# Data deduplication technique for optimized storage

G. Sai Divya K.Sai Sri Thanya E.Sumasree K.Keerthana

[*] Indian Institute of Information Technology, Sri City, Chittoor[*]

***Abstract*- These days a lot of problems in storage and performance are appearing in digital data. There are many techniques used for eliminating the redundant data in the storage. Data deduplication is one of the technologies for detecting data redundancy and it reduces the storage space and network bandwidth. In this paper, we summarized various chunking algorithms of data deduplication. Finally we proposed a method based on checking the data, whether it is in the database before storing it .**

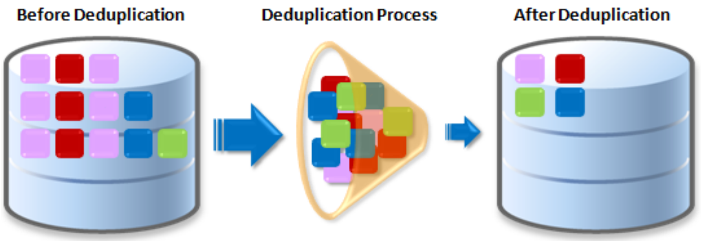
***Keywords*-Data deduplication,storage, cloud computing, Hashing , MD5**

## I. INTRODUCTION

The growth in technology is increasing the amount of storage.There are many data storage devices like computers, mobile,tablets etc. The old technology was to connect the devices and physically transfer data between the devices. There is a need for new technology that can be faster and easier and the solution for it is cloud storage. After the cloud became popular and a huge number of people started using it many problems are appearing with the cloud.According to statistics, 60% of the data is reductant.The duplicated data effects on the storage and performance of cloud. To address this, many techniques like data compression, data deduplication can be used to improve the storage capacity thereby reducing the replications of data.Data deduplication has been proposed as the best technique to solve this problem. After uploading the data ,apply an algorithm to analyze and check if any data are matched, keep one and delete the others or we can analyze the data at the time of uploading and see whether it matches anything that is stored in storage , if it matches just ignore it.In this paper we will focus on the second technique of checking the data at the time of uploading.Hash algorithm is used to check the data if it is matched, ignore the uploading else count it as a unique and continue uploading to store it .This paper consists of various techniques for data deduplication.

## II. DATA DEDUPLICATION

The data deduplication technique works by tracking each data file and eliminating each file that it found more than one copy of it in the storage. It is one of the most popular techniques in saving storage. The data deduplication is used by many vendors. The deduplication is very important for the shared storage The deduplication is also a data reducing technique. Unlike the compression which is compressed and kept all data. There is more than one way to deduplicate the data. There is more than one way to detect the duplicated data and eliminate it. By the end of the day all leads to the same point: reduce the size to save storage. Figure 1. shows the strategies that are used for data deduplication.



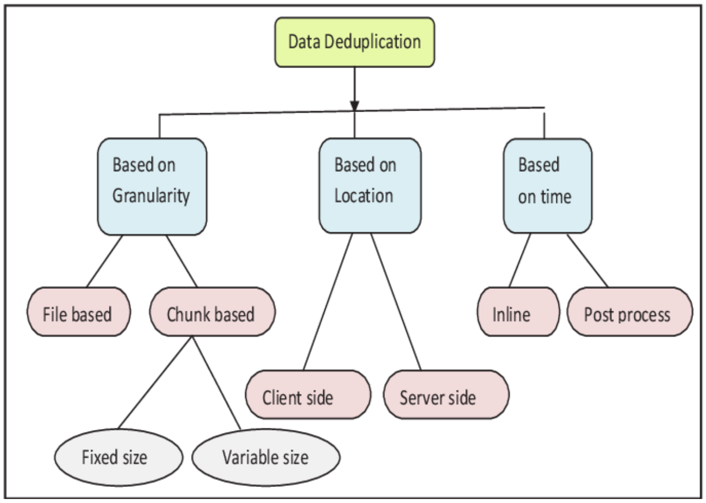


Figure 1.Strategies of data deduplication

## Data chunking

## 1.File level deduplication:

The single instance storage or whole file chunking does not break files into a smaller chunk, rather than it treats the whole file as a chunk. It finds the hash value for the entire chunk which is the file index. If a new incoming file matches with the file index, then it is considered as duplicate and it points to the existing file index. We can use MD5 or SHA-1 as hashing algorithms. Whole file chunking is simple and fast, but it can only detect exact file duplication.

## 2. Chunk level deduplication:

- **Fixed- size chunking:**

In this data deduplication algorithm, it breaks the files into equals sized chunks in which the chunk boundaries are fixed such as 4KB, SKB, etc. for example if a chunk size is defined as 4KB, a file is chunked at SKB,12KB,16KB,20KB continuously.The checksum technique is used to check if there is any duplication. Only the unique checksum is stored in the storage.

- **Variable size chunking:**

Variable block chunking is different from fixed block chunking. Here, chunking boundaries are determined based on the contents of the file, so it is more resistant to the insertion and deletion.

This has three important steps to follow which are as follows:

1) Dividing the file into variable blocks based on the chunk boundaries.

2) Generating the hash values for each block.

3) Identify the redundant data from the hash values.

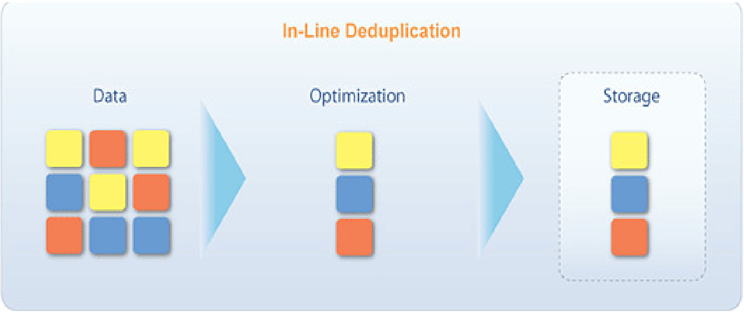
## Location based deduplication

1.**Source based deduplication** process happens at the client side. This process is done on this side by applying a special program to detect the duplication on the database of the client himself. The advantage of doing the deduplication at the client side is saving bandwidth, because only the unique data will be stored in the cloud.

2.**Target based deduplication**: In this deduplication process is done on the target server. The procedure for this type is by storing all the data into the cloud or backup, then the server will handle and sort the data. Then find the duplications and eliminate them.

## III. PROPOSED TECHNIQUE

Our proposed technique depends on reducing the data before it is stored in the storage. This type of process is done on the client side.The system analyzes the data and divides into chunks, then hash codes are given for the chunks by using a hash function. Whenever a new data or file comes, it is divided into chunks and the hash codes are checked, if the system found the same it means that data is stored already, and it ignores the data block.



Here hashing is done by using MD5(Message digest) algorithm. MD5 has less processing time so this algorithm is chosen. There are many advantages by using this technique. No need for extra storage space. The data domain is less and consumes less bandwidth.

The other idea of improving this technique is to use trie for storing the hash code of chunks.We represent each hash code as a individual node and .If the input hash code is new or an extension of the existing one, we need to construct non-existing nodes of the previous one, and mark end of the word for the last node.With the help of trie we need not store all the hash codes of a file , if some part of it is present we can add new nodes for other hashes. This implementation reduces space complexity and time complexity also as searching in trie takes less time.

## IV. CONCLUSION

In this study, the way of optimizing cloud storage is discussed. Deduplication is one of the various techniques that is used for optimization. Deduplication has many strategies depending on data size, the location of the data processing and the time of data processing. In this a new method of implementing fixed-size chunking based on trie is proposed.This technique reduces the space and time complexity. In future, more research works will be done and implementing the proposed method.

## V. REFERENCES

1. Zuhair S. Al-sagar, Mohammad S. Saleh, Aws Zuhair Sameen "Optimizing the Cloud Storage by Data Deduplication: A Study" IRJET Volume: 02 Issue: 09 | Dec-2015

2. Cai Bo,Zhang Feng Li,Wang Can "Research on chunking algorithms of data deduplication"

3. E. Manogar, S. Abirami"A study on data deduplication techniques for optimized storage" Conference Paper · December 2014

4. B.Tirupati Reddy, U.Ramya, Dr.M.V.P Chandra Sekhar,”A comparative study on data deduplication techniques in cloud storage”