iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Storage Optimization Using De Duplication: A Better Approach

Anju Bala Malhotra[1], Jasbeer Narwal[2]
[1]Department of Computer Science, Haryana Engineering College, Jagadhri, Kurukshetra University, Haryana, India

Abstract: Cloud storage is one of the services provide in cloud computing which has been increasing in reputation. With the growing data size of cloud computing, a decrease in data volumes could help provider reducing the costs of running large storage system and saving energy use. So data de-duplication techniques are brought to recover storage competency in cloud storages. In this paper, we propose a dynamic de-duplication system for cloud storage, that aim to advance storage competency and maintain redundancy for fault acceptance.
Keywords: Cloud computing, De-duplication, Cloud Storage,

I. INTRODUCTION

Data de-duplication identifies a new type of approaches in which slow up the storage capacity necessary to store info or maybe the number of files which needs to be moved over a linkage. These kinds of approaches discover coarse-grained redundancies inside a data set, at the. g. a new file structure; Data de duplication not only decreases this storing space necessities by abolishing redundant files but also lowers the system transmitting of replicate files from the network storage systems.

A Methods of De- Duplication

Determined by data de-duplicated you can find a two strategies throughout De-duplication

1) *File Level De- duplication*: Using this method 1st picks up identical data which is taken off. Only one replicate of file will be kept.
   A Pointer is utilized to point an original file for the next subsequent replicates. In this method, it doesn't consider the items present inside the file.
   One example is two file data together with easy title change are stored two different files. The advantage of this method is quite simple and also quickly. Using this method will be often known as Single Instance Storage [33].
2) *Block or Sub File De-duplication:* This File will be separated in several chunks termed blocks and also replicate blocks are usually diagnosed using special hash formula. If the file is uniqely written in hard disk drive different only pointer is utilized to point this hard disk drive area. According to size of block you can find that there are two procedures throughout block De-duplication.
3) *Fixed Length Block:* De-duplication stops your information in repaired size pieces. This weakness on this method will it be neglects to search the repetitive documents to be a littler change. Despite the fact that this technique will be rapidly, direct furthermore negligible CPU cost.
4) *Variable-Length block:* De-duplication stops your information in changing size pieces. The advantage of this strategy will be when essentially any change occurs this limit of the block will be changed without change all through consequent blocks. Utilizing this strategy includes much more CPU menstrual cycles to recognize restrictions furthermore with respect to general record.

B. Dictated by authorization systems you can discover two techniques all through De-duplication

1) *Source/Client organized De-duplication:* The entire De-duplication procedure is finished at source/customer part before conveying your information towards a reinforcement gadget.
2) *Target organized De-duplication:* This De-duplication procedure is finished at backup device.

C. Determined by in the occasion the De-duplication is finished all through back gadget you can discover two methodologies:-

1) *inline De-duplication:* Allows De-duplication soon after securing your information at go down system. Utilizing this system includes less capacity required for go down.

2) *Post-process De-duplication:* Allows De-duplication taking after your got documents will be distributed in hard drive that is, De-duplication will be planned later on. Utilizing this technique includes much more space for hiding away stockpile go down records.

D. Determined by precisely how De-duplication is finished, you can discover two methods

1) *Hash based De- duplication:* With the aforementioned methods that is, record furthermore block level De-duplication is used to perceive regardless of whether two information documents or possibly blocks are typically indistinguishable. Any Hash will be made by using calculations simply like SHA-1, MD5, for information or possibly squares [34].
2) *Content or Application-mindful De-duplication:* The thing parts your information in colossal sections by method for making sense of the data of the things simply like information documents, programing articles, and database objects. Consequently the thing finds these repetitive fragments furthermore stores only this bytes transformed from the a few sections, henceforth alluded to as byte level De-duplication.

II. PROBLEM FORMULATION AND OBJECTIVE

A. Problem Formulation

It makes system a network efficient and storage optimization systems. Today, in the perspective of customer data sharing frameworks the issues for enormous scale, vastly redundant internet data storage is high. Because, of this redundancy stowing price is decreased. Storage for this gradually integrated Net information can be attaining by this one de duplication. The problems with existing data storage system.

If we consider a case in which user update one same file on multiple time, it take space a lot on server memory.

If server have large amount of data than searching technique become slow.

Unwanted space consumption is a very costly when user are in billion.

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 thing in the system so we need accurate fingerprint generator algorithm which finds files fast and accurate and current system having this type of functions but it not proposed 100% accuracy.

B. Objectives

1) To check numerous solutions to storage information on-line.
2) To propose A De-Duplication Method pertaining to Decreasing Ram Ingestion inside cloud Calculating SHA-2 protocol hash code generation.
3) To analyze this proposed approach using the current approach.

III. METHODOLOGY

As the data is getting bigger in size continuously, it is important to have some mechanism to have only legitimate data. Most of the cloud computing companies are using deduplication techniques to save storage space, which requires continuous improvement to the already present algorithms.

To have some improvement over MD5 algorithm used by hash function to generate fingerprints is the main focus of this research proposal.

The problems that are proposed to solve in this research are hash generation performance enhances using MD5.Accuracy to generate unique fingerprints to find duplicate data is increased.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

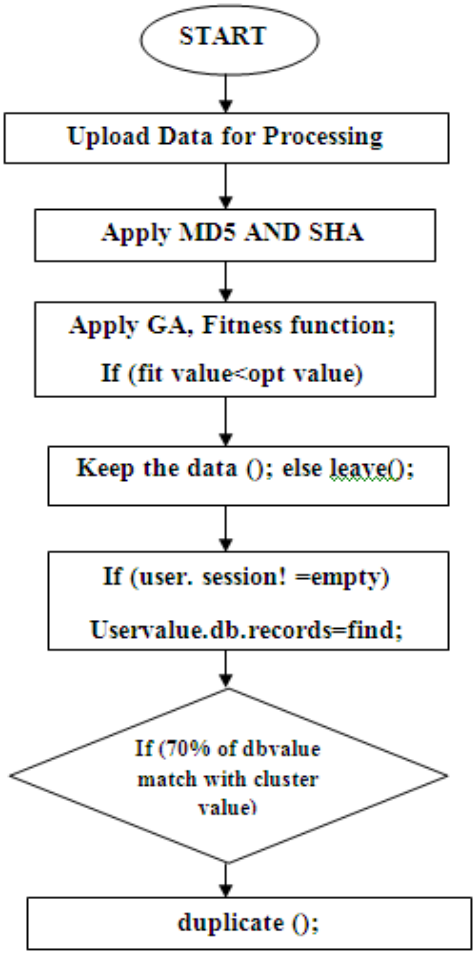


Figure: 1 Flowchart of work

IV. RESULTS

The accuracy is measured in terms of matching errors that consist of false rejection rate and false acceptance rate.

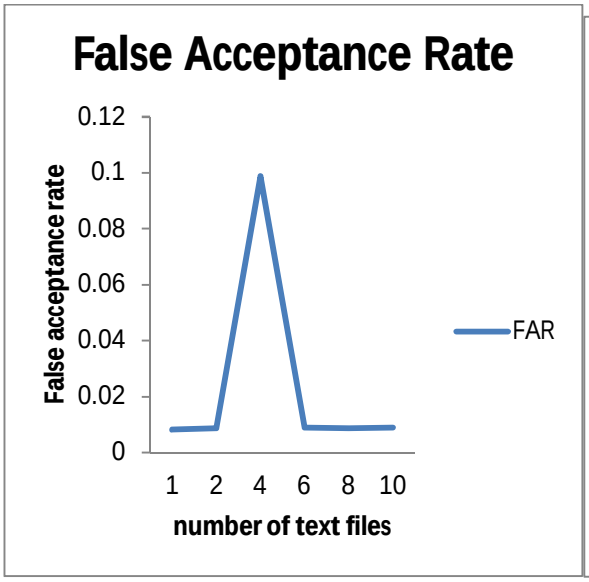


Figure . 2 False Accepatance rate

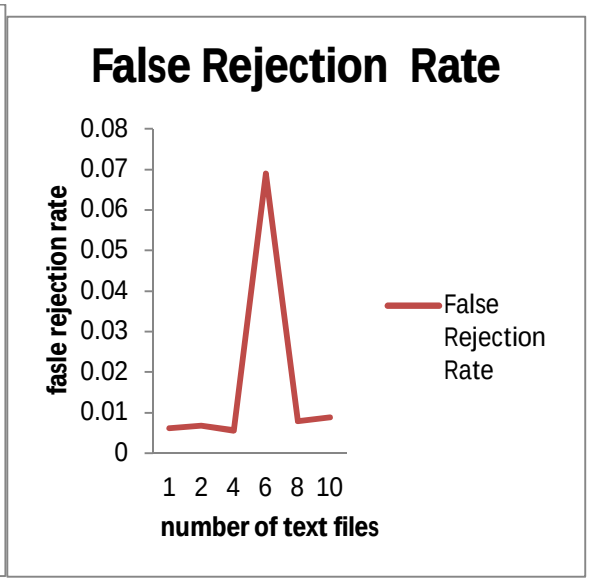


Figure . 3 False Rejection rate

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

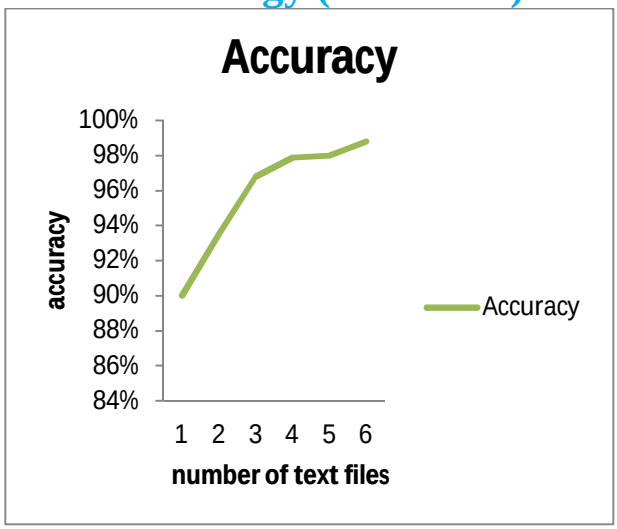


Figure . 4 Accuracy

Far shows that how false acceptance our system works to detect the duplicate files. From this graph we can conclude that duplicator detect the duplicate in less time and perform it better. Frr shows that how false rejection our system works to detect the duplicate files. From this graph we can conclude that duplicator detect the duplicate in less time and perform it better.
Accuracy shows that how accurately our system works to detect the duplicate files. From this graph we can conclude that duplicator detect the duplicate in less time and perform it accurately.

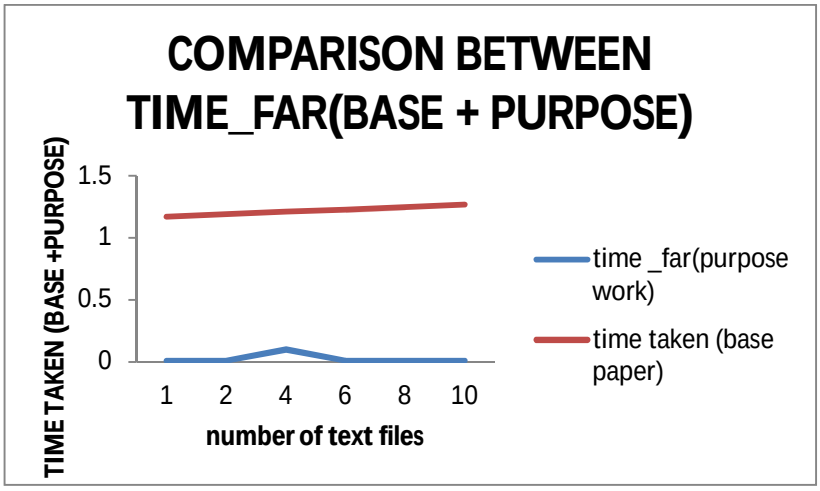


Figure no: 5 Comparison between time _far (base+purpose)

The time_far is that which is used by de-duplicator to detect the duplication of file. In which we calculate the detection time in micro second. The base paper consumes the more time, but my purpose work consume the less time after deduplication.

V. CONCLUSION AND FUTURE WORK

In this paper, we have discussed about storage issues in the cloud computing and shows the de-duplication method for solving the problem of storage at cloud. These techniques are general methods to improve throughput performance of de-duplication storage systems. In our research, MD5 and SHA1 algorithms have been used. MD5 encrypt the original data and decrypt the data. After, decryption, SHA1 is applied that highly secure the data. Accuracy came out to be better to save the time for processing that mostly occur during far_time,frr_time and accuracy.
In our research work, MD-5 and SHA with optimization SHA has been used. MD-5 uploads original data and change to encrypted form so that no else can use it. With SHA , more security can be applied . Optimized SHA optimize more data of SHA.

A. Future Scope

In future, we will work on finding possible optimizations in terms of bandwidth, storage space and computation and as we have worked on texts and audios , so in future we can work on video and pdf files.BFO can be used instead of GA and this research can be used to perform the same De-duplication on real cloud storage using NOSQL Database which contains different kinds of data.

VI. ACKNOWLEDGEMENT

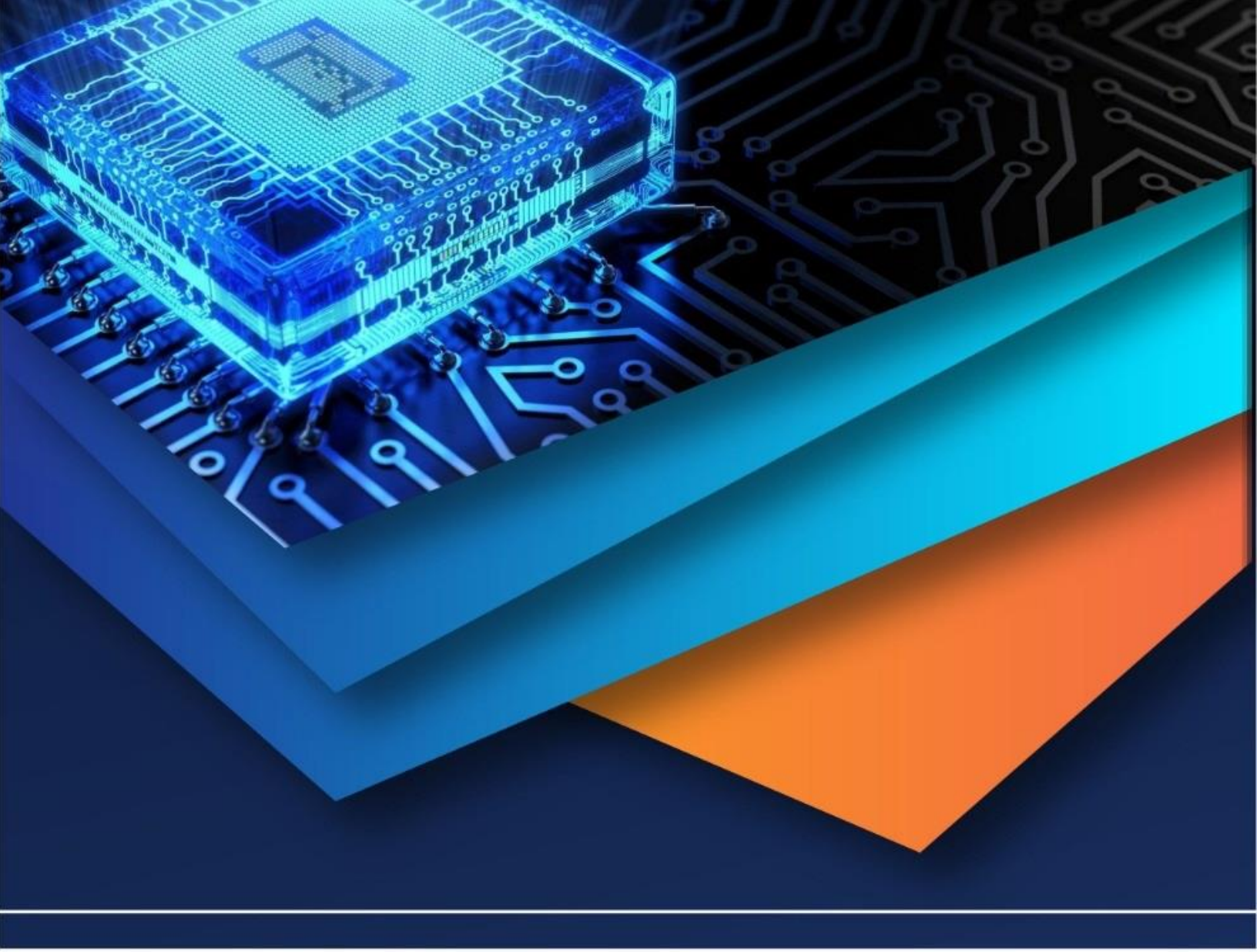
I would like to express deepest gratitude to my adviser Mr. Jasbeer Narwal for his full support,expert guidance, understanding and encouragement throughout my study without whom it would have been impossible to attain success.

REFRENCES

[1] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. (In Technical Report, 2013).
[2] S. Quinlan and S. Dorward. Venti: a new approach to archival storage. In Proc. USENIX FAST, Jan 2002.
[3] C. Ng and P. Lee. Revdedup:A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.
[4] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441–446. ACM, 2012
[5] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 201
[6] PATTERSON, R. H., GIBSON, G. A., GINTING, E., STODOLSKY, D., AND ZELENKA, J. Informed pre fetching and caching. In Proceedings of the Fifteenth ACM Symposium on Operating Systems Principles (Dec. 1995), ACM Press, pp. 79–95
[7] HSU, W. W., SMITH, A. J., AND YOUNG, H. C. The automatic improvement of locality in storage systems. ACM Transactions on Computer Systems 23, 4 (2005), 424–473
[8] B. Zhu, K. Li, and H. Patterson. Avoiding the disk bottleneck in the data domain deduplication file system. In FAST, 2008.
[9] N. Mandagere, P. Zhou, M. A. Smith, and S. Uttamchandani. Demystifying data deduplication. In Companion '08: Proceedings of the ACM/IFIP/USENIX Middleware '08 Conference Companion, pages 12–17, New York, NY, USA, 2008. ACM.
[10] D. Harnik, B. Pinkas, and A. Shulman-Peleg. Side Channels in Cloud Services: Deduplication in Cloud Storage. IEEE Security and Privacy, 8(6):40–47, 2010
[11] Y. Tan, H. Jiang, D. Feng, L. Tian, and Z. Yan. CABdedupe: A causality-based deduplication performance booster for cloud backup services. IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2011.
[12] A . Sabina ,Information Security through Normalization in Cloud Computing, IJSRP, Vol.3, 2011.
[13] Yinjin Fu, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu,AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Service, IEEE International Conference on Cluster Computinges in the Personal Computing Environment (2011).
[14] Zhe SUN, Jun SHEN. DeDu: Building a Deduplication Storage System over Cloud Computing. In Proceedings of the 15th International Conference on Computer Supported Cooperative Work in Design, 2011
[15] Dirk Meister, Jürgen Kaiser,Block Locality Caching for Data Deduplication**.** In 11th USENIX Conference on File and Storage Technologies(FAST). USENIX, February 2013.
[16] A. Wildani, E. L. Miller, and O.Rodeh. HANDS: A heuristically arranged non-backup in-line deduplication system. University of California, Santa Cruz, March 2012
[17] Waraporn Leesakul, Paul Townend, Jie Xu, Dynamic Data Deduplication in Cloud Storage, IEEE, 2013.
[18] Dongfang Zhao, Kan Qiao, Ioan Raic,y ,HyCache+: Towards Scalable High-Performance Caching Middleware for Parallel File Systems, Office of Science of the U.S. Department of Energy under contract DE-AC02-06CH11357, 2014.
[19] Jyoti Malhotra, Novel Way of Deduplication Approach for Cloud Backup Services Using Block Index Caching Technique, IJIRCCE, Vol.3 , 2014.
[20] M.Shyamala ,Enhanced Dynamic Whole File De-Duplication (DWFD) for Space Optimization in Private Cloud Storage Backup, ICAST, Vol.4, 2014
[21] Pasquale Puzio, Melek O¨ nen, Sergio Loureiro, ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage, IEEE, 2013.
[22] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In EUROCRYPT, pages 296–312,2013. [7] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.
[23] M. Bellare, C. Namprempre, and G. Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009.
[24] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing.Workshop on Cryptography and Security in Clouds (WCSC 2011),
[25] J. Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management.(In IEEE Transactions on Parallel and Distributed Systems, 2013).
[26] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.
[27] K. Zhang, X. Zhou, Y. Chen, X.Wang, and Y. Ruan. Sedic: privacyaware data intensive computing on hybrid clouds. In 18th ACM conference on Computer and communications security, CCS"11, pages 515–526, New York, NY, USA, 2011. ACM.
[28] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.
[29] M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller. Secure data deduplication. In Proc. of StorageSS, 2008
[30] Waykole, A Survey Paper on Deduplication by Using Genetic Algorithm Alongwith Hash-Based Algorithm, Vol. 4, Issue 1(Version 1), January 2014, pp.343-346.
[31] Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui.,(2011) ,A secure cloud backup system with assured deletion and version control, (In 3rd International Workshop on Security in Cloud Computing).
[32] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer.(2002), Reclaiming space from duplicate files in a serverless distributed file system,In

Volume 5 Issue V, May 2017
ISSN: 2321-9653

ICDCS.
[33] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg.(2011), Proofs of ownership in remote storage systems. ACM.
[34] Elhadj Benkhelifa , Dayan Fernando.(2013) A Novel cloud hybrid access mechanism for highly sensitive data exchange, The Fourth International Conference on Cloud Computing



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY