Special Issue on Proceedings of International Virtual Conference, Recent Materials for, Engineering Applications and Sustainable Environment, December 2020. International Journal of Aquatic Science, Volume 11, Issue 1, 2020.

# De-duplication in Centralized Cloud Server using Third party Auditing

S. Pathur Nisha[1], L.Nithya[2] and A.Geetha[3]

Dept. of Computer Science and Engineering

Nehru institute of technology, Coimbatore

Correspondent e-mail: nithi.be@gmail.com

***Abstract***

*Cloud computing is the fundamental for the next computing generation. Cloud computing is the collection of the resources, all the resources of cloud that are provided across the internet. For avoiding the unwanted space in the cloud storage the file system is necessary.de-duplication is most commonly used technique in cloud storage which allows only one copy of the file to store. De duplication is most useful in reducing the communication and the storage expenses in the client perspective as well as in the server. In this paper the main objective is to avoid the duplicated files in the centralized cloud server. A secure auditing method will convince the client that the storage of data will be secured in the cloud. Some security problems are raised between the cloud service provider and client, For that third party auditing has been introduced, the files are classified into four categories they are sensitive files, confidential files, private clouds, public clouds this will enlarge the maintenance and management of the large data in the cloud. While uploading the files in the cloud the most important files will be stored in the sensitive and confidential files and the less important files will be stored into the private and public cloud. The file which is stored in the sensitive and in the confidential file will be stored permanently, the files that are uploaded in the pubic cloud will be deleted after 6 months and files that are stored in the private files will be deleted after 3 months. The file segmentation is the technique has been involved in this for the maintenance of the files.*

***Keywords*****:** *Thirdparty auditing, file system segmentation, cloud computing,*

## 1. Introduction

The term cloud derived from telephony. The name cloud computing was motivated by the symbol of cloud that are often used to have the vote of diagrams and the flow charts The various computing technologies such as utility computing, grid computing, virtualization and parallel computing are collectively called as cloud computing. "One or more people using the computer simultaneously through internet is called cloud computing", the applications and the services that are advanced to the internet is cloud computing. It is not suddenly come into sight overnight, remotely the computer systems which enable the sharing of the resources to achieve the consistency and economies scale. The cloud computing can be accessed by the users with the network connected devices such as smart phones, tablets, laptops, desktops in that some are purposeless without it. Cloud computing is an innovative information system architecture, visualized as what may be the future of computing , a driving force demanding from its audience to rethink their understanding of operating systems, client–server architectures, and browsers. Cloud computing has leveraged users from hardware requirements, while reducing overall client side requirements and complexity.

In 1950's the mainframe computers came into existence, at the time so many users acquire the central computer through the dummy terminals, that may affect the cost of the

organization at that time the concept of purveying of distributed access to the single computer has been established.

In 1960's the idea of emerging the utility and grid computing was lasted up to the pre-internet era. Joseph Carl RobnettLicklider who is the founder of cloud computing. The cloud computing phase was originated in 1999. But in the year of 2007 cloud computing really classified as Iaas, Paas, and Saas. The cloud computing have two types of models, service models which cover the all kind of services like Platform as a service, Infrastructure as a service, and Software as a service. And the another model is the deployment model which consist of public cloud and the private cloud, hybrid cloud.

De-Duplication is removal of unwanted or redundant information from the storage. It is the endeavor of finding if there any same entries in to the storage. The major pros of the de duplication is it will decrease the traffic in the network by sending the distinctive data at the time of backup, And also it will optimize the media of storage by deleting the unwanted blocks of data.

Trusted third party auditing means a CSP (cloud service Provider) has been called and it will provide the storage and another CSP will provide the Security. The CSP which provides the Security will not been known about the data which is stored.

The major advantage of the cloud storage is the disaster recovery, reduction in expenses, and enlarge the storage will enhance the business.

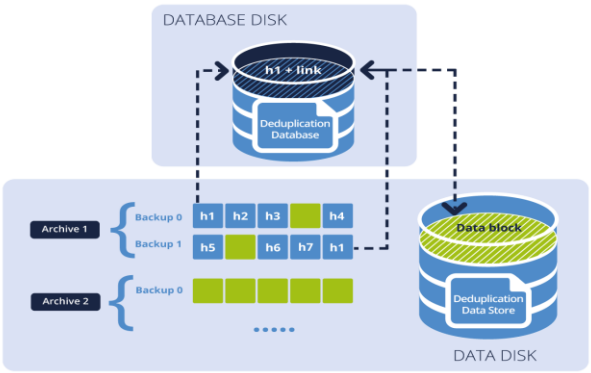



**Figure 1. De Duplication in Cloud**

## 2. MOTIVATION OF THE PROBLEM

In the recent days the companies like AMAZON, GOOGLE and etc, which is based on the cloud storage server are getting increased. The company is need to keep storage server for their own. Face book daily using minimum of 50TB storage, The organizations needs the cloud storage as 100 to 200 TB minimum, so much of the companies do not increase their cloud space because that is quiet expensive process but at the same time they need the maximum storage, since they be in need of maintaining the large files. So many cloud service for the storage will have the particular bandwidth allocation if there is any company that will extend the given allocation will be charged significantly.

## 3. LITERATURE REVIEW

D. Zissis and D. Lekkas, (2011) they had researched that the often transition in the cloud will Fuelled on critical issues. Many risks that are uncharted been introduced from the

relocation of the clouds, the main objective of this work is to evaluate the security of the cloud by recognizing the requirements of the security and the another is to present feasible solution that reduces these threats. In this they will be providing the trusted third party characterizes and also the regulate the cryptography, especially the Public key Infrastructure is utilizing with SSO and LDAP, for safeguard the authentication, Confidentiality and the integrity.

Q. Wang, C. Wang, K. Ren, W. Lou and J. Li,(2011) they focused on certain integrity of the cloud storage. The third party auditor will be placed on behalf of the cloud of the client, to check the integrity of the data stored in the cloud. Cloud computing are not only restricted for the backup of data. In this work they have improved the existed proof of the storage models by utilizing the Merkle Hash Tree Construction for the block authentication. Also they explored the method of bilinear cluster signature to enlarge the main result in many user setting, TPA will perform the auditing tasks simultaneously, so this will enhance the security and the performance of this proposed scheme. The scheme works efficiently and it is secure.

R. Buvya, C.S.Yeo, J. Broberg and I. Brandic, the great advancement in the ICT (Information and Communication Technology) computing will be one of the $5^{th}$ utility. The computing services are considered to meet the everyday needs of the community. The paper proposed to make the market oriented resource allocation by leveraging technologies as virtual machines. This is will insight on market based resource management strategies that maintain both the customer driven service management an the computational risk management to strengthen the service level agreement. They also highlight the differences between HPC workload internet based service workload. The data centers are mentioned and maintained across the clock by content providers. facility of commodity hardware to run the applications inward virtual machines very effectively by the Micro processor technology and the software, the isolation of the applications from the underlying hardware and the other VM. The VM will allow the end user themselves as a service to provide the best consumer service like the consumers can install their own applications.

Armbrust M, Fox A,Griffith R, Joseph A D, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stocia I, Zaharia M, “A View of Cloud Computing” the view that they have mentioned here as the innovation ideas doesn’t have enough capital to deploy the services and the human expenses to operate it, the best way is getting service from cloud. This paper makes the clear idea about the difference between the cloud and the conventional computing and also the different opportunities of the cloud computing. The cloud used to provide the services on the basis of pay per use, known as public cloud. The different computing offer the level of abstraction for the management of resources.

Zissis. D, Lekkas. D, “Addressing Cloud Computing Security Issues,” Future Generation Computer System, in main objective of this paper is that, it has two phases first is Cloud security identifying the unique security requirements and second is to a viable solution that will ignore the potential threats. The solutions are made by cryptography, a horizontal level of service, available to all suggested entities, an essential trust is maintained.

Wang. C, Wang. Q, Ren. K and Lou.W, “Privacy –Preserving Public Auditing For data storage security in Cloud Computing”, in this paper, Extensive security and performance analysis show the proposed schemes are provably secure and highly efficient proposed a secure cloud storage system supporting privacy-preserving public auditing. Further extend result to enable the TPA to perform audits for multiple users simultaneously and efficiently. Due to its long list of unprecedented advantages in the IT history: on-demand self-service, ubiquitous network access, location independent resource pooling, rapid resource elasticity, usage-based pricing and transference of risk.

Special Issue on Proceedings of International Virtual Conference, Recent Materials for, Engineering Applications and Sustainable Environment, December 2020. International Journal of Aquatic Science, Volume 11, Issue 1, 2020.

## 4. EXISTING METHODOLOGY

### 4.1 New Technology File System:

This is the NF fie system of Windows NT operating System this is used mainly for storing and retrieving. There are various drawbacks in the existing system. The transactions that are done by all the storage server is like retrieving, storing, management of data and the storage management. This file system is integrated file compression. The cloud storage is acquired by most of the companies to maintain and the security of the data. The cloud is more expensive to maintain, but the personal storage for the user are limited to 1 GB to 2 GB due to the cost. In that storage, most of the users are storing the duplicate data so this may lead the storage most complex and inefficient. The major problem with this system is replicated file storage, the usage of different file names to store the same data and older data or the unwanted data storage

The lack of file system may cause the problem with this existing method. Storing the entire files in the same location will create the most retrieval problems so at that time the complexity of searching will occur. In the case the users and the business organization need to buy more storage spaces and cloud servers.

This method is quite expensive for the middle level company or the individuals. And also in this multiple user interface is not allowed and also very important thing is not much secured.The major disadvantage of existing method is: file naming limitations, operating system incompatibility, space overhead.

Various duplication techniques had been introduced and there is no solution that has been developed for the avoidance of all kind of duplication. The duplication techniques been developed on the basis of the different characteristics of the data sets and also the capacity of the system. The data sets check the parts of the file content along with the previously stored file, this is done for the better storage of space, along with the system designs. If the system capacity is low then the system will have low quality de duplication design that will affect the performance of the system. The internet service providers are necessary for the fast processing of data at the routers. It is better to remove the data before storing it, if a system is idle it will store the data temporarily and the duplicated data will be removed within the temporary storage. In this work they used detect the duplication using the chunk index caches and bloom filters.

## 5. PROPOSED METHODOLOGY

The dynamic de duplication techniques has been introduced in the proposed system, the third party auditing has been established only when the organization decided to create a management system. Which will provides the formal assurance to the web applications that shares data centers. To point out the issue, this will merge the online measurements and the maintenance of the data. To arrest the impermanent behavior of the application workloads, modeling the resource of the server using the time-domain explanation of a generalized processor sharing server.

The arguments of this model are constantly upgrade using online monitoring and the remaining storage will accord the servers. The framework which used here will uses time sequential analysis methods to prognosticate the expected call of duty parameters from the plump system values. The constrained non-linear optimization method to forcefully allocate the resources of the server based on the requirements of the application. This transient

behavior of the application will shows the nonlinearity in the system model. The results of this method will assign the resources of the system, in over load conditions. Periodical execution done by the predictor for the resource allocation status used for the future demands.

The drawbacks in the existing system overcome in the proposed method. File system creation taken place in this project to make the server more powerful and efficient, this is done through the TPA (Third Party Auditing). All the file transactions are done through the TPA. TPA will be playing a vital role as a middleware between cloud storage and the user. All the data that is stored will be observed by the TPA, this will not allow any duplicated files to store. This method is applicable for images, document files and note pad files. Image will be calculated by the RGB calculator and document and note pad files are under gone the file pixel validation. It is introduced in the storage server. All details of the image will be taken as the reference file, in case of replicated files means, the TPA doesn't allow to store, also the file system is separated into 4 major types they are:

- Sensitive files
- Confidential files
- Private files
- Public files

Due to these file system segmentation, the owner can upload the files according to the importance. The sensitive and the confidential files are more secure, it will not delete the file automatically. But in case of public and private files will delete the files automatically after some period of time.

## 6. ARCHITECTURE DIAGRAM

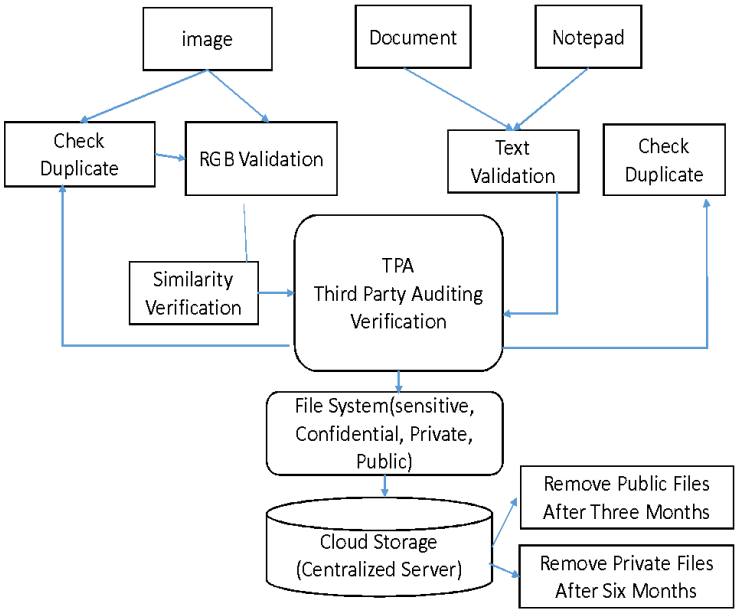



**Figure 2. Architectural diagram for de-duplication**

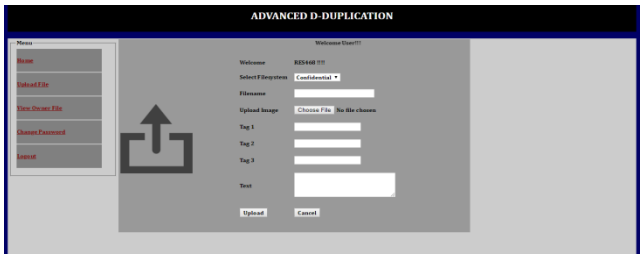
If the owner uploading a image, using MAT lab calculator it will identify the RGB value of that image, if another image is uploaded it will check the RGB value of the image with the previously uploaded image. And also in the image watermarking will be applied, if someone is needed the picture that we uploaded they want to send their

Special Issue on Proceedings of International Virtual Conference, Recent Materials for, Engineering Applications and Sustainable Environment, December 2020. International Journal of Aquatic Science, Volume 11, Issue 1, 2020.

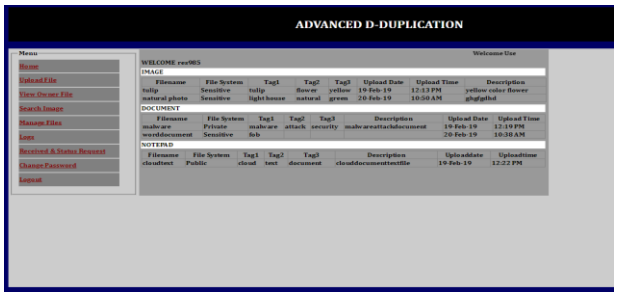
email id and phone number along with the message then the owner can make decision whether to send the image or not. In case of the document, it will convert the alphabets into 0's and 1's and it will cross check with the every document.

## 7. EXPERIMENTAL RESULTS

### 7.2 Uploading files




### 7.3 Management of uploaded files




### 7.4 Downloading the image with the permission of the owner


## 8. CONCLUSION:

In this work, the implementation of the de duplication has been implemented. This will be useful for the network of the public storage. Most of the companies were doing the storage manually and that may lead so many obstacles in the growth of the organization, this is the main objective of the project. Initially, gathered all information from the different real time server. The list of data that we collected has been cross checked with another group of data to find out the similarities between the file types. The data can be accessed simultaneously at the same time if it is stored in the centralized data base format.

At the time of uploading itself the duplicated files will be getting rejected. This has tested in different circumstance the output verified successfully.

## 9. REFERENCES

[1] Bellare.M, Keelveedhi.S, and T.Ristenpart, "Message-locked encryption and secure deduplication",2013, pp. 296-312

[2] Jiang.T, Chen.X, Wu.Q, Ma.J, Susilo.W, and Lou.W."Secure and efficient cloud data deduplication with randomized tag," IEEE Transactions on information Forensics and security, Vol.12, no. 3, pp. 532-543,2017.

[3] Kang Yang ,XiaohuaJia "Third-Party Storage Auditing Service",2014.

[4] Li, Jin, Yan Kit Li, XiaofengChen,MingqiangLi,Jingwei Li, Patrick PC Lee, and Wenjing Lou. "Secure deduplication with efficient and reliable convergent key management." Parallel and distributed systems, IEEE Transactions on 2, no.6(2016) 138-150.

[5] Meyer .D .T and Bolosky .J .W, "A study of practical deduplication," ACM Trans. Storage , Vol. 7, no. 4, pp. 1-20, 2012 .

[6] Shivas Mishra, Sujith Singh, and Syed Taqi Ali, "RSA based cross domain secure deduplication on cloud storage" Vol.29, Nov 18, October 2018

[7] Yan. Z, Ding.W, Yu.X, Zhu.H and Deng.R.H, "Deduplication on encrypted big data in cloud", IEEE transactions on big data, Vol.2, no.2, pp. 138-150, 2016.

[8] Zhou. Y, Feng. D, Xia.W, Fu.M, Huang.F, Zhang. Y, and Li. C, "Secdep: A user-aware efficient fine-grained secure deduplication scheme with multi-level key management", in mass storage systems and technolohgies(MSST), 2015 31st symposium on. IEEE,2015,pp. 1-14.

[9] Vishnu Sekhar. R, Nandhini. N, Bhanumathy. D, Hemalatha. M, "Identity Based Authentication for Data Stored in Cloud", Volume 5,Issue 3,March 2015.

[10] ChippyJaccob, Rekha V.R, "Secure and Reliable File Sharing System With De-Duplication using Erasure Correction Code", IEEE , July 2017.

[11] Nupoor M. Yawale, Gadichha. V. B, "Third Party Auditing(TPA) for data storage security in cloud with RC5algorithm", Volume 3, Issue 11, November 2013.

[12] Bellare, Mihir, SriramKeelvadhi, and Thomas Ristenpart, "Dupless: Server-aidded encryption for deduplicated Storage" USENIX Association, 2013.