International Institution for Technological Research and Development
Volume 2, Issue 1, 2017

Optimizing Secure Deduplication for Big Data in Cloud

Dipeeka Shah[1], Prof. Gayatri S. Pandi (Jain)[2]

[1] CE, LJIET, GTU
Ahmedabad, Gujarat, India
shahdipeeka27@gmail.com

[2] CE, LJIET, GTU
Ahmedabad, Gujarat, India
gayatree.jain@gmail.com

Abstract

The trend of cloud has increased with the increasing data in internet. Cloud is based on virtualization which is raising the virtual machine images in data centers, and maintaining the original and backups create the requirement of space in several TB. This results in creation of replicas of existing data or new data at time of entry, resulting duplication. For this, deduplication of data is required. Previously, we have done survey on different deduplication technique and studied that which approach suits the best. In this paper, we have developed dynamic deduplication technique in that we have check the deduplication of files according to their similar or different types. As per the analysis it reveals that this dynamic deduplication technique is more efficient and less time consuming, with the help of chunks and metadata of file.

Keywords: *Data Deduplication, Cloud, Big data, secure deduplication*

1. Introduction

The use of cloud computing to store, share their data for the different purposes has increased now a days. Different users are uploading and storing their data for multiple times. It needs a very large amount of storing space to store that data. It may happen that different users upload same data and many times the same user uploads the data more than once knowingly or unknowingly. If the data is stored again and again, it needs a very large amount of storage space. To save the storage space checks can be applied at the time when the data is uploaded by the user. If the data already exists then it will notify user that there is a duplicate data otherwise it will store in the cloud server. Deduplication is the technique of removing the redundant data [17].

With the rapid adoption of cloud services, more and more volume of data is stored at remote servers, so techniques to save disk space and network bandwidth are needed. A key concept in this context is deduplication, in which the server stores only a single copy of each file, regardless of how many clients want to store that file. All clients possessing that file only use the link to the single copy of the file stored at the server [17].

2. Data Deduplication

Data **Deduplication** is a specialized data compression technique for eliminating duplicate copies of repeating data [17].

2.1. Benefits

Basically, it can reduce the storage space occupied by the data. This will bring the following benefits:

- IT savings funds (do not need the extra space needed to increase investment)
- Reduce the backup data, data snapshots of the size of the (cost-saving, saving time, etc.)
- Less power pressure (because of less hard, less tape, etc.)
- Save network bandwidth (because only fewer data)
- Saving time
- Because of the need less storage space, disk backup possible.

2.2. Data Deduplication Process

The basic steps to delete duplicate data consists of five stages [17]:

1) The first phase of data collection phase, by comparing the old and the new backup data backup, reducing the scope of the data.
2) The second phase of the process of identifying data, in bytes, of the data collection phase marks a similar data objects.

3) The data is re-assembled, new data is saved, the previous stage was marked duplicate data is saved data pointer replacement. The end result of this process is to produce a copy of the deleted after the backup group view.
4) Actually remove all the duplicate data before performing a data integrity check efficacy.
5) Finally remove the redundant storage of data, the release of previously occupied disk space for other uses.

2.3. Data Deduplication Techniques

Data de-duplication technology to identify duplicate data eliminate redundancy and reduce the need to transfer or store the data in the overall capacity [15].

At present mainly the file level, block-level and byte-level deletion strategy, they can be optimized for storage capacity.

A. File-level data deduplication

It is also known as Single Instance Storage (SIS). It checks the index of the file with stored files and doing the comparison. If the files are different than it will store and update the index; otherwise, the only deposit pointer to an existing file. So, the file will be stored only once and then copy all the "stub" alternative, while the "stub" pointing to the original file [15].

B. Block-level data deduplication

The data of files will be divide into the blocks and do the comparison of the data blocks. If the data blocks are different than t will be store and update the index, otherwise, the only deposit pointer to store the same data block's original location [15].

C. Byte-level data deduplication

The new data stream and have stored more bytes of data stream one by one, to achieve higher accuracy. With byte-level technology products are usually able to "identify the content," In other words, the supplier of the backup process the data flow implementation of the reverse engineering to learn how to retrieve the file name, file type, date / time stamp and other information[15].

3. Related Work

In [1], Puzio, P., Molva, R., Onen, M., & Loureiro, S has proposed ClouDedup, which is a secure and efficient storage service which assures block-level deduplication and data confidentiality at the same time. Although based on convergent encryption, ClouDedup remains secure by additional encryption operation and an access control mechanism. Furthermore, as the requirement for deduplication at block-level raises an issue for key management, it is suggested to add a new component in order to implement the key management for each block together with the actual deduplication operation and it never impacts the overall storage or operational cost. Block-level deduplication instead of file level deduplication has been preferred as storage space are not affected by the overhead of metadata management Additional layers of encryption are added by the server and the optional HSM.

In [2], Zhou, Ruijin, Ming Liu, and Tao Li. have presented a way to reduce the overhead of extra CPU computation (hash indexing) and IO latency introduced by deduplication and for this, they characterized the redundancy of typical big data workloads to justify the need for deduplication. Analysis and characterization of performance and energy impact brought by either local or global deduplication has been done under various big data environments. They identified three sources of redundancy in big data workloads as 1) deploying more nodes, 2) expanding the dataset, and 3) using replication mechanisms.

In [3], Luo, Shengmei, Guangyan Zhang, Chengwen Wu, Samee Khan, and Keqin Li. presented Boafft that uses Min-Hash to estimate two similar super blocks, a cloud storage system with distributed deduplication that achieves scalable throughput and capacity using multiple data servers to deduplicate data in parallel, with a minimal loss of deduplication ratio. Similarity is shown using Jaccard Similarity Coefficient. Firstly, Boafft uses an efficient data routing algorithm based on data similarity that reduces the network overhead by quickly identifying the storage location. Secondly, the Boafft maintains an in-memory similarity indexing in each data server that helps avoid a large number of random disk reads and writes, which in turn accelerates local data deduplication. Thirdly, the Boafft constructs hot fingerprint cache in each data server based on access frequency, to improve the data deduplication ratio. Boafft can provide a comparatively high deduplication ratio with a low network bandwidth overhead, along with better usage of the storage space, with higher read/write bandwidth and good load balance.

In [4], Yan, Zheng, Mingjun Wang, Yuxiang Li, and Athanasios V. Vasilakos has proposed a scheme to deduplicate encrypted data at CSP by applying PRE to issue keys to different authorized data holders based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. For ownership verification it has used ownership verification

protocol based on crypto GPS identification scheme. For security of the proposed scheme has been ensured by use of PRE theory, symmetric key encryption and ECC theory.
In [5], Wen, Mi, Kejie Lu, Jingsheng Lei, Fengyong Li, and Jing Li have proposed an efficient secure deduplication scheme for big data outsourcing in the cloud computing called BDO-SD that implies Convergent encryption. The overall BDO-SD consists of four phases: Registration phase, File upload phase, File download phase, and File recovery phase. With Convergent encryption , data owner can upload files to the cloud with data deduplication by using the convergent keys, while data users can retrieve required files by giving file tags or file keywords. Security analysis demonstrates that BDO-SD can achieve data confidentiality, key security and query privacy.

In [6], Waghmare, V., & Kapse, S have proposed secure authorized deduplication using token generation mechanism in cloud service and service providers employ data deduplication technique without giving access to either user's plain text or user's decrypted data. For authorized deduplication system, they have used Symmetric Encryption, De-duplication using Hashing function, Convergent Encryption and Token Generation.

In [7], Chen, Ming et al. proposed a local deduplication method for speeding up the operation progress of virtual machine image deduplication and reduce the operation time. For this, improved k-means clustering algorithm has been used that classify the metadata of backup image to reduce the search space of index lookup and improve the index lookup performance.

In [8], Xu, J., Zhang, W., Ye, S., Wei, J., & Huang, T. has proposed a novel server-side deduplication scheme for encrypted data that allows the cloud server to control access to outsourced data even when the ownership changes dynamically by exploiting randomized convergent encryption and secure ownership group key distribution, preventing data leakage for revoked users even though they previously owned that data and is honest-but-curious cloud storage server, guarantying data integrity against any tag inconsistency attack.

In [9], Hur, J., Koo, D., Shin, Y., & Kang, K. proposed a cloud based model for implementing deduplication of a large amount of data available. The model comprises of both the deduplication of data before uploading to the cloud storage and the reverse deduplication of data when downloading the necessary data. The two deduplication techniques namely Preprocessing and Chunking process are used on client side whereas on server side, deduplication techniques the hash generation and Duplicate identification are used. Performance of deduplication process is measured based on Deduplication Rate, Lookup Latency, Collision Rate, and Security Measure.

In [10], Kirubakaran, R., Prathibhan, C. M., & Karthika, C. has proposed a dynamic deduplication scheme for cloud storage that also maintains QoS of Cloud environment, which aiming to improve storage efficiency and maintaining redundancy for fault tolerance. In this, after identifying the duplication, the Redundancy Manager then calculates an optimal number of copies for the file based on number of references and level of QoS. The numbers of copies are dynamically changed based on the changing number of references, level of QoS and demand for the files.

4. Proposed Work:

4.1 Problem Statement

In orthodox methodology of deduplication of data, the data is divided in blocks and compare it with other blocks. This approach is simple yet time consuming because it compares all the blocks of data with other blocks of data. Efficiency of data deduplication is also one of the limitations due to dividing data into various blocks.

4.2 Proposed System

We have made a survey on different techniques of deduplication and which approach suits the best. From that, we have developed dynamic deduplication technique in that we have check the deduplication of files according to their similar or different types. As per the analysis it reveals that this dynamic deduplication technique is more efficient and less time consuming, with the help of chunks and metadata of file.

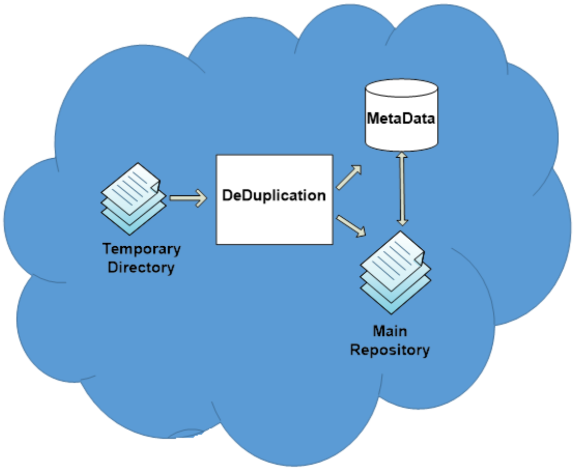

Figure: 1. Proposed Model

The user will upload the file in cloud storage server. That file will be stored in the metadata in it is not uploaded previously and if not than it will be end. After storing the file in metadata it will check that the upload files are similar types of files or not. The comparison is based on file size, type, name etc. If the files are same type of files than it will be divide in the chunks. The number of chunks will be calculated. The comparison of same type of file will be done by comparing the numbers of the chunks. For the same numbers of chunks will be comparing the first. Each chunk will be comparing individually. If the duplicate data is found than notify message will be sent to the user and it's up to on that user to save that duplicate data or not. Now, the second approach for the different types of files, we will compare the whole block of the file. If the duplicate data is found than notify message will be sent to the user and it's up to on that user to save that duplicate data or not.

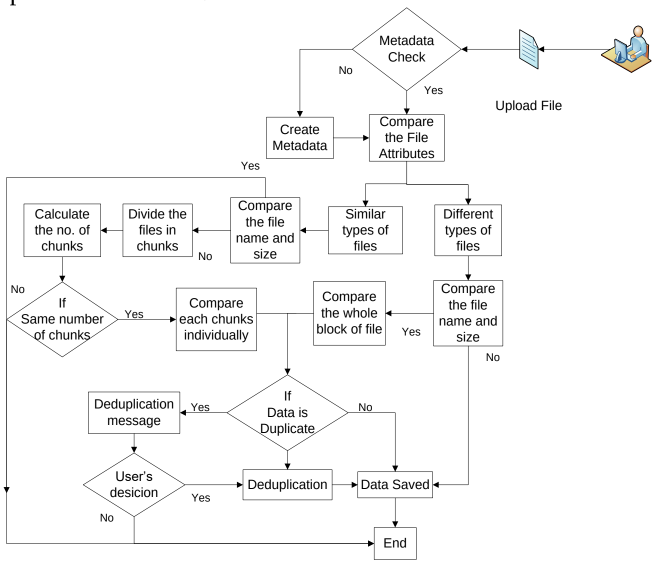

Figure: 2. Flow Diagram of the Proposed System

4.3 Algorithm:

The algorithm is as following:

```
Step 1: Upload file.
Step 2: Metadata check
        If yes
          Go to Step 3
        Else
           Create metadata
Step 3: Compare the file attributes (File name, type, size)
Step 4: Similar types of files
Step 5: Divide the file in chunks
Steps 6: Calculate the no. of chunks
Step 7: If the same no. of chunks
Step 8: Compares each chunks individually
Step 9: Different types of files
Step 10: Deduplication check
Step 11: Deduplication message to user
Step 12: Store/ Discard the data (As per user's decision)
Step 13: End
```

5. Experimental Evaluation:

As we are checking the deduplication content so its need to find out the accuracy of checked dedup content. Accuracy refers to the closeness of a measured value to a standard or known value. We can measure the accuracy for the textual data by using the following equation:

$$Accuracy = \frac{\sum tf(no.of\ dedup\ lines)}{\sum(total\ lines)}$$

Where, tf = term frequency

We have done the analysis for the performance and throughput**.** As per the analysis we can stated that the performance and throughput are increasing by using the dynamic deduplication techniques.

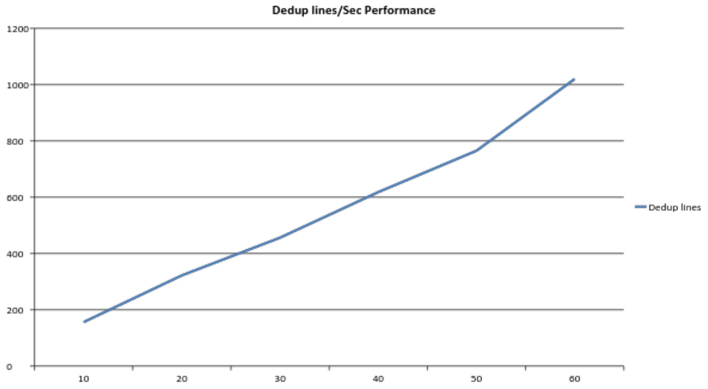

Figure: 3. Proposed System's Performance (Dedup lines/Sec)

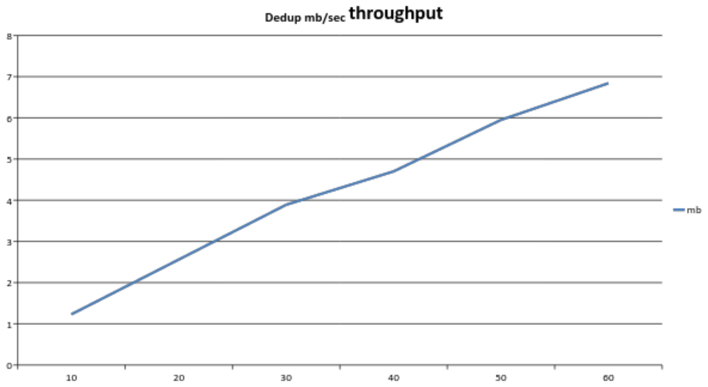

Figure: 4. Proposed System's Throughput (Dedup lines/Sec)

6. Conclusion:

Cloud storage services offers on demand virtualized storage resources and customers only pay for the space they actually consumed. As the increasing demand and data store in the cloud, data deduplications one of the techniques used to improve storage efficiency. Data deduplication is a specialized data compression technique for eliminating duplicate copies of data in storage. Here, we have developed dynamic deduplication technique in that we have check the deduplication of files according to their similar or different types. As per the analysis it reveals that this dynamic deduplication technique is more efficient and less time consuming, with the help of chunks and metadata of file.

Acknowledgments

I would like to thanks my faculty member who guided me when I needed help, also like to thank every person who help me and guide me in this work by their experience and knowledge.

References

[1]. Puzio, P., Molva, R., Onen, M., & Loureiro, S. (2013). ClouDedup: Secure Deduplication with Encrypted Data for Cloud Storage. 2013 IEEE 5th International Conference on Cloud Computing Technology and Science,, PP No:363-370. doi:10.1109/cloudcom.2013.54

[2]. Zhou, Ruijin, Ming Liu, and Tao Li. "Characterizing the Efficiency of Data Deduplication for Big Data Storage Management." *2013 IEEE International Symposium on Workload Characterization (IISWC)* (2013) , PP No: 98-108. Web.

[3]. Luo, Shengmei, Guangyan Zhang, Chengwen Wu, Samee Khan, and Keqin Li. "Boafft: Distributed Deduplication for Big Data Storage in the Cloud." *IEEE Transactions on Cloud Computing* (2015) , PP No: 1-13. Web.

[4]. Yan, Zheng, Mingjun Wang, Yuxiang Li, and Athanasios V. Vasilakos. "Encrypted Data Management with Deduplication in Cloud Computing." *IEEE Cloud Computing* 3.2 (2016), PP No: 138-150. Web.

[5]. Wen, Mi, Kejie Lu, Jingsheng Lei, Fengyong Li, and Jing Li. "BDO-SD: An Efficient Scheme for Big Data Outsourcing with Secure Deduplication." *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (2015), PP no: 214-219. Web.

[6]. Waghmare, V., & Kapse, S. (2016). Authorized Deduplication: An Approach for Secure Cloud Environment. *Procedia Computer Science, 78*, PP no: 815-823. doi:10.1016/j.procs.2016.02.063

[7]. Chen, Ming et al. "A Duplicate Image Deduplication Approach Via Haar Wavelet Technology". 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems (2012): n. PP no: 624 - 628.doi:10.1109/CCIS.2012.6664249. Web. 10 Feb. 2017.

[8]. Xu, J., Zhang, W., Ye, S., Wei, J., & Huang, T. (2014). A Lightweight Virtual Machine Image Deduplication Backup Approach in Cloud Environment. *2014 IEEE 38th Annual Computer Software and Applications Conference*, 503-508. doi:10.1109/compsac.2014.73

[9]. Hur, J., Koo, D., Shin, Y., & Kang, K. (2016). Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage. *IEEE Transactions on Knowledge and Data Engineering, 28*(11), 3113-3125. doi:10.1109/tkde.2016.2580139

[10]. Kirubakaran, R., Prathibhan, C. M., & Karthika, C. (2015). A cloud based model for deduplication of large data. *2015 IEEE International Conference on Engineering and Technology (ICETECH)*. doi:10.1109/icetech.2015.7275007

[11]. Leesakul, W., Townend, P., & Xu, J. (2014). Dynamic Data Deduplication in Cloud Storage. *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, 321-325. doi:10.1109/sose.2014.46

[12]. Chen, R., Mu, Y., Yang, G., & Guo, F. (2015). BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication. *IEEE Transactions on Information Forensics and Security, 10*(12), 2643-2652. doi:10.1109/tifs.2015.2470221

[13]. Fu, Y., Jiang, H., & Xiao, N. (2012). A Scalable Inline Cluster Deduplication Framework for Big Data Protection. *Lecture Notes in Computer Science Middleware 2012,* 354-373. doi:10.1007/978-3-642-35170-9_18

[14]. Li, J., Li, J., Xie, D., & Cai, Z. (2016). Secure Auditing and Deduplicating Data in Cloud. *IEEE Transactions on Computers, 65*(8), 2386-2396. doi:10.1109/tc.2015.2389960

[15]. Pietro, R. D., & Sorniotti, A. (2016). Proof of ownership for deduplication systems: A secure, scalable, and efficient solution. *Computer Communications*, *82*, 71-82. doi:10.1016/j.comcom.2016.01.011

[16]. Q., Z., & X. (2010, December 3). Data Deduplication Techniques. *Future Information Technology and Management Engineering (FITME), 2010 International Conference on*, 430-433. doi:10.1109/FITME.2010.5656539

[17]. Dipeeka Shah, Gayatri S. Pandi (Jain). “A Comparative Survey of Optimizing Secure Deduplication for Big Data in Cloud”, 2016, International Institution for Technological Research and Development, Volume 1, Issue 6.

[18]. SearchStorage. (2016). what is data deduplication (intelligent compression or single-instance storage)? - Definition from WhatIs.com. [online] Available at: http://searchstorage.techtarget.com/definition/data-deduplication [Accessed 19 Sep. 2016].

[19]. Data Deduplication Overview- from technet.microsoft.com. Retrieved September 23, 2016, from https://technet.microsoft.com/en-us/library/hh831602(v=ws.11).aspx

[20]. Understanding Data Deduplication — and Why It’s Critical for Moving Data to the Cloud – from Dhruva.com. Retrieved September 29, 2016, from

http://www.druva.com/blog/a-simple-definition-what-is-data-deduplication/

Dipeeka Shah received the diploma and bachelor degree in information technology from Gujarat Technological University, in 2012 and 2015 respectively. Perusing master degree in computer engineering from Gujarat Technological University.

Gayatri S. Pandi (Jain). Presently, she is professor and head of department in PG Department of LJIET, Ahmedabad.