IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

# A Prototype Of Role Based And Attribute Based De-duplication

**Dinesh Mishra[1*], Sanjeev Patwa[2]**

[1,2] CSE, SOET, Mody University, Rajasthan, India

[*]*Corresponding Author: dmishra1475@gmail.com, Tel.: +91-9406761494*

DOI: https://doi.org/10.26438/ijcse/v7si10.912 | Available online at: www.ijcseonline.org

***Abstract***— Cloud storage services enable individuals and organizations to outsource data storage to remote servers. Cloud storage providers generally adopt data de-duplication, a technique for eliminating redundant data by keeping only a single copy of a file, thus saving a considerable amount of storage and bandwidth. However, an attacker can abuse de-duplication protocols to steal information. At present, there is a considerable increase in the amount of data stored in storage services, along with dramatic evolution of networking techniques. In storage services with huge data, the storage servers may want to reduce the volume of stored data, and the clients may want to monitor the integrity of their data with a low cost, since the cost of the functions related to data storage increase in proportion to the size of the data. To achieve these goals, secure de-duplication and integrity auditing delegation techniques have been studied, which can reduce the volume of data stored in storage by eliminating duplicated copies and permit clients to efficiently verify the integrity of stored files by delegating costly operations to a trusted party, respectively. This paper present study and development of a prototype for role and attribute based de-duplication.

***Keywords***— Data de-duplication, data reduction, de-duplication approaches, role based, attribute based de-duplication

## I. INTRODUCTION

Cloud storage provides customers with highly available, scalable and low cost data storage with elastic provisioning and usage-based pricing [1]. It achieves economy of scale from multi-tenancy in which multiple customers are served from a shared pool of resources; each customer, which could be an enterprise or government, is referred to as a tenant [2]. Indeed, many enterprises are moving to cloud storage services such as Google Drive [3], Dropbox [4] or Mozy [5] rather than using their on-premise storage server. Individual users, usually the employees of a company, are thereby provided with ubiquitous file-sharing and backup services by the outsourcing of data to the cloud storage. The fast growth of data volumes from users raises a challenging issue to minimize costs of storing outsourced data in the cloud storage. Data de-duplication techniques [6] can achieve this goal by allowing a cloud storage service provider (CSP) to eliminate redundant data on the storage. Since this technique saves more than 90% of resources, most CSPs take advantage of it [7][8].

Data de-duplication (henceforth de-duplication), is the process of eliminating copies of repeating data, thus reducing both the intra-file- and inter-file data redundancies [9]. *"By identifying common chunks of data both within and between files and storing them only once, de-duplication can yield cost savings by increasing the utility of a given amount of storage[10]."*

The effectiveness of de-duplication varies widely across the different de-duplication algorithms and different data sets. The governing cooperation of large naval, fishing, oil, etc. Fleets requires information to flow from their management system(s) to their fleet in a robust manner, likewise the fleets have information required by the governing entity. The information necessitates the need to differentiate between different recipients; the captain requires some documents, while crewmembers have differing needs. The research work proposes a framework for role and attribute based de-duplication for cloud content. In this work we will perform de-duplication on the basis of roles assigned to users along with attributes of contents. It will support for heterogeneous type of contents like text, images, etc. Proposed prototype of the system will apply own encryption scheme on the basis of various parameters like roles, attributes and type of data.
Rest of the paper is organized as follows, Section II contains the related work, Section III contain proposed prototype of the research and, section IV concludes research work.

## II. RELATED WORK

There exists a myriad of de-duplication methodologies, which involves invoking several processes to both chunk and restore the files. Irrespective of framework, application or

 

algorithm, de-duplication can be categorized into four major steps [11]:
1. Identifying the unit of comparison
2. Creating smaller unique identifier of these units to be compared
3. Match for duplicates
4. Saving unique data blocks.

Therefore, the de-duplication process itself can be divided into three generic steps as seen in figure [13]:

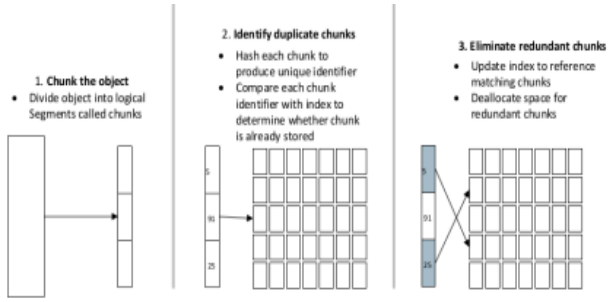


Figure 1: De-duplication process

In step one; a given file is divided into individual chunks of either fixed or variable size. In step two, each chunk is hashed to produce a unique identifier, which we will denote as the *checksum* for that chunk. The checksums are subsequently compared with an index to determine whether that chunk is already stored in the system. In step three, the actual de-duplication takes place i.e. where redundant chunks are eliminated by updating the indexes referencing matching chunks, and de-allocating space/deleting for the redundant ones.
De-duplication can work at file level, block-level and byte-level deletion strategy [12].

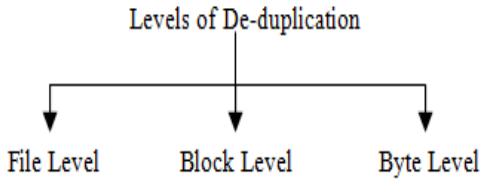


Figure 2: Levels of de-duplication

Comparison of de-duplication levels are shown below:

| File Level | Block Level | Byte Level |
|---|---|---|
| Whole file is considered as a single instance and hash value for whole file is generated. | More fine grain level de-duplication, dividing each file into blocks. | Compares data chunk byte by byte & checks for redundant fragments. |
| Index is very small. | Index size is greater than file level. | Index size is huge. |
| Less computation overhead. | More computation overhead. | More computation overhead. |

Figure 3: Comparison of levels[12]

De-duplication approaches [15][14] are basically divided into three:

- *Location based de-duplication*
- *Time based de-duplication*
- *Chunk based de-duplication.*

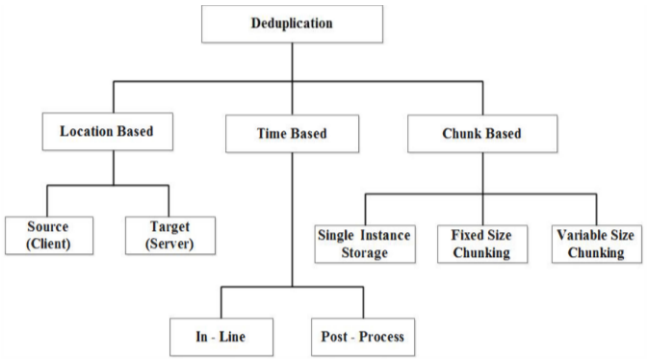


Figure 4: De-duplication approaches [15][14]
Comparison of different approaches are in table below[16]:

Table 1: Approaches [16]

| Approaches | Efficiency | Cost | Throughput | De-duplication ratio |
|---|---|---|---|---|
| Source based | Medium | Low | Medium | Medium |
| Target based | Medium | High | Medium | Medium |
| In line | Medium | Low | Low | Low |
| Post process | High | High | Medium | High |

## III. PROPOSD PROTOTYPE

De-duplication is a process often specialized to solve specific *Service Level Objectives*, therefore based on the objective of the application and its needs; the methodology of the frame-works varies in terms of how they achieve the de-duplication. The methods for de-duplication are a trade-off between i/o usage, processing time and storage needs. The research work proposes a framework for role and attribute based de-duplication for cloud content. In this work we will perform de-duplication on the basis of roles assigned to users along with attributes of contents. It will support for different type of contents like text, images, etc. Proposed system will apply encryption scheme on the basis of various parameters like roles, attributes and type of data.
Research objectives are:

- to improve de-duplication ratio
- to improve storage efficiency
- to reduce communication and computation overhead
- to perform de-duplication on heterogeneous content like multimedia contents
- perform Attribute based de-duplication
- perform Role based secure de-duplication

File naming encryption with fast indexing. Proposed system is shown below:

 

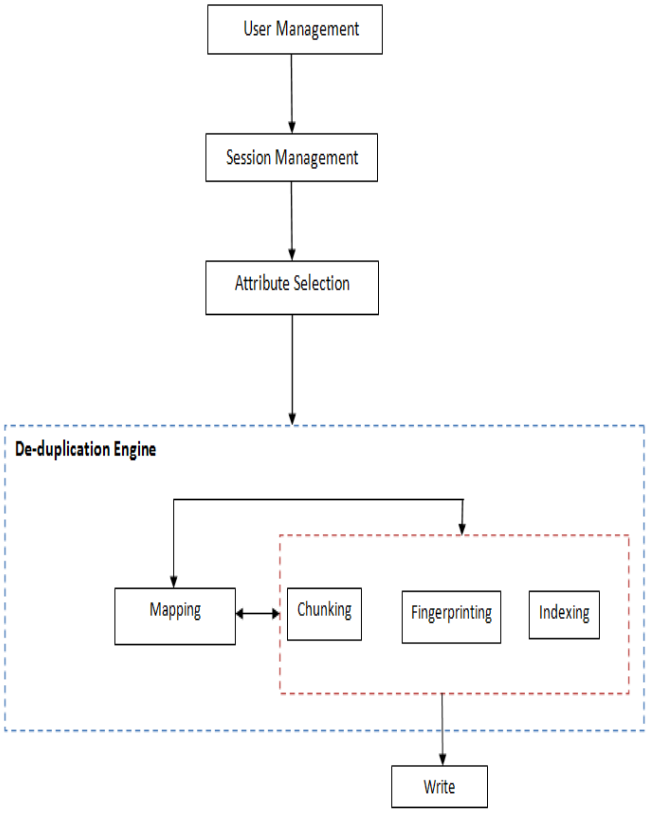


Figure 5: Proposed system

**User management** is responsible for creation of users, assigning roles and privileges to users . It also includes unique token generation for user.

**Session management** will perform creation of session and managing all session variables to make system integrity.

**Attribute selection** selects attributes like size of file, type of file, user type etc. On the basis of this selection a particular type of de-duplication method will be selected. Fixed size chunking will be used for text content with smaller size and variable size chunking will be used for larger size text file and other content types also.

**De-duplication engine** is responsible for removing redundancy by providing storage space optimization, less overhead and security. It also check redundant contents according to roles assigned to user. It involves mapping, chunking, fingerprinting and indexing operations. System will generate hash using SHA2 method for every chunk. That hash will be mapped with user token and stored on metadata. For security issues system will perform naming encryption method for content. It will reduce searching overhead while checking for duplicate data. Detailed working of de-duplication engine is shown below:

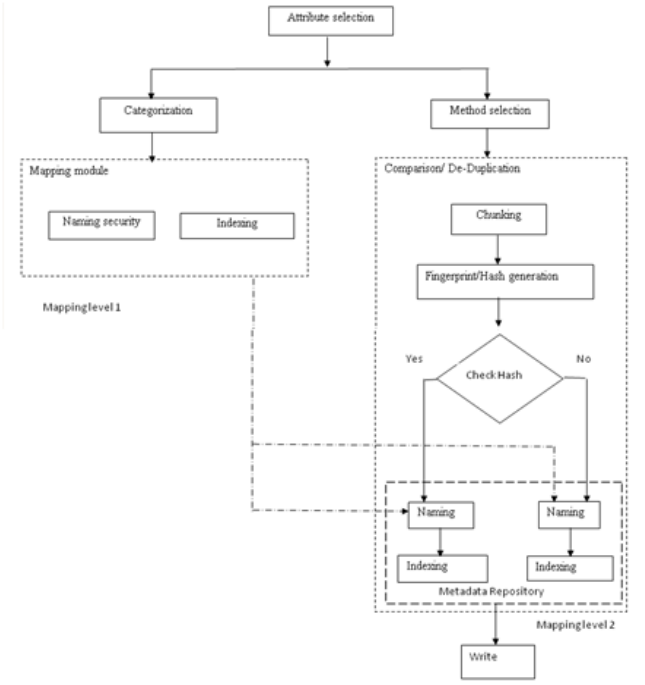


Figure 6: Detailed system

## IV. CONCLUSION AND FUTURE SCOPE

The main design features of our proposed enterprise model can be summarized as follows:

- Enterprise Level Data De-duplication
- Secure Indexing Scheme
- Multi-user Searchable Encryption for Data De-duplication
- Secure and Efficient File Sharing viewpoint

System is under development, so this paper does not contain result and evaluation.

## REFERENCES

[1] F. Durao, J. F. S. Carvalho, A. Fonseka, and V. C. Garcia, "*A systematic review on cloud computing*," The Journal of Supercomputing, vol. 68, no. 3, pp. 1321–1346, 2014.

[2] P. Mell and T. Grance, "*The NIST definition of cloud computing*," National Institute of Standards and Technology, Information Technology Laboratory, Tech. Rep., 2009.

[3] "*Google Drive*," 2017. [Online]. Available: https://www.google.com/drive/

[4] "*DropBox, a file-storage and sharing service*." [Online]. Available: http://www.dropbox.com

[5] "*Mozy, cloud backup solutions*." [Online]. Available: https://www.mozy.com

[6] D. T. Meyer and W. J. Bolosky, "*A study of practical deduplication*,", ACM Transactions on Storage, vol. 7, no. 4, pp. 1–20, 2012.

[7] N. Mandagere, P. Zhou, M. A. Smith, and S. Uttamchandani, "*Demystifying data deduplication*," in Proceedings of the ACM/IFIP/USENIX Middleware '08 Conference Companion (Companion'08), 2008, pp. 12–17.

 

[8] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side *Channels in Cloud Services: Deduplication in Cloud Storage*," IEEE Security & Privacy Magazine, vol. 8, no. 6, pp. 40–47, 2010.
[9] Nagapramod Mandagere, Pin Zhou, Mark A Smith, and Sandeep Uttamchandani. *"Demystifying data deduplication*". In *Proceedings of the ACM/IFIP/USENIX Middleware'08 Conference Companion*, pages 12–17. ACM,2008.
[10] Mark W Storer, Kevin Greenan, Darrell DE Long, and Ethan L Miller. "*Secure data deduplication*". In Proceedings of the 4th ACM international workshop on Storage security and survivability,pages 1–10. ACM, 2008.
[11]Ravindra Mahabaleshwar. *"Effective data deduplication implementation"*, Whitepaper 2011.
[12]Qinlu He, Zhanhuai Li, Xiao Zhang, "*Data De-duplication Techniques*", International Conference on Future Information Technology and Management Engineering, IEEE 2010
[13] Dave Cannon. Data deduplication and tivoli storage manager. *Tivoli Storage, IBM Software Group (September 2007)*, 2009
[14] AndrejTolic, AndrejBrodnik, "*Deduplication in unstructured-data storae systems*", ELEKTROTEHNISKI VESTNIK 82(5): 233–242, 2015
[15]E. Manogar, S. Abirami, "*A Study on Data Deduplication Techniques for Optimized Storage*", 2014 Sixth International Conference on Advanced Computing(lCoAC) IEEE
[16] Vruti Satish Radia, Dheeraj Kumar Singh, *" Secure deduplication Techniques: A Study***",** International Journal of Computer Applications (0975 – 8887) Volume 137 – No.8, March 2016

**Authors Profile**

*Mr. Dinesh Mishra* pursed Bachelor of Technology and Master of Technology in Coputer Science & Engineering from RGPV, Bhopal. Currently pursuing Ph. D. from Mody University, Rajasthan. He has having teaching experience of 12 years. His field of intrest is cloud computing.

*Dr. Sanjeev Patwa* is currently working as assistant profesor and Ph. D. coordinator in CSE department, SOET, Mody University, Rajasthan. He has published more than 20 research papers in reputed international journals including and conferences including IEEE and it's also available online.

 