RESEARCHERID
THOMSON REUTERS

[Varshini, 5(6): June 2018]
DOI- 10.5281/zenodo.1262287

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

EFFICIENT DATA STORAGE MECHANISM IN CLOUD COMPUTING

Vasam Varshini*[1] & B.Sridhara Murthy[2]

*[1]M.Tech Student, Department of CSE, Kakatiya Institution Of Technology and Science, Warangal District, Telangana, India

[2]Assistant Professor, Department of CSE, Kakatiya Institution Of Technology and Science, Warangal District, Telangana, India

ABSTRACT

As data graduallygrows within data storage areas, the cloud storage systems nonstopface challenges in saving storage capacity and providing capabilities necessary to move big data within an acceptable time frame. In this paper, we propose the Boafft, a cloud storage system with distributed deduplication. The Boafft achieves scalable throughput and capacity using multiple data servers to de- duplicate data in parallel, with a minimal loss of deduplication ratio. Initially, the Boafft uses an effective data routing procedure based on data similarity that decrease the network burdenby quickly identifying the storage location. Secondly, the Boafft maintains an inmemory similarity indexing in each data server that helps avoid a large number of random disk reads and writes, which in turn accelerates local data deduplication. Thirdly, the Boafft build hot fingerprint cache in each data server based on accessoccurrence, so as to improve the data deduplication ratio. Our comparative analysis with EMC's statefull routing algorithm reveals that the Boafft can provide a comparatively high deduplication ratio with a low network bandwidth overhead. Moreover, the Boafft makes best usage of the storage area, with higher read/write bandwidth and good load balance.

Keywords: *Big data, cloud storage, data deduplication, data routing, file system.*

I. INTRODUCTION

As data is at a riotous growth, the redundancy in data is also escalating and with this comes apre- condition for anorderly method to classify and organize such a huge amount of data. Big data is an embryonic term for datasets that are extremely huge which are not capable to manage by conventional techniques. Since data is being outsourced to cloud storage, the effectual management of storage space asks for more attention. Deduplication [1]turns up to be asuitableway out for data detonation in big data epoch by decelerating the data expansion speed by wiping out the redundant data. The olddeduplication methods work on primary storage only. The management of such massive data turns out to be very intricate. Data deduplication is a lossless compression technique that averts the replicated data from being stored into the storage devices. Hasty increase in data is momentous challenge to be conquered. Data deduplication fundamentallyorientationsthe data previouslystored on the disk by a phenomenon that supersedes the indistinguishable data in a file or identical regions of file (similar data).

[Varshini, 5(6): June 2018]
DOI- 10.5281/zenodo.1262287
ISSN 2348 – 8034
Impact Factor- 5.070

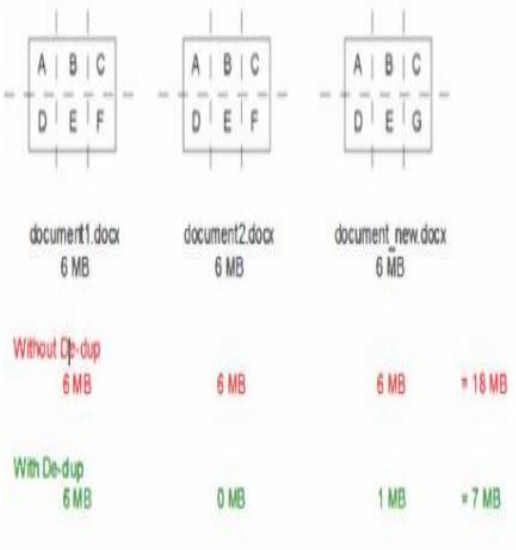
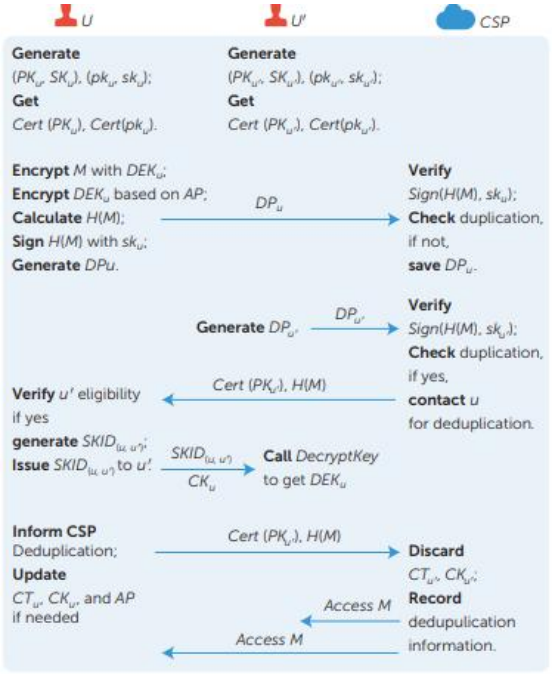


Fig 1: Data de duplication process

Data deduplication [3] technique implicates the categorization of types of file, ripping of file data into lumps, calculation of fingerprints of chunks using MD5 or SHA-1 that helps in categorizing data as unique or identical. The inimitable data is stored into the storage space where as the replicated data is removed from the database and its reference to the original copy is passed using a pointer. The distributed data deduplication arise challenges regarding scalable and data reduction ratio throughout the system nearly to that of centralized system that queries and compares data globally, heading towards the goal of best data deduplication ratio. The main confront of data deduplication methods is to detectrepeated data segments swiftly. Data deduplication [4] could work at file level, which guarantees no duplicate file, or work at block level, which make certain that duplicate data segments contained by a file could be detected andByte level necessitate too much I/O operation. An exemplary data deduplication schemes demands the process flow of chunking being the initial step followed by fingerprinting, indexing and lastly storage management [5-6]. Chunking: The first captious step in the viable proceeding of data deduplication that demands partitioning of file or data stream into inconsequential chunks, in a way each can be duplicate identified. Fingerprinting: reckons the cryptographic secured hash signatures (as SHA1) of data block (chunk), which is a compute radical task nevertheless can be accelerated by numerous parallelizing or pipelining strategies [7-8]. Indexing: designates the procedure diagnostic ate for the identical fingerprints in prodigious storage systems [9]. Storage Management: creditsto storage as well as probable post-deduplication scheduledof unique chunks along with their metadata, counting such courses as German to further compression [10], locating non-contiguous fragment [11], reliability [12], and security [13]. In this paper, the focus is around file level deduplication where the file is divided into the fixed size chunks using MD5 algorithm, the hash code of 128 bit is calculated and compared; and only unique hash valued data contents are saved in the database and the duplicated ones are reported. Indexing will store unique hash values bucket wise. For new data stream hash values will be generated and compared bucket wise by considering left most bit of these hash values, if it exists then the data stream blocks are duplicates, otherwise, it will be stored as a new data if hash values are not matched. The massive data addressed above is proficiently handled by Hadoop [14]. Hadoop is a distributed open source programming framework that is used for processing the large data sets based on its distributed file system (HDFS). Hadoop was developed by Google's Map Reduce, a user defined function. Redundancy removal at primary storage leads to collision and incompetent use of storage space.

[Varshini, 5(6): June 2018]
DOI- 10.5281/zenodo.1262287
ISSN 2348 – 8034
Impact Factor- 5.070



II. RELATED WORK

One vital challenge of today's cloud storage services is that the management of the ever increasing volume of information. to create information management scalable , deduplication has been a widely known technique to scale back cupboard space and transfer information measure in cloud storage. Rather than keeping multiple information copies with an equivalent content, deduplication eliminates redundant information by keeping only 1 physical copy and referring alternative redundant information to it copy. Every such copy will be outlined supported totally different granularities: it should discuss with either an entire file or a additional fine-grained fixed-size or variable-size information block (i.e., block-level deduplication). Today's industrial cloud storage services, similar to Drop box, Mozy, and note pal, are applying deduplication to user information to save lots of maintenance value. in step with the user's read, information outsourcing raises security and privacy issues. we tend to should trust third-party cloud suppliers to properly enforce confidentiality, integrity checking, and access management mechanisms against any corporate executive and outsider attacks. However, deduplication, whereas up storage and information measure potency, is incompatible with ancient cryptography. Specifically, ancient cryptography needs totally different users to write their information with their own keys. Thus, identical information copies of totally different users can result in different cipher texts, creating deduplication not possible. a replacement construction Dekey is planned to produce economical and reliable oblique key management through oblique key deduplication and secret sharing. Dekey supports each file-level and block level de-duplications.

III. PROPOSED SYSTEM

In this paper, we tend to show the way to style secure deduplication systems with higher dependability in cloud computing. we tend to introduce the distributed cloud storage servers into deduplication systems to produce higher fault tolerance. To more shield knowledge confidentiality, the key sharing technique is used, that is additionally compatible with the distributed storage systems. in additional details, a file is initial split and encoded into fragments by victimisation the technique of secret sharing, rather than secret writing mechanisms. These shares are going to be distributed across multiple freelance storage servers. what is more, to support deduplication, a brief scientific discipline hash worth of the content will be computed and sent every storage server because the fingerprint of the fragment hold on at each server.

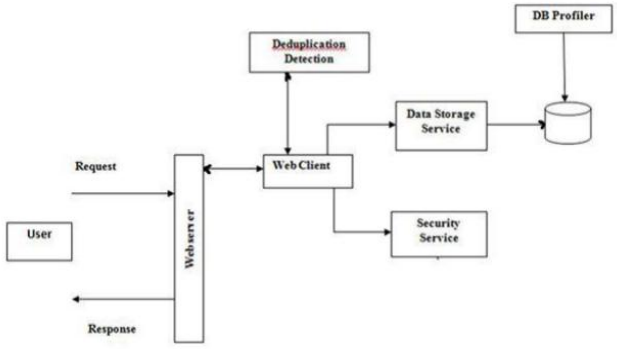
[Varshini, 5(6): June 2018] ISSN 2348 – 8034
DOI- 10.5281/zenodo.1262287 Impact Factor- 5.070

Only the information owner WHO initial uploads the information is needed to reckon and distribute such secret shares, whereas all following users WHO own identical knowledge copy don't have to be compelled to reckon and store these shares to any extent further. To recover knowledge copies, users should access a minimum variety of storage servers through authentication and acquire the key shares to reconstruct the information. In different words, the key shares of knowledge can solely be accessible by the approved users WHO own the corresponding data copy. Four new secure deduplication systems area unit planned to produce economical deduplication with high dependability for file-level and block-level deduplication, severally. the key cacophonic technique, rather than ancient secret writing strategies, is used to guard knowledge confidentiality. Specifically, knowledge area unit split into fragments by victimisation secure secret sharing schemes and hold on at completely different servers

The planned system edges below
Distinguishing feature of our proposal is that knowledge integrity, together with tag consistency, are often achieved.

- To our data, no existing work on secure deduplication will properly address the dependability and tag consistency drawback in distributed storage systems.
- Our planned constructions support each file-level and block-level deduplications.
- Security analysis demonstrates that the planned deduplication systems area unit secure in terms of the definitions laid out in the planned security model. in additional details, confidentiality, dependability and integrity are often achieved in our planned system. 2 forms of collusion attacks area unit thought-about in our solutions. These area unit the collusion attack on the information and also the collusion attack against servers. particularly, the information remains secure even though the somebody controls a restricted variety of storage servers.

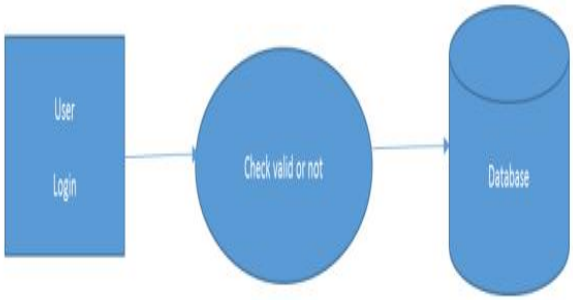
We tend to implement our deduplication systems victimisation the Ramp secret sharing theme that permits high dependability and confidentiality levels. Our analysis results demonstrate that the new planned constructions area unit economical and also the redundancies area unit optimized and comparable the opposite storage system supporting identical level of dependability.



IV. SYSTEM MODULES

User registration
In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first. At the very least, you need toprovide an email address, username, password, display name, and whatever profile fields you have set to required. The display name is what will be used when the system needs to display the proper name of the user

ISSN 2348 – 8034
Impact Factor- 5.070



Server start and upload file

The user can start up the server after cloud environment is opened. Then the user can upload the file to the cloud **SECURE DEDUPLICATION SYSTEM:** To support authorized de duplication the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call is file token instead. To support authorized access a secret key KP will be bounded with a privilege p to generate a file Token. De duplication exploits identical content, while encryption attempts to make all content appear random; the same content encrypted with two different keys results in very different cipher text. Thus, combining the space efficiency of de duplication with the secrecy aspects of encryption is problematic

V. CONCLUSION

Data deduplication may be a technique for eliminating duplicate copies of knowledge, and has been wide employed in cloud storage to scale back cupboard space and transfer information measure. This project tries to formally address the matter of achieving economical and reliable key management in secure deduplication. It introduces a baseline approach during which every user holds associate freelance passkey for encrypting the focused keys and outsourcing them to the cloud. Additionally, the users will revoked from the given cluster at any time. Likewise, session based mostly deduplication is taken into account so it will increase the safety for the outsourced knowledge. The secure deduplication with heterogeneous knowledge storage management technique provides versatile cloud knowledge storage while not duplication and higher access management. To at the same time handle multiple audit sessions from totally different users for his or her outsourced knowledge files, any it's extend into multi-user setting, wherever the TPA will perform multiple auditing tasks during a batch manner for higher potency.

REFERENCES

1. *D.T. Meyer and W.J. Bolosky, "A Study of Practical Deduplication," ACM Trans. Storage, vol. 7, no. 4, 2012, pp. 1–20.*
2. *M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-Locked Encryption and Secure Deduplication," Advances in Cryptology (EUROCRYPT 13), LNCS 7881, 2013, pp. 296–312.*
3. *J. Li et al., "A Hybrid Cloud Approach for Secure Authorized Deduplication," IEEE Trans. Parallel Distributed Systems, vol. 26, no. 5, 2015, pp. 1206–1216.*
4. *Z. Wan, J. Liu, and R.H. Deng, "HASBE: A Hierarchical Attribute-Based Solution for Flexible and Scalable Access Control in Cloud Computing," IEEE Trans. Information Forensics and Security, vol. 7, no. 2, 2012, pp. 743–754.*

5. *Gluster File System [Online]. Available: http://www.gluster.org/ community/documentation/index.php, visited in 2015.*
6. *J. Gantz, and D. Reinsel, "The digital universe decade »are you ready?" IDC White Paper, May 2010[J]. 2011.*
7. *H. Biggar, "Experiencing data de-duplication: Improving effi- ciency and reducing capacity requirements," White Paper, the Enterprise Strategy Group, Feb. 2007[J]. 2012.*
8. *A. Jas, J. Ghosh-Dastidar, M. E. Ng, et al., "An efficient test vector compression scheme using selective Huffman coding[J]," IEEE Trans. Comput.-Aided Des. Integrated Circuits Syst., vol. 22, no. 6, pp. 797–806, Jun. 2003.*
9. *J. H. End III, "Hardware-based," LZW data compression Co-processor: U.S. Patent 6624762[P]. Sep. 9, 2003.*
10. *H. Che, Z. Wang, K. Zheng, et al., "DRES: Dynamic range encoding scheme for tcam coprocessors," IEEE Trans. Comput., vol. 57, no. 7, pp. 902–915, Jul. 2008.*
11. *L. P. Deutsch, "DEFLATE compressed data format specification version 1.3," RFC Editor, 1996*