OPTIMIZING INFORMATION LEAKAGE IN MULTICLOUD STORAGE SERVICES

Nanbon Abera Tolasa, Master of Technology, Department of Software Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), School of Information Technology, Hyderabad, Telangana

Dr. G.Venkata Rami Reddy, Professor, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University Hyderabad (JNTUH), School of Information Technology, Hyderabad, Telangana

Abstract - There are many advanced strategy to store data on multiple clouds storage providers. We can control information leakage which occur on single cloud by distributing our data over different multiple clouds. But unplanned distribution of data would produce high probability of information leakage in multiple clouds.

In this project work, I found the problem of information leakage caused by unplanned distribution of data chunks in multicloud. Then, I produced StoreSim system which aware leakage of information storage system in multiple clouds. The aim of this system is syntactically store the similar data on the same cloud, which minimize information leakage. Therefore, I designed MinHash algorithm to efficiently generate similarity-preserving signatures for data chunks and designed a function to compute the information leakage based on these signatures which hashed by fingerprinting algorithms such as SHA-1 and MD 5.

Lastly, provided optimal information leakage storage system based on planned distribution of data chunks across multiple clouds**.**

1. INTRODUCTION

1.1 Introduction

Now-a-days technology is growing and everyone is using different devices such as mobile phones, tablets and computers to store their massive critical data. There are many cloud storage providers such as Microsoft OneDrive, iCloud, and Dropbox which used to store users data over the cloud. These storage services have good demands due to the simplicity and cheap storage price. However, these storage providers are taking the control of user's data which may leak the user's data by different reasons such as trap door, hack bribe and coercion. [1]

The proper way to reduce the degree of information loss is using multiple clouds, which used to reduce one point failure in single cloud. The recent cloud storage providers like, Dropbox is used rsync-like protocols to operate the local file to remote file in their storage. In rsync-like protocols each user file is divided into chunks and hashed with fingerprinting algorithms such as SHA-1, MD5. Hence when a local file is modified, the changed hash will be uploading to the cloud. In fact, now a days service providers like Dropbox, Google Drive are used data deduplication methods to check similarity between data chunks by their fingerprints, but this fingerprint will check only as data nodes are duplicate or not. It is simple to check identical chunks, but to efficiently find out similarity between chunks is a complex task due to lack of similarity preserving signatures. [7]

Therefore, to address this problem I provided StoreSim which aware information leakage storage system by storing similar data to same cloud and I designed MinHash algorithm to efficiently generate similarity-preventing signatures for data chunks and designed function to control information leakage.

1.2 Objective of the project

Most of the time people are storing their data's over the centralized multiple clouds storage providers such as Microsoft Azure, Dropbox and iCloud, which are lead to control their data and one point of failure in multiple clouds. [7] However, unplanned distributions of data on multicloud storage providers would produce high degree of information leakage and one point failure in multiple clouds. Hence, the objective of this project is to find out the optimal techniques to control information leakage in multicloud. Then, I presented StoreSim which is information leakage conscious storage system and store the similar data to same cloud. I designed MinHash algorithm to efficiently generate similarity-preserving signatures for data chunks and designed a function to compute the information leakage based on these signatures which hashed by fingerprinting algorithms such as SHA-1 and MD 5. Therefore, I provided optimal information leakage in multicloud storage system which used to store users data in efficient and secure manner.

2. Literature Survey

2.1 A Review of the technique used

2.1.1 DEPSKY: Dependable and Secure Storage in a Cloud-of Clouds

Now-a-days the popularity of using multiple clouds are very rapidly growing and most of the companies, governments, and individual person are using different cloud storage providers such as Microsoft Azure, Dropbox, and Amazon S3. These cloud storage services are used to store different critical data such as company database, different higher institution database, medical database and personal information to the cloud. However, the reliability, security, and availability of the data which stored in cloud are remain concerns. According to this paper, they produce DEPSKY system which improve the security, availability and reliability of the information which moved to the cloud using data encoding, encryption and duplication on different cloud that form virtual cloud. This system is secured and dependably virtual storage that used to reduce loss of availability, loss and corruption of data, and loss of privacy through distributing encrypted dissimilar data to the different cloud. However, if the user wants to modify data which distributed over the cloud there is no similarity checking techniques and might be all data present in one cloud which produce one point of failure in cloud. [1]

Therefore, to address this problem I produced StoreSim system which is information leakage conscious storage system through storing similar data to same cloud. I designed MinHash algorithm to efficiently generate similarity-preserving signatures for data chunks and designed a function to compute the information leakage based on these signatures, which hashed in fingerprinting algorithms such as SHA-1, MD5 before modification of cloud data. Finally, I provided optimal storage system which used to store data across multiple clouds to improve the security, availability and reliability of information.

2.1.2 Scalia: An Adaptive Schemes for Efficient Multi-Cloud Storage

While increasing the amount of data which stored in cloud would grow the demands of cloud providers. There are a number of public cloud providers such as Amazon S3, Dropbox and Microsoft Azure which offers storage placement to market and the service fees among providers may change overtime to adapt customer. The customers are distributes their data over different multicloud based on access pattern of data items to prevent vendor lock-in, loss of availability and durability of data. But, choosing the best place of storage provider for our data is it needs to know access pattern of the data (i.e. rarely accessed data should be stored in low cost storage providers and a very popular data may be placed on high cost storage providers). Most of the time it is complex to know the access pattern of the data items, therefore it needs adaptive solution to choose the most cost efficient provider. However, getting a suitable cloud providers based on access pattern is not enough, because the providers may end the business, increase service fees or transfer owner to another cloud storage providers and while our file is transfer into another storage providers, then vendor lock-in will happen. [3]

According to this paper, they produced Scalia system which used to store data chunks for a long time continuously in multiple clouds based on access time which minimize storage costs. In this system the user can distribute the data chunks over different cloud storage providers which have availability, durability, no extra service payment, no single point failure, user is accessed as hosted service, and user can only pay service fees. However, many customers are distribute their data chunks to multicloud but, while user distribute their data chunks based on access pattern to the multicloud there is no techniques which checks the similarity between customer's data chunks and the original file might place in one cloud which produce one point failure.

Hence, to solve this problem I produced StoreSim which is information leakage conscious storage system using distribute similar data over same cloud and I designed MinHash algorithm to efficiently generate similarity-preserving signatures for data chunks and designed a function to compute the information leakage based on these signatures. Therefore, I produced more adaptive multicloud storage system which is more efficient, secured, high durable, less cost and efficient storage providers.

3. OVERVIEW OF THESYSTEM

3.1 Existing System

In our daily tasks, we are storing our critical data over different cloud storage providers such as Google Drive, Dropbox, and iCloud, but unplanned distribution of data chunks over multiple clouds storage providers will produce one point failure in cloud data. [7]

In the existing system, data deduplication techniques are adopted during distributing data over multiple clouds, which used to identifies the same data chunks by their fingerprints which generated by fingerprinting algorithms such as SHA-1, MD5, but any change to the data will produce a different fingerprint. However, these fingerprints can only detect whether or not the data nodes are duplicate, which is only good for exact equality testing. [1] [7] Determining the similarity of data chunks is relatively simple, but it is difficult tasks to determine efficiently due to the absence of similarity preserving signatures.

3.1.1 Disadvantages of Existing System

- ✓ Unplanned distribution of data chunks can lead to high information disclosure even while using multiple clouds.
- ✓ Redundancy of similar data during frequently modification of multiple clouds data.
- ✓ No enough security for multiple clouds data.
- ✓ There is a chance of all information to be loss.

3.2 Proposed System

To minimize the information leakage during unplanned distributions of data, I presented StoreSim system, which aware an information leakage storage system in the multicloud. It used to store similar data over the same cloud syntactically to minimize losing of information in multiple clouds. This system is achieved their goal by using novel MinHash algorithm, to efficiently generate similarity-preventing signatures for data chunks and designed a function to compute the information leakage based on these signatures.which given by fingerprint algorithms such as SHA-1,MD5. Now a user’s can store data in multiple clouds, modify each cloud data as they need and download the data from multiple clouds in secure manner.

Therefore, I developed advanced system which used to store data in secure, effective and efficient manner over centralized multicloud storage system to minimize information leakage.

3.2.1 Advantages of Proposed System

- ✓ Each data chunk which stored across multiple clouds is encrypted and no one can see without decryption.
- ✓ There is no vendor-lock-in problem and user can stored his multiple cloud data easily for along a period of time continuously.
- ✓ There are cost optimization, data consistency and availability in the proposed system;
- ✓ Reduce the chance to loss all the information in the same time.

3.3 System Modules

In this project work, I used three modules and each module has own functions, such as:

1. Data Owner(Client) module
2. Metadata Servers module
3. Cloud Service Providers module

3.3.1 Data Owner module

This module is used to pre-process the users' data for purpose of optimization such as chunking, deduplication, delta encoding and binding. The owner must register to the system and then login by his/her own username and password. After that they will do the following functionalities, such as:

- ✓ Uploading their own files from local machine to the system, then dividing files into individual chunks of a maximum size of data unit and encrypt that files after that uploading to multiple clouds.
- ✓ The data owner can modify their own data which uploaded to the multiple clouds. In this process the users' can modify both cloud data's by checking similarity between data chunks and jaccard similarity would specify where you upload the modified data.
- ✓ The owner can download the data's which stored in multiple clouds by sending request to the storage providers to get cloud keys.

3.3.2 Metadata servers module

This module is used to store information's about files, Cloud Storage Provider's, and user's, which represented the files, stored in the multiple cloud storage providers. But, before viewing this page the user must login to the page by using username and password which are manually assigned by developer of the system.

3.3.3 Cloud Service Providers module

In this module, I presented the cloud functionality. The cloud user must login to the page to views files, view requests and approve the requests which sending from other users or owner of data to get both cloud keys. I used DriveHQ cloud service which integrated to the local system to store different data chunks after chunking.

3.4 Data Set

In this project I used IEEEDataPort for data sets which is easily accessible data platform that enable users to store, search, access, and managed data. It accept all data formats and the sizes of datasets up to 2TB. Moreover users can downloading the datasets in the clouds. Hence I used a dataset which is submitted by **Subbaiah Gorala Bala** (http://ieee-dataport.org/2131, http://dx.doi.org/10.21227/dxh0-hn41, http://ieeedataport.org/documents/enterprise-master-data-hub-architecture) [8] that specify about metadata information for a business purpose. Jaccard similarity is finding the similarity of the documents textually such as the web or a collection of new articles. Therefore I examine StoreSim system by using distributing some articles information into multicloud in secure manner.

4. RESULTS

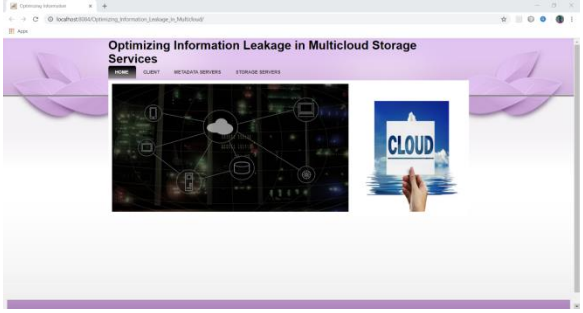
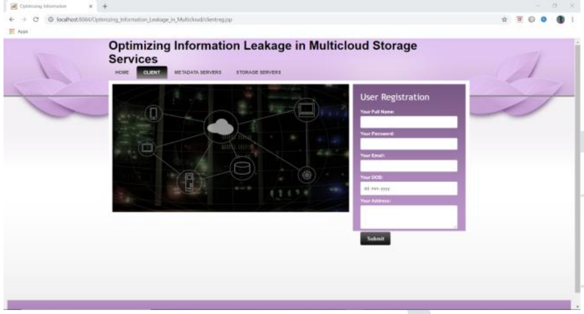

Optimizing Information Leakage in Multicloud Storage Services

Fig 4.1: **Home Page**


Optimizing Information Leakage in Multicloud Storage Services

User Registration

Your Full Name:

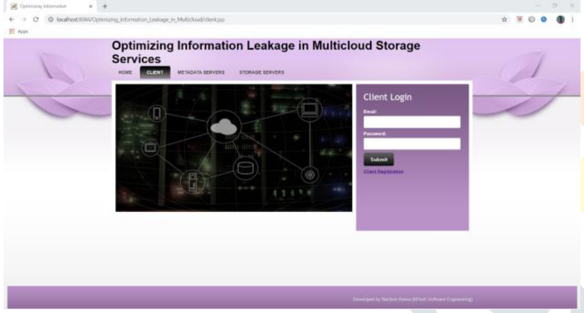
Your Password:

Your Email:

Your DOB: dd-mm-yyyy

Your Address:

Fig 4.2: Client Registration

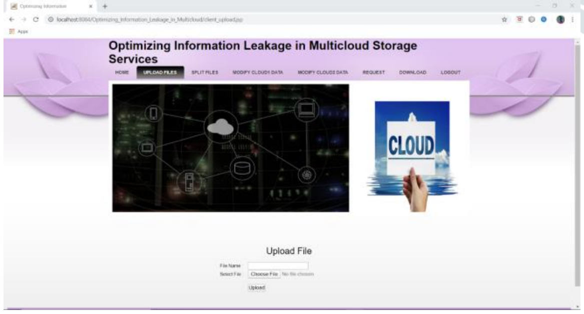

Optimizing Information Leakage in Multicloud Storage Services

Client Login

Email:

Password:

Fig 4.3: Client Login

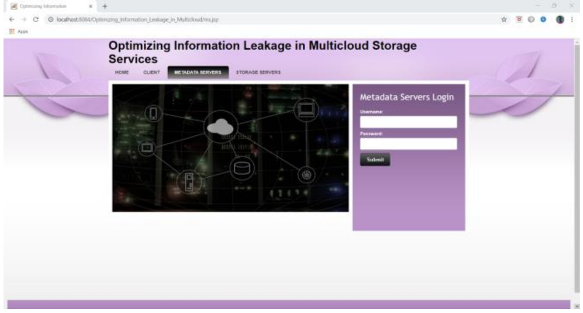

Optimizing Information Leakage in Multicloud Storage Services

Upload File

File Name

Select File Choose File No file chosen

Fig 4.4: Upload Files


Optimizing Information Leakage in Multicloud Storage Services

Metadata Servers Login

Username:

Password:

Fig 4.5: Login Metadata Servers

5. CONCLUSION

Most of the people are storing their data over the multiple clouds for security purpose. However, unplanned distribution of data over the multiple clouds may produce the probability to loss all data, which increase degree of information leakage in multicloud. To optimize information leakage, I provided StoreSim system, which is a storage system, that aware information loss in multiple clouds and storing the similar data to same cloud. I designed MinHash algorithm to efficiently generate similarity-preserving signatures for data chunks and designed a function to compute the information leakage based on these signatures which are hashed by fingerprint algorithms such as SHA-1, MD5.

Finally, I provided the optimal multicloud storage providers which used to minimize information leakage efficiently.

Future Enhancement

It is impossible to develop software which satisfies all user requirements. However this system has some future enhancement, such as:

- ✓ When cloud data is modified a user will get notification automatically through his/her Email.
- ✓ As a security technology updated, then system security will also update.
- ✓ Adding the number of clouds more than two to control information leakage in advanced way.

REFERENCES

[1] A. Bessani, M. Correia, B. Quaresma, F. Andr´e, and P. Sousa, “Depsky: dependable and secure storage in a cloud-of-clouds,” ACM Transactions onStorage (TOS), vol. 9, no. 4, p. 12, 2013.

[2] H. Chen, Y. Hu, P. Lee, and Y. Tang, “Nccloud: A network-coding-based storage system in a cloud-of-clouds,” 2013.

[3] T. G. Papaioannou, N. Bonvin, and K. Aberer, “Scalia: an adaptive scheme for efficient multi-cloud storage,” in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society Press, 2012, p. 20.

[4] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha,“Spanstore: Cost-effective geo-replicated storage spanning multiple cloud services,” in Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles. ACM, 2013, pp. 292–308.

[5] U. Manber et al., “Finding similar files in a large file system.” in UsenixWinter, vol. 94, 1994, pp. 1–10.

[6] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin, and M.Walfish,“Depot: Cloud storage with minimal trust,” ACM Transactions on Computer Systems (TOCS), vol. 29, no. 4, p. 12, 2011.

[7] J.-M. Bohli, N. Gruschka, M. Jensen, L. L. Iacono, and N. Marnau, “Security and privacy-enhancing multicloud architectures,” Dependable and Secure Computing, IEEE Transactions on, vol. 10, no. 4, pp. 212–224, 2013.

[8] subbaiah Bala, "Enterprise Master Data Hub Architecture", IEEE Dataport, 2020. [Online]. vailable: http://dx.doi.org/10.21227/dxh0-hn41. Accessed: Jun. 15, 2020.