A Survey Paper on Removal of Data Duplication in a Hybrid Cloud

Prakash Gapat [1] ,Snehal Khillare [2], Akshay Khiste [3], Mrs. Rohini Pise [4]

[123]BE IT, Department of Information Technology, PCCOE, Maharashtra, India
[4] Assistant Professor, Department of Information Technology, PCCOE, Maharashtra, India

---***---

Abstract—*Data deduplication is an effective compression tech-nique to eliminate duplicate copies of repeting data. It has been widely used in cloud storage to reduce the amount of storage valume and save bandwidth. For the protection of sensitive data while supporting deduplication, the convergent encryption technique is being used to encrypt the data before stored it to cloud storage,In order to make system more secure, the different privileges to users are again considered while checking duplicate content.But the problem which occurs in this approch is that even if some contents of both files are different ,it stores them as two different files which leads to reduction in cloud storage space.The solution for this problem is to perform apply technique for deduplication(block level).In this approch the file which is to be stored on cloud storage is divided into number of different blocks based on contents and deduplication is performed on these blocks.*

Keywords - Deduplication, authorized duplicate check, confidentiality, cloud

Introduction

Cloud computing technique which is most widely used everywhere .In that, computing is done over the large communication network like Internet. This is an important solution for business storage in low cost. Cloud computing provide vast storage in all sector like government, enterprise,also for storing our per-sonal data on cloud.Without background implementation de-tails,platform user can access and share different resources on cloud.The most important problem in cloud computing is that large amount of storage space and security issues.One critical challenge of cloud storage to management of ever-increasing volume of data.To improve scalability,storage problem data deduplication[8] is most important technique and has attracted more attension recently.It is an important technique for data compression.It simply avoid the duplicate copies of data and store single copy of data.Data deduplication take place in either block level or file level.In the file level approach duplicate files are eliminate,and in block level approach duplicate blocks of data that occur in non-identical files.Deduplication reduce the storage needs by upto 90-93 % for backup application,68% in standard file system.Important issues in data deduplication that security and privacy to protect the data from insider or outsider attack.For data confidentiality,encryption is used by different user for encrypt there files or data ,using a secrete key user perform encryption and decryption operation. For upload file to cloud user first generate convergent key,encryption of file then load file to the cloud.To prevent unauthorize access proof of ownership protocol is used to provide proof that the

user indeed owns the same file when deduplication found.After the proof,server provide a pointer to subsequent user for accessing same file without needing to upload same file.When user want to download file he simply download encrypted file from cloud and decrypt this file using convergent key. To make information administration versatile in distributed computing, deduplication has been an understood system and has pulled in more consideration as of late. Information deduplication is a particular information pressure procedure for disposing of copy duplicates of rehashing information away.The procedure is utilized to enhance stockpiling usage and can likewise be connected to network information exchanges to diminish the quantity of bytes that must be sent.Rather than keeping various information duplicates with the same substance, deduplication wipes out excess information by keeping one and only physical duplicate and alluding other repetitive information to that duplicate. Deduplication can occur at either the document level or the square level.For file level deduplication,it wipes out copy duplicates of the same document. Deduplication can likewise happen at the piece level, which dispenses with copy squares of information that happen in non-indistinguishable records.

II. CLOUD COMPUTING SYSTEM

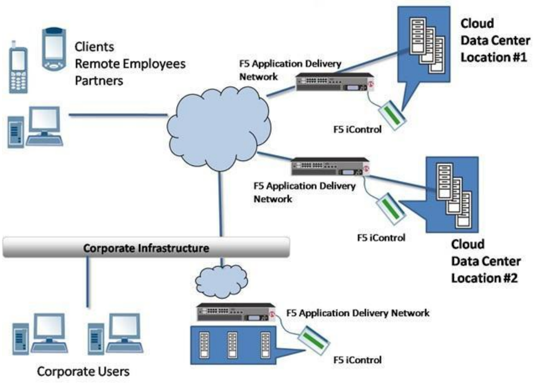

Clients
Remote Employees
Partners
F5 Application Delivery Network
Cloud Data Center Location #1
F5 iControl
F5 Application Delivery Network
F5 iControl
Cloud Data Center Location #2
Corporate Infrastructure
F5 Application Delivery Network
F5 iControl
Corporate Users

Fig. 1. Cloud Computing System

A. Types of cloud storage

1) Public Cloud: Public clouds are made available to the public by a service provider who provide the cloud infrastructure. Most common public cloud providers are namely Google , Amazon and Microsoft own and operate their infrastructure and offer access all over the Internet. In this model, customer has no control or visibility over where the infrastructure is located. It is most important to consider that all the customers on public cloud storage share the same infrastructure with availability variances ,limited configuration and security protections .Public Cloud customers generally benefit from economies of scale, because of the infrastructure costs are spread across all users which allows each and every client to operate on a low-cost, pay-as-you-go model. The main advantage of public cloud infrastruc-tures is that public cloud storages are typically larger in scale than an in-house enterprise cloud, which in turn provides clients with seamless, on demand scalability. These clouds offer one of the greatest level of efficiency in shared resources; however, at the same time these stor-ages are also more insecure than private cloud storages.

2) Private Claud: Private clouds are those that are specially designed for a business. Many companies consider private clouds as a good starting point. They allow the organization to host applications, infrastructure and development environments in a cloud.Also addressing concerns regarding data security and control that can arise in the public cloud environment.In general there are two types of private clouds: First and most common private cloud is a Private Cloud: This cloud also known as an Internal Cloud.It is hosted within an organizations data center. The benefits of a this private cloud scalable,virtualized, flexible.Private cloud infrastructure are non accepted. The another type of private cloud is an externally Hosted Virtual Private Cloud:It is being hosted by a third party Cloud Service Providers. This provider an exclusive private cloud environment .This takes responsibility for implementing, managing, and securing the Cloud infrastructure.

3) Hybrid Cloud :A hybrid cloud is best-of-breed. It com-bines the comfort level of a private cloud with the versatility and flexibility of the public cloud. Hybrid platforms use either public clouds or Hosted Virtual Private Clouds for some processesand applications. Both of these merge with on premises private clouds for high-security application environments . As with the private model, in a hybrid cloud, an organization may choose to continue to use their data center equipment and sensitive data secured on their own network. And like the public cloud, a hybrid model let an organization take advantage of a clouds accessibility , backup, scalability and disaster recovery. There are some of the limitations of the public cloud while still gaining many of the public clouds benefits.

B. Cloud services

1) Software as a Service(SaaS): The SaaS is not suitable for applications that require real-time response or those for which data is not allowed to be hosted externally. The most likely candidates for SaaS are applications for which

2) Platform as a Service(PaaS): Platform-as-a-Service (PaaS) gives the capability to deploy consumer-created or acquired applications using programming languages and tools supported by the provider. The user does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, or storage. The user has control over the deployed applications and, possibly, over the application host-ing environment configurations. Such services include session management, device integration, sandboxes, instrumentation and testing, contents management, knowledge management, and Universal Description, Discovery, and Integration (UDDI), a platform-independent Extensible Markup Language (XML)-based registry providing a mechanism to register and locate Web service applications. PaaS is not particulary useful when the application must be portable, when proprietary program-ming languages are used, or when the underlaying hardware and software must be customized to improve the performance of the application. The major PaaS application areas are in software development where multiple developers and users collaborate and the deployment and testing services should be automated.

3) Infrastructure as a Service(IaaS): Infrastructure-as-a-Service (IaaS) is the capability to provision processing, stor-age, networks, and other fundamental computing resources; the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, stor-age, deployed applications, and possibly limited control of some networking components, such as host firewalls. Services offered by this delivery model include: server hosting, Web servers, storage, computing hardware, operating systems, vir-tual instances, load balancing, Internet access, and bandwidth provisioning. The IaaS cloud computing delivery model has a number of characteristics, such as the fact that the resources are distributed and support dynamic scaling, it is based on a utility pricing model and variable cost, and the hardware is shared among multiple users. This cloud computing model is particulary useful when the demand is volatile and a new business needs computing resources and does not want to invest in a computing infrastructure or when an organization is expanding rapidly.

4) RELATED WORK

Data de-duplication

Data de-duplication is divided into many types. There is no best way to implement data de-duplication across an organization. For maximize the benefits, organizations may introduce more than one de-duplication strategy every decade. It is very simple to understand the backup and backup challenges, when selecting de-duplication as a solution. Data de-duplication has mainly divided into three forms. According to definitions, some forms of data de-duplication, such as compression, have been around for decades. Lately, single-instance storage has introduce the removal of redundant files from storage environments and now, we have seen the introduction of sub-file de-duplication. Three types of data de-duplication are:

1) Data Compression The main use of Data compression is reducing the size of files. Data compression is works within a file to identify and remove empty space that present in file as repetitive forms. This form of data de-duplication is only local to the file.Data Compression-does not take into consideration of other files and data segments within those files. Data compression technique is occur many years ago,but being isolated to each single file, the benefits are limited as comparing to other forms of de-duplication. For example, data compression will not be effective in eliminating and recognizing duplicate files, but will independently compress each of the files.
2) Single-Instance Storage The main use of Single-Instance Storage is removing multiple copies of any file.Single-instance storage (SIS) environments are de-

   tect and remove redundant copies of identical files. After a file is stored in a single-instance storage system then, all those other references to same file, will refer to the original file,only single copy. Every time Single-instance storage systems compare the content of files to determine the incoming file is identical to an existing file in the storage system. Content-addressed storage is typically completed with single-instance storage func-tionality.In file-level de-duplication,it avoids storing files that are a dublicate of another file, many files that are considered unique or identical by single-instance storage measurement may have a more amount of redundancy within the files or between files. For example,take one small element (e.g., a new name inserted into the title slide of a file) for single-instance storage for two large files as being different and requiring them to be stored without further de-duplication.

3) Sub-file De-Duplication Sub-file de-duplication detects same or redundant data within and across files as opposed to finding identical files as in Single-Instance Storage implementations.With the help of sub-file de-duplication, redundant copies of data are detected and are eliminated even after the duplicated data exist, within different files. This form of de-duplication finds the unique data elements within an organization and detects when these elements are used oe find within other files. As a result, sub-file de-duplication removes the storage of duplicatedata in organization. Sub-file data de-duplication has many benefits even where files are not identical, but have data elements that are already recognized somewhere in the organization. Sub-file de-duplication implementation is mainly divided into two forms.First form is Fixed-length sub-file de-duplication uses an arbitrary fixed length of data to search for the duplicate data within the files.If file is simple in design, fixed-length data segments miss many opportunities to produce redundant sub-file data. (Consider the case where an addition of a date or name or title is added to a document’s title pagethe whole content of the document will shift, causes the failure of the de-duplication tool to detect same data). Seecond form isVariable-length implementations are usually not locked to any of

arbitrary segment length. Variable-length implementations match data segment sizes to the naturally occurring duplication within files, vastly increasing the overall de-duplication ratio as compare to fixed length (In the example above, variable-length de-duplication will catch all duplicate segments in the document, no matter where the changes occur).

B. Data De-duplication types:

1) File-level de-duplicationIt is commonly known as single-instance storage, file-level data de-duplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index. The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by ”stub” which points to the original file.

2) Block-level de-duplicationBlock-level data de-duplication operates on the basis of sub-file level. As the name implies, that the file is being broken into segments blocks or chunks that will be examined for previously stored information vs redundancy. The popular approach to determine redundant data is by assigning identifier to chunk of data, by using hash algorithm for example it generates a unique ID to that particular block. The particular unique Id will be compared with the central index. In case the ID is already present, then it represents that before only the data is processed and stored before .Due to this only a pointer reference is saved in the location of previously stored data. If the ID is new and does not exist, then that block is unique. After storing the unique chunk the unique ID is updated into the Index. There is change in size of chunk as per the vendor. Some will have fixed block sizes, while some others use variable block sizes

3) Variable block level de-duplication It Compares vary-ing sizes of data blocks that can reduce the chances of collision, stated Datalinkss Orlandini.[7]

Proposed Method

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.

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

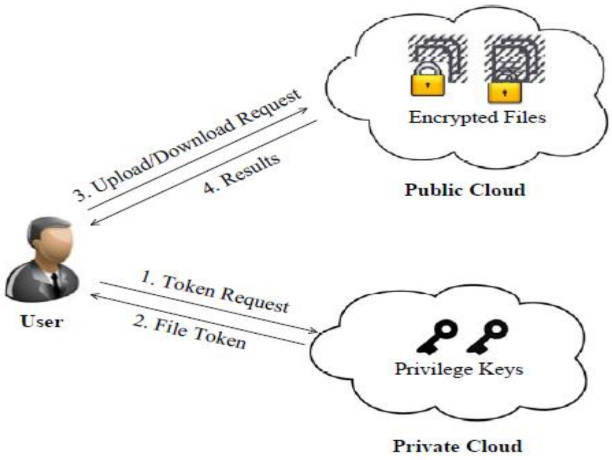
2) We present an advanced scheme to support stronger security by encrypting the file with differential privilege keys

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

1)* Traditional Encryption algorithm *: Although it is known that data deduplication gives more benefits, se-curity and privacy concerns arise because the users sensitive data is susceptible to both the outsider as well as insider attacks. So, while considering the traditional encryption techniques to secure the users sensitive data there are many issues are associated. Traditional encryp-tion provides data confidentiality but it is not compatible with deduplication. As in traditional encryption different users encrypt their data with their own keys. Thus, the identical data of the different users will lead to different ciphertext which is making the data deduplication almost impossible in this traditional approch.

2)Convergent Encryption algorithm: The convergent encryption[1] techniques are those which provide the data confidentiality to the users outsourced data stored on the public clouds. These techniques while providing the confidentiality to the data are also compatible with the data deduplication process. In this algorithm the encryption key is itself derived from the message. So it supports data deduplication also, because the same file will give the same encryption key so it will generate the same ciphertext irrespective of users which makes data deduplication possible.

SYSTEM ARCHITECTURE



Conclusion

In this paper, the idea of authorized data deduplication was proposed to protect the data security by including differential authority of users in the duplicate check. In public cloud our data are securely store in encrypted format,and also in private cloud our key is store with respective file.There is no need to user remember the key.So without key anyone can not access our file or data from public cloud.

References :

[1]Jin Li, Yan Kit, 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]P. Anderson and L. Zhang. Fast and secure laptop back-ups with encrypted de-duplication. In Proc. of USENIX LISA, 2010.

[3]M. Bellare, Keelveedhi, and T. Ristenpart. Dup-less: Serveraided encryption for deduplicated storage. In USENIX Security Symposium, 2013.

[4]Li, X. Chen, M. Li, J. Li, P. Lee, andW. Lou. Secure deduplication with efficient and reliable convergent key management. In IEEE Transactions on Parallel and Dis-tributed Systems, 2013.

[5]Neal Leavitt, ”Hybrid Clouds Move to the Forefront. Published by the IEEE Computer Society, MAY 2013.