IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Authorized Data Availability with Secure Deduplication Framework using Hybrid Cloud

Ms. Rajani Sajjan[1], Ms. Gayatri Chavan[2], Dr. Vijay Ghorpade[3]

Pursuing Ph.D. in Computer Science & Engineering Shivaji University, Kolhapur[1]

Pursuing M.E in Computer Science & Engineering Department VVPIET, Solapur University[2]

Completed Ph.D. from STRM University, Nanded[3]

Abstract: Since the demand for data storage is increasing day by day and by the industry analysis we can say that digital data is increasing gradually, but the storage of redundant data is excess which results in most of the storage used unnecessary to keep identical copies. So the technology de-duplication is introduced to efficiently utilize the cloud storage system. It is one of the important techniques used for eliminating duplicate data copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. To maintain confidentiality for the sensitive data while associating deduplication, the encryption technique has been used to encrypt the data before outsourcing to the users. To enhance data security this paper makes the first attempt to formally address the problem of authorized data deduplication for which the differential privileges of users are further deliberated in duplicate check besides the data itself. Security will be analysed in terms of four aspects that is making the data securely available, authorization of duplicate check, maintaining integrity and also to make the data confidential. The usage of hybrid cloud architecture is used which supports large cloud user by efficiently storing their data in the cloud environment by using the combination of both public cloud and private application server, So that it provides the facility to store sensitive data in private application server and less critical data on to the public cloud where huge savings can be made.

Keywords: cloud computing, de-duplication, data availability,data compression, cloud service provider, private application server, public storage server.

I. INTRODUCTION

Cloud computing delivers massively scalable computing resources as service with Internet based technologies. As digital data is growing tremendously, cloud storage service is gaining popularity since they provide convenient and efficient storage service that can be accessed anytime from anywhere. Cloud computing integrates the computing storage, networking and other computing resources and leases to users; the cloud storage is designed in the form of virtualized computing environment. According to the definition by NIST [1] (National Institute of standards and Technology), "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." This cloud model is composed of four deployment models.

1) "Private cloud" used only in one organization.
2) "Community cloud" used in multiple organizations sharing concern.
3) "Public cloud "used by general public.
4) "Hybrid cloud" composed of two or more distinct deployment model.

Cloud Computing provides several means of interaction between cloud servers and users through the service layer provided in cloud architecture such as:

(i) Software as service (SaaS): This provides complete application as service.
(ii) Platform as a service (PaaS): This provides business clients with independently maintained platform for developing other application on top of it.
(iii) Infrastructure as a service (IaaS): This provides a complete environment for deploying, running and managing virtual machine and storage.

Despite the significant advantages that cloud computing has there are still many security obstacles, factors on security of cloud computing are: data confidentiality, integrity, and availability (CIA). Data confidentiality means that only authorized persons can use the data. Data integrity refers to information that has not been modified or remains untouched.

Data availability refers to use of data in time whenever needed and also to the availability of cloud service provider (CSP) on demand. Authentication refers to the process of verifying whether the incoming user is authorized or not. As cloud computing becomes prevalent, information are made available by virtualized resources to user as service across the whole Internet by hiding the platform and implementation details. Recently cloud based storage service such as Drop-box, Google drive, Apple icloud, Mozy, Microsoft SkyDrive competitively offer

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

easy to access, secure, reliable and low cost remote storage space for file-sharing, document suites and online backup services for their users. As they enable easy data access from anywhere anytime, the main quality characteristics of such services are how efficiently they can handle the large amount of network bandwidth requirement from user to cloud storage and how effectively they can reduce the storage space usages. However storage of high redundant data makes inefficient use of cloud storage resource and upload bandwidth due to which the volume of data stored in cloud increase quickly. To solve this problem, the data deduplication method is used and the objective is to improve the storage efficiency. In this proposed work data compression is focused mainly which is among the forms of data deduplication to avoid excess storage space at cloud. Data Deduplication 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 deduplication can be simplified .Data de-duplication has mainly three forms.

A. Data Compression

Data compression is a method of reducing the size of files. Data compression works within a file to identify and remove empty space that appears as repetitive patterns.

B. Single-Instance Storage

Removing multiple copies of any file is one form of the de-duplication. Single-instance storage (SIS) environments are able to detect and remove redundant copies of identical files. After a file is stored in a single-instance storage system than, all the other references to same file, will refer to the original, single copy. Single-instance storage systems compare the content of files to determine if the incoming file is identical to an existing file in the storage system.

C. Sub-file De-Duplication

Sub-file de-duplication detects redundant data within and across files as opposed to finding identical files as in SIS implementations. Using sub-file de-duplication, redundant copies of data are detected and are eliminated—even after the duplicated data exist, within separate files. This form of de-duplication discovers the unique data elements within an organization and detects when these elements are used within other files. As a result, sub-file de-duplication eliminates the storage of duplicate data across an organization.

II. LITERATURE SURVEY

Cloud storage service like dropbox and google drive offer convenient file accessibility, sharing and collaboration. These service are popular, however many enterprise have been vary to adopt them for business document because of security, privacy, ownership. Cloud storage service performs deduplication to save space by uploading each file clients conventionally encrypt their files. Message-locked encryption the most convergent encryption resolves this issue, public storage server deduplication module proposed an architecture that provides secure de-duplicated storage resisting brute-force attacks, and releases it in a system called dupless [7]. In dupless, clients encrypt under message-based keys obtained from a key-server. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf and yet achieves strong confidentiality. It shows that encryption for de-duplicated storage can achieve performance and space savings close to that of using the storage service with plaintext data. The substantial increase in security comes at a modest price in terms of performance, and small increase in storage requirements relative to the base system. Research has been focused on the combination of private cloud and cloud storage services [2].

The infrastructure of cloud storage that can hide the complexity of it management from its user [3]. In order to solve the security issue of cloud storage service, their numerous approaches in the field, to keep the security in public cloud would cost more effort to build some programming framework. [5] proposed architecture to allow users to securely store data on public cloud, while allowing for search ability through the user's encrypted data. The similar approach in private cloud architecture can be found. [4] compared private cloud storage and traditional storage model, and is compared and analysed. Feasibility of private cloud storage, presents mass based Hadoop. Client side deduplication attempts to identify deduplication opportunities already at the client and save the bandwidth of uploading copies of existing files to the server. Here it is identified that attacks that exploit client-side deduplication allow an attacker to gain access to arbitrary size files of other users based on very small hash signatures of these files. More specifically, an attacker who knows the hash signature of a file can convince the storage service that it owns that file. Hence the server lets the attacker download the entire file. To overcome such attacks [9] author introduce the notion of proofs of ownership (pows) is introduced which lets a client efficiently prove to a server that the client holds a file, rather than just a short information about it.

III. PROPOSED SYSTEM

The convergent encryption has been proposed to enforce data confidentiality while making deduplication feasible. It encrypts/decrypts a data copy with a convergent key, which is obtained by computing the cryptographic hash value of the content of the data copy. After key generation and data encryption, users retain the keys and send the cipher text to the cloud. Since the encryption operation is deterministic and is derived from the data content. Identical data copies will generate the same convergent key and hence the same cipher text. To prevent unauthorized access, a secure privilege to access protocol is also needed to provide the proof that the user indeed owns the same file when a duplicate is found; a secure

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

"privilege to access" notion is used. After the proof, subsequent users with the same file will be provided a pointer from the server without needing to upload the same file. A user can download the encrypted file with the pointer from the server, which can only be decrypted by the corresponding data owners with their convergent keys. Thus, convergent encryption allows the cloud to perform deduplication on the cipher texts and the proof of ownership prevents the unauthorized user to the file.

In such an authorized deduplication system, each user is issued a set of privileges during system initialization. Each file uploaded to the cloud is also bounded by a set of privileges to specify which kind of users is allowed to perform the duplicate check and access the files. Before submitting his duplicate check request for some file, the user needs to take this file and his own privileges as inputs. The user is able to find a duplicate for this file if and only if there is a copy of this file and a matched privilege stored in cloud. For example, in a company, many different privileges will be assigned to user. In order to save cost and efficiently management, the data will be moved to the storage - cloud service provider (S-CSP) in the public cloud with specified privileges and the deduplication technique will be applied to store only one copy of the same file. Because of privacy consideration, some files will be encrypted and allowed the duplicate check by employees with specified privileges to release the access control. As files are sensitive and needed to be fully protected against both public cloud and private. For making the data available by using cloud data service and to make the data service trustworthy only authorized user should be able to access the data. Additionally integrity and confidentiality of data should be maintained by using data deduplication.

In the proposed system, the following system objectives are to be considered:

1) To make the data available to the authorized user by eliminating duplicate copies.
2) To preserve data Integrity and confidentiality.

The implementation of project idea will be employed at various cloud computing platforms, in which the data stored in the cloud assures the user with the essentials of security aspect by implication of data integrity for the data available at the cloud and preserving confidentiality of data by using authorized users.

IV. SYSTEM DESIGN

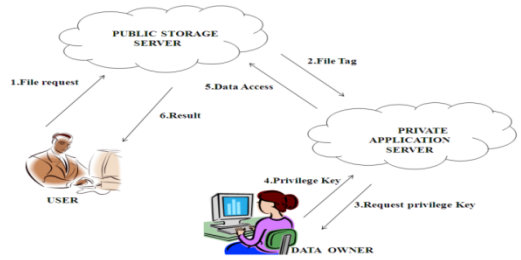

Fig 1: System Architecture

The system design includes four entities (I) public storage server (II) private application server (III) Data owner (IV) User. As shown in Figure 1, user request for the file at public storage server (PSS) here the storage server sends file request to the private application server (PAS). PAS stores the privilege key and tag of the file it request for the privilege key to access the data from the data owner, here data owner is the entity which owns the data. Data owner sends its access permission with the intermediate PAS and make the data available at PSS here the authorization of file accessibility is performed so that user can access the data.

The "PSS" is the entity that provides the deduplication and stores the data on behalf of the user authorized duplicate check is carried out at PSS, which keeps only unique data, maintains a map between existing files and associated tag with hash map. The "user" is the one which outsources their data to the public storage server and as to undergo authorization before uploading and downloading files at public storage server. "PAS" provides with secure usage of cloud services, provides execution environment and infrastructure working as interface between user and public storage server. PAS generates tag associated with its privilege's for the authorization purpose and maintains a key storage with hash map. The "data owner" is the entity which makes its data available in "PSS" so that various user can access the data by the authorized privileges

A. Key Generation

As shown in Figure 2. Key generation model, the encryption technique is used to encrypt the data before it is outsourced in the PSS by using 256- bit AES algorithm in cipher block chaining (CBC) mode. As it is concerned with security aspect user has been authorized with different privileges to further considered the duplicate check besides the data itself which as to be uploaded in the PSS.

Cipher block chaining (CBC) is a mode of operation for a block cipher(one in which a sequence of bits are encrypted as a single unit or block with a cipherkey applied to the entire block). Cipher block chaining uses what is known as an initialization vector (IV) of a certain length. One of its key characteristics is that it uses a chaining mechanism that causes the decryption of a block of cipher text to depend on all the preceding cipher text blocks.

As a result, the entire validity of all preceding blocks is contained in the immediately previous cipher text block. A single bit error in a cipher text block affects the decryption of all subsequent blocks. Rearrangement of the order of the cipher text blocks causes decryption to become corrupted. Basically, in cipher block chaining, each plaintext block is XORed (XOR) with the immediately previous cipher text block, and then encrypted. Identical cipher text blocks can only result if the same plaintext block is encrypted using both the same key and the initialization vector, and if the cipher text block order is not changed. .Ideally, the initialization vector should be

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

different for any two messages encrypted with the same key.

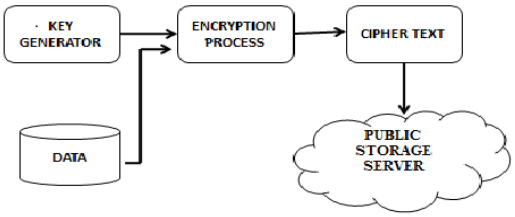

Fig 2: Key Generation Model

B. Tag Generation

With reference to Figure 3 Tag generation model, the user derives the encryption key from each original data copy and encrypt the data copy, during which tag is also been derived using SHA-1 algorithm from the encrypted data.SHA-2 (Secure Hash Algorithm 2) is a set of cryptographic hash functions SHA stands for Secure Hash Algorithm. Cryptographic hash functions are mathematical operations run on digital data; by comparing the computed "hash" (the output from execution of the algorithm) to a known and expected hash value, a person can determine the data's integrity. For example, computing the hash of a downloaded file and comparing the result to a previously published hash result can show whether the download has been modified or tampered with. a key aspect of cryptographic hash functions is their collision resistance: nobody should be able to find two different input values that result in the same hash output.The derived tag will be used to detect duplicates generated by PAS. User computes and sends duplicate check tag to the PAS for authorized duplicate check. Every tag holds the correctness property, firstly user sends the tag to the PAS to check if the identical copies have been already stored, and here in this the encryption key and tag are independently derived. The notion“ privilege to access “ and “ identity check ” protocol is proposed as an interactive algorithm that enables user to prove their ownership of data copies to the PSS and identity check is used to verify whether the accessibility of the particular client is accepted or rejected.

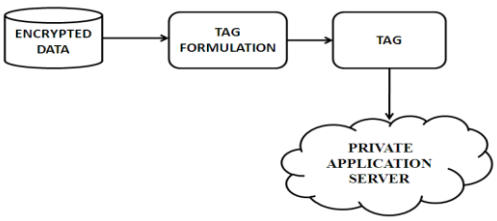

Fig 3: Tag Generation Model

V. RESULT ANALYSIS

As described in the above sections the generation of duplicate copies at cloud space is detected an avoided by generating keys and tag depending upon on the accessibility and authorization of users. The results are been analysed on the basis of file encryption time and space utilized by specific file before and after compression focusing the reduced space of file at cloud space after compression.

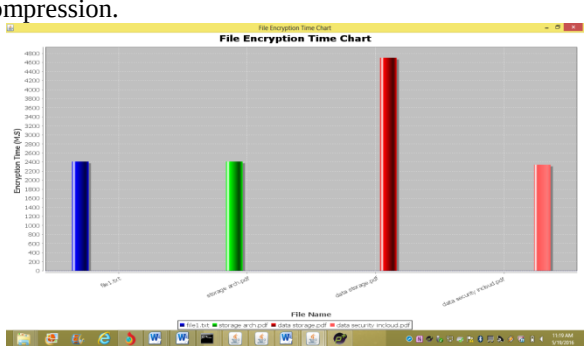

Fig 4 File Encryption Time Chart

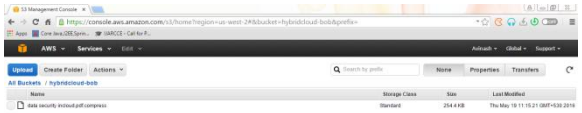

Fig 5: Compressed file at amazon Cloud

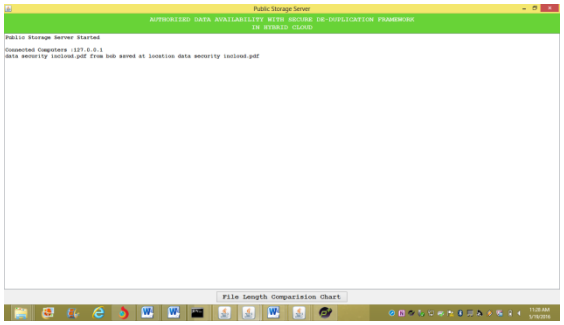

Fig 6; Public storage server

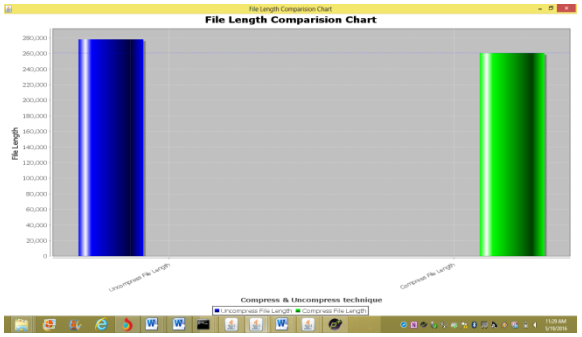

Fig 7: File before and after compression

VI. CONCLUSION

In this paper the concept of authorized data deduplication is anticipated to protect the data by including distinct privileges of users and deduplication technique which avoids, detect duplicate file and reduce the storage space supporting authorized duplicate check in public storage server. In which the duplicate-check tag of files are

DOI 10.17148/IJARCCE.2016.5590

generated and stored by the private application server with private keys. As per the basic security model this scheme is secured in terms of insider and outsider attacks. As an evidence of concept, a design of proposed authorized duplicate check scheme will be consider at the time of implementation and accompanied by stabilizing the security for the data stored in public storage server. Authorized duplicate check scheme gains minimal overhead and also overcome the wastage of unnecessary storage of repetitive data in public storage space by compressing file.3

REFERENCES

[1]. NIST Cloud Computing Standards Roadmap Working Group NIST Cloud Computing Program Information Technology Laboratory.
[2]. J. Deng, J. Hu, A.C.M, Liu and J.Wu , "Research and application of cloud storage" Proc. 2nd International workshop and Intelligent System and Application (ISA 10),IEEE Press ,May 2013 , pp-1-5.
[3]. J. Wu, L.Ping, X. Ge, Y.Wang and J.Fu, "Cloud storage as the Infrastructure of Cloud Computing" Proc. International Conference on Intelligent Computing and Cognitive Informatics(ICICCI 10),IEEE Press, June 2013,pp,380-383
[4]. D.Zhang, F.Sun,X.Cheng, and C.Liu, "Researcher and Hadoop – based enterprise file cloud storage system" Proc.3rd international conference on Awareness Science and Technology (Icast 11),IEEE Press, Sept 2014,pp380-3833
[5]. Koletka and A, Hutchison "Architecture for secure searchable cloud storage" Proc. International Conference on Informatics Security South Africa (ISSA 11),IEEE Press Aug, 2014, pp.1-7
[6]. L. Hao and D. Han, "The study and design on secure –cloud storage system" Proc. First International Conference on Electrical and Control Engineering (ICECE 11),IEEE Press ,Sept 2011 ,pp.5126-5129
[7]. M. Bellare, S. Keelveedhi, and T. Ristenpart. Dupless: Server aided encryption for deduplicated storage. In USENIX Security Symposium 2014.
[8]. M. Bellare, C. Namprempre and G.Neven. Security proofs for identity-based identification and signature schemes. J. Cryptology, 22(1):1–61, 2009..
[9]. M. Bellare and A. Palacio. Gq and schnor identification schemes: Proofs of security against impersonation under active and concurrent attacks. In CRYPTO, pages 162–177, 2002.
[10]. Jin Li, Yan Kit Li, XiaofengChen, PtrickP.C.Lee, Wenjing Lou. "A hybrid cloud approach for secure authorized deduplication" IEEE Transaction on Parallel and Distributed System ,May 2015.
[11]. J. Yuan and S. Yu. Secure and constant cost public cloud storage auditing with deduplication .IACR Cryptology ePrint Archive, 2013
[12]. Neal Leavitt, "Hybrid Clouds Move to the Forefront.‖ Published by the IEEEComputerSociety,MAY2013.
[13]. B. Mao, H. Jiang, S. Wu, and L. Tian. POD: Performance Oriented I/O Deduplication for Primary Storage Systems in the Cloud. In IPDPS'14, May 2014.
[14]. T. Matsumoto, T. Seito, A. Kamoshita, T. Shingai and A. Sato, "High-Speed Secret Sharing System for Secure Data Storage Service" SCIS 2012 The 29th Symposium on Cryptography and Information Security
[15]. Amazon EC2 SLA, http://aws.amazon.com/ec2-sla/
[16]. S.Rance, Defining Availability in the Real World, Hewlett Packard, 2013.

BIOGRAPHIES

Ms. Rajani S. Sajjan completed B.E. (Computer Science & Engineering) from Walchand College of Engineering from Sangli in 1999. Completed M.Tech(Computer Science & Engineering)from PDA College of Engineering, Gulbarga. Pursuing Ph.D. in Computer Science & Engineering from Shivaji University, Kolhapur.

Ms. Gayatri .S.Chavan completed B.E. (Computer Science &Engineering) from Solapur University in 2012 and pursuing M.E. (Computer Science & Engineering) from Solapur University.

Dr. Vijay R. Ghorpade completed Ph.D. from STRM University, Nanded. Specialized in Mobile Ad-hoc networks, Data Mining & Cloud Computing.