Enhanced De-Duplication Methods to Improve Performance in Cloud Storage Environment

Manreet Kaur[1], Jaspreet Singh[2]

[1]*M.tech Scholar, Department of Information Technology, CEC Landran, Mohali, Punjab, India*

[2]*Assistant Professor, Department of Information Technology, CEC Landran, Mohali, Punjab, India*

Abstract - Cloud computing plays a main role in the business domain current as computing processes are delivered as utility on demand to client over the internet. Cloud storage is one of the services offered in cloud computing which has been increasing in famous. The major benefits of using cloud storage their time in buying and maintaining storage environment while only giving for amount of storage requested, which can be scaled-up and done upon demand. With the increasing data size of cloud computing, a reduction in data density could help providers decreasing the costs of running large storage system and convertible energy consumption. De duplication Data is the method which compresses the data by deleting the numerous copies of inner data and it is widely used in cloud storage to save bandwidth and minimum the storage space. To secure the private of sensitive data during De duplication, the hashing method is used to generate the hash for save data before outsourcing.

Keywords - *Cloud Computing, De-duplication, Private, Hashing Techniques and Secure.*

I. INTRODUCTION

Cloud Computing is the novel emerging trends in the novel generation technology. Each client has big amount of data to share to store in a quickly available protected place. The concept of De duplication is reached here to efficiently utilize the bandwidth and circle disc usage on cloud computing. To escape the De duplication copies of the similar data on cloud may cause loss of time, bandwidth consumption and space [1]. Cloud computing is the internet based, a network of remote servers associated over the internet to store, share, change, add and resourcing of data. The benefits of cloud computing: No longer have to pay for someone (or a team of someone's) to do things such as install and update software, install and manage email servers and/or fine servers, run backups – the loveliness of cloud computing is that all of the business of maintaining the service or application is the accountability of the cloud vendor, not yours .No longer have to buy software. Besides the convenience of not consuming to buy software programs and install them on your own servers/computers, using cloud applications instead can be cheaper. One may be able to consolidate your separate application needs into one multi-application cloud computing service. For illustration, Google Apps for Business includes email, a calendar scheduling application, Google Docs for generating documents, presentations and forms and using online file storage and Google Sites for creating websites, all for only $5/month for each person on your account. Now think about the price of, let's say, Microsoft Office including Microsoft Outlook for email Able to cut back on system hardware. File storage, data backup and software packages all take up a lot of space on servers/computers. With cloud computing, you use someone else's servers to store all this data instead, liberation up your in-house [2] computer equipment for other purposes or even letting you get rid of some of it. A cloud computing application may make integration easier. Because many cloud computing applications contain an Application Programming Interface (API) you may be able to find "compatible" applications rather than having to pay to have the submissions you want to be integrated customized for you. Cloud computing applications are habitually updated, so you don't have to spend time and money doing it – and giving you the advantage of always having access to an application's latest features and purposes. Cloud computing allows you and your employee's easy access to applications and data from different computers and devices. "As more consumers and businesses adopt tools such as smart phones and tablets, the ability to cloud data in the cloud and access it from just about anywhere on the planet is quickly becoming vital. Cloud computing lets you start up or grow your small business quickly. It's a lot easier and faster to sign up for a [3] cloud computing application than to buy a server, get it up and running and install software on it. And since you don't need to buy hardware and software, your start up or expansion is cheaper, too.

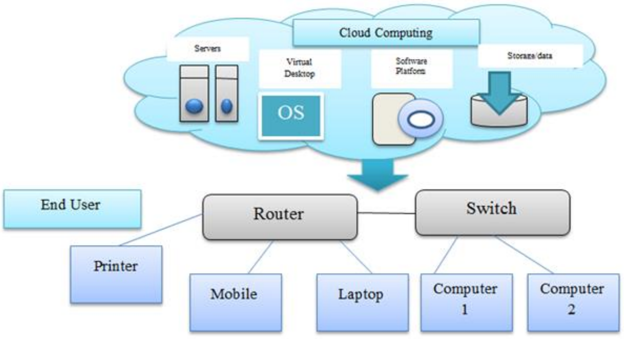

Fig. 1 Cloud Computing Architecture

II. DE-DUPLICATION

De-duplication is aim for highly redundant operations like backup, which needs again and again copying and storing the same data set multiple times for recovery purposes over 30-90 day periods. Consequences [4], enterprises of all sizes rely on backup and retrieval with Deduplication for fast, reliable and cost-effective backup and recovery. Deduplication segments an incoming data stream, uniquely verifies data segments, and then compares the segments to past stored data. If the segment is unique, it's stored on circle disc. However, if an incoming data segment is a copy of what has already been stored, created to it and the segment is not stored again.

Advantages of Data De-duplication are eliminating redundant data can important shrink storage requirements and improve bandwidth efficiency. Since primary storage has gotten inexpensive over time, enterprises typical store many versions of the same sequence so that new workers can reuse past done work. Some processes like backup store extremely redundant information.

Deduplication less stored over and over again, consuming redundant storage space on disk or tape, electricity to power and cool the disk or tape drivers, and bandwidth for duplication. This creates a chain of cost and resource inefficiencies within the arrangement [5].

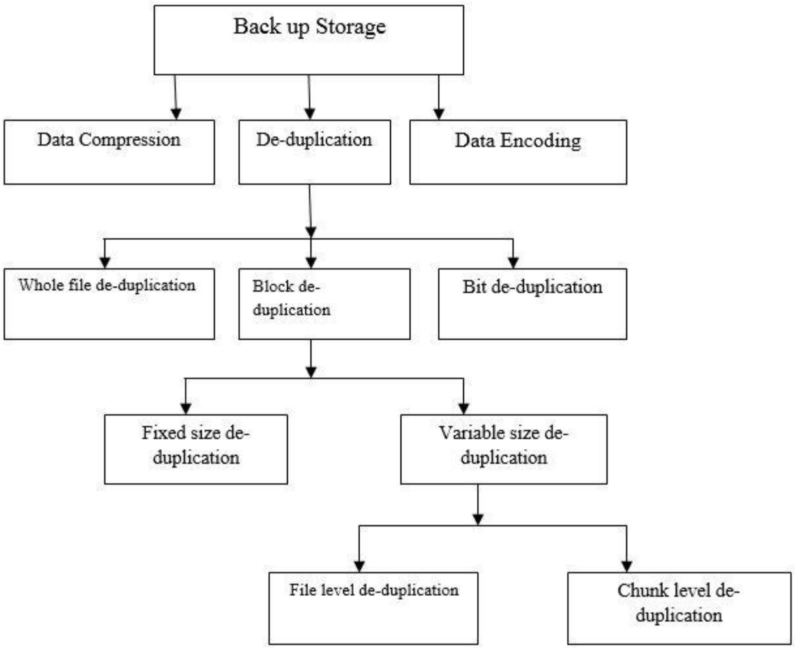
Back up Storage
Data Compression
De-duplication
Data Encoding
Whole file de-duplication
Block de-duplication
Bit de-duplication
Fixed size de-duplication
Variable size de-duplication
File level de-duplication
Chunk level de-duplication

Fig 2. De- Duplication Techniques

III. RELATED WORK

Vasilios et.al.2013 [6]presents a migration support network, in which fundamental elements are cost effective system. They proposed a three level framework that satisfies al the necessity in view of cost assumption. They utilized the windows azure policy as a part of creating prototyping model. Besides, the ability to consolidate necessities for numerous administration sorts, e.g., information stockpiling and systems administration, is imagined to be given, encouraging the choice making in relocation sorts past the off-stacking of the application stack on a VM.**Haitao et.al. 2011 [7]**proposed relocation methods taking into account (dynamic, receptive and shrewd procedures), albeit basically in light of the present data, can make the mixture cloud-helped VoD organization set aside to 30% transmission capacity cost contrasted and the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost and server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size and cloud substance upgrade system assume the key parts in the client experience change **.C. Ward et.al. 2010[8]**Acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers workload movement for this situation and talk about the effect that computerized relocation has on the expense and dangers ordinarily connected with relocation to cloud.**Kang et.al.2013[9]** Proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capacity. Then, if the migration constraint is gratified, we transfer another VM from this PM to oblige the new VM. In addition, we study a hybrid scheme where a batch is working to accept upcoming VMs for the on-line development. Evaluation results prove the high efficiency of our algorithms.**Xian Xinet.al.2013[10]**proposed a dynamic prototype system termed Cyber Live App to support application sharing and migration on demand among various users. Cyber Live App gives two key administrations: a safe multi-client sharing administration for the virtual desktop of a VM and multi-VM application sharing and movement.

IV. PURPOSE ALGORITHMS (SHA (SECURE HASH ALGORITHM)

Various types of SHA:-

● SHA 0
● SHA 1
● SHA 256
● SHA 512

The Secure Hash Algorithm is one of a number of cryptographic hash functions. There are currently three generations of Secure Hash Algorithm:
● SHA-1 is the original 160-bit hash function. The similar to the earlier MD5 algorithm.
● SHA-2 is a relation of two similar hash functions, with dissimilar block sizes, known as SHA-256 and SHA-512. They differ in the word size; SHA-256 uses 32-bit words where SHA-512 uses 64-bit words.
● SHA-3 is a future hash function standard still in development [11].

SHA-1 Algorithm: 1-The SHA algorithm uses 5 state variables, each of which is a 32 bit integer (an unsigned long on most systems). These variables are sliced and dice and are (eventually) the message digest. The variables are initialized as follows:

h0 = 0x67452301

h1 = 0xEFCDAB89

h2 = 0x98BADCFE

h3 = 0x10325476

h4 = 0xC3D2E1F0

There are 80 rounds in SHA Algorithm.The hash value created by the sha hash function.

MD-5 Algorithm

MD5 is an algorithm that is used to confirm data integrity through the creation of a 128-bit message summary from data input (which may be a message of any length) that is demanded to be as unique to that specific data as a fingerprint is to the specific separate. MD5, which was developed by Professor Ronald L. Rivest of MIT, is intended for use with digital signature requests, which require that large files must be trampled by a secure method before being encrypted with a secret key, under a public key cryptosystem. MD5 [12] is currently a normal, Internet Engineering Task Force Request for Comments 1321. Allowing to the standard, it is "computationally infeasible" that any two messages that have been input to the MD5 algorithm could have as the output the same communication digest, or that a false message could be shaped through uneasiness of the message digest. MD5 is the third message digest algorithm created by Rivest[12]. All three (the others are MD2 and MD4) have similar constructions, but MD2 was optimized for 8-bit machines, in assessment with the two later formulas, which are optimized for 32-bit machines. The MD5 algorithm is an allowance of MD4, which the critical review found to be fast, but possibly not unconditionally secure. In comparison, MD5 is not quite as fast as the MD4 algorithm, but offers much more declaration of data security.

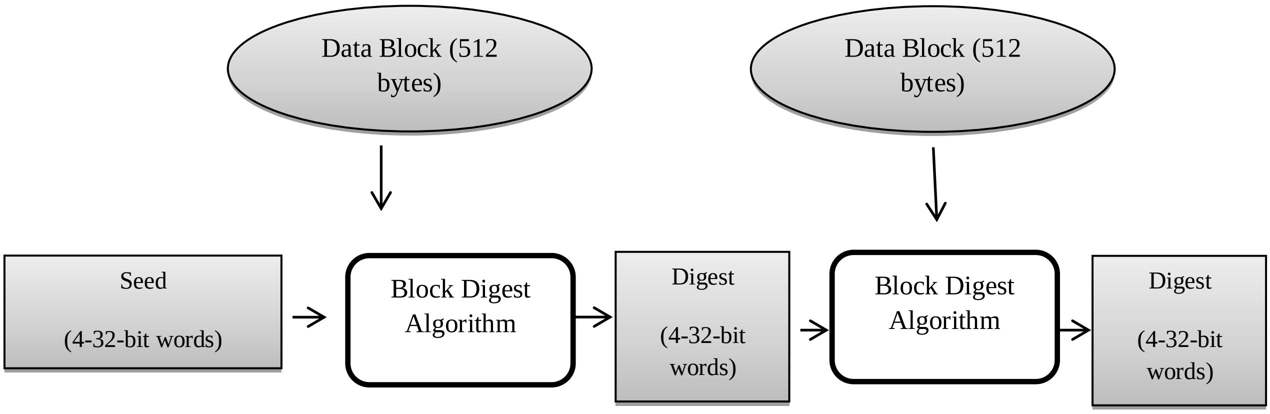

Fig. 3 MD-5 Digest Message

V. RESULT AND DISCUSSIONS

It is clear from this graph that the time taken by SHA2 algorithm is quite less than MD5 and SHA1 . Time taken by SHA2 for uploading file , updating file and deleting file is less than other algorithms . Hence SHA2 performs better as it lowers the time consumption rate in cloud .

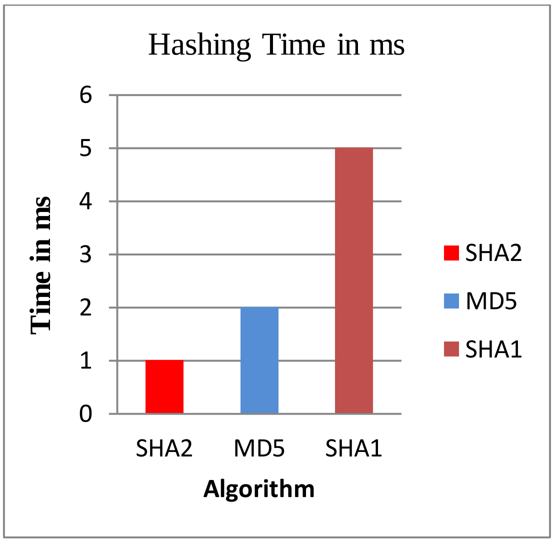

Fig. 4 Hashing Time in millisecond

Table 1: Hashing Time in ms

SHA2	MD5	SHA1
0.5	0.5	1.0
0.5	1	2.1
0.8	1.5	3.3
1	1.8	4
1.1	2	5

The below graph shows the memory used before file upload and after file upload. It is clear from the graph that the memory space is increased when we upload the new file in database. But when duplicate file is detected by using hashing algorithm then there is no effect on memory space it is same as before. In this way by using De-duplication memory space is less consumed.

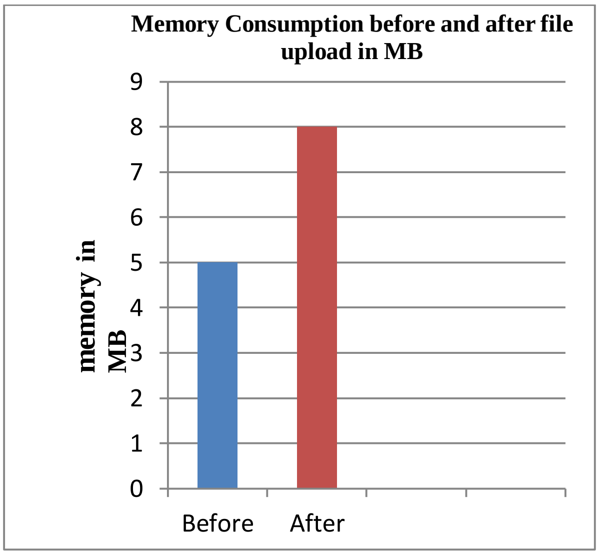

Fig. 5 Memory Consumption before and after file upload in MB

Table 2: Memory Consumption Before and After

Before	After
0.5	0.8
0.9	1.3
1.4	1.8
2.5	2.5
3.7	3.9
4	4.8
5	5.4
0	6.7
0	7.9
0	8

Detection time is that which is used by de-duplicator to detect the duplication of file. In which we calculate the detection time in micro second.

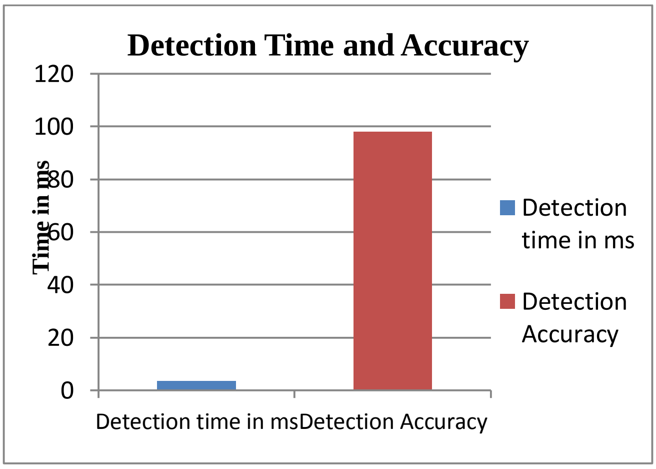

Fig.6 Detection Time and Accuracy

Table 3: Detection time and accuracy

Detection Time	Detection Accuracy
0.9	20
3	40
5	60
8	80
10	98

Accuracy shows that how accurately our system works to detect the duplicate files. From this graph we can conclude that duplicator detect the duplicate file in less time and perform it accurately.

VI. CONCLUSION

Cloud is a costly storage provider, so the motivation is to use its storage area efficiently. De-duplication has proved to reduce memory consumption by removing the useless duplicate files. So far from the previous studies file level de-duplication is the better method to be used, the focus of the proposed work will be on file level de-duplication. In this work, a dynamic data De duplication scheme for cloud storage is proposed, in order to fulfil a balance between changing storage efficiency and fault tolerance requirements, and also to improve performance in cloud storage systems. A lot of research has been carried out over this by means on hashing algorithm. SHA2 will perform better than MD5 and SHA1.Our aim is to choose a well-built algorithm which will generate a good hash value in time reducing cloud storage. In this proposed work the use of Microsoft azure provides the replica of the cloud computing environment which is used by many companies. Thus the work can easily be accomplished by the use of cloud framework without any cost consumption usage.

VII. REFERENCES

[1] Kaaniche, Nesrine, and Monique Laurent. "A secure client side deduplication scheme in cloud storage environments." *New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on*. IEEE, 2014.

[2] Li, Jin, et al. "A hybrid cloud approach for secure authorized deduplication."*Parallel and Distributed Systems, IEEE Transactions on* 26.5 (2015): 1206-1216.

[3] Leesakul, Waraporn, Paul Townend, and JieXu. "Dynamic data deduplication in cloud storage." *Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on*. IEEE, 2014.

[4] Backialakshmi, N., and M. Manikandan. "Data de duplication using N0SQL Databases in Cloud." *Soft-Computing and Networks Security (ICSNS), 2015 International Conference on*. IEEE, 2015.

[5] Nagarajaiah, Harsha, ShambhuUpadhyaya, and VijiGopal. "Data De-duplication and Event Processing for Security Applications on an Embedded Processor." *Reliable Distributed Systems (SRDS), 2012 IEEE 31st Symposium on*. IEEE, 2012.

[6] VasiliosAndrikopoulos, Zhe Song, Frank Leymann, “Supporting the Migration of Applications to the Cloud through a Decision Support System”, Institute of Architecture of Application Systems, IEEE, pp. 565-672, 2013.

[7] Haitao Li, LiliZhong, Jiangchuan Li, , Bo Li, KeXu, “ Cost-effective Partial Migration of VoD Services toContent Clouds”, 2011 IEEE 4th International Conference on Cloud Computing, pp. 203-110, 2011.

[8] C. Ward, N. Aravamudan, K. Bhattacharya, K. Cheng, R. Filepp, R. Kearney, B. Peterson, L. Shwartz, C. C. Young, “Workload Migration into Clouds – Challenges, Experiences, Opportunities”, 2010 IEEE 3rd International Conference on Cloud Computing, pp. 164-171, 2010.

[9] Kangkang Li, HuanyangZheng, and JieWu .“Migration-based Virtual Machine Placement in Cloud Systems”, 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet, IEEE, pp. 83-90, 2013.

[10] Jianxin Li, Yu Jia a, Lu Liub, TianyuWoa, “ CyberLiveApp: A secure sharing and migration approach for live virtual desktop applications in a cloud environment, Elsevier, Vol. 29, pp.334-340, 2013.

[11] Jung, Ho Min, et al. "Efficient data deduplication system considering file modification pattern." *International Journal of Security and Its Applications*6.2 (2012): 421-426.

[12] Zhang, Yang, Yongwei Wu, and Guangwen Yang. "Droplet: a distributed solution of data deduplication." *Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing*. IEEE Computer Society, 2012.