# Survey on Data Storage Management using Deduplication

*Pooja Pinjare* [1]
*Poonam Gupta* [2]
*Pundalik Gawandi*[3]

[1] *PG Student, Department Of Computer Engineering, G.H Raisoni College of Engineering and Management (Autonomous)Wagholi Pune, India affiliation by Savitribai Phule University,*
[2]*Head Of Department, Department Of Computer Engineering, G.H Raisoni College of Engineering and Management (Autonomous) Wagholi Pune, India affiliation by Savitribai Phule University,*
[3] *PG Student, Department Of Computer Engineering, G.H Raisoni College of Engineering and Management (Autonomous)Wagholi Pune, India affiliation by Savitribai Phule University.*

## *Abstract*

***Data Deduplication is one of the popularly used data compression techniques. It is used for finding and deleting repeated copies of data and has been mostly used now in a big storage system to reduce the storage space of the system. Day by day the storage system problem is increased tons to avoid the repeated storing of file Deduplication concept is used.in that we did the survey on deduplication concept where deduplication is The process of finding duplicate data from the storage systems to avoid the storage problem which happens due to the repeated storage of files is called Deduplication. This system is employed to enhance storage management. based on the survey we proposed the system which provides efficient storage management with finding the duplicates data which are having a similar hash code i.e. the data which are repeated within the personal computer system. To seek out duplicate data from the system MD5 hashing is used. MD5 is the hashing algorithm. The proposed system takes the file hash code to urge the precise duplicate file. Firstly the database creation is completed during which all the folders get scan and hash code of its store within the database. Once the database is prepared then the user can upload the new file within the system then the hash code of that file is calculated and generated hash code is matched with the database record if the hash code matches then the system will give the message file is duplicate with the file name and path of that file. If the file isn't duplicate then the path and hash code of that file is a store within the database and file store at a specified location.***

**Keywords**- Data Deduplication, Storage Management, MD5.

## 1. INTRODUCTION

As knowledge and network technology is rapidly developed and rapidly increases the size of the data center is the same as population growth has affected the world in an undesirable way. Growing population increases the problem of accommodation, pollution, etc. this analogy with our computer world; an increasing amount of data also increases the problem of Wastage of storage space, performance, cost, etc. In the computer world data comes from Different sources and in different forms. A data source can be a single computer, mobiles, and servers. Data as a different form of files structured/unstructured files, compressed files, images, audio, video, log files, exe, bin files, etc.

As the data is increase at faster and it is doubling in a less span of time, arranging data is a most difficult and huge task. Considering these problems experts are looking for different techniques to get the best technique to manage Duplicate data. Their different techniques one such technique is Data Deduplication. Deduplication is a Data compression technique that finds duplicate data and saves a large amount of space on the storage system. Figure 1. Shows the process of deduplication 1. Research shows that 75% of data of the digital world is copy data from different sources and 90% of data has come from backup data.

In the Data Deduplication process, we analyze data and compute a unique identifier of data. There are different ways of classifying the deduplication method inline and post-process. Post-process method checks the data after writing the data to the Storage disk and then identifying duplicates unlike inline deduplication, where data is deduplication done before the data being stored to the Storage disk.

Our proposed system manages the data by using the inline deduplication process where we check the data before storage in the storage system where we get the file from where we want the data and get the location where we want to store the data before that once we get the data we calculate the hash code of the file which we want to store in the system after that we check file hash code with the create a hash database of all files if the file is duplicate then we display the message that file is present and location of that file else the file hash and its location is store in the database.

- *General Deduplication process involves*

(1) Identifying file types
(2) Dividing file data into chunks
(3) Calculating fingerprints of chunks, and
(4) Identifying and storing non-identical data.

Files are segmented into pieces (chunks) using chunking methods such as file-level chunking and sub file chunking. File-level chunking considers the entire file as one chunk. In sub file chunking, the file is divided into fixed-size static chunks of variable size dynamic chunks [1]. With whole file level chunking, there is the lowest overhead of CPU, I/O, and indexing. In our Computer system, we have the number of files which are the same. The duplicate files are removed easily using file Deduplication.

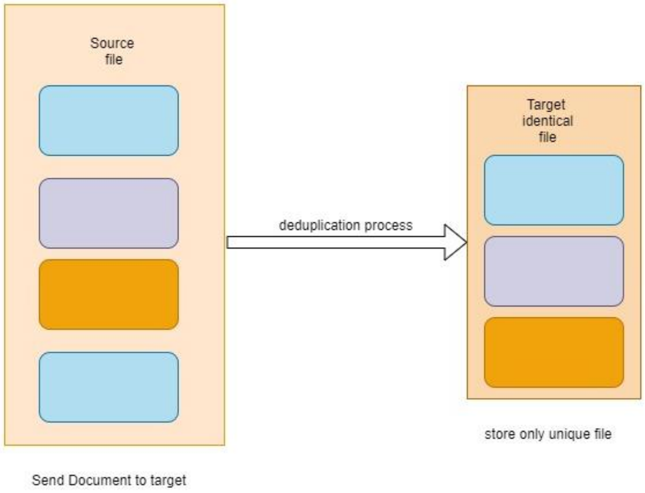

Fig.1. Schematic diagram of Data Deduplication

File-level chunking, there is the lowest overhead of CPU, I/O, and indexing. In file-level chunking deduplication, only one unique instance of the data is actually on the server and redundant data is replaced with a pointer to the unique data copy. To get the duplicate file from the Computer system we can use the different hashing algorithm for example RSA, SHA-1, SHA-2 etc. here we will use the MD5 hashing algorithm. The MD5 is an extended version of MD4.

- *Why is Data Deduplication useful*?

Data Deduplication helps storage administrators reduce costs that are related to duplicated data. Large datasets often have tons of duplication, which increases the prices of storing the info. For example: User file shares may have many copies of an equivalent or similar files. Virtualization guests might be almost identical from VM-to-VM .Backup snapshots may need minor differences from day to day.

The space savings that you simply can gain from Data Deduplication depend upon the dataset or workload on the quantity. Large amount of similar kind of dataset has the optimize up to 95%, or a 20x reduction in storage utilization.

The MD5 message-digest algorithm is mostly for hash function producing a 128-bit hash code. The generated MD5 hash code of the files is used to find the duplicate files on the system by comparing the generated hash code and stored hash code.

## 2. RELATED WORK

This paper [1] developed the Characteristics of backup workloads in production systems. The author presents a whole characterization of backup workloads by analyzing statistics and content information collected from an outsized set of EMC information Domain backup systems in production use. This analysis is complete (it covers the statistics of over ten systems) and exhaustive (it uses elaborated traces of the information of various production systems that store nearly 700TB of backup data). They tend to compare these systems with a close study of Microsoft's primary storage systems and incontestable that back-up storage differs considerably from the first storage employment in terms of knowledge quantities and capability necessities, similarly because of the quantity of knowledge storage capability redundancy Among the info. These properties provide distinctive challenges and opportunities once coming up with a disk-based filing system for backup workloads.

This paper [2] developed Primary information Deduplication large scale study and system storage the author presents a huge study about primary information Deduplication and uses the results to manage the design of a replacement primary information Deduplication system used within the Windows Server 2012 Operating System. The file information was analyzed by fifteen servers of worldwide distributed files that host information for over 2000 users in a very massive international company. The results area unit won't accomplish a fragmentation and compression approach that maximizes Deduplication savings by minimizing the information generated and manufacturing an even distribution of the portion size. Deduplication process resizing with information size is achieved by a stinting hash index of RAM and information partitioning, so that memory, central processing unit, and disk search resources stay accessible to satisfy the most employment of the IO service.

Redundancy [3] elimination among massive collections of files. Propose a replacement storage reduction theme that reduces information size with comparable potency to the foremost high-ticket techniques, however, at a value love the quickest however least effective. This process, referred to as REBL (Block Level Redundancy Elimination), provided benefits of compression, finding of duplicate blocks, and delta coding to delete the repeated data in huge amounts, eliminating a large spectrum of redundant information in a very ascendible and economical means. REBL typically encodes a lot more succinctly than compression (up to an element of 14) and a mixture of compression and suppression of duplicates (up to an element of half dozen.). REBL is additionally coded equally to a method supported delta coding that considerably reduces the area in a very case. Additionally, REBL uses super fingerprint, a method that reduces the info required to spot similar blocks by drastically reducing the process necessities of the matching blocks: it converts the comparisons of O (n2) into searches of hash tables. As a result, the employment of super fingerprints to avoid enumerating the corresponding information objects decreases the calculation within the REBL alikeness section of a few orders of magnitude.

This paper, the author proposed the [4] Encrypted information Storage with Deduplication process which applied to the Twin Cloud. The info and also the personal cloud wherever the token generation is going to be generated for each of one file. Before uploading the file to the normal public cloud, the sender can send the file to the personal cloud for the token value that is unique and value generated automatically. The generation that is unique to every file. Hash and token values are created by the clouds at the personal level and send the value which is created by the cloud to the receiver. The value which is created by the personal cloud is a unit unbroken within the personal cloud itself so that whenever a successive token generation file arrives, the personal clone will discuss with a similar token. When the receiver gets the token value for a given file, the public cloud has the token similar or not. If the value of the token is available on the public cloud, then it will point to a pointer to the available file, else it will give the message to load a file on the cloud. A system provides confidentiality and does the block-level Deduplication method to get the file for finding at a similar time. Block-level deduplication is the describe In this paper where the data that can be in the file format is divided into the number of the parts that parts having the same static or dynamic way of the partitioning the data of the file the static method of the block level Deduplication where the part is already specified in the coding that can be 4*4, 2*2, etc. and in the dynamic method the based on the data changing partitioning is done.

[5] In that paper, the author tends to have gotten information Deduplication by providing information proof from the info owner. This takes a look at and is employed once the file is uploaded. Every file stored on the cloud for that there is restricted by a group of permission to know the kind of users who will perform duplicate verification for finding the similarity and user of that files who can access that files. New duplication constructs compatible with approved duplicate verification within the cloud hybrid design wherever the personal cloud server generates duplicate file verification keys. The projected system includes an information owner take a look at, therefore it'll facilitate implementing higher security problems in cloud computing.

6] In this paper the author approaches the Block locality Cache (BLC), which captures the previous backup execution significantly above existing approaches and forever uses up-to-date data regarding things and is so less susceptible to aging. Tend to evaluate the approach using a simulation supporting the detection of multiple sets of real backup info. The simulation compares the Block locality Cache with the approach of Zhu et al. And provides an in-depth analysis of the behavior and conjointly the IO pattern. To boot, a paradigm implementation is used to validate the simulation.

7] During this paper authors discovered the optimized WAN replication of backup knowledge sets exploitation delta compression rumored by the stream off-site knowledge replication is very important for big recovery, but this tape transfer method is cumbersome and error-free. Replication in AN extremely wide space network (WAN) can be promising numerous, but fast network connections are expensive or impractical in many remote locations, so higher compression is needed to create WAN replication wise. They tend to give a replacement technique for replicating backup knowledge sets through a WAN that not only removes duplicate file regions (Deduplication) but together compresses similar file regions with delta compression that's on the market as a feature of EMC knowledge Domain systems.”

In this paper the author collects data [8] from the file system content of 857 desktop computers in Microsoft for an amount of four weeks. They tend to analyze the info to work out the relative efficiency of information deduplication, particularly considering the elimination of complete file redundancy against blocks. They found that full file Deduplication reaches approximately 3 quarters of the area savings of a lot of aggressive block Deduplication for live file system storage and eighty-seven of backup image savings. They tend to conjointly investigate file fragmentation and locate that it doesn't prevail, and that they have been updated on previous studies on file system data, and that they have found that file size distribution continues to affect massive unstructured files.

The author [9] has developed a generic model of file system changes supported properties measured in terabytes of real and completely different storage systems. Their model connects to a generic framework to emulate changes within the file system. Supported observations from specific environments, the model will generate associate initial file systems followed by continuous changes that emulate the distribution of duplicates and file sizes, realistic changes to existing files, and file system growth.

In this paper author [10] builds the system that achieves confidentiality and permits block-level deduplication at an identical time. This method is built on prime of oblique encoding with LFSR (Linear Feedback Shift Register) encoding technique. A shift register's distinctive performance is shifting its contents into adjacent positions among the register or, within the case of the position on the highest, out of the register. The position on the other finish is left empty unless some new content is shifted into the register. The contents of a register are sometimes thought of as being binary that is, ones And zeros. They did the block-level deduplication rather than file-level deduplication since the gains in terms of space for storing. Block-level deduplication is that the describe during this paper wherever the information which will be within the file format is split into the amount of the components that components having an equivalent static or dynamic means of the half partitioning the information of the file the static methodology of the block level Deduplication wherever the part is already per the cryptography which will be 4*4, 2*2, etc. and within the dynamic methodology, the supported information ever-changing partitioning is finished. The token is generated and also assigned the non-public key for that information.

[11]In this paper author analyzes the ghost cache approach that was originally designed for addressing weak locality incurred high memory overhead in fingerprint caching. to handle this, they projected TLE-LRU, a cache replacement rule that may estimate the temporal neighborhood of the info streams and range the cache allocation consistent with the estimation. They designed a hybrid Deduplication system named SLADE with TLE-LRU. SLADE achieved high inline cache potency and reduced the Deduplication employment within the post-processing Deduplication part. SLADE improved the inline Deduplication quantitative relation by up to thirty-nine.70with idedup in their Experiments.

[12] In this paper author is making an associate application that will perform fixed-size block-level Deduplication checks on the fragmented blocks so encrypting the blocks before uploading it to the cloud. Also, it performs policy-based mostly file assured deletion to firmly delete young lady from the cloud. As a result, it tries to realize accuracy, security, and alternative style goals for the system.

## 3. EXISTING SYSTEM:-

System store a lot of files in the storage system that is already out there therefore this place is wastage. Therefore the projected system takes away the duplicate file using the Deduplication construct. Deduplication takes advantage of information similarity to realize storage reduction. The advantage of examination blocks or files bit to bit is that it's correct, however, it's also time overwhelming. The advantage of comparing files by hash value is that it's very fast. The previous system uses the block level deduplication which is also time-consuming and CPU overhead is more so we use the file level chunking for finding the duplicate file easily. Before every user uploads associate encrypted file, he calculates its hash value and compares that hash value with the hold on hash values is there's no same hash value then the user can get ` no duplicates', the user stores his file to the system.

## 4. PROPOSED SYSTEM

In this paper, we propose a confidence scheme during which the system finds all duplicate files.
Our proposed deduplication scheme involves the following steps

(1) Identifying file types
(2) Calculating fingerprints of the file, and
(3) Identifying and storing the non-identical file.

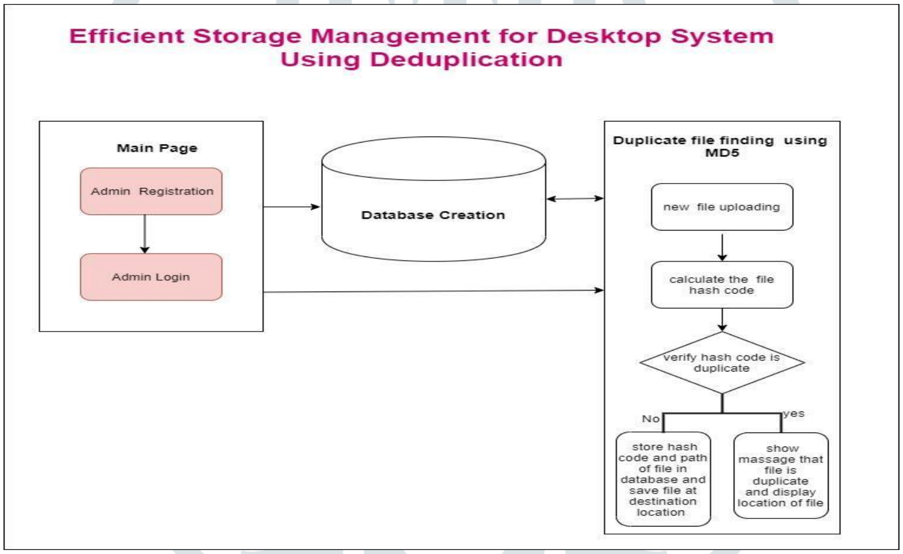
Efficient Storage Management for Desktop System Using Deduplication
Main Page
Admin Registration
Admin Login
Database Creation
Duplicate file finding using MD5
new file uploading
calculate the file hash code
verify hash code is duplicate
No
yes
store hash code and path of file in database and save file at destination location
show massage that file is duplicate and display location of file

Fig.1. System architecture.

Our proposed system manages the data by using the inline deduplication process where we check the data before store within the storage system where we get the file from where we would like the information and find the situation where we would like to store the data before that when we get the data we calculate the hash code of the file which we would like to store within the system subsequently we check file hash code with the create a hash database of the all files if the file is duplicate then we display the message that file is present and site of that file else the file hash and its location is a store within the database. When users want to store the file on the system then the file may be chosen using the 'Browse' button. It uses a straightforward hash code for locating the duplicates. Which can be unique for unique files and can be the same for duplicate ones. Although it doesn't account for files which can have the identical name and therefore the same file size but have different content. Here we use the MD5 to find the duplicate files. Else hash code and path is stored in the database and file is saved at specified location.

## 5. CONCLUSION

In this paper we did the survey on deduplication concept and understand how data deduplication works, what are ways of the data deduplication and algorithm used in data deduplication. Data Deduplication is very important and significant within the follow of data storage, especially for the management of massive information. In this paper we proposed the inline-level deduplication method that offers versatile data Deduplication within the desktop storage system. Schema are often custom-made to completely different scenarios and application requests and offer cost-efficient management of data storage for a desktop system.

## 6. REFERENCES

[1] D. Meister, J. Kaiser, and A. Brinkmann, “Block locality caching for data deduplication,” in Proc. 6th Int. Syst. Storage Conf., 2013, pp. 1–12.
[2] M. Lillibridge, K. Eshghi, and D. Bhagwat, “Improving restore speed for backup systems that use inline chunk-based deduplication,” in Proc. 11th USENIX Conf. File Storage Technol, Feb. 2013, pp. 183–197.
[3] V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuenning, and E. Zadok, “Generating realistic datasets for Deduplication analysis,” in Proc. USENIX Conf. Annu. Tech. Conf., Jun. 2012, pp. 261–272.
[4] D. T. Meyer and W. J. Bolosky, “A study of practical deduplication,” ACM Trans. Storage, vol. 7, no. 4, p. 14, 2012.
[5] G. Wallace, F. Douglis, H. Qian, P. Shilane, S. Smaldone, M. Chamness, and W. Hsu, “Characteristics of backup workloads in production systems,” in Proc. 10th USENIX Conf. File Storage Technol., Feb.2012,pp.33–48.
[6] El-Shimi, R. Kalach, A. Kumar, A. Ottean, J. Li, and S. Sengupta, “Primary data deduplication-large scale study and system design,” in Proc. Conf. USENIX Annu. Tech. Conf., Jun. 2012, pp.285–296.
[7] P. Shilane, M. Huang, G. Wallace, and W. Hsu, “WAN optimized replication of backup datasets using stream-informed delta compression,” in Proc. 10th USENIX Conf. File Storage Technol.,Feb.2012,pp.49–64.
[8] P. Kulkarni, F. Douglis, J. D. Lavoie, and J. M. Tracey, “Redundancy elimination within large collections of files,” in Proc.usenixannu.Tech.Conf. Jun.2012, pp.59–72.
[9] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou “A Hybrid Cloud Approach for Secure Authorized Deduplication” IEEE Transactions on Parallel and Distributed Systems: PP Year 2014.
[10] Shweta D. Pochhi, Prof. Pradnya V. Kasture “Encrypted Data Storage with Deduplication Approach on Twin Cloud “International Journal of Innovative Research in Computer and Communication Engineering.
[11] Huijun Wu, Chen Wang, Yinjin Fu, Member, IEEE, Sherif Sakr, Kai Lu,Liming Zhu ,
Differentiated Caching Mechanism to Enable Primary Storage Deduplication in Clouds " in IEEE Trans,2018.
[12] Sneha C. Sathe,Nilima M.Dongre , Block Level based Data Deduplication And Assured Deletion in Cloud", in ICSSIT,2018.