International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 03 - Issue 01 || January 2018 || PP. 44-49

Data Deduplication in Cloud Environment – A Survey

Sarah Prithvika P.C.[1], Ramani S.[2], Jakkulin Joshi J.[3] and Sindhu K.[4]
[1,2,3,4] *Assistant Professor, Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India*

Abstract: Data deduplication prevents duplicate copies of the data from being stored. A single copy of the data is stored and will be provided to all the authorized users. Data deduplication helps to use the cloud storage in a more efficient manner and also improves the bandwidth, but there are certain considerations like security, availability and integrity that must be kept in mind while choosing a particular data deduplication technique. To provide security, encryption is employed. Performing deduplication on encrypted data is a challenge. This paper attempts to discuss some of the data deduplication techniques.

Keywords: Availability, Deduplication, Integrity, Security

1. Introduction

Cloud computing is used to provide services using distributed and virtual technology. Users are provided with storage and computing services. The benefits of using the cloud server include savings in cost, better accessibility and improved scalability. Cloud computing technology provides services such as Software as a service (SaaS), Platform as a service (PaaS), Infrastructure as a service (IaaS) to the user on demand. Data is growing at an alarming growth. So data deduplication is a fast growing technology because it reduces storage needs by storing only a single copy of the data. Data deduplication is provided by various cloud service providers like Amazon, Dropbox and Google Drive. As more and more users use the cloud storage, the need for security and optimized storage arises and must be addressed. Encryption and deduplication are conflicting technologies and are challenging to implement. Also, deduplication offers the benefit of reduced replication, since only unique data is present in cloud.

2. Background

Data Deduplication can be performed at the source or target side. Performing data deduplication at target side [1], [2] is less intrusive than doing it at source side. Data deduplication can be done using inline method or post-processing method. In the former technique, redundancy is removed while the data is written and requires less data storage and in the latter, redundancies are removed after the data is written and so needs more storage. Data deduplication can be done at different levels like file level, block level, variable sized chunk level and at binary level. If two similar files exist, then only one copy of the file is saved. Block level deduplication works on a block of data. It is more flexible, but it requires more processing power. Blocks are fixed sized, so to increase effectiveness, the data can be divided into variable sized chunks. Data Deduplication can be done at binary level and for two identical files the bit pattern is the same, so it is saved only once. When the system does deduplication based on the contents of the file, it is called as content aware systems. By understanding the content of the data, we can deduplicate more efficiently.

If there are three files File1, File2 and File3, then if deduplication is not performed then the total space occupied by the files would be 18 KB, considering each file to be 6 KB as shown in Fig 1. If deduplication is performed, then only one copy of data is stored and others will point to it. Even if a portion of the file changes, only the changed data is saved because the rest of the data remains the same and has already been stored. File2 is the same as File1 so it is not stored, since File1 is already stored. In File3 the c has been changed to X, so only the change is stored and in this case it takes only 1 KB.

a b c d e f **File1**	a b c d e f **File2**	a b X d e F **File3**	
Size of files			**Total**
6 KB	6 KB	6 KB	18 KB
Size of file after deduplication			**Total**
6 KB	0 KB	1 KB	7 KB
Space savings after deduplication is 11 KB.			

Fig 1. Space savings after deduplication

Data deduplication and compression work on data to reduce the storage requirements. Compression looks at redundancies within a file and data deduplication looks at redundancies in segments across files.

2.1 Working of Deduplication

An input file or chunk of data is provided by the user. The hash value is calculated. Hashing is a CPU intensive process. If the hashing occurs in the production server itself it may reduce the performance of the system. So it is better to offload it or do it in the target. The problem of hash collision may also occur. If two different data produce the same hash value it is known as hash collision. This problem may occur but is rare. If the hash is already present in the hash index, then a pointer is set to the existing data, otherwise the data is stored and hash value is updated in the hash index table as shown in Fig 2.

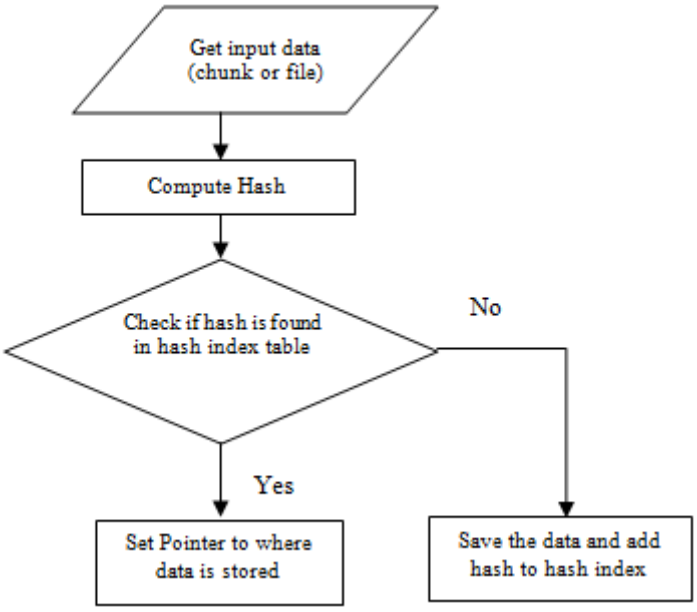

Fig 2. Deduplication Process

Ensuring proof of ownership is seen in [3], [4] and [5]. Proof of ownership issues arise when an attacker lays hold of the hash value and claims rights to the file. Convergent Encryption produces a key from the data itself. This way, different users will encrypt the same message with the same key and so the same cipher text will be produced. So deduplication and encryption can be done. But this raises the problem that if the attacker generates a key from a plaintext and encrypts it, then it can be checked if the resulting cipher text is already present in the cloud. So to avoid this, we can have an additional layer that will perform encryption and decryption.

2.2 Applications of Data Deduplication

Data Deduplication can be applied within a file, across files, across applications, across clients over a period of time [6]. It is used in archiving and backups of file systems, databases that have a low change rate, VMware environment, NAS, LAN and SAN.

3. Related Work

The idea discussed in [7] deals with privacy preserving in a cross-domain architecture. If the information is leaked, then the cloud storage provider is held accountable. Here, there are 3 tiers; a key distribution center, a cloud service provider, clients belonging to different domains and the corresponding managers. The key distribution center deals with the distribution and management of the keys. The cloud service provider provides storage and performs deduplication on messages from different domains. The domains have a contract with the cloud service provider. Each domain has a local manager who is responsible for deduplication in that domain. Each client is affiliated with a domain.

The client encrypts the message and sends it to the domain manager. Clients need to upload encrypted data and also two tags. This scheme ensures data confidentially, privacy-preserving, availability and accountability, while resisting brute-force attacks. It also improves the time complexity of duplicate search. In [8], a decentralized server-aided encryption technique is discussed for data deduplication. In a centralized key distribution center a single key is used, this is not a very secure method, so a decentralized approach is used. A

decentralized approach also avoids having to pre-share the keys. Deduplication is performed across tenants and keys are produced from different key distribution centers. Each tenant has a key server at its premises which generates the convergent key. The cloud service provider checks if two cipher texts encrypted using different keys have the same plain text. A blind signature scheme is used, so only the cipher text is revealed, sensitive information is not revealed. The signature is obtained from the user, but the signer will not know the message. So, unauthorized access is prevented. Brute force attack is not possible. Parallelism of modern CPUs can be used to improve the effectiveness. The problem with this approach is that it requires more storage space.

In [5], a randomized tag is used. The time taken to identify the duplicate cipher text is greatly reduced. Decision tree structures are created for static and dynamic data. The static deduplication is built based on the random elements from the client, which does not allow the tree to be altered. Dynamic deduplication tree is built, based on the self generation tree, which allows the server to do updates, inserts and some other optimizations. A decision tree is a tree like model for decision support. The first decision is indicated by the root. The nodes are connected by branches. The termination node indicates the final outcome of that particular branch. In this data duplication scheme only one element is found or the element is absent. Elements can be inserted and deleted. It is desirable to keep the tree small. This scheme does not leak anything other than the data for the deduplication path choosing in the tree. An adversary cannot differentiate between two sequences as long as they have the same tree path. In [9] mutual ownership verification happens and proof of ownership is checked. This scheme provides dynamic ownership management. A user may wish to remove data from the cloud. After the request is processed by the cloud server, the revoked user must not be given access to the data. Similarly, if the data uploaded by user is already present in the cloud, the user must be authorized to access the data. This scheme supports inter user file level and intra user block level deduplication. Data consistency is also ensured.

4. Multimedia Deduplication

Multimedia data is used in fields such as military communication, medical imaging, multimedia systems and much more. It is not a good idea to encrypt multimedia data, because the size of multimedia is usually large and so a lot of time is needed to encrypt data. It is not necessary that the encrypted audio/video file must be exactly same as a decrypted file, because of the nature of human perception. Bit by bit comparison is no longer a good method for multimedia deduplication. Watermark can be used in images to provide copyright information. It is not suitable for large images. In [10] a secure image deduplication occurs through compression. It uses embedding of a partial encryption and image hashing into the compression algorithm to perform deduplication. The encryption scheme is used to secure the data and the hashing is done to classify the images so that deduplication can be performed. In [11] we see that computer vision technique is used in the image retrieval problem. Deduplication of electricity bills has been attempted using block truncation coding which is applied to gray scale images and it is lossy compression technique. Images are put into clusters based on block size and duplicates in clusters are removed.

A perception hash can be used to produce a small digest from an image file where even if the image changes a little bit, it creates a small change in the hash value. This is different from the cryptographic hash function, where a small change in the data may have a huge change in the hash value. This is used to compare the images. If the features are similar, then the perception hashes are close. It extracts some features from the image and calculates a hash value. Authentication is performed by looking at the hashed value of original image and image to be authenticated. It accepts content preserving manipulations and rejects malicious manipulations.

Perception hashing has 4 stages. They are transformation stage, feature extraction stage, quantization stage and compression and encryption stage. The compression and encryption stage is usually ignored because the quantization stage is very difficult if it precedes the compression and encryption stage because we do not know the behavior of the extracted features after content preserving/changing manipulations. The quantization stage enhances robustness and increases randomness to lessen collisions in the system. Some examples of content preserving manipulations are noise addition, scaling, rotation, cropping and contrast adjustment. Some examples of content changing manipulations are removing/adding objects, moving objects, changing the color or texture. In the transformation phase, the image undergoes special/frequency transformations to make all the features that have been extracted depend on the image pixels or its frequency coefficients. In the extraction stage a continuous hash vector is generated. The quantization step converts the continuous perceptual hash vector into binary perceptual hash string. Finally the binary perceptual hash string is compressed and encrypted.

5. Record Deduplication

This refers to the finding of records in a dataset that refer to the same object across different sources. By avoiding duplicates we can ensure repositories with concise data and without any replicas. It also enables high quality data retrievals, savings in computational time and efficient storage utilization. This occurs when there are no common identifiers in the data sets because of differences in record shape and storage location.

Record deduplication usually arises when data is obtained from different sources using different ways. It can also be found in creating OCR documents, digital libraries and e-commerce sites. This may be because the way in which structured and unstructured data is stored.

So it is necessary to find a deduplication function that finds duplicate data in the repository. In the data preparation stage the structural heterogeneity problem is avoided by storing it in a uniform manner in the database. Structural heterogeneity occurs when the fields are structured differently. In one database the address field may be one and in another database the address may be stored in more than one field as address line 1,2,3,4. Lexical heterogeneity occurs when the data representation is different. For example in one database the address may be represented as 1st street and in another database it may be represented as first street. The preparation phase consists of data parsing, transformation of data and standardization step.

After this stage any one technique can be applied to find duplicates. In the active learning technique [12] the learner has to automatically come up with potentially confusing pairs of data. The user has to label it as duplicate or not. The task of the user is easier than that of learner. This method is not always suitable because it requires training model. An alternative is to use distance based techniques. This technique does not need training data. Each record is considered to be a field and field matching techniques are used to find similarity between the records. The problem with this method is that the value for matching threshold has to be defined. Rule based approach can also be taken and is basically a distance based approach with the distance between records as either 0 or 1. Unsupervised learning approach is based on the fact that similar vectors relate to the same class. Algorithms group together similar comparison vectors.

Genetic programming (GP) approach can also be used. It is the evolution of programs or algorithms for supervised learning. This helps to avoid dirty data. The optimal solution may not be found, but we can find a near optimal solution. The most commonly used GP representations are trees and graphs. The two main operations in GP are crossover and mutation. Crossover is performed by choosing a node in each sub tree and then exchanging the sub trees. Mutation is randomly performed operations on an individual item. This finds a point in the GP tree randomly and replaces the existing sub tree with the newly generated sub tree. The whole process is repeated until the maximum number of generations is reached or target is achieved. In the end the best item is chosen as the solution to the problem.

6. Deduplication for Big Data

A lot of data is frequently generated and updated and this voluminous data is called as big data. Big data handles structured data, unstructured data and semi-structured data. Doing this in real time is a challenging task. A distributed system is one way to handle it. Even though data is distributed across the systems, it is not known to the user. Some of the challenges in such a system are scalability, reliability, replication, availability and cost-effectiveness. To overcome these problems, we use Hadoop. Hadoop is a open source framework that is distributed and is used for handling large quantities of data. The main advantages in using Hadoop are robustness, accessibility, scalability, cost effectiveness and simplicity. It is designed using low cost hardware and is fault tolerant. Data is stored redundantly in systems so as to provide fault tolerance. Replication is done in the Hadoop system to increase the availability of data, so the amount of storage needed also increases. To implement deduplication hash values are computed using MD5 and SHA1 algorithms [13]. A hash is a bit that is 128 bit for MD5 and 160 bit for SHA-1. The hash value is passed to the HDFS system. If the hash value is already present, then the data need not be uploaded, otherwise it is uploaded. Memory utilization is handled by Hadoop distributed file system (HDFS). Hadoop uses the mapreduce paradigm. The named node is the node that manages the HDFS. The named node is called as master. The data node is the node where the data is present before it is processed. The data node is called as slave. The master node is the node where the job tracker runs and it accepts requests from the client. The slave node is where the mapreduce function happens. The job tracker will schedule and assign jobs to task tracker. The task tracker tracks the task and reports status to job tracker. The secondary node is used for having checkpoints in HDFS.

Mapreduce consists of two tasks; the map and the reduce step as shown in Fig 3. Map takes a set of data and converts it into some other data and data is represented as a tuple with key/value pair. The reduce step takes the input from the map step and reduces the number of tuples. In the map step, the input in a file or directory from the HDFS system is processed to produce several small units of data. The reduce step works on the output of the map phase and produces a new output that is stored in the HDFS. The main advantage is that scaling can be done over several nodes. Once the application is written for the mapreduce framework, then it becomes very easy to scale it.

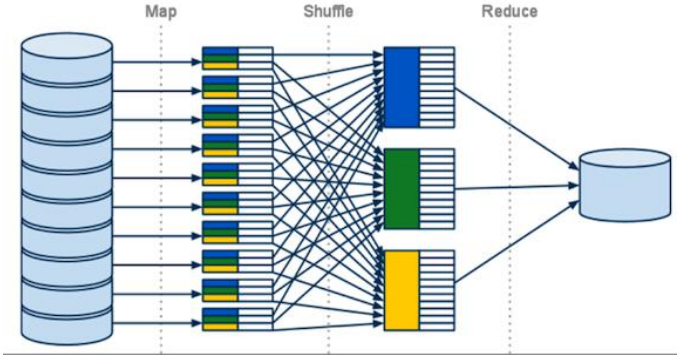

Fig 3.Map Reduce Framework

In [14], we see a approach to do deduplication on big data using a method called Boafft. It is a cluster based deduplication scheme. The server has storage as well as computational capability. It uses a data routing algorithm based on data similarity to reduce the network overhead. It avoids random number of disk reads and writes which in turn increade the deduplication process. To achieve this it maintains a similarity index on each server. It achieves good load balance. The data deduplication ratio measures the effectiveness of the deduplication process. So this method improves the ratio based on the access frequency.

There are many inline cluster deduplication techniques used to improve reliability and efficiency. Some of the challenges faced in these systems are reduced deduplication ratio as the number of nodes increases, high communication overhead and difficulty in load balancing. AR-Dedupe system proposed in [15] discusses about cluster deduplication scheme that uses an application-aware mechanism which further improves the performance. It has three divisions; the backup client, the metadata and the deduplication server nodes. It splits the bulky data into pieces and finds the fingerprint and it can be represented in the backup client. Based on the routing mechanism it sends the chunks to the server nodes. Thus this scheme balances the load, improves deduplication ratio and reduces the communication overhead.

7. Conclusion

Thus, some of the techniques used for data deduplication while considering security, availability and integrity in cloud environment have been studied. While performing deduplication a compromise must not be made on the security, availability or the integrity of the data.

References

[1]. SNIA,"*Advanced Deduplication Concepts*,"[online] 2011.Available from *http://www.snia.org/sites/default/education/tutorials/2011/fall/DataProtectionManagement/ThomasRiveria_Advanced_Dedupe_Concepts_FINAL.pdf*

[2]. *http://searchdatabackup.techtarget.com/tip/Where-and-how-to-use-data-deduplication-technology-in-disk-based-backup*

[3]. Yang, Chao, Jianfeng Ma, and Jian Ren. *"Provable Ownership of Encrypted Files in De-Duplication Cloud Storage." Ad Hoc & Sensor Wireless Networks 26.1-4(2015): 43-72*

[4]. Li, Jin, Yan Kit Li, Xiaofeng Chen, Patrick PC Lee, and Wenjing Lou. *"A hybrid cloud approach for secure authorized deduplication." Parallel and Distributed Systems, IEEE Transactions on 26, no. 5 (2015): 1206-1216.*

[5]. Tao Jiang, Xiaofeng Chen, Qianhong Wu, Jianfeng Ma, Member, Willy Susilo and Wenjing Lou. "*Secure and Efficient Cloud Data Deduplication With Randomized Tag*", *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 12, NO. 3, MARCH 2017.*

[6]. Dutch T Meyer and William J Bolosky."*A study of practical Deduplication". ACM Transactions on Storage (TOS),7(4):14, 2012.*

[7]. Xue Yang, Rongxing Lu, Kim Kwang Raymond Choo, Fan Yin, and Xiaohu Tang. *"Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud", 10.1109/TBDATA.2017.2721444, IEEE Transactions on Big Data*

[8]. Youngjoo Shin, Dongyoung Koo, Joobeom Yun and Junbeom Hur. "*Decentralized Server-aided Encryption for Secure Deduplication in Cloud Storage*", *10.1109/TSC.2017.2748594, IEEE Transactions on Services Computing*

[9]. Shunrong Jiang, Tao Jiang and Liangmin Wang. *"Secure and Efficient Cloud Data Deduplication with Ownership Management", 10.1109/TSC.2017.2771280, IEEE Transactions on Services Computing*

[10]. Fatema Rashid, Ali Miri and IsaacWoungang, *"Secure image deduplication through image compression", Elsevier, Volumes 27– 28, April–May 2016, Pages 54-64*

[11]. Anum Javeed Zargar, Ninni Singh, Geetanjali Rathee and Amit Kumar Singh, *"Image data-deduplication using the block truncation coding technique", IEEE, 10.1109/ABLAZE.2015.7154986*

[12]. Kristi Morton, Hannaneh Hajishirzi, Magdalena Balazinska, Dan Grossman, "*View-Driven Deduplication with Active Learning", arXiv:1606.05708v1*

[13]. S.Ranjitha, P.Sudhakar, K.S.Seetharaman, *"A Novel and Efficient De-duplication System For HDFS", ELSEVIER, ScienceDirect, Procedia Computer Science 92 (2016) 498 – 505*

[14]. ShengmeiLuo, Guangyan Zhang, Chengwen Wu, Samee U. Khan *"Boaftt: Distributed Deduplication for Big Data Storage in the Cloud" IEEE, 2015, pp. 1-13*

[15]. Xing Yu-Xuan, XlaoNong, Liu Fang, "AR-Dedupe: An Efficient Deduplication Approach for Cluster Deduplication System", Shanghai Jiaotong University and Springer-Verlag Berlin Heidelberg, 2015,pp. 76 -81.