ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Implementation of Deduplication Scheme to Minimize Cloud Storage Requirement

Sulakshana S. Patange, Ganesh V. Kadam

PG Scholar, Department of Computer Engineering, JSPM NTC, RajashriShahu School of Engineering and Research, Narhe, Pune, India

Assistant Professor, Department of Computer Engineering, JSPM NTC, RajarshiShahu School of Engineering and Research, Narhe, Pune, India

ABSTRACT: A service model where data is provided to many users over the world is called as Cloud storage. Data in this model is stored in several logical pools. Cloud storage is a concept of storage beyond an interface where the storage is controlled and managed on demand. This technology is benefitted by many features like data backup and archival, no need of maintaining hardware resources, grater data accessibility, etc. Data deduplication is an important feature in cloud storage. Deduplication process recognizes and gets rid of the repeated data in the backup storage, indirectly improving the network bandwidth. Providing security along with the deduplication process is a challenging task. In our proposed work we use a new method known as convergent encryption for giving confidentiality to the data before it is outsourced. We use the cryptographic tuning and domain separation method to overcome the drawback of convergent encryption and to achieve privacy with higher extends than traditional system. In the proposed system we have utilized hybrid cloud architecture.

KEYWORDS: Convergent encryption, hybrid cloud, deduplication, cloud computing

I. INTRODUCTION

The ubiquity of the cloud computing has resulted in the widespread availability of cluster-based services and applications accessible through the Internet. Examples include online storage services, big data analytics, and e-commerce websites. In such a cluster-based cloud environment, each physical machine runs a number of virtual machines as instances of a guest operating system to contain different kind of user applications, and their data is stored in virtual hard disks which increases the security issues.

Considering an example, the Aliyun cloud, which is the largest cloud service provider by Alibaba in China, automatically conducts the backup of virtual disk images to all active users every day The cost of supporting a large number of concurrent backup streams is high because of the huge storage demand. If such VM snapshot data is plainly backed up without any duplicate reduction, storage waste would be extreme high, So In the recent decade, data deduplication techniques are emerged to solve this data redundancy problem in backup systems.

There are multiple dedicated storage servers in a cloud storage model [1]. There may be multiple third parties hosting these storage servers. There are many challenges in the computing domain, such as the existence of the redundant data.This redundant data can be handled by a technique known as data deduplication technique [2]. The important data should be backed-up to avoid data loss. The backup applications are used to back up the data. The application creates a big backupfile,which is stored on the space. Each and every backup file of same data can create redundant copy of the data. Transmitting the big backup file needs a lot of network resources. The redundant copy of the data must be removed, and this is done by the data deduplication technique. In this technique, the copied data or duplicate data blocks are removed, and one a single copy iskept on the server [3]. This approach reduces the storage space, the cost of maintaining the storage space and also indirectly decreases the network bandwidth usage.

The rest of the paper is arranged as follows. In section II the related work is discussed. In section III, we propose the system model for our deduplication system. We support the discussion by considering problem definition, barriers of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

existing systems,mathematical model. In the next section the experimental results have been discussed and then we conclude paper in section .

II. RELATED WORK

Cloudstorage and data deduplication techniques have pulled the attention of many researchers recently. He et al. [4] have talked over various cloud storage techniques and have recommended some techniques for reducing the storage volumes.The authors have proposed data deduplication engine that generated an index of digital signatures. This index also provides pointers for knowing the presence of data blocks. A new concept of data migration is emerging now-a-days. It is nothing but the relocation of data from one storage to other storage. Both the storages are geographically separated.New concept known as Proof of Ownership is implemented for deduplication process by Halevi et al. [5]. Here a client can manifest based on Merkle-Hash Tree [5]. But the proposed system doesnt take into attention providing the data security. The client-side deduplication has many limitations which are discussed by Harnik et al. [6]. An excellent survey on various deduplication techniques is done in [7]. The concept of deduplication is very simple, it says only a single copy of the duplicate data must be maintained and there should be a pointer for pointing the duplicate blocks. And this process can be attained in three levels: file level, byte level and block level.The researcher [8] given a framework comprising of twin cloud for protected outsourcing of information and subjective processing to an untrusted service cloud. The researcher [10] proposed a deduplication system in cloud to reduce the storage size of the tags for integrity check. To upgrade the security of deduplication and secure the information secrecy. The researcher [9] demonstrated to secure the information by transforming the predictable message into unpredictable message.The researcher [11] proposed the "Sparse Indexing" deduplication system uses a different approach to avoid the chunk lookup disk bottleneck. Here, the chunks are sequentially grouped into segments. These segments are then used to search similar existing segments using a RAM based index, which stores only a small fraction of the already stored chunks. In contrast to other approaches, Sparse Indexing allows to store a chunk multiple times if the similarity based system is not able to detect the segments, which already have stored the chunk. Therefore, Sparse Indexing is a member of the class of approximate data deduplication systems.

III. PROPOSED METHODOLOGY AND DESIGN

The propounded system is presented for carrying out secured authorized deduplication process. It has many significant features. The architecture used is hybrid cloud architecture. This section will describe the system design, the entities, assumptions made while developing the system and the implementation plan.

A. Problem Definition

To increase the amount of information that can be stored on cloud by saving bandwidth and to eliminate duplicate copies of redundant data to preserve confidentiality of sensitive data while supporting deduplication.

B. System Overview

The proposed system is based on the hybrid cloud architecture. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server instead. In this way, the users cannot share these private keys of privileges in this proposed construction, which means that it can prevent the privilege key sharing among users in the above straightforward construction. To get a file token, the user needs to send a request to the private cloud server. The intuition of this construction can be described as follows. To perform the duplicate check for some file, the user needs to get the file token from the private cloud server. The private cloud server will also check the users identity before issuing the corresponding file token to the user. The authorized duplicate check for this file can be performed by the user with the public cloud before uploading this file. Based on the results of duplicate check, the user uploads this file.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

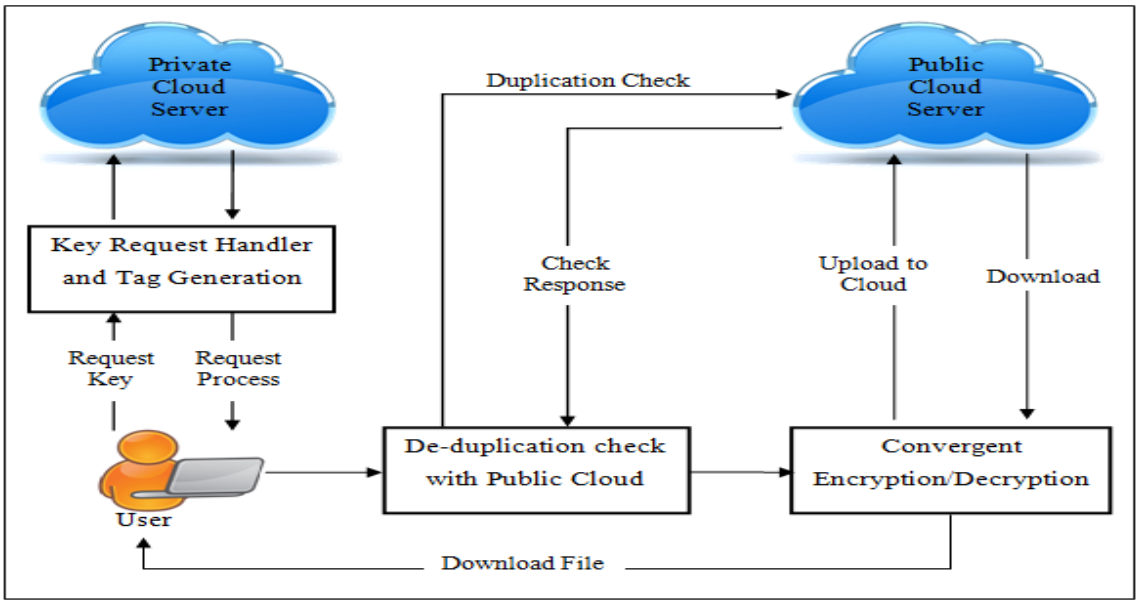

Figure 1.Proposed System Architecture

C.Assumptions and entities

The proposed system is based on some assumptions made. The very first assumption is that the S-CSP has plentiful storage capacity and computational power. Also it is online every time. The second assumption is that, while the system is setup, the user is provided with a certain set of privileges. The third assumption includes the sensitivity of the files and the need of protecting them from both the clouds.There are three entities defined in our system as shown in figure 2.

1.Data Users - A user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth, which may be owned by the same user or different users. In the authorized deduplication system, each user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges. The user is issued a set of privileges in the setup of the system. Each file is protected with the convergent encryption key and privilege keys to realize the authorized deduplication with differential privileges.

2.Private Cloud - Compared with the traditional deduplication architecture in cloud computing, this is a new entity introduced for facilitating users secure usage of cloud service. Specifically, since the computing resources at data user/owner side are restricted and the public cloud is not fully trusted in practice, private cloud is able to provide data user/owner with an execution environment and infrastructure working as an interface between user and the public cloud. The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users. The interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

3.S-CSP: This entity is one of the established parts into the public cloud. It is used for providing data outsourcing services, data storage services for the user. S-CSP performsdeduplication for storing only the distinctive data instead of storing the redundant data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

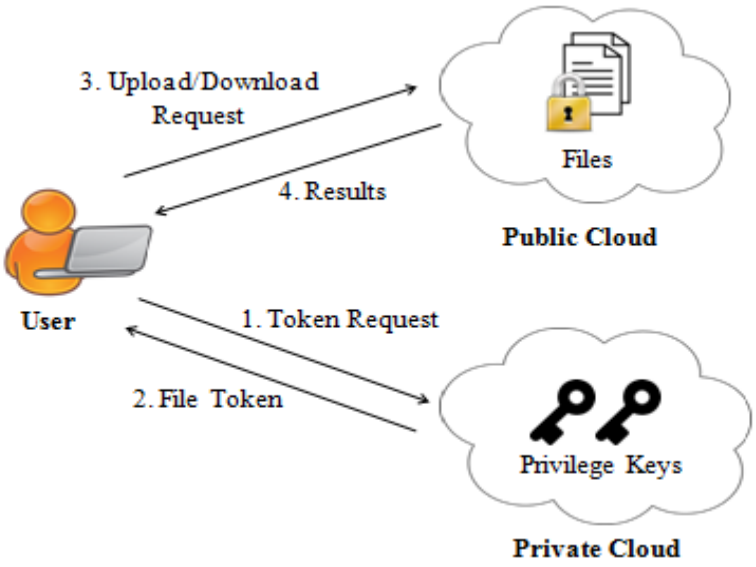

Figure 2. Authorized Deduplication

D. Implementation Plan

The implementation plan is based upon the notion of authorized deduplication as shown in figure 2. There is a tag generated for a file *F*, and this tag is based on the privileges of the file *F*. In our proposed work, the tags are called as file tokens. Tag is an conventional notation [12]. The process of authorized access requires a secret key $_{Kp}$ joined with a privilege *p*. Then the token generated can only be used by the user with privilege *p*[13][14]. In other words, the user with privilege *p* can only compute the token denoted by $_{TF,p}$ and is given as $_{TF,p}$= *TagGen(F,*$_{Kp}$). There are many ways of token creation methods, one of which can be *H(F,Kp)*, where H(.) is nothing but a function of hash. We have used convergent encryption method, discussed in the sub section.

1) Convergent Encryption: The encrypted data contradicts the deduplication process, that is for instance consider that there are two similar files, encryption of both the files with a different set of keys produces different encrypted output,which cannot be shared anymore. This problem can be solved by a method introduced as convergent encryption technique. This methods states that the key for encryption should be created from the data of the file itself by means of function similar to a hash function. Then from the above example, encryption of both the files will produce same encrypted output, hence, indirectly supporting the deduplication process. The encryption keys generated in this way are separate. But the owner of the file has to make some attempts for receiving the keys created from the data of the file. The managing and storing of the keys is a bit challenging. It's a overhead for maintaining the metadata of the keys. Hence, the encryption and deduplication process is carried out remotely in our proposed work. The method of convergent encryption [15][16][17][18] must be implemented correctly so that it produces a secured environment for the data like any other encryption scheme.

2) Domain Separation: Domain separation and Cryptographic tuning approaches are used for solving the drawback of convergent encryption as the encryption fails to deal with many security aspect. .

E.ProposedAlgorithm
Input: File *F*
Output: Encrypted file stored over cloud.

1. The object to be encrypted is validated to ensure it is suitable for this type of encryption. This generally means, at a minimum, the file is sufficiently long. (There is no point in encrypting, say, 3 bytes this way. Someone could trivially encrypt every 3-byte combination to create a reversing table.)
2. Some kind of hash of the decrypted data is created. Usually a specialized function just for this purpose is used,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

not a generic one like SHA-1. (For example, HMAC-SHA1 can be used with a specially-selected HMAC key not used for any other purpose.)

3. This hash is called the 'key'. The data is encrypted with the key (using any symmetric encryption function such as AES-CBC).
4. The encrypted data is then hashed (a standard hash function can be used for this purpose). This hash is called the 'locator'.
5. The client sends the locator to the server to store the data. If the server already has the data, it can increment the reference count if desired. If the server does not, the client uploads it. The client need not send the key to the server. (The server can validate the locator without knowing the key simply by checking the hash of the encrypted data.)
6. A client who needs access to this data stores the key and the locator. They send the locator to the server so the server canlookup the data for them, then they decrypt it with the key. This function is 100 percentage deterministic, so any clients encrypting the same data will generate the same key, locator, and encrypted data.

F.Mathematical model

Let C be the set of clients that is it may have one or more clients. *CSP* is the cloud service provider. *GUI* is the graphical user interface. *F* denotes a file. Hence, the system *S* is given as follows:
S = {C, CSP, GUI}

Global set of entities involved are described below:
G ={D, U, PC, PRC, VM}
Where, *D* is the data owner, *U* is the end user, *PC* denoted as public cloud, *PRC* as the private cloud, *VM* as computational virtual machine. There are two functions p1 and p2 which are described as verification of clients and verification of end users respectively.

The processes that are carried out are: Verification, Encryption, Decryption and Locator. In the verification process, the user needs to login with a username and password. The inputs to the encryption process are the contents to encrypted and thekey. Similarly, the inputs to the decryption process are key and the cipher text to be decrypted. Locator contains H(C).
Process:
Verification[Y/N] = login(Username, Password)
Encryption = (Key, Content)
Decryption = (key, Cipher)
Locator = H (C)

IV.RESULT AND ANALYSIS

A. *Dataset*
This system does not use any precise dataset. Input to the system is a file.

B. Experimental Setup:
The Proposed system has been divided into following modules
UI Implementation- It will focus on creation of UI for entire application. This module will create the platform for data owner to browse the data and store it on cloud. The data will be first encrypted and tags will be generated for the same. Once we will create basic application to setup data storing we will focus on encryption and data tagging in next module.

Convergent Encryption and data tagging- In this module we will focus on encryption techniques to be applied for data to be encrypted. Apart from this for every data blocks, the tag needs to be generate which can be then used to verify the identity of blocks or files.
The Upload and Download module- will allow the user to upload files on cloud and will allow its end users to decrypt the files from cloud and access it. This module will also allow to data owner to put access control mechanism which can be consider as a contribution to the proposed system. This module will also checked the proof of ownership and verify the data by comparing hash value.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Analysis - This module will allow us to compare our approach with other existing encryption and tagging concept.This will also allow us to improve the runtime system after comparing it with other existing data security systems.

C.Experimental Evaluation:

We have performed the given experiment on java platform with mysql as a database to keep track of users and files. The system is evaluated with the windows machine with latest configuration with 4GB RAM and I series processor. We have uploaded files with various formats and noted down different execution time with j-profiler and timer by running application with debug mode. The following parameters are checked.

1. Execution Time:

The overall execution time for file upload including de-duplication check and encryption for particular file is as mentioned in below table2. The graph analysis for execution time shownin figure 3..

Table 2.Execution time and File size

Approach	Execution Time (ms)	File Size(Kb)
Without Cryptographic tuning and domain separation	78	156
Our Approach	48	156

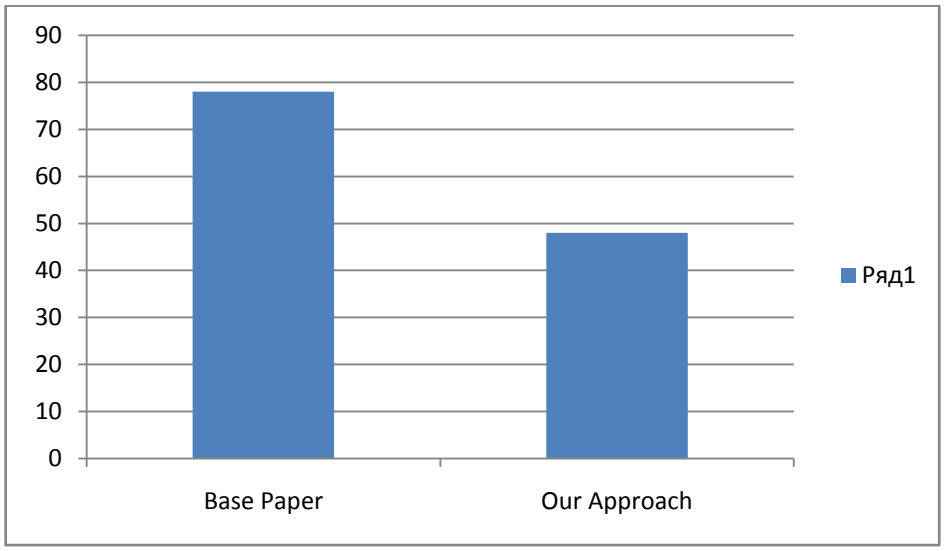

Figure.3 Execution time

2. Throughput

Based on the execution time and file size we have also calculated the throughput of the system and which clearly shows that our system improves throughput with cryptographic tuning and domain separation. Throughput of proposed system is as shown in figure 4 estimated with Throughput Calculator Tool.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

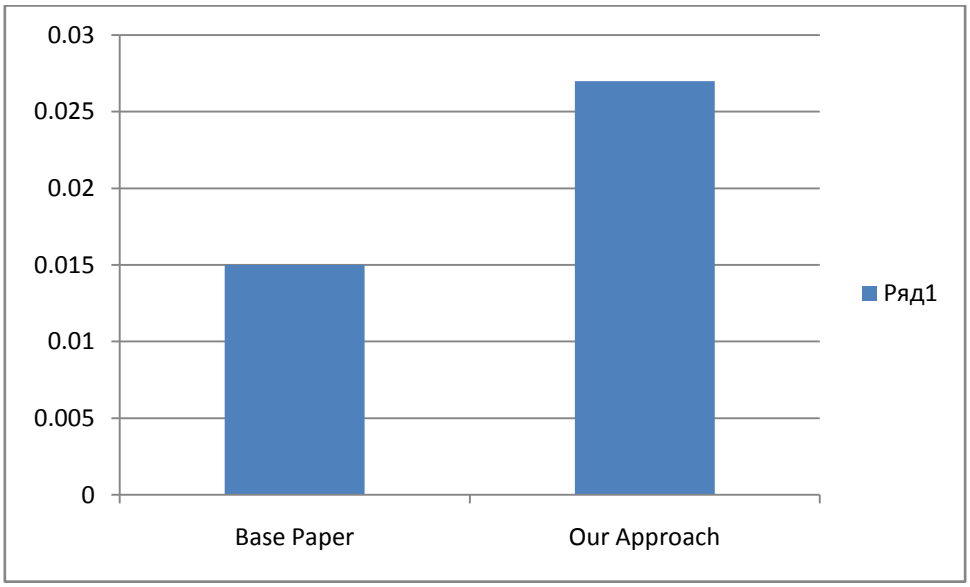

Figure 4.Throughput Graph

V.CONCLUSION

The notion of authorized data deduplication was proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. We used convergent encryption with modification version to deal with brute force attack using Domain Separation and Cryptographic tuning to make better authorized deduplication technique.

REFERENCES

[1] Panzieri, Ozalp, Babaoglu1, Stefano, Ferretti, Vittorio, Ghini,MorenoMarzolla, Distributed Computing in the 21st Century:of Cloud Computing Fabio Technical Report UBLCS-2011-03 May 2011.
[2] DeepavaliBhagwat, KaveEshghi, Darrell D. E. Long, Mark Lillibridge., Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup.
[3] Deduplication and Compression Techniques in Cloud Design"by Amrita Upadhyay, pratibha R Balihalli, ShashibhushanIvaturi and ShrishaRao 2012 IEEE.
[4] Z. Li, X. Zhang, and Q. He, Analysis of the key technology on cloud storage, in international Conference on Future Information Technology and Management Engineering, 2010, pp. 427428.
[5] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg.Proofs of ownership in remote storage systems. In Y. Chen, G. Danezis, and V. Shmatikov, editors, ACM Conference on Computer and Communications Security, pages 491500. ACM, 2011.
[6] D. Harnik, B. Pinkas, and A. Shulman-Peleg. Side channels in cloud services: Deduplication in cloud storage. IEEE Security & Privacy, 8(6), 2010.
[7] Q. He, Z. Li, and X. Zhang, Data deduplication techniques, in International Conference on Future Information Technology and Management Engineering,pp.431-432,2010.
[8] Bugiel, Sven, et al. "Twin clouds: An architecture for secure cloud computing." Proceedings of the Workshop on Cryptography and Security in Clouds Zurich. 2011.
[9] Bellare, Mihir, SriramKeelveedhi, and Thomas Ristenpart. "DupLESS: server-aided encryption for deduplicated storage." Proceedings of the 22nd USENIX conference on Security.USENIX Association, 2013.
[10] Yuan, Jiawei, and Shucheng Yu. "Secure and constant cost public cloud storage auditing with deduplication." Communications and Network Security (CNS), 2013 IEEE Conference on.IEEE, 2013.
[11] D. Ferraiolo and R. Kuhn. Role-based access controls. In *15th NIST-NCSC National Computer Security Conf.*, 1992
[12] M. Bellare, S. Keelveedhi, and T. Ristenpart.Messagelockedencryption and secure deduplication. In UROCRYPT,pages 296 312, 2013.
[13] D. Ferraiolo and R. Kuhn. Role-based access controls. In 15th NIST-NCSC National Computer Security Conf., 1992.
[14] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E.Youman.Role-based access control models. IEEE Computer,29:3847, Feb 1996.
[15] K. Bennett, C. Grothoff, T. Horozov, and I. Patrascu. Efficient sharing of encrypted data. In In Proceedings of ASCIP 2002,pages 107120. Springer-Verlag, 2002.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

[16] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M.Theimer.Reclaiming space from duplicate files in a serverless distributed file system.In In Proceedings of 22nd International Conference on Distributed Computing Systems ICDCS, 2002.
[17] W. J. Bolosky, J. R. Douceur, D. Ely, and M. Theimer. Feasibility of a serverless distributed file system deployed on an existing set of desktop PCs. In Proceedings of the international conference on measurement and modeling of computer systems (SIGMETRICS), 2007.
[18] M. W. Storer, K. Greenan, D. D. Long, and E. L. Miller.Secure data deduplication. In StorageSS 08: Proceedings of the 4th ACM international workshop on Storage security and survivability, pages 110, New York, NY, USA, 2008.

BIOGRAPHY

Sulakshana S. Patangeis P. G. Scholar in the Computer Engineering Department, JSPM's RajashriShahu School of Engineering and Research, Narhe, Pune, India She has received Bachelor of Engineering (B.E.) in Computer Engineering from Pune Institute Of Computer Technology (Pune University),India. She is currently working as Lecturer in MIT Polytechnic, Kothrud, Pune. Her research interests are Securityand software engineering.

Prof. Ganesh V.Kadamis a full time Assistant Professor at Department of Computer, JSPM's RajashriShahu School of Engineering and Research, Narhe, Pune, India.