ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

An Approved Data Deduplication Mechanism for Removing Identical Copies of Repeating Data in Cloud Environment

G.Veena
M.Tech (CSE)
Vignana Bharathi Institute of Technology.

Rajasekhar Jelli
Assistant Professor,
Vignana Bharathi Institute of Technology.

Abstract:
In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred can be greatly reduced.In order to improve data security, in this paper we propose a system, where the privileges of users are also considered along with the data itself. We examine and implement An authorized duplicate check scheme for removing duplicate copies of repeating data in the cloud Environment to Reduce Amount of Storage Space.

Keywords: ***Data deduplication, User Privileges, Storage, Cloud, Security.***

Introduction:

Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most organizations, the storage systems contain duplicate copies of many pieces of data. For example, the same file may be saved in several different places by different users, or two or more files that aren't identical may still include much of the same data. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Companies frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage as well.

In its simplest form, deduplication takes place on the file level; that is, it eliminates duplicate copies of the same file. This kind of deduplication is sometimes called file-level deduplication or single instance storage (SIS). Deduplication can also take place on the block level, eliminating duplicated blocks of data that occur in non-identical files. Block-level deduplication frees up more space than SIS, and a particular type known as variable block or variable length deduplication has become very popular. Often the phrase "data deduplication" is used as a synonym for block-level or variable length deduplication.

Deduplication Benefits

The primary benefit of data deduplication is that it reduces the amount of disk or tape that organizations need to buy, which in turn reduces costs. NetApp reports that in some cases, deduplication can reduce storage requirements up to 95 percent, but the type of data you're trying to deduplicate and the amount of file sharing your organization does will influence your

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

own deduplication ratio. While deduplication can be applied to data stored on tape, the relatively high costs of disk storage make deduplication a very popular option for disk-based systems. Eliminating extra copies of data saves money not only on direct disk hardware costs, but also on related costs, like electricity, cooling, maintenance, floor space, etc.
Deduplication can also reduce the amount of network bandwidth required for backup processes, and in some cases, it can speed up the backup and recovery process.

Deduplication vs. Compression

Deduplication is sometimes confused with compression, another technique for reducing storage requirements. While deduplication eliminates redundant data, compression uses algorithms to save data more concisely. Some compression is lossless, meaning that no data is lost in the process, but "lossy" compression, which is frequently used with audio and video files, actually deletes some of the less-important data included in a file in order to save space. By contrast, deduplication only eliminates extra copies of data; none of the original data is lost. Also, compression doesn't get rid of duplicated data -- the storage system could still contain multiple copies of compressed files.

Deduplication often has a larger impact on backup file size than compression. In a typical enterprise backup situation, compression may reduce backup size by a ratio of 2:1 or 3:1, while deduplication can reduce backup size by up to 25:1, depending on how much duplicate data is in the systems. Often enterprises utilize deduplication and compression together in order to maximize their savings.

Dedupe Implementation

The process for implementing data deduplication technology varies widely depending on the type of product and the vendor. For example, if deduplication technology is included in a backup appliance or storage solution, the implementation process will be much different than for standalone deduplication software.

In general, deduplication technology can be deployed in one of two basic ways: at the source or at the target. In source deduplication, data copies are eliminated in primary storage before the data is sent to the backup system. The advantage of source deduplication is that is reduces the bandwidth requirements and time necessary for backing up data. On the downside, source deduplication consumes more processor resources, and it can be difficult to integrate with existing systems and applications.

By contrast, target deduplication takes place within the backup system and is often much easier to deploy. Target deduplication comes in two types: in-line or post-process. In-line deduplication takes place before the backup copy is written to disk or tape. The benefit of in-line deduplication is that it requires less storage space than post-process deduplication, but it can slow down the backup process. Post-process deduplication takes place after the backup has been written, so it requires that organizations have a great deal of storage space available for the original backup. However, post-process deduplication is usually faster than in-line deduplication.

Deduplication Technology and overview

Data deduplication is a highly proprietary technology. Deduplication methods vary widely from vendor to vendor, and many of those methods are patented. For example, Microsoft has a patent on single instance storage. In addition, Quantum owns a patent on variable length deduplication. Many other vendors also own patents related to deduplication technology.
Deduplication may occur "in-line", as data is flowing, or "post-process" after it has been written.

Post-process deduplication

With post-process deduplication, new data is first stored on the storage device and then a process at a later time will analyze the data looking for duplication. The benefit is that there is no need to wait for the hash calculations and lookup to be completed before storing the data thereby ensuring that store performance is not degraded. Implementations offering policy-based operation can give users the ability to defer

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

optimization on "active" files, or to process files based on type and location. One potential drawback is that you may unnecessarily store duplicate data for a short time which is an issue if the storage system is near full capacity.

In-line deduplication

This is the process where the deduplication hash calculations are created on the target device as the data enters the device in real time. If the device spots a block that it already stored on the system it does not store the new block, just references to the existing block. The benefit of in-line deduplication over post-process deduplication is that it requires less storage as data is not duplicated. On the negative side, it is frequently argued that because hash calculations and lookups takes so long, it can mean that the data ingestion can be slower thereby reducing the backup throughput of the device. However, certain vendors with in-line deduplication have demonstrated equipment with similar performance to their post-process deduplication counterparts.

Post-process and in-line deduplication methods are often heavily debated.

Source versus target deduplication

Another way to think about data deduplication is by where it occurs. When the deduplication occurs close to where data is created, it is often referred to as "source deduplication". When it occurs near where the data is stored, it is commonly called "target deduplication".

Source deduplication ensures that data on the data source is deduplicated. This generally takes place directly within a file system. The file system will periodically scan new files creating hashes and compare them to hashes of existing files. When files with same hashes are found then the file copy is removed and the new file points to the old file. Unlike hard links however, duplicated files are considered to be separate entities and if one of the duplicated files is later modified, then using a system called copy-on-write a copy of that file or changed block is created. The deduplication process is transparent to the users and backup applications. Backing up a deduplicated file system will often cause duplication to occur resulting in the backups being bigger than the source data.

Target deduplication is the process of removing duplicates of data in the secondary store. Generally this will be a backup store such as a data repository or a virtual tape library.

Deduplication methods

One of the most common forms of data deduplication implementations works by comparing chunks of data to detect duplicates. For that to happen, each chunk of data is assigned an identification, calculated by the software, typically using cryptographic hash functions. In many implementations, the assumption is made that if the identification is identical, the data is identical, even though this cannot be true in all cases due to the pigeonhole principle; other implementations do not assume that two blocks of data with the same identifier are identical, but actually verify that data with the same identification is identical.[6] If the software either assumes that a given identification already exists in the deduplication namespace or actually verifies the identity of the two blocks of data, depending on the implementation, then it will replace that duplicate chunk with a link.

Once the data has been deduplicated, upon read back of the file, wherever a link is found, the system simply replaces that link with the referenced data chunk. The deduplication process is intended to be transparent to end users and applications.

Chunking

Between commercial deduplication implementations, technology varies primarily in chunking method and in architecture. In some systems, chunks are defined by physical layer constraints (e.g. 4KB block size in WAFL). In some systems only complete files are compared, which is called single-instance storage or

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

SIS. The most intelligent (but CPU intensive) method to chunking is generally considered to be sliding-block. In sliding block, a window is passed along the file stream to seek out more naturally occurring internal file boundaries.

Client backup deduplication

This is the process where the deduplication hash calculations are initially created on the source (client) machines. Files that have identical hashes to files already in the target device are not sent, the target device just creates appropriate internal links to reference the duplicated data. The benefit of this is that it avoids data being unnecessarily sent across the network thereby reducing traffic load.

Primary storage and secondary storage

By definition, primary storage systems are designed for optimal performance, rather than lowest possible cost. The design criteria for these systems is to increase performance, at the expense of other considerations. Moreover, primary storage systems are much less tolerant of any operation that can negatively impact performance. Also by definition, secondary storage systems contain primarily duplicate, or secondary copies of data. These copies of data are typically not used for actual production operations and as a result are more tolerant of some performance degradation, in exchange for increased efficiency.

To date, data deduplication has predominantly been used with secondary storage systems. The reasons for this are two-fold. First, data deduplication requires overhead to discover and remove the duplicate data. In primary storage systems, this overhead may impact performance. The second reason why deduplication is applied to secondary data, is that secondary data tends to have more duplicate data. Backup application in particular commonly generate significant portions of duplicate data over time.

Data deduplication has been deployed successfully with primary storage in some cases where the system design does not require significant overhead, or impact performance.

EXISTING SYSTEM:

1. Data deduplication systems, the private cloud is involved as a proxy to allow data owner/users to securely perform duplicate check with differential privileges.
2. Such architecture is practical and has attracted much attention from researchers.
3. The data owners only outsource their data storage by utilizing public cloud while the data operation is managed in private cloud.

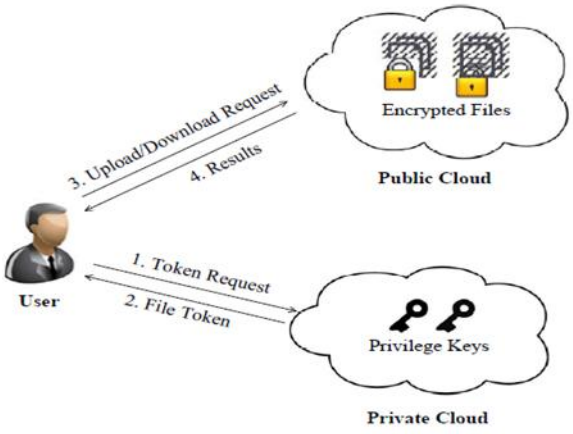
DISADVANTAGES OF EXISTING SYSTEM:

1. Traditional encryption, while providing data confidentiality, is incompatible with data deduplication.
2. Identical data copies of different users will lead to different ciphertexts, making deduplication impossible.

PROPOSED SYSTEM:

In this paper, we enhance our system in security. Specifically, we present an advanced scheme to support stronger security by encrypting the file with differential privilege keys. In this way, the users without corresponding privileges cannot perform the duplicate check. Furthermore, such unauthorized users cannot decrypt the cipher text even collude with the S-CSP. Security analysis demonstrates that our system is secure in terms of the definitions specified in the proposed security model.

SYSTEM ARCHITECTURE:



ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

ADVANTAGES OF PROPOSED SYSTEM:

1 The user is only allowed to perform the duplicate check for files marked with the corresponding privileges.

2 We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys.

3 Reduce the storage size of the tags for integrity check. To enhance the security of deduplication and protect the data confidentiality,

Conclusion:

Whenever data is transformed, concerns arise about potential loss of data. By definition, data deduplication systems store data differently from how it was written. As a result, users are concerned with the integrity of their data. The various methods of deduplicating data all employ slightly different techniques. However, the integrity of the data will ultimately depend upon the design of the deduplicating system, and the quality used to implement the algorithms. As the technology has matured over the past decade, the integrity of most of the major products has been well proven. We also presented several new deduplication constructions supporting authorized duplicate check in hybrid cloud architecture, in which the duplicate-check tokens of files are generated by the private cloud server with private keys. Security analysis demonstrates that our schemes are secure in terms of insider and outsider attacks specified in the proposed security model. As a proof of concept, we implemented a prototype of our proposed authorized duplicate check scheme and conduct testbed experiments on our prototype. We showed that our authorized duplicate check scheme incurs minimal overhead compared to convergent encryption and network transfer.

References:

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, “A Hybrid Cloud Approach for Secure Authorized Deduplication”, IEEE Transactions on Parallel and Distributed Systems, 2014.

[2] David Geer, "Reducing the Storage Burden via Data Deduplication.computer.org‖, December 2008.

[3]https://www.daniweb.com/images/attachments/0/WP_Deduplication_US_Letter_090702.pdf

[4]http://en.wikipedia.org/wiki/Data_deduplication#Deduplication_overview

[5]http://www.computerworld.com/article/2474479/data-center/data-deduplication-in-the-cloud-explained--part-one.html

[6] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P. C. Lee, Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEETransactions on Parallel and Distributed Systems,Volume:PP,Issue:99,Date of Publication :18.April.2014

[7] Yang Zhang, Yongwei Wu and Guangwen Yang, Droplet: a Distributed Solution of Data Deduplication, ACM/IEEE 13th International Conference on Grid Computing,2012

[8] PasupuletiPavani& Mohammed Alisha, Secure Data Deduplication with Efficient Key Management in Cloud Databases, International Journal & Magazine of EngineeringTechnology Management and Research (IJMETMR) , Vol: 3 (2016), Issue No: 2 (February), http://www.ijmetmr.com/olfebruary2016/PasupuletiPavani-MohammedAlisha-33.pdf

[9] Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and AkhileshSaxena, Intelligent Workload Factoring for A Hybrid Cloud Computing Model , Published by the IEEE Computer Society ,2009

[10] Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster, Virtual Infrastructure Management in Private and Hybrid Clouds, Published by the IEEE Computer Society, 2009

[11] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, Duplicate Record Detection: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007

[12] Srivatsamaddodi, Girija V. Attigeri, DrkarunakarA.k, Data Deduplication Techniques and Analysis. Third International Conference on Emerging Trends in Engineering and Technology IEEE, 2010