Attribute Based De-Duplication Method

Dinesh Mishra, Dr. Sanjeev Patwa

Abstract: In the current scenario huge amount of data will be generated from different sources for storage. Cloud is a cheapest source for such storage. Data may be of heterogeneous type and with different attributes. Duplicate data produces a problem for cloud service provider. Deduplication is a technique to provide optimal storage solution. Main challenges in deduplication are optimizing storage with reliability and accuracy along with reducing computation overhead and recoverability. The proposed research describes a framework to provide attribute based deduplication on cloud contents. Developed framework will provide attribute based deduplication on heterogeneous contents. System has been evaluated on various parameters like deduplication ratio, computation overhead etc.

Index Terms: Attribute based deduplication, Content based deduplication, Deduplication, engine, Fast indexing , Heterogeneous contents, Storage optimization, Data compression .

1. INTRODUCTION

DATA generation rate is growing explosive today. Huge amount of digital data is generated from different applications. Best solution for storing such data is cloud. Cloud service provider is responsible for providing storage of data which can be accessed and provided by different users from all over the world[3,4]. Today in IT budgets, huge money being invested on storage capacity. Data growth rate is explosive, that is mentioned in IDC's Digital Universe study [6]. This huge amount of data produces more problems, like performance degradation, quality degradation and increase in operational costs. Deduplication is derived as solution to overcome such problems. Cloud storage is a major source that provides customers with availability, scalability and low cost data storage with elasticity and pay as per use based pricing [1]. 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. 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.

There are various problems in cloud data storage offered by Cloud Service Providers (CSPs). Firstly How to make cloud data access control adapt to various scenarios and satisfy different user demands becomes a practically important issue. Second, flexible cloud data de-duplication with data access control is still an open issue. Duplicated data could be stored at the cloud [2][1] in an encrypted form by the same or different users, in the same or different CSPs. From the standpoint of compatibility, it is highly expected that data de-duplication can cooperate well with data access control. Solution to these problems can be data deduplication technique [5]. Data de-duplication stores only one unique instance of the data type on the disk or tape. In this method redundant data is replaced with a pointer to the unique data copy. This reduces the hardware used to store data and the bandwidth costs required for transmitting and receiving purposes. De-duplication belongs to intelligent data compression technique for redundant data reduction [7]. The paper proposes a holistic and heterogeneous data storage management scheme in order to solve the above problems. The proposed scheme is compatible with the access control scheme. It further realizes flexible cloud storage management with both data de-duplication and access control that can be operated by either the data owner or a trusted third party or both or none of them.

2 BACKGROUND

De-duplication process mainly has four stages that is chunking, Fingerprinting, Indexing and Writing [7]. Method flow is shown in fig 1.

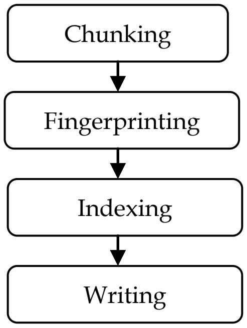


***Fig. 1**. Deduplication process.*

The steps in de-duplication process are explained below: Chunking: It is division of input data stream into chunks. Fingerprinting: In this hash values for each chunk are generated using hashing methods like SHA-1, MD5. These hash values are called as fingerprints of chunks. Indexing: In this comparison of previously stored hash values with newly generated hash values for finding duplicated data chunks will be performed. Writing: Storing the unique copy of data into storage media. Data de-duplication techniques based on data are as follows:

- *Dinesh Mishra is currently pursuing Ph. D. in department of Computer Science & Engineering from Mody University, Rajasthan India Phone +919406761494 E-mail: dmishra1475@gmail.com*
- *Dr. Sanjeev Patwa is currently Assistant Professor in department of Computer Science &Eengineering, Mody University, Rajasthan India E-mail: sanjeevpatwa.cet@modyuniversity.ac.in*

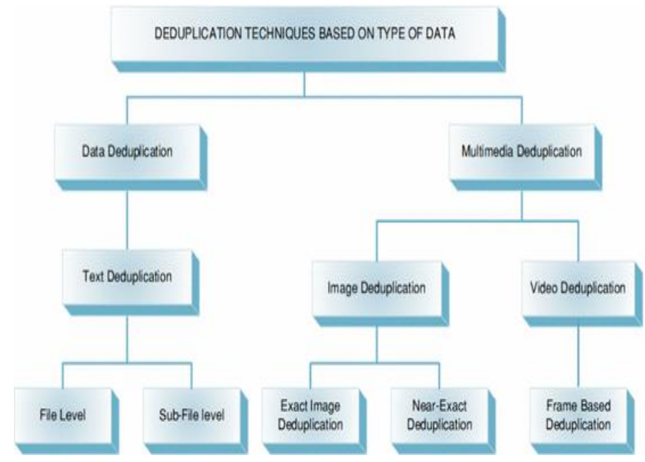


***Fig. 2**.Content based deduplication techniques*

Data de-duplication is done by the following methods [9][11] as shown in Fig. 3.

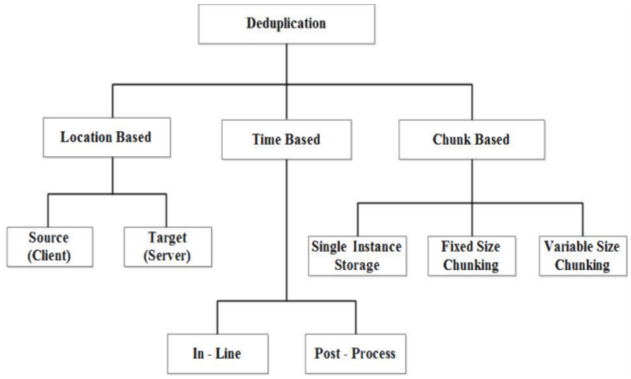


***Fig. 3**. Deduplication approaches.*

The entire de-duplication is carried out either at the source side (Client) or target side (Server). Data can be processed at three places, before being written into a disk (Inline) or after writing to the disk (Post), or both before and after written to the disk (Hybrid) [5]. An Inline de-duplication can be done the client side or when the data is transferring from the client/source to the server. Different chunk based de-duplication strategies are available in the literature to remove the redundant data which is present in the disk or backup system as discussed below:

- Single Instance Storage or Whole file chunking
- Fixed Size Chunking
- Variable size Chunking

Here, chunking boundaries are determined based on the contents of the file, so it is more resistant to the insertion and deletion.

2.1 Related Work

3 PROPOSED SYSTEM

The research work proposes a framework for attribute based de-duplication for cloud content. In this work we will perform de-duplication on the basis of attributes of contents. It will support for different type of contents like text, images, etc.

TABLE 1

COMPARISON OF RELATED WORK

Paper	Methodology	Advantage	Issues
[13]	Secure and efficient encrypted EMRs deduplication scheme for cloud-assisted eHealth systems, namely Health Dep. It uses MLE Keys.	Able to resist brute force attacks. Strong Security. Reduce the storage costs by more than 50% in small samples.	Communication overhead increases with number of key servers.
[14]	Performance oriented I/O de-duplication	Improves the performance and saves capacity of primary storage systems.	Security of Data
[15]	Request based selective deduplication protocol.	Reduces the bursty traffic on network & provide security by encryption.	User feedback and choice based method. Work on reducing the bursty traffic only.
[16]	It integrates cloud data de-duplication with access control.	De-duplication based on ownership, Security improved	It Maps the user with their encrypted hash value.
[17]	Scheme uses variable-size block-level deduplication based on the technique of Rabin finger printing.	Secure against offline brute-force, dictionary attacks & reduces computation overheads	Use of trusted third party server
[18]	A lazy method dedupswift is introduced to reduce the disk I/O bottleneck.	DedupeSwift is able to save 65.24% and 89.84% space overhead	The write and read performance is not so high.

Proposed system contains following components:

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.

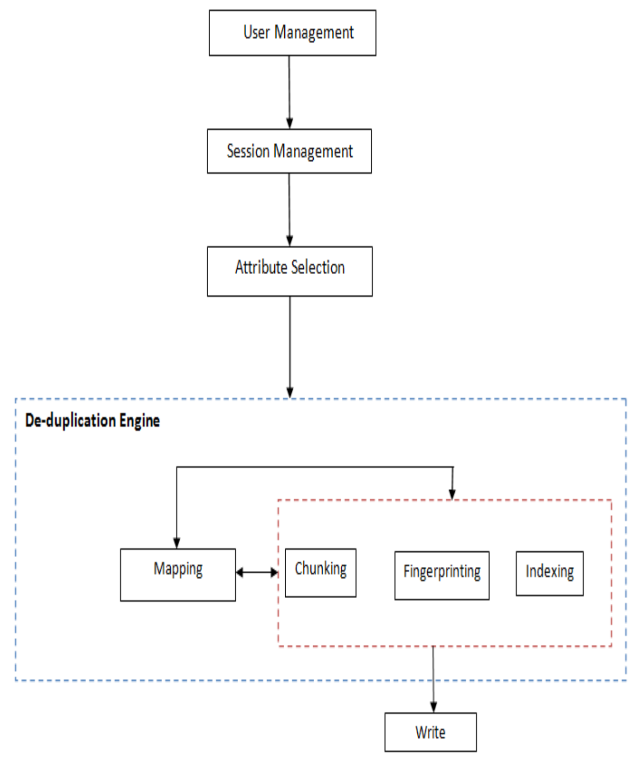


Fig. 4. *Proposed* System

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. Proposed system will check for attribute of uploaded file first. If file is of text type then system uses fixed size chunking method while if file type is image than system reserves memory size for chunk. This may vary according to file size. In chunking, if new chunk has been found than hashing will be performed, it will be uniquely stored in database and indexing has been performed otherwise only indexing has been done. For retrieving file, reading index for that file and according to that index all chunks has been merged sequentially to make complete file ready to download. Decoding of contents is also performed in the process of retrieving of a file.

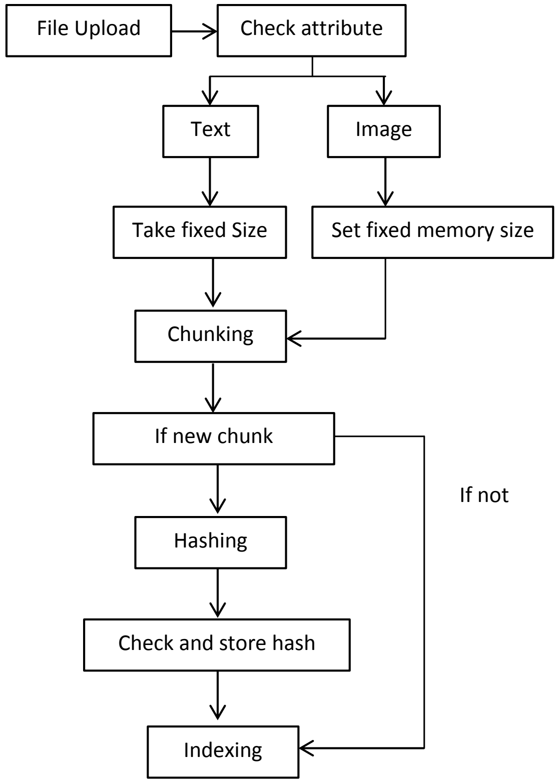


Fig. 5. Attribute *based* methodology

4 EXPERIMENTAL SETUP

Proposed framework is developed on real time cloud. Cloud type is infra as a cloud for this framework. System is developed with python. Heroku cloud has been used for implementation. It is shared cluster based cloud. Operating system for development is Linux. SQLite database has been used.

5 RESULT AND DISCUSSION

Snapshot of main system is shown in figure below:

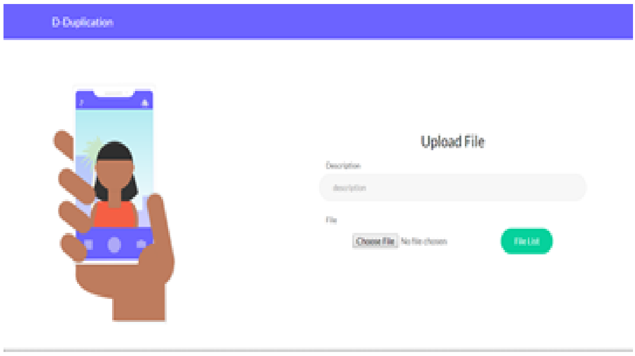
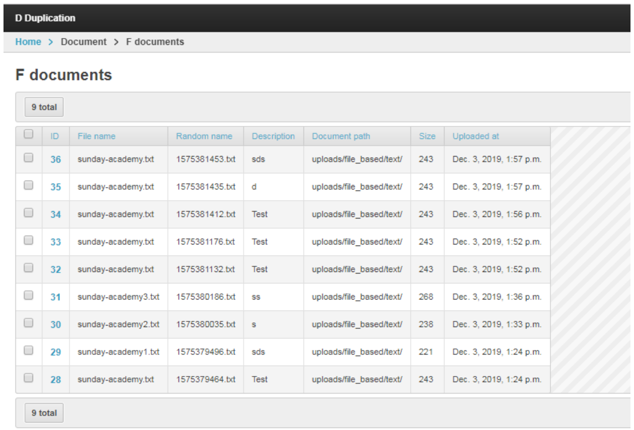


Fig. 6. *User interface*



D Duplication

Home > Document > F documents

F documents

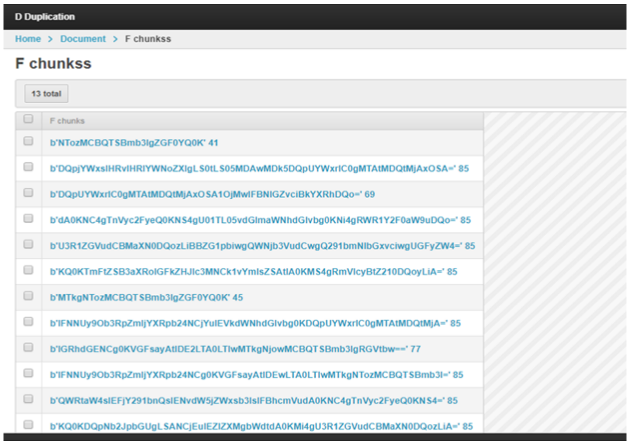
9 total

□	ID	File name	Random name	Description	Document path	Size	Uploaded at
□	36	sunday-academy.txt	1575381453.txt	sds	uploads/file_based/text/	243	Dec. 3, 2019, 1:57 p.m.
□	35	sunday-academy.txt	1575381435.txt	d	uploads/file_based/text/	243	Dec. 3, 2019, 1:57 p.m.
□	34	sunday-academy.txt	1575381412.txt	Test	uploads/file_based/text/	243	Dec. 3, 2019, 1:56 p.m.
□	33	sunday-academy.txt	1575381176.txt	Test	uploads/file_based/text/	243	Dec. 3, 2019, 1:52 p.m.
□	32	sunday-academy.txt	1575381132.txt	Test	uploads/file_based/text/	243	Dec. 3, 2019, 1:52 p.m.
□	31	sunday-academy3.txt	1575380186.txt	ss	uploads/file_based/text/	268	Dec. 3, 2019, 1:36 p.m.
□	30	sunday-academy2.txt	1575380035.txt	s	uploads/file_based/text/	238	Dec. 3, 2019, 1:33 p.m.
□	29	sunday-academy1.txt	1575379496.txt	sds	uploads/file_based/text/	221	Dec. 3, 2019, 1:24 p.m.
□	28	sunday-academy.txt	1575379464.txt	Test	uploads/file_based/text/	243	Dec. 3, 2019, 1:24 p.m.

9 total

Fig. 7. *Stored file record*

Stored chunk details are shown below:



D Duplication

Home > Document > F chunkss

F chunkss

13 total

□	F chunks
□	b'NTozMCBQTSBmb3IgZGF0YQ0K' 41
□	b'DQpjYWxsIHRvIHRlYWNoZXIgLS0tLS05MDAwMDk5DQpUYWxrIC0gMTAtMDQtMjAxOSA=' 85
□	b'DQpUYWxrIC0gMTAtMDQtMjAxOSA1OjMwIFBNIGZvciBkYXRhDQo=' 69
□	b'dA0KNC4gTnVyc2FyeQ0KNS4gU01TL05vdGlmaWNhdGlvbg0KNi4gRWR1Y2F0aW9uDQo=' 85
□	b'U3R1ZGVudCBMaXN0DQozLiBBZG1pbiwgQWNjb3VudCwgQ291bmNlbGxvciwgUGFyZW4=' 85
□	b'KQ0KTmFtZSB3aXRoIGFkZHJlc3MNCk1vYmlsZSAtIA0KMS4gRmVlcyBtZ210DQoyLiA=' 85
□	b'MTkgNTozMCBQTSBmb3IgZGF0YQ0K' 45
□	b'IFNNUy9Ob3RpZmljYXRpb24NCjYuIEVkdWNhdGlvbg0KDQpUYWxrIC0gMTAtMDQtMjA=' 85
□	b'IGRhdGENCg0KVGFsayAtIDE2LTA0LTIwMTkgNjowMCBQTSBmb3IgRGVtbw==' 77
□	b'IFNNUy9Ob3RpZmljYXRpb24NCg0KVGFsayAtIDEwLTA0LTIwMTkgNTozMCBQTSBmb3I=' 85
□	b'QWRtaW4sIEFjY291bnQsIENvdW5jZWxsb3IsIFBhcmVudA0KNC4gTnVyc2FyeQ0KNS4=' 85
□	b'KQ0KDQpNb2JpbGUgLSANCjEuIEZlZXMgbWdtdA0KMi4gU3R1ZGVudCBMaXN0DQozLiA=' 85

Fig. 8. *Chunk records*

Performance is evaluated on the basis of deduplication ratio. Proposed system has been evaluated by taken different scenarios of same file under consideration and improvement has been shown.

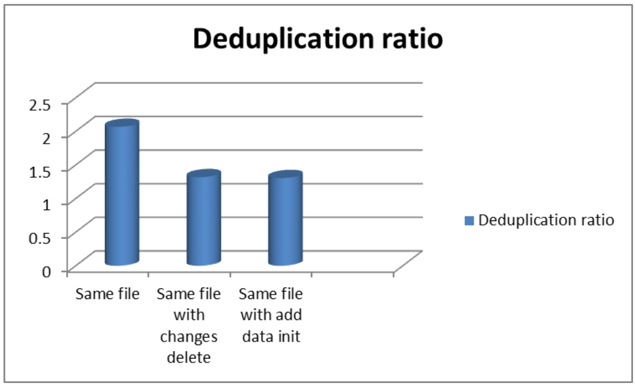


Fig. 9. *Deduplication ratio in different file processing*

ile and after editing on it, proposed system also reduces chunking time as shown in fig. 7.

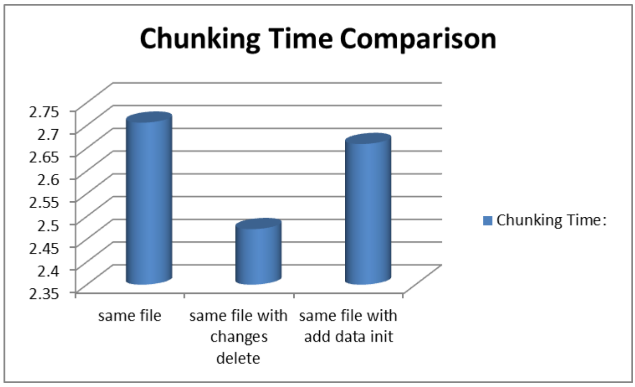


Fig. 7. *Chunking time in different file processing*

Chart shown below represent hashing time comparison for file during deduplication and found improved.

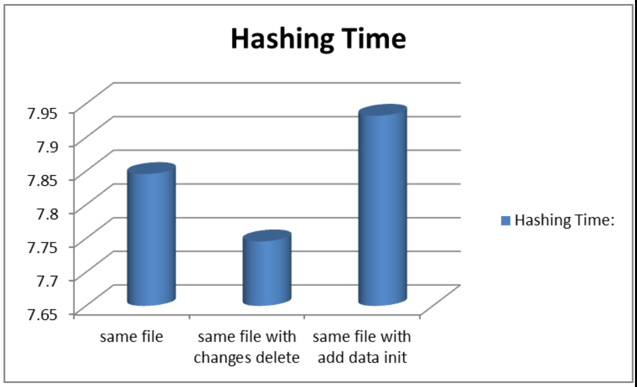


Fig. 8. *Hashing time in different file processing*

6 CONCLUSION

Data de-duplication is a scalable and efficient redundant data reduction technique for large-scale storage systems, which addresses the challenges imposed by the explosive growth in demand for data storage capacity. Proposed system will be checked for text and image files. After implementing it we found that proposed system performs better in hashing time, chunking time and deduplication ratio. Further work can also be done with taking consideration of roles of different users also. There are other interesting issues in de-duplication, such as de-duplication ratio estimation, file recipe compression, and video/image de-duplication, etc., that are worthy of future research and development attention.

REFERENCES

[1] Z. Yan, X. Y. Li, M. J. Wang, A.V. Vasilakos, “Flexible data access control based on trust and reputation in cloud computing,” IEEE Trans. Cloud Comput., 2015. Doi: 10.1109/TCC.2015.2469662.

[2] Q. Liu, G. J. Wang, and J. Wu, "Consistency as a service: Auditing cloud consistency" IEEE Trans. Netw. Serv. Manage., vol.11, no.1, pp. 25-35,

2014.

[3] Q. Duan, “Cloud service performance evaluation: Status, challenges, and opportunities – A survey from the system modeling perspective”, Digital Commun. Netw., Available online 23 December 2016, ISSN 2352-8648, http://dx.doi.org/10.1016/j.dcan.2016.12.002.

[4] Q. Liu, C. C. Tan, J. Wu, and G. J. Wang, "Towards differential query services in cost-efficient clouds," IEEE

[5] AndrejTolic, AndrejBrodnik, “Deduplication in unstructured-data storage systems”, ELEKTROTEHNISKI VESTNIK 82(5): 233–242, 2015

[6] C.Policroniades and I.Pratt, “Alternatives for detecting redundancy in storage systems data”, in Proceedingsof the General Track: 2004 USENIX Annual Technical Conference, 2004, pp. 73-86.

[7] Yukun Zhou, Dan Feng, Wen Xia, Min Fu, Fangting Huang, Yucheng Zhang, Chunguang Li,“SecDep: A User-Aware Efficient Fine-Grained Secure Deduplication Scheme with Multi-Level Key Management”, IEEE Mass Storage Systems and Technologies (MSST) 2015 31st Symposium, Year - 2013

[8] Sarah Prithvika P.C. , Ramani S. , Jakkulin Joshi J. and Sindhu K, “Data Deduplication in Cloud Environment – A Survey”, International Journal of Latest Engineering and Management Research (IJLEMR) ISSN: 2455-4847 Volume 03 - Issue 01January 2018,PP. 44-49

[9] Qinlu He, Zhanhuai Li, Xiao Zhang, “Data De-duplication Techniques”, International Conference on Future Information Technology and Management Engineering, IEEE 2010

[10] Ravneet Kaur,Inderveer Chana, Jhilik Bhattacharya, “ Data deduplication techniques for efficient cloud storage management: a systematic review”, Journal of Supercomputing, Springer, May 2018, Volume 74, Issue 5, pp 2035–2085

[11] E. Manogar, S. Abirami, “A Study on Data Deduplication Techniques for Optimized Storage”, 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE

[12] Wen Xia, Hong Jiang, Dan Fen, “A Comprehensive Study of the Past, Present, and Future of Data De-duplication”, Vol. 104, No. 9, IEEE 2016

[13] Yuan Zhang et al, “HealthDep: An Efficient and Secure Deduplication Scheme for Cloud-Assisted eHealth Systems”, Transactions on Industrial Informatics, IEEE 2018

[14] Bo Mao et al, “Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud” IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 6, JUNE 2016

[15] Nishant N.Pachpor et al, “Improving the Performance of System in Cloud by Using Selective Deduplication”, ICECA, IEEE 2018

[16] Vidhya R et al, “Elimination of Redundant Data in Cloud with Secured Access Control”, ICTACC.2017 IEEE

[17] Haonan Su et al, “An Efficient and Secure Deduplication Scheme Based on Rabin Fingerprinting in Cloud Storage”, CSE-EUC.2017, IEEE

[18] Jingwei Ma et al, “DedupeSwift: Object-oriented Storage System based on Data Deduplication”, TrustCom, 2016 IEEE