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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Performance Measure of Deduplication in Hadoop Framework

K.Dhivya[1], Dr.K.Latha[2]

PG Scholar, Department of Software Engineering, Anna University, Trichy, Tamilnadu, India[1]

Assistant Professor, Department of Computer Science and Engineering, Anna University, Trichy, Tamilnadu, India[2]

ABSTRACT: Data are generated and updated tremendously fast by users through any devices in anytime and anywhere in big data. The exponential growth of data has brought a tremendous challenge on the storage system in Cloud computing. However, data duplicates need a lot of extra storage space and funding in infrastructure. So in addition to make the data secure and the delete the redundant copies of repeating data the data deduplication is used .Data deduplication is the technology which detects and eliminates redundant data in the dataset. Using the deduplication technique can improve utilization of the storage space effectively the proposed system can formulate a proper deduplication strategy to sufficiently utilize the storage space with Hadoop Distributed File System and it uses the data duplicates to increase data reliability. For deduplication the homomorphic algorithm is used .the homomorphic algorithm allows complex mathematical operation to be performed on encrypted data. And supports to store the encrypted data in public cloud and provides high security and this mechanism mainly allows operations to be performed on cipher text. Our deduplication strategy deletes useless duplicates to increase the storage space. The experimental results show that our method can efficiently improve the storage utilization of a data using the Hadoop Distributed File System.

KEYWORDS: Cloud storage, deduplication, Hadoop framework Storage Utilization.

I. INTRODUCTION

Cloud offers various services to the user. Data storage service provided by cloud is the most commonly used service given by cloud. Users are mainly used to upload data onto the cloud and let cloud to manage that data. Data may be in the form of files which can be personal or private. As user is storing data onto the cloud, user has to pay rent for storing data. Data storage rent may vary from different cloud service provider but as user is paying rent and if user is storing the same copy of data on cloud multiple times then the rent will go on increasing. In this paper, cloud based deduplication scheme is proposed. In which user will store the data on cloud only once same copy of data will not get stored again. Deduplication mechanism is proposed in this paper. As user is storing data onto cloud then user must be requiring security to data. Mainly For the security issues purpose data is get being stored on the cloud in the encrypted format. So deduplication checking has to be performed on the encrypted data. User will lead to store the data on the cloud in encrypted mechanism format and then it is checked that the data is already present on the cloud or not. If the data that user want to store data in the cloud is already present then duplication occur. User will also store data in encrypted format and with that data is in duplicate manner duplication check will also be performed on the data present on the cloud which is also the encrypted data. It is complicated for the data holder to check the deduplication on encrypted data. In this scheme the Homomorphic algorithm is used to manage encrypted data storage with deduplication. By using this technique the rent that user has to pay for data will reduced as the same copy of data will not be allowed to store onto the cloud. With reducing the cost it also provide security to the user as data will store in encrypted format.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

II. RELATED WORK

Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng [1], proposed a system the cloud computing offers many services by rearranging the resources, and the most important characteristics of cloud computing is data storage. And in this paper it suffers from security weakness, here the deduplication is mainly done by proxy re encryption and ownership challenge. In this we cannot access the data when the data holder is offline and it is more complicated to perform deduplication and further still difficult in generating the encryption keys.

Pasquale Puzio, Refik Molva, Melek Onen, Sergio Loureiro [2] proposed a system depends on block level based deduplication mechanism and providing data confidentiality at the same time. Aim of the system is to identify identical data and store them only once. The result of encryption is mainly to make the encrypted data copy which cannot be differentiate after being encrypted. In the process of deduplication it is difficult to identify the same data segment. So they have used convergent encryption in which encryption key is usually the result of hash of data. In this process the block-level deduplication and data confidentiality is mainly focused. Using Block-level deduplication makes the system more flexible and efficient. ClouDedup preserves confidentiality and privacy even against potentially malicious cloud storage providers thanks to an additional layer of encryption. ClouDedup offers an efficient key management solution through the metadata manager; the new architecture defines several different components and a single component cannot compromise the whole system without colluding with other components. ClouDedup works transparently with existing cloud fully compatible with standard storage API.

Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee [3], and Wenjing Lou proposed a system, in which they have used convergent encryption technique to encrypt the data, Data duplication is performed with rendering confidentiality of data. They also mainly present several new process deduplication schemes to perform duplicate data check in a cloud. In their system, they have used hybrid cloud architecture consisting of a public cloud and a private cloud. The private cloud is involved as a proxy to allow data owner to securely perform duplicate check with differential privileges. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud. The system mainly supporting the different duplicate check is proposed under this hybrid cloud architecture, mainly to eliminate redundant copies where the S-CSP resides in the process of public cloud. The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.

Z. Sun, J. Shen, and J. M. Yong [4] proposed a system, which consist of a front-end deduplication and Hadoop Distributed File System at the front end, it has a deduplication application. At the back end, there are two main components, which are HDFS used as a mass storage system and HBase, used as a fast index. Executed results were obtained from simulation using VMware to simulate a cloud environment and execute the application on the cloud environment.

Mihir Bellare1, Sriram Keelveedhi2, Thomas Ristenpart [5] proposed, a new cryptographic primitive method, Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is derived from the message. On the practical side, they provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. On the side the challenge is base on the standard model solutions, and they make certain process of connections with deterministic encryption mechanisms, hash functions secure on correlated inputs and the then extract the paradigm to deliver the schemes under different assumptions and for different classes of message sources.

T. Y. Wu, J. S. Pan, and C. F. Lin [6] proposed a system, in which to reduce the workload due to duplicate file, proposed the index name server (INS) to manage not only file storage, data de-duplication, optimized no selection, and server load balancing, but also file compression, chunk matching, real-time feedback control, IP information, and busy level index monitoring. To manage and optimize the storage nodes based on the client-side transmission status by proposed INS; all nodes must give optimal performance and offer suitable resources to clients. In this way, not only can the performance of the storage system be improved, but the files can also be reasonably distributed, decreasing the workload of the storage nodes.

Ruay-Shiung Chang, Chih-Shan Liao, Kuo-Zheng Fan, and Chia-MingWu[7] proposed a system in which to reduce the duplicated copied by using Hadoop distributed file computing technique has brought some great benefits. However, with the arrival of big data, some difficult issues have come to light. In this paper, they focused on one of the important problems which is about storage space in which HDFS plays a principal role in cloud computing, but which is also a cause for concern because the copy rule of HDFS needs quadruple storage space just for saving a file, and the added replicates will occupy most of the storage space. Actually, this is an expensive cost, especially in a petabyte or greater

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

scale, because the retained data; the probability of redundant data becomes higher.

III. SYSTEM DESCRIPTION

In the real world, the amount of data is growing exponentially; small and medium enterprises or educational organization will suffer from insufficient space problem. Therefore, in this paper, to face this challenge for storage systems, deduplication mechanism to improve the utility of storage space in HDFS, which is currently one of the best solutions for big data. Then, the proposed system architecture will be presented as the following. In this section, we will make a description of the system framework and the details of the components, algorithms separately. It will make a description of the system framework and the details of the components, algorithms separately. This study mainly describes the mechanism of homomorphic algorithm, which is mainly work on encrypted data. Hadoop Distributed File System At the front end, it has a deduplication application. At the back end, there are two main components, which are HDFS used as a mass storage system and HBase, used as a fast index. All results were obtained from simulation using VMware to simulate a cloud environment and execute the application on the cloud environment based on the Hadoop file system. Thus the deduplication is mechanism which is used to reduce the deduplicated of redundant data. Moreover and increase the bandwidth efficiency, and improves the storage Utilization.

A. The Overview of the System Framework:

The users want to upload the data into the storage system from their personal devices or other servers, and then the storage system stores this data via two-tier deduplication, which runs on two different components, which are prefilter and postfilter, respectively. In fact, they also run different levels of deduplication. The prefilter is responsible for the deduplication in the file level. The reason is that when a file is already copied with triple duplication in HDFS to provide data reliability and to further save the storage space, storing redundant data repeatedly in the storage system is unnecessary and we can also gain the great benefit of time saving, because the duplicate file will not be processed in detail. However, the different files may include the same blocks (after chunking with HDFS). The deduplication on a block level in our system, which runs on Postfilter and which is responsible for eliminating the redundant blocks. Due to the triple duplicates in HDFS keeping per block of a file in quadruplicate (every block with one original and triple copies), even though the files are homoplastic (only a few blocks are different), HDFS still writes triple copies per block of every file. Therefore, the storage space can be saved by removing the same blocks of similar files. By doing so, the storage space can be improved more efficiently. Next, only the data from the data center to the other additional storage spaces which supports the offsite backup should be backed. To develop a cloud based application in which documents will be uploaded only once to reduce cloud server space as well as platform rent. To implement security of document on cloud server with the help of encryption techniques To implement document destruction technique to enhance security of important documents as well as to reduce the rent of cloud platform. To integrate cloud data de-duplication with access control. Existing solutions of encrypted data de-duplication suffer from security weakness. It cannot mainly support the data access control and revocation. Here the scheme is mainly enhance de-duplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

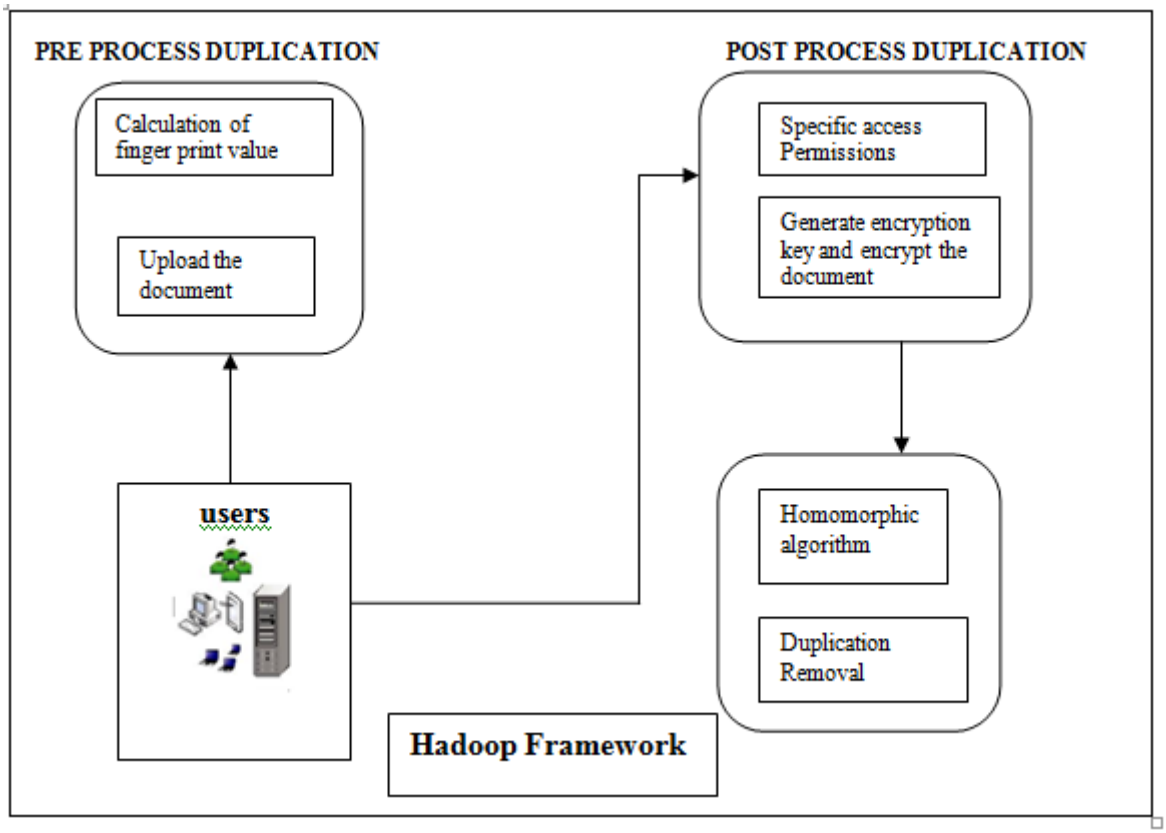

Fig1. Proposed system Architecture

B. The Situation of Operation in the Prefilter:

To avoid storing the redundant files in the storage space, the main task in prefilter is to find the same file which has been saved in the storage system. In fact, the system is divided the HDFS into two tiers, with the prefilter as the first level which combines the chunk component of HDFS. Therefore, during the operation of the prefilter, all the data is forwarded in the file type. However, the remaining components in the prefilter are file filter and metadata. The file filter is responsible for the input data which comes from users, and the metadata component records the information on the data in a table for identifying whether the file is redundant or not. Then, before the file is forwarded to the data center, the file will be divided into several blocks by the chunk component. When users want to save the data in the storage system by inputting the data, then, in the prefilter, the file filter will first check the storage to see whether saving this data will cause an overflow. If there is not enough memory to save this data, then the file filter will return the message to deny the request from users. Otherwise, the file filter will examine the data to see whether there is a duplicate with the attributes of a file from the metadata table. If it is redundant, then the data will be dropped from the file filter. Otherwise, the data will be divided into several blocks, and, at the same time, the attributes of the file will be updated in the metadata table. Finally, the data is forwarded to the data center in blocks. As in the above design, our deduplication of file level is according to the attributes of the data, whether any collisions will take place or not, rather than relying on the fingerprint of the file. The reason for this is to forward the new data without any delay. Comparing the fingerprint with the attribute, for the fingerprint, the system needs to first calculate the fingerprint, so the data has to wait until the computation has finished. On the other hand, for the attribute, no matter how large the data is, we only need to fetch the properties such as file name, user, built time, modified time, and size. Therefore, the computing time for the attribute is smaller than the time for the fingerprint. However, users will not create the same file with the same name at the same time theoretically. Thus, the deduplication of file level using identification of the properties of a file is feasible. By doing so, a new file will be written in our storage system only once.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

C. The Function of the Postfilter:

Mainly By deduplication, system can save some computing time before the redundant files are written into the system and thus make the usage of storage space more effective. However, only the file-level deduplication is still insufficient, because there is a high probability of redundant data that takes place between different but similar files; the intention is especially for updating the data. Traditionally, for the acceptable data, the storage system will follow the copy rule of HDFS, where the system can adopt the original data only for 25% of the total storage system with no extra storage space, because quadruple storage space is needed for saving a file in HDFS. The triple duplicate method in HDFS is designed for providing data reliability, but it also influences the usage of the storage system. Therefore, we need the further block-level deduplication. the components after the prefilter are data center, Postfilter, and metadata server. The data center is responsible for storing data and all the metadata are collected in the metadata server. To avoid the redundant blocks being kept in the storage space, the principal duty of the Postfilter is to make the same block to be wiped out which has been stocked in the storage space. However, the Postfilter is the second level of the divided HDFS structure which is responsible for the management of blocks. Besides, we examine the duplicate block with the fingerprint which is calculated by hash function. when the data received from the prefilter has been divided into several blocks, the fingerprint of the block in the Postfilter will be calculated with the hash function T hen, we seek out all of the f fingerprints in the metadata server to confirm whether they are duplicate blocks or not. If there is any block with the same fingerprint, then the duplicate block is dropped. Otherwise, it is forwarded to the data center and the blocks are saved in the Data Nodes. Next, to offer data reliability, the duplicate blocks will be created as HDFS. Finally, according to the current situation of the storage space, there will be distinct deduplication decisions. the deduplication of block level is according to the fingerprint of the data to see whether any collisions will take place or not. If there are any collisions, then the block will be dropped out by the Postfilter.

D. Homomorphic Encryption

Homomorphic encryption scheme allows one or more plain-text operations (e.g. addition and multiplication) to be carried out on the cipher texts. If the addition operation is allowed, then the scheme is known as additive homomorphic encryption. If the multiplication operation is allowed, then the scheme is known as the method multiplicative homomorphic encryption. In an additive homomorphic encryption scheme, the cipher-text of the sum of two plaintexts, $m_1 + m_2$, can be obtained using some computation " •" on the cipher texts of m_1 and m_2, without first decrypting m_1 and m_2 or requiring the decryption key. Additive homomorphic encryption also allows the user to obtain the cipher text of $m_1 \times m_2$ by performing m_2 times of "•" computation on m_1's cipher text. let $E_{P\,K}$ () be the function of encrypting with the public key, and " •" is modular multiplication .

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Homomorphic Algorithm for Deduplication:

```
//Input: data in the type of block from pre-filter
//Output: several blocks or no output to data center
//all the fingerprint of stored blocks in the metadata server
While receive the data from pre-filter
Compute the fingerprint of data with hash function;
For seek out the index table
If there are any collision between the fingerprint and index table
Then drop out the request of save data;
  Else forward the block to a Data Node;
If there are any successful for saving the block
 Then fetch the utility of storage space;
 Switch (the utility of storage space)
 Case of the utility storage space is greater or equal to h h 1
If the number of copies of any cold data is greater or equal to one
 Then remove all of the copies of per cold data
                Break;
 Case of the utility storage space is greater or equal to h h 2
If the number of copies of any hot data is greater or equal to one
 Then eliminate one of the copies of those hot data
                Break;
                Default
        Following the copy rule of HDFS
    But run the De-duplication of file and block level;
                Break;
            End While
```

Fig2. Algorithm for Homomorphic encryption

IV. SIMULATION& RESULTS

In the proposed method, the deduplication is implemented in the Hadoop framework and mainly the performance of the deduplication is observed over the different operating Systems like Windows 7, Windows 8, and Windows 10.And the deduplication mechanism based on Hadoop framework using the homomorphic encryption algorithm, and certainly works on the original form of the data. Mainly these mechanisms help the storage space of the user and reduce the bandwidth utilization and reduce the cost of the user.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Performance Measure of Deduplication

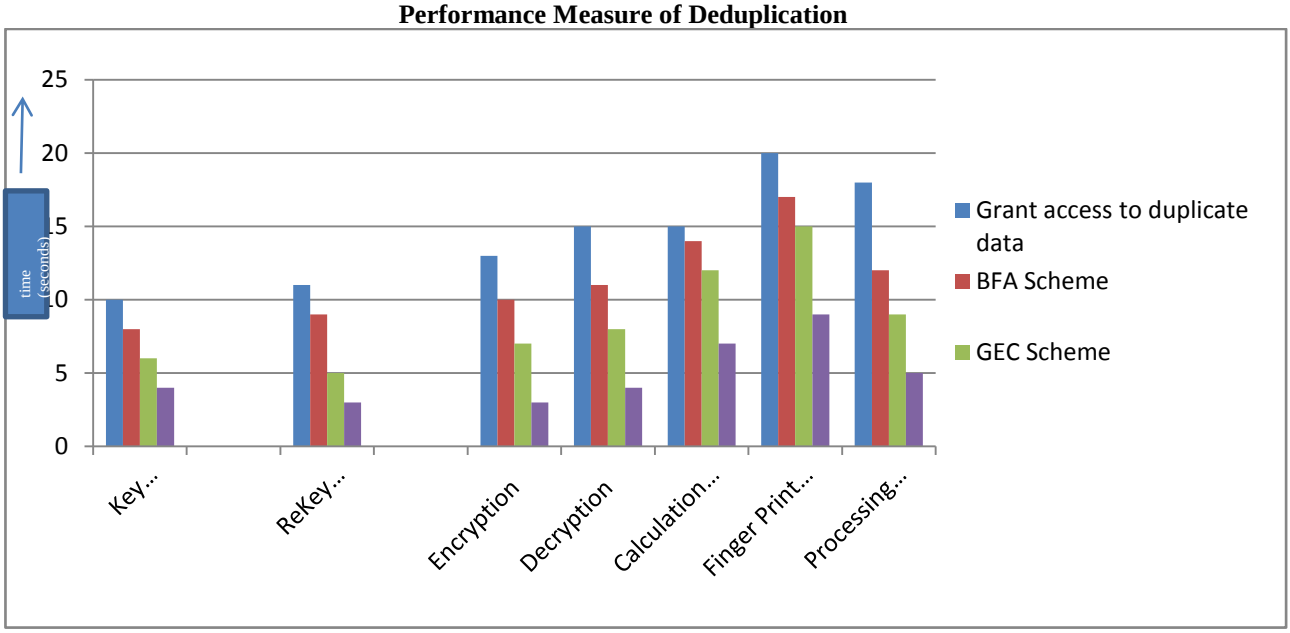

Table1. Performance Measure of Deduplication

V. CONCLUSION AND FUTURE WORK

In the past few years, the cloud computing technique has brought some great benefits. However, with the arrival of big data, some difficult issues have come to light. In this paper, one of the important problems which is about storage space in which HDFS plays a principal role in cloud computing, but which is also a cause for concern because the copy rule of HDFS needs quadruple storage space just for saving a file, and the added replicates will occupy most of the storage space. Actually, this is an expensive cost, especially in a petabyte or greater scale, because the retained data; the probability of redundant data becomes higher. Therefore, proposed a dynamic deduplication decision maker to improve the usage of storage space, proposed method, mainly according to the current situation of systems, forms a suitable solution. Our system is aimed at small enterprises and organizations, especially people with interests or educational environments that can more effectively use storage space by removing duplicates without investing any additional funding in infrastructure. However, the evaluation of our proposed system shows that our method can save more data than the others.

REFERENCES

[1] "Deduplication on Encrypted Big Data in Cloud", Zheng Yan, Senior Member, IEEE, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H. Deng, Fellow, IEEE, IEEE transactions on big data, vol. 2, no. 2, april-june 2016 .
[2] "ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage", Pasquale Puzio, Refik Molva, Melek Onen, Sergio Loureiro, 2013 IEEE International Conference on Cloud Computing Technology and Science .
[3] "A Hybrid Cloud Approach for Secure Authorized Deduplication", Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou, IEEE transactions on parallel and distributed systems, vol. 26, no. 5, may 2015.
[4] Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative Work Des., 2011, pp. 348–355.
[5] "Message-Locked Encryption and Secure Deduplication", Mihir Bellare1, Sriram Keelveedhi2, Thomas Ristenpart3, proceedings of Eurocrypt 2013.
[6] T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes, "IEEE Syst. J., vol. 8, no. 1, pp. 208–218 Mar.2014.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

[7] W. K. Ng, Y. Wen, and H. Zhu, "Private data deduplication protocols in cloud storage," in Proc 27th Annu. ACM Symp. Appl. Comput.,2012, pp. 441–446.
[8] Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data dedup scheme in cloud computing," in Proc.Int. Conf. Syst., 2014, pp. 85–90.
[9] C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp.250–262.
[10] C.Yang, J. Ren, and J. F. Ma, "Provable Ownership file based duplication", in pro IEEE Global Commun,2013, pp.283-290..