International Journal of Emerging Technology and Innovative Engineering
Volume 2, Issue 2, February 2016 (ISSN: 2394 – 6598)

SECURE DATA POSSESSION AND DEDUPLICATING DATA IN CLOUD FOR STORAGE AS A SERVICE

K.Vivekrabinson
Computer Science and Engineering
Kalasalingam Institute of Technology
Anand Nagar,Krishnankoil
Mail-id: vivekrabinson1993@gmail.com

R.Parvadha Devi M.Tech
Assistant Professor, Computer Science and Engineering
Kalasalingam Institute of Technology
Anand Nagar,Krishnankoil
Mail-id: parvadha.ramar@gmail.com

ABSTRACT

Cloud storage means the storage of data online in the cloud. In this paper, we show how to securely store data in cloud and also how to remove duplicated data from cloud storage. Specifically, aiming at provide efficient and reliable storage to user. For that we propose two technique use Diff algorithm to eliminate duplicated data by comparing the uploaded file with existing files. If the file match with some other then the file get discarded otherwise it will moved to next process called encryption. By the help of duplicate elimination user can use their storage space efficiently. The second technique is advanced encryption standard for encrypting the user files to prevent from intruders. After duplication check gets completed encryption process automatically triggered without any interaction of client. It will produce a encrypted text based on AddRoundKey and Lookup tables. So that no one can easily break the file. Similar to encryption we can reverse the process to get original files. So we can easily avoid unauthorized modification. Thus the cloud confirms security.

Keywords— cloud storage,advanced encryption,deduplication

I. INTRODUCTION

A. Cloud Storage

Cloud storage is an industry term for managed data storage through hosted network (typically Internet-based) service. Several types of cloud storage systems have been developed supporting both personal and business uses.

The most basic form of cloud storage allows users to upload individual files or folders from their personal computers to a central Internet server. This allows users to make backup copies of files in case their originals are lost. Users can also download their files from the cloud to other devices, and sometimes also enable remote access to the files for other people to share. Businesses can utilize cloud storage systems as a commercially-supported remote backup solution. Either continuously or at regular intervals, software agents running inside the company network can securely transfer copies of files and database data to third-party cloud servers. Cloud storage services may be accessed through a co-located cloud computer service, a web service application programming interface (API) or by applications that utilize the API, such as cloud desktop storage, a cloud storage gateway or Web-based content management systems.

*B. **Duplicate Elimination***

Duplicated data are important problem now a day likes storing the same data into storage space. 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.

It will drain the storage provided by cloud provider. For example, Consider Google drive a user registered with Google account can use particular amount of storage space provided by Google. But the main drawback is it doesn't have the facility of duplicate elimination. A user can save same file many times into Google drive, thus it will reduce the efficiency of Google drive.

In order to overcome this problem we propose cloud storage with duplicate elimination facility. It was achieved by the help of Diff Algorithm.

C. *Encryption*

Security is the major concern in our modern life. In cryptography, encryption is the process of encoding messages or information in such a way that only authorized parties can read it. Encryption does not of itself prevent interception, but denies the message content to the interceptor. In an encryption scheme, the intended communication information or message, referred to as plaintext, is encrypted using an encryption algorithm, generating cipher text that can only be read if decrypted. An authorized recipient can easily decrypt the message with the key provided by the originator to recipients, but not to unauthorized interceptors.

In olden days there is no security provided for user data. They mainly concentrated on authentication like login modules. There is no security for user data stored in that cloud. To achieve security, use encryption technique to encrypt user data and avoid modification of user data.

In order to overcome this problem we propose cloud storage with encryption facility. It was achieved by the help of Convergent encryption technique.

II. RELATED WORK

Halevi et al [2] have proposed technology that keeps their cost down is deduplication. In order to remove duplicated data client generate hash signature and attached to particular file. In this paper use proof of ownership protocol to eliminate duplicated by comparing the hash signature which was attached to that particular file. In this method user have full access on creating hash signature and attaching to particular file. But the major disadvantage is the sizes of hash signature are too small. The possibility of occurrence of same signature is high because of these are happened based on user knowledge. So it will produce inconsistency like assign same hash signature for two different files. Thus the reliability of cloud storage gets failed.

Li et al [12] have proposed ensuring the integrity of data storage in Cloud Computing. At first the file uploaded by user sends to cloud server. Cloud server redirects the file to cloud auditor for encryption. Once auditor encrypt the file he send back to cloud server. He verifies the encryption which was done by auditor is correct or not. If it was correct then upload the file otherwise once again repeat the process. This method achieves high integrity for the files because of the two step verification process. But the major disadvantage is it takes more amount of time for single process.

Yuan et al [13] have proposed POR(Proof Of Retrivability) and PDP(Proof Of Data Possession) to achieve security for the files stored in the cloud storage. And also used POW(Proof Of Ownership) protocol to eliminate duplicated data in public cloud. Late endeavors to this issue present huge computational and correspondence costs and have additionally been demonstrated not secure. It requires another answer for backing productive and secure information integrity auditing with storage deduplication for cloud storage. In this paper they proposed Public and Constant cost storage integrity Auditing scheme with secure Deduplication (PCAD).In this scheme use polynomial based authentication scheme. And the integrity checking and duplicate elimination are done by third party auditor. Anyone can register their identity as auditor. The main disadvantage of this method is the untrusted auditor can easily modify our source files without any knowledge of the users. So it will provide major inconvenience to users those who are registered with particular cloud storage.

Our approach various from the above mentioned work and overcome the problems like strengthen the encryption technique and also done encryption by cloud server. At first duplicated data are eliminated by the help of Diff algorithm and then they are moved to encryption phase. It will gradually reduce the time by avoid encryption of the duplicated file

III. SYSTEM DESIGN

A. *Block Diagram*

The below **Fig 1** represents overall block diagram for how user registered and how they are store and retrieve the files.

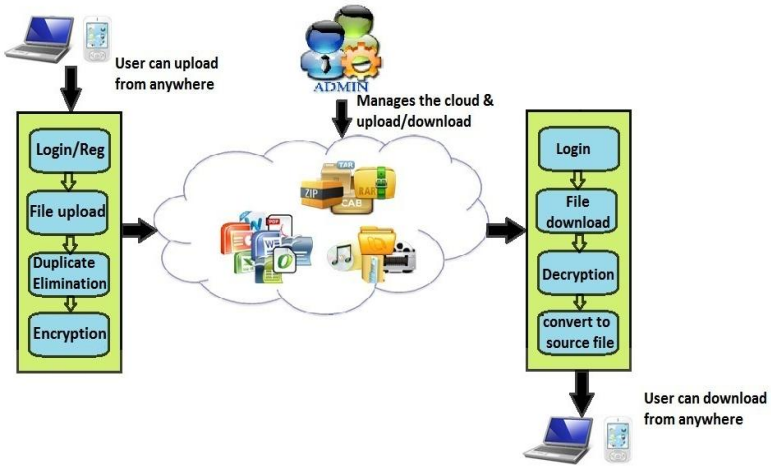

Fig. 1. OVERALL BLOCK DIAGRAM

In Secure data possession, similar to all cloud user first register their identity via registration. Once user gets registered, admin process the registration request and send user id and password to user via email. After that, user can login into cloud. User can upload their files using file upload options. To achieve reliability once the user uploaded the file diff algorithm check the uploaded file with existing file stored in cloud for whether the uploaded data already present at cloud or not. If the is not a duplicated one then the encryption process automatically triggered otherwise the new file override the existing file. If user request for particular file while encrypting, we use temporary cache memory to support the user. It helps to achieve reliability. We achieve reliability, using advanced encryption technique. After the completion of these two processes the file gets stored into cloud.

If the user wants to download the file once again he have to login to cloud and click download option. Once the user click download button file get automatically decrypted and converted to original file. Now the user can view the file which he has uploaded without any inconvenience.

The role of admin is to manage, create, delete all the users as well as manage the cloud.

B. *Diff Algorithm*

The program diff [6],[7] reports differences between two files, expressed as a minimal list of line changes to bring either file into agreement with the other. Diff has been engineered to make efficient use of time and space on typical inputs that arise in vetting version-to-version changes in computer-maintained or computer-generated documents. Time and space usage are observed to vary about as the sum of the file lengths on real data, although they are known to vary as the product of the file lengths in the worst case.

The central algorithm of diff solves the 'longest common subsequence problem' to find the lines that do not change between files. Practical efficiency is gained by attending only to certain critical 'candidate' matches between the files, the breaking of which would shorten the longest subsequence common to some pair of initial segments of the two files.

C. *Advanced Encryption Standard*

The Advanced Encryption Standard or AES [8],[9] is a symmetric block cipher used by the U.S. government to protect classified information and is implemented in software and hardware throughout the world to encrypt sensitive data.

It has the following steps to encrypt the user data

- KeyExpansion : It require 128 bit key for each encryption
- InitialRound :
  1. AddRoundKey : Each byte of file compared with key using bitwise xor.
- Rounds :
  1. SubBytes : Each byte is replaced with another byte based on lookup table.
  2. ShiftRows : Based on the offset value if shift the bytes in each row.
  3. MixColumns : Combine the bytes of each column with one another using invertible linear transformation.
  4. AddRoundKey
- Final Round (no MixColumns) :
  1. SubBytes
  2. ShiftRows
  3. AddRoundKey

Similarly the decryption process consist the reverse process of encryption and the steps are

- Inverse ShiftRows
- Inverse SubBytes
- Inverse AddRoundKey
- Inverse MixColumns

IV. SYSTEM IMPLEMENTATION

The system [7] is implemented by the help of openstack tool and base operating system as ubuntu. Some of the screen shot of the system are given below.

Fig. 2. User registration page

Fig 2 shows user registration page for new users. By the help of this site user register their identity and the details are stored into database.

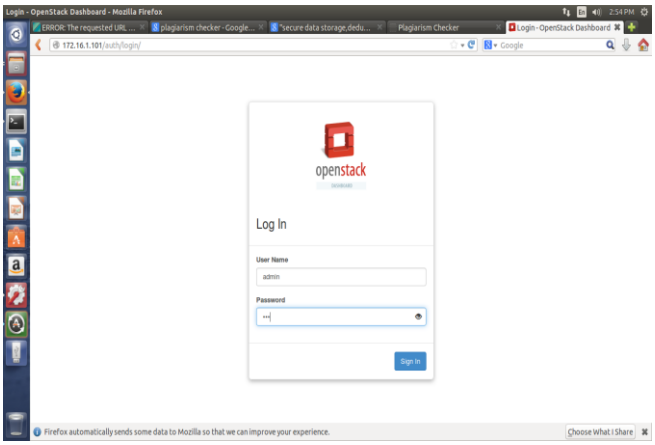

Fig. 3. Cloud login page

Fig 3 shows login page for all users. And it was accessed by the help of clicking sign in link in the website.

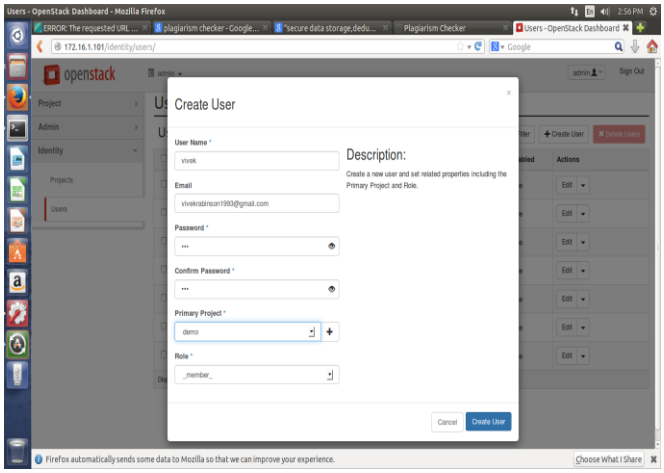

Fig. 4. User creation process for registered users

Fig 4 shows user creation process and which was done by the help of admin. Here admin create the account based on the information which was provided by the users via registration and send their userid and password to users through mail.

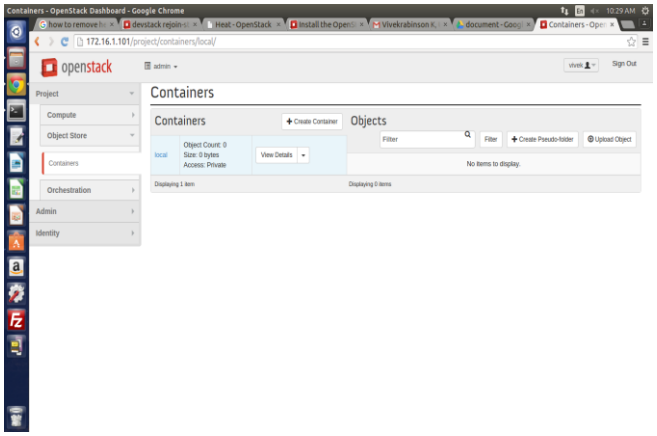

Fig. 5.Storage details for particular user

Fig 5 shows storage information for particular users. Once user receives userid and password they can login into cloud and store their data into cloud.

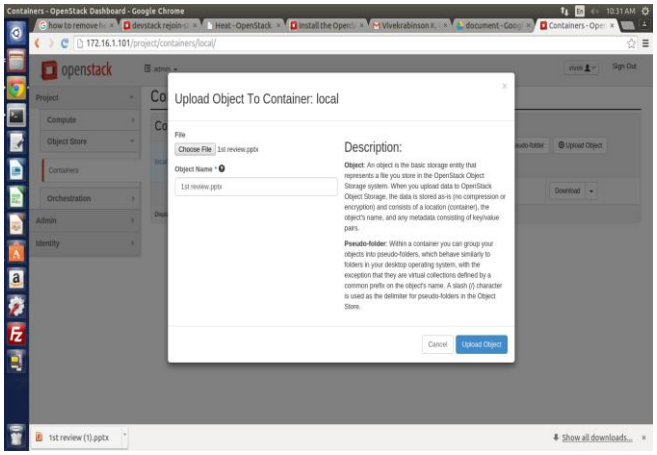

Fig. 6. Uploading data into cloud

Fig 6 shows file uploading process. To upload files into cloud user need to click upload button and choose file from local machine. And provide name for file we are going to upload.

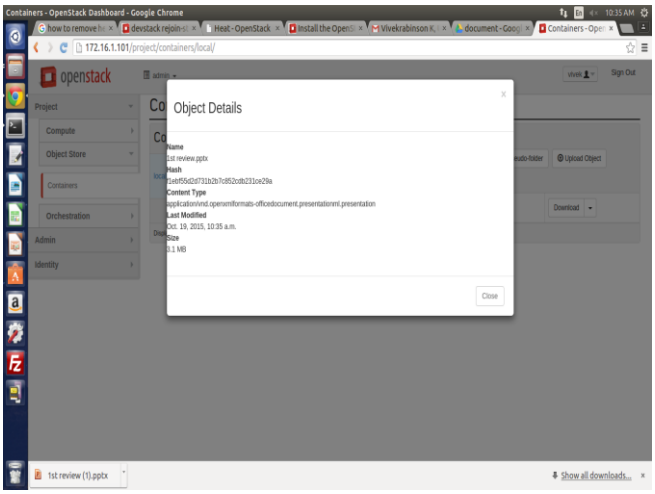

Fig. 7. Uploaded file information

Fig 7 shows uploaded file information. And these include time and date of uploaded, name and hash key. Once we click upload button it will automatically check duplicates and encrypts the file.

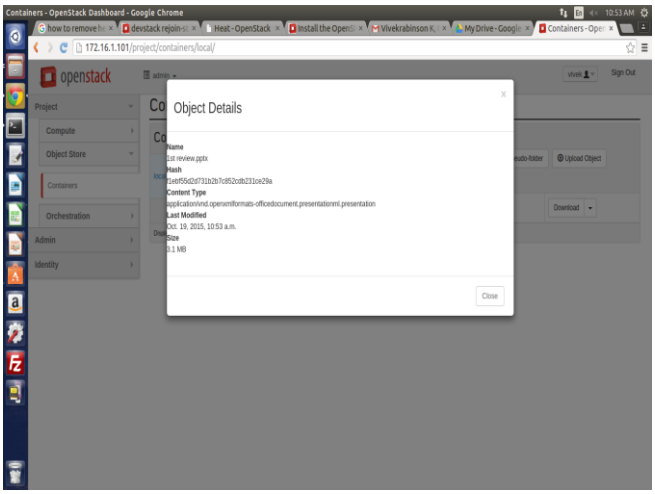

Fig. 8. Uploading same file

Fig 8 shows the process of uploading the same file which was already present in the cloud. If we upload the same file it caught on duplication check and the older file will get eliminated. Because the new file may have some new information. And the details of newly uploaded shown in above figure.

V. RESULT ANALYSIS

In this section we attempt to analyze the result of our proposed scheme. In this project the cloud is configured by the help of the guide which was downloaded from openstack website [3]. After that have to configure swift which was provide storage services by the help of same guide.

After configuring all the components we have to create website for new user registration and feedback and contact information. At next connect website with openstack to complete the process. The registered user information will get stored in website alone. Once the connection established successfully we can access the cloud from anywhere.

At next we have to set break point [4] to cloud in order to add duplication check and encryption to cloud. Upload Diff algorithm into cloud for eliminate duplicated data. Using security guide, [5] enable security to data which are passed in duplication check. And the security is done by encryption process. While downloading the process get reversed like decrypt user files and bring back to them. Using this cloud user can upload 5GB of data at a time. The benefit of this tool is, it will automatically replicate data into multiple partition and will avoid single point failure.

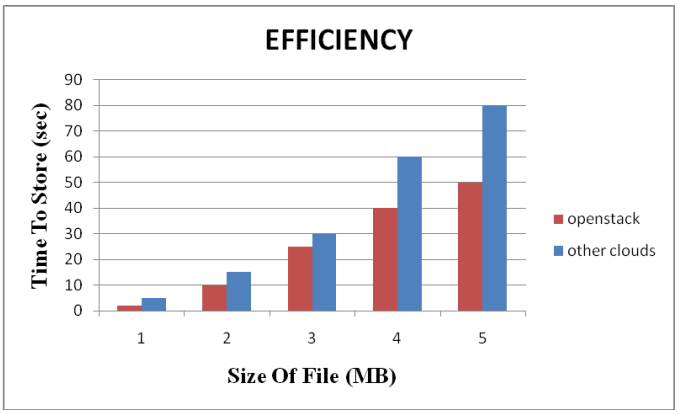

Fig. 8. Efficiency chart

In this chart shows the efficiency of the cloud. The graph is plotted for size of file (X-axis) and time to store (Y-axis).compared to some other cloud like open nebula and eucalyptus, openstack provide faster data storage. The main difference between openstack with some other cloud are, it segregate the file into number of pieces and then stored them. Thus it improve the efficiency by reduce the amount of time required to store the file.

The main advantage of this project is mobile friendly. User can access their cloud using their handheld device with the help of internet. Due to this we no need to buy mobile with larger internal storage. Simply buy a basic mobile with internet connection and upload the files whenever we need.

VI. CONCLUSION & FUTURE WORK

Thus the system created, provide cloud environment to users and user can store and retrieve their files into cloud from anywhere and any device. Thus the reliability of the system and eliminate duplicated files from the cloud achieved by diff algorithm. Security, which is the main drawback of the cloud system is well managed and tightened by using the AES encryption standard.

Even though this process works well the following works are still unrevealed and can be concentrated in the near future.

- Provide Software As A Service to users
- Combine Storage As A Service and Software As A Service
- Allocate instance to users
- Increase number of users

REFERENCES

[1] Ateniese G, Burns R, Curtmola R, Herring J, Kissner L, Peterson Z, and Song .D,(2007) "Provable data possession at untrusted stores," in Proceedings of the 14th ACM Conference on Computer and Communications Security, ser. CCS '07. New York, NY, USA: ACM, pp. 598–609.

[2] Halevi S, Harnik D,Pinkas B, and Shulman-Peleg A, (2011)"Proofs of ownership in remote storage systems," in Proceedings of the 18th ACM Conference on Computer and Communications Security. ACM,pp. 491–500.

[3] http://docs.openstack.org/juno/install-guide/install/apt/content/

[4] http://ais.seecs.nust.edu.pk/colonial/siteData/Documentation%20on%20adding%20encryption%20to%20OpenStack%20Swift_v1_4.pdf

[5] http://docs.openstack.org/security-guide/tenant-data/data-encryption.html

[6] http://c2.com/cgi/wiki?DiffAlgorithm

[7] https://en.wikipedia.org/wiki/Diff_utility

[8] https://en.wikipedia.org/wiki/Advanced_Encryption_Standard

[9] http://www.tutorialspoint.com/cryptography/advanced_encryption standard.htm

[10] Jingwei Li, Jin Li, Dongqing Xie and Zhang Cai,(2015)"secure auditing and deduplicated data in cloud", at IEEE transaction of computers,doi10.1109.

[11] Li J,Chen X,Li M,Li J,Lee P, and Lou W,(2014) "Secure deduplication with efficient and reliable convergent key management," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp.1615-1625.

[12] Li .J, Tan .X, Chen .X, and Wong .D, (2013)"An efficient proof of retrievability with public auditing in cloud computing," in 5th International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 93–98.

[13] Yuan J and Yu S,(2013) "Secure and constant cost public cloud storage auditing with deduplication," in IEEE Conference on Communications and Network Security (CNS), pp. 145–153.