Research Article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Data Deduplication in Cloud Storage System

P. Priya Ponnusamy*, K. Divya Preethaa and K. Yazhini

[1]CSE, PSG Institute of Technology and Applied Research, Coimbatore, India

ABSTRACT

As the technology is increasing day-to-day, people are storing their regular data in the cloud. The cloud computing services avoid the owing of complexity, upfront cost and maintaining their own IT infrastructure. Cloud computing services include Gmail, video streaming services in Netflix and backup photos. Cloud computing services have the basics of storage and networking. The cloud is the platform, where many users can use and store the data in an efficient way. While storing their data in the cloud, people may save the similar file content many times without their knowledge. This may cause confusion and also memory consumption. To avoid this, the proposed method helps in detecting files, which have similar contents. This method helps the users to make use of the memory space efficiently. By default, the files with same name are identified and notified. In this method, the files having the similar content are also identified and informed to the user. That is, content level duplication is carried in this proposed method. One of the best methods is the data deduplication technique. The data deduplication does not require any additional hardware. The network bandwidth, maintenance and the storage use get reduced and achieve good performance. This cloud computing is now being used in many business fields because it doesn't require any investment cost. The files are stored in the cloud by the users and checks for any duplication. The method is done using the Openstack cloud platform. By considering the benefits, the data deduplication is done effectively in cinder storage.

KEYWORDS: Data Deduplication, Openstack, Cinder, Storage space, Cloud Computing.

***Corresponding Author**

P. Priya Ponnuswamy

Assistant Professor, Department of Computer Science and Engineering

PSG Institute of Technology and Applied Research

2/12Kumaran nagar, Near Cheran Maanagar ,Coimbatore - 641 035 , Tamilnadu, India

Email -Id : priyabaskii@gmail.com

Ph: 9788923780

I. INTRODUCTION

Due to the advancement in technology, the data is produced rapidly and getting stored on time-to-time basis. The data sources are in the form of camera, tablets, mobile phones, laptops, computer systems, etc. The millions of data are stored in a fraction of time. Data Sharing is done mainly through the network. It can be done between different systems. Researchers found a solution to store these tons of data in the Cloud. Cloud Computing provides all the services, resources and data to the computer users. Cloud Computing provides Infrastructure as a Service, Software as a Service and Platform as a Service. Based on the accessibility and functionality of the Cloud services, the cloud has four Deployment models such as Private cloud, Public cloud, Hybrid cloud and Community cloud. Cloud Computing helps in creating more virtual machines for the user, as per pay and use basis. As more users are accessing the cloud, the data is shared and stored in the cloud may be repeated, which leads to store redundant data in the cloud.

To use the storage capacity of the cloud in an efficient manner, Data Deduplication method is proposed. The main goal of the Data Deduplication technique is to detect the redundant copies of data, which has been stored in the cloud storage system[3]. It ensures only one unique copy of data or file being stored. It checks content inside the file. Data Deduplication in cloud storage system provides efficient and secure file detection in the cloud. It manages storage system and it optimizes storage space in the cloud efficiently. It reduces the processing time of the system and thus enhances the system performance. For the file level Data duplication, the files with same name are detected. For the content level, the files with similar contents are detected. Many cloud service providers like Drive only checks File level duplication while uploading files. The files uploaded with same content but different names are not detected and thus getting up stored in the cloud. So, the main aim is to detect the recopied files in the cloud. The figure 1 represents data deduplication in the storage space[1], in which the total memory space before deduplication is 400 GB and after deduplication is 100 GB. The total storage memory space saved is 300 GB.

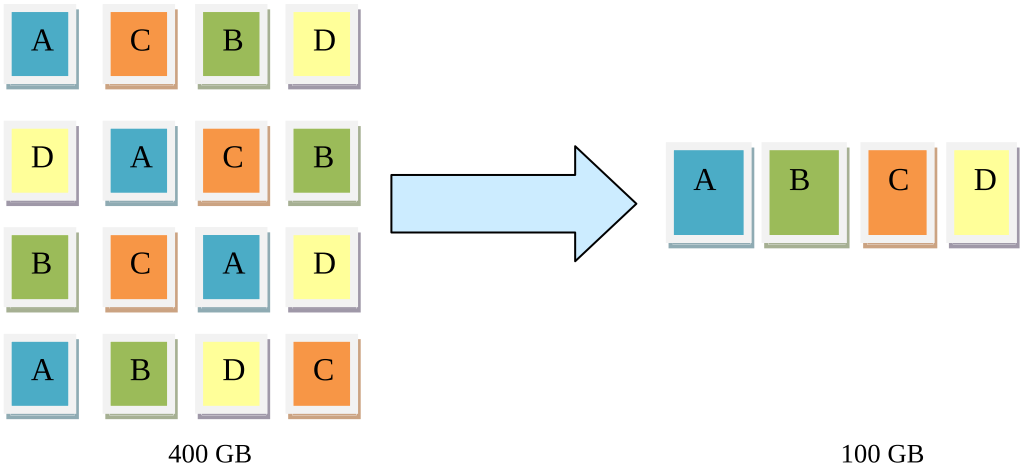


Figure 1. Data Deduplication[1]

Data Deduplication can be done in reference to chunking, location and time[2]. Based on chunking, there is file level chunking, whole file is consider as one chunk and block level chunking, in which whole file is partitioned into blocks of chunks. Based on location, Duplicate data are removed in source side before sending to users and or at client side after receiving. In reference to time, there is inline data deduplication and post-process data deduplication. In In-line data deduplication, it is done at client side, so that only unique copies of data are stored in disk before sending to server. In Post proccess data deduplication, the method is done at server side, the data's are stored on disk first, then duplicates are removed. As Data Deduplication in cloud storage system deals with the data, there is concern in security. For the security purpose, hashing on the data is done. The Data Deduplicaion in Cloud storage system is implemented in Openstack Cinder storage. Openstack is set of software tools for managing cloud computing platforms. Cinder is the block storage volume in Openstack Cloud platform. It provides virtualised block storage, in which users can use resources either through LVM or Plug-in drivers[5]. Cinder block storage provides volume to the virtual machines or instances.

II. PROBLEM STATEMENT

In recent years, the critical task is to store the fast growing data. These data's are stored in the cloud. There might be similar data found in the cloud. Due to this, redundant copies of data are being stored. This might not be realized, as the data has been recopied and resaved again and again. This takes more storage space in the cloud system and thus it leads to slow the system performance and increase the processing time. The main challenge in the cloud is the duplicated data storage.

III. LITERATURE SURVEY

For the duplication process, the files are divided into chunks, and the chunks are selected based on the threshold value and the sliding window and the performance is done on the chunk files[10]. The chunk files are compared to one another to find the duplicate files. The main problem is there may be some information leakage while performing the duplication operation, this has done to reduce the space efficiency as a compromise[8]. For the deduplication operation, convergent encryption is introduced on the file contents which produce the similar cipher text for the same plain text but it needs more computational costs in storage space[10], which is difficult for the normal users. In another case, the cloud makes the users to allow the files, which are accessed[13] and checks for the duplication. The problem is the usage of MD5 hash algorithm[5], as it identifies the duplicate files faster, but it is not much secure. At the same time, while finding the duplicate files, the consistency of the data is maintained through the commitment scheme, but this method does not support and satisfy the semantic network. This makes other clients and servers to take an important role in sharing the resources[6]. So this network should be maintained properly. The file in the cloud is protected, only known user with the re-encryption key can access the file. When the key is provided, that has to be used for access, but here updating the key is not possible.

In another terms, for the easy access of the file, the files are stored in the public cloud. This would invites a problem that anonymous user can access the file without the need of key and there is a possibility in modification of data’s in the file. The duplicate files are identified for many users, for the space management and the effective resource and time utilization, where the files are divided into different types of chunks and compared using the stream matching algorithm[7]. Here the problem is that when many users while restoring the data, there is a possibility of forming duplicate files in the backup storage. The security should be achieved while finding the data duplication by dividing the users into authenticated one and the anonymous one[8] and the map is used for the reconstruction of chunks after finding the duplication process, as the chunk may change its order after dividing process, but the time taken to compare each chunk with the other is time consuming. The duplicate files are identified at the client side. This method of duplication reduces the storage and the network bandwidth[9]. As many users transfer the files to the server, there can arises a situation that duplicate files may present in the server side, but this method identifies the duplication only in the client side. The cloud provides the best way to identify the files for making resource utilisation effectively and also taken all the measures to protect the files[14]. The main problem here is that action and steps are to be taken for the new changing environment, when the new problem arises[15].

IV. PROPOSED SYSTEM

The proposed model works in the cloud platform. In the cloud service many users can access the file and can also store the file. When the user storing the file, the same file can be stored multiple times or many users can store the same data. These are the reasons for the presence of duplicate files in the cloud system. Most of the cloud platforms, identify the duplicates only in the file name level, but in this method the file name duplicate along with the content level also checked for duplication. In the proposed system, the duplicated files, which are resaved, recopied, or redundant copies are detected and displayed. The files with same content but different file names are also checked. The proposed Data Deduplication in cloud storage system is done in Openstack Cloud platform. In Openstack, all the files are stored in cinder Volume, which is the storage area. The Data Deduplication is done in this storage area.

For this project, VMware workstation is installed, which helps in running multiple Operating system in a single machine. The Openstack is installed and the nodes used for this model are Controller and Compute nodes[19], which is described below. The login helps in authentication for both nodes.

Controller

The controller node runs controlling services of Openstack, which includes the image service, the identity service, the management service, the networking service and the dashboard. It also has supporting services, which includes database and message broker. It has block storage, object storage, orchestration and telemetry services as optional.

Compute

The compute node runs the hypervisor of compute that operates the virtual machine or instance. Hypervisor is the process which helps to create and to run virtual machines. It gives control to instances and network. It helps in connecting the instances to the network and provides security to the instances.

The figure 2 shows the flow chart for the proposed method. The dedupication method helps in managing the storage space in the cloud efficiently. The Hashing method is the performance metric for the deduplication methodology. Hash is a function, which takes input value and converts it into the compressed output value. The output value is called hash codes, hash values, hashes and digests. The input can be of any length, but the output is of fixed size length. The hash functions are used for data integrity checks, password storage. The algorithm for the proposed method is follows as

1. Specify the Openstack Cloud cinder storage directory, in which the deduplication has to be done.
2. Take each file in the directory specified.
3. Calculate size of each file and same size files are considered, since similar content files will have same size.

4. The File based chunking is done, so that whole file is considered as one chunk.
5. Find hash value for same size files using SHA-512 function.
6. The same hash values of files are considered, as they have similar contents.
7. Display the files.

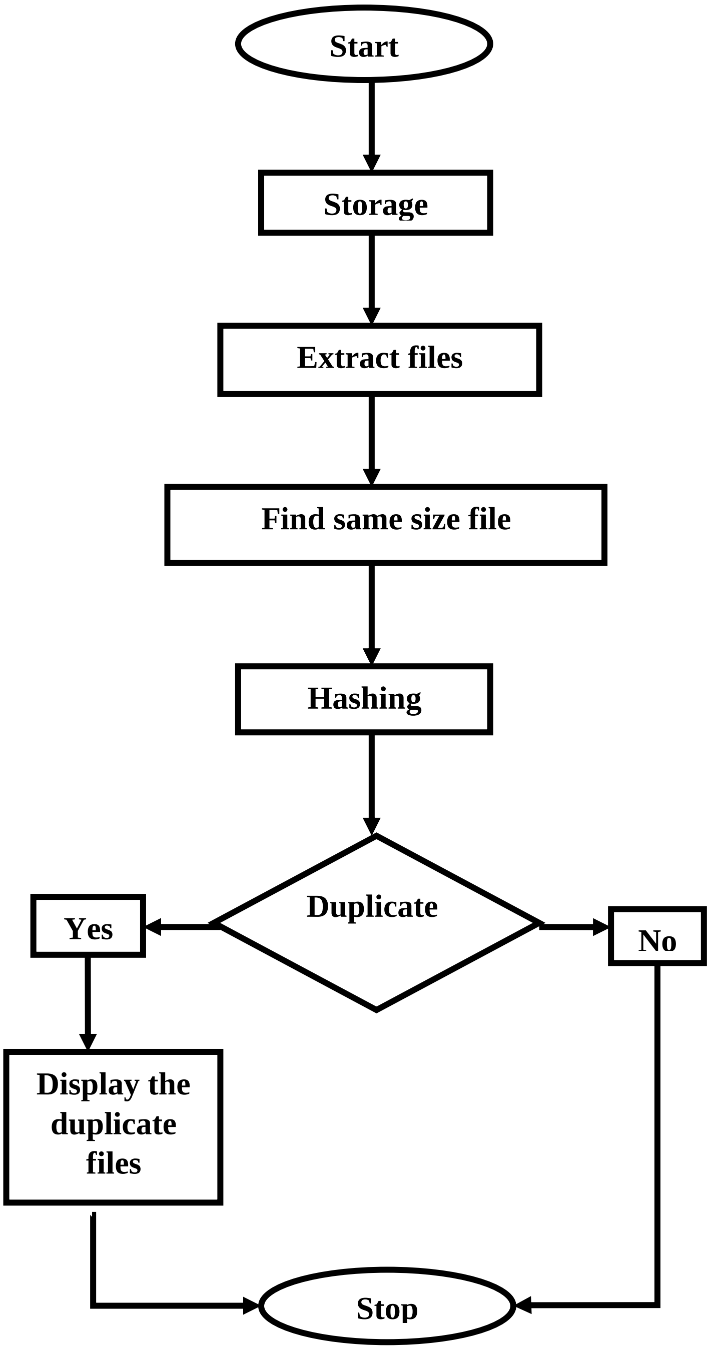


Figure 2. Flow chart for Data deduplication in cloud storage system

The hash functions known commonly are message digest and secure hash algorithm. The SHA-512 algorithm is used to get hashcode value. Secure hash algorithm is more secured while considering the security purpose and no serious attack happens. SHA protects the data and it can't be changed or modified easily[16]. SHA-512 produces the output of 512 bit hash code. The hash code value is found for similar size files. The file having same hash code value is updated in the list. The list displays the duplicate files.

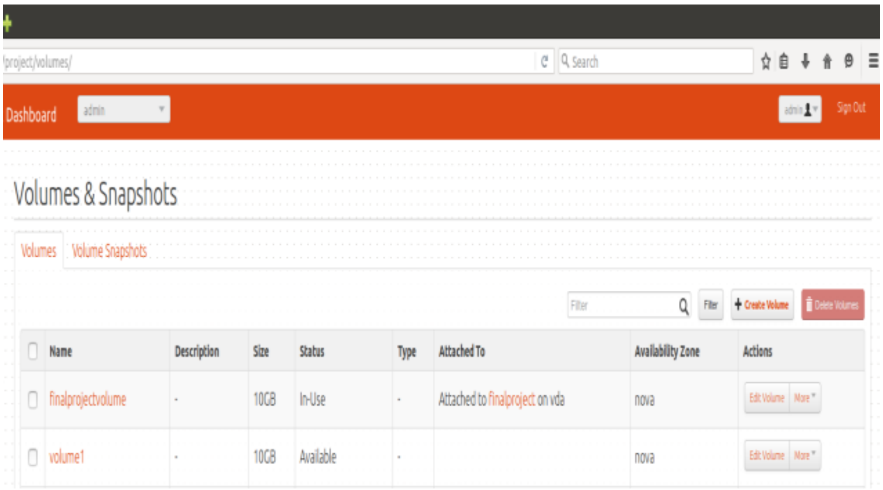
V. IMPLEMENTATION DETAILS

The data deduplication in cloud storage system is implemented in Openstack. This project is implemented with controller and compute machines. The services like image creation, volume creation, virtual machine creation, attachment of virtual machine to the volume are used and the method FTP transfer is done for this implementation.

Services used

In Openstack, Horizon is the web-based graphical user interface, which is used to manage storage, compute and networking services. The Virtualisation tool VMware Workstation Pro helps in converting the file into raw one. After creating the new image, it is connected to the Horizon graphical interface which is the display, helps in interacting with guest Operating System. Image creation is done in Horizon Dashboard in Controller node. The image is created with the name 'final'. The image source is from the system location which is the ISO file. The QCOW2 image format is used, which is one of the disk image format supported by QEMU processor emulator.

Cinder is the block storage service which is used for storing data permanently in the Openstack cloud. It provides storage resources to the users. Data Deduplication in cloud storage system is done in this storage area. For this, Volume can be created using either CLI or Horizon Dashboard. In this model, in Controller, the cinder volume is created using CLI. The Volume is created using bootable image. The volume name is 'finalprojectvolume'. The volume size used for creation is 10 GB. While creating, the status is displayed as creating. After creation of volume, the command cinder show is used, which displays current status as available, which clearly shows that the volume is created. Cinder Volume takes its space in compute node, after its creation. The Cinder stores all files from the Controller node. The files stored here are recopied file, redundant file, and same content with different file name. Data Deduplication is done in this storage area. The figure 3 shows the cinder volume, which has been created in the Openstack.



□	Name	Description	Size	Status	Type	Attached To	Availability Zone	Actions
□	finalprojectvolume	-	10GB	In-Use	-	Attached to finalproject on vda	nova	Edit Volume More
□	volume1	-	10GB	Available	-		nova	Edit Volume More

Figure 3. Screenshot of the Cinder volume creation

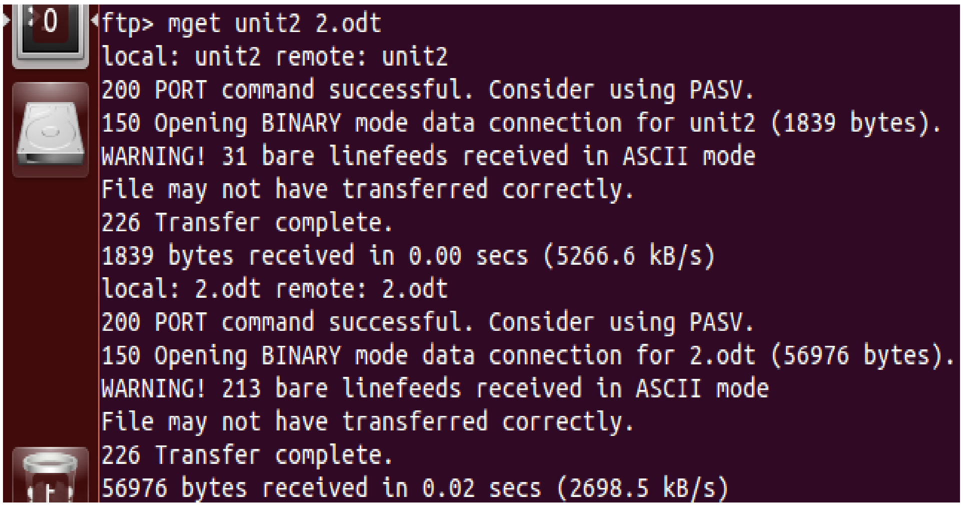
Virtual machine instance run inside the cloud. Virtual machines allow users to store multiple OS on the same system. It helps in switching between OS effectively. It improves productivity and reduces the cost. Virtual machine instance can be created either using CLI or using Horizon dashboard. The source of the instance is from image or snapshot or from block storage volume that contains the image or the snapshot. The instance is created using flavour, availability zone. The instance name created as ‘finalproject’. The flavour used is m1.small. The availability zone is compute1. The instance is created from block storage volume finalprojectvolume that contains image final. After creating the instance, the status is converted from build to Active.

While creating instance from bootable block storage volume, the instance is automatically attached to the volume. Thus the virtual machine is attached to the volume successfully. The attachment of volume is verified using CLI and in the Horizon interface. The cinder list command displays Volume id, display name, status, size, volume type, bootable source and attachment. The cinder volume finalprojectvolume is attached to virtual machine finalproject.

FTP Transfer

FTP stands for File Transfer Protocol which transfers files between two systems. FTP users are authenticated in the form of username and password, and also users can connect anonymously, if the server permits it. FTP connection taken place either through active mode or passive mode. Active mode allows open communication between server and the device over both channels. In Passive mode, where the server concentrates, but not maintain connection actively, allow other device to do all the work. FTP Transfer is done between Controller and Compute machines. For the FTP Transfer, the FTP is installed on both Controller and Compute machines.

In order to transfer, both the systems are connected to each other with ipaddress. The command ftp is used along with other machine ipaddress and connection is made between controller and compute systems. After Connection, each machine is authenticated using user name and password. The Login is made successful between controller and compute systems. As Cinder storage is in Compute node, the files are transferred to compute node from controller node. The figure 4 represents the ftp transfer done at compute machine.



```
ftp> mget unit2 2.odt
local: unit2 remote: unit2
200 PORT command successful. Consider using PASV.
150 Opening BINARY mode data connection for unit2 (1839 bytes).
WARNING! 31 bare linefeeds received in ASCII mode
File may not have transferred correctly.
226 Transfer complete.
1839 bytes received in 0.00 secs (5266.6 kB/s)
local: 2.odt remote: 2.odt
200 PORT command successful. Consider using PASV.
150 Opening BINARY mode data connection for 2.odt (56976 bytes).
WARNING! 213 bare linefeeds received in ASCII mode
File may not have transferred correctly.
226 Transfer complete.
56976 bytes received in 0.02 secs (2698.5 kB/s)
```

Figure 4. Screenshot of the FTP Transfer

VI. RESULTS

After the successful FTP File Transfer, all the files are stored in cinder volume storage. The storage is in compute machine. The algorithm is implemented in this storage location to manage storage space. The main aim is to find duplicate files using data deduplication technology. In general, hashing is done for all the files, and then the duplicate files are found[17]. In this project, the goal is to display files with same content, but different file name. If the files have same content, then their size will be same.

In this case, the hashing is needed not to be done for all the files, but for the files having same size. So, first, list the files having same sizes. Here, File based chunking deduplication is done, where the hashcode is generated for the whole file. Hashing is done only for the files listed based on size. The duplicate files found based on the hash value are displayed from the list. In the figure 5 the input files in cinder storage in the compute machine are shown.

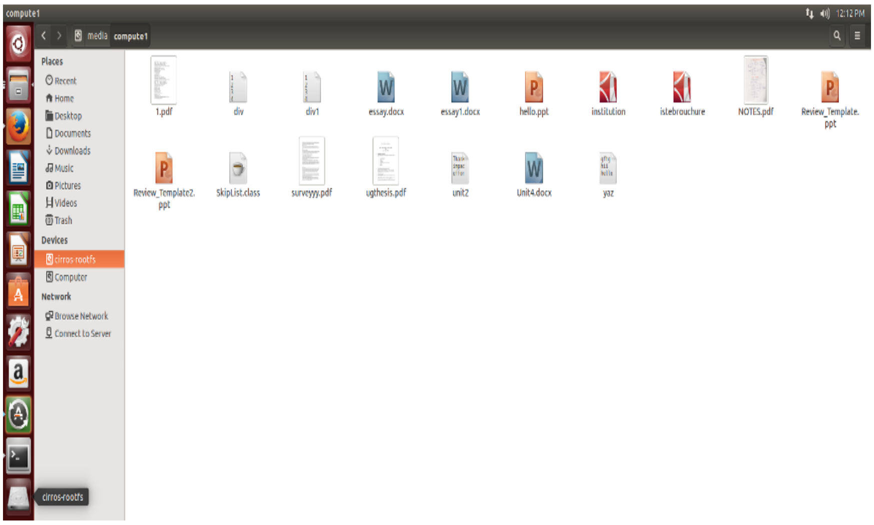
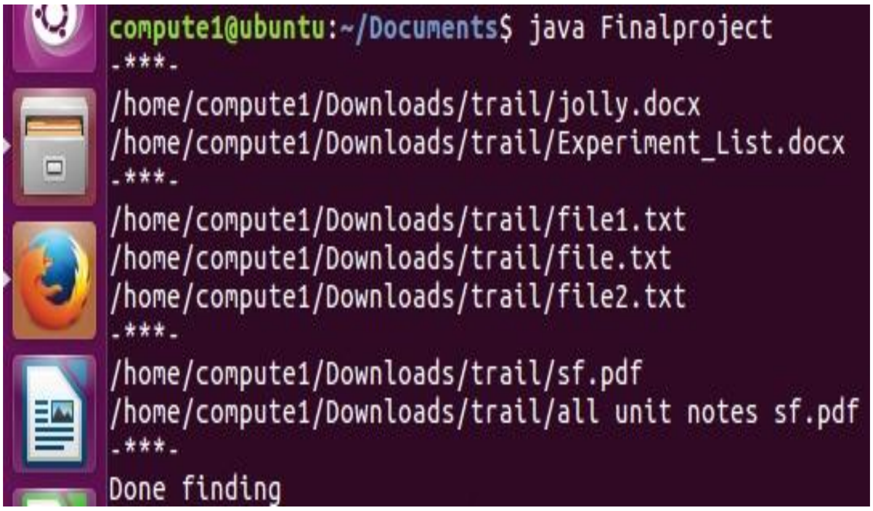


Figure 5. Files in Volume

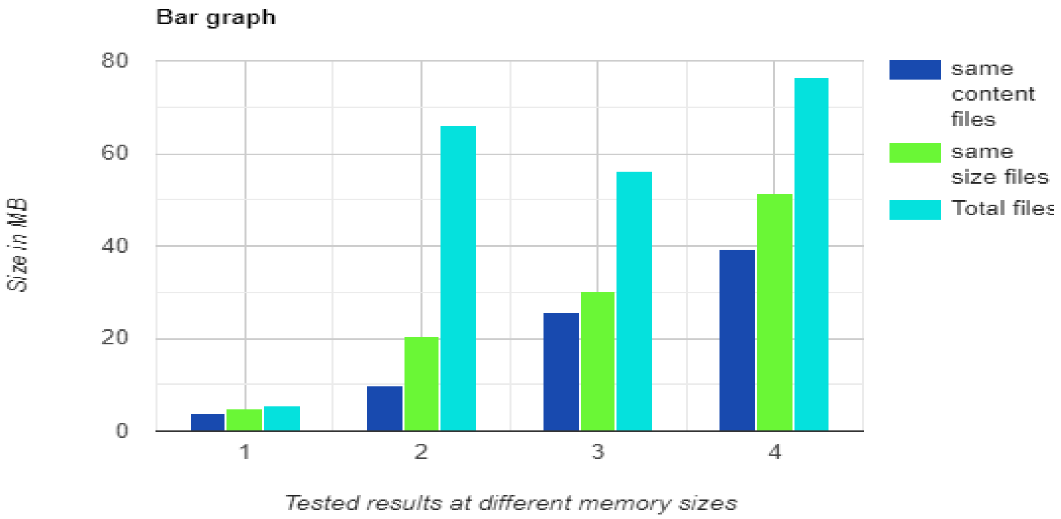
Once the code is executed in the linux terminal in compute node, the deduplicated files stored in the volume storage are displayed. The file names are shown, so that the duplicate files can be find easily. The figure 6 shows the output duplicate file names.



```
compute1@ubuntu:~/Documents$ java Finalproject
-***-
/home/compute1/Downloads/trail/jolly.docx
/home/compute1/Downloads/trail/Experiment_List.docx
-***-
/home/compute1/Downloads/trail/file1.txt
/home/compute1/Downloads/trail/file.txt
/home/compute1/Downloads/trail/file2.txt
-***-
/home/compute1/Downloads/trail/sf.pdf
/home/compute1/Downloads/trail/all unit notes sf.pdf
-***-
Done finding
```

Figure 6. Display of files

The bar graph in figure.7 is represented using size in MB which is tested at memory size of the storage at different levels. The three bars represent the size of same content files, the size of same size files, and the size of total files in the memory space. There are four tests done at storage by implementing the method and results obtained as based on the same content file and same size file[18]. There is variation in the graph, when more files are stored in the location. The bar graph shows result as if memory size decreases, the amount of deduplication that is same content files are also decreasing.



Graph 1. Bar graph

The different files like documents, text files are stored in the cinder storage in Openstack. For testing, duplicate files with same content and different file names are stored. Other parameters like same content files, same size file with different file names and different content files, same size file with different file names are tested. Different content, different size file with different file name are also tested.

VII. CONCLUSION

In this paper, Data Deduplication in cloud storage system is developed to manage cloud storage space effectively by identifying the data redundancy without comprising the data security. The duplicate files are mainly arises due to multiple shares, multiple users, backups, e-mail attachments sending to many users. The proposed system helps in finding deduplication not only by the file name, but also by the contents inside the file with data integrity. This system will solve the cloud storage capabilities problem. The process of checking each and every file in the location, identifying the standard measures of the file reduces the time and the work amount. The proposed method helps in increasing the performance in the cloud. Due to the use of Deduplication method in this system, it helps in saves storage space, saves time, optimisation, high capacity possible. So, the output productivity increases and more users will involve in it. In future, more work can be done in the security areas by concerning the data privacy of the user's, who are accessing the cloud.

ACKNOWLEDGEMENT

We express our deep sense of gratitude to Shri. L. Gopalkrishnan, Managing Trustee, PSG & SONS Charities for his blessing in the project work. We are thankful to Dr. P. V. Mohanram, our Principal for providing us with an environment to complete our project successfully. We are grateful to Dr. G. Chandramahan, our Vice principal for reaching out his hands in all aspects. We are deeply indebted to Dr. R. Manimegalai, Head of the Department, Computer Science and Engineering, who moulded us both technically and morally for achieving great success in life. We are grateful to Ms. P. Priya Ponnusamy, our Project Guide, for being instrumental in the completion of our project with her guidance. Also, we express our sincere thanks to Mr. C. P. Shabariram and Dr. S. Bhuvana, our Project Coordinators for their constant encouragement and support throughout the course. Finally, we take this opportunity to extend our deep appreciation to our family and friends, for all that they meant to us during the crucial times of our project.

REFERENCES

[1] Michael bertuit, Architect Expert Microsoft. “Amazing data deduplication!!! How to optimize your free space on windows 8.1 & windows 2012 [online]”. 2014 [cited 2014 April]. Available from: URL: http://www.system-center.fr/wp-content/uploads/2014/04/microsoft-deduplication-windows-bertuitm.png

[2] N. Chhabra and M. Bala, "A Comparative Study of Data Deduplication Strategies," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018; 68-72, doi: 10.1109/ICSCCC.2018.8703363.

[3] D. Geery, "Reducing the Storage Burden via Data Deduplication," in Computer, Dec. 2008; 41(12): 15-17, doi: 10.1109/MC.2008.538.

[4] Openstack training course, “Openstack Block Storage with Cinder” [online] , Available from: URL: https://www.theskillpedia.com/course/openstack-training-course.

[5] M. V. Maruti and M. K. Nighot, "Authorized data Deduplication using hybrid cloud technique," 2015 International Conference on Energy Systems and Applications, Pune, 2015; 695-699. doi: 10.1109/ICESA.2015.7503439.

[6] H. Cui, R. H. Deng, Y. Li and G. Wu, "Attribute-Based Storage Supporting Secure Deduplication of Encrypted Data in Cloud," in IEEE Transactions on Big Data, Sept. 2019; 5(3): 330-342. doi: 10.1109/TBDATA.2017.2656120.

[7] Nagapramod Mandagere, Pin Zhou, Mark A Smith, and Sandeep Uttamchandani, 2008, “Demystifying data deduplication”, In Proceedings of the ACM/IFIP/USENIX Middleware ’08 Conference Companion (Companion ’08), Association for Computing Machinery, New York, NY, USA, 12–17. DOI: https://doi.org/10.1145/1462735.1462739.

[8] Mark W. Storer, Kevin Greenan, Darrell D.E. Long, and Ethan L. Miller, 2008, “Secure data deduplication”, In Proceedings of the 4th ACM international workshop on Storage security and

survivability (StorageSS ’08), Association for Computing Machinery, New York, NY, USA, 1–10. DOI: https://doi.org/10.1145/1456469.1456471.

[9] N. Kaaniche and M. Laurent, "A Secure Client Side Deduplication Scheme in Cloud Storage Environments," 2014 6th International Conference on New Technologies, Mobility and Security (NTMS), Dubai, 2014, pp. 1-7, doi: 10.1109/NTMS.2014.6814002.

[10] J. Wu, Y. Li, T. Wang and Y. Ding, "CPDA: A Confidentiality-Preserving Deduplication Cloud Storage with Public Cloud Auditing," in IEEE Access, vol. 7, pp. 160482-160497, 2019, doi: 10.1109/ACCESS.2019.2950750.

[11] A. S. Ibrahim, J. Hamlyn-Harris, J. Grundy and M. Almorsy, "CloudSec: A security monitoring appliance for Virtual Machines in the IaaS cloud model," 2011 5th International Conference on Network and System Security, Milan, 2011, pp. 113-120, doi: 10.1109/ICNSS.2011.6059967.

[12] Z. Xiao, W. Song and Q. Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," in IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 6, pp. 1107-1117, June 2013, doi: 10.1109/TPDS.2012.283.

[13] Peter Mell Timothy Grance, “The NIST Definition of Cloud Computing”, U.S. Department of Commerce, National Institute of Standards and Technology, Special Publication 800-145, September 2011, Available from: URL: https://csrc.nist.gov/publications/detail/sp/800-145/final.

[14] Sean Quinlan and Sean Dorward, “A new approach to archival storage”, In Proceedings of the FAST 2002 Conference on File and Storage Technologies, Monterey, California, USA January 28-30, 2002.

[15] Mohamed Almorsy, John Grundy and Ingo Muller, “An Analysis of the Cloud Computing Security Problem,”, 2016 [cited 2016 September], arXiv, cs.SE.

[16] Sundeep Saradhi Kanthety. “Network Security – SHA 512” [online]. 2018 [cited 2018 Feb 6]. Available from: URL: https://www.youtube.com/watch?time_continue=1&v=VtHSyoJkDXw&feature=emb_logo.

[17] Vytautas Jakutis. “To recursively find all duplicate files in the directory” [online]. 2015 [cited 2015 Mar 15].

[18] N. N. Pachpor and P. S. Prasad, "Improving the Performance of System in Cloud by Using Selective Deduplication," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 314-318, doi: 10.1109/ICECA.2018.8474932.

[19] P.PriyaPonnuswamy et. al. "File retrieval and storage in the source cloud tool using digital bipartite and digit compact prefix indexing method" 2018 Concurrency computation practices and experiences doi10.1002/cpe5307.