IJRECE VOL. 4 ISSUE 3 JULY.-SEPT. 2016 ISSN: 2393-9028 (PRINT) | ISSN: 2348-2281 (ONLINE)
DOI: 10.13140/RG.2.1.2594.7127

Enhanced and optimized data de-duplication technique to eliminate duplicate data with Hybrid hashing in cloud environment

Navdesh kaur[1], Amandeep Verma[2]
Student (M.Tech), Assistant Professor
Department of Computer Science of Engineering
Punjabi University Regional Centre for Information Technology and Management, Mohali

Abstract- Cloud computing offers an on demand, self-service, speedy, adaptable and universal access to several computing properties and resources. Deduplication is the procedure of eliminating redundancy to alleviate storage needs, only one distinctive instance of the information is really preserved on storage media, like disk or tape. In the method of de-duplication, additional copies of constant information are deleted, leaving only 1 copy to be stored. Information is analysed to spot duplicate byte patterns to make sure the single instance is indeed the only file. Then, duplicates are replaced with a reference which points to the stored single instance of file. Data de-duplication is widely used in cloud storing to save bandwidth and minimalize the storage space on Server. However, current explores on data Deduplication, which mostly focus on the static sections such as the backup and store systems, are not appropriate for cloud storage system due to the dynamic nature of data. Data Deduplication strategies are conveyed to recover storage potency in cloud storages. With the dynamic setting of information in cloud storage, data procedure in cloud changes actively, some information portions could also be delivered usually in period of time, however might not be utilized in another time period. Some datasets could also be ordinarily accessed or updated by multiple users at a similar time, whereas others may need the high level of redundancy for reliability demand. To secure the privacy of sensitive data during Deduplication, the convergent hashing technique SHA3 is used to find duplicity in the data before Storage. For optimisation the Artificial Bee Colony Optimization Algorithm is used to ensure the unique hash data. For better data protection, this dissertation talks about the issue of data Deduplication authorization.

Keywords- *Deduplication, authorization, Cloud Computing, Block Level and Artificial Bee Colony Optimization Technique.*

I. INTRODUCTION

Cloud computing is a long time back envisioned vision of registering as an utility, where information proprietors can remotely store their information in the cloud to acknowledge on-demand choice applications and administrations from a common pool of configurable system resources. Cloud is another arrangement of action wrapped around new advances, for instance, server virtualization that exploit economies of scale and multi-inhabitancy to diminish the cost of using information innovation resources. It in like manner conveys new and challenging security threats to the outsourced information [1]. Since cloud administration suppliers (CSP) are specific administrative substances, information outsourcing truly surrenders the proprietor's conclusive control over the fate of their information. Cloud computing and stockpiling courses of [2] action outfit customers and tries with various capacities to store and maintain their information in outcast information centres. It relies upon sharing of resources to achieve soundness and economies of scale, similar to a utility (like the power matrix) more than a framework. At the foundation of distributed computing is the more broad thought of consolidated foundation and shared administration.

Cloud computing, or basically "the cloud", concentrates on expanding the adequacy of [3] the common resources. Cloud resources are for the most part utilized and shared by various customers and in addition rapidly and responsibly reallocated per demand. When anyone use the term "moving to cloud" they generally are promoting the idea of stepping towards the OPEX [4] model in which the uses utilizes a common framework and pay as they uses from the traditional CAPEX model in which users purchase the required equipment and then use it more than a time [2].

Data de-duplication is information compression techniques which reduces the storage capacity by eliminating duplicate copies of information or reduce the sum of information that has to be transfer over a complex [5].

Data De-duplication – Also identified as ‘Single Case Storage’.Data de-duplication not simply reduce the storeroom gap necessities by eliminating redundant information but minimizes the system broadcast of photocopy information in the system storeroom system. It is a means of dropping storage wants by eliminate disused statistics the optimization of storage is identified as de – duplication storing space. Only one exclusive illustration of the information is really retain on storage medium

Data de-duplication done at client side & server side [6]. In client side de-duplication is done previous to distribution the information to a storage space device. Only unique information is transferred to the mechanism with the

DOI: 10.13140/RG.2.1.2594.7127

minimum available band measurement & it requires less time. At server side de-duplication is done after sending the information to storage scheme. De-duplication is as well used in back up services to reduce network bandwidth.

It minimizes the complex programming of redundant information in the complex storage scheme.

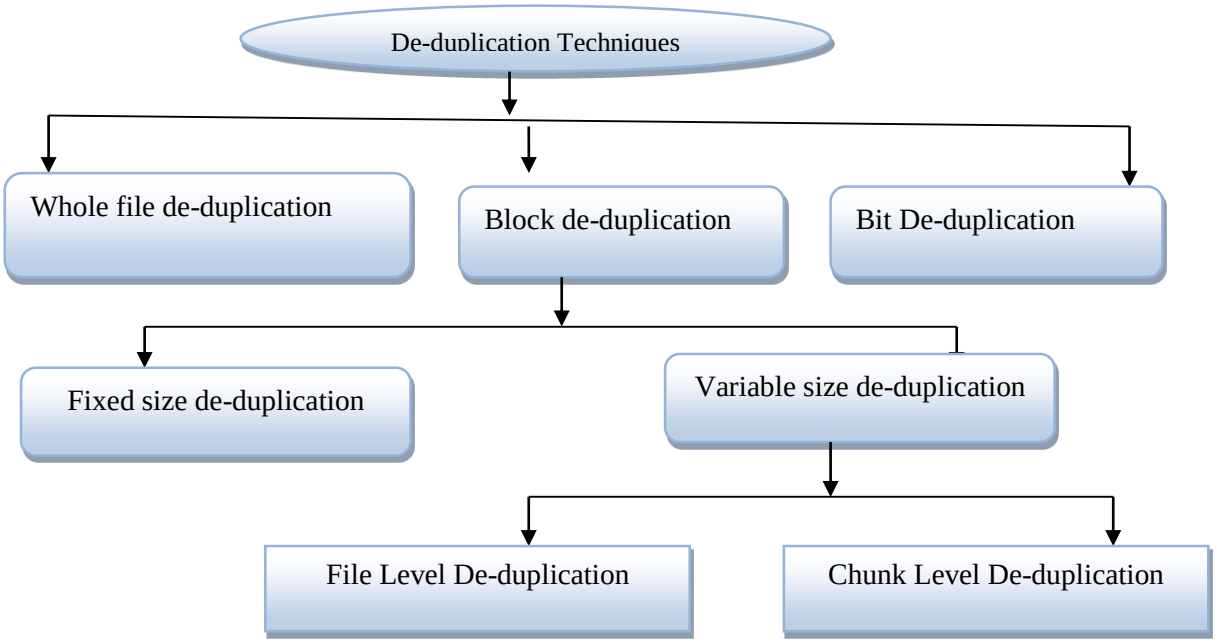


Fig no: 1 Methods of Deduplication

A. Techniques of de-duplication

1) Hashing Technique

In hashing technique hashing the information means creating a hash value or number of the file, block & byte which guarantee to be unique for all the above types [7]. In hashing technique some hashing algorithms are used. These hashing algorithms have their own properties like their output size, block size, rounds & performance. Hashing technique is used after uploading the file. When fingerprint of the file is generated then it is stored in the metadata & used for the comparison purpose. From the above block diagram the file to be uploaded is fed to the hashing method which generates the hash worth. The hash value is compared with already existing hash values. If a matching hash value is found the particular file will not be added to the cloud storage, else server will store the file. If two files have same hash value then it is said to be a similar file otherwise it is considered as a different file. By using this technique storage space is reduced & time is also saved to find the duplicate files. Searching of files is also easy when we have customized information storage.

2) Application Aware de-duplication

This technique of de-duplication is known as Byte Level De-duplication, because in which deepest level of de-duplication is performed. It is also known as Content Aware De-duplication where all the content of the objects, files, applications. Data is divided into blocks & then check the bytes & stored only those bytes which are not unique. This method is so time consuming method & some loss of data is possible.

II. SHA-3 ALGORITHM

SHA-3 is an associate of the Secure Hash Algorithm . The SHA-3[10] archetypal was unconfined by NIST in 2015.Dubbed Keccak the secluded hash method, which will confidently be recognized as SHA-3, beat 63 other proposal after NIST dispersed an unlock call for a SHA-2 substitute in 2007. That shift was ambitious by doubts which so far haven't come to pass--that SHA-2 might be susceptible to being fractured. Hashing method are significant in sequence safety tool, & new to corroborate communication, as well as digital signature & entrance permit. "A noble hash process has a lesser number of essential kind," according to NIST. "Some modify in the unique message, though little, must reason a modify in the concentration, & for some known file & assimilate, it should be infeasible for a falsifier to make a varied file with the similar assimilate." SHA-3 uses the sponge construction, within which information is "engrossed" into the scrubber, & then the consequence is "hug" unavailable. In the fascinating stage, communication tablets are XORed into a separation of the condition, which is before transmuted as a complete. In the "clutch" stage, manufacture blocks are taken from the similar separation of the condition, alternate by formal transformation.

DOI: 10.13140/RG.2.1.2594.7127

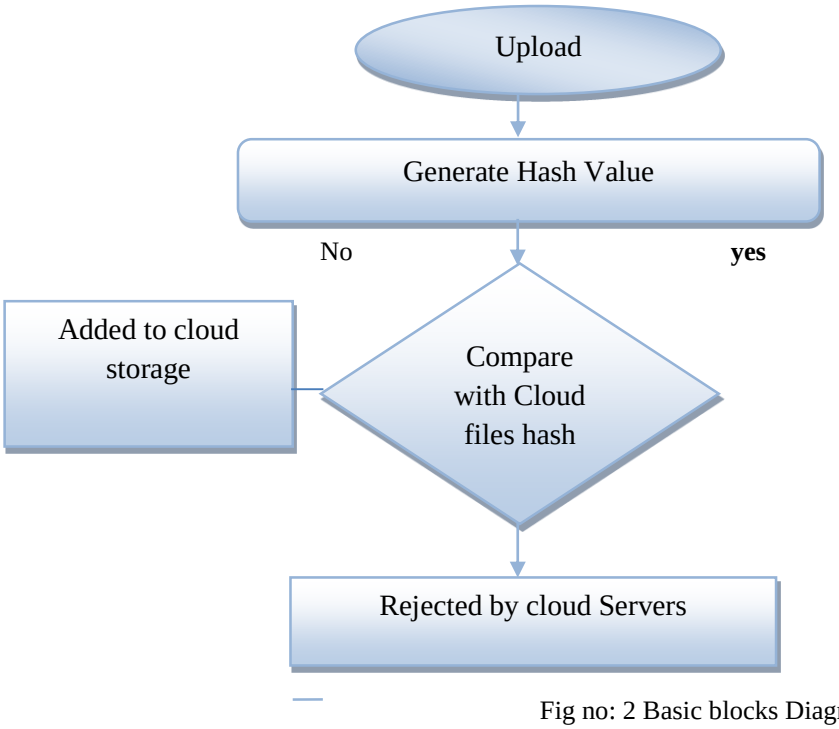


Fig no: 2 Basic blocks Diagram

The dimension of the piece printed & convert is entitled "rate", & the fraction that is unspoiled by contribution/productivity is entitled "capability". The capability decides the safety of the format. The utmost sanctuary equal is partial the capability.

III. RELATED WORK

Vasilios et al [11] 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 & 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.** [12] 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. **C. Ward et al**. [13] 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. **Kang et al.** [14] proposed the migration algorithm .The VM to its best PM specifically, with the proviso that it has adequate capability. Then, if the relocation restriction is gratified, we transfer a different VM after this PM to oblige the novel VM. In addition, we are learning a mixture system where a lot is working to recognize forthcoming VMs for the on-line expansion. Assessment upshots establish the great competence of our method. **Xian Xin et al.** [15] planned a lively model scheme expression CyberLiveApp to carry request contribution & relocation on command amongst a variety of consumer CyberLiveApp gives 2 key management: a safe multi-client contribution management for the practical desktop of a VM & multi-VM request distribution & group.

IV. EXISTING PROBLEM

Deduplication makes system a network efficient and storage optimization systems. At present, in the background of customer information allocation proposal the contests for great scales, extremely unnecessary internet information storing space is great. Due to this idleness storing space charge is decreases. Loading for this gradually central network information can be receiving by its de-duplication. The problems with existing information storage system.[1]

1. If we consider a case in which user update one same file on multiple time, it take space a lot on server memory.
2. If server have large amount of information then searching technique become slow.

DOI: 10.13140/RG.2.1.2594.7127

3. Unwanted space consumption is a very costly when user are in billion.
4. Current hashing function or searching technique is not much better It is a more time consuming process to search any of records or de- duplicate any new content.

Data is most important in the system so we need accurate verification generator algorithm which finds files fast and accurately and current systems have this type of functions but the accuracy is less.

V. SIMULATION ANALYSIS

The objective of this thesis is to propose hybrid approach expending De-Duplication Approach for Reducing Memory Consumption in Cloud Computing using ABC (Artificial Bee's Colony Optimization) and SHA-3 algorithm hash code generation. To propose hybrid approach expending De-Duplication Approach for Reducing Memory Consumption in Cloud Computing using hybrid approach (ABC+SHA3) algorithm hash code generation. To compare proposed method in terms of various parameter matrixes with existing de-duplication schemes.

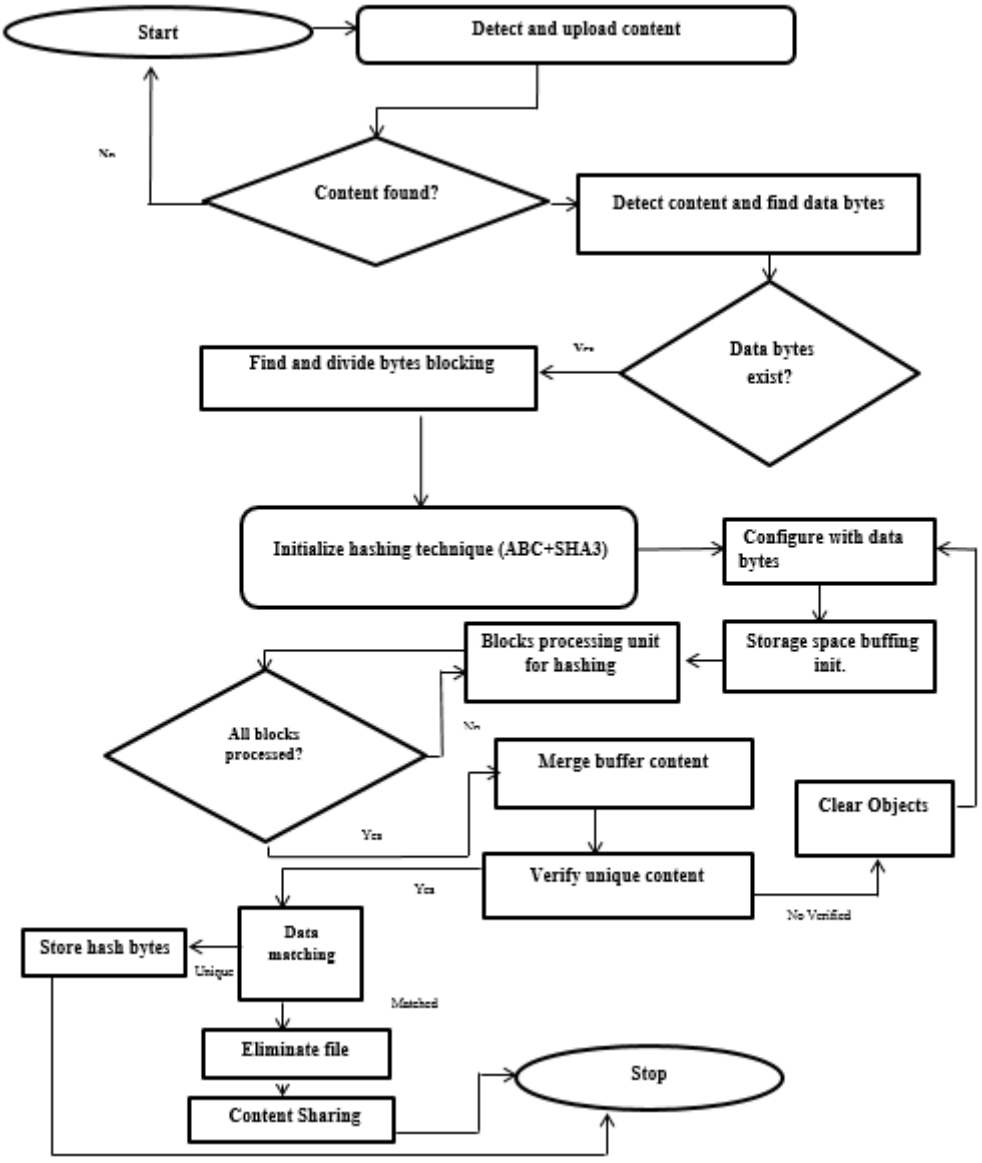


Fig.3 Proposed Work Flow Chart

VI. RESULT ANALYSIS

The comparison is done between the three techniques MD5, SHA1 and the proposed technique have been compared on

DOI: 10.13140/RG.2.1.2594.7127

basis of detection time, detection accuracy, memory consumption, Hash Time. The result has been extracted for previous techniques and the proposed technique.

A. Accuracy and Detection time

The performance evaluation table shows different algorithms performance in terms of detection accuracy and detection time of hashing from a dataset.

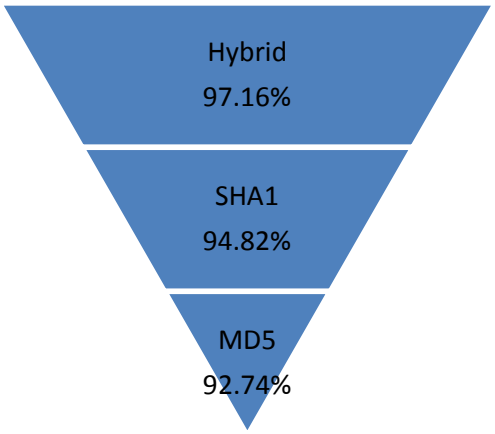


Fig.4: Accuracy

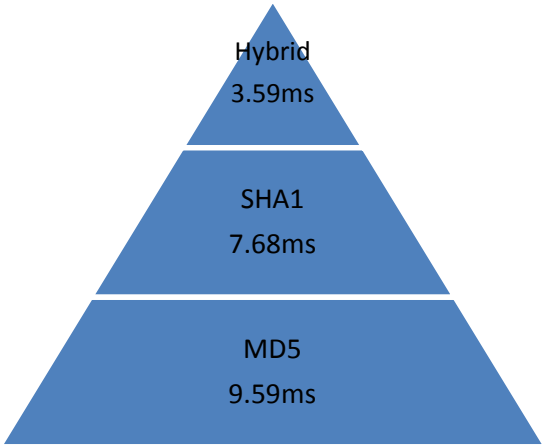


Fig.5: Detection Time

Figures 4 shows that proposed Deduplication technique gives better accuracy that the previous techniques (97.16%). In figure 5 Detection Time of hybrid algorithm is significantly better than the other two techniques.

The various test cases showed below which defines a better performance calculations.

Table 1 Detection Accuracy

	test1	test2	test3	test4	test5
MD5	92.82	91.56	91.32	92.32	90.86
SHA1	94.82	94.65	93.45	94.01	93.21
Proposed	97.16	98.26	97.66	97.68	98.22

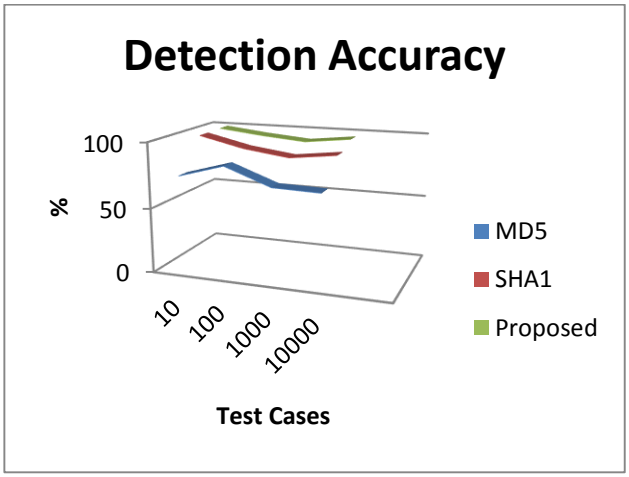


Fig.6 Performance Chart of Detection Accuracy

Here in all the cases the Detection Accuracy is always better than other techniques. This process shows the better performance of proposed technique in the cloud environment.

Table 2 Detection Time

	test1	test2	test3	test4	test5
MD5	9.59	9.23	10.32	9.25	8.88
SHA1	7.68	7.82	7.65	7.6	7.99
Proposed	3.59	3.65	2.96	3.98	3.24

DOI: 10.13140/RG.2.1.2594.7127

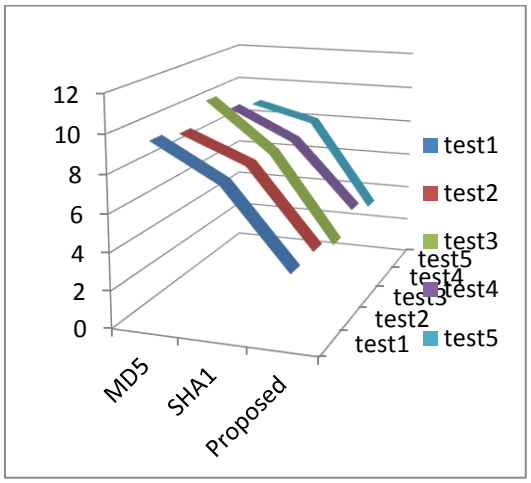
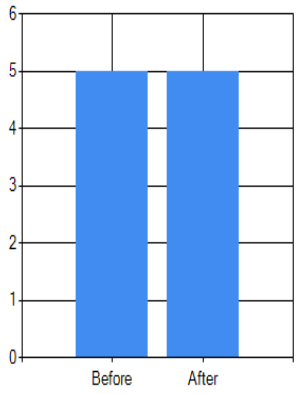


Fig.7 Performance Chart of Detection Time

It shows the detection of content from the dataset in terms of time (in millisecond). As the accuracy parameter, the detection time also lead the other two. The parameter with different test cases showing the better results of proposed technique in the whole processing.

B. Memory Consumption

Another parameter deals with memory consumption which shows the results of proposed approach that how much data is stored and uploaded inside the disk drive. In the graph below both bars are same that means file is not uploaded because of duplicate content.



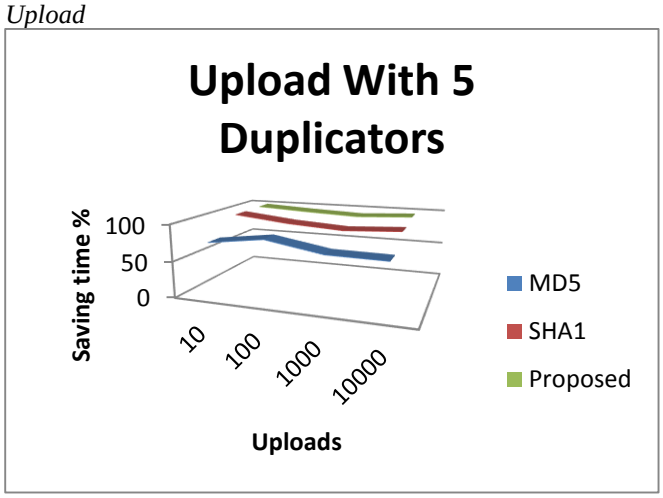
Before: 5.07 Mega Byte ... After: 5.07 Mega Byte.

Fig.8: Memory Consumption Graph

C. Results using Multiple Duplicators

After performing the working of algorithm's inner structure, system create some duplicators to check the upload, update and delete time saving of algorithm with multiples files.

Upload



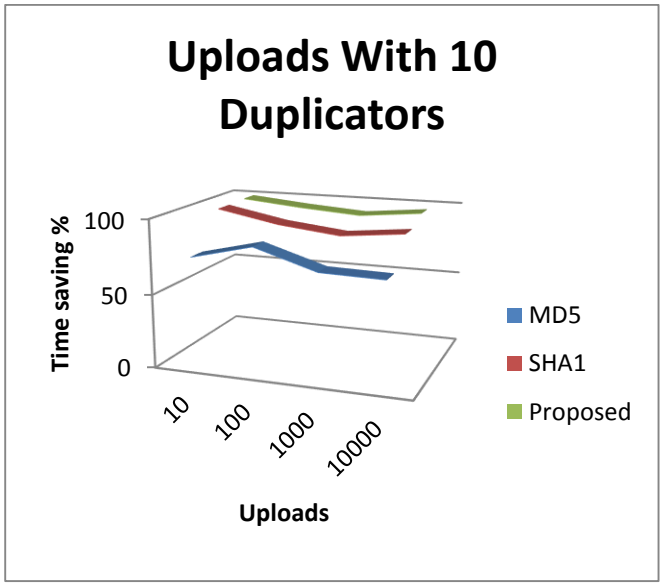


Fig.9: Upload time Graphs for 5 and 10 duplicators

The performance of both 5 and 10 duplicators with different number of upload is better with the use of proposed technique. It is saving more time than other algorithms as shown in test cases.

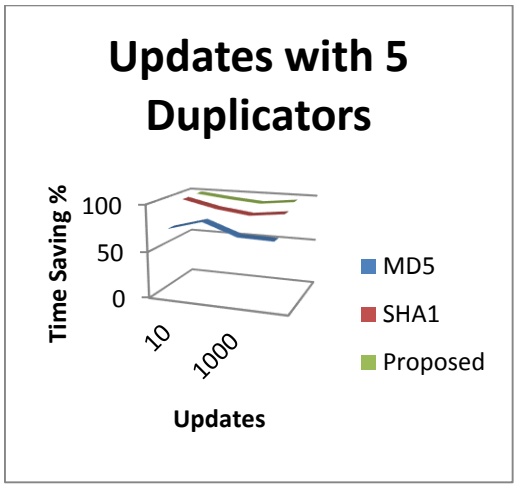
DOI: 10.13140/RG.2.1.2594.7127

Update

FILE UPDATE PARAMS

Number of Uploads	5 Duplicators			10 Duplicators		
	md5	Previous	Hybrid Algorith	md5	Previous	Hybrid Algorith
Time saving using 10 Updates	34	41.78	58.22	56	75.02	90.15
Time saving using 100 Updates	48	61.79	71.16	67	75.34	77.18
Time saving using 1000 Updates	56	61.78	73.28	75	82.09	84.18
Time saving using 10000 Updates	60	75.25	82.22	79	96.17	97.18

Fig.10 File Update Parameters



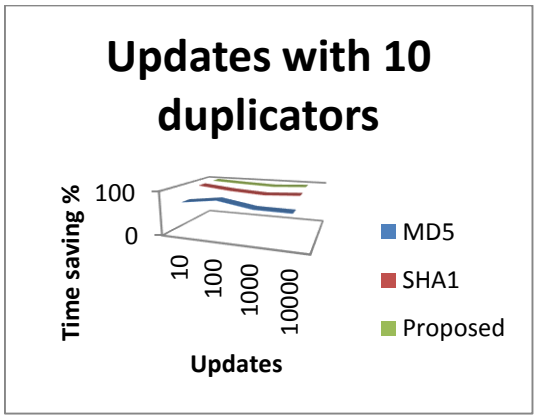


Fig.1 Time Saving graphs for file update

The performance of both 5 and 10 duplicators with different number of update is better with the use of proposed technique. It is saving more time than other algorithms as shown in test cases.

Delete

FILE DELETE PARAMS

Number of Uploads	5 Duplicators			10 Duplicators		
	md5	Previous	Hybrid Algorith	md5	Previous	Hybrid Algorith
Time saving using 10 DELETE	77	93.42	95.16	74	98.68	99.15
Time saving using 100 DELETE	59	69.31	74.18	84	90.59	95.17
Time saving using 1000 DELETE	32	40.74	56.15	71	85.87	92.15
Time saving using 10000 DELETE	71	90.28	94.16	70	90.03	96.17

Fig.12 File Delete Parameters

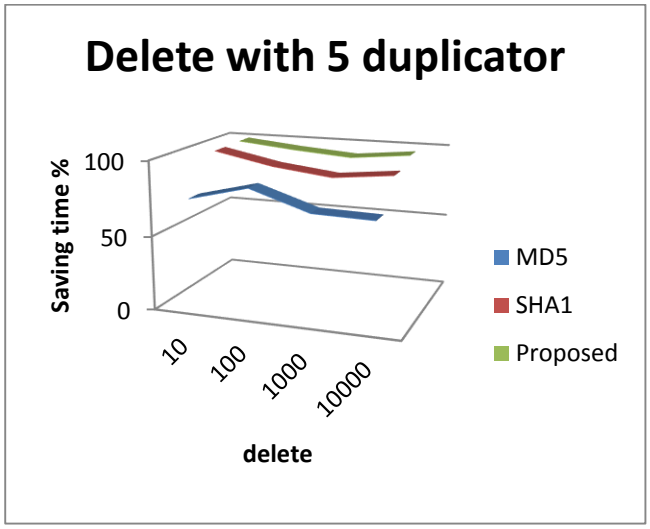


Fig.132 Time Saving graph for file deletion (5 Duplicators)

DOI: 10.13140/RG.2.1.2594.7127

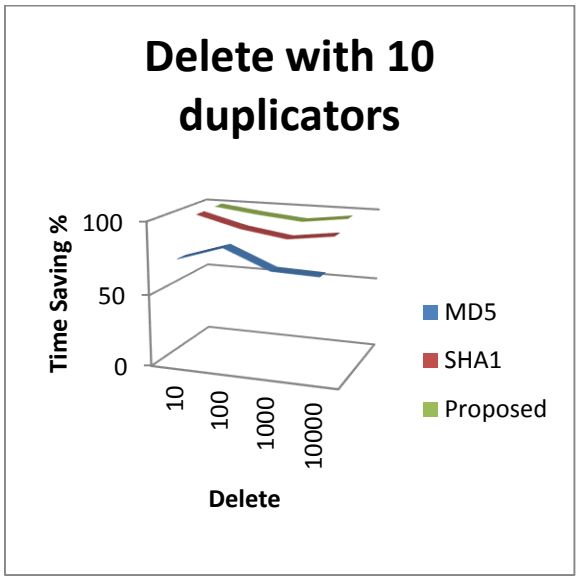


Fig.143 Time saving graph for file deletion (10 Duplicators)

The performance of both 5 and 10 duplicators with different number of Delete is better with the use of proposed technique. It is saving more time than other algorithms as shown in test cases.

D. Hashing Time

Sr. No.	Algorithm	Time in Milli second
1	MD5	3.913
2	SHA1	2.524
3	Hybrid Algorithm	0.613

Fig.15 Hashing Time

This table shows the execution time complexity of different algorithm to find a unique content from a particular file. The less complexity defines a good quality of algorithm in this case. Here the working of proposed algorithm is better than the other algorithms. It detects the uniqueness of a file in less than one millisecond.

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 fulfil stability between varying storage effectiveness & mistake tolerance desires, & also to pick up presentation in cloud storage systems. A lot of analysis has been applied out over this by means on hashing algorithm. From the previous hashing algorithms SHA2 will perform better than SHA3 and Artificial Bee Colony Optimization Technique. 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.

VIII. REFERENCES

[1]. Andrikopoulos, Zhe Song, Frank Leymann, "Supporting the Migration of functions to the Cloud through a Decision Support System", Institute of Architecture of function Systems, IEEE, pp. 565-672, 2013.

[2]. 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.

[3]. 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.

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

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

[6]. Jyoti Malhotra1,Priya Ghyare2, "A Novel Way of De-duplication Approach for Cloud Backup Services Using Block Index Caching Technique", Vol. 3, Issue 7, July 2014 DOI: 10.15662/ijareeie.2014.0307040

[7]. Upadhyay, Amrita, et al. "Deduplication & compression techniques in cloud design." Systems Conference (SysCon), 2012 IEEE International. IEEE, 2012

[8]. Monjur Ahmed & Mohammad Ashraf Hossain, "CLOUD COMPUTING & SECURITY ISSUES IN THE CLOUD" International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.1, January 2014.

DOI: 10.13140/RG.2.1.2594.7127

[9]. Vasilios Andrikopoulos, 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.

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

[11]. 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.

[12]. 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.

[13]. 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.

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

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