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

Prof. Brijendra Gupta, Akshay Salunkhe, Vrushali Darandale, Prashant Kaldoke, Dilip Sharma

Siddhant College of Engineering, Pune

Abstract: Secure information deduplication will considerably scale back the communication and storage overheads in cloud storage services, and has potential applications in our massive data-driven society. Existing information deduplication schemes area unit usually designed to either resist brute-force attacks or make sure the potency and information accessibility, however not each conditions. we tend to also are not conscious of any existing theme that achieves responsibility, within the sense of reducing duplicate data revealing (e.g., to work out whether or not plaintexts of 2 encrypted messages area unit identical). during this paper, we tend to investigate a three-tier cross-domain design, associated propose an economical and privacy-preserving massive information deduplication in cloud storage (hereafter mentioned as EPCDD). EPCDD achieves each privacy-preserving and information accessibility, and resists brute-force attacks. additionally, we tend to take responsibility into thought to supply higher privacy assurances than existing schemes. we tend to then demonstrate that EPCDD outperforms existing competitive schemes, in terms of computation, communication and storage overheads. additionally, the time quality of duplicate search in EPCDD is power.

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

1. Introduction:

Cloud storage usage is probably going to extend in our huge information driven society. for instance, IDC predicts that the quantity of digital information can reach forty four ZB in 2020. different studies have additionally urged that concerning seventy fifth of digital information square measure identical (or duplicate), and information redundancy in backup and repository storage system is considerably quite ninetieth. whereas price of storage is comparatively low cost and advances in cloud storage solutions enable U.S. to store increasing quantity of information, there square measure associated prices for the management, maintenance, process and handling of such huge information. It is, therefore, expected that efforts are created to cut back overheads because of information duplication. The technique of information of knowledge of information deduplication is intended to spot and eliminate duplicate data, by storing solely one copy of redundant information. In different words, information deduplication technique will considerably scale back storage and information measure necessities. However, since users and information homeowners might not absolutely trust cloud storage suppliers, information (particularly sensitive data) square measure doubtless to be encrypted before outsourcing. This complicates information deduplication efforts, as identical information encrypted by totally different completely different completely different} users (or even constant user mistreatment different keys) can end in different ciphertexts. Thus, the way to expeditiously perform information deduplication on encrypted information may be a topic of current analysis interest. In recent times, variety of information deduplication schemes are planned within the literature. These schemes square measure designed to understand encrypted information deduplication. However, the theme suffers from brute-force attacks, the foremost standard attack in secure information deduplication schemes. planned another economical secure deduplication theme SecDep to resist brute-force attacks. However, this theme solely deals with small-sized information, and isn't appropriate for large information deduplication. to resolve this downside, planned a theme to deduplicate encrypted huge information keep within the cloud supported possession challenge and proxy re-

encryption. though this theme is economical, it's at risk of 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 associate economical and privacy-preserving massive knowledge deduplication in cloud storage (here when named as EPCDD). EPCDD achieves each privacy-preserving and knowledge accessibility ,and resists brute-force attacks. In recent times, type of data deduplication schemes is planned inside the literature. These schemes area unit designed to understand encrypted data deduplication. we tend to projected a economical and privacy conserving cross domain design within which user send request to key distribution server for key then key distribution server send public key to knowledge user. exploitation secret key, knowledge user write in code file and send to native domain manager. Then native domain manager check this enter self domain, if file is already hold on on native domain then doesn't have to be compelled to send file to cloud for storing. native domain manager send responses to knowledge user that file is already hold on on domain and provides reference of file to knowledge user. If file isn't accessible on native domain then domain manager send file to cloud for storing. Then cloud service supplier sign up native domain B. if file isn't accessible on domain B then cloud server store this distinctive and send response to native domain manager otherwise cloud solely offer reference of file to specific domain manager. Our systems deliver the goods each privacy conserving and knowledge accessibility and resist rumor force attack. Main advantage of our system is that knowledge accessibility and responsibleness. In accessibility, the duplicated knowledge has been deleted, as long because the consumer has uploaded the ciphertext cherish the particular knowledge, it should make sure that this consumer will transfer and decipher the hold on ciphertext to get this knowledge. additionally to achieving potency in storage, communication and computation, dependability, security and privacy ought to even be taken into thought once coming up with a deduplication theme.

4. Literature Survey:

1) Paper name: Big forensic data reduction: digital forensic images and electronic evidence.

Author: Darren QuickKim-Kwang Raymond Choo

Year: 2015

Description: the data reduction process outlined can be applied using common digital forensic hardware and software solutions available in appropriately equipped digital forensic labs without requiring additional purchase of software or hardware. The process can be applied to a wide variety of cases, such as terrorism and organised crime investigations, and the proposed data reduction process is intended to provide a capability to rapidly process data and gain an understanding of the information and/or locate key evidence or intelligence in a timely manner.

2) Paper name: Message-locked encryption and secure deduplication

Author: Mihir BellareSriram KeelveedhiThomas Ristenpart

Year: 2013

Description: We formalize a new cryptographic primitive that we call Message-Locked Encryption (MLE), where the key under which encryption and decryption are performed is itself derived from the message. MLE provides a way to achieve secure deduplication (space-efficient secure outsourced storage), a goal currently targeted by numerous cloudstorage providers. We provide definitions both for privacy and for a form of integrity that we call tag consistency. Based on this foundation, we make both practical and theoretical contributions. On the practical side, we provide ROM security analyses of a natural family of MLE schemes that includes deployed schemes. On the theoretical side the challenge is standard model solutions, and we make connections with deterministic encryption, hash functions secure on correlated inputs and the sample-then-extract paradigm to deliver schemes under different assumptions and for different classes of message sources. Our work shows that MLE is a primitive of both practical and theoretical interest.

3) 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

4) Paper name: Side channels In cloud services: Deduplication in cloud storage

Author: Danny Harnik ; Benny Pinkas

Year: May2014

Description: As the volume of data increases, so does the demand for online storage services, from simple backup services to cloud storage infrastructures. Although deduplication is most effective when applied across multiple users, cross-user deduplication has serious privacy implications. Some simple mechanisms can enable cross-user deduplication while greatly reducing the risk of data leakage. Cloud storage refers to scalable and elastic storage capabilities delivered as a service using Internet technologies with elastic provisioning and usebased pricing that doesn't penalize users for changing their storage consumption without notice.

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

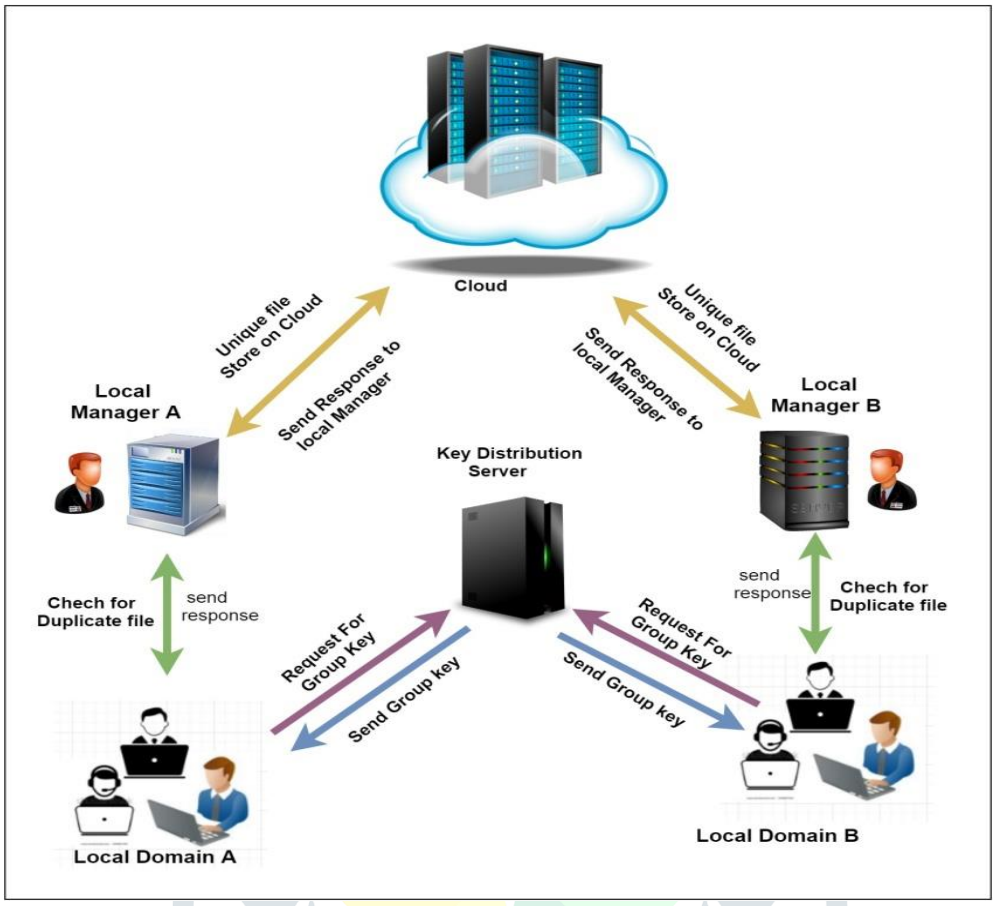
Author: Jin Li, Xiaofeng Chen, Mingqiang Li, Jingwei Li,

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.}

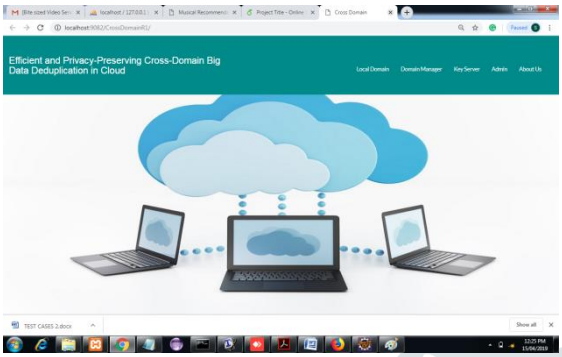
5) Identify success as s.

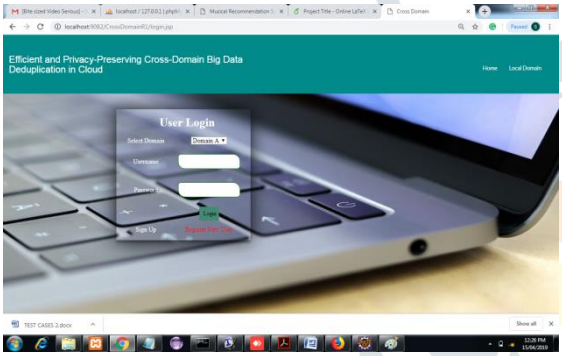
- s= {check duplicate file that is already store on Local manager server or Cloud server and I file already exist then duplicate file is not stored on cloud only give reference to new file.}

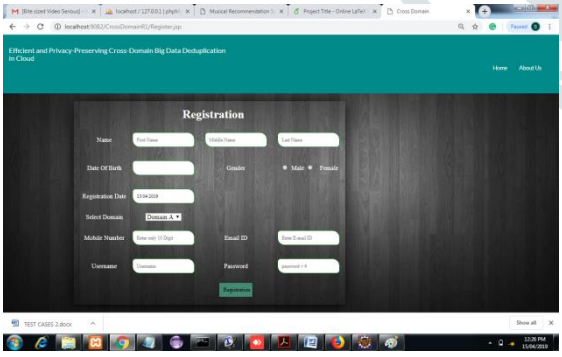
6) Identify the initial condition as Ic

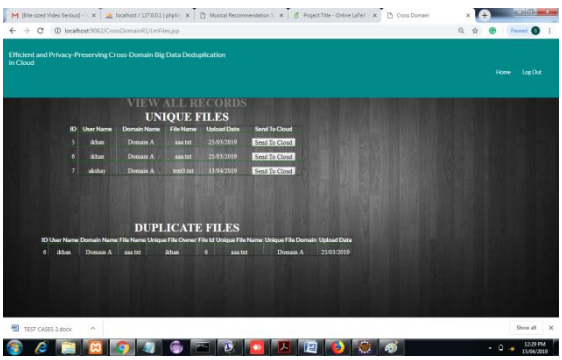
Ic={Out sourced data with its privacy privileges to be maintain)

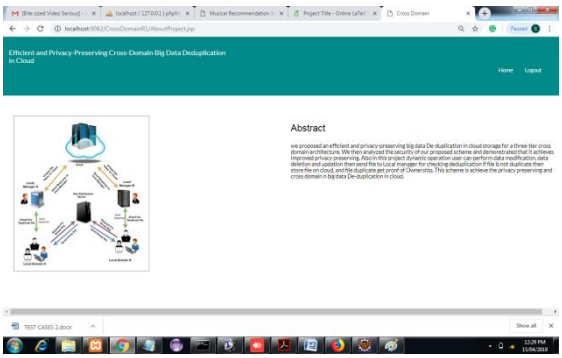
Results











7. Conclusion:

Cloud storage adoption, notably by organizations, is probably going to stay a trend within the predictable future. This is, unsurprising , because of the digitisation of our society. One associated analysis challenge is a way to effectively scale back cloud storage prices because of knowledge duplication. during this paper, we tend to planned AN economical and privacy-preserving huge knowledge deduplication in cloud storage for a three-tier cross domain design. we tend to then analyzed the protection of our planned theme and incontestible that it achieves improved privacy protective, responsibility and knowledge convenience, whereas resisting brute-force attacks. we tend to additionally incontestible that the planned theme outperforms existing progressive schemes, in terms of computation, communication and storage overheads. additionally, the time quality of duplicate search in our theme is AN economical exponent 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