INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1439 - 1445
RECEIVED : 24/11/2016. PUBLISHED: 06/12/2016
December 6, 2016

MAXIMIZE THE STORAGE SPACE IN CLOUD ENVIRONMENT USING DEDUPLICATION

[1]M.Gokila, [2]N.Radha,
[1]PG Scholar, Computer science and engineering, Mahendra Engineering College,
[2]Assitant professor, Dept of computer science and engineering, Mahendra Engineering college.

Abstract

Data deduplication is a technique for eliminating redundant copies of data, and has been widely used in cloud storage to reduce storage space and upload bandwidth. However, there is only one copy for each file stored in cloud even if such a file is owned by a huge number of users. Furthermore, the challenge of privacy for sensitive data also arises when they are outsourced by users to cloud. Aiming to address the above security challenges, this project makes the first attempt to formalize the notion of distributed reliable deduplication system. It proposed new distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems.

Index terms: Deduplication, Cloud computing, File level and Trusted third party.

1. INTRODUCTION

The cloud storage arrivals motivate enterprises and organizations to outsource their data storage to third-party cloud providers, as evidenced by many real-life case studies [10]. The management of the ever growing volume of data is the important challenge of today's cloud storage services. Data de-duplication is a precise data compression technique in cloud computing environment. It eliminating duplicate or redundant copies of repeating data to backup data and minimize network and storage overhead. It is not only used to develop storage utilization and can also be useful to network data transfers to trim down the number of bytes that must be sent. As a substitute of maintaining multiple data copies with the similar content, it eliminates duplicate data by keeping only one physical copy and referring other duplicate data to that copy [7]. De-duplication has received more concern from both academia and industry because it can considerably improves storage utilization and save storage space, particularly for the applications with high de-duplication percentage such as archival storage space. Different types of de-duplication systems have been proposed based on various types of de-duplication approach such as server side or client side de-duplications, block-level or file-level de-duplications [11]. In particular, with the introduction of cloud storage data de-duplication techniques become more attracting and important for the management of greater volumes of data in cloud storage services which motivates enterprises and organizations to outsource data storage to third-party cloud providers, as proof by many real-life case studies. Today's commercial cloud storage services, such as Google Drive, Mozy and Dropbox have been applying de-duplication to save the network bandwidth and the storage cost with client-side de-duplication [1]. Though deduplication technique can save the storage space for the cloud service providers, it decreases the reliability of the system [2]. Most of the previous deduplicationsystems have only been considered carefully in a single-server setting. As lots of deduplication systems and cloud storage systems are anticipated by users and applications for

higher reliability, especially in archival storage systems where data are critical and should be preserved over long time periods.

Cloud environment, the providers offer their services to several fundamental models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Network as a Service (Naas) and Software as a Service (SaaS). *Infrastructure as a service:* the providers of IaaS offer computers – physical or virtual machines and other resources [12]. *Platform as a service:* In this model, cloud service providers distribute a computing platform, typically including operating system, database, execution environment of programming languages and web server. Application developers can develop and run their software resolutions on a cloud platform without the complexity and price of buying and managing the underlying hardware and software layers [12]. *Software as a service (SaaS):* It is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. This provider generally price applications using a subscription fee [13]. *Network as a service (Naas):* It involves the optimization of resource allocations by considering network and computing resources as a unified whole. Naas concept materialization also includes the provision of a virtual network service by the network infrastructure owners to a third party [12].

2. RELATED WORK

John R. Douceur, William and Atul Adya, they present a mechanism in [4] to reclaim space from this incidental duplication to make it accessible for controlled file replication. Their mechanism includes convergent encryption, which enables replica files to coalesce into the space of a single file, even if the files are encrypted with different users' key, and SALAD [4], a self arranging and associative Database for aggregating file content and location information in a decentralized, scalable, fault-tolerant manner. Relocating the replicas of files with identical content to a common set of storage machines. T. Ristenpart, provides an architecture in [3] which provides secure deduplicated storage counter brute-force attacks, and realizes it in a system called DupLESS. In this, clients encrypt underneath message based keys received from a key server via a PRF protocol [3]. It enables clients to store encrypted data with an existing service, have the service perform deduplication on behalf of them, and still achieves strong confidentiality guarantees. Several works have looked at the common problem of enterprise network security, but none provide solutions that meet all requirements from the above threat model.

M. Bellare and S. Keelveedhi describe a concept of MLE in [5]. Message-Locked Encryption (MLE) provides a way to get secure deduplication, a goal presently targeted by numerous cloud-storage providers. We provide definitions both for a form of integrity and for privacy that we call tag consistency. Based on this foundation, we make both practical and theoretical contributions. On the realistic side, we present ROM security analyses of a natural family of MLE schemes that includes deployed schemes [5].However, in attempting to build MLE from these primitives, several problems arise. A related problem is that it is not clear how an MLE scheme might decrypt.

Nikhil O. Agrawal, S.Kulkarni and Prof.S describe about secure deduplication in [10]. It can eliminate replica copies of storage data and limit the harm of stolen data if we reduce the value of that stolen information to the attacker. This paper addresses the problem of achieving reliable and efficient key management in secure deduplication. They first introduce a baseline approach for the users, which holds an independent master key for encrypting the convergent keys and outsourcing them. Masquerade attacks such as identity theft and fraud are a serious computer security problem.

Paul Anderson and Le Zhang developed an algorithm in [6] which takes advantage of the data. It is common between users to raise the speed of backups, and decrease the storage requirements. This algorithm supports client-end per-user encryption which is essential for confidential personal data. It also supports a unique feature which allows immediate detection of common sub trees; it has to query the backup system for every file.This algorithm does have some disadvantages. In particular, a change to any node implies a change to all of the ancestor nodes up to the root.

Zooko Wilcox-O'Hearn andBrian Warner discus a concept of Tahoe [8], it is a system for secure, distributed storage. It uses ability for access control, cryptography for confidentiality and integrity [8], and erasure coding for fault-tolerance. It has been arranged in a commercial backup service and is currently operational. The implementation is Open Source. If the integrity test out fails, the client necessarily needs to know which erasure code share or shares were wrong, so that it can reconstruct the file from other shares. If the integrity check applied only to the cipher text, then the client wouldn't know which share or shares to replace.

Henry C. H. Chen, Yang Tang, they presents FadeVersion in [9], it is a secure cloud backup system that serves as a security layer with an advantage of today's cloud storage services. It follows the standard version-controlled backup design, which reduces the storage of redundant data across different versions [9] of backups. On top of this, FadeVersion submit an application cryptographic protection to data backups. Particularly, it enables fine-grained assured deletion; cloud clients can assuredly delete specific backup versions or files on the cloud and make them permanently inaccessible to anyone. Deleting an old version may make the future versions unrecoverable.

3. PROBLEM FORMULATION

In this work, we refer a data copy to be a whole file and this leads to file-level deduplication, which eliminates the storage of any redundant files. We deploy our deduplication mechanism in file levels. Specifically, to upload a file, a user needs to perform the file level duplicate check. If the file is a duplicate, then all its blocks must be duplicates as well.

User: A user is an entity that wants to outsource data storage to the S-CSP and access the data later. It is mainly used to save the upload bandwidth; the user simply uploads unique data but does not upload any duplicate data, it may or may not be owned by the same user or different users.*Trusted Third party:* The third party gives the data outsourcing service and stores data on behalf of the users. To reduce the storage cost, the S-CSP eliminates the storage of duolicate data via deduplication and keeps only unique data.

In previous work users are keeping multiple data copies with the same content, it leads to data redundancy. In existing work file is eliminated base on the file name, the content of the particular file is not verified. In the proposed work, instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Proposed work describes the concept of file-level deduplication, which discovers redundancies between different files and removes these redundancies to reduce capacity demands.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1439 - 1445
RECEIVED : 24/11/2016. PUBLISHED: 06/12/2016
December 6, 2016

4. SYSTEM MODEL

Owner Registration

In this module an owner has a permission to upload his/her files in a cloud server; he/she should register first. Then only he/she is able to do registration. For registration he/she needs to fill the details in the registration form. These details have been maintained in a database. In this module, any of the above mentioned person have to login, they must login by their user-id and password.

User Registration and User Login

This module describes the concept user registration, if a user requests to access the data which is stored in a cloud; he/she must register their details first. These details have been maintained in a Database. If the user authentication is verified, he/she can download the file by using file id which has been stored by data owner when it was uploading. Owner can allow access or refuse access for accessing the data. So users able to access his/her account by the corresponding data owner. If owner does not permit the user, the user can't get the data.

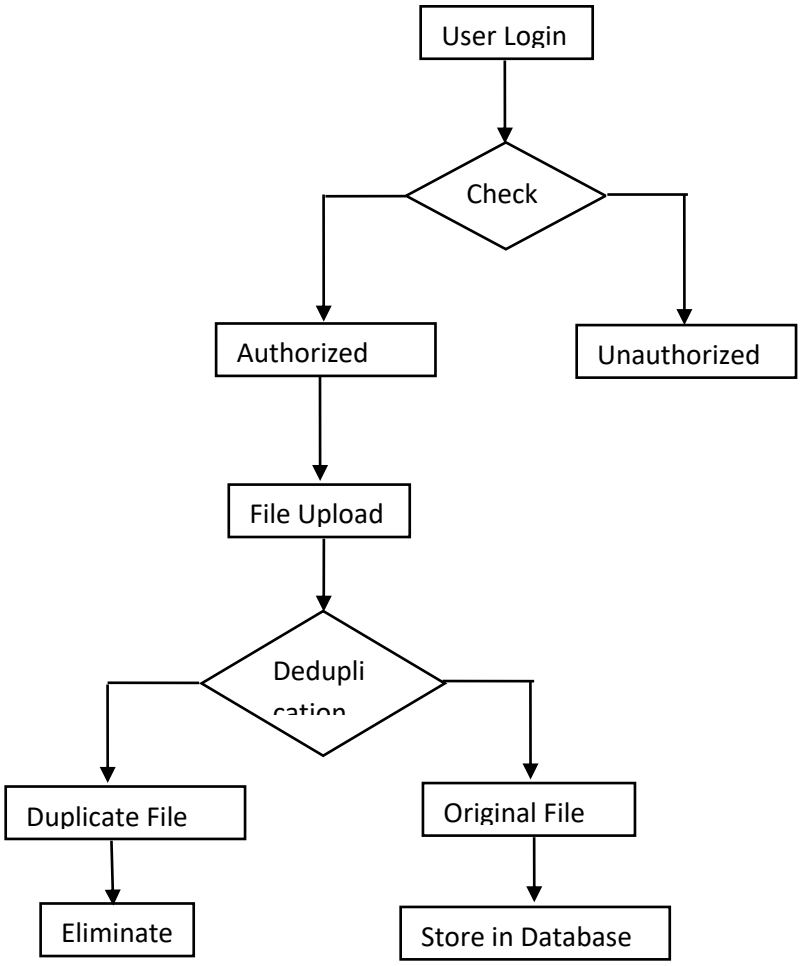

Fig 1.1 Flow diagram

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 6 (2016) PAGES 1439 - 1445
RECEIVED : 24/11/2016. PUBLISHED: 06/12/2016
December 6, 2016

Cloud Service Provider

This module supports the development of Cloud Service Provider. This is an entity that provides a data storage service in public cloud. It provides the data outsourcing service and stores data on behalf of the users. To reduce the storage cost it eliminates the storage of redundant data via deduplication and keeps only unique data.

Data Users Module

A user is an entity which requests to outsource data storage to the S-CSP and access the data later. In a cloud storage system presenting deduplication, the user can only upload unique data but could not upload any redundant data to save the bandwidth of uploaded data, which may or may not be owned by the same user or different users. In the authorized deduplication system, each user has been issued a set of permission in the setup of the system.

File Upload

Owner can upload the file into database along with metadata, with the help of this metadata and its contents; the end user has the permission to download the file. In the cloud uploaded file was in encrypted form, only registered user can decrypt the data.

Fig: 1.2 File Uploading

Trusted Third Party (TTP) Login

After the completion of logging procedure cloud provider gets the permission. Cloud provider can view the files uploaded by their authorized clients. Also upload this file into separate Cloud Database. In this module TTP has monitors the data owners file with help of verifying the file of data owner file and stored the files in a database. The TTP checks the cloud service provider (CSP), and find out whether the CSP is authorized one or not. In this module, if a cloud service provider or maintainer of cloud wants to do some cloud offer, they should register first. The authorized users can upload/ download the file from cloud database.

5. CONCLUSION

In this paper we develop a file level deduplication technique, which reduce the bandwith and storage space in cloud. It is a is a efficient data compression technique in cloud computing environment. It is not only used to develop storage utilization and can also be useful to network data transfers to trim down the number of bytes that must be sent. The security requirements of data confidentiality and tag consistency are also achieved.

REFERENCES

1. Jin Li, Xiaofeng Chen, Xinyi Huang, Shaohua Tang, Yang Xiang, "Secure Distributed Deduplication Systems with Improved Reliability" IEEE TRANSACTIONS ON COMPUTERS, VOL. 64, NO. 12, DECEMBER 2015

2. S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in Proc. ACM Conf. Comput. Commun. Secur., 2011, pp. 491–500.

3. M. Bellare, S. Keelveedhi, and T. Ristenpart, "Dupless: Server aided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur. Symp., 2013, pp. 179–194

4. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," in Proc. 22nd Int. Conf. Distrib. Comput. Syst. , 2002, pp. 617–624.

5. M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Proc. EUROCRYPT, 2013, pp. 296–312.

6. P. Anderson and L. Zhang, "Fast and secure laptop backups with encrypted de-duplication," in Proc. USENIX 24th Int. Conf. Large Installation Syst. Admin., 2010

7. . S. Plank and L. Xu, "Optimizing Cauchy Reed-Solomon Codes for fault-tolerant network storage applications," in Proc. 5th IEEE Int. Symp. Netw. Comput. Appl., Jul. 2006, pp. 173–180.

8. Z. Wilcox-O'Hearn and B. Warner, "Tahoe: The least-authority file system," in Proc. ACM 4th ACM Int. Workshop Storage Security Survivability, 2008, pp. 21–26.

9. A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui, "A secure cloud backup system with assured deletion and version control," in Proc. 3rd Int. Workshop Secur. Cloud Comput., 2011.

10. Li, X. Chen, M. Li, J. Li, P. Lee, and W. Lou, "Secure deduplication with efficient and reliable convergent key management," IEEE Trans. Parallel Distrib. Syst. , vol. 25, no. 6, pp. 1615–1625, Jun.2014

11. C. Liu, Y. Gu, L. Sun, B. Yan, and D. Wang, "R-admad: High reliability provision for large-scale de-duplication archival storage systems," in Proc. 23rd Int. Conf. Supercomput., 2009, pp. 370–379.

12. M. Li, C. Qin, P. P. C. Lee, and J. Li, "Convergent dispersal: Toward storage-efficient security in a cloud-of-clouds," in Proc. 6th USENIX Workshop Hot Topics Storage File Syst., 2014, pp. 1–1

13. A. Rahumed, H. C. H. Chen, Y. Tang, P. P. C. Lee, and J. C. S. Lui, "A secure cloud backup system with assured deletion and version control," in Proc. 3rd Int. Workshop Secur. Cloud Comput., 2011.

14. M. W. Storer, K. Greenan, D. D. E. Long, and E. L. Miller, "Secure data deduplication," in Proc. 4th ACM Int. Workshop Storage Security Survivability , 2008, pp. 1–10.

15. J. Stanek, A. Sorniotti, E. Androulaki, and L. Kencl, "A secure data deduplication scheme for cloud storage," in Tech. Rep., 2013.

16. D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," IEEE Secur. & Privacy, vol. 8, no. 6, pp. 40–47, Nov./Dec. 2010.