Throughput	Low	High
Computational complexity	Low	High
Level of efficiency in terms of storage space utilization	Less	High.

The various deduplication techniques are analysed using various parameters and metrics. From the above results it is seen that applying in-line deduplication with chunk-level technique may results with good efficiency in the deduplication process.

VI. CONCLUSION

The cloud storage provides the storage space for the individual and the organization that are all in need of storage space. The cloud storage can be efficiently utilized with the help of deduplicaiton techniques. There are various types of deduplication techniques and every technique has its own advantages and disadvantages and that can be efficiently utilized for suitable applications. In this paper we analysed the benefits of applying deduplication technique for efficient utilization of cloud storage space and discussed about the different deduplication techniques in various dimenstions. Among various techniques, In-line with variable-sized chunk level deduplication will help us to achieve efficient utilization of cloud storage space since it does not require additional storage space as it is required in Post-process method, because it does not store any duplicate data and in case of variable – sized chunk level deduplication, it looks for even the part of the data is duplicate or not instead of checking the duplication for complete data. The major reason of this research work is to compare the various available deduplication technique. So that the various groups of cloud storage providers can be benefitted by applying the suitable deduplication techniques in their respective scenarios.

VII. FUTURE WORK

The limitation of our work is, we did not consider various types of digital data for comparing the deduplication techniques. In future research work, we are interested in analyzing the various techniques of deduplication with the respect to the different types of digital data.

REFERENCES

[1] KamalaKannan, T., Sharmila, K., Shanthi, M. C., & Devi, M. R.(2019) Study on Cloud Storage and its Issues in Cloud Computing, International Journal of Management, Technology And Engineering, Volume IX, Issue I, 2019, P(976-981).

[2] R. Arokia Paul Rajan , S. Shanmugapriyaa (2012) Evolution of cloud storage as cloud computing infrastructure service. IOSR Journal of Computer Engineering, Volume 1, Issue 1, PP 38-45.

[3] Tang, Y., Yin, J., & Wu, Z. (2016, June). Try Managing Your Deduplication Fine-Grained-ly: A Multi-tiered and Dynamic SLA-Driven Deduplication Framework for Primary Storage. In 2016 IEEE 9th International Conference on Cloud Computing (CLOUD) (pp. 859-862). IEEE.

[4] Paulo, J., & Pereira, J. (2014). A survey and classification of storage deduplication systems. ACM Computing Surveys (CSUR), 47(1), P(1-30).

[5] Nurshafiqah, M. Z., Yoshii, H., Enomoto, F., Koike, I., & Kinoshita, T. (2017, April). Data Deduplication for Audio Data Files. In Proceedings of 32th International Conference on Computers and Their Applications (CATA2017) (pp. 17-21).

[6] Sawant, A. A., & Game, P. S. (2018, August). Deduplication of Audio Files to Remove Redundancy in Cloud storage. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-4). IEEE.

[7] Manogar, E., &Abirami, S. (2014, December). A study on data deduplication techniques for optimized storage. In 2014 Sixth International Conference on Advanced Computing (ICoAC) (pp. 161-166). IEEE.

[8] Anand Bhalerao, Ambika Pawar(2017), A survey: On data deduplication for efficiently utilizing cloud storage for big data backups on International Conference on Trends in Electronics and Informatics ICEI 2017.

[9] D.Viji, Dr.S.Revathy(2019) , Various data deduplication techniques of primary storage on Proceedings of the Fourth International Conference o n Communication and Electronics Systems (ICCES 2019) IEEE Conference Record # 45898; IEEE Xplore ISBN: 978-1-7281-1261-9.

[10] Fu, Y., Jiang, H., Xiao, N., Tian, L., & Liu, F. (2011, September). AA-Dedupe: An application-aware source deduplication approach for cloud backup services in the personal computing environment. In 2011 IEEE International Conference on Cluster Computing (pp. 112-120). IEEE.

[11] Deepu, S. R. (2014). Performance Comparison of Deduplication techniques for storage in Cloud computing Environment. Asian Journal of Computer Science And Information Technology, 4(5) (pp. 42-46).

[12] Kim, Y., Kim, C., Lee, S., & Kim, Y. (2016). Design and Implementation of Inline Data Deduplication in Cluster File System. KIISE Transactions on Computing Practices, 22(8), 369-374.

[13] Wildani, A., Miller, E. L., &Rodeh, O. (2013, April). Hands: A heuristically arranged non-backup in-line deduplication system. In 2013 IEEE 29th International Conference on Data Engineering (ICDE) (pp. 446-457). IEEE.

[14] Srinivasan, K., Bisson, T., Goodson, G. R., &Voruganti, K. (2012, February). iDedup: latency-aware, inline data deduplication for primary storage. In Fast (Vol. 12, pp. 1-14).

[15] Kathpal, Atish, Matthew John, and Gaurav Makkar. "Distributed duplicate detection in post-process data de-duplication." HiPC. 2011.

[16] Pal, S., More, K., &Pise, P. (2018, February). Content-Based Deduplication of Data Using Erasure Technique for RTO Cloud. In 2018 International Conference On Advances in Communication and Computing Technology (ICACCT) (pp. 109-113). IEEE.

[17] https://pibytes.wordpress.com/2013/02/17/deduplication-internals-content-aware-deduplication-part-3/.

[18] Jyoti Malhotra, JagdishBakal "FiLeD: File Level Deduplication Approach". International Journal of Computer Trends and Technology (IJCTT) V44(2):74-79, February 2017. ISSN:2231-2803.

[19] Bhalerao, Anand, and AmbikaPawar. "Two-Threshold Chunking (TTC): Efficient Chunking Algorithm For Data Deduplication For Backup Storage.", International Journal of Scientific and Technology Research Volume 8, Issue 09, September 2019 ISSN 2277-8616.

[20] Venish, A., &Sankar, K. S. (2016). Study of chunking algorithm in data deduplication. In Proceedings of the International Conference on Soft Computing Systems (pp. 13-20). Springer, New Delhi.

[21] Sharma, N., AV, K. P., &Kakulapati, V. (2019). Data deduplication techniques for big data storage systems. Int. J. Innov. Technol. Explor. Eng., 8(10), 1145-1150.

[22] Yoon, M. (2019). A constant-time chunking algorithm for packet-level deduplication. ICT Express, 5(2), 131-135.