IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 4, April 2017

An Effective Approach for Data De-duplication using Hybrid Cloud

Soumya Dath G

Assistant Professor, Department of Information Science and Engineering, GSSS Institute of Engineering and Technology for Women, Mysore, India

Abstract: Data de-duplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To protect the confidentiality of sensitive data while supporting de-duplication, the convergent encryption technique can be used to encrypt the data before outsourcing. To better protect data security, this paper is attempting to make a survey for data security methods that accounts for both security and storage efficiency. Different from traditional de-duplication systems, the differential privileges of users are considered in authorized duplicate check besides the data itself in the proposed scheme. The methods can be included in the hybrid cloud approach for effective de-duplication.

Keywords: de-duplication, authorized duplicate check, confidentiality, hybrid cloud.

I. INTRODUCTION

Cloud computing has been deployed in a variety of data storages, data centers, network communications, data managements. Researchers introduce and define cloud computing in different aspects and terms. The US National Institute of Standards and Technology defined cloud computing as a model for enabling access to a pool of resources such as servers, networks, applications, and services with low cost and minimal management. The characteristics consist of on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured pay-as-you-go services. Meanwhile the deployment models are highlighted with private cloud, public cloud, and hybrid cloud.

II. PROBLEM STATEMENT

In our system we implement a project that includes the public cloud and the private cloud and also the hybrid cloud which is a combination of the both public cloud and private cloud. In general by if we used the public cloud we can't provide the security to our private data and hence our private data will be loss. So that we have to provide the security to our data for that also provides the data de-duplication . which is used to avoid the duplicate copies of data.User can upload and Download the public cloud.for that user generates the key and storeed that key onto the private cloud. at the time of downloading user request to the private cloud for key and then access that particular file. we make a use of private cloud also. When we use a private clouds the greater security can be provided. In this system we

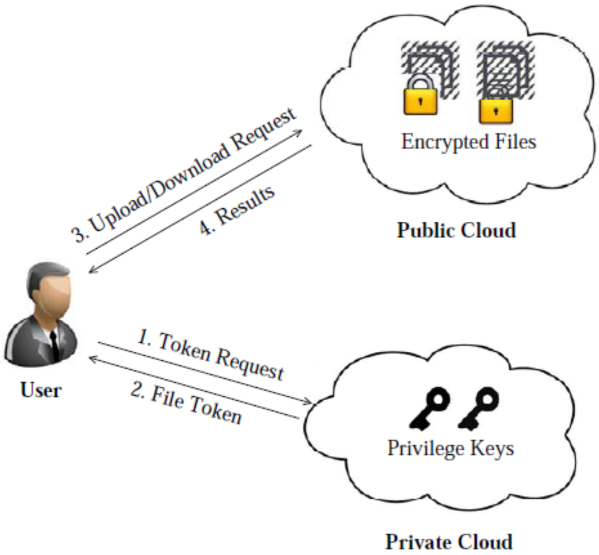


Figure 2.1 system architecture

DOI10.17148/IARJSET.2017.4409

III. RELATED WORK

Hybrid Cloud is the architecture that provides the Organization to efficiently work on both the private and public cloud architecture in combination by providing the scalability to adopt. Here some of the basic concepts and idea proposed by authors and how best and easy to adopt this environment is explained by Neal Leavitt. [1]

An intelligent workload factoring, service for organization customers which makes the best use of the present public Cloud services including their private owned data centers. It allows the organization to work between the off-premises and the on-premises infrastructure. The efficient core technology that is used for intelligent workload factoring is a fast redundant data element detection algorithm, that helps us factoring all the incoming requests based on the data content and not only on volume of data, Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena. [4] The term ―Cloud‖ has many definitions one among them is to provide infrastructure as a service system where the IT infrastructure will be deployed in the particular cloud service provider, data center as virtual machine. The growing popularity of laas will help us to transform the organization present infrastructure into the required hybrid cloud or private cloud. OpenNebula Concept is being used that will provide the features that are not present in any other cloud software, Borja Sotomayor ,Rubén S. Montero and Ignacio M. Llorente, Ian Foster. [5] Data De-duplication 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 de-duplication can be simplified, David Geer. [3] In the real world more often we tend to see the data that are two or more in database. The records which are duplicate will share the different keys that will make the duplicates matching task difficult and will result in errors. Errors will usually occur due to lack of standard formats, incomplete information or transcription errors. The through analysis of duplicate record detection literature survey is done in this paper. The duplicate detection algorithm is used which detects the duplicate records and also some of the metrics are considered that will help us to detect the similar field entry of data that is done. Multiple techniques are presented that will help us to improve the efficiency and the existing tools that is present is being covered, Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios. [6] De-duplication is the technique that is most effective most widely used but when it is applied across the multiple users the cross-user de-duplication tend to have to many serious privacy implications. Simple mechanisms can be used which can enable the cross-user de-duplication which will reduce the risks of the data leakage and also some of the security issues are discussed with how exactly to identify the files and to encrypt them while sending is discussed, Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg. [2] Data that is being collected from several data sources are being stored in the repositories called data warehouse. During the ETL (Extraction, transformation, loading) or OLTP (On Line Transaction Processing) in the data warehouse we often tend to find the duplicate copies of data in the table. Since the quality of data is very essential to gain the confident of users, more amount of money and time is being spent in obtaining the high quality data. Data cleaning is the process where the dirty data is removed. Here they have discussed some methods and strategies to remove duplicate data by, Srivatsa maddodi, Girija V. Attigeri, Dr karunakar A.k. [7].

IV. METHODS

1. User Module In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

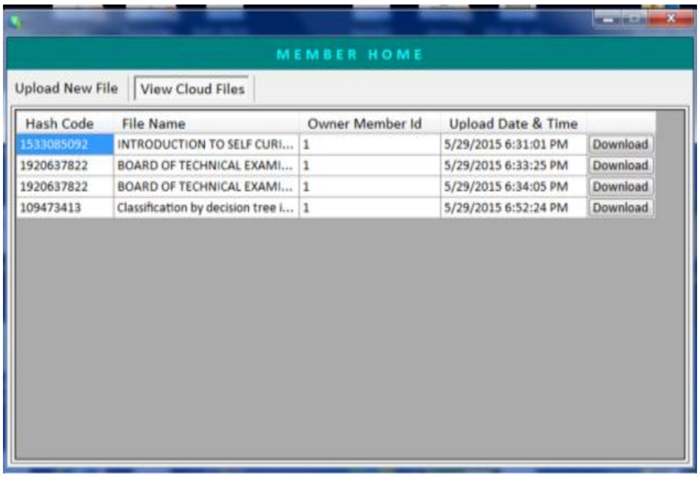
2. Secure De-duplication System To support authorized de-duplication, the tag of a file F will be determined by the file F and the privilege. To show the difference with traditional notation of tag, we call it file token instead. To support authorized access, a secret key kp will be bounded with a privilege p to generate a file token. Let ϕ′ F;p = TagGen(F, kp) denote the token of F that is only allowed to access by user with privilege p. In another word, the token ϕ′ F;p could only be computed by the users with privilege p. As a result, if a file has been uploaded by a user with a duplicate token ϕ′ F;p, then a duplicate check sent from another user .

3. Security of Duplicate Check Token We consider several types of privacy we need protect, that is, i) unforgeability of duplicate-check token: There are two types of adversaries, that is, external adversary and internal adversary. As shown below, the external adversary can be viewed as an internal adversary without any privilege. If a user has privilege p, it requires that the adversary cannot forge and output a valid duplicate token with any other privilege p′ on any file F, where p does not match p′. Furthermore, it also requires that if the adversary does not make a request of token with its own privilege from private cloud server, it cannot forge and output a valid duplicate token with p on any F that has been queried.

4. Send Key Once the key request was received, the sender can send the key or he can decline it. With this key and request id which was generated at the time of sending key request the receiver can decrypt the message.

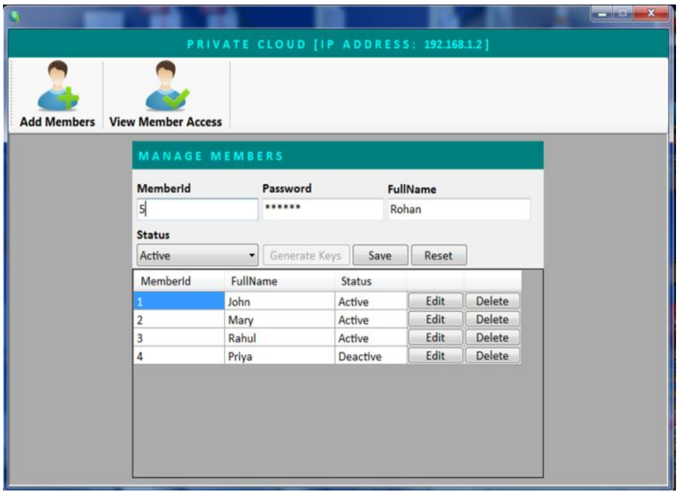
DOI10.17148/IARJSET.2017.4409

V. RESULTS AND DISCUSSION



Hash Code	File Name	Owner Member Id	Upload Date & Time	
1533085092	INTRODUCTION TO SELF CURI...	1	5/29/2015 6:31:01 PM	Download
1920637822	BOARD OF TECHNICAL EXAMI...	1	5/29/2015 6:33:25 PM	Download
1920637822	BOARD OF TECHNICAL EXAMI...	1	5/29/2015 6:34:05 PM	Download
109473413	Classification by decision tree i...	1	5/29/2015 6:52:24 PM	Download

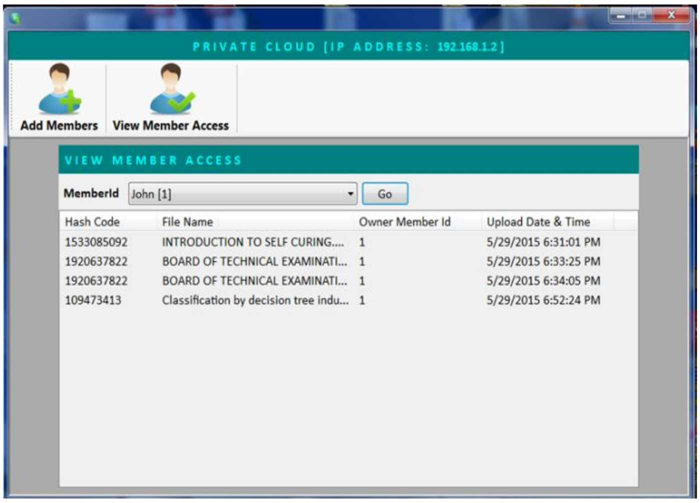
Figure 1 Files Accessed by a particular user. The user can also view the files with which the user has access. If desired, file can be downloaded.



MemberId: 5
Password: ******
FullName: Rohan
Status: Active

MemberId	FullName	Status		
1	John	Active	Edit	Delete
2	Mary	Active	Edit	Delete
3	Rahul	Active	Edit	Delete
4	Priya	Deactive	Edit	Delete

Figure 2 Adding members at the private cloud. The administrator at the private cloud server provides the user with user-name and password as credentials.

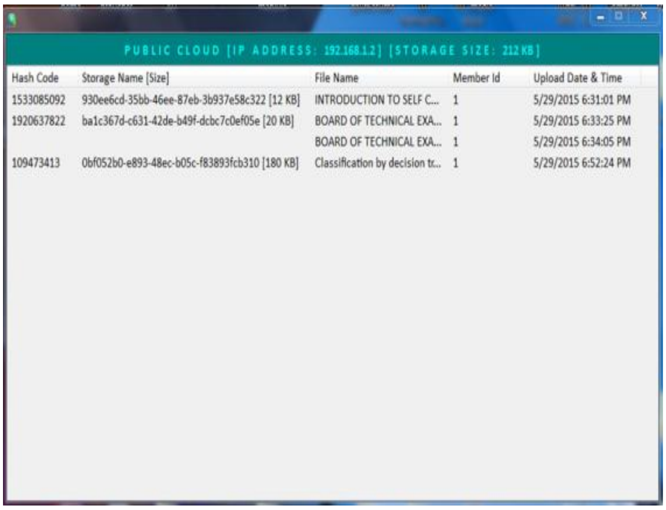


MemberId: John [1]

Hash Code	File Name	Owner Member Id	Upload Date & Time
1533085092	INTRODUCTION TO SELF CURING....	1	5/29/2015 6:31:01 PM
1920637822	BOARD OF TECHNICAL EXAMINATI...	1	5/29/2015 6:33:25 PM
1920637822	BOARD OF TECHNICAL EXAMINATI...	1	5/29/2015 6:34:05 PM
109473413	Classification by decision tree indu...	1	5/29/2015 6:52:24 PM

Figure 3 Access to the files by a user. The private cloud server can maintain the list of files the user is eligible to access

DOI10.17148/IARJSET.2017.4409



PUBLIC CLOUD [IP ADDRESS: 192.168.1.2] [STORAGE SIZE: 212 KB]

Hash Code	Storage Name [Size]	File Name	Member Id	Upload Date & Time
1533085092	930ee6cd-35bb-46ee-87eb-3b937e58c322 [12 KB]	INTRODUCTION TO SELF C...	1	5/29/2015 6:31:01 PM
1920637822	ba1c367d-c631-42de-b49f-dcbc7c0ef05e [20 KB]	BOARD OF TECHNICAL EXA...	1	5/29/2015 6:33:25 PM
		BOARD OF TECHNICAL EXA...	1	5/29/2015 6:34:05 PM
109473413	0bf052b0-e893-48ec-b05c-f83893fcb310 [180 KB]	Classification by decision tr...	1	5/29/2015 6:52:24 PM

Figure 4 Hash value used as index. The public cloud uses the hash value as the index. If the duplicate file exists, it need not be stored again and again.

VI. CONCLUSION

The well-known data de-duplication algorithms are divided into fixed-length chunking and variable- length chunking. The fixed-length chunking is very fast for processing data de-duplication but degrades the de-duplication performance. However, the variable length chunking can achieve significant data de-duplication performance with high computation overhead and longer processing time. In this paper, we suggest a dynamic chunking approach that overcomes the inherent problem of fixed-length chunking by adapting file similarity technique. The key idea of this work is to find several duplicated point by comparing hash key value and file offset within file similarity information. Several issues remain open. First, our work has limitations on supporting simple data file which has redundant data blocks with spatial locality; therefore, if the file has several modifications then overall performance will be degrade. For future work, we plan to build a massive de-duplication system with huge number of files. In this case, handling file similarity information needs more elaborated scheme

REFERENCES

[1] Neal Leavitt, "Hybrid Clouds Move to the Forefront.‖ Published by the IEEE Computer Society, MAY 2013.

[2]. Danny Harnik, Benny Pinkas, Alexandra Shulman- Peleg "Side Channels in Cloud Services De-duplication in Cloud Storage.‖ copublished by the ieee computer and reliability societies, november/ december 2010.

[3] David Geer, "Reducing the Storage Burden via Data De-duplication computer.org‖, December 2008.

[4] Hui Zhang, Guofei Jiang, Kenji Yoshihira, Haifeng Chen and Akhilesh Saxena, Intelligent Workload Factoring for A Hybrid Cloud Computing. Model, Published by the IEEE Computer Society ,2009

[5] Borja Sotomayor, Rubén S. Montero and Ignacio M. Llorente, Ian Foster, Virtual Infrastructure Management.

[6] Private and Hybrid Clouds, Published by the IEEE Computer Society, 2009.

[7] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, Duplicate Record Detection: A Survey, ieee transactions on knowledge and data engineering, vol. 19, no. 1, january 2007.

[8] Mokadem, R., Hameurlain, A.: An efficient resource discovery while minimizing maintenance overhead in sdds based hierarchical dht systems. International Journal of Grid and Distributed Computing 4(3), 1–23 (2011)

DOI10.17148/IARJSET.2017.4409