ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

Effective De - Duplication Method in Secure Cloud Storage

Dr.Sunil Tekale
Professor-CSE
Department of Computer Science
Malla Reddy College Of Engineering

***Abstract:*-- Cloud computing is a promising computing standard which resources of the computing infrastructure are provided as services over the Internet. There is limited source of storage and networking in cloud system. So data de-duplication takes an important role in the cloud structure. In this paper data de-duplication process will be discussed in detail. In data de-duplication there are several methods available that makes it easy to implement. In this paper we will examine about all methods, processes that are used in data de-duplication.**

Keywords: **cloud computing, Scalability, virtual machine, de-duplication.**

I. INTRODUCTION

Clouds are large pools of easily usable and reachable resources. In cloud all resources connected virtually to create single system image. These resources can be dynamically reconfigured to adjust to a flexible load (scale), allowing optimum resource utilization. Cloud storage refers to scalable and elastic storage capabilities that are delivered as a service using Internet technologies with elastic provisioning and use based pricing that does not penalize users for changing their storage consumption without notice. There are five basic characteristic of any cloud system that are:

- On-demand self-service
- Broad network access
- Resource pooling
- Measured service
- Rapid elasticity

Cloud computing contains both hardware and applications provided to users as a service by the Internet. With the fast development of cloud computing, ever more cloud services have emerged, such PaaS (platform as a service), as SaaS (software as a service), and IaaS (infrastructure as a service). Computing resources [9] are limited and eventually any system which grows in data or usage will saturate the resources available to it.

The resources in question may be e.g. processing capacity for computationally intensive systems or storage capacity for data intensive systems. Network Capability is an important scalability point in distributed systems. Structural scalability alarms the interior design of a system and provides methods to it to manipulate the data model. It means it provides methods to shrink data model and expand data model. It can understand like its deployment.

Cloud storage is a paradigm in which the online storage is networked and records are stored on several committed storage servers. Sometimes these storage servers can be maintained by other third parties. The concept of cloud storage is derived from cloud computing. It denotes to a storage device accessed over the Internet via Web service application program interfaces (API). For example: HDFS (Hadoop Distributed File System, hadoop.apache.org) is a distributed file system that runs on commodity hardware; it was introduced by Apache for managing huge data.

Data de-duplication is also known as single instancing or intelligent compression. It essentially points to the removal of replicate data. In the de-duplication process, duplicate data is deleted, only one copy or single instance of the data to be stored in the database. Data de-duplication is a term used to describe an algorithm or technique that eliminates duplicate copies of data from storage. Data de-duplication is commonly performed on secondary storage systems such as archival and backup storage.

II. RELATED DATA

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

A. Data De-Duplication

The term data de-duplication points to the techniques that [1] saves only one single instance of replicated data, and provides links to that instance of copy in place of storing other original copies of this data. By the evolution of services from tape to disk, data de-duplication has turn into a key element in the backup process. It specifies that only one copy of that data is saved in the datacenter [10]. Every user, who want to access that copy linked to that single instance of copy. So it is clear that data de-duplication help to decrease the size of datacenter. So it could say that de-duplication means that the number of the replication of data that were usually duplicated on the cloud should be controlled and managed to shrink the physical storage space requested for such replications.

The basic steps for de-duplication are:

a) In first step files are divided into small segments.
b) After the segment creation new and the existing data are checked for similarity by comparing fingerprints created by SHA-1 algorithm (another methods can also applicable).
c) Then Metadata structures are updated. d) Segments are compressed.
e) All the duplicate data is deleted and data integrity check is performed.

B. Requirements for data De-Duplication

There is only one necessary condition for the data de-duplication is that data de- duplication should be scalable [2]. It means that de-duplication should be elastic. It doesn't affect to the overall storage structure. To handle scalable de- duplication, two methods have been proposed,

1. Sparse indexing: Sparse indexing is a method used to solve the chunk lookup blockage caused by disk access, by using sampling and manipulating the inherent locality inside backup streams. It picks a small slice of the chunks in the stream as samples; then, sparse index maps these samples to the existing sections in which they occur. The arriving streams are fragmented into relatively big segments, and each segment is de-duplicated against only some of the most similar previous sections (segments).

2. Bloom filters with caching. The Bloom filter exploits Summary Vector. Basically summary vector is a compact in-memory data structure, for discovering new segments; and Stream-Informed Segment arrangement, which is a data arrangement method to improve on-disk locality, for consecutively accessed segments; and Locality Well-maintained Caching with cache fragments, which maintains the locality of the impressions of duplicated segments, to achieve high cache hit ratios.

C. Types of data de-duplication

There are two major categories [8] of data de-duplication: Offline Data de-duplication: [7] in an offline de-duplication state, first data is written to the storage disk and de- duplication process takes place at a later time. Online Data de-duplication: In an online de-duplication state, replicate data is deleted before being written to the storage disk. Once the timing of data de-duplication has been decided then there are numbers of existing techniques that can be apply. The most used de-duplication approaches are: whole file hashing (WFH), sub file hashing (SFH), and delta encoding (DE). Whole File Hashing: In a whole file hashing (WFH) technique, the whole file is directed to a hashing function. The hashing function is always cryptographic hash like MD5 or SHA-1. The cryptographic hash is used to find entire replicate files. This approach is speedy with low computation and low additional metadata overhead. It works very well for complete system backups when total duplicate files are more common. However, the larger granularity of replicate matching stops it from matching two files that only differ by one single byte or bit of data. Sub File Hashing: [2] Sub file hashing (SFH) is Appropriately named. Whenever SFH is being used, it means the file is broken into a number of smaller sections before data de-duplication. The number of sections depends on the type of SFH that is being used. The two most common types of SFH are fixed size chunking and variable-length chunking. In a fixed-size chunking approach, a file is divided up into a number of fixed-size pieces called "chunks". In a variable-length chunking approach, a file is broken up into "chunks" of variable length. Some techniques such as Rabin fingerprinting [28] are applied to determine "chunk boundaries". Each section is passed to a cryptographic hash function (usually MD5 or SHA-1) to get the "chunk identifier". The chunk identifier is used to locate replicate data. Both of these SFH approaches find replicate data at a finer granularity but at a price.

Delta Encoding: The term delta encoding (DE) is Comes from the mathematical use of the delta

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

symbol. In math and science, delta is used to calculate the "change" or "rate of change" in an object. Delta encoding is applied to show the difference between a source object and a target object. Suppose, if block A is the source and block B is the target, the DE of B is the difference between A and B that is unique to B. The expression and storage of the difference depends on how delta encoding is applied. Normally it is used when SFH does not produce results but there is a strong enough similarity between two items/ locks / chunks that storing the difference would take less space than storing the no duplicate block.

III. CATEGORY OF DATA DE- DUPLICATION STRATEGIES

Data de-duplication strategies can be categorized according to their operational area. In this respect there are two main data de-duplication strategies:

File-level De-duplication: File level de-duplication is performed over a single file. In this type of de-duplication two or more files are identified as similar if they have the same hash value.

Block-level de-duplication: Block level de-duplication is performed over blocks. It first divides files into blocks and stores only a single copy of each block. It could either use fixed-sized blocks or variable-sized chunks. It can be further divided on the basis of their targeted area. Target based de- duplication: This type of de-duplication performed on the target data storage center. In this case the client is unmodified and not aware of any de-duplication. This technology improves storage utilization, but does not save bandwidth Source based de- duplication: This type of de-duplication performed on the data at the source before it's transferred. A de-duplication aware backup agent is installed on the client who backs up only unique data. The result is increased bandwidth and storage efficiency. But, this enforces extra computational load on the backup client. Replicates are changed by pointers and the actual replicate data is never sent over the network.

IV. THE PROPOSED SYSTEM

In the proposed system we are achieving the data de-duplication by providing the proof of data by the data owner. This proof is used at the time of uploading of the file. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files.

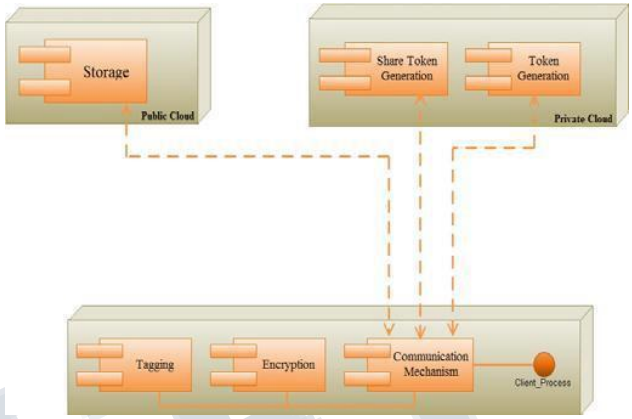

Fig1. Proposed System Framework Communication

Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. Fig1. Shows the Proposed System Framework Communication between clients, Private Cloud, Public Cloud.

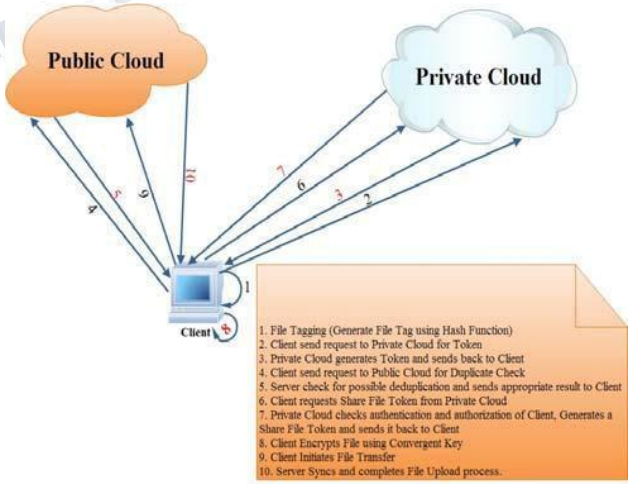

Fig2.Proposed System of Operations

In this work Public Cloud is used for storage data ,Private Cloud is used for performance the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 12, December 2016

operations like share token generation ,token generation, also Client is performed operations like tagging of file ,encryption of file, communication between private cloud and public cloud.

Encryption of Files

Here we are using the common secret key k to encrypt as well as decrypt data. This will use to convert the plain text to cipher text and again cipher text to plain text. Here we have used three basic functions:

KeyGenSE: k is the key generation algorithm that generates κ using security parameter 1.
EncSE (k, M): C is the symmetric encryption algorithm that takes the secret κ and message M and then outputs the ciphertext C;

DecSE (k, C): M is the symmetric decryption algorithm that takes the secret κ and ciphertext C and then outputs the original message M.

V. CONFIDENTIAL ENCRYPTION

It provides data confidentiality in de- duplication. A user derives a convergent key from each original data copy and encrypts the data copy with the convergent key. In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. The user has to prove that the data which he wants to upload or download is its own data. That means he has to provide the convergent key and verifying data to prove his ownership at server.

VI. CONCLUSION

Cloud computing has reached a maturity that leads it into a productive phase. This means that most of the main issues with cloud computing have been addressed to a degree that clouds have become interesting for full commercial exploitation. This however does not mean that all the problems listed above have actually been solved, only that the according risks can be tolerated to a certain degree. Cloud computing is therefore still as much a research topic, as it is a market offering. For better confidentiality and security in cloud computing we have proposed 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. Proposed system includes proof of data owner so it will help to implement better security issues in cloud computing.

REFERENCES

[1] P. Anderson and L. Zhang. "Fast and secure laptop backups with encrypted de- uplication". In Proc. of USENIX LISA, 2010.
[2] M. Bellare, S. Keelveedhi, and T. Ristenpart. "Dupless: Server aided encryption for deduplicated storage". In USENIX Security Symposium, 2013.

[3] Pasqualo Puzio, Refik Molva ,MelekOnen ,"CloudDedup: Secure De-duplication with Encrypted Data for Cloud Storage", SecludIT and EURECOM,France.

[4] Iuon –Chang Lin, Po-ching Chien ,"Data De-duplication Scheme for Cloud Storage" International Journal of Computer and Control(IJ3C),Vol1,No.2(2012)

[5] Shai Halevi, Danny Harnik, Benny Pinkas,"Proof of Ownership in Remote Storage System", IBM T.J.Watson Research Center, IBM Haifa Research Lab, Bar IIan University,2011.

[6] M. Shyamala Devi, V.Vimal Khanna,Naveen Balaji "Enhanced Dynamic Whole File De- Duplication(DWFD) for Space Optimization in Private Cloud Storage Backup",IACSIT, August,2014.

[7] Weak Leakage-Resilient Client –Side de-duplication of Encrypted Data in Cloud Storage" Institute for Info Comm Research,Singapore,2013

[8] Tanupriya Chaudhari , Himanshu shrivastav, Vasudha Vashisht, "A Secure Decentralized Cloud Computing Environment over Peer to Peer",IJCSMC,April,2013

[9] Mihir Bellare, Sriram keelveedhi,Thomas Ristenart ,"DupLESS: Server Aided Encryption for Deduplicated storage" University of California, San Diego2013

[10] Dave Russell: Data De-duplication Will Be Even Bigger in 2010, Gartner, 8 February 2010.