ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

"Survey of Various Data Deduplication Method and Security Algorithm Issue in Cloud Computing"

Gauravkumar Pralhad Patel[1], Prof. Kailash Patidar[2],

PG Scholar, Department of CSE, SSSIST, Sehore, India[1]

HOD CSE/IT, Department of CSE,SSSIST, Sehore, India[2]

Abstract- Cloud computing has become a widely adopted platform for large-scale data storage and management. However, the rapid growth of data has led to the problem of duplicate data storage, which increases storage requirements, reduces server efficiency, and degrades overall system performance. This paper proposes a de-duplication framework that enhances cloud storage performance by identifying and eliminating redundant data before it is stored on the server. The proposed approach employs the Term Frequency–Inverse Document Frequency (TF-IDF) technique for feature extraction and an inverted index for efficient duplicate detection and retrieval. By storing only unique data instances, the system optimizes storage utilization, reduces redundancy, and improves search efficiency. Experimental results demonstrate that the proposed method improves server performance by reducing de-duplication time, increasing storage efficiency, and enhancing search accuracy in terms of precision and recall. The proposed framework provides an effective solution for cloud data management and can be extended in future work by integrating advanced de-duplication techniques with secure data transmission and encryption mechanisms to further improve security and time efficiency.

Keywords- Cloud Computing, Data De-duplication, TF-IDF, Inverted Index, Cloud Storage, Information Retrieval, Precision, Recall, Storage Optimization, Search Performance.

I. INTRODUCTION

Cloud term is denoted as such type network where information hide such as internet service provider network .and computing term is denoted such as data processing and computing to solve the problem in computer. Data processing apply any where in computer or any computing device shuch as single computer or any server device.In network number of server availability and different category severs is dependent on website data. We Computers Internet services and communication peripheral devices application are deplovping very fast and use of cloud computing has increased rapidly in today's world. Reducing cost is major concept of internet and using single server cost is so high so we use shred hosting ,virtual private server and dedicated server to reduce cost of internet connectivity .Content play a major role in internet a lot of content will connected more number of people and increases services load of internet and to improve content we introduce free internet service to allowed user for content writing by using email, blogs, Wikis, Forums , Chat, Social Networking, online shopping.Cloud model provide on demand network services to servers, storage, applications, and services. Today scenario almost data is stored in digital form due to enhancement in networking and storage technology. to be utilize data we stored in the cloud space because disk space is costly to stored and maintain. Cloud Computing is the latest trend in computing service provision. This kind of service has

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

seen this technological in addition to cultural adjust of computing service supply from being provided in your community to being provided remotely by third-party companies . Functionality such as hardware, processing and also other functionality has become offered on demand, as a service in addition to both freely and from cost. The consumer has effectively lost control over how their data is being stored, shared and used, plus over this security utilized to protect the data.

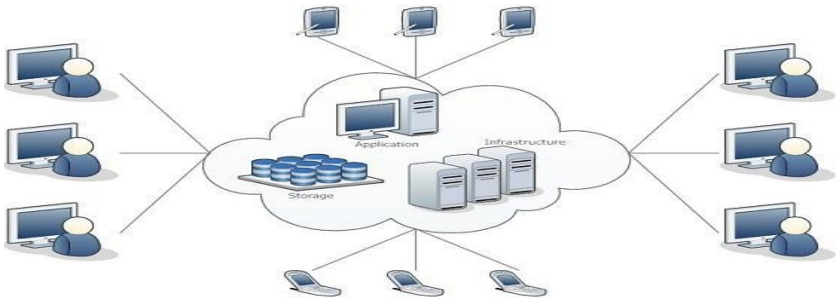

Fig.1 Cloud Computing Model

Impair computing is frequently a computing model where hardware, platform, communications and also software are defined and also delivered being a service rather than product. Cloud research takes advantage of hardware virtualization to be able to securely and also dynamically allot physical resources including computational energy, storage, and networks on the users. Cloud assets are delivered to the end-users through Web services.

II. LITERATURE REVIEW

Data de duplication in public impair by division of cloud good geographical location". Data de duplication can be a method of controlling the traffic in a cloud environment. Cloud applications hunt for resources for execution. The resources are usually storage, processing, bandwidth, etc. Allocation these resources efficiently to all the competing jobs. In this kind of paper, we describe data deduplication and balancing in a public cloud by partitioning this cloud into several sub-clouds. This division of public impair into several sub-clouds is done good geographical location

[1]. Cloud provide great achievement, it also introduces a many of security threats to the information and data which is now being ported from on-premises to off-premises. This paper discusses the various service models such as IaaS, PaaS, SaaS. Some products offer Internet-based services—such as storage, middleware, collaboration, and database capabilities—directly to users such as Iaas,Paas,Saas.Most cloud computing systems in operation today are proprietary**,** rely upon infrastructure that is invisible to the research community, or are not explicitly designed to be instrumented and modified by systems researchers. This paper also explores the types of cloud computing environment. Paper also discuss the cloud database including deployment model, characteristics.Also it identifies some technological and legal issues in cloud computing**.**

[2]. Compute clouds are enormous server farms packed with computing power and storage space accessible through the Internet. Instead of having to manage one's own infrastructure to run applications, server time and storage space can be bought from an external service provider. From the customers' point of view the benefit behind this idea is to

be able to dynamically adjust computing power up or down to meet the demand for that power at a particular moment. This kind of flexibility not only ensures that no costs are incurred by excess processing capacity, but also enables hardware infrastructure to scale up with business growth. But security is very big concern in the data sharing in the network. Same as in cloud computing data security is also a very sensitive matter. This paper discusses the concept of Cloud computing to achieve a complete definition of what a Cloud is and its security issues, More than 20 definitions have been studied allowing for the extraction of a consensus definition as well as a minimum definition containing the essential characteristics. This paper pays much attention to the security issues of Cloud computing..

[3] The Attribute based accessto extensible media in cloud to help sharing the contents to the network.The data encryption standard cipher text policy algorithm is use. It's a important to protect the data from unauthorized access. Media contents such as images, audio, video, and text etc..These contents are protected by this security method. . All the storage is done on cloud so the security of data is threatened. So there is need to secure the data on the cloudThe public cloud services are sent to the client by internet. The load balancing technic is use in cloud computing.Cloud Load Balancing simply means doing load balancing, i.e. allocate workload across different servers, network channel, CPU, disk drives, on the cloud, i.e. inside the Internet itself. This method is use helps to obtain , as other load balancing methods, helps to achieve optimal system utilization, maximumthroughput, minimum response time, and prevent overload of system.,

[4] Privacy security is a key issue for cloud storage. To solve this problem, the paper1 proposes a privacy-preserving cloud storage framework, which includes the design of data organization structure, the generation and management of keys,the treatment of change of users' access right and dynamic operations of data, and the interaction between participants. Wedesign an interactive protocol and an extirpation-based key derivation algorithm, which are combined with lazy revocation,multi-tree structure and symmetric encryption to form a privacy-preserving, efficient framework for cloud storage. Asystem is realized which is based on the framework. The paper analyzes the effectiveness of extirpation-based key derivation algorithm, the overhead of the system and the privacy security of the framework. Finally, we summarize our work and introduce our future research directions.

III. PROPOSED WORK

In our proposed system multiple data and documents are stored by data owner. Documents where stored in cloud server.CSP cloud server provider where detect duplicate data using inverted index technique. Once data de duplication is checked then markle hash tree generate hash function to generate secure key and encryption and decryption perform at user end.our proposed system contain following term describe below

- **Inverted Index:** We proposed a inverted index based data de duplication technique. In this data de duplication scheme overcome the problem of searching duplicate data in data set. Our scheme support conjunctive multi key world searching. Inverted index is a key data structure underlying modern Information Retrieval. For each term t , we must store a list of all documents that contain term – identify each documents by a docID , a documents serial number .We used fixed size array for this.

So we need variable size postings is normal and best .on disk , a continuous run of postings is normal and best . in memory can use linked list or variable length array .

Inverted index construction :-

Documents, tokenizer (TOKEN STREEM), logistic modules(MODIFIED TOKEN) indexer (INVERTED INDEX)

Initial stage of text processing:-

Tokenization:-cut character sequence in to world tokens .

Normalization:- map tex and query term to same form ex you want U.S.A and USA

Stemming:- we may wish to different forms of a root to match ex authorize authorization

Stop world : we may omit very common words (or,not) ex the, a,to of .

- **TF-IDF:** Tf-idf is best known weighting scheme in information retrieval scheme. In our proposed algorithm we are applying TF-IDF technique in Inverted index to find out number of times of a world appears in the documents .TF-IDF stands for term frequency-inverse document frequency

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- **Merkle hash tree :** Our proposed system work with Merkle Hash Tree is a well-studied authentication structure [7]. It is used to efficiently prove that a set of elements are undamaged and unaltered. It helps greatly in reduction of server time [9]. It is used by cryptographic methods to authenticate the file blocks. The leaf nodes of the MHT are the hash values of the original file blocks. The idea behind generating MHT is to break the file into a number of blocks. Apply hashes to the authentic data values i.e. the original file blocks and combine iteratively. Now, rehash the result hash nodes and combine in a tree-like fashion and repeat this procedure till we get a tree with a single root. The MHT is generated by the client and is stored at both the client and the server side. Fig 2 depicts an example of MHT. The tree has four leaf nodes viz. m1, m2, m3 and m4. Initially, we apply hash on each of these file blocks and obtain h(m1), h(m2), h(m3) and h(m4). Then, h(m1) and h(m2) are hashed and combined together to get ha. Similar process happens with blocks m3 and m4 and here, we get hb. Here, h is a secure hash function. This can be expressed as ha = h(h(m1)|| h(m2)) and hb = h(h(m3)|| h(m4)) Further, ha and hb are combined and rehashed to obtain the root as hr. This can be expressed as hr= h(ha|| hb)
- **AES:** AES is a block cipher. The algorithm supports a variety of key sizes as 128,192 or 256. The default size is 256 bits. The encryption of data blocks is done in 10, 12 and 14 rounds depending on the size of the key used. It provides fast and flexible encryption and can be easily implemented on various platforms. In this paper, AES-128 is used and so encryption is done in 10 rounds. This algorithm is used for both encryption and decryption. For encryption, it takes data blocks and the secret key as the input and outputs the encrypted data blocks. For decryption, encrypted data blocks and key are given as inputs and original file blocks are the output.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

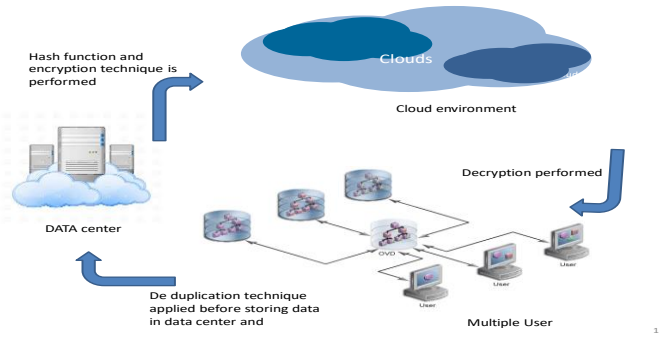

Fig.2 work flow diagram

Proposed Algorithms:

Calculations: Setup, Key Generation, encryption and decoding.

Setup: The setup algorithmic project takes no info barring the understood security parameter.

Work flow process:-

Input:Web page documents is provided as a input to read Document is denotes as term D. data is store in data center and all the web pages is available in cloud storage so for all documents we required Cloud storage S .and Inverted map IM as a input used to calculate inverted index which is designed to allow very fast full-text searches. An inverted index consists of a list of all the unique words that appear in any document, and for each word, a list of the documents in which it appears.

1 We have to read R= read Document(D) from the documents which is available in cloud storage .

2 Now we extract Text Features (R)= E

Tf-Idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

3 Encrypt data (R)= E AES encryption is performed over data.

4 Sp[] = E.splite File (E)

Generate has function and check the value of has function to performed this function till end loop,in these step massage is encrypted and secure for communication

5 $Sp[] = E.splitFile(E)$

6 $for(i = 0; i \leq Sp.length; i++)$

- a. $H = genrateHash(Sp[i])$
- b. $if(H != HashTree.node)$
  - i. $HashTree.createNode(H)$
- c. Else
  - i. Remove H
- d. End if

7 End for

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

IV. RESULT ANALYSIS

The implementation of the proposed Multiple Replica Cryptography Technique is described in Chapter 3. Thus it includes the detailed understanding about experimental evaluation and performance of the proposed system. The essential parameters are used for evaluation are listed with observed values.

Uploading Memory Consumption

The amount of main memory required to execute the algorithm with the input amount of data is known as the memory consumption or space complexity. The total memory consumption of the algorithm is computed using the following formula.

$$\text{Consumed Memory} = \text{Total Memory} - \text{Free Memory}$$

Table.1 Memory Consumption

Number of Runs	Proposed Technique
1	112712
2	101772
3	109178
4	117089
5	97833
6	12257

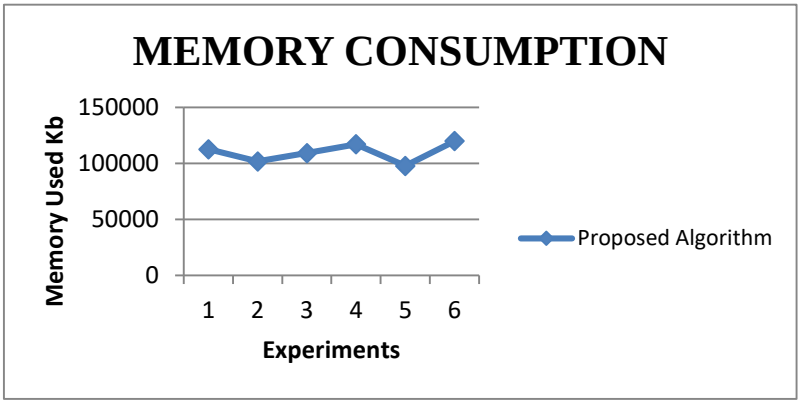

Fig.3 Memory Used

The figure 3 and the table 1 show the memory or space complexity of proposed cryptographic approach. In this diagram the amount of main memory consumed in terms of kilobytes (KB) is given in Y axis and the number of experiments are reported at X axis. According to the obtained results the proposed algorithm consumes lesser resources and gives better performance of the encrypted and decrypted file.

Uploading Time Consumption

The amount of time required to develop the upload a data file on the server for cryptographic model is termed as the time complexity of the algorithm or time consumption of system. The time consumption is given using following formula:

$$\text{Time consumption} = \text{End time of file put on Server} - \text{Start time of File put on data model}$$

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

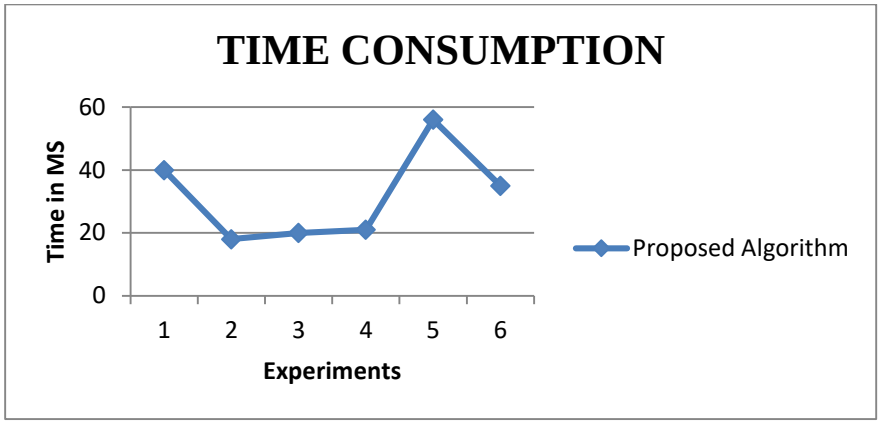

Fig.4 Uploading Time Consumption

Table .2 Uploading Time Consumption

Number of Runs	Proposed technique
1	40
2	18
3	20
4	21
5	56
6	35

The figure 4 and table 2 shows the amount of time consumed during uploading a file on server. In this graph the blue line shows the performance of proposed multiple replica cryptographic approach. The X axis show the different number experiments to analysis the model performance with respective to choose different data files and Y axis shows the amount of time consumed during the encryption process.

4.3 Downloading Memory Consumption

The algorithms need a significant amount of main memory to store the data for processing. This storage requirement is termed as the memory consumption or the space complexity of the system. Here the downloading based memory consumption is computed.

Table.3 Memory Consumption

Number of Runs	Proposed technique
1	**112712**
2	**101772**
3	**109178**
4	**117089**
5	**97833**
6	**12257**

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

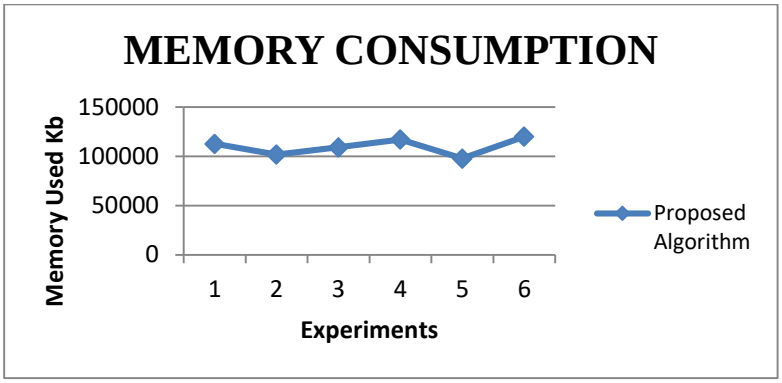

Fig.5 Downloading Memory Used

The memory consumption of the proposed multiple replica cryptography is demonstrated using the figure 5 and table 3. In this diagram the X axis shows the list of experiments and Y axis shows the memory consumption of the implemented system in terms of KB (kilobytes). According to the obtained results the performance of system is depends on the amount of file for process. Thus as the file size increases the required memory is also increases. According to the system performance the proposed technique is adoptable due to less amount of memory consumption.

Downloading Time Consumption

The computational algorithms need an amount of time for producing the outcomes. Here downloading time is the time required by the server to do download the data file on the user system. That is computed using the following formula.

$$\text{Time Consumption} = \text{End time of File put on system} - \text{Start Time to download file by Server}$$

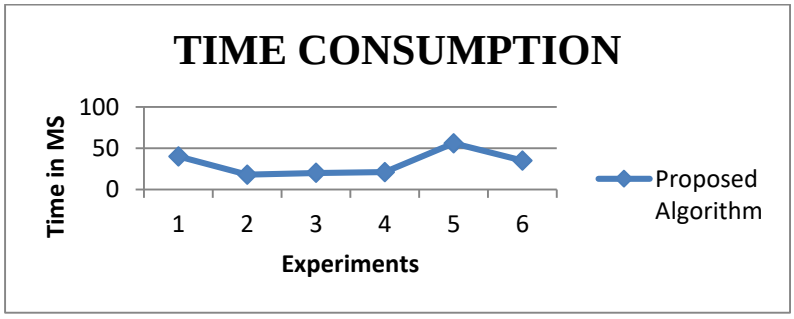

Fig.6 Downloading Time Consumption

Table.4 Downloading Time Consumption

Number of Runs	Proposed technique
1	40
2	18
3	20
4	21
5	56
6	35

The downloading time complexity of the proposed cryptographic technique is given using figure 6 and table 4. In this diagram the X axis represent the number of experiments and the Y axis shows the amount of time consumed for downloading the file on system. According to the obtained results the performance of the system is depends on the amount of file size. As the amount of file size is vary the amount of time for downloading is vary.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT

Server Response

The amount of time required to produce the outcome after making the request from the server is termed as the server response time. The response time not included the encryption or decryption activity during these measurements.

Table.5 Server Response Time

Number of Runs	Proposed Technique
1	5
2	4
3	1
4	0.48
5	0.49
6	2

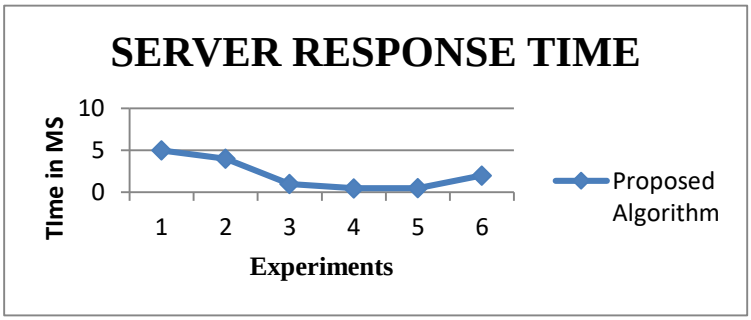

Fig.7 Response Time

The computed response time of the proposed technique for cloud based secure communication is demonstrated using the figure 4.5 and table 4.5. The X - axis of this diagram contains the amount of experiments performed using the system and the Y axis shows the amount of time required for generating the response through the server for traverse the hash tree. That can also term as the communication overhead for the system. According to the computed results the response time is not depends on the amount of file size or other parameters. That is directly depends on the amount of work load on the target server where the data is stored or the application is hosted.

V. CONCLUSION AND FUTURE SCOPE

Cloud computing enhance its performance using this proposed system. This proposed system help to improve performance of data storing by removing data duplication on server. Major solution provided by proposed solution to be find duplicate data. Duplicate data storage is a major issue for any data base system. Duplicate data store multiple copy of same information so storage is not use by some other user required for storage more information in to server or any data storage. In our propose system same copy of data is removed buy using tf-idf technique and using inverted index in our proposed work.We simulate that when we use inverted index to prevent duplicate data . Performance of system increases more as before. Server access time is increases . Time to make search with their relevancy in terms of precision and recall and De-duplication time required with the amount of files In future we work with more

duplication technique with secure data transmit ion . we will focus to apply de duplication with efficient security and focus on time efficiency .

REFERENCES

[1] **Rahul Bhoyar Prof. Nitin Chopde M.E (Scholar) M.E (Computer Engineering) "Cloud Computing:Service models,Types,Database and ssues" IJACCSEE Volume 3, Issue 3, March 2013.].**

[2] **Neeraj Shrivastava and Rahul Yadav IES, IPS, Academy Indore, MP, INDIA. " A Review of Cloud Computing Security Issues"[International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 1, July 2013**

[3] **Tejashri Khandve, Megha Talekar , SheetalDhiwar ." Security and Load Balancing In Cloud Computing" [International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 10, October 2015**

[4] **RuWei Huang1,2 Si Yu1, "Design of Privacy-Preserving Cloud Storage Framework", (2010 Ninth International Conference on Grid and Cloud Computing IEEE**

[5] **M.Thamizhselvan R.Raghuraman, "A NOVEL SECURITY MODEL FOR CLOUD USING TRUSTED THIRD PARTY ENCRYPTION", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15**

[6] **DUTCH T. MEYER, The University of British Columbia, Microsoft Research WILLIAM J. BOLOSKY, "A Study of Practical Deduplication" , ACM Transactions on Storage, Vol. 7, No. 4, Article 14, Publication date: January 2012..**

[7] **Mr. Avinash R. Dhok Ms. Ashwini P. Kolhe, "A Survey on Scalable Data Security and Load Balancing in Multi Cloud Environment", IJIRST –International Journal for Innovative Research in Science & Technology| Volume 1 | Issue 8 | January 2015 ISSN (online): 2349-6010**

[8] **Riddhi Movaliya Department of Computer Engineering PIET, Harshal Shah "A Survey of Secure Data DeduplicationInternational Journal of Computer Applications (0975 – 8887) Volume 138 – No.11, March 2016.**

[9] **Mr. Yendhe A.1, Ms. Dumbre T.2, Ms. Mahadik S.3, Ms. Gholap A.4, Prof. Gunjal A.5 "SURVEY ON SECURE PRIVILEGED BASED DATA DEDUPLICATION IN CLOUD USING TWIN CLOUD", Vol-1 Issue-4 2015 IJARIIE-ISSN(O)-2395-4396**

[10] **M. Karthigha1* and S. Krishna Anand2, "A Survey on Removal of Duplicate Records in Database", Indian Journal of Science and Technology | Print ISSN: 0974-6846 | Online ISSN: 0974-564 Aprial 2013 IJST.**

[11] **Akhila Ka*,Amal Ganesha,Sunitha Ca, "A Study on Deduplication Techniques over Encrypted Data", Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016 doi: 10.1016/j.procs.2016.05.123.**

[12] **Fatema Rashid, Ali Miri, Isaac Woungang A Secure Data Deduplication Framework for Cloud Environments2012 Tenth Annual International Conference on Privacy, Security and Trust 978-1-4673-2326-0/12/$31.00 ©2012 IEEE 81**

[13] **International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438 Volume 4 Issue 8, August 2015 www.ijsr.net Licensed Under Creative Commons Attribution CC BY A Hybrid Cloud Approach for Secure Authorized Deduplication Jagadish 1, Dr.Suvarna Nandyal2**

[14] **Mr. Antony Xavier Bronson Research Scholar, Department of Computer Science and Engineering International Journal of Applied Engineering Research "Enhanced De-key Approach to Reduce Data De-Duplication in Cloud Storage"ISSN 0973-4562 Volume 11, Number 7 (2016) pp 5316-5320 © Research India Publications. http://www.ripublication.com**

[15] **JADAPALLI NANDINI1 , RAMIREDDY NAVATEJA REDDY2, IMPLEMENTATION OF HYBRID CLOUD APPROACH FOR SECURE AUTHORIZED DEDUPLICATION Volume: 02 Issue: 03 | June-2015 www.irjet.net p-ISSN: 2395-0072**

[16] **Sejun Song Baek-Young Choi Daehee Kim Selective Encryption and Component-Oriented Deduplication for Mobile Cloud Data Computing 2016 International Conference on Computing, Networking and Communications, Cloud Computing and Big Data IEEE 2016**