Data Protection with De-duplication in Cloud Computing

Kandasamy.V[1], Siva Alagesh.S[2], Pradeepraj.C[3]
[1]Assistant Professor, Department Of Information Technology,
Loyola Institute Of Technology, Chennai.
[2,3]UG Students, Department Of Computer Science and Engineering,
Loyola Institute Of Technology, Chennai.

***Abstract*— Attribute-Based encryption technique is used to encrypt the Data. This differs from asymmetric (or public-key) encryption, which uses one key to encrypt a message and another to decrypt the message. In ABE system does not tolerate secure de-duplication. To solve this problem we are proposing a new encryption technique called Cipher Text Policy-based Encryption (CP-ABE). CP-ABE uses file access tree structure (folder inside the folder) to encrypt data. In this paper, we proposed Equality Checking Algorithm to check the files/data whether it's duplicate or not. Any duplication files present it will intimate the data owner. The Symmetric Algorithm is used to encrypt the files/data for security purpose and this project implemented for AmazonS3 Cloud. An Amazon cloud did not detect duplicate files, its only check file names and if you upload same name and format but different content during an Amazon cloud is replaced in existing file. But, If you upload different name and same content then, your file uploaded. It will not check the content of the file. In other clouds (Dropbox, cloud, etc.) not check duplicate files, it changes the name of the file(1), file(2), etc. But, we proposed system check content also. A user uploads a file with different content then it can be duplicated and highlight your duplicate files.**

***Index Terms*— CP-ABE, storage, De-Duplication, AmazonS3-cloud**

1) INTRODUCTION

The cloud computing is the widely used technology that enables the virtual storage, it means renting the space from unknown places. Also Mining technique is applied for checking the matching of contents presents on a cloud.

Cloud computing greatly facilitates data providers who want to outsource their data to the cloud without disclosing their sensitive data to external parties and would like users with certain credentials to be able to access the data. This requires data to be stored in encrypted forms with access control policies such that no one except users with attributes (or credentials) of specific forms can decrypt the encrypted data. An encryption technique which meets the Cipher Text attribute-based encryption (CP-ABE) a user's private key is associated with an attribute and a message is encrypted under an access structure over a set of attributes. Also, a user can decrypt a cipher text with his/her private key if his/her set of attributes satisfies the access policy associated with this cipher text. In spite of, the accepted ABE system decline to manage highly secure duplication, which is a to save storage, space, network, and bandwidth by eliminating redundant copies of the encrypted data which is stored in the cloud. To the best of our knowledge, the existing constructions for secure de-duplication are not built on attribute-based encryption. Nevertheless, since ABE and secure de-duplication have been widely applied in cloud computing, it would be desirable to design a cloud storage system possessing both properties.

With the help from Amazon S3 cloud, bulk amount of data (Big-data) can be handled for storing on it. Sometimes it may causing with storage running out properties based on the same data replication. In previous they are work with the avoidance of duplication based on the attributes of the particular data/file it may be name, extension of the file.

For the secure de-duplication we may go with ciphering techniques. Here the another concern for space after de-duplication is compression of files based on Huff-Mann compression technique.

Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. Promising as it is, an arising challenge is to perform secure deduplication in cloud storage. Although convergent encryption has been extensively adopted for secure deduplication, a critical issue of making convergent encryption practical is to efficiently and reliably manage a huge number of convergent keys[2]. Here in this paper security plays the major role for protecting the data/file over the Amazon s3 cloud.

2) LITERATURE SURVEY

MLE brings a plan to manage protected de-duplication (space-efficient secure outsourced storage), a goal presently targeted by numerous cloud-storage providers. Based on this infrastructure, we make both practical and theoretical supplements in 2013[1]. Data deduplication is a technique for eliminating duplicate copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. Promising as it is, an arising challenge is to perform secure deduplication in cloud storage. Although convergent encryption has been extensively adopted for secure deduplication, a critical issue of making convergent encryption practical is to efficiently and reliably manage a huge number of convergent keys in 2014 [2]. In our system, a user's private key will be associated with an arbitrary number of attributes expressed as strings. On the other hand,

when a party encrypts a message in our system, they specify an associated access structure over attributes.[3] Cloud computing promises a more cost-effective enabling technology to outsource storage and computations. The hardware-based solutions are not scalable, and fully homo morphic encryption is currently only of theoretical interest and very inefficient in 2011[4].

3) EXISTING SYSTEM

A) Convergent Encryption-The process of calculating the Hash-Key value from the original file is known as Convergent Encryption Techniques. They are considering the private key as the pre-calculated hash-key then they encrypt the rest of files. At last using password, encrypted hash key value should get stored in somewhere else. The one and only way to access the password again for file by holding the original file.

This Convergent encryption is open for a "Fire attack" therefore an attacker can effectively confirm whether a target arrests a particular file by plain-text or not and then simply the comparing output with files captured by the target source of file. This type attack only solve the problem for a non-unique information which the user stored. The confirmation of an argument that could be made by the fire attack easily ineffective simply by adding the random characters to the plain text such that unique piece of Data therefore we should get the unique encrypted file. Make the file as unique by adding the broken blocks of Plaintext independently, convergent, encrypted by using this Convergent Encryption Methodology.

4) PROPOSED SYSTEM

A) Cipher Text-Policy Attribute-Based Encryption (CP-ABE) Algorithm- In the CP-ABE scheme, each user's decryption key is a set of aspects describing that user's permissions. When a cipher text is encrypted, a set of attributes is nominated for the encryption, and only users tied to the relevant attributes are able to decrypt the cipher text.

The example given on the website presents a cipher text encrypted such that only employees with the aspects "Human Resources" UNION "Executive" is to able decrypt it. HR employees have the "Human Resources" attribute tied to their private keys, and Executive employees have the "Executive" attribute tied to their private keys. Therefore, it is able to decrypt the encrypted message. Unlike other Role-Based Access Control (RBAC) systems, CPA does not require a trusted authority or any form of storage. The encryption itself serves as the RBAC mechanism.

5) SYSTEM ARCHITECTURE

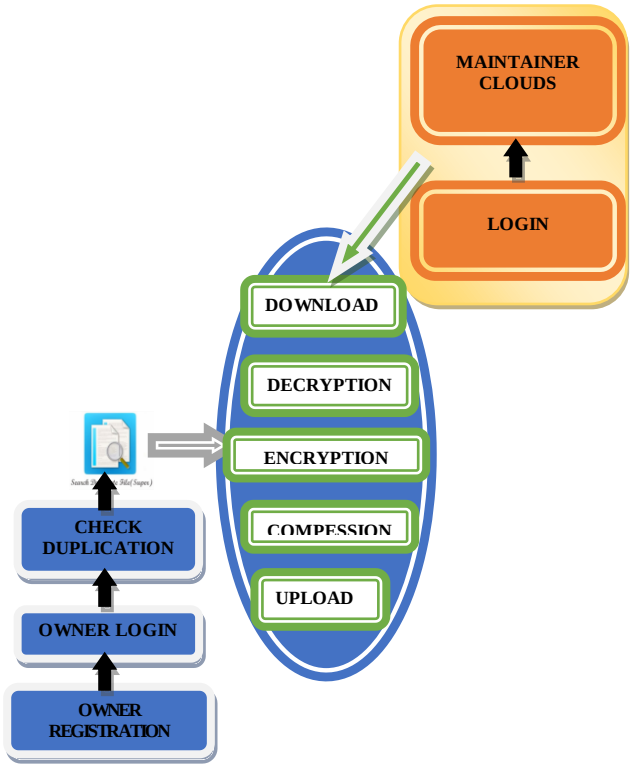

Fig (1): Architecture Diagram

6) MODULE DESCRIPTION

A. Authorization control creation and Key Generation

B. Owner uploading and Built Hybrid Cloud

C. Detect De-duplication

D. Key Exchanging

E. Verification and File Retrieving

A) Authorization control creation and Key Generation:

An Authorized user is to access their particular files for uses they must use their own private keys for executing their queries. Also authorization provides highly authenticated secure sharing of resources over the cloud from private to public cloud and vice versa. Key is mainly used for creating the password for secure authentication for every files. Here in this module there are two main concerns are developed one is for secure authorization control and another concern for generating key values for the every files for providing high protection. All the authorized values are provided by CA (Certificate Authority). After getting the CA values it may get the tokens for every transaction of file it may uploading/downloading it is to be stored in the S-CSP. By using this unauthorized users access can be eliminated by the duplicate check of key value stored in the S-CSP.

B) Owner uploading and Built Hybrid Cloud:

Owner of the cloud has to upload their data/files to cloud. By using the hybrid cloud architecture then only the user or owner of the file should get highly protection. Here the private keys are not shared to users these things are kept by private cloud server. By sending the request to private cloud server then only user able to get the file token. Check for

duplication only done with the help of this file tokens.

C) Detect De-duplication:

Data confidentiality helps the avoidance of de-duplication in Convergent Encryption. Convergent key can be derived original copy of data and then encrypts the inherited copy of original file with the convergent key. Duplicated copies from the cloud can be detected by the inherited tag generated by the user. If the tag may holds the correctness between the original copy of file and another file should consider as the duplicate one based on their percentage of content it is directly shown to the user, deletion is based on the user's choice. If they maintains to cloud means they may carried out or otherwise by using the compression technique they may compress it. No both two convergent keys promises with the data confidentiality. After the process both two files should be stored over the server side.

D) Key Exchanging:

Private cloud manages the private keys for requesting the tokens from users. Private cloud's interface provides the way to allow the user to submit the both files and queries these are computed and stored securely. The server belongs to the private cloud will automatically check the user's query based on the identity depends on their reservation. Checking of duplicate files can be performed before the operation of uploading.

E) Verification and File Retrieving:

Private cloud receives the set of keys from each user as symmetric key. Verification is done with the help of identification protocol by using the individual proof and verification algorithms respectively. Assume the key *U* for perform the identification with the servers as a secret key. Privilege for the particular user as *PU*. Ownership protocol for the defined user as *POW* based on the information. Table of information for public cloud details should be stored by the private cloud server as *PKU* for its privilege. S-CSP is used to check the details of file whether it is already available or not if not we can download the file or otherwise it sends the abort signal to the user as download failed. Original file can be recovered by using ciphering key which is used in S-CSP.

7) RESULTS AND DISCUSSIONS

Results are based on the user inputs i.e Data/files from the user-end to the cloud end. Every transaction of data must need a secure authentication so here we are securing our cloud storage by username & password. Every user must log-in the system every time.

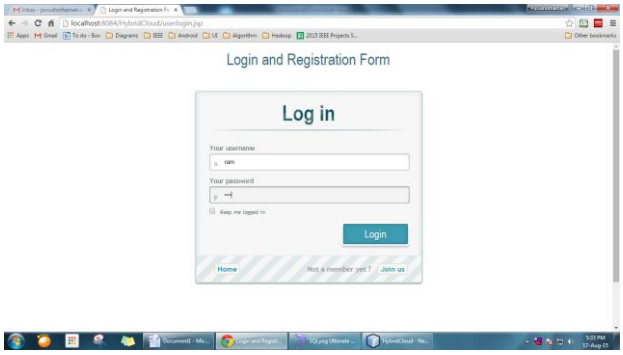

Fig (2): Log-in Window

After the successful login, user may need to upload file to cloud for storing purpose. Here the notification sends to user about the file if it is already available or not. Based on this user has to de-duplicate their files.

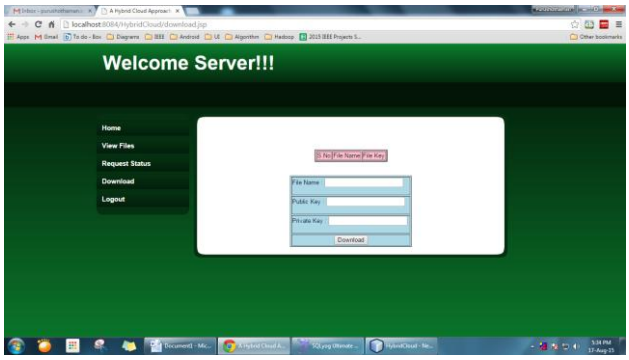

Fig (3): Check for file availability

At last all files to be listed before the user after the successful de-duplication from user's private cloud.

Fig (4): List of Files After de-duplication

8) CONCLUSION

Cipher Text Policy Attribute-based encryption (CP-ABE) has been widely used in cloud computing where data providers resource with their encrypted data to the cloud and can share the data with users possessing specified credentials. On the other hand, de-duplication is an important technique to save the storage space and the bandwidth of network, which dismisses the duplicate copies of same data. In this paper, we presented a novel approach to realize cipher text policy attribute-based storage system supporting secure de-duplication. 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 the proof is valid, the private cloud runs a tag matching algorithm to see whether the same data underlying the cipher text has been stored. 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.
From this paper it only supports word documents with any type of extensions and also replicated images also. In future we extend the work for all the types of multimedia files over the cloud such that videos, audios and 3Dimensional video also. By using the efficient multimedia algorithm for reducing the Band width.

REFERENCES

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart. Message-locked encryption and secure deduplication. In *EUROCRYPT*, 2013.

[2] N.O.Agrawal, S.S.Kulkarni Secure Deduplication And Data Security With Efficient And Reliable CEKM, 2014.

[3] John Bethencourt, Amit Sahai, Brent Waters, Cipher text-Policy Attribute-Based Encryption

[4] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider. Twin clouds: An architecture for secure cloud computing. In *Workshop on Cryptography and Security in Clouds (WCSC 2011)*, 2011.

[5] P. Anderson and L. Zhang. Fast and secure laptop backups with encrypted de-duplication. In *Proc. of USENIX LISA*, 2010.

[6] M. Bellare and A. Palacio. Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In *CRYPTO*, pages 162–177, 2002.

[7] Jingwei Li Chuan Qin, Patrick P. C. Lee, and Jin Li, Rekeying for Encrypted Deduplication Storage

[8] Jian Liu, N. Asokan, Benny Pinkas, Secure Deduplication of Encrypted Data without Additional Independent Servers

[9] Mihir Bellare, Sriram Keelveedhi, Thomas Ristenpart, DupLESS: Server-Aided Encryption for Deduplicated Storage

[10] Paul Anderson, Le Zhang, Fast and Secure Laptop Backups with Encrypted De-duplication.

AUTHORS BIBLIOGRAPHY

Mr. V. KANDASAMY, M.E.,
Assistant Professor,
Department of Information Technology,
Loyola Institute of Technology,
Chennai.

S. SIVA ALAGESH,
Final Year UG Student,
Department of Computer Science and Engineering,
Loyola Institute of Technology,
Chennai.

C. PRADEEPRAJ,
Final year UG Student,
Department of Computer Science and Engineering,
Loyola Institute of Technology,
Chennai.