Hybrid SHA3 Algorithm for Reduction of Memory Consumption over Cloud Network

Harpreet Kaur[1], Maninder Kaur[2]
[1]M.Tech (Student), [2]Head of Department
Department of Computer Science, Department of Computer Science, Doaba Institute of Engineering and Technology, Kharar

Abstract - Cloud computing become more popular in providing fast, secure services and reduce the amount of storage space and save bandwidth. Data deduplication is one of the significant data looseness techniques for eliminating duplicate copies of repeating data. The proposed scheme in this paper not simply the decreases the cloud storage size but also improves the speed ,accuracy of data deduplication. The focus of the proposed work will be on file level de-duplication. In this work, we suggest a dynamic information De-duplication method for shade storage; in direct to fulfill stability between varying storage effectiveness & mistake tolerance desires, & also to pick up presentation in cloud storage systems. In this thesis hybrid algorithm is used to overcome or enhance the capacity of parameters. In hybrid (SHA1 & SHA3) algorithm some features of SHA1 & SHA3 are combined to obtain new algorithm. SHA3 was designed to be very efficient in hardware but is relatively slow in software.

Keywords: Cloud Computing, Data Deduplication, SHA-1 algorithm, SHA-3 algorithm, hybrid algorithm recovers the speed, complexity etc.

I. INTRODUCTION

CLOUD computing support data storage, processing, and management of backup data. Cloud computing provide different facilities to the individual user such as seemingly unlimited storage space and availability and accessibility of data anytime and anywhere. Cloud computing provides a big resource pool by linking network resources together. Cloud computing is increasingly become a commercial trend because of its desirable properties, such as scalability, elasticity, fault-tolerance, and pay-per-use. The most important and popular cloud service is data storage service. Cloud users upload personal or confidential data to the data center of a Cloud Service Provider and allow it to maintain and secure these data. Cloud computing offers a new service such as re-arranging various resources and data online over the Internet. Data storage is the most important and popular cloud service. In order to preserve the memory consumption of cloud, there is a need of data deduplication [1].

Data deduplication reduces the storage space requirements by eliminating redundant data. Although data deduplication removes data redundancy and data replication. There are some issues such as major data privacy and security are introduced when the user use it. Data deduplication is more popular due to space-efficient approach in backup storage systems. Data deduplication has been demonstrated to be an effective technique in Cloud backup and archiving applications to reduce the backup window, improve the storage-space efficiency. Existing solutions of data deduplication suffer from security weakness [2]. A hash value is formed by transformation that takes a variable size input m and returns a fixed-size string. A cryptographic hash function aims to guarantee a number of security properties. In this case collisions must exist but such collision should be hard to find the output appears random. An important application of secure hashes is verification message integrity. The hash value provides a digital fingerprint of a message's contents, which ensures that the message has not been altered by an intruder, virus, or by other means. Hash algorithms are effective because of the extremely low probability that two different plain text messages will yield the same hash value. HASH functions are common and critical cryptographic primitives. By far the most widespread hash functions are SHA-1 (Secure Hash Algorithm- 1), and MD5 (Message Digest) .It is so weak to a newly refined attack that it may be shattered by real-world hackers.SHA1 has long been measured supposedly broken, and all major browsers had already planned to stop accepting sha1 based signatures. To select a particular algorithm i.e SHA-3 standard is based on the strength of security and efficiency in Hardware implementation for wide variety of platforms [3].

II. LITERATURE SURVEY

R Maggiani et al. [4] proposed the Saas infrastructure for the improvement of administrations. Distributed computing can be a solitary capacity application, a framework on which these applications (and numerous others) can run, an arrangement of administrations that offer the benefits of enormous measures of processing assets, & the capacity to store a lot of information remotely. Numerous organizations & instructive infrastructures are simply starting to understand the advantages of cloud-based applications that have generally obliged site permitting, establishment, & support.

Ali Khaje et al. [5]proposed the outcomes demonstrate that the framework base for the situation learning would have price 38% fewer more than 6 years on EC2, & utilizing distributed computing could have conceivably wiped out 22% of the bolster requires this framework. These discoveries appear to be sufficiently huge to require a movement of the framework to the cloud yet our partner sway investigation uncovered that there are noteworthy dangers connected with this.

M. W. Storer, [6] built up two models for secure deduplicated storage: validated and anonyms. These two plans show that security can be consolidated with deduplication in a manner that gives an scope of security attributes. In the models they introduce, security is given through the utilization of merged encryption.

C. Ward et al. [7] represented acquainted the augmentations with a coordinated mechanization capacity called the Darwin structure that empowers on load movement for this situation & talk about the effect that computerized relocation has on the expense & dangers ordinarily connected with relocation to cloud.

Haitao et al. [8]proposed relocation methods taking into account (dynamic, receptive & shrewd procedures), albeit basically in light of the present information , can make the mixture cloud-helped VoD organization set aside to 30% transmission capacity cost contrasted & the Clients/Server mode. They can likewise handle unpredicted the glimmer group activity with little cost. It likewise demonstrates that the cloud cost & server transmission capacity picked assume the most essential parts in sparing expense, while the distributed storage size & cloud substance upgrade system assume the key parts in the client experience change.

Deepu, S. R. [9] proposedTwo methods were used for de-duplication, namely, chunk level & file level. For these levels hash values were computed by using MD5 algorithm. This application helped in easy maintenance of information on the cloud platform so that no duplicate files were saved in the Cloud.

III. SCOPE OF STUDY

This survey presents an extensive overview of deduplication in saving memory consumption in storage system. This survey describes new improved techniques that reduce the storage capacity needed to store data or the amount of data that has to be transferred over a network. These different types of techniques detect coarse-grained redundancies within a data set, e.g. a file system, and remove them [10]. One of the most important applications of data deduplication is backup storage systems where these approaches are able to reduce the storage requirements to a small fraction of the logical backup data size. As a first contribution, we extend the existing technique by modifying it and added new algorithm so that their accuracy and detection time could be better than earlier. For each of them, we describe the distinct approaches taken to address the main challenges of deduplication. The second contribution is the analysis of deduplication system in different levels: chunk level, file level etc. Each level type has distinct assumptions that impact the deduplication system's design. In this survey a hash is generated for every stream or file and compared it with stored hash so that duplicate data could be eliminated. We focus on deduplication systems that eliminate both intrafile and interfile redundancy over large datasets without any knowledge about data versions.

IV. IMPROVED TECHNIQUE DEDUPLICATION

Deduplication is an effective technique for optimization of instances of data stored in cloud storage . Deduplication can be further divided into chunk level and file level deduplication. Chunk level deduplication method results the storage of unique chunks by eliminating duplicate chunks. This method achieves better deduplication efficiency because it does exact deduplication. At the end, the throughput is low as it checks every incoming chunk for duplication. File level deduplication method easily eliminate duplicate files by calculating single checksum of complete file data. Files who has same properties are referred to as similar files. This method achieves better throughput because it compares every chunk only with chunks of similar files. However, the deduplication efficiency is comparatively low as some duplicate chunks may be found among different groups. Hence, this technique performs only approximate deduplication. In order to achieve more utilization of storage area, then identify the duplicate data in files. Every incoming file is divided into chunks. Based on how the incoming chunk is checked against duplicates, deduplication can be categorized into two types, namely, chunk and file level.

1. *Chunk level Deduplication :*

Whenever a data stream has to be written,it firstly identified the chunk from stream so that it check against stored chunk for deduplication. This is termed as chunk level deduplication. Since every incoming chunk is checked for eliminating duplication, only unique chunks occupy the cloud storage. Therefore, chunk level deduplication has better deduplication efficiency. However, as each incoming chunk is checked against a large list of chunk indices, the number of disk I/O operations is large. This has a significant impact on deduplication throughput. Previously used storage backup workload demands good deduplication efficiency as it involves large data redundancy among different workloads. Hence, this deduplication approach is best suited for such workloads [11].

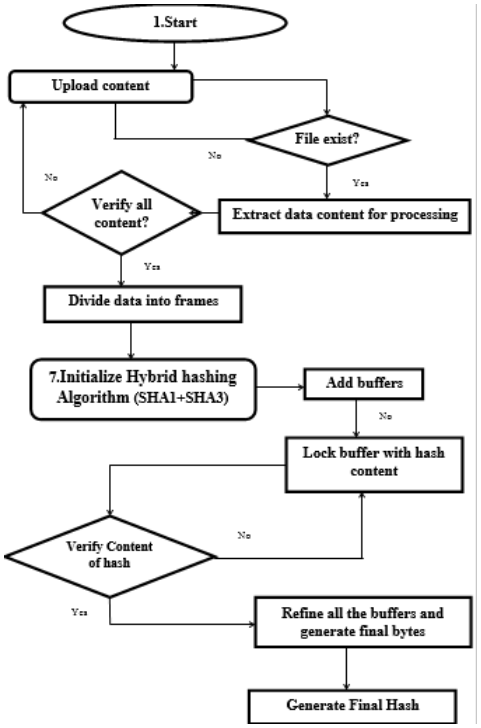
2. *File Level Deduplication:*

Whenever a data stream has to be written, every chunk in the stream is checked against the chunks of similar files. This is termed as file level deduplication. This approach provides a scalable solution with the division of chunk index into two parts i.e Primary and Secondary index .

In this approach, all the Chunk that constitute a file and the minimum Chunk among them are found. This minimum Chunk is termed as representative Chunk_ID. According to Broder's Theorem, two files are similar, when their representative Chunk_IDs of both the files are same. Primary chunk index consists of representative Chunk_ID, whole file hash and the address of the secondary index or bin. Bin is made up of three fields namely, Chunk_ID, chunk size and the storage address of the chunk. Whenever there is a need to bake up the file, it is chunked and both the representative Chunk_ID and the hash value for the entire file are found. In the primary index, Representative Chunk_ID is searched and if it is not present, the incoming file is considered as new. In the disk, a new bin is created and all Chunks, their corresponding size and a pointer to the actual chunks are stored. To update the primary index after new bin is created, Representative Chunk_ID, hash value of a new file and the pointer to the newly created bin are added. If the hash value of the whole file does not match but representative Chunk_ID of the incoming file is already present in the primary index then the incoming file can be considered to be nearly similar to the one that is already on the disk. Most of the chunks of this file will probably be available in the disk. The whole file hash value is not modified in the primary index and the updated bin is written back to the disk. There is no need to update the bin if the whole file hash value of the incoming file is found in the primary index, then the incoming file is considered as a duplicate. Since every incoming chunk is checked only against the indices of similar files, this approach achieves better throughput as compared to the chunk level deduplication [12].

V. PROPOSED WORK

For analysis of the proposed work , thorough study of existing deduplication technique is done and compared with the proposed method.


1.Start
Upload content
File exist?
No
Yes
Verify all content?
No
Extract data content for processing
Yes
Divide data into frames
7.Initialize Hybrid hashing Algorithm (SHA1+SHA3)
Add buffers
No
Lock buffer with hash content
Verify Content of hash
No
Yes
Refine all the buffers and generate final bytes
Generate Final Hash

After Generating Final Hash, Verify the contents as shown below:

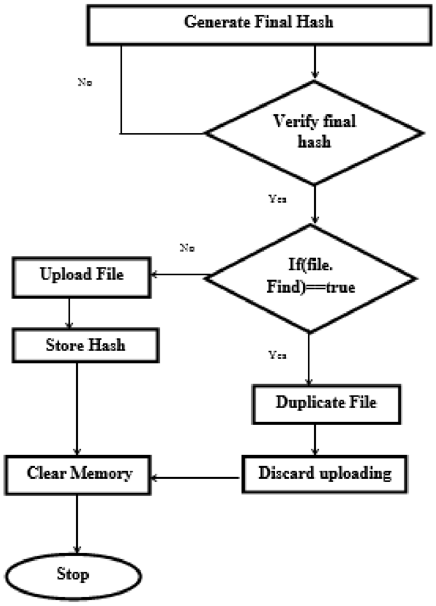


Fig.1: Proposed Work

VI. RESULT AND DISCUSSION

In this research paper some of the parameters are to be discussed and their comparison with previous algorithm used.

Some parameters are memory consumption, accuracy, detection time, hashing time and complexity.

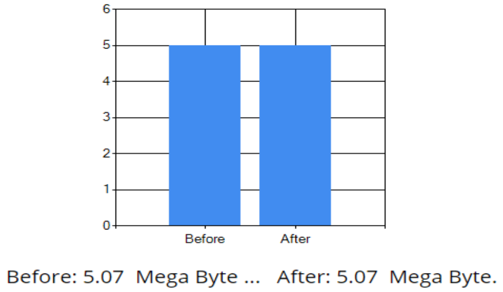
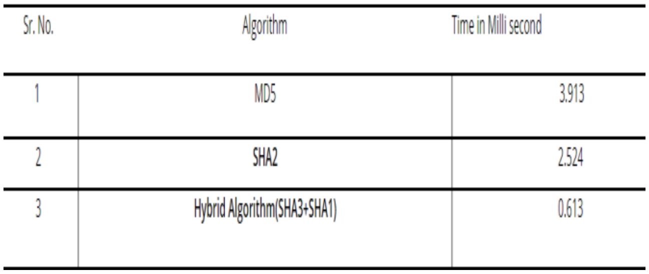


Fig.2: Memory Consumption

The Fig.2 shows memory structure over a cloud server. Here the both values before and after same because the file uploaded is same as upload before by someone else. So in this case the file will shared with this user and system discard the upload process to save memory consumption.



Sr. No.	Algorithm	Time in Milli second
1	MD5	3.913
2	SHA2	2.524
3	Hybrid Algorithm(SHA3+SHA1)	0.613

Fig.3: Time Complexity

The Fig.3 shows that there are various other algorithms which are used to generate hash of a particular file. Here this diagram shows three different algorithm's time complexity for hash generation. The proposed hybrid algorithm is working better than others as shows in the table.

Other parameters are as accuracy for hashing and detection time over cloud storage. The accuracy parameter used to define the performance of algorithm in terms of their output. And detection time is used to shows the searching optimization of a cloud storage. It shows how efficiently it detect duplicate file over a cloud network in terms of time. The various test cases are as shown below

SHA1 algorithm: Accuracy: 92.52% and Detection Time:9.51ms

SHA2 algorithm: Accuracy: 94.88% and Detection Time: 7.85ms

HYBRID algorithm: Accuracy: 97.18% and Detection Time: 2 .63ms

Times saving in % with various file tests with the use of 5 duplicators. It is more time saving process when files uploaded by proposed technique. Here all the cases define better performance by proposed technique as shown in Fig.4.

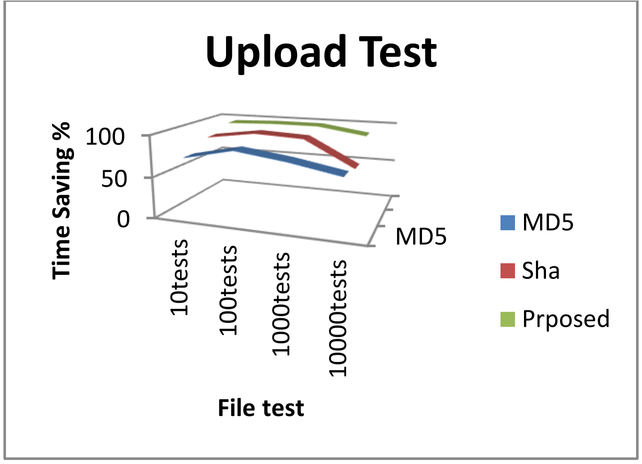


Fig.4: Detection Time

The same technique also tests in the 10 duplicators and here also it saving more time than other algorithms as shown in Fig.5.

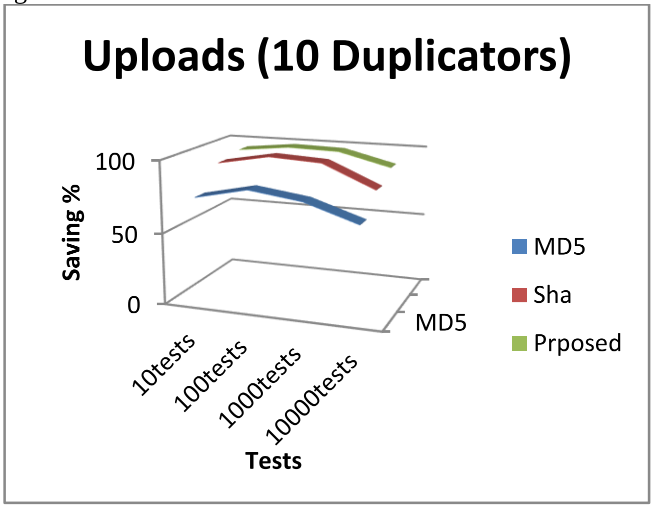


Fig.5: Detection Time

Times saving in % with various file tests with the use of 5 duplicators. It is more time saving process when files updated by proposed technique. Here all the cases define better performance by proposed technique as shown in Fig.6.

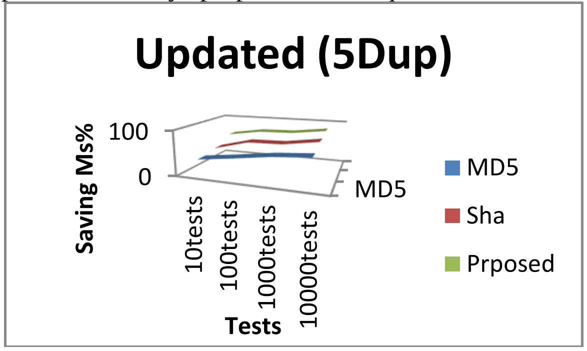


Fig.6: Detection Time

The same technique also tests in the 10 duplicators and here also it saving more time than other algorithms as shown in Fig.7.

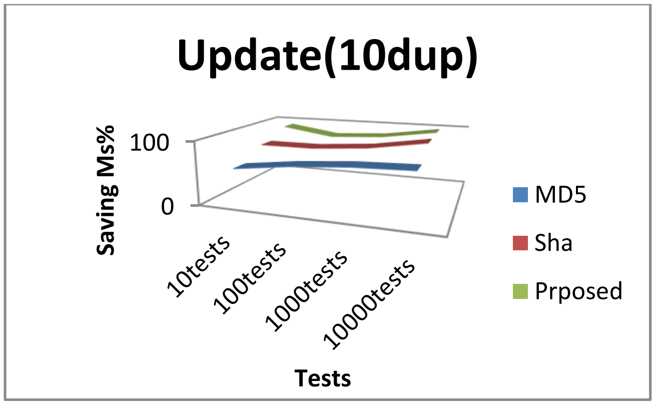


Fig.7: Detection Time

Times saving in % with various file tests with the use of 5 duplicators. It is more time saving process when files delete by proposed technique. Here all the cases define better performance by proposed technique as shown in Fig.8.

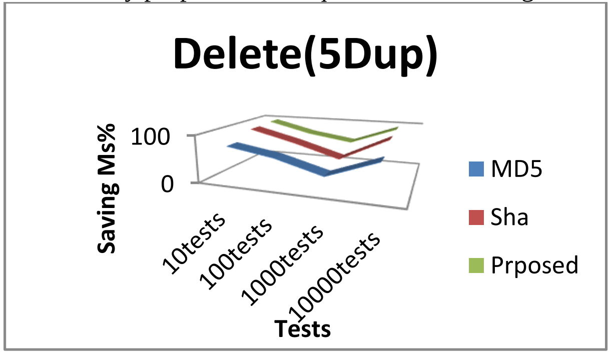


Fig.8: Detection Time

The same technique also tests in the 10 duplicators and here also it saving more time than other algorithms as shown in Fig.9.

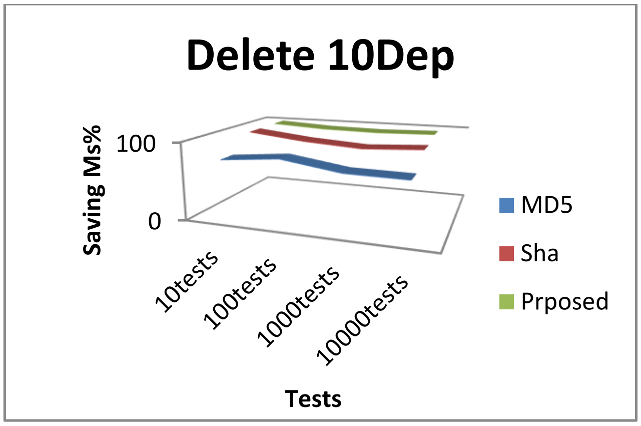


Fig.9: Detection Time

VII. CONCLUSION

Cloud is the costly storage provider, so the motivation is to use its storage area efficiently De-duplication has been proved to reduce memory consumption by removing the useless duplicate files. So far from the previous studies file level de-duplication is the better approach to be used, the focus of the proposed work will be on file level de-duplication. In this work, we suggest a dynamic information De-duplication method for shade storage; in direct to fulfill stability between varying storage effectiveness & mistake tolerance desires, & also to pick up presentation in cloud storage systems. A lot of research has been carried out over this by means on hashing algorithm.

From the previous hashing algorithms Hybrid will perform better than SHA-1 and SHA-3 Techniques.

We achieved the best performance according to the SHA-3 Hashing Algorithm. Then evaluate the performance parameters like utilization, accuracy and memory consumption etc.

VIII. REFERENCES

[1]. Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE, “Deduplication on Encrypted Big Data in Cloud”, IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 2, APRIL-JUNE 2016.

[2]. Wen Xia, Hong Jiang, Senior Member, IEEE, Dan Feng, Member, IEEE,and Yu Hua, Senior Member, IEEE, “Similarity and Locality based Indexing for High Performance Data Deduplication” DOI 10.1109/TC.2014.2308181, IEEE Transactions on Computers.

[3]. Muhammad Arsalan, Dr. Arshad Aziz, “Comparative Analysis of high speed and low area architectures of Blake SHA-3 candidate on FPGA”, 2012 10th International Conference on Frontiers of Information Technology.

[4]. R. Maggiani, "Cloud computing is changing how we communicate," 2009 IEEE International Professional Communication Conference, IPCC 2009,Waikiki, HI, United

states ,pp 1, July 2009.

[5]. Khajeh-Hosseini, A., Greenwood, D., Sommerville, I., (2010). Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS. Submitted to IEEE CLOUD 2010.

[6]. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008.

[7]. 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.

[8]. 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.

[9]. Deepu, S. R. "Performance Comparison of De-duplication techniques for storage in Cloud computing Environment." Asian Journal of Computer Science & Information Technology 4.5 (2014).

[10]. Singh, deepika, and preetika singh. “New challenges for Security against Deduplication in cloud Computing” International Journal 2.1(2014).

[11]. Amrita Upadhyay, Pratibha R Balihalli, Shashibhushan Ivaturi and Shrisha Rao, “Deduplication and Compression Techniques in Cloud Design" by 2012 IEEE .

[12]. Deepu S R1,Bhaskar R2,Shylaja B S3 “Performance comparison of Deduplication Techniques for Storage in Cloud Computing Environment”, Asian Journal of Computer Science And Information Technology 4 : 5 (2014) 42 - 46.

[13]. Deepu, S. R. "Performance Comparison of Deduplication techniques for storage in Cloud computing Environment." Asian Journal of Computer Science & Information Technology 4.5 (2014).