Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI:10.21474/IJAR01/4604
DOI URL: http://dx.doi.org/10.21474/IJAR01/4604
ISSN NO. 2320-5407

RESEARCH ARTICLE

A SURVEY ON SECURE DEDUPLICATION SCHEMES.

Anita A.Kundgir[1], Kalyani P. Daberao[1] and S. S. Hatkar[3].
1. Dept. of CSE, (CNIS), SGGS IE & T, Nanded, India.
2. Dept of CSE(CNIS), SGGS IE & T, Nanded, India.

Manuscript Info

Manuscript History

Received: 24 April 2017
Final Accepted: 26 May 2017
Published: June 2017

Key words:-
Cloud storage; Deduplicatinn; Integrity;

Abstract

In today's world as we all know the cloud computing performs a number of operations on data. The valuable information from the set of data is extracted as per their need. Nowadays it becomes trend to be a part of digitalization. Digital libraries contains huge amounts of data. Most of the times while data storage the number of copies of the same data are stored, again and again, so to remove such duplicate data copies we need a Deduplication technique. The deduplication removes unnecessary data, but at the same time, it is beneficial if it retrieves it with more reliability.Thispaper, represents the detail study of various approaches to improve the reliability of data after removing duplicate copies in data storage. The use of such deduplication schemes also reduces the data transfer rate to upload and download data .The time consumption should be less which will effect on its accessibility.

Introduction:-

As number of huge applications are performing the online transaction and various operations, the necessary amount of data is stored online. The online data size is now increased tremendously. Most of time the same data saved in different places due to which the data volume is unnecessarily increased. Nowadays we can see there are a number of companies which offering cloud storage service. Like yahoo, google Microsoft, its companies offering cloud service for data users. As the data storage is becoming famous in online commercials the service providers, improving the tools and application to make simple and secure data service.

As the multiple data copies of the same content are stored in the cloud to remove such repeated data the Deduplication scheme is introduced. And it has received very much attention in the educational field as well as the business field. The data, researchers in the area are trying to make it as improved as better than earlier one.

Several ways to perform Deduplication are based on their content,locationtype of data,static and dynamic runtime. Now if we think about location.it categories whether the data is deduplicated either client side or server side. In source-based deduplication, the customer initially hashes every information portion he wishes to transfer and sends these outcomes to the capacity supplier to check whether such information are as of now put away: in this way just "undeduplicated" information sections will be really transferred by the client.While deduplication at the client side can accomplish transfer speed, it sadly can make the framework powerless against side-channel attacks whereby intruder can quickly find whether a specific information is put away or not.[6].

Corresponding Author:-Anita A.Kundgir.
Address:-Dept. of CSE, (CNIS), SGGS IE & T, Nanded, India.

Types Of Data Deduplication:-
Data deduplication type is classified on the basis of deduplication timing. As the deduplication has to perform data deduplication, it creates a record for new files or data items with respect to the time[13].

Offline Deduplication:-
In this case the data deduplication algorithm is performed on the all data which is to be stored in storage system. The advantage of this scheme is that it is performed on all static data which has been already stored is storage system, and it will improve the efficiency .only the disadvantage is that it effect on performance and system will be little slower.

Online Deduplication:-
As comparing to offline data deduplication, it is performed when the data is uploaded at the time. The advantage of this scheme is that it allows space reallocation but the problem is that it increases the waiting time since the write operation of file is stopped until the duplicate files are removed.

Whole File Hashing:-
In this method the whole file is sent to hashing function. The MD5 and SHA1 hashing functions are used basically hash function is used to map the data to fixed size .the cryptographic hash allows data to map the hash values .hash functions are used for building caches of large data sets. Advantage is that due to full data backup the efficiency of system increased. Only drawback is that increasing granularity of duplicate data, it prevents the duplicate data it prevents the duplicate copies which differ by data byte.

Sub File Hashing:-
In this method before mapping the hash value the file is divided into several chunks i.e. called subfiles .The subfiles are divided with fixed length chunking and variable length chunking. Various fingerprinting schemes are used to determine chunk size, the broken files are transferred to cryptographic hash function [6].

Related Work:-
The data Deduplication scheme is used for neglecting the replicas of similar data types. This scheme provides homogeneous results which reduces the unnecessary space occupied by replicas of the same data type.

In 2013 mihirbellare.SriramKeelveedhi,T. Ristenpart ,"Dupless Server aided encryption for deduplicated storage ".they introducesd the concept of security and confidentiality by the scheme of symmetric encryption thry explained symmetric encryption as well as convergent encryption.the system enables client to store encrypted data with an service which checks it for deduplicated storage ,can reach the goal of perfoemance with more reliability.[2]

In 2013 at EUROCRYPT ,Author formalized a new cryptographic encryption called Message Locked Encryption for deduplication.

In 2002 J. R. Douceur,WilliumJ.B,D.Simon ,M. Theimer ,"Reclaiming space from Duplicate Files in a Serverless Distributed File System",provides a Farsite distributed file system for the purpose of reclaiming storage spaces from the storage. The system enables the identification and collect together all data files when they are provided with encryption with different users. They provided two schemes convergent encryption and other is SALAD self-arranging ,losssy,associative Database for aggregating file contents and location information scalable and fault tolerant manner. The system removes the copies of files with ideal content to storage systems.

Data Deduplication is a technique that is mainly used for reducing the redundant data in the storage system which will unnecessarily use more bandwidth and network. So here some common technique is being defined which finds the hash for the particular file and with that the process of duplication can be simplified, David Geer.

The concept of proof of ownership in cloud storage system is explained by D. Harnik. They identified the security issues related to the dropping out of date and time consumption by using the client side encryption

In 1997Author Adi Shamir proved the Secure Sharing Schemethey shown that how to divide the data D into the n piecessuch that they can be easily recovered by key Where every k-1 piece does not have the information about data. They proved the scheme for robust key management for system which can be secure and reliable for working.

In 1989 Author Michel O.R developed an information dispersal algorithm which breaks the file into blocks or small pieces it has numerous application to secure the storage information of computer networks for fault tolerance and efficient transmission of network.

s:-

Convergent Encryption:-

It is used to provide a data confidentiality in secure deduplication. It performs encryption and decryption operations by using the convergent key which is obtained from a cryptographic hash value. In data encryption process the user regains the keys and forwards the cipher text. Since the copies of similar, identical data will generate the similar convergent key and same cipher text. Gather for the decryption the cipher text and convergent key are used to get the original data. The convergent Encryption is defined with four functions. [3]

- Keygen (M) →K

In the key generation algorithm, it takes original data M as input. And maps it into a convergent key K.

- Encrypt (K, M) →C

In Encryption algorithm like symmetric encryption takes both key K and Data M as input and then outputs ciphertext c.

- Decrypt (K, C) → M

In decryption algorithm, it takes input as ciphertext C and the same convergent key K which is used for encryption and outputs the original data M.

- TagGen(M) →T(M)

The Target generation algorithm target the original data and outputs a tag to the data. By using this method two users having ideal data copies can get two ciphettexts with same encryption key hence the cloud service provider will easily able to perform deduplication scheme on this data.As the encryption keys are randomly generated from the data provided .so no need to communicate between data owner and users.

The Disadvantage of the scheme is that as encryption key is derived from the data itself. so if the intruder get access to data storage can break security such attacks are called as Dictionary Attacks. In such attack shared secret is compared with the data.

Symmetric Encryption:-

This encryption uses the similar secret key for the purpose of encryption and decryption. They can share the secret between two or more parties, which maintains private resource information. This encryption defined with the three functions.[10]

- Key Generation (KeyGen())→k

This algorithm generates the key for encryption. The random key ‘k' is said to be distinctive and nonspecific

- Encryption(Enc(m,k)) →C

This algorithm takes an input of the identical data ‘ M ' and random key ‘ K' which produce the output ciphertext. due to the unique key k the output C is also identical every time.

- Decryption(Dec(C,K)) → M

This algorithm generates the output as identical data plaintext M by taking input as a Key and Cipher thext.the same key must be used which is used at the time of encryption.

Proof Of Ownership:-

The aim of this proof of ownership is to prove their own data copies to cloud storage server. This is useful when we have implemented a system for only authorized users. A significantly more extreme and direct security danger from utilizing deduplicated distributed storage is that the enemy may pick up the responsibility for by just listening in on document hashes. Customer side deduplication can find that anybody possessing the record hash can pick up responsibility for document by transferring the document hash. In the storage system unwanted file operations are prohibited and then it adds authentic user. There are numerous measurements to assess the proficiency of a POW plan, for example, the data transfer bandwidth necessity, I/O overhead at both sender and receiver side must be computed.[1]

Basically cloud storage is of two type storage either primary memory or secondary is disk of various size .The proof of ownership design requires that file must be accessed by user as well as cloud service provider. The server must have to verify the users demand after that it will allow.

Message Locked Encryption:-
MLE is a symmetric encryption scheme in which the encryption and decryption keys are same and are derived from the message shared. As it enables duplication of cipher text, it lets key to the message. The encryption algorithm allow the cipher text and key to recover the message M. The tag generation algorithm maps the cipher text to a tag used for the server to detect deduplication in files .this scheme accepts the tag generation to cpphertext.it is resistant to fake attacks as practically they provide ROM security Analysis .they make connections with deterministic encryption hash function secure on correlated input for different message sources.[9]

Implementation of Deduplication:-
Deduplication term is defined as removing the same data copy or replicas of data. Data deduplication can be measured by following two function. [12].

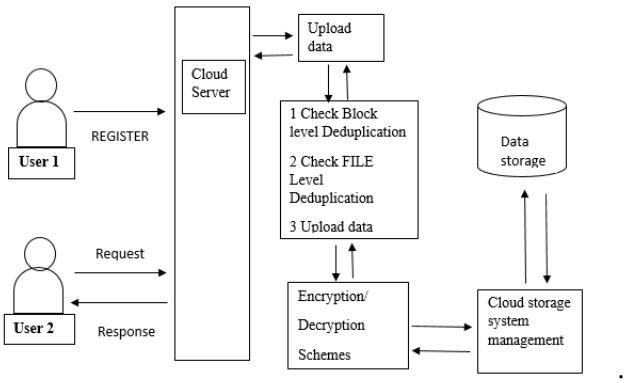

Fig 1:- System Architecture.

Deduplication Ratio:-
This is calculated as size of data before applying deduplication algorithm over size of data after applying deduplication algorithm.

Throughput:-
It is calculated as rate of data deduplication per second.as less time for performing it will increase its efficiency.

The term Data Deduplication is measured with functions as which type of data is to be deduplicated, where the duplication scheme has been taking place either source or destination device, timing for deduplication i.e. preprocessing or post processing of deduplication.

As number of approaches are present for Deduplication we describe them below.

Fingerprint Based Deduplication:-
The fingerprint based schemes involves the data chunks, it divides the data blocks into small size chunks. several chunks of same block are independent in processing. there are two types of data chunking they are static chunking as well as content defined chunking.

Static chunking involves splitting of data into same size chunks as well as fixed size chunking .

The content based chunking is preferred in backup system than splitting data in same identical size it reduces the repetitions in data deduplication [6].

Delta Based Deduplication:-
In this scheme it performs deduplication while writing the data blocks.it checks the redundancy at runtime if they found similar block as earlier one then they just encode its reference and call it in. The example of this is the IBM data backup system.
There are basically two types of data deduplication on basis of Implementation follows.

File Level Deduplication:-
In this type of Deduplication it compares at file storage .which are stored with number of records which has necessary elements. The index of file record is maintained. The index is changed only if the file is particular and single instance if not then only a pointer to previous file store references. The only single copy is saved and the other are replaced by the reference to the original file.

In this sort of thinks about a record that must be filed or reinforcement that has as of now been put away by checking all its vital properties against the list. The file is the record of the documents. The list is refreshed and put away just if the record is exceptional if not then just a pointer to existing document store references. The main single occurrence is spared and the other are supplanted by the stub to the first record.

Block level Deduplication:-
This type of duplication operates on the basis of the blocks. The time is broken into segments. The blocks of chunks will be examined for previously stored information. The popular approach to determine redundant data is assigning identities to the chunk of data by using various algorithm, which generates the unique ID of that particular block, the particular id is compared to the central index.

Data Deduplication scheme allows compression of data for removing the replicas of same data type. To improve the storage performance the scheme is applied to the online data to reduce the size of data. As in block level deduplication the blocks and chunks are compared to storage data size .the chunks are used to store the index values. The index maintains the records of uploaded data to system. As Compress which is a program based on LZW algorithm, which performs fast in less storage space. Another is DEFLATE a lossless data compression algorithm which is a combination of LZ77 and Huffman coding.

Secret Sharing Scheme:-
The scheme contains two algorithms Share and Recover. By using Share the user shares the key to access the file and by using Recover they can access any file on a cloud. In RSSS no information about the file reduced by any no. of shares.[7]

System with Tag Consistency:-
As we are preventing cipher text to be getting duplicate, because the main aim is to maintain only single copies of identical data. So tag provides security guarantees against the fake duplication attacks. The tag generation is purely performed by the data owner that's why chances of to be suffering from malicious attacks are minimized[2].

Conclusion And Future Scope:-
In this survey article we study the several approaches to maintain data with identical copies, deduplication schemes are very useful in today's life. As most of IT industries now turning towards data security and integrity various technologies are implemented. By stating all this scheme we expect the better enhancement.

References:-
1. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems." in ACM Conference on Computer and Communications Security, Y. Chen, G. Danezis, and V. Shmatikov, Eds. ACM, 2011, pp. 491–500.
2. Mihir. Bellare, Sriram. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage," in USENIX Security Symposium, 2013.
3. "Secure deduplication with efficient and reliable convergent key management" by J. li,Chen, jingwei,parteic and W.Lou ,in IEEE 2013.
4. " A Survey on Deduplication Scheme in Cloud Storage" of author Deepa D.,Revathi M.
5. "An Efficient and Secure Dynamic Auditing Protocol for Data Storage in Cloud Computing
6. "in IEEE transactions in 2012 ,Author K. yang and X.jia.
7. "Fingerprinting by random polynomials",by M. Rabin in 1981.
8. "How to share secret" by Adi Shamir in ACM commun.,1979.
9. Multiple Ramp schemes",by A. santis and B.Masuci in IEEE Transactions july 1999.
10. "Message Locked Encryption in secure deduplication"in Eurocrypt,2013.

11. “A Secure data deduplication scheme for cloud storage”in technical Report,2013 by J.Stanek,A. Sorniotti,Androulaki,kenel L.
12. ”Efficient dispersal of information for security ,loadbalancing,fault tolerance” in ACM journal 1989,M.Rabin.
13. “Reclaiming Space from Duplicate Files in a Serverless Distributed File System”. John R.Douceur, AtulAdya, William J. Bolosky, D.Simon, M. TheimerMicrosoft .
14. “A Survey on Data Deduplication in Cloud Storage Environment”,in IJSRET 2015.by Mani U.V.,G. Mohan