International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Deduplication of Encrypted Textual Data in Cloud Environment

Mukhid Lashkari[1], Arpan Suntnure[2], Siddheshwar[3], Shubham Kathale[4], Hansaraj Wankhede[5]
[1,2,3,4]*Student, Dept. of Computer Engineering, G. H. Raisoni College of Engineering & Management, Pune, India*
[5]*Professor, Dept. of Computer Engineering, G. H. Raisoni College of Engineering & Management, Pune, India*

Abstract: **Data storage is the most important cloud service. The privacy of the data holders, and textual data are sometimes stored in cloud in an encrypted form. Still the encrypted textual data faces new challenges for cloud data deduplication, and this becomes important problem for big data storage and processing in cloud. The data stored in the cloud should be secured to don't access the unauthorized user. There comes a data security method known as Encryption form. In order to maintain the users, Privacy and the security of the data is stored in the cloud in the encrypted format. These problems cannot be flexibly support data access control and revocation. In this paper, we implement a scheme to deduplicate encrypted textual data stored in cloud based on ownership challenge and re-encryption. It integrates cloud data deduplication with access control. The implement and its performance based on extensive analysis and computer replication. The results can be predicted that the superior efficiency and effectiveness of the scheme for potential practical deployment to data deduplication in cloud storage.**

***Keywords*: Cloud computing, Java, Data deduplication**

1. Introduction

Data storage service is the most valuable in cloud service. The personal data is stored in the cloud service provider (CSP) and that is maintained by the cloud service provider. In existing research to maintain data privacy only outsource encrypted textual data are proposed to the cloud. Especially for shared data, the duplicated data in an encrypted form are stored by the same or different users which lead to wastage of storage. Moreover, the lost of control over the own personal data, leads to high data security risks, especially data privacy loss. The same or different users may stored duplicated data in encrypted form to CSP, specially for scenarios where data are shared among many users.

To deduplicate textual encrypted data stored in the cloud [5] and support secure data access control at the same time and efficient a scheme based on Advance encryption standard(AES) is proposed. AES is one type of private -key encryption in which the hidden key of a user and the encrypted key depends upon the statement. The decryption of the encrypted data is possible only if the set or group of statement of the private key matches with the statement of the encrypted data. Cloud storage services commonly use deduplication, which eliminates duplicate data by storing only a one copy of each file. [1] [12] Deduplication reduces the space and bandwidth requirements of data storage services, and is most effective when applied over the multiple users, a common practice by cloud storage. In source-based de-duplication, the client first hashes each data segment he wishes to upload and sends these results to the storage provider to check whether such textual data are already uploaded. Thus only "not de-duplicated" data segments will be actually stored by the client.[4]

2. Problem statement

We are going to build an application to deduplicate textual encrypted data stored in cloud based on ownership challenge and data encryption. We evaluate its performance based extensive analysis and computer simulation.

3. Literature survey

Clouds are large pools of easily usable and reachable resources. In cloud all resources connected virtually to create single system image. These resources can be dynamically reconfigured to adjust to a flexible load (scale), allowing optimum resource utilization.[14] Cloud storage refers to scalable and elastic storage capabilities that are delivered as a service using Internet technologies with elastic provisioning and use based pricing that does not penalize users for changing their storage consumption without notice. Two cloud approaches is suggested for data security and privacy of the users in cloud. [7]

In the paper named "Sedic: privacy-aware data intensive computing on hybrid clouds"[14], presents solution based on Merkle Trees and Specific encoding we identify attacks that exploit client side deduplication attempts to identify reduplication. The implemented prototype of the new protocol and ran it to evaluate performance and asses the POW scheme benefits. Sedic schedules Map's such that tasks on private clouds operate on sensitive data while tasks on public clouds operate on non-sensitive data.[14] In spite of these things it is impossible to verify experimentally the assumption about the input distribution.

In the year 2013 the paper named "Weak leakage –resilient client side Deduplication of encrypted data in cloud storage"[8] [13], propose a secure client – side deduplication scheme. It

addressed an important security concern in cross-user client – side deduplication but the convergent encryption and custom encryption methods are not semantically secure.

4. Proposed system

The proposed system will be given a feature based on data ownership challenge to manage textual encrypted data files with deduplication. To propose an effective approach to check file data ownership and check duplicate file contents with secure challenge and big data support. We integrate cloud data deduplication with data access control in a simple way. The proposed system consists of following models:

- *User:* First of all, user have to create a Login ID and password for the privacy of the individual data. If the user is not registered then user is not allowed to access to the uploaded data.
- *Upload:* After login of registered user, the user uploads the data to the cloud on which the operations are to be performed.
- *Deduplication check:* During the file is being uploading the file is verified that the duplication is not present. Here the files are categorize into file formats, like the file is txt file or in docs format. The files are being uploaded are compared. Then it sees whether the contents in the file are not present in the file on the earlier file are present on the cloud. If match found is greater than 75% of saved files in cloud, the server denies to upload. If it is same content as saved in cloud, the server will show the duplicate data in the percentage by comparing with all files in the whole cloud. If the duplication is found then server asks that user is wants to upload or not. If the user wants to upload file then the file is granted to upload.
- *Key generation:* The key generation performed on the basis of contents present in file. After getting all the credentials the file is sent for key generation. In this part the file is encrypted for the secure access of file.
- *Server:* The encrypted file is then stored on the cloud. The file can be accessed from the cloud.
- *File sharing and access:* If any user want to access the file uploaded by the other user then the file key is shared with the particular user. Then and only then the user can access the particular file.
- *Admin:* Admin keeps track of the all activities of the all users.

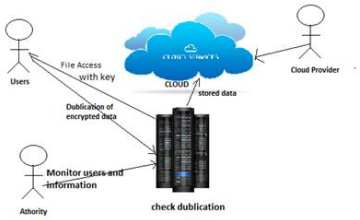

Fig. 1. System architecture

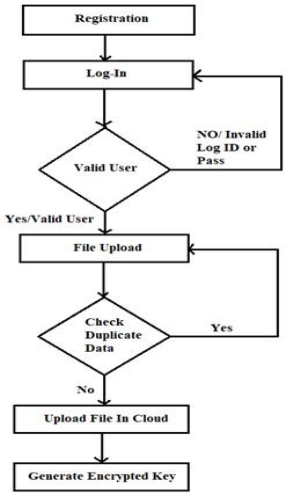

Fig. 2. Flowchart

A. Advantages

1. To manage the textual encrypted data efficiently.
2. Security can be achieved.
3. Easy to find duplicate data in textual files.
4. Easy to keep the Access Control in Users.

B. Disadvantages

1. This mechanism cannot work Offline.
2. It requires a High Speed Internet Connection.
3. It finds the duplication in only Textual Files.

5. Conclusion

We have developed the scheme that provides the deduplication of textual data in the cloud environment. And the all the files are stored after deduplication check-up are in form of encrypted format. We have also studied and practiced that Management of duplicate textual data in the Cloud and Encryption Standard Mechanism and Algorithms. The system can efficiently manage the encrypted data with deduplication. This is though important and significant in practice for achieving a successful cloud storage service. The data which is in encrypted form can be securely accessed because only authorized data holders can obtain the symmetric keys used for data decryption.

6. Future work

We will be develop the deduplication mechanism for all extensions formats of textual data. We will build this Cloud Deduplication Service for Android Devices. We will work on increasing the Response Time.

References

[1] M. Bellare, S. Keelveedhi, and T. Ristenpart, "DupLESS: Serveraided encryption for deduplicated storage," in Proc. 22nd USENIX Conf. Secur., 2013, pp. 179–194.
[2] Mozy, Mozy: A File-storage and Sharing Service. (2016). [Online]. Available: http://mozy.com/

[3] PasqualoPuzio, RefikMolva, Melek Onen, "Cloud Dedup: Secure Deduplication with Encrypted Data for CloudStorage", Seclud IT and EURECOM, France.
[4] Weak Leakage-Resilient Client –Side deduplication of Encrypted Data in Cloud Storage" Institute for Info Comm Research, Singapore, 2013
[5] Riddhi Movaliya, Harshal Shah, "A Survey of Secure Data Deduplication". International Journal of Computer Applications, Volume 138, No.11, March 2016.
[6] Khurud Pallavi B, Late Rohini A, Lawar Shreya A, Kurhade Jyoti B," Deduplication of Encrypted Big Data in Cloud", IJARIIE, Vol-3 Issue-2 2017.
[7] B Lalitha and G Murali. Implementing deduplication technique for rdf files with enhanced security using multi cloud servers. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pages 3618–3621. IEEE, 2017.
[8] Chang Liu, Chi Yang, Xuyun Zhang, and Jinjun Chen. External integrity verification for outsourced big data in cloud and iot: A big picture. Future generation computer systems, 49:58–67, 2015.
[9] Pasquale Puzio, Refik Molva, Melek Onen, and Sergio Loureiro. Cloudedup: secure deduplication with encrypted data for cloud storage. In Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, volume 1, pages 363–370. IEEE, 2013.
[10] Hyungjune Shin, Dongyoung Koo, Youngjoo Shin, and Junbeom Hur. Privacy preserving and updatable block-level data deduplication in cloud storage services. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), pages 392–400. IEEE, 2018.
[11] Prachi D. Thakar and D. G Harkut. Cloud based hybrid model for authorized deduplication. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2015.
[12] Rajashree Shivshankar Walunj, Deepali Anil Lande, and Nilam Shrikrushna Pansare. Secured authorized deduplication based hybrid cloud. International Journal of Engineering and Sciences, 3(11), 2014.
[13] Jibin Wang, Zhigang Zhao, Zhaogang Xu, Hu Zhang, Liang Li, and Ying Guo. I-sieve: an inline high performance deduplication system used in cloud storage. Tsinghua Science and Technology, 20(1):17–27, 2015.
[14] Zheng Yan, Wenxiu Ding, Xixun Yu, Haiqi Zhu, and Robert H Deng. Deduplication on encrypted big data in cloud. IEEE transactions on big data, 2(2):138–150, 2016. 22 Deduplication Techniques for Managing Textual Data in Cloud Environment.
[15] Taek-Young Youn, Ku-Young Chang, Kyung Hyune Rhee, and Sang U. K Shin. Efficient client-side deduplication of encrypted data with public auditing in cloud storage. IEEE Access, 2018.