Achieving Efficient and Privacy-Preserving Cross-Domain Big Data Deduplication in Cloud

Suraj Hole, Pooja Patil, Kishore Nair, Prof. Shamal Kashid
SIDDHANT COLLEGE OF ENGINEERING,PUNE

Abstract: Data deduplication will significantly reduce the communication and storage overheads in cloud storage services, and has potential applications in our huge data-driven society. Existing data deduplication schemes are generally designed to either resist brute-force attacks or make sure the efficiency and data availability, however not each conditions. We have a tendency to also be not conscious of any existing theme that achieves responsibility, within the sense of reducing duplicate data revelation in this paper, we have a tendency to investigate a three-tier cross-domain design, associate degree propose an efficient and privacy-reserving huge information deduplication in cloud. EPCDD achieves each privacy-preserving and data deduplication, and resists brute-force attacks. Additionally, we have a tendency to take responsibility into thought to supply higher privacy assurances than existing schemes. We have a tendency to then demonstrate that EPCDD outperforms existing competitive schemes, in terms of computation, communication and storage overheads. Also in this project dynamic operation user can perform data modification, data deletion and updating .After updating it sends file to local manager for checking deduplication if file is not duplicate then store file on cloud, and duplicate file will get proof of Ownership. This scheme will achieve the privacy preserving and cross domain big data deduplication in cloud. Additionally, the time complexness of duplicate search in EPCDD is index.

Keywords: Secure data deduplication, big data, brute-force attacks, data availability, and accountability.

1. Introduction:

Cloud storage usage is likely to increase in our big data driven society. For example, IDC predicts that the amount of digital data will reach 44 ZB in 2020. Other studies have also suggested that about 75% of digital data are identical (or duplicate), and data redundancy in backup and archival storage system is significantly more than 90%. While cost of storage is relatively cheap and advances in cloud storage solutions allow us to store increasing amount of data, there are associated costs for the management, maintenance, processing and handling of such big data. It is, therefore, unsurprising that efforts have been made to reduce overheads due to data duplication. The technique of data deduplication is designed to identify and eliminate duplicate data, by storing only a single copy of redundant data. In other words, data deduplication technique can significantly reduce storage and bandwidth requirements [6]. However, since users and data owners may not fully trust cloud storage providers, data (particularly sensitive data) are likely to be encrypted prior to outsourcing. This complicates data deduplication efforts, as identical data encrypted by different users (or even the same user using different keys) will result in different ciphertexts. Thus, how to efficiently perform data deduplication on encrypted data is a topic of ongoing research interest. In recent times, a number of data deduplication schemes have been proposed in the literature. These schemes are designed to realize encrypted data deduplication. However, the scheme suffers from brute-force attacks, the most popular attack in secure data deduplication schemes. Proposed another efficient secure deduplication scheme SecDep to resist brute-force attacks. However, this scheme only deals with small-sized data, and is not suitable for big data deduplication. To solve this problem, proposed a scheme to deduplicate encrypted big data stored in the cloud based on ownership challenge and proxy re-encryption. Although this scheme is efficient, it is vulnerable to brute-force attacks.

2. Existing System;

Existing data deduplication schemes are generally designed to either resist brute-force attacks or ensure the efficiency and data availability, but not both condition. We are also not aware of any existing scheme that achieves accountability, in the sense of reducing duplicate information disclosure (e.g., to determine whether plaintexts of two encrypted messages are identical).

Another efficient secure deduplication scheme SecDep to resist brute-force attacks. However, this scheme only deals with small-sized data, and is not suitable for big data deduplication. Existing system is not suitable to achieve big data deduplication and work with only small scale data. previous deduplication system not properly resist bruit force attack so we proposed a efficient and achieving big data deduplication.

3. Proposed System:

We propose an efficient and privacy-preserving big data deduplication in cloud storage (here after referred to as EPCDD). EPCDD achieves both privacy-preserving and data availability ,and resists brute-force attacks. In recent times, variety of knowledge deduplication schemes is planned within the literature. These schemes are designed to appreciate encrypted knowledge deduplication. We proposed a efficient and privacy preserving cross domain architecture in which user send request to key distribution server for key then key distribution server send public key to data user. Using secret key, data user encrypt file and send to local domain manager. Then local domain manager check this file in self domain, if file is already stored on local domain then does not need to send file to cloud for storing. Local domain manager send responses to data user that file is already stored on domain and give reference of file to data user. If file is not available on local domain then domain manager send file to cloud for storing. Then cloud service provider check in local domain B. if file is not available on domain B then cloud server store this unique and send response to local domain manager otherwise cloud only give reference of file to specific domain manager. Our systems achieve both privacy preserving and data availability and resist bruit force attack. Main advantage of our system is that data availability and accountability. In availability, the duplicated data has been deleted, as long as the client has uploaded the ciphertext corresponding to the specific data, it must ensure that this client can download and decrypt the stored ciphertext to obtain this data. In addition to achieving efficiency in storage, communication and computation, reliability, security and privacy should also be taken into consideration when designing a deduplication scheme.

4. Literature Survey:

1) Paper name: A Survey and Classification of Storage Deduplication Systems.

Author: JO˜ AO PAULO and JOS´E PEREIRA.

Year: May2014

Description: We define deduplication as a technique for automatically eliminating coarse-grained and unrelated duplicate data. Unlike traditional compression techniques that eliminate interfile redundancy or redundancy over a small group of files, typically stored together in the same operation, deduplication aims at eliminating both intra file and interfile redundancy over large datasets, stored at different times by uncoordinated users and activities, and possibly even across multiple distributed storage servers.

2) Paper name: Secure Deduplication of Encrypted Data without Additional Independent Servers.

Author: Jian Liu, N. Asokan, Benny Pinkas

Year: 2015

Description: In this paper we use the term to refer to both _les and blocks. Deduplication strategies can also be categorized according to the host where deduplication happens. In server-side deduplication, all _les are uploaded to the storage

server, which then deletes the duplicates. Clients are unaware of deduplication. This strategy saves storage but not bandwidth.

3) Paper name: DupLESS: Server-Aided Encryption for Deduplicated Storage.

Author: Mihir Bellare, Sriram Keelveedhi

Year: 2013

Description: We propose an architecture that provides secure deduplicated storage resisting brute-force attacks, and realize it in a system called DupLESS. In DupLESS, clients encrypt under message-based keys obtained from a key-server via an oblivious PRF protocol. It enables clients to store encrypted data with an existing service, have the service perform deduplication on their behalf, and yet achieves strong confidentiality guarantees. We show that encryption for deduplicated storage can achieve performance and space savings close to that of using the storage service with plaintext data.

4) Paper name: A Hybrid Cloud Approach for Secure Authorized Deduplication.

Author: Sunita S. Velapure1, S. S. Barde2

Year: 2014

Description: In this paper makes the primary plan to formally address the matter of licensed knowledge deduplication. Completely different from ancient deduplication systems, the differential privileges of user's area unit more thought-about in duplicate check besides the info itself. Addition to this we present many new deduplication constructions supporting licensed duplicate check in a hybrid cloud design. Security analysis demonstrates that our theme is secure in terms of the definitions as per the planned security model. As a proof of construct, we have a goal to implement a paradigm of our planned licensed duplicate check theme and conduct test bed experiments using our paradigm. We have a goal to show that our planned licensed duplicate check theme incurs comparatively less overhead compared to traditional operations. Secure Deduplication with Efficient and Reliable Convergent Key Management

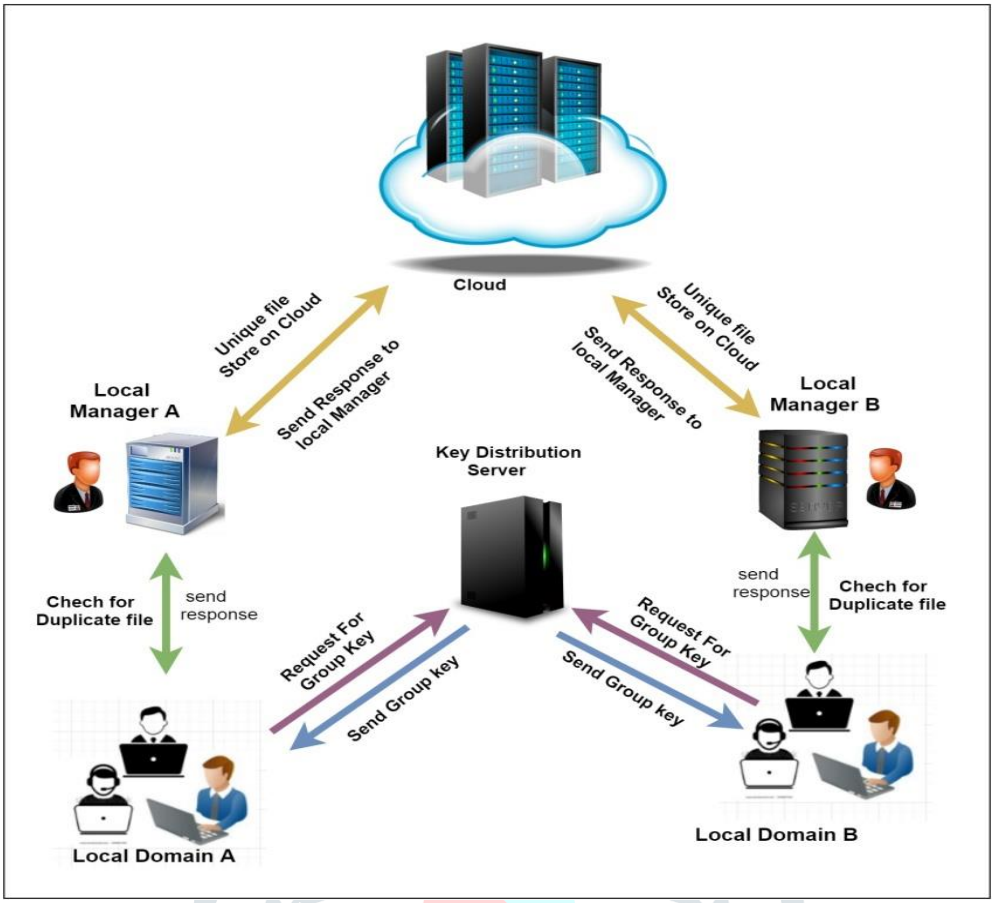
5) Paper name: A Hybrid Cloud Approach for Secure Authorized Deduplication.

Author: Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li,

Year: 2013

Description: This paper makes the first attempt to formally address the problem of achieving efficient and reliable key management in secure deduplication. We first introduce a baseline approach in which each user holds an independent master key for encrypting the convergent keys and outsourcing them to the cloud. However, such a baseline key management scheme generates an enormous number of keys with the increasing number of users and requires users to dedicatedly protect the master keys. To this end, we propose Dekey, a new construction in which users do not need to manage any keys on their own but instead securely distribute the convergent key shares across multiple servers. Security analysis demonstrates that Dekey is secure in terms of the definitions specified in the proposed security model.

5. Architecture Diagram:



6. Mathematical Model:

System S as a whole can be defined with the following main components.

S={ I , O , P , F , s , Ic)

1) Identify set of input as I

Let I={Set of outsourced data sets by corresponding data user}

2) Identify set of output as O

Let O={store unique file on cloud server .}

3) Identify the set of processes as P

- KDC=Key Distribution center.
- LM=Local manager.
- Uo=set of owners.
- Tg=Tag generation.
- DDT=Deduplication Decision Tree (searching the duplicate data.)
- Sk=Symmetric Key
- Gen(k)=Key Generator - bilinear parameter generator algorithm .
- Op= Output of System

4) Identify failure cases as F

- F=store duplicate file on Local manager server and cloud server.}

5) Identify success as s.

- s= {check duplicate file that is already store on Local manager server or Cloud server and I file already exist then duplicate file is not stored on cloud only give reference to new file.}

6) Identify the initial condition as Ic

Ic={Out sourced data with its privacy privileges to be maintain)

7. Conclusion:

Cloud storage adoption, particularly by organizations, is likely to remain a trend in the foreseeable future. This is, unsurprising, due to the digitization of our society. One associated research challenge is how to effectively reduce cloud storage costs due to data duplication. In this paper, we proposed an efficient and privacy-preserving big data deduplication in cloud storage for a three-tier cross domain architecture. We then analyzed the security of our proposed scheme and demonstrated that it achieves improved privacy preserving, accountability and data availability, while resisting brute-force attacks. We also demonstrated that the proposed scheme outperforms existing state-of-the-art schemes, in terms of computation, communication and storage overheads. In addition, the time complexity of duplicate search in our scheme is an efficient logarithmic time.

References:

[1] IDC, "Executive summary: Data growth, business opportunities, and the it imperatives,," http://www.emc.com/leadership/digital-universe/ 2014iview/executive-summary.htm, 2014.

[2] J. Gantz and D. Reinsel, "The digital universe decadeare you ready," https://hk.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf.

[3] H. Biggar, "Experiencing data de-duplication: Improving efficiency and reducing capacity requirements," The Enterprise Strategy Group., 2007. [Online]. Available: http://journals.sagepub.com/doi/abs/10.1177/ 000944550704300309

[4] D. Quick and K.-K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," Digital Investigation, vol. 11, no. 4, pp. 273–294, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.diin.2014.09.002

[5] ——, "Big forensic data reduction: digital forensic images and electronic evidence," Cluster Computing, vol. 19, no. 2, pp. 723–740, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10586-016-0553-1

[6] M. Dutch, "Understanding data deduplication ratios," http://www. chinabyte.com/imagelist/2009/222/l3pm284d8r1s .pdf, 2009.

[7] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels In cloud services: Deduplication in cloud storage," IEEE Security & Privacy, vol. 8, no. 6, pp. 40–47, 2010. [Online]. Available: http://dx.doi.org/10.1109/MSP.2010.187

[8] J. Paulo and J. Pereira, "A survey and classification of storage deduplication systems," ACM Comput. Surv., vol. 47, no. 1, pp. 11:1–11:30, 2014. [Online]. Available: http://doi.acm.org/10.1145/2611778

[9] S. Keelveedhi, M. Bellare, and T. Ristenpart, "Dupless: Server-aided encryption for deduplicated storage," in Proceedings of the 22th USENIX Security Symposium, Washington, DC, USA, August 14-16, 2013, 2013, pp. 179–194. [Online]. Available: https://www.usenix.org/conference/ usenixsecurity13/technical-sessions/presentation/bellare

[10] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings, 2013, pp. 296–312. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-38348-9 18