ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (600-606), Month: April - June 2015, Available at: www.researchpublish.com

Secure Authorized Data De-Duplication Using Convergent Encryption Technique

[1]Pradeep H R, [2]Yamuna U

[1] MTech in Computer networks and engineering, SJB Institute of technology Bangalore, Karnataka, India
[2] Assistant professor, Department of Information science and engineering, SJB Institute of technology Bangalore, Karnataka, India

***Abstract*: Data de-duplication is the important data compression techniques which is used in eliminating the duplicate copies of repeating data in the cloud storage environment, and has been widely used in cloud storage environment to reduce the amount of storage space needed and used to save bandwidth. The convergent encryption technique has been used to encrypt the data before outsourcing, it is used protect the confidentiality of the data which is sensitive in nature while supporting de-duplication. Authorized data de-duplication has been introduced to better protect data security. This proposed system is different from the traditional de-duplication systems, the differential privileges of the users are further considered while processing duplicate check besides the data itself. Presented several new de-duplication constructions supporting authorized duplicate check in a hybrid cloud architecture. Security analysis proves that this scheme is secure in terms of the definitions specified in the proposed security model. Implementation of prototype for the proposed authorized duplicate check scheme has been done and also conducted the testbed experiments using this prototype. Compared to the normal operation the proposed authorized duplicate check scheme incurs minimum overhead.**

***Keywords:* De-duplication, authorized duplicate check, confidentiality, hybrid cloud.**

I. INTRODUCTION

Cloud computing is the terminology where the computational services takes place over the Internet. Cloud services will offer individuals and businesses organisation to use software and hardware that are managed by third parties service providers at remote locations. For example cloud computing service includes online file storage, social networking sites etc. Using cloud computing model accessibility to information and computer resources from anywhere in the internet can be done. Cloud computing provides a shared pool of resources that includes storage space, networks, processing power, and specialized corporate and the user applications. Figure 1.1.shows the cloud architecture. Cloud computing infrastructure can be classified as public cloud, private cloud or hybrid cloud. Public cloud is based on standard cloud computing model, which services are offered over the Internet and are owned and operated by a cloud service provider. Some examples for public cloud services are online photo storage services, e-mail services, or social networking sites and services for enterprises can be offered in a public cloud. Public cloud services can be free as well as offered on a pay as you go service. A private cloud is designed to offer the same features and benefits as that of the public cloud systems, but it includes control over enterprise and customer data, worries about security and data confidentiality, and issues connected to the regulatory compliance.

There will be a privacy concerns in the cloud computing environment because the cloud service provider can access the data that is stored on the cloud storage. There is a possibility of accidentally or deliberate attempt of alteration or even deletion of information. Many cloud service providers can share information with the third parties if necessary for purposes of law and order without any warrant. Before the user starts using the cloud storage environment privacy policies should be agreed upon. Solutions to the privacy includes policies and legislation as well as end user's choices for

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (600-606), Month: April - June 2015, Available at: www.researchpublish.com

how the data is to be stored. Users can encrypt the data using various data encryption scheme and store it in the cloud storage environment.

In computing, data de-duplication is also called as the data compression technique for eliminating multiple copies of same data. Related and somewhat synonymous terms are intelligent data compression and single-instance data storage. This data de-duplication technique is used to improve the storage utilization very efficiently and can also be applied to network data transfers to reduce the number of bytes that is sent over the internet or the network. In the de-duplication process, unique identity of the data, or byte pattern, are identified and then stored during a process of analysis. As the analysis continues, other chunks of data are compared to the stored copy which is present in the storage area and whenever a match occurs, the redundant chunk is replaced with a small reference point that points to the stored data chunk. Given that the same byte pattern may occur several times, the amount of data that must be stored or to be transferred can be reduced efficiently.

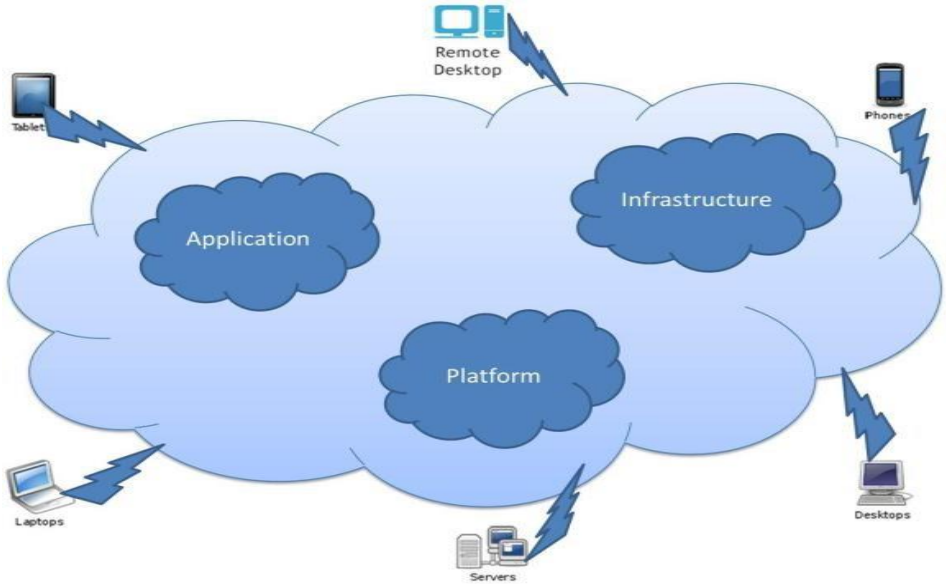

Figure 1.1 Cloud Architecture

Data de-duplication or Single Instancing essentially refers to the elimination of the multiple instances of similar data copy. In the de-duplication process, duplicate data is deleted, leaving only one single instance of the data to be stored in the storage environment. However, indexing of all data is still retained that the data ever be required to the user in future. In general, data de-duplication technique eliminates the duplicate copies of repeating data in the storage area

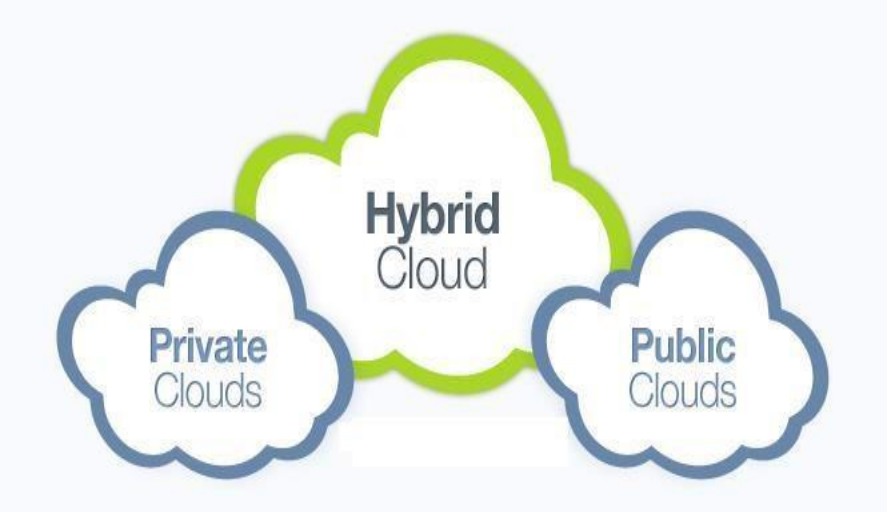

Figure 1.2 Hybrid Cloud computing Architecture

Hybrid cloud computing approach is used to avoid this duplicate data and to maintain the confidentiality of the data in the cloud storage environment. This hybrid cloud computing approach is a combination of both public and private cloud. Advantage of using hybrid cloud storage is scalability, reliability, rapid deployment and potential cost savings of public cloud storage with the security and full control of private cloud storage. Figure 1.2. Shows the hybrid cloud architecture.

II. LITERATURE SURVEY

1. Data Encryption:

Encryption is the process of encoding the information from one format to other. Information is encrypted in such a way that only authorized users can access and read it. In an encryption process information is referred to as the plaintext. Encrypted information is the cipher text. The output of encryption is called the cipher text. Plaintext is encrypted using encryption algorithm that can be symmetric or asymmetric algorithm. Plaintext is encrypted using some keys and it is not possible to decrypt the cipher text without these keys. There are two types of encryption process Symmetric key encryption process.

 Public key encryption process

Symmetric key encryption: In symmetric key encryption process, the encryption and decryption key is same. Public key encryption: In public key encryption schemes, the encryption key is published for anyone to use it as encryption key indeed. Only receiving parties will be having accessibility to the decryption key that enables messages to be read.

2. Symmetric Encryption process:

This is the type of encryption process in which the same key is used to encrypt and decrypt the message content. This differs from asymmetric encryption, which uses two keys. One key is used to encrypt the message content and another key is used to decrypt the message content. An encryption is the process in which the sender and receiver of a message share the single key and this key is used to encrypt and decrypt the message content.

Symmetric key encryption is simpler easier and faster when compared to asymmetric encryption process. Symmetric key encryption is also called secret-key encryption method. The most popular symmetric key encryption is Advanced Encryption Standard (AES), Data Encryption Standard (DES) and 3DES.

A symmetric encryption consists of three primitive functions, They are

- KeyGener.(1y) :- It is the key generation algorithm i.e SHA-1 that generates key k using the security parameter 1y.
- Encrypt.(k,M) :- It is the symmetric key encryption algorithm. It takes the key k as encryption key and encrypts the plaintext M message content. It produces cipertext C of message M.
- Decrypt.(k,C) :- It is the symmetric key decryption algorithm. It takes the key k as decryption key and decrypts the cipher text C. It produces the original plaintext M from cipher text C indeed.

Disadvantages:

The main drawback is that the two parties must share the keys in a secure manner.

3. Asymmetric Encryption process:

Symmetric-key cryptosystems use the same key for encryption and decryption of the message content, though a message or group of the messages may have the different keys than others. The main disadvantage of symmetric ciphers is the key management necessary to use them securely indeed. Each distinct pair of communicating parties must ideally share different keys, and perhaps each ciphertext is swapped as well. The number of keys that required increases as the square of the number of the network associates, which very quickly requires complex key management system to keep them all consistent and secret. The difficulty of securely founding a secret key amid two communicating parties, when a secure channel does not previously exist between them, also presents a chicken-and-egg problem which is a considerable practical problem for cryptography users in the actual world.

Whitfield Diffie and Martin Hellman projected the notion of *public-key* cryptography in which two dissimilar but mathematically linked keys are used—a *public* key and a *private* key. A public key system is so built that the calculation

of one key (the 'private key') is computationally infeasible from the other (the 'public key'), even though they are necessarily related with each other. Instead, both keys are engendered secretly, as an unified pair.

In public-key cryptosystems, the public key can be liberally dispersed, while its paired private key must persist secret. In the public-key encryption system, the *public key* is used for the encryption purpose, while the *private* or *secret key* is used for decryption purpose. While Diffie and Hellman may well not find such a system, they showed that public-key cryptography was certainly possible by bestowing the Diffie– Hellman key exchange protocol, a answer that is now widely used in protected communications to allow two parties to secretly decide on the shared encryption key.

Diffie and Hellman's publication flashed widespread academic exertions in finding an applied public-key encryption system. This contest was finally attained in 1978 by Ronald Rivest, Adi Shamir, and Len Adleman, whose answer has since become known as the RSA algorithm.

The Diffie–Hellman and RSA algorithms, in accumulation to being the first publicly known examples of high superiority public-key algorithms, have been amid the most extensively used. Others embrace the Cramer–Shoup cryptosystem, ElGamal encryption, and various elliptic curve techniques.

4. Data De-duplication Technique:

Data de-duplication is a specialized technique, which is used to eliminate the duplicate copies of the same data content. It is used to improve the storage space utilization and also it can be applied to network data transfer to reduce the amount of data that is to be transferred

Post process de-duplication:

In this de-duplication process, new data is first stored on the storage environment and then it will analyse the data looking for duplication which is present or not.

Advantages:

The benefit is that here no need to wait for the hash value calculations.

Disadvantages:

One of the potential drawbacks of post process de-duplication is that it may store the duplicate data unnecessarily for a short period of time which is an issue if the storage system is near the full capacity.

Inline de-duplication:

In this process where the de-duplication hash value calculations are created on the target device as the data enters the device in the real time. If the device spots a block of data that it already stored on the system then it does not store the new block of data.

Advantages

It requires less storage space for data storage as data is not duplicated.

Disadvantages

Hash value calculations are performed in the client side. It increases the overhead in the client side.

5. Convergent Encryption Technique:

Convergent encryption technique is also known as content hashing keyword method. It is a cryptosystem that produces identical ciphertext from the identical plaintext data. This is mainly used in cloud computing to remove duplicate elements or files from storage environment without the provider having to access to the encryption keys. It generates the file tag, which is used to detect the duplicate files in the storage environment.

The system first computes the cryptographic hash value of the plaintext message. It then encrypts the plaintext by using its hash value that is generated as a key. Finally, the hash value (key) itself is encrypted with the key chosen by the user indeed.

Convergent encryption scheme can be defined in with four primitive functions as shown below.

- KeyGeneration.(M) :- It is the key generation algorithm i.e SHA-1 that generates key k using the data copy message M.
- Encryption.(k,M) :- It is the symmetric key encryption algorithm i.e AES. It takes the key k as encryption key and encrypts the plaintext message M. It produces C as the ciphertext of message M.
- Decryption.(k,C) :- It is the symmetric key decryption algorithm. It takes the key k as decryption key and decrypts the ciphertext C. It produces the original plaintext M from ciphertext C.
- TagGen.(M) :- It is the file tag generation algorithm that maps the original data copy M and outputs a tag T(M).

6. Proof of Ownership Protocol:

The proof of ownership is the protocol which is used to notice that the particular user is the owner of the particular. Halevi et al. proposed the notion of ―proofs of ownership‖ (PoW) for the de-duplication systems, such that a client ie user can efficiently prove to the cloud storage server that he owns a file without necessity of uploading the file itself. Several PoW constructs based on Merkle-Hash .Tree are proposed to enable client-side data de-duplication, which include the bounded leakage of setting. Pietro and Sorniotti proposed another efficient PoW scheme by choosing the projection of the file onto a randomly selected bit-positions as the file proof indeed. Note that all the above schemes that is given, do not consider the data privacy. Recently, Ng et al. for encrypted files, but they did not addressed how to minimal the key management overhead.

III. PROPOSED SYSTEM

Convergent encryption is proposed in this project to overcome the drawbacks of existing traditional encryption scheme. It enforces the data confidentiality and also data deduplication feasible. It encrypts or decrypts the file using the convergent key. This convergent key is obtained by computing the hash value using content of the file as unit. User encrypts the file using convergent key and sends the cipher text to the storage cloud i.e public cloud. The identical data copies will produce same cipher texts.

In the proposed system secure authorized duplicate check concept is used. In the authorized duplicate check, each user's authorization has been checked. Each authorized user will get the file token from the ticket generating server which is present in private cloud. Using the file token each user will request for the duplicate check in the public cloud storage environment.

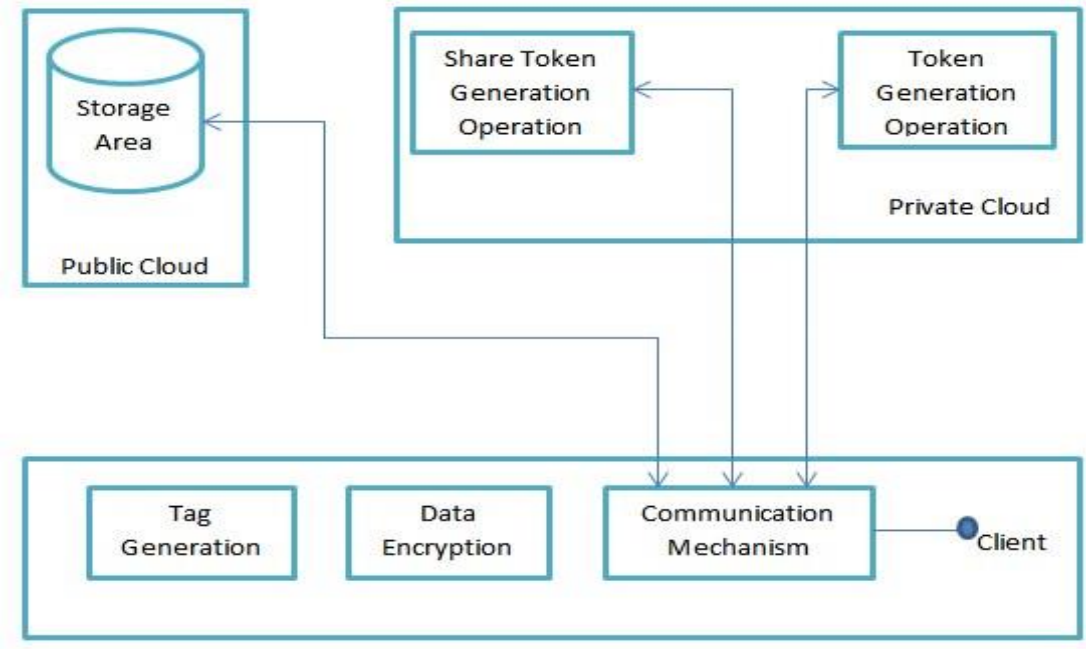

Fig1.Proposed Framework for communication

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (600-606), Month: April - June 2015, Available at: www.researchpublish.com

To prevent unauthorized access to the data present in the storage area, Proof of Ownership (PoW) is also proposed. If the PoW is success then the user will get the pointer to the existing file without uploading the same file to the storage environment again.

ENCRYPTION OF FILES:

A symmetric encryption consists of three primitive functions, They are

- KeyGener.(1y) :- It is the key generation algorithm i.e SHA-1 that generates key k using the security parameter 1y.
- Encrypt.(k,M) :- It is the symmetric key encryption algorithm. It takes the key k as encryption key and encrypts the plaintext M message content. It produces cipertext C of message M.
- Decrypt.(k,C) :- It is the symmetric key decryption algorithm. It takes the key k as decryption key and decrypts the ciphertext C. It produces the original plaintext M from ciphertext C indeed.

CONFIDENTIAL ENCRYPTION: It delivers data confidentiality in deduplication. A user derives a convergent key from each unique data copy and encrypts the data copy with the convergent key. In addition, the user also stems a tag for the data copy, such that the tag will be used to sense duplicates.

PROOF OF DATA: The proof of ownership is the protocol which is used to notice that the particular user is the owner of the particular. Halevi et al. proposed the notion of ―proofs of ownership‖ (PoW) for the de-duplication systems, such that a client ie user can efficiently prove to the cloud storage server that he owns a file without necessity of uploading the file itself. Several PoW constructs based on Merkle-Hash .Tree are proposed to enable client-side data deduplication, which include the bounded leakage of setting. Pietro and Sorniotti proposed another efficient PoW scheme by choosing the projection of the file onto a randomly selected bit-positions as the file proof indeed. Note that all the above schemes that is given, do not consider the data privacy. Recently, Ng et al. for encrypted files, but they did not addressed how to minimal the key management overhead.

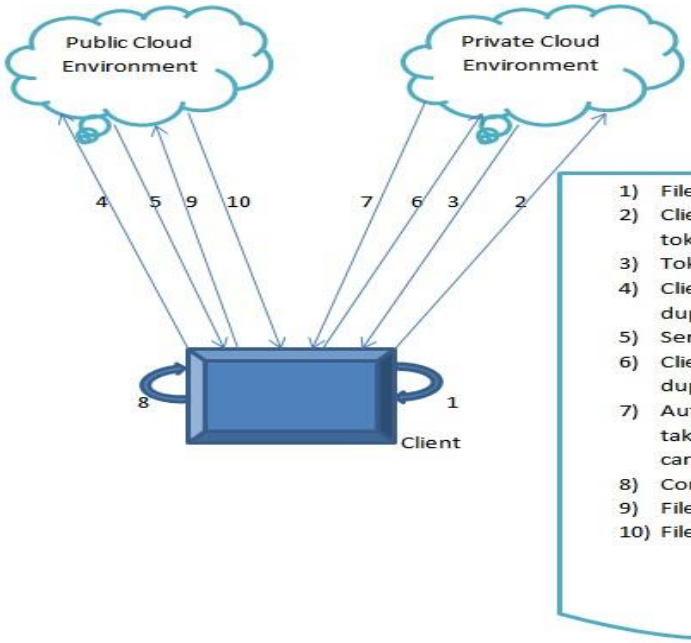

Fig1. Shows the Proposed System Framework

1) File tag generation using SHA-1
2) Client sends request to private cloud for file token
3) Token generation and send to client
4) Client sends request to public to check duplicate file
5) Server checks and sends appropriate result
6) Client requests share token from private is duplicate file is found
7) Authentication and authorization operation takes place and share token operation is carried out and sends the token to the client
8) Convergent key is used the encrypt the data
9) File uploads
10) File upload process complete

Fig2. Proposed system of operation

IV. CONCLUSION

Cloud services and its usage are mounting in a radical scale. As the number of users increases the amount of data that is stored in the storage environment increases and also risk to the data increases. In this proposed system the notion of authorized data de-duplication process has been proposed to protect confidentiality of the user's data. For better confidentiality and security in cloud computing the new data de-duplication construct supporting authorized duplicate check has been introduced in hybrid cloud computing architecture, in which the duplicate check tokens for the files are

generated by the ticket generating server present in the private cloud. In proposed system proof of ownership protocol has been applied, it will help to implement better security issues in cloud computing environment. Analysis validates that the proposed authorized duplicate check scheme incurs minimal overhead compared to other processes.

This Hybrid cloud module can be implemented in massive cloud storage environment with large amount of data and user access. Access control mechanism can be implemented to increase the security for the user data and network. By using access control mechanism unauthorized access to the cloud storage environment can be restricted.

REFERENCES

[1] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In Workshop on Cryptography and Security in Clouds (WCSC 2011), 2011.

[2] 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 ICDCS, pages 617– 624, 2002.

[3] J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Distributed Systems, 2013.

[4] W. K. Ng, Y. Wen, and H. Zhu. Private data deduplication protocols in cloud storage. In S. Ossowski and P. Lecca, editors, Proceedings of the 27th Annual ACM Symposium on Applied Computing, pages 441– 446. ACM, 2012.

[5] J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl. A secure data deduplication scheme for cloud storage. In Technical Report, 2013.

[6] J. Xu, E.-C. Chang, and J. Zhou. Weak leakage-resilient client-side deduplication of encrypted data in cloud storage. In ASIACCS, pages 195–206, 2013.

[7] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. IACR Cryptology ePrint Archive, 2013:149, 2013.

[8] J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication. IACR Cryptology ePrint Archive,2013:149, 2013.

[9] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[10] C. Ng and P. Lee. Revdedup: A reverse deduplication storage system optimized for reads to latest backups. In Proc. of APSYS, Apr 2013.

[11] Iuon –Chang Lin, Po-ching Chien ,‖Data Deduplication Scheme for Cloud Storage‖ International Journal of Computer and Control(IJ3C),Vol1,No.2(2012)

[12] Shai Halevi, Danny Harnik, Benny Pinkas,‖Proof of Ownership in Remote Storage System‖, IBM T.J.Watson Research Center, IBM Haifa Research Lab, Bar IIan University,2011.

[13] M. Shyamala Devi, V.Vimal Khanna,Naveen Balaji ‖Enhanced Dynamic Whole File De Duplication n (DWFD) for Space Optimization in Private Cloud Storage Backup‖,IACSIT, August,2014.

[14] Weak Leakage-Resilient Client –Side deduplication of Encrypted Data in Cloud Storage‖ Institute for Info Comm Research, Singapore, 2013.

[15] Tanupriya Chaudhari , Himanshu shrivastav, Vasudha Vashisht, ‖A Secure Decentralized Cloud Computing Environment over Peer to Peer‖,IJCSMC,April,2013.

[16] Mihir Bellare, Sriram keelveedhi,Thomas Ristenart,‖DupLESS: Server Aided Encryption for Deduplicated storage‖ University of California, San Diego2013.

[17] M. Bellare, S. Keelveedhi, and T. Ristenpart. ―Dupless: Server aided encryption for deduplicated storage‖. In USENIX Security Symposium, 2013.