An Efficient Framework and Techniques of Data Deduplication in Data Center

Bhashmitha J
Department of Computer Science and Engineering
Srinivas School of Engineering,
Mangalore, India

Chaithra P K
Department of Computer Science and Engineering
Srinivas School of Engineering,
Mangalore, India

Laxmi Manohari
Department of Computer Science and Engineering
Srinivas School of Engineering,
Mangalore, India

Anusha
Department of Computer Science and Engineering
Srinivas School of Engineering,
Mangalore, India

Santhosh S
Asst. Professor
Department of Computer Science and Engineering
Srinivas School of Engineering
Mangalore, India

Abstract:- Here we use deduplication techniques in order to store the files in data center. In deduplication we check a particular file is duplicate or not. If that file is not duplicated then only auditor allow file to store on data center by using different method based file storage. In our project we face challenges on encrypted data storage and management with deduplication. The access control mechanism is also included in this section. Here overall checking of deduplication with heterogeneous data storage management process across multiple cloud service providers with respect to analyzing performance of data center.

Keywords*:- Data deduplication,data center,Deduplication framework.*

I. INTRODUCTION

Data deduplication works by eliminating the redundant copies of same data. Data deduplication stores only one of unique copies. Data center allows storage of data and access to the computer services . So this data deduplication technique is used to eliminate duplicatecopies of data . Data storage service is one of the most widely consumed cloud services. Deduplication technique in data center is commonly used to reduce the space and bandwidth requirements of the services. The benefits of data center are: cost efficient, provides security, disaster recovery.

II. LITERATURE SURVEY

C. Yang, J. Ren, and J. F. Ma,There is an issue that many redundant copies of files are stored in the remote storage servers, which is waste of bandwidth and increases cost. Solution to this client-side deduplication was introduced to avoid uploading same files which already exists.

T.-Y.Wu,J.-S. Pan, and C.-F. Lin,we can access any file which is stored in cloud from any place if we have strong internet connection. Cloud storage provides offsite backups of data. This helps in reduction of cost. Cloud storage has greater accessibility and reliability and lower overall storage costs.

J. W. Yuan, and S. C. Yu,In this paper, we come across two main issues that is privacy and efficiency. Here we make use of a private keyword-based file retrieval scheme where user can get the files from untrusted server without any leakage of information.

Z. Yan, W. X. Ding, X. X. Yu, H. Q. Zhu , and R. H. Deng, . Existing solutions of encrypted data deduplication suffer from weak security issues. Here we use proxy encryption which integrates cloud data deduplication with access control. Here we also reduplicate the encrypted data.

J. Li, X. F. Chen, M. Q. Li, J. W. Li, P. P. C. Lee, To reduce space and bandwidth deduplication technique was introduced which can eliminate the redundant copies. But here we face some problems related to security and ownership. Many schemes have been introduced to sort these issues which involves sharing of same encryption key for the same data for each and every user.

M. Wen, K. Ota, H. Li, J. S. Lei, C. H. GU; and Z. Su, This paper is about secure deduplication. The most challenging issue of secure deduplication is how to manage data and convergent key when user updates it again and again. This problem is solved by using a key method called bilinear pairing. The implementation of this method costs more.

III. METHODOLOGY

A. Data Deduplication

Data Deduplication works by eliminating redundant copies of same data and has been widely used to reduce storage space and bandwidth. Hash value is calculated for the uploaded data. It is then compared with the existing hash value. Data is not uploaded if there are any duplicate values. It is replaced with the pointer to the particular data. If there are no any duplicate copies then the data is uploaded to the server.

B. How does deduplication works?

Here the objects are compared in the form of files or blocks and it removes the copies which are already existed. Non unique processes are removed in this method. Data deduplication mainly works by dividing the input data into blocks and files. For these files and blocks hash values are calculated. By calculating the hash value we can easily determine whether another block of same data has been stored or not. If the same copies are found then replacement is done for that duplicate data with reference to the object already present.

C. Hash based algorithm

Certain algorithms are used to determine the chunks of data for the created hash the data is identified as a duplicate and is not stored in the database. The algorithm used here is Secure Hash Algorithm 1[SHA-1].

- *SHA-1*

Cryptographic signatures are created for security purpose. The 160-bit value created by SHA-1 is different for each set of data. Here it breaks the data in to chunks. These chunks are either fixed or variable in length. This processes the chunk with the hashing algorithm to create a hash value. If the hash value already existed then the data is considered as duplicate and is not stored. If the hash value doesn't exist then the data is stored.

D. Advanced Encryption Standard Algorithm

AES is a cryptographic algorithm which is mainly used to protect the electronic data. AES is symmetric block cipher which encrypts and decrypts the data. This algorithm is capable of using the keys of 128,192 and 256 bits to encipher and decipher the data in the blocks of 128 bits.

E. Data Encryption Standard

DES is a cryptographic algorithm applied to a block of data simultaneously rather than 1 bit at a time. It will encrypt the plaintext message by grouping it into 64 bit blocks. Then these blocks are enciphered using the secret key into a 64 bit cipher text. Decryption is nothing but inverse of encryption.

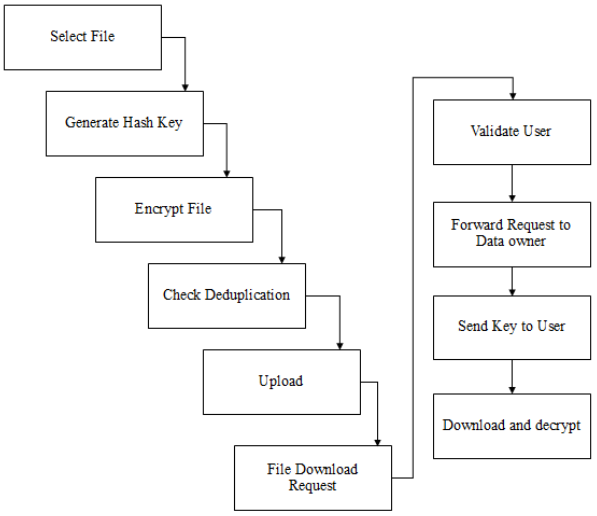
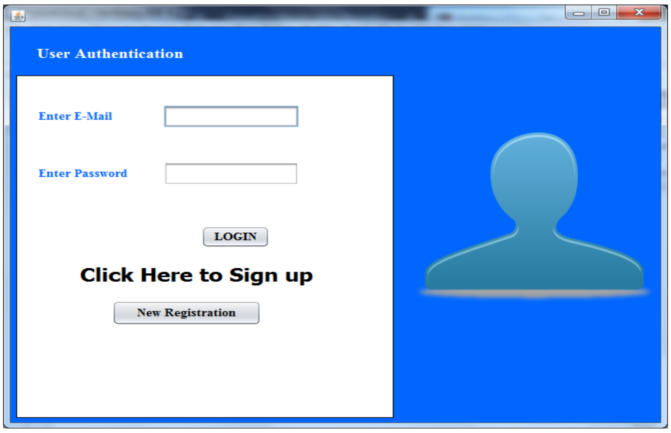


Fig 1:- Diagrammatic representation of deduplication technique.

IV. RESULTS


User Authentication

Enter E-Mail

Enter Password

Click Here to Sign up

Fig 2:- Window to login.

The main window contains two buttons. New Registration button is used to register the user to the database. Login button is used tologin the user.

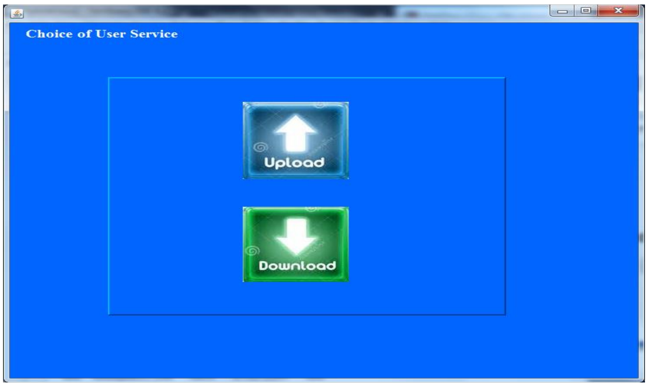
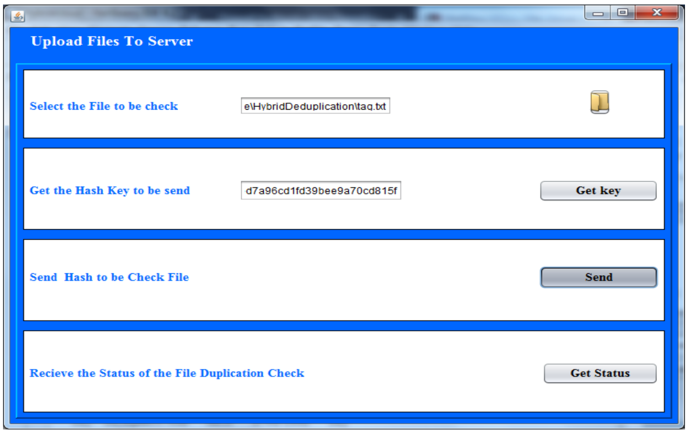

Choice of User Service

Fig 3:- Choice of user service window.

From this window user have to selectupload or download option.


Upload Files To Server

Select the File to be check e\HybridDeduplication\tag.txt

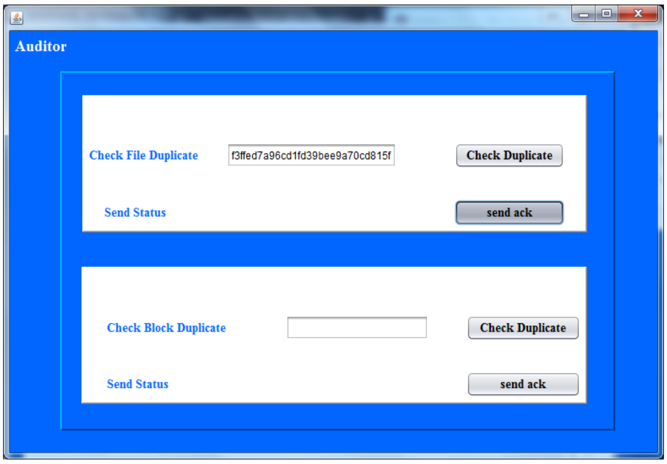
Get the Hash Key to be send d7a96cd1fd39bee9a70cd815f

Send Hash to be Check File

Recieve the Status of the File Duplication Check

Fig 4:- Upload file.

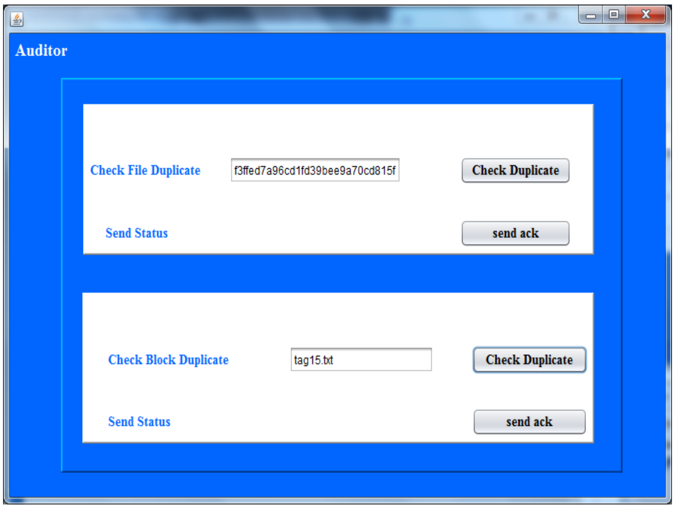
In this window we will upload the file and check the hash value.



Check File Duplicate: f3ffed7a96cd1fd39bee9a70cd815f
Send Status
Check Block Duplicate:
Send Status

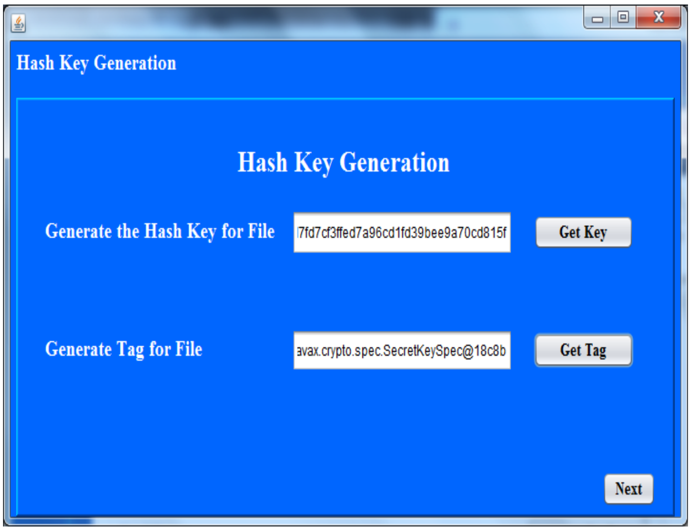
Fig 5:- Check for duplication.

It will check for file level and block level deduplication.



Check File Duplicate: f3ffed7a96cd1fd39bee9a70cd815f
Send Status
Check Block Duplicate: tag15.txt
Send Status

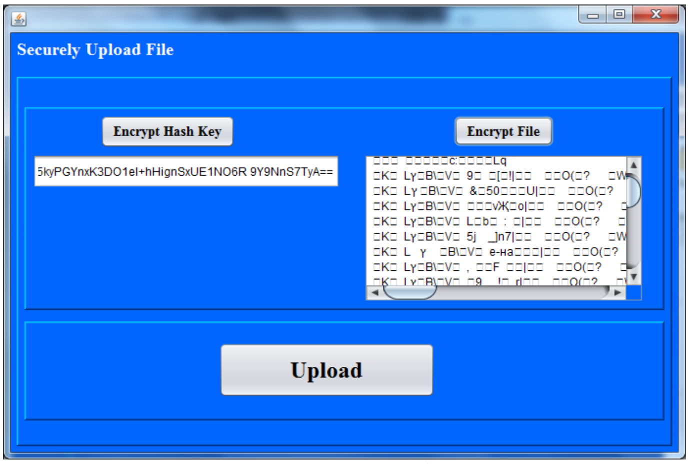
Fig 6:- Check for block level deduplication.



Generate the Hash Key for File: 7fd7cf3ffed7a96cd1fd39bee9a70cd815f
Generate Tag for File: avax.crypto.spec.SecretKeySpec@18c8b

Fig 7:- Hash key generation.

Hash key values and tag is generated for the files.



5kyPGYnxK3DO1el+hHignSxUE1NO6R 9Y9NnS7TyA==

Fig 8:- Upload the file

.

Encrypt the file and securely upload the file by clicking “UPLOAD” button.

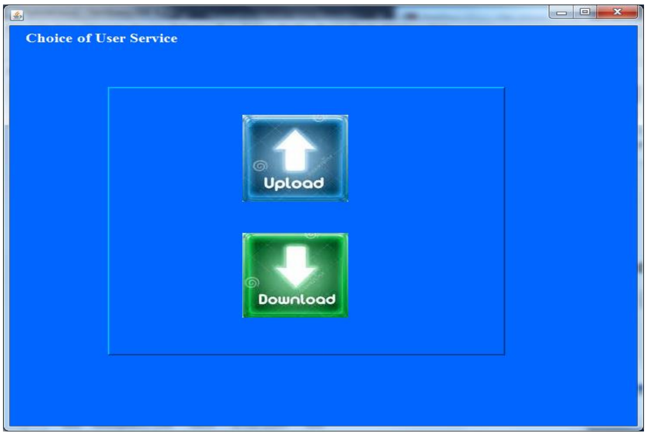
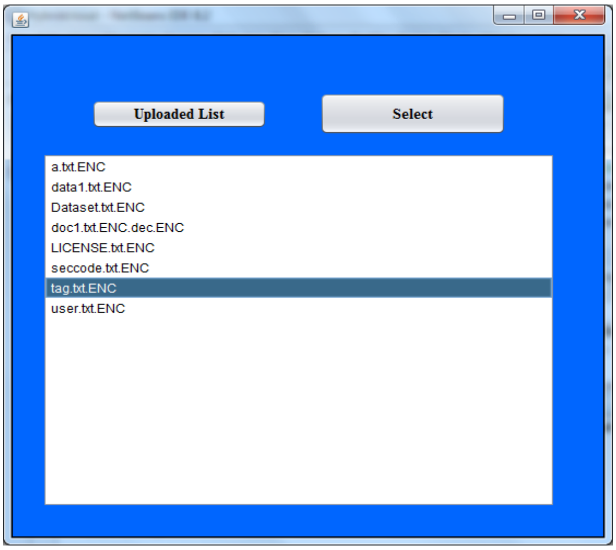


Fig 9:- Selecting the file.

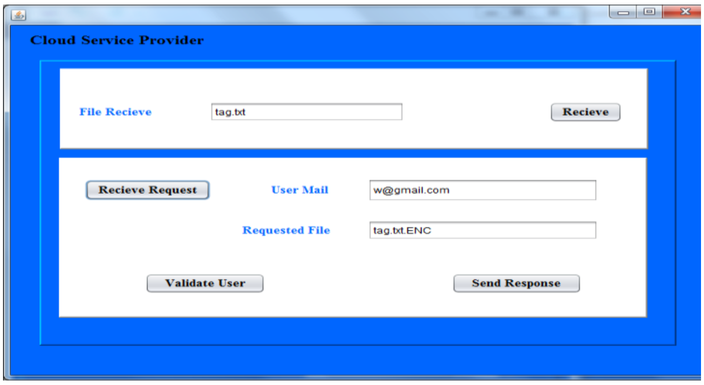
Once the file is uploaded we can download it by clicking on by “DOWNLOAD” icon.



a.txt.ENC
data1.txt.ENC
Dataset.txt.ENC
doc1.txt.ENC.dec.ENC
LICENSE.txt.ENC
seccode.txt.ENC
tag.txt.ENC
user.txt.ENC

Fig 10:- selecting the file.

We can select the file from uploaded list to download.The encrypted file can be selected.



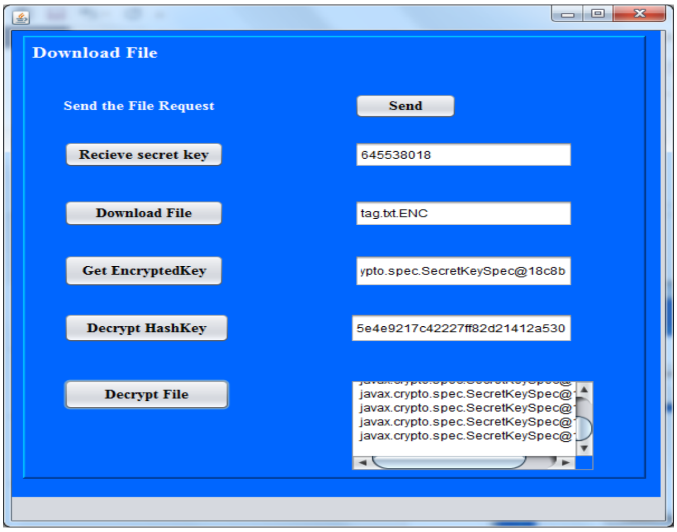
File Recieve: tag.txt

User Mail: w@gmail.com

Requested File: tag.txt.ENC

Fig 11:- CSP

We can receive the request from the user for the file then it will validate for the user identity and it will send the response.



Send the File Request

645538018

tag.txt.ENC

ypto.spec.SecretKeySpec@18c8b

5e4e9217c42227ff82d21412a530

```
javax.crypto.spec.SecretKeySpec@
javax.crypto.spec.SecretKeySpec@
javax.crypto.spec.SecretKeySpec@
javax.crypto.spec.SecretKeySpec@
```

Fig 12:- Download the file.

It will validate the user identity and send the file to download. It will decrypt the file at the last by using the encrypted, secret keys.

V. CONCLUSIONS

In this paper we focus on checking duplication of data and eliminating the redundant data. Here we are using file and block level deduplication techniques to check the similar data copies. If there are no any similar copies then the data will be stored, if not data will be eliminated. Encryption and decryption algorithms are used to encipher and decipher the data. By this technique we can efficiently use the storage space.

VI. ACKNOWLEDGMENT

I would like to express my deep sense of gratitude to our institution Srinivas School of Engineering for providing the means of attaining our most cherished goal. I would like to express our deep gratitude to our project coordinator Mr. Santhosh S, Asst. Professor, Dept of CSE for his wholehearted support, guidance and assistance for the successful completion of the project.

REFERENCES

[1]. C. Yang, J. Ren, and J. F. Ma, "Provable ownership of file in de-duplication cloud storage," in Proc. of IEEE Global Common. Conf. (GLOBECOM), pp. 695-700, 2013.

[2]. T.-Y. Wu, J.-S. Pan, and C.-F. Lin, “Improving accessing efficiency of cloud storage using de-duplication and feedback schemes,” IEEE Systems J., vol. 8, no. 1, pp. 208-218, 2014.

[3]. C.-I. Fan, S.-Y. Huang, and W.-C. Hsu, “Hybrid data deduplication in cloud environment,” in 2012 Int. Conf. Inf. Secure. Intel. Control (ISIC), pp. 174-177, 2012.

[4]. J. W. Yuan, and S. C. Yu, "Secure and constant cost public cloud storage auditing with deduplication,” in IEEE 2013 Conf. Common. Newt. Secure. (CNS), pp. 145-153, 2013. E.-J. Goh, H. Sachem, N. Modadugu, and D. Bone, “Sirius: securing remote untrusted storage," in Proc. Newt. Diatribe. Syst. Secure. Sump. pp. 131-145, 2003.

[5]. JZ. Yan, W. X. Ding, X. X. Yu, H. Q. Zhu, and R. H. Deng, “Deduplication on encrypted big data in cloud,” IEEE Trans. on Big Data, vol. 2, no. 2, pp. 138-150, April-June 2016.

[6]. Z. Yan, M. J. Wang, Y. X. Li, and A. V. Vasilakos, “Encrypted data management with deduplication in cloud computing,” IEEE Cloud Comput. Mag., vol. 3, no. 2, pp. 28-35, 2016.

[7]. 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.

[8]. J. Hur; D. Koo; Y. Shin; and K. Kang, “Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage,” IEEE Trans. Know. Data Eng., vol. 28, no. 11, pp. 3113-3125, 2016.

[9]. J. Li, X. F. Chen, M. Q. Li, J. W. Li, P. P. C. Lee; and W. J. Lou, “Secure Deduplication with Efficient and Reliable Convergent Key Management,” IEEE Trans. Parallel Diatribe. Syst., vol. 25, no. 6, pp. 1615-1625, 2014.

[10]. M. Wen, K. Ota, H. Li, J. S. Lei, C. H. GU; and Z. Su, “Secure Data Deduplication With Reliable Key Management for Dynamic Updates in CPSS,” IEEE Trans. Compute. Social Syst., vol. 2, no. 4, pp.137-147, 2015Sanai, and B. Waters, “Fuzzy identity-based encryption,” in Proc. of 24th Int. Conf. Theory App. Cryptographic Tech., pp. 457–473, 2005.

[11]. J. Li, Y. K. Li, X. F. Chen, P. P. C. Lee, and W. J. Lou. “A hybrid cloud approach for secure authorized deduplication,” IEEE Trans. Parallel Diatribe. Syst., vol. 26, no. 5, pp. 1206-1216, 2015S. C. Yu, C. Wang, K. Ren, and W. J. Lou, “Attribute based data sharing with attribute revocation,” in Proc. ACM Asia Conf. Compute. Common. Secure. pp. 261–270, 2010.

[12]. G. J. Wang, Q. Liu, and J. Wu, "Hierarchical attribute-based encryption for fine-grained access control in cloud storage services," in Proc. of 17th ACM Compute. Common. Secure. pp. 735-737, 2010.

[13]. M. Zhou, Y. Mu, W. Soil, M. H. Au, and J. Yan, “Privacy-preserved access control for cloud computing,”

in Proc. of IEEE 10th Int. Conf. Trust, Secure. Privacy Compute. Common. pp. 83-90, 2011.

[14]. "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", http://www. emc.com/leadership/digitaluniverse/2014iview/executivesummary.htm, April 2014, EMC Digital Universe with Research & Analysis by IDC.

[15]. Li, X. Chen, M. Li, J. Li, P. Lee, W. Lou,"Secure deduplication with efficient and reliable convergent key management", IEEE Transactions on Parallel and Distributed Systems, Vol. 25(6), Year – 2014.

[16]. Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, Mohammad Hassan, AbdulhameedAlelaiwi,"Secure Distributed Deduplication Systems with Improved Reliability", IEEE Transactions on Computers Volume: PP, Year – 2015.

[17]. Amazon Inc., "Amazon Elastic Compute Cloud," http://aws. amazon.com.

[18]. Amrita Upadhyay, Pratibha R Balihalli, Shashibhushan Ivaturi and Shrisha Rao,"Deduplication and Compression Techniques in Cloud Design", 2012 IEEE.