ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

SECURE DATA DEDUPLICATION IN HYBRID APPROACH

Krishnapraseeda V
Department of Computer Science And Engineering
United Institute of Technology,Coimbatore,TN,India
praseedasuren@gmail.com

ABSTRACT. The cloud computing technology develops during the past decade; outsourcing data to cloud service for storage becomes an attractive trend, which benefits in sparing efforts on heavy data maintenance and management. Also the security for the cloud is still a big task. However, since the outsourced cloud storage is not fully trustworthy, it raises security concerns on how to realize data deduplication in cloud while achieving integrity auditing. In this work, the study was done about the problem of integrity auditing and secure deduplication on cloud data. Specifically, aiming at achieving both data integrity and deduplication in cloud, the propose system consist of two secure systems, namely SecCloud and SecCloud+ also the process of sending and receiving in case of secure by means of same file content with same file name. SecCloud introduces an auditing entity with maintenance of a MapReduce cloud, which helps clients generate data tags before uploading as well as audit the integrity of data having been stored in cloud. Compared with previous work, the computation by user in SecCloud is greatly reduced during the file uploading and auditing phases. SecCloud+ is designed motivated by the fact that customers always want to encrypt their data before uploading, and enables integrity auditing and secure deduplication on encrypted data.

Index terms : Cloud Computing, Deduplication, Integrity, SecCloud, MapReduce, Auditing.

I. INTRODUCTION

Cloud computing is the term used to share the resources globally with less cost .we can also called as "IT ON DEMAND". It provides three types of services i.e., Infrastructure as a service (IaaS), Platform as a service (PaaS) and Software as a service (SaaS)[1],[4],[2]. The ever cheaper and more powerful processors, together with the software as a service (SaaS) computing architecture, are transforming data centres into pools of computing service on a huge scale. End users access the cloud based applications through the web browsers with internet connection. Moving data to clouds makes more convenient and reduce to manage hardware complexities[2],[7]. Data stored at clouds are maintained by Cloud service providers (CSP) with various incentives for different levels of services.

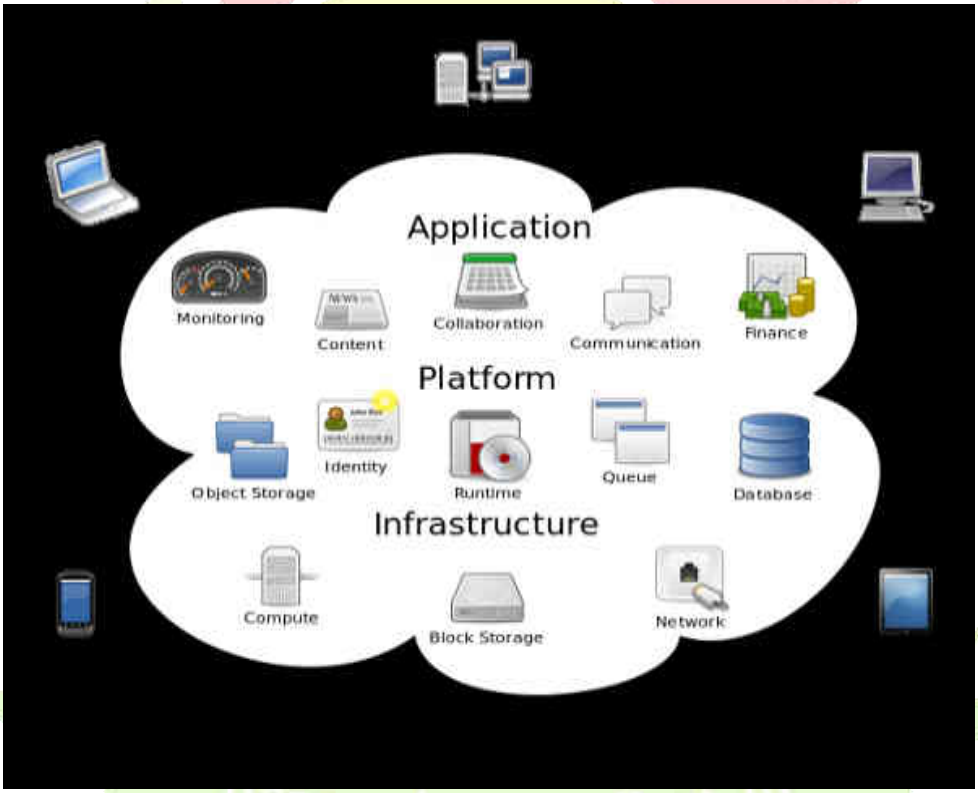
ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

Architecture of Cloud Computing

End users access the cloud based applications through the web browsers with internet connection. Moving data to clouds makes more convenient and reduce to manage hardware complexities. Data stored at clouds are maintained by Cloud service providers (CSP) with various incentives for different levels of services. However it eliminates the responsibility of local machines to maintain data, there is a chance to lose data or it effects from external or internal attacks.



To maintain the data integrity and data availability many people proposed several algorithms and methods that enable on demand data correctness and verification. So Cloud servers are not only used to store data like a ware house , it also provides frequent updates on data by the users with different operations like insert, delete , update and append.

Deduplication

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage[5],[6]. 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.

Data deduplication is a technique for reducing the amount of storage space an organization needs to save its data. In most organizations, the storage systems contain duplicate copies of many pieces of data. For example, the same file may be saved in several different places by different users, or two or more files that aren't identical may still include much of the same data. Deduplication eliminates these extra copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the original copy. Companies frequently use deduplication in backup and disaster recovery applications, but it can be used to free up space in primary storage as well.

Related Work. Considering the process of Deduplication, Data deduplication 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[5],[8]. To protect the confidentiality of sensitive data while supporting deduplication, Cloud computing provides seemingly unlimited "virtualized" resources to users as services across the whole Internet, while hiding platform and implementation details[2],[7],[4] and [3]. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively low costs. As cloud computing becomes prevalent, an increasing amount of data is being stored in the cloud and shared by users with specified *privileges*, which define the access rights of the stored data.

SecCloud introduces an auditing entity with maintenance of a Map Reduce cloud, which helps clients generate data tags before uploading as well as audit the integrity of data having been stored in cloud[3],[6].

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

Disadvantages :-

One critical challenge of cloud storage services is the management of the ever-increasing volume of data. Whenever data is transformed, concerns arise about potential loss of data. By definition, data deduplication systems store data differently from how it was written.

Our Contribution. In this paper, we solve this open problem and propose SecCloud+, which allows for integrity auditing and deduplication on encrypted files. Cloud Clients have large data files to be stored and rely on the cloud for data maintenance and computation. They can be either individual consumers or commercial organizations. Cloud Servers virtualize the resources according to the requirements of clients and expose them as storage pools. Typically, the cloud clients may buy or lease storage capacity from cloud servers, and store their individual data in these bought or rented spaces for future utilization. Auditor which helps clients upload and audit their outsourced data maintains a MapReduce cloud and acts like a certificate authority. This assumption presumes that the auditor is associated with a pair of public and private keys. Its public key is made available to the other entities in the system.

Advantages :-

- It provides the Integrity auditing by Clustering the files with removing the duplicate files.
- The duplicate files are mapped with a single copy of the file by mapping with the existing file in the cloud.
- Integrity Auditing: The first design goal of this work is to provide the capability of verifying correctness of the remotely stored data.

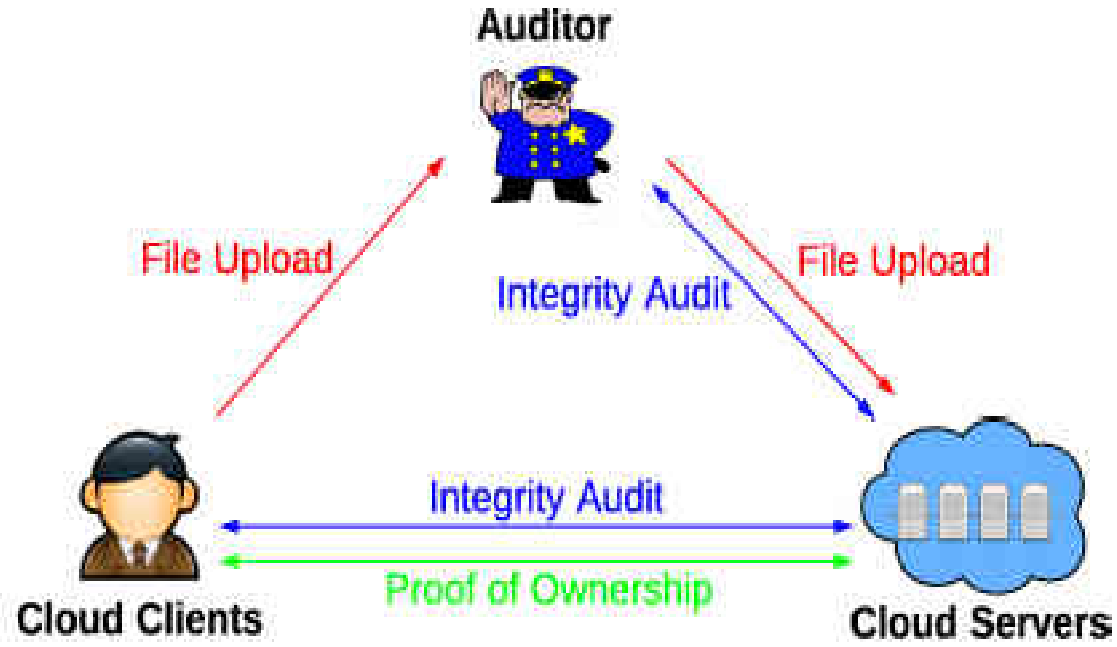
Convergent encryption:

Convergent encryption [4], [8] provides data confidentiality in deduplication. A user (or data owner) derives a convergent key from each original data copy and encrypts the data copy with the convergent key.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016



In addition, the user also derives a tag for the data copy, such that the tag will be used to detect duplicates. Here, we assume that the tag correctness property [4] holds, i.e., if two data copies are the same, then their tags are the same. To detect duplicates, the user first sends the tag to the server side to check if the identical copy has been already stored. Note that both the convergent key and the tag are independently derived, and the tag cannot be used to deduce the convergent key and compromise data confidentiality. Both the encrypted data copy and its corresponding tag will be stored on the server side.

Proof of ownership:

The notion of proof of ownership [8] enables users to prove their ownership of data copies to the storage server. Specifically, PoW is implemented as an interactive algorithm (denoted by PoW) run by a prover (i.e., user) and a verifier (i.e., storage server). The verifier derives a short value fðMÞ from a data copy M.

II. MODELS AND GOALS

A. SYSTEM MODEL

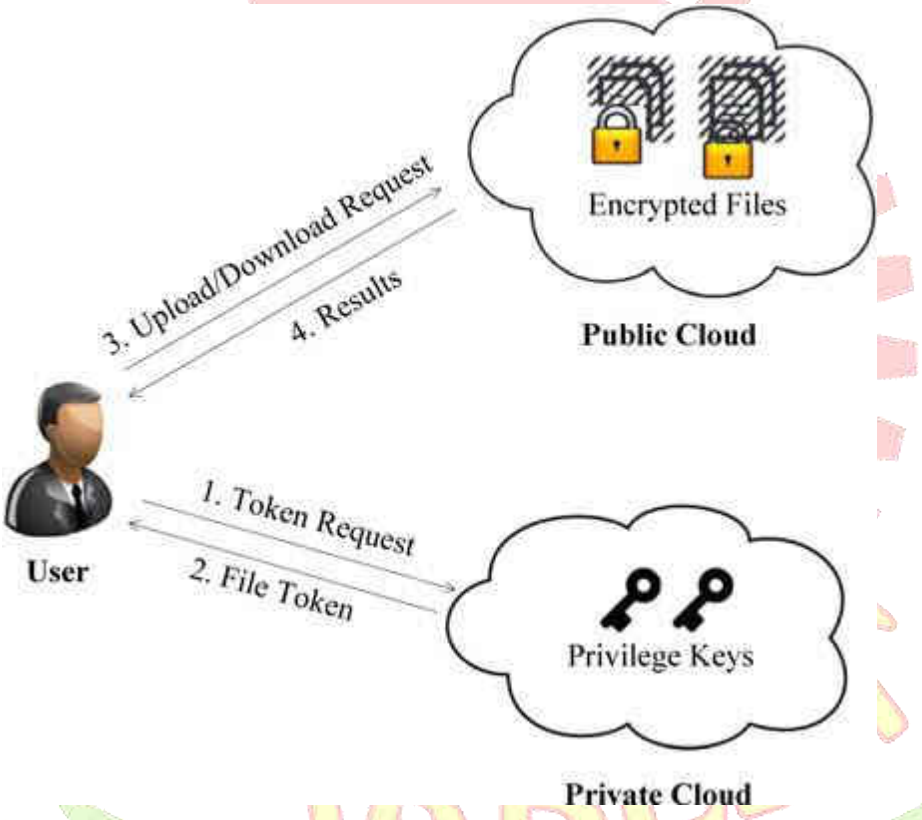
At a high level, our setting of interest is an enterprise network, consisting of a group of affiliated clients (for example, employees of a company) who will use the S-CSP and store data with deduplication technique. In this setting, deduplication can be frequently

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

used in these settings for data backup and disaster recovery applications while greatly reducing storage space. Such systems are widespread and are often more suitable to user file backup and synchronization applications than richer storage abstractions. There are three entities defined in our system, that is, users, private cloud and S-CSP in public cloud as shown in Fig. 1. The S-CSP performs deduplication by checking if the contents of two files are the same and stores only one of them.



The access right to a file is defined based on a set of privileges. The exact definition of a privilege varies across applications. For example, we may define a role-based privilege [7], [1] according to job positions (e.g., Director, Project Lead, and Engineer), or we may define a time-based privilege that specifies a valid time period (e.g., 2016-01-01 to 2016-01-31) within which a file can be accessed.

Authorized Deduplication

In this paper, we will only consider the file-level deduplication for simplicity. In another word, we refer a data copy to be a whole file and file-level deduplication which eliminates the storage of any redundant files. Actually, block-level deduplication can be easily deduced from file-level deduplication, which is similar to [5]. Specifically, to

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

upload a file, a user first performs the file-level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well; otherwise, the user further performs the block-level duplicate check and identifies the unique blocks to be uploaded. Each data copy (i.e., a file or a block) is associated with a token for the duplicate check.

B. DESIGN GOALS

In this paper, we address the problem of privacy-preserving deduplication in cloud computing and propose a new deduplication system supporting for

_ Differential authorization. Each authorized user is able to get his/her individual token of his file to perform duplicate check based on his privileges. Under this assumption, any user cannot generate a token for duplicate check out of his privileges or without the aid from the private cloud server.

_ Authorized duplicate check. Authorized user is able to use his/her individual private keys to generate query for certain file and the privileges he/she owned with the help of private cloud, while the public cloud performs duplicate check directly and tells the user if there is any duplicate. The security requirements considered in this paper lie in two folds, including the security of file token and security of data files. For the security of file token, two aspects are defined as unforgeability and indistinguishability of file token.

III. OUR CONSTRUCTION

We implement a prototype of the proposed authorized deduplication system, in which we model three entities as separate C++ programs. A Client program is used to model the data users to carry out the file upload process. A Private Server program is used to model the private cloud which manages the private keys and handles the file token computation. A Storage Server program is used to model the S-CSP which stores and deduplicates files. We implement cryptographic operations of hashing and encryption with the OpenSSL library

1. User:

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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 DeDuplication System:

To support authorized deduplication, 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 $\phi' F;p = \text{TagGen}(F, kp)$ denote the token of F that is only allowed to access by user with privilege p. In another word, the token $\phi' F;p$ 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 $\phi' F;p$, then a duplicate check sent from another user will be successful if and only if he also has the file F and privilege p. Such a token generation function could be easily implemented as $H(F, kp)$, where $H(\_)$ denotes a cryptographic hash function.

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.

IV. PERFORMANCE EVALUATION

We conduct testbed evaluation on our prototype. Our evaluation focuses on comparing the overhead induced by authorization steps, including file token generation

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

and share token generation, against the convergent encryption and file upload steps. We break down the upload process into six steps, 1) Tagging, 2) Token Generation, 3) Duplicate Check, 4) Share Token Generation, 5) Encryption, 6) Transfer. For each step, we record the start and end time of it and therefore obtain the breakdown of the total time spent.

V. CONCLUSION

To achieve both data integrity and deduplication in cloud, the proposed concept has done a literature survey in this phase. Also an auditing entity with maintenance of a MapReduce cloud, which helps the clients generate data tags before uploading as well as audit the integrity of data having been stored in cloud. In addition, the proposed enables the secure deduplication through introducing a Proof of Ownership protocol and preventing the leakage of side channel information in data deduplication. Compared with previous work, the computation by user is greatly reduced during the file uploading and auditing phases and also an advanced construction motivated by the fact that customers always want to encrypt their data before uploading, and allows for integrity auditing and secure deduplication directly on encrypted data.

VI. REFERENCES

[1] Jin Li, Yan Kit Li, Xiaofeng Chen, Patrick P.C. Lee, and Wenjing Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 26, NO. 5, MAY 2015.

[2] OpenSSL Project, (1998). [Online]. Available: http://www.openssl.org/

[3] P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. 24th Int. Conf. Large Installation Syst. Admin., 2010, pp. 29–40.

[4] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Sec. Symp., 2013, pp. 179–94.

[5] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. 32nd Annu. Int. Conf. Theory Appl. Cryptographic Techn., 2013, pp. 296–312.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

[6] M. Bellare, C. Namprempre, and G. Neven, "Security proofs for identity-based identification and signature schemes," J. Cryptol., vol. 22, no. 1, pp. 1–61, 2009.

[7] M. Bellare and A. Palacio, "Gq and schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks," in Proc. 22nd Annu. Int. Cryptol. Conf. Adv. Cryptol., 2002, pp. 162–177.

[8] S. Bugiel, S. Nurnberger, A. Sadeghi, and T. Schneider, "Twin clouds: An architecture for secure cloud computing," in Proc. Workshop Cryptography Security Clouds, 2011, pp. 32–44.