# Optimizing the Cloud Storage by Data Deduplication: A Study

**Zuhair S. Al-sagar[1], Mohammad S. Saleh[2], Aws Zuhair Sameen[3]**

*[1] Department of Electrical, Baqubah Technical Institute, Middle Technical University, Baqubah, Iraq*
*[2] Department of Electronic, collage of Engineering, Diyala University, Baqubah, Iraq*
*[3] Department of Electrical, Electronic and Systems, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), Selangor, Malaysia*

---***---

**Abstract -** ***In the last few years, the digital data, such images, audio, video and files are exploding. A lot of problems in storage and performance are appearing. Lots of money is spent on these problems. Until the scientists focus on the problems and the needs to solve these problems. Now days there are many techniques used for eliminating the redundant data in the storage. One of the best technique is the deduplication data. In this paper a study on previous researches will do. Focus on the gaps and the problems that they could not solve it or introduce them as a future work. Finally, we propose a new method depending on the literature to fill the gaps. This method based on the checking the data, whether it is in the cloud storage before storing it.***

***Key Words:*** *Data Deduplication, Storage, Cloud Computing, Hashing, MD5.*

## 1. INTRODUCTION

Now days with the huge increasing of population and the using of technology, it leads to many problems [1]. The growth in technology is increasing the amount of storage or communication and technique devices. There are many data sources like cameras, mobiles, tablets and computers etc. Sometimes one person has more than one device and storing data from all of them or he/she wants to use his/her data on any device that have at any time. The old technology was to connect the devices to each other physically and start transferring data between them. The need for new technology that should be easier and faster appears. The scientist and researchers start thinking for solution until they found the cloud storage [7]. Many clouds are offered from different brands. After the cloud become popular and used from a huge number of users many problems are appearing again with cloud. Some of these problems are external and other internal. Security was one of the big challenges of the cloud service providers. The implement and applied algorithms to keep the customer's information secure [4]. The data duplication is one of the big challenges in cloud computing [6]. The duplicated data effects on the storage and the performance of the cloud. Previous researches show that about 90% of the data that are stored in cloud backup are duplicated [1]. The researchers start studying for deduplication techniques to optimize the storage [5]. There is more than one way to deduplicate the data. Either by analyzing the data within the uploading and see whether it matches anything that is stored in storage, if it matches just ignore it or by uploading the data, then apply an algorithm to analyze and check if any data are matched keep one and delete the others [1]. This study focus on the first technique which ignore uploading the duplicated data. 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 [4].

The rest of this paper is organized as follows: section two literatures on various techniques for storage optimization. Section three explains the deduplication and the proposed algorithm will be explained in section four. Section five will be the conclusion of this paper.

## 2. STORAGE OPTIMIZATION

Many techniques are used to optimize the storage. Some of them are Compression, Snapshots and Deduplication. The Deduplication is one of the most popular technique that is used in this field [5, 6]. In this section the three types of optimization that are introduced will explain in details.

### 2.1 Compression

The data compression mechanism work by reducing the size of the file to save the storage. The word reducing means removing some binary digits from the file. The compression technique focusing only on the important information in the data. The compression technique compresses all files even if it is duplicated. Because of the data size are reduced so the processing speed decrease, that means the overall speed will increase and the time to load or store data are decreasing. The compression technique does not work only on saving more storage, it can be used for security [3]. To control the security term in these acts hash algorithm like MD5 used to authority verifier. The hash algorithm is used to test the integrity of the files.

## 2.2 Snapshot

The snapshot technology applied only on the data that are reached multiple times. It is almost used in the operating system to enable multiple access to that system. The snapshot is implemented by many vendors for read only while the other used it for writing also [5].

## 2.3 Deduplication

The data deduplication technique works by tracking each data file and eliminate each file that it found more than one copy of it in the storage. It is one of the most popular technique in saving the storage. The data deduplication used from many vendors. The deduplication very important for the shared storage [6]. 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. This point will explain in details in the next section.

## 3. DATA DEDUPLICATION

There is more than one way to detect the duplicated data and eliminate it. By the end of the day all leads to same point reduce the size to save storage. Figure 1. Shows the strategies that are used for data deduplication.

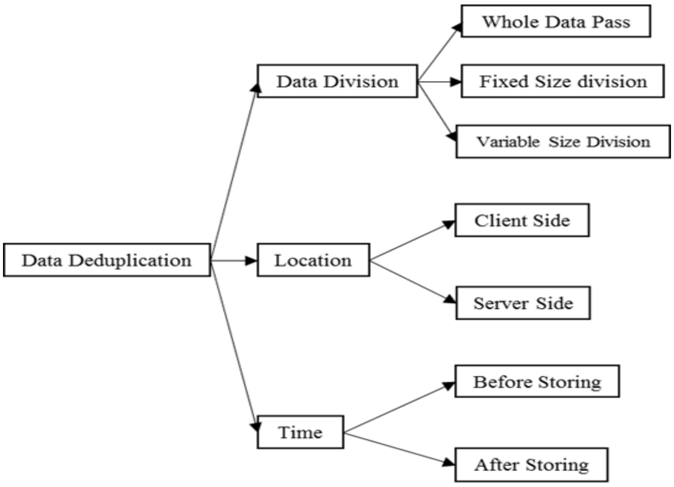

**Fig -1:** Strategies of Data deduplication

## 3.1 Data Division

This method is done by dividing the data into a sequence of bytes, then the divided blocks are used to test the redundancy. The deduplication done by store only the unique block. There are different types of data division strategy to deduplicate the data. These strategies are:

1- Whole file pass. This procedure is done by pass whole data without dividing it into smaller blocks. The compression is done with the hash indexed in the file if it matches, it counts it as a duplication.
2- Fixed size dividing. The procedure for this algorithm is done by dividing the data into equal block sizes, which means that the boundaries of the blocks are fixed for example 4Kbytes, 8Kbytes, etc. The checksum technique is used to check if there is any duplication. Only the unique checksum is stored in the storage. The weakness of this method is if large data are stored. It will divide it to a big number of segments or blocks, the ability of errors will be bigger.
3- Variable size dividing. The difference between this method and the previous fixed size is in this method the boundaries are not fixed. It is determined according to the data size. This method is more efficient compared to the previous two. This algorithm is the best for backup [5].

## 3.2 Location

In cloud which is type of networks. The data could be stored in two locations the first one in at the client side and the second one is on the server side. Depending on the location the deduplication process is done.

1- The client side which also called the source. The deduplication process done in 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 stored in the cloud.
2- The second location is the server side. The deduplication process happens to on the cloud servers. 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. The advantage of this process is reducing the number of overheads from clients.

### 3.3 Time

Time is one of the most important criteria in the field of processing and computing. Eliminate the duplicated files, make the speed is higher, that means less time processing. There are two types of deduplication techniques depending on the time. The first one is before storing the data to storage and the second one is after storing the data.

1- Before storing the data which is known as an inline process. This type of process done on the client side. Figure 2 shows the mechanism of deduplication data before storing them to the cloud.

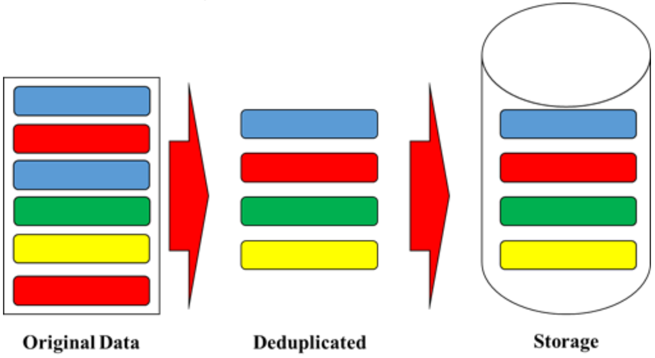

**Fig -2:** Before storing process

Basically the procedure is the system analyze the data before storing it by one of mechanism for checking like hash and so on. If the system found the same data is stored already, ignore the data block or packet. Else store the data and save its analysis for future processing. There are many advantages by using this technique. No need for extra storage space. The data domain is less. Less bandwidth. This process is new and a few cloud service providers offering this type of algorithms for deduplications.

2- After storing or called a post process. This process is done on the server side. Figure 3. Shows the mechanism of this process.

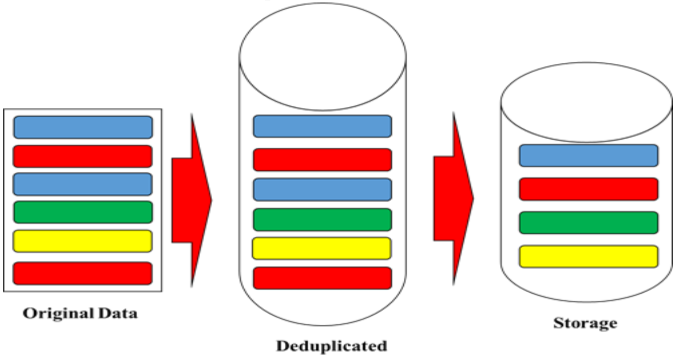

**Fig -3:** After storing process

All the data are stored in the storage then the deduplication algorithm will do its job by eliminating the duplicate data. The advantage of this method compares to previous ones is this method has higher performance. The other point is this method has the ability to share the index. The disadvantages are needing more storage space that has ability to hold all the data. Faster processing that means higher cache. This method is more expensive, but for huge data it is more efficient.

## 4. PROPOSED TECHNIQUE

From previous studies of the way to optimizing the storage. We found that one of the best solutions is by deduplication or in another word eliminate the duplicated files, keep only unique data in storage or backup. There are many techniques for deduplication. This paper focuses on three types of it as explained in section three. From this method we see that the time is an important factor that should be in focus. Our proposed algorithm depends on reducing the data before it's stored in the storage or backup. Figure 4. Shows the flowchart of the proposed algorithm.

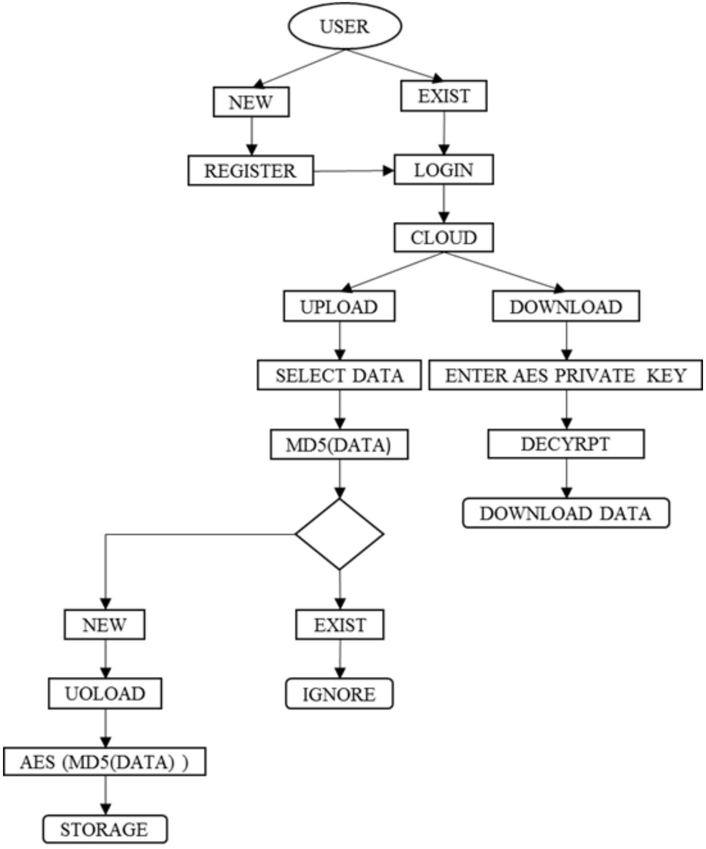

**Fig -4:** Flowchart of the proposed algorithm

The idea is using a hash algorithm attached to each data block. A SHA algorithm is used for this purpose. In this paper Message digest (MD5) is chosen. The reason of choosing MD5 because it 128bit length. Sha is more secure than MD5 because it is 160bit length, but the longer length means slower processing and more storage space is needed. For the first time when someone looks into the difference between the SHA and MD5 which is only 32bits in length he will think it is not a big difference, why we did not use the more secure algorithm. Actually for a small number of data it is true, but in reality because of this is a public service the number of data transfer will be huge. The small difference between the two hashes will be effected on the storage capacity. The second reason is the processing time. Because the MD5 is used for checking the data, is it unique or not, so the term of security is not the main of objectives of choosing this hash algorithm. For security Advanced Encryption Algorithm (AES) will use. The AES will use to encrypt the data. For downloading the data from the cloud the user should key in his/ her private key to have an access to the original data and download it.

## 5. CONCLUSION

In this study, the way of optimizing the cloud storage are discussed. Deduplication is one of the various techniques that is used for optimization. This survey focuses on the deduplication methods among other types of techniques. Deduplication has many strategies depending on data size, the location of the data processing and the time of data processing. In this study a new method is proposed depending on the time of data arrival to the cloud. Hence, this technique improves storage capacity and improve the performance by comparing the data before storing it using MD5 hash algorithm and store only the unique data file. In future, more research works will done and implementing the proposed method. And also to develop an efficient method to enhance the way of classifying the stored data.

## REFERENCES

[1] Jyoti Malhotra and Jagdish Bakal, "A Survey and Comparative Study of Data Deduplication Techniques", *IEEE International Conference on Pervasive Computing (ICPC),* 2015.

[2] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", *IEEE Transactions on Parallel and Distributed Systems*, 2014.

[3] R. Saranya, G. Indra and Dr. N. Sankar Ram, "Data Compression Technique to Eliminate Duplicates in Cloud Computing", *International Journal for Scientific Research & Development (IJSRD), Vol. 3, Issue 03*, 2015.

[4] Mathias Grawinkel, Michael Mardaus, Tim Süß and André Brinkmann, "Evaluation of Hash-Compress-Encrypt Pipeline for Storage System Applications", *IEEE,* 2015.

[5] E. Manogar and S. Abirami, "A study on Data Deduplication Techniques for Optimized Storage", *IEEE Sixth International Conference on Advanced Computing (ICoAC),* 2014.

[6] Qinlu He, Zhanhuai Li and Xiao Zhang, "Data Deduplication Techniques", IEEE International Conference on Future Information Technology and Management Engineering, 2010.

[7] The future of Cloud Computing, *European Commission International Society and Media*, 2010.

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.