ISSN:2348-2079

Volume-7 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

A hybrid cloud approach for preserve authorized deduplication

Mr.P.Ganesh[1], Mrs. C. Navamani MCA., M.Phil., ME.,[2]

[1]Final M.C.A, Department of M.C.A, Nandha Engineering College (Autonomous), Erode-52

[2]Assistant Professor, Department of M.C.A, Nandha Engineering College (Autonomous), Erode-52

ABSTRACT

Information deduplication is one of vital information pressure procedures for killing copy duplicates of rehashing information, and has been broadly utilized in distributed storage to decrease the measure of storage room and spare data transmission. To ensure the classification of touchy information while supporting deduplication, the focalized encryption procedure has been proposed to scramble the information before re-appropriating. To all the more likely ensure information security, this paper makes the main endeavor to formally address the issue of approved information deduplication. Unique in relation to conventional deduplication frameworks, the differential benefits of clients are additionally considered in copy check other than the information itself. We likewise present a few new deduplication developments supporting approved copy check in half and half cloud design. Security examination shows that our plan is secure as far as the definitions determined in the proposed security demonstrate. As a proof of idea, we execute a model of our proposed approved copy check plan and direct proving ground tests utilizing our model. We demonstrate that our proposed approved copy check conspire brings about insignificant overhead contrasted with ordinary activities.

Keywords: Deduplication, Authorized duplicate check, Confidentiality, Hybrid cloud.

INTRODUCTION

Domain description

Distributed computing is a web innovation that uses both focal remote servers and web to deal with the information and applications. This innovation enables numerous organizations and clients to utilize the information and application without an establishment. Clients and organizations can get to the data and records at any PC framework having a web association. Distributed computing gives considerably more successful registering by unified memory, preparing, stockpiling and data transmission. Distributed computing is a term to portray an innovation that disseminates PC benefits from a nearby customer.

"just in case" network power, to "pay for usage" and thereby maximizing the ability of their users while avoiding idle network processing. Cloud computing is an internet technology that utilizes both central remote servers and internet to manage the data and applications. This technology allows many businesses and users to use the data and application without an installation.

Users and businesses can access the information and files at any computer system having an internet connection. Cloud computing provides much more effective computing by centralized memory, processing, storage and bandwidth. The process of cloud computing technology is broken down into three segments. Platform, Applications and Infrastructure are three segments of this technology. Each part serves many functions and provides applications for both individuals and businesses all around the world

Segments

- Applications- On Demand

Author for correspondence:
Department of M.C.A, Nandha Engineering College (Autonomous), Erode-52

- Infrastructure
- Platform

APPLICATIONS- ON DEMAND

Application segment is the only part of internet technology that has been proved as a useful business model. By accessing many business and individual applications all over the cloud from the central server, most of the businesses can cut some very serious costs. On the other side, on demand applications come in dissimilar varieties of pricing schemes and how the internet applications are delivered to the customers.

Infrastructure

Infrastructure is the backbone of the whole concept of this technology. All infrastructure vendors allow users to make their own cloud applications. Amazon's S3 is considered as a segment of the infrastructure segment.

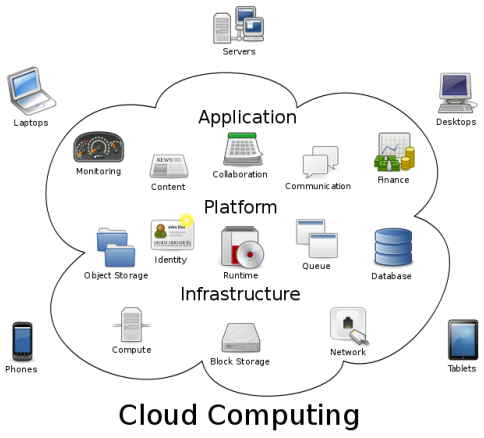
Platform

Most of the companies, who are providing several on demand services and applications, have been developed three platform services. These are Infrastructure As a Service, Platform As a Service and Software As a Service.

There are many types of public cloud computing:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Storage as a Service (STaaS)
- Security as a Service (SECaaS)
- Data as a Service (DaaS)
- Test environment as a Service (TEaaS)
- Desktop as a Service (DaaS)
- API as a Service (APIaaS)

Cloud computing provides much more effective computing by centralized memory, processing, storage and bandwidth. Cloud computing is a term to describe a technology that distributes computer services away from a local client. Cloud computing is an internet technology that utilizes both central remote servers and internet to manage the data and applications. This technology allows many businesses and users to use the data and application without an installation. Users and businesses can access the information and files at any computer system having an internet connection Cloud computing is an internet technology that utilizes both central remote servers and internet to manage the data and applications.



In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage. 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.

A Hybrid Cloud is a combined form of private clouds and public clouds in which some critical data resides in the enterprise's private cloud while other data is stored in and accessible from a public cloud. Hybrid clouds seek to deliver the advantages of scalability, reliability, rapid deployment and potential cost savings of public clouds with the security and increased control and management of private clouds. As cloud computing becomes famous, an increasing amount of data is being stored in the cloud and used by users with specified privileges, which define the access rights of the stored data.

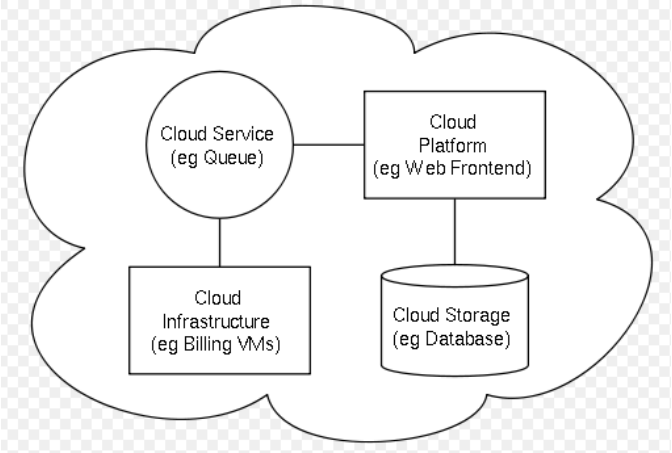


Figure: 1. Architecture of cloud computing.

The critical challenge of cloud storage or cloud computing is the management of the continuously increasing volume of data. Data deduplication or Single Instancing essentially refers to the elimination of redundant data. In the deduplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. However, indexing of all data is still retained should that data ever be required. In general the data deduplication eliminates the duplicate copies of repeating data.

The data is encrypted before outsourcing it on the cloud or network. This encryption requires more time and space requirements to encode data. In case of large data storage the encryption becomes even more complex and critical. By using the data deduplication inside a hybrid cloud, the encryption will become simpler.

As we all know that the network is consist of abundant amount of data, which is being shared by users and nodes in the network. Many large scale network uses the data cloud to store and share their data on the network. The node or user, which is present in the network have full rights to upload or download data over the network. But many times different user uploads the same data on the network. Which will create a duplication inside the cloud. If the user wants to retrieve the data or download the data from cloud, every time he has to use the two encrypted files of same data. The cloud will do same operation on the two copies of data files. Due to this the data confidentiality and the security of the cloud get violated.

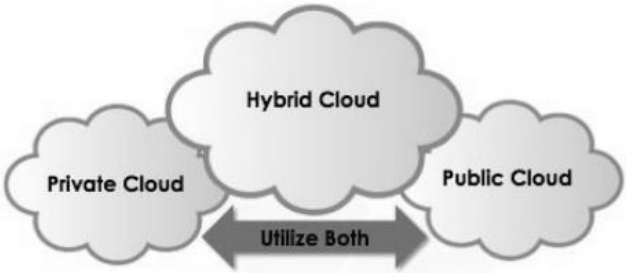


Figure: 2. Architecture of Hybrid cloud.

To avoid this duplication of data and to maintain the confidentiality in the cloud we using the concept of Hybrid cloud. It is a combination of public and private cloud. Hybrid cloud storage combines the advantages of scalability, reliability, rapid deployment and potential cost savings of public cloud storage with the security and full control of private cloud storage.

LITERATURE SURVEY

In previous deduplication systems cannot support differential authorization duplicate check, which is important in many applications. In such an authorized deduplication system, each user is issued a set of privileges during system initialization. The overview of the cloud deduplication is as follow:

Post-process deduplication

With post-process deduplication, new information is first put away on the capacity gadget and after that a procedure sometime in the not too distant future will break down the information searching for duplication. The advantage is that there is no compelling reason to sit tight for the hash figurings and query to be finished before putting away the information along these lines guaranteeing that store execution isn't corrupted. Executions offering arrangement based task can enable clients to concede streamlining on "dynamic" documents, or to process records dependent on sort and area. One potential downside is that you may pointlessly store copy information for a brief span which is an issue if the capacity framework is close full limit.

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.

Source versus target deduplication

Another approach to consider information deduplication is by where it happens. At the point when the deduplication happens near where information is made, it is frequently alluded to as "source deduplication." When it happens close where the information is put away, it is generally called "target deduplication ." Source deduplication guarantees that information on the information source is deduplicated. This by and large happens straightforwardly inside a record framework. The record framework will occasionally examine new documents making hashes and contrast them with hashes of existing documents.

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. 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.

METHODOLOGY

The basic objective of this work is the problem of privacy preserving deduplication in cloud computing and a proposed System focus on these aspects:

- Differential Authorization**:** Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges.
- Authorized Duplicate Check: Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate.

Proposed system

In the proposed framework we are accomplishing the information deduplication by giving the confirmation of information by the information proprietor. This confirmation is utilized at the season of transferring of the document. Each document transferred to the cloud is additionally limited by a lot of benefits to indicate which sort of clients is permitted to play out the copy check and access the records [4]. Before presenting his copy check ask for some record, the client needs to take this document and his very own benefits as information sources. The client can locate a copy for this document if and just if there is a duplicate of this record and a coordinated benefit put away in 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.

Confidential encryption

It provides data confidentiality in deduplication. 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.

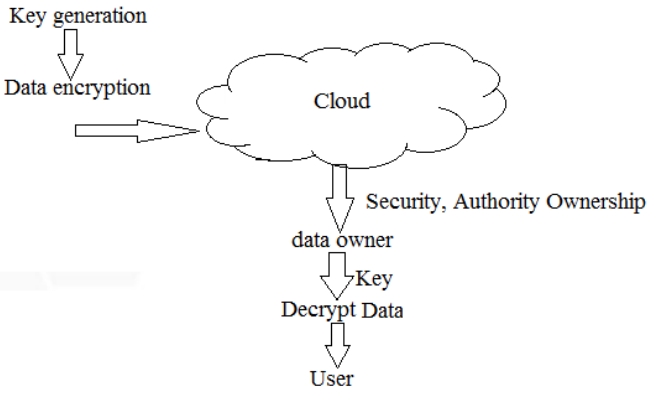


Figure: 3 confidential data encryption

Proof of Data

The user have to prove that the data which he want to upload or download is its own data. That means he have to provide the convergent key and verifying data to prove his ownership at server.

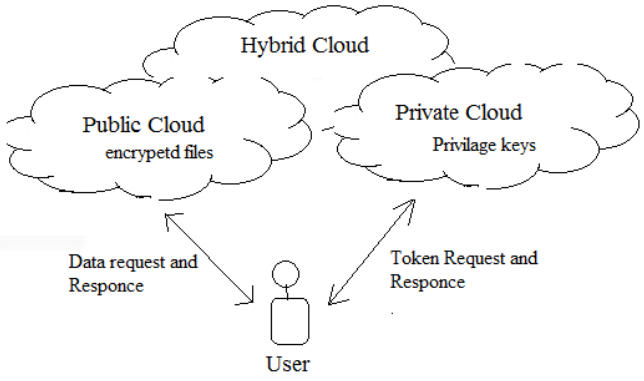


Figure: 4 System Architecture.

SHA1 Algorithm Description

In the proposed system convergent key for each file is generated by using secure hashing algorithm-1 the steps of this algorithm is given below Step1: Padding

- Pad the message with a single one followed by zeroes until the final block has 448 bits.
- Append the size of the original message as an unsigned 64 bit integer.

Step2: Initialize the 5 hash blocks (h0,h1,h2,h3,h4) to the specific constants defined in the SHA1 standard. Step3:Hash (for each 512bit Block) ○ Allocate an 80 word array for the message schedule . Set the first 16 words to be the 512bit block split into 16 words. The rest of the words are generated using the following algorithm step4: word[i3] XOR word[i8] XOR word[i14] XOR word[i16] then rotated 1 bit to the left. Loop 80 times doing the following. Calculate SHAfunction() and the constant K (these are based on the current round number[5]. e=d, d=c c=b (rotated left 30) b=a ,a = a (rotated left 5) + SHAfunction() + e + k + word[i] ○ Add a,b,c,d and e to the hash output. step5: Output the concatenation (h0,h1,h2,h3,h4) which is the message digest.

FLOW CHART

The private keys for the privileges are managed by the private cloud, who answers the file token requests from the users and this interface offered by the private cloud allows user to submit files and queries to be securely stored and computed respectively.

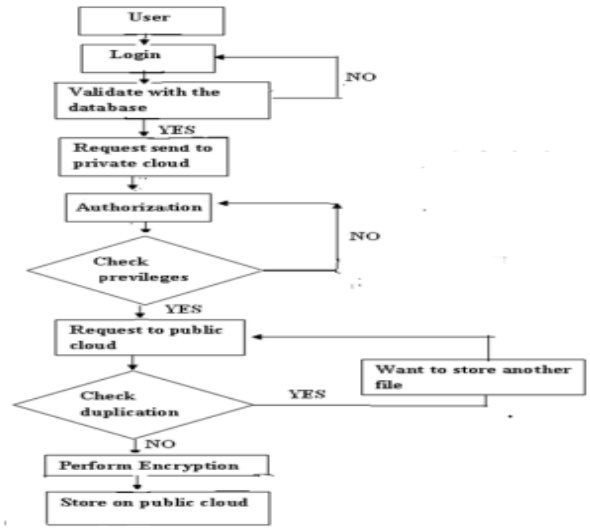


Figure 2: Flow Diagram of the proposed work.

In deduplication system, hybrid cloud architecture is introduced to solve the problem of unauthorized deduplication of file. The private keys for privileges will not be issued to users directly, which will be kept and managed by the private cloud server. The user needs to send a request to the private cloud server to get a file token. The user needs to get the file token from the private cloud server to perform the duplicate check for some file. The user either uploads this file or prove their ownership based on the results of duplicate check . If it is passed, the private cloud server will find the corresponding privileges of the user from its stored table list and send to the user then user can upload his files. The same way user can download his file from storage cloud.

CONCLUSION

There is scope for future development of this project. The world of computer fields is not static; it is always subject to be dynamic. The technology which is famous today becomes outdated the very next day. To keep abstract of technical improvements, the system may be further refined. So, it is not concluded. Yet it will improve with further enhancements.

Enhancements can be done in an efficient manner. We can even update the same with further modification establishment and can be integrated with minimal modification. Thus the project is flexible and can be enhanced at anytime with more advanced features.

REFERENCES

[1]. OpenSSL Project, 1998. [Online]. Available: http://www. openssl.org/
[2]. P. Anderson and L. Zhang, “Fast and secure laptop backups with encrypted de-duplication”, in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, 29–40.
[3]. M. Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Serveraided encryption for deduplicated storage”, in Proc. 22nd USENIX Conf. Sec. Symp., 2013, 179–194.
[4]. M. Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication”, in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, 296–312.

[5]. M. Bellare, C. Namprempre, and G. Neven, "Security proofs for identity-based identification and signature schemes", J. Cryptol., 2(2), 2009, 11–61.
[6]. M. Bellare and A. Palacio, "Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks", in Proc. Annu. Int. Cryptol. Conf. Adv. Cryptol., 2002, 162–177.
[7]. S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, "Twin clouds: An architecture for secure cloud computing", in Proc. Workshop Cryptography Security Clouds, 2011, 32–44.
[8]. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system", in Proc. Int. Conf. Distrib. Comput. Syst, 2002, 617–624.
[9]. D. Ferraiolo and R. Kuhn, "Role-based access controls", in Proc. NIST-NCSC Nat. Comput. Security Conf., 1992, 554–563.
[10]. GNU Libmicrohttpd, 2012. Available: http://www. gnu.org/software/libmicrohttpd
[11]. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems", in Proc. ACM Conf. Comput. Commun. Security, 2011, 491–500.
[12]. J. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management", in Proc. IEEE Trans. Parallel Distrib. Syst., http:// doi.ieeecomputersociety.org/10.1109/TPDS.2013.284.
[13]. libcurl, 1997. [Online]. Available: http://curl.haxx.se/libcurl/
[14]. C. Ng and P. Lee, "Revdedup: A reverse deduplication storage system optimized for reads to latest backups", in Proc. 4th Asia- Pacific Workshop Syst., http://doi.acm.org/10.1145/2500727. 2500731 2013.