ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

A COMPARATIVE STUDY ON DATA DEDUPLICATION TECHNIQUES IN CLOUD STORAGE

B.Tirapathi Reddy*[1], U.Ramya[2], Dr.M.V.P Chandra Sekhar[3]
KLUniversity,
Email: tirapathireddyb@kluniversity.in

Received on 06-08-2016 *Accepted on 10-09-2016*

Abstract:

Now A days with the wide popularity of cloud computing many users of the computer system are using cloud computing for variety of services. Cloud offers various services to its users on a pay as you use basis. With the rapid growth of data rates with every user, users are attracted towards the cloud storage to store huge volumes of data. But with this there is a possibility that most of the data available in cloud storage is redundant. We need to eliminate the redundant copies of the data to utilize the cloud storage space effectively; thereby we can reduce the storage costs of a user or organization. There are various DE duplication techniques available to eliminate the redundant data items so that we can keep only one copy of the data item. This paper deals with various mechanisms available to eliminate redundant copies of the data, and also addresses the drawback and advantages of all these mechanisms.

Keywords: DE duplication, Cloud storage, Block level, file level, Client side, Server side.

Introduction:

In the present scenario of the technological world, each and every user handles lot of data and information. Each user has data for his/her personal as well as professional usage. But the internal storage capacity of each user PC doesn't support much to this cause. Many researchers tried to find a solution for this storage problem as a result finally termed a solution where any user can store his unlimited data and information in a centralized distributed storage location which is named as a Cloud. A cloud is a combination of many services: Infrastructure-as-a-service, Platform-as-a-service and Software-as-a-service. Cloud provides a user with the required hardware and software tools that is required for him/her or any application that the user wants over the internet. This is done using platform-as-a-service. Infrastructure-as-a-service is a cloud computing model where a third force hosts various set of virtualized computing resources over the internet. Software-as-a-service is a software licensing model where the trending and legacy software's are hosted at a central location and is licensed on a subscription basis. These are accessed most often via a

thin client namely web browser. As lot of user's information has to be outsourced onto the cloud, there is a huge requirement of transmission bandwidth and storage capacity. These two aren't a big concern in the present scenario but in the years to come the data will be increased by many folds as a result the transmission bandwidth and the storage capacity are going to be a huge concern. Cloud Computing is a term based on utilities and consumption of virtualized resources. It provides deployment models and services to the cloud. It allows many kinds of information resources to be uploaded for existing process to render computing reports without any needto store data on the cloud. Cost efficient method to use, maintain and upgrade and it has unlimited space to store data on cloud and has automatic software integration, very easy to access the data, last and important is we have quick deployment even though it takes time to complete. Cloud is used as pay-as-you-go. Cloud computing is a new environment that deploys on infrastructure deployment which support on-demand services and software. Cloud storage refers to store the information online in the cloud. It offers you to store, manage and access data very easily on cloud's infrastructure. Data can be stored in digital, physical and logical pools with multiple spans of servers. Cloud storage improves the applications performance to keep costs low. Storage has greater elasticity and scalable to backup data, no need to purchase any storage from the cloud. We can access it for free to store our data in the cloud.

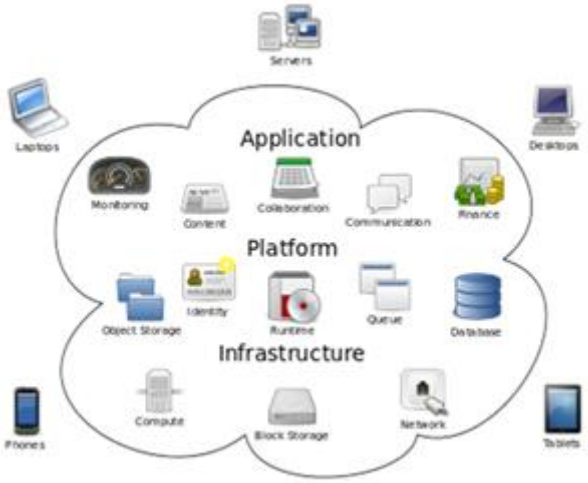


Fig: Cloud Architecture.

Introduction of cloud storage encourages organizations to upload huge amount of their data to the third party cloud storage providers. One of the challenges of the current day cloud storage providers is the overhead involved in managing the ever greater volume of data. Cloud storage is the data storage in which the cardinal data is stored in coherent mere, where the physical cache compasses multiple servers, and the physical environment is customarily owned and primed by a hosting company [6]. Cloud storage providers are merely responsible for keeping the data accessible and available to the outside world. People and management buy or charter repository capacity from the

CSP's to store the user's or the organization's data. Cloud storage services may be amassed through a co-located cloud computer provision, a web service (API)application programming interface, such as cloud desktop storage, a cloud storage access or web-based gratified management systems [8]. Cloud storage is 1) Serene of multiple distributed resources, but still acts as one, 2) Highly fault tolerant through dispersal of data and redundancy, 3) Vastly resolute through the creation of versioned copies, 4) Typically consistent with regard to data replicas.Cloud storage is also known as Data storage of a service. It is more scalable due to rapid elasticity to store capabilities that are delivered through internet. Today's cloud service provider's offers huge storage space and massive parallel computing resources at low priced. It is related to data cleaning approach in data management systems. In data management systems remove the duplicates in organized data records. Previous concise query reduces the some duplicate records and display the results with computational saving and resources. Compare to previous approach Using Block level display the effective with better evidence results. These results are best suitable results. After removing the duplicate records, apply the suggested function. It will contain less computational resource's utilization in implementation. Compared to all previous approaches, this approach provides fewer burdens, efficient and accurate results display. Cloud service provider is the top brass of ever increasing in volume of data. With high increasing amount of data produced worldwide, multiple users are becoming very popular, Data security is provide to prevent all the users from migration information to current storage. The data need to encrypt before it leaves the owner's state. DE-duplication uses as a compression mechanism in the cloud storage. File storage is a frequent storage used on hard drives and network attached storage systems. File level storage is quite opposite to the block level storage. In file level the size is based upon the usage of message and whereas block level stores as per the memory size. File level is very easy to use and implement and cost is less compared to block level storage. Each user has to get his/her own authorization to check permissions based on tokens for duplicate check in private cloud server through differential authorization. Authorized duplicate check checks for the duplicate directly it doesn't allow any users without duplicate check.

Existing System:

To make data management feasible and flexible, DE duplication has been the widely used mechanism. DE duplication is the process of excluding redundant copies of data and maintaining only one copy of the data on the cloud storage. DE duplication can be performed on the client-side as well as the server side. A client side DE duplication mechanism save the upload bandwidth costs but require lot of computing capacity and consumes so much time. Server side DE duplication saves the time at the client end but incurs lot of bandwidth costs. There are many

client side as well as server side DE-duplication mechanisms. In turn the DE duplication can be applied at the block level or file level. Block level DE duplication mechanism helps to eliminate redundant copies of the data.

DE-duplication Techniques are in compatible with conventional encryption mechanisms.

The major DE-duplication techniques that were being implemented nowadays are the "Popular-unpopular file strategy" and "Weak Leakage-Resilient Client-side DE-duplication of Encrypted Data in Cloud Storage" [2][3]. In the popular-unpopular technique of DE duplication, basically the files are classified into two types- one is the popular file and another is the unpopular file. A file comes under popular file if the number of uploads of the same file is greater than a threshold value (say 100). The popular files will be given only the first (and the only one) layer of security namely the symmetric encryption where the file is encrypted by using any one of the encryption algorithm. If the file is not a popular file i.e. if the uploads of the file is below the threshold value, the file will be secured under two layers of security. The first layer of security is same as in the popular file case and the second layer of security is the convergent encryption where the plaintext will be encrypted using the key that was obtained from the plaintext. This convergent encryption provides greater level of security to the file so that the plaintext will not be looked upon by any other intruder. The disadvantage with this system is that privacy is maintained only for the users who own unpopular files, whereas the privacy is degraded for the owners of popular files. Another issue related with this system is that a malicious storage provider may refuse to remove the uploaded file even after the user's request to delete his copy of the file. Data Integration applies for removing the duplicate records. In Data integration time apply the different kinds of concepts like record matching, record linkage and concise query data. It's not possible to remove the total number of duplicate records. It can take more amount time for extraction of results. There are four types of DE duplication, depends on DE duplication happens at client side before upload. It is more effective when triggers at the client side. As per security concerns of DE duplication we have, 1. Harnick- leads to data leakage in cloud a storage system which takes place in client side DE duplication. 2. Convergent Encryption key- when two persons are trying to upload same copy of a file it doesn't allows and it attacks the duplicate copies of unwanted data, combines confidentiality of data with DE duplication as possible. Convergent encryption key consists of encryption message with plaintext using symmetric scheme with a key. It keeps changing to guess the content of a message. 3. Message locking system- Highlights the authentication analysis of encrypted message which provides confidentiality for predictable information and gains semantic security. It is confidential for unused data. POW scheme- provides client side DE duplication in abounded leakage setting and allows a security proof for their solutions in a random oracle model. IT does not addresses entropy files. POW acts as an interactive method used by the users and storage

server. Dupless- It's a server encryption aided message for DE duplicated cloud storage. It uses the functions of convergent encryption key for key generation to get secure components. Dupless provides more security to server side data, but we targets for secure client side DE duplication. The second technique is "Weak Leakage-Resilient Client-side DE duplication of Encrypted Data in Cloud Storage" [10] where the first user A wants to upload a file F to the cloud storage, he chooses a random AES key and produces double cipher texts. By encrypting file F with encryption key using AES method which produces the first cipher text, and the second cipher text is generated by encrypting the short AES key with file F as the encryption key using an encryption method. Finally the user A will send a hash value and the two cipher texts generated to the cloud storage server. Finally, the storage server will compute the hash of the first cipher text and inserts it into its database. If another user B wants to upload the similar file F into the cloud which has been already uploaded, B needs to send the hash (F) to the storage server and if it finds this value in its database the server requests few computations by user B to verify whether he is the owner of the file. To verify the cloud server sends a short cipher text to user B. The user B decrypts the text sent by the cloud server using F and obtains the AES key. The user B now encrypts file F with the obtained AES key and computes the hash of the first cipher text and sends it back to the cloud. If the hash value existing in the storage server and the value computed by the user are equal, then the user B will be given access to that file and will be allowed to download it from the cloud storage. The possible attacks with this system are the intruder may somehow find out the hash value stored in the cloud and fool the cloud server as if he is the actual owner of the file and may try to grab access to the file stored in the cloud. Another attack is possible if the cloud server don't verify whether there exists a file which contains the hash that is stored in the cloud. The intruder may substitute the encrypted text with a file of equal size which is manipulated before uploading it to the cloud

DE duplication checks for the confidential data and provides duplicate check for hybrid cloud storage approach. DE duplication based system needs solutions to thedata of types which they are expected to handle, while using convergent encryption key DE duplication is more feasible to enforce the confidential data, convergent key generates identical copies to produce same cipher text. Here, convergent encryption key uses for both encryption and decryption of data copy. Our analysis is mainly focused on the output source data set which comprises instances of others and instances of some data items. Like AAMY to backup tools which links and clones of virtual machine images. Server points to the encrypts file so that user can download the file and provides decrypted file to the data owners. Unauthorized uses can access a file through power of ownership which allows you to DE duplicate on convergent encryption. For example, in an organization much permission will be assigned to the employees.

Symmetric keys shares same key to encrypt and decrypt the message. Convergent encryption acts as owner to the DE duplication which provides confidential data to the user. If two copies are identical then their tags should be same and the tag cannot compromise the confidentiality in reducing the encrypted copy. Only authorized parties should access our data. Metadata leaks data about the file and provides security for unpopular files (F2). File F2 covers with 2 layers where as file(F1) removes the outside layer, In cryptosystem outer layer is semantically secure. Data DE duplication is based on location, disk placement and data unit.DE duplication uses to restore and recovery applications. We are using very huge amount of data leakage where as to project our cost, storage space and network utilization, both the network utilization and storage space have its own cost usage. If we have lost any data there a great solution to find through data DE duplication to decrease its restore time. DE duplication takes place in both file level and block level. Mainly in hybrid cloud approach data duplication is done under block level. Based on the runtime and costs we have many DE duplication techniques.

1. Unique file- Having multiple files that runs on the same operating system with fixed dates.
2. Downloading same file multiple times- Even though the user can download the same file multiple times when the error occurs.

   Example: EMAIL
3. Different files in common blocks- Same file is stored in different user's operating system.

Block Level Deduplication:

Block level is also known as "CHUNK LEVEL" .When the DE duplication is at Block level the file is divided into number of blocks and of variable length or fixed length, it means if we make any modifications to the file in the file itwill only changes and saves the particular segment. Block level needs more access power than the file level DE duplication, many of backup and storage space provides block based DE duplication. Each block is divided into segments and the files are saved only once, the main advantage of using block level is it can hold any type of file. Very large number of usage and memory can be resource intensive based on checksums on block storage. It has mainly three block based trade-offs. They are:

1. Server side DE duplication- all the blocks of data is to be restored from the source and sent to the destination at server side.
2. Post processing- first the server is to be written in blocks of data to double the server side storage space and consumes more storage space in IO operations.

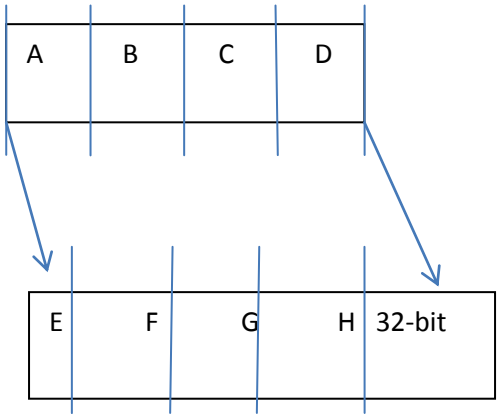
3. Hardware acceleration- Server moves from particular positions to the dedicated, and hardware with high cost storage.

In block level segments are assigned using hash code which generates uniqueness to identify chunks. Block level supports file, divide the segments into streamline and save the difference between every file must be efficient.

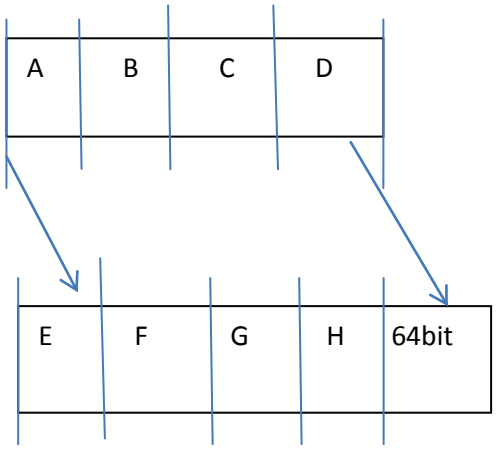
File Level Deduplication:

File level deduplication can be performed easily, requires less amount of accessing power to obtain and generate hash code easily. In file level if any one byte may be changed its hash value changes. Finally the result of hash value is saved in storage. File level stores less amount of space than block level. The stored amount can be saved in disk or archeive. File level deduplication reults in backup on file name, size and type. File level deduplication is easy to perform and requires less amount of processing power files. File level DE duplication checks for multiple copies of a file. Save one instance of it and supply memory location of that file for any resultant concern storage."

Differences Betwwen Block Leve Vs File Level:



Fixed size data sequence blocks
We need to store 8 unique blocks
Using single instance



Variable length data sequence length
we need to store 5 unique blocks
using DE duplication

Client Side Deduplication:

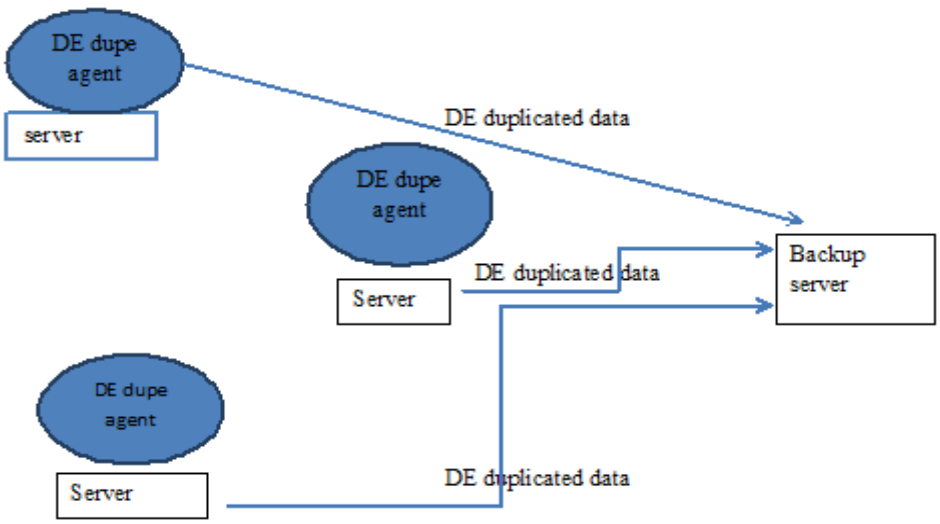


Fig:Client side deduplication.

A client is basically a user, a system or a piece of element or package that accesses a service created accessible by a server The Client-side duplication techniqueis used on backup archive client to remove redundancy data while backup and archive processing before data transfers to Tivoli Storage manager server. DE dupe reduces the amount of data sent over the local area network using Client-side DE duplication. With this method, the exclusion of some files on the client side can be processed. In order to reduce network traffic the cache will be enabled.

The client creates extents which are parts of files and those are compared with other files in order to find duplicates. The client and server work together to find duplicate extents. If client finds the duplicate extents, it will be sent to the server. Some new extents that are found will be sent to the server. And if there is any matching, then that will not be sent to the server.

Through this we can effectively save the bandwidth and storage including drop Box, Wuala and Mozyhome. The data-leakage mechanism is found by harnik et al and Halevy et al where they found that the users files are leaked to outside adversaries and they called this issue as Proofs of Ownership (POW). In this mechanism the owner of the file f proves that the file is used in an effective way and in bounded leakage system the data will be leaked of the file F.

Advantages of client-side DE duplication:

The data that is sent over the local area network will be reduced. Duration of the processing time will be offloaded from server to client node.Inorder to allow duplicate data on the server; the processing power is eliminated to allow space savings to be done immediately in the server.

Disadvantages of Client-side DE duplication:

Until the backup of primary storage pools of non-duplicated copy, the server will not have all of those files. Client extends are reassembled into contiguous files during the storage pool backup.

Server Side Deduplication:

A server is a active instance of an appliance or a software which is able of accepting requests from the applicant and gives the responses appropriately. By connecting to several servers the computer can provide large number of services to user. In this method the data DE duplication is done by server. The Tivoli Storage management administrator can specify the data DE duplication location in order to use register node and update node server commands with dedupe.

In this additional processing is not required that were DE duplicated by the client. The data DE duplication is always enabled for DE duplication enabled storage pools. This reduces the number of disk have to be compelled to store statistical data, however it doesn't change thenumber informationmeasure that's needed so as to induce the backups

from the server. The de-dup IDT/VTL willreplicate the dedupe information to a different location. By that we are able to have on-siteand off-site copy without creating an actual tape.

Advantages of Server side Data DE duplication:

- Few implementations do very fast (100s of MB/s to 1000s of MB/s).
- Does not need modification in backup software.

Disadvantages of Server Side Data DE duplication:

- Considered a "Band-Aid" by some to assist backup software system that was designed to use disk.
- Needs hardware at each remote website to be protected via de-dupe.
- Onsite & offsite copies could also be outside of knowledge of backup software.

Conclusion:

In future cloud storage is essential for every mobile/pc users, as it reduces the data management overhead and minimizes storage costs. Various widely used DE duplication mechanisms were proposed to protect the data security by including differential privileges of users in the duplicate check. We also presented several new DE duplication 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.

References:

1. B.Tirapathi Reddy, M.V.P Chandra sekhar, L.S.S Reddy, V.Krishna Reddy, P.Sai Kiran, “A Survey on Assured File Deletion in Cloud Environment", International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 9, Number 23 (2014) pp. 19899-19907
2. Mark W. Storer Kevin Green an Darrell D. E. Long Ethan L. Miller, “Secure Data Deduplication”, StorageSS’08,October 31, 2008, Fairfax, Virginia, USA
3. Jan Stanek, Alessandro Sorniottiy, Elli Androulakiy, and Lukas Kencl, “ A Secure Data Deduplication Scheme for Cloud Storage”, Czech Technical University in Prague, IBM Research - Zurich, Ruschlikon, Switzerland
4. John Harauz, Lori M. Kaufman and Bruce Potter, “Data Security in the World of Cloud Computing”, IEEE Computer and Reliability Societies, July/August 2009, 1540-7993/09
5. Bibin K Onankunju, “Access Control in Cloud Computing”,International Journal of Scientific and Research Publications, Volume 3, Issue 9, September 2013

6. Peter Brudenall , Bridget Treacy and Purdey Castle, " Outsourcing to the cloud: data security and privacy risks", Hunton and Williams, FW March 2010
7. Yang Tang, Patrick P. C. Lee, John C. S. Lui, Radia Perlman, "Secure Overlay Cloud Storage with Access Control and Assured Deletion"
8. Abhishek Mohta ,Ravi Kant Sahu,Lalit Kumar Awasthi, "Robust Data Security for Cloud while using Third Party Auditor", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 2, February 2012
9. Yang Tang , Patrick P. C. Lee , John C. S. Lui , and Radia Perlman, "FADE: Secure Overlay Cloud Storage with File Assured Deletion", The Chinese University of Hong Kong, Intel Labs
10. Jia Xu, Ee-Chien Chang, Jianying Zhou, "Weak Leakage-Resilient Client-side Deduplication of Encrypted Data in Cloud Storage", ASIA CCS'13, May 8–10, 2013, Hangzhou, China
11. Weichao Wang, Rodney Owens, Zhiwei Li, Bharat Bhargava, "Secure and Efficient Access to Outsourced Data", CCSW'09, November 13, 2009, Chicago, Illinois, USA
12. Zhifeng Xiao and Yang Xiao, "Security and Privacy in Cloud Computing", IEEE Communications Surveys & Tutorials, Vol. 15, No. 2, Second Quarter 2013
13. PradnyeshBhisikar and Prof. Amit Sahu, "Security in Data Storage and Transmission in Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013
14. Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, "Proofs of ownership in remote storage systems", In: CCS '11, New York, NY, USA, ACM (2011) 491–500
15. Y.G.Min, Y.H.Bang, "Cloud Computing Security Issues and Access Control Solutions", Journel of Security Engineering, vol.2, 2012.
16. NesrineKaaniche, Maryline Laurent, "A Secure Client Side Deduplication Scheme in Cloud Storage Environments"

Corresponding Author:

B.Tirapathi Reddy*,

Email: *tirapathireddyb@kluniversity.in*