ISSN No:-2456 –2165

A Secure Access Policies Based Data Deduplication System

R. Hari Shankar Prasad
Department of CSE
SRM University

S. P. Maniraj
(Asst.prof)
Department of CSE
SRM University

Abstract—This paper presents a survey on an attribute-based storage system with secure deduplication in a hybrid cloud setting with higher confidentiality and reliability. Deduplication is a technique which is widely employed to backup data and minimize network and storage overhead by detecting and eliminating redundancy among data. It eliminates duplicate copies of identical data in order to save storage space and network bandwidth. Private cloud is responsible for duplicate detection and a public cloud manages the storage. Instead of keeping multiple data copies with the same content, in this system eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Each such copy can be defined based on user access policies, the user will upload the file with access policies and then file type question with answer. Then same file with different access policies to set the particular file to replace the reference. Applying deduplication to user data to save maintenance cost in HDFS storage system.

Keywords— Cloud Computing, Encryption Techniques, Classification, Data security, Authentication, Security in Cloud

I. INTRODUCTION

Cloud Computing is a combination of IT services provided by many service providers. The term cloud was originated from the internet and is also a platform that gives people the opportunity for sharing resources, services and information globally. In general, cloud computing has diverse definitions obtained by several important organizations. With infinite storage space provide by cloud service provider users tend to use as much space as they can and vendors constantly look for techniques aimed to minimize redundant data and maximize space savings. Users will access information according to their needs and most users access same information again and again, the cost of computation, application hosting, content storage and delivery is reduced significantly.

A technique has been introduced for reducing the amount of storage space an organization needs to save its data which is the data deduplication system. It helps in eliminating the storage systems contain duplicate copies of many pieces of data. For example, the same file may be saved in several different places by different users, or two or more files that aren't identical may still include much of the same data. Along with low ownership costs and flexibility, users require the protection of their data and confidentiality guarantees through encryption. To make data management scalable deduplication we are use Encryption for secure deduplication services. Both the deduplication system and encryption system are two conflicting technologies present. The cipher text and the user's private key depend on attributes. If the features of a user key counterpart to those of the cipher text, then decryption is permitted. ABE utilizes four algorithms: encryption, decryption, setup and key generation.

The DES algorithm ensures data security in cloud. The security design architecture of the system is planned by using DES cipher block chaining, which eradicates the hackers. The data which is sent, being interrupted and replaced has no danger. The system with encryption is adequately secure but for that, the level of encryption has to be amplified, as computing power upsurges. To secure the results, a symmetric key is used to encrypt the communication system between the modules. The cloud data security must be utilized to analyse the data security requirements, the data security risk, the data security process through encryption and disposition of security functions.

This encryption algorithm is used to address the security and privacy issues in cloud storage to shield the data from illicit access.

II. EXISTING SYSTEM

The previous deduplication systems which we have been discussed is only been considered having a single-server setting.

However, as lots of deduplication systems and Storage systems are intended by users and applications for higher reliability, especially in archival storage systems where data

are critical and should be preserved over long time periods. The deduplication storage systems requires and provides reliability higher when compared to other high-available systems. Specifically, each user must associate an encrypted cipher key with each block having its outsourced encrypted data copies, which is used later to restore the data copies., An own set of encrypted cipher keys is required, even though different users may share the same data copies so that no other users can access their files. Encryption mechanisms are often used to protect the confidentiality before forwarding the data into cloud. Commercial storage service providers are hesitant to apply encryption over the data because it makes de duplication impossible. Both the public key encryption and symmetric key encryption, require different users to encrypt their data with their own keys. As a result, identical data copies of different users will lead to different cipher texts As a result, the number of encrypted cipher keys being introduced in a linearly scales with the total number of blocks being stored and the number of users. Reliability is one of the other issue which is been faced in the present system . If the master key is accidentally lost or changed, then the user data cannot be recovered;, then the user data will be leaked. Cost increases to the storage of content as well as for the keys storage. Increase bandwidth with upload time.

III. DISADVANTAGES

- Single server system.
- De-duplication is not scalable.
- Many users with a various number of keys .
- Cost increases to the storage of content as well as for the keys storage.
- Security lacks

IV. PROPOSED SYSTEM

A. System Architecture

The deduplication process is carried out in the present system. The process is carried out and defines to provide the security to the file which are been uploaded in the cloud. To enable the deduplication and distributed storage of the data across HDFS, we have shown the concept of deduplication effectively and security is achieved by means of Proof of Ownership of the file. The convergent keys are been outsourced to the slave machines in a secure manner. The file-level deduplication and block level deduplications are been supported by decryption key. Avoidance of Key Management overhead and it provides fault tolerance guarantees for managing the key. .We have used the Triple DES Technique as the plain text is encrypted more than two times with the convergent key so that our data will be secured. Scalability is achieved as the decryption key is been achieved efficiently. Multiple users of same data is only referred and not added. This results in achieving the Cost Efficiency. Deleting content of shared file will allow deleting only convergent keys references but not content stored in HDFS file storage. User registers to the HDFS storage and login the page for uploading the file with required information. User chooses the file and uploads to Storage where the HDFS store the file in rapid storage system and file level de-duplication is checked. The MD5 message-digest algorithm is cryptographic hash function which is used to tag files. It produces a 128-bit hash value typically expressed in text format as 32 digit hex value so that files of same are de-duplicated. Shrinking the file chosen of fixed size and generating tags for each blocks which is compressed. After that generate convergent keys for each blocks split to verify block level de-duplication. Here we provide filename and password for file authorization in future. Encryption is done by the Triple Data Encryption Standard (3DES) algorithm. Finally the original content is encrypted as cipher text and stored in slave system. Blocks are stored in Distributed HDFS Storage Providers. After encryption the convergent keys are securely shared with slave machines provider to Key Management machines. Key management slave checks duplicate copies of convergent keys in KMCSP. The Comma Separated Values (CSV) file is used to check proof of verification and store keys in a secured manner. The proof of ownership is required for a user to delete own contents. The download request needs proper ownership verification of the document he requires to download. The ownership for the file is created by unique tag which is been generated by MD5 algorithm and it helps in verifying the existing tag of user. After the verification the original content is decrypted by requesting the Distributed HDFS storage. The HDFS storage request the key management slave for the decrypted keys and finally the original content is received by the user. When the User deletes the content only the reference content is deleted but not the original content.

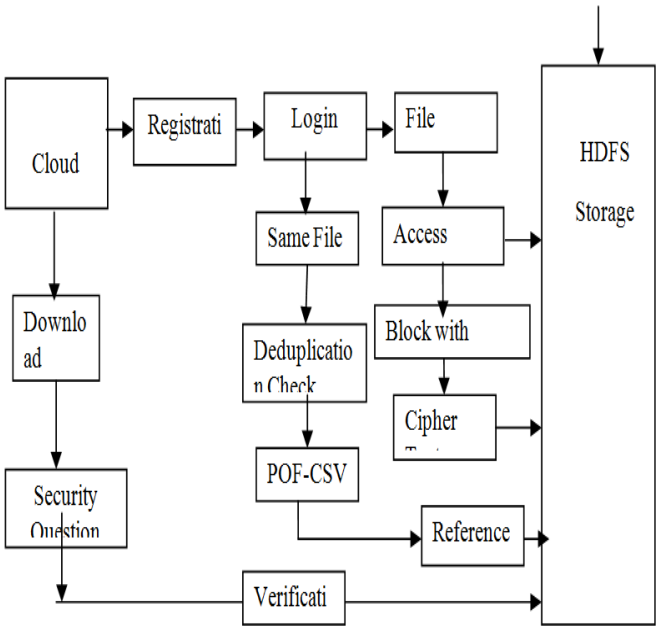

Fig. System Architecture

V. SYSTEM MODULE

A. User Registration

The User has to upload his/her files in a cloud server and should register first before uploading. Then only he/she can be capable of doing it. For that he requires to fill the details in the registration form. The user details are maintained in a database.

B. File upload with access policies

In this module User will chooses the file and uploads to Storage where the HDFS storage system .In the system will generate a signature in particular file and then split into multiple block. Each block will be generate signature with key. In the signature by using MD5 message-digest algorithm is cryptographic hash function producing a 128-bit hash value typically expressed in text format as 32 digit hex value so that files of same are de-duplicated. After that generate convergent keys for each blocks splitting to store CSV file like filename, file path, blocks, username, password and block keys. Encrypt the blocks by RSA algorithm is asymmetric cryptography algorithm. As the name describes that the Public Key is given to everyone and Private Key is kept private. Here the plain text is encryption to cipher text and stored in slave system. Blocks are stored in Distributed HDFS Storage Providers. After upload the file to set the access policies with set security question

C. Detection Deduplication Method

File-level data deduplication compares a file to be backed up or archived with copies that are already stored and it is done by checking the attributes against an index. If the file is unique, it is been stored .And the index is updated; if only a pointer to the existing file is stored. The result is that only one instance of the file is saved, and subsequent copies are replaced with a reference that points to the original file. Another one signature match checking looks within a file and saves unique Iterations of each block. All the blocks are broken into compressed with the same fixed length. Each compressed data is processed using a hash algorithm such as MD5 or SHA-1.

D. Download User File

The final model user request for downloading their own document which they have uploaded in HDFS storage. In this download request will analysis the user attribute once it will matched then ask the security questions for particular file. After complete the process needs proper ownership verification. After verification the original content is decrypted by requesting the Distributed HDFS storage where HDFS storage request key management slave for keys to decrypt and finally the original content is received by the user.

E. Advantages

- Improve the reliability of data
- The confidentiality of the user's data is achieved.
- Unique feature of the proposal is that data integrity, as well as tag consistency, can be achieved.
- Here we proposed the secured system and data owner can decide whether the user can access the system or not.

F. Mathematical Model

Let S be the system object.

It consist of following

S={U,F,TPA,CSP}
U= no of users

U={u1,u2,u3,…..un}

F= no of files

F={f1,f2,f3,…..fn}

B=no of blocks.

B{B1,B2,…,Bn}

TPA= Third Party Auditor
TPA={C,PF,V,POW}

C=challenge

PF =proof by CSP

V= verification by TPA

POW= proof of ownership

CSP= Cloud Service provider

CSP={PF,F}

PF=proof

F=files

G. Used Algorithms

- *KeyGen(F)*

The key generation algorithm gives a input as file content F and generate outputs as the convergent key ckF for F.

- *Encrypt (ckF;F)*

The encryption algorithm gives input as the convergent key ckF with file content F and generate the ciphertext ctF as output.

- *Decrypt (ckF; ctF)*

The decryption algorithm gives input, the convergent key ckF with ciphertext ctF and generate the plain file F as output.

- *TagGen(F)*

The tag generation algorithm gives input, a file content F and generate the tag tag F of F.

H. Advanced Encryption Standard:

The following AES steps of encryption for a 128-bit block are given below:

- Derive the set of round keys from the cipher key
- Initialize the state array with the block data (plaintext).
- Add the initial round key to the starting state array.
- Perform nine rounds of state manipulation.
- Perform the tenth and final round of state manipulation.

Copy the final state array out as the encrypted data (ciphertext)

VI. CONCLUSION AND FUTURE SCOPE:

We have proposed to enable de-duplication in convergent keys and distribute the convergent keys across multiple Key Management slave machines. And chunks in the various cloud storage providers and downloaded securely by receiving keys from Key Management slaves and Chunks from the Distributed HDFS with improved reliability.An attribute-based storage system which employs cipher text-policy attribute-based encryption (CP-ABE) and supports secure deduplication. The deduplication and distributed storage of the data across HDFS is achieved .And then using two way cloud in our storage system is built under a hybrid cloud architecture, where a private cloud manipulates the computation and a public cloud manages the storage.. If so, whenever it is necessary, it regenerates the cipher text into a cipher text of the same plaintext over an access policy which is the union set of both access policies.like public cloud and private cloud. We have shown the concept of deduplication effectively and security is achieved by means of Proof of Ownership of the file. That is attribute-based storage system cipher text-policy attribute-based encryption (CP-ABE) and supports secure deduplication. User privacy is enhanced by access requests to privately inform the cloud server about the users access desires .Forward security I realized by the session identifiers to prevent the session correlation. It is possibly applied for enhanced privacy preservation in cloud applications

REFERENCES

[1]. Gore Swapnali, Gore Supriya, Tengale Kanchan, Tengale Varsha, Asst.Prof. S. B. Bandgar " Modern Secure Distributed Deduplication Systems with Improved Reliability" International Journal of Advanced Research in Computer Science and Software Engineering October-2015.
[2]. Mr.Kulakarni Harish, Mr.Ravi kumar chandu " Shared Authority Based Privacy-preserving Authentication Protocol in Cloud Computing" International Journal of Engineering Research and Applications (ijera national conference on Developments, Advances & Trends in Engineering Sciences (January 2015)
[3]. R.Bindu, U.Veeresh, CH. Shashikala "Provable Multicopy Dynamic Data Possession in Cloud Computing Systems" International Journal of Computer Engineering In Research Trends January-2016.
[4]. Waghamare Amol Arjun "A Secure Hybrid Cloud Approach to Avoid Deduplication" International Journal of Computer Science and Mobile Computing April 2015.
[5]. Zodge Kalyani1, Amruta Amune "review on secure distributed deduplication systems with improve reliability" Global Journal of Advanced Engineering Technologies 2016.nd Taamoto