A Survey Of Secure And Deduplication Frameworks For Cloud Based Application

Dr.G.Sushmitha Valli, Harika Arete

Abstract-In the contemporary era, data is growing rapidly and it is assuming characteristics of big data. Such data is outsourced to cloud infrastructure to have benefits like availability, scalability and affordability. However, there are issues related to storing duplicates copies of data. When data is duplicated, it causes storage overhead. Deduplication is the process of identifying and elimination of redundant copies of data in cloud. It allows one instance of data to be saved permanently and its duplicate copies will not have actual data but a reference to the same copy of saved data. It is widely used in cloud computing for backup technology to improve efficiency. Data compression and deduplication are two important techniques used by cloud service providers (CSPs) to optimize utilization of space in storage media. Data deduplication may take place at file level or block level. Deduplication may be made at source or the target end. Source deduplication consumes more processing power and it becomes difficult to handle it with existing resources. The target deduplication takes place in the backup system, probably in cloud storage which will be easier to get deployed. In many cloud based applications, there is need for performance optimization with secure deduplication. In this paper, a survey of different deduplication techniques is made to provide useful insights.

Keywords – Cloud computing, security, deduplication, secure deduplication

1. INTRODUCTION

Storage is essential for maintain valuable data of an organization. Traditionally data is stored in HDD of a system. There are many other storage media as well. Of late, cloud based storage also came into existence. There are mechanisms to reduce storage space such as compression techniques. However, compression techniques leave duplicates. There are other reasons for havingduplicates. There is need for removing duplicates if not, they lead to loss of storage space and performance of applications go down. As studied in [1]- [20], there are many techniques for removal of duplicates. There are techniques that can be employed at block level, byte level and file level. The data duplication methods are widely used in backup and recovery procedures as well. In the context of cloud and other storage media, in this paper investigation is made on the existing techniques used for de-duplication process. Our contributions in this paper include the review of literature on secure de-duplication techniques and their mechanisms. The remainder of the paper is organized as follows. Section 2 provides fundamentals of data de-duplication. Section 3 provides block-level data deduplication techniques. Section 4 presents byte-level data de-duplication methods. Section 5 presents the general de-duplication process. Section 6 presents secure de-duplication for cloud based healthcare systems.

2. DATA DEDUPLICATION

Data duplication as explored in [2] and [3] commonly occurs in many storage systems. Duplication of data causes many issues. For instance, it causes repetition of work or consumes more space in storage media. Even compression techniques leave duplicate copies of data. Duplication may occur in a single file, cross-document, cross-time, cross-client and cross-application. De-duplication is the process of eliminating duplicate copies. It is mainly used by backup systems and storage system in order to optimize storage facilities. Duplication files may occur in different time periods and in different locations and different size of data is encountered. When duplicate blocks are removed, it is done with an indicator. As studied in [4] and [5], there are highly redundant datasets that gain benefit from de-duplication techniques. Such techniques save time and effort besides making the storage place economical. When compression techniques are used, it leads to reduction of data but chances are there to increase duplications.

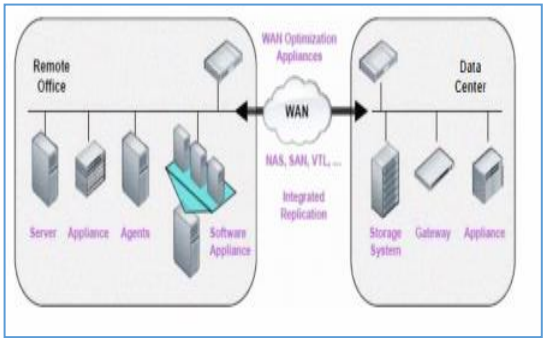

Figure 1: Locations where de-deduplication is needed

As presented in Figure 1, when it comes to normal compression techniques, the data de-duplication techniques are different as investigated in [13] and [14]. When data compression is employed, for each fie data is duplicated in one way or other. Though compression can reduce file size, it leaves a duplicate copy unless duplication is eliminated explicitly. Incremental backup also eliminates duplicate copies. Data de-duplication technologies are thus widely used in backup systems to ensure that the space is not wasted due to unnecessary duplications. Therefore, data de-duplication techniques verify and detects data duplications and eliminates them without actually causing loss of data.

- *Dr.G.Sushmitha Valli, Professor, Department Of Computer Science and Engineering, MLR Institute Of Technology. E-mail: susmitagv@gmail.com.*
- *Harika Arete, Student, Department Of Computer Science and Engineering, MLR Institute Of Technology. E-mail: areteharikareddy@gmail.com.*

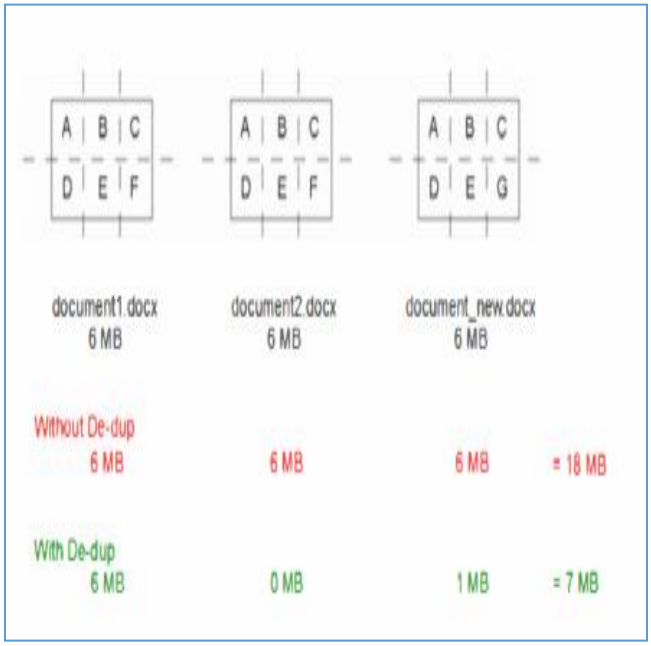

Figure 2: Data deduplication

As presented in Figure 2, there are there documents written with de-duplication technique and without de-duplication technique. When de-duplication technique is employed, the resultant storage needed by the files is 7 MB while the result of without de-duplication technique used is 18 MB. Thus de-duplication improves storage performance and leverages performance of backup and restore mechanisms that are widely used in business continuity and disaster recovery.

3.BLOCKLEVELDATADEDUPLICATIONTECHNOLOGY

The de-duplication techniques can work at block level as well as investigated in [10] and [11]. When data stream is divided into number of blocks, it is essential to check each block for identification of duplicates. Hash functions and digital signatures can be employed for data verification. When any block is found unique, it is stored in the permanent storage media and its index is updated. Otherwise only a pointer is updated to reflect the original location of the block in the storage media. The maintenance of a point to existing data is far better than maintaining a physical copy of data. In order to judge duplicates hash based algorithms are used. The popular hash based algorithms are SHA series and MD5 series.

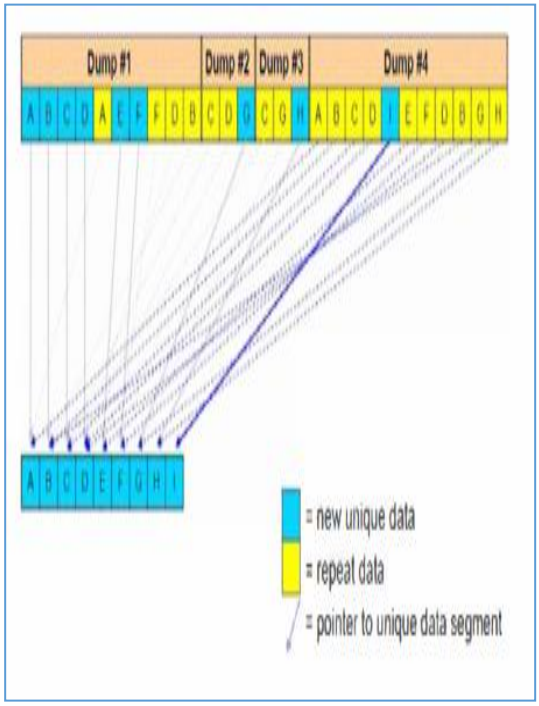

Figure 3: Shows data de-duplication technique in action

As can be seen in Figure 3, it is understood that by maintain a point to already existing data, the storage system avoids a duplicate. However, a pointer is maintained to the original data to have accessibility to the desired data. Thus it can reduce storage space and improve performance of algorithms and applications [12]. It also saves resources and money besides improving systems and promoting certainty. The de-duplication technique can reduce backup data and the size of data snapshots in order to save time and cost. It also consumes less power and improves network bandwidth utilization. Besides saving time, it saves disk space and backup and restore mechanisms will provide optimal performance. As backup storage contains duplications, it is essential to employ de-duplication mechanisms. It reduces the need for storage space and improves flexibility and does not cause issues to parties that need access to the data.

4. BYTE-LEVEL DATA DEDUPLICATION

Data de-duplication can occur at byte level. There might be data stream from which data arrives byte by byte for higher level of accuracy. The de-duplication products can identify byte level duplicates as well. As explored in [14], it is understood that the backup process provided a vendor may have specific implementation that can be reserve engineered in order to have better control over the de-duplication process. With determination of duplicates and eliminating them the computational load is reduced. The backup post-processing is essential to ensure that there are no duplicates kept unnecessarily. There needs to be disk cache in order to complete de-duplication process effectively [15].

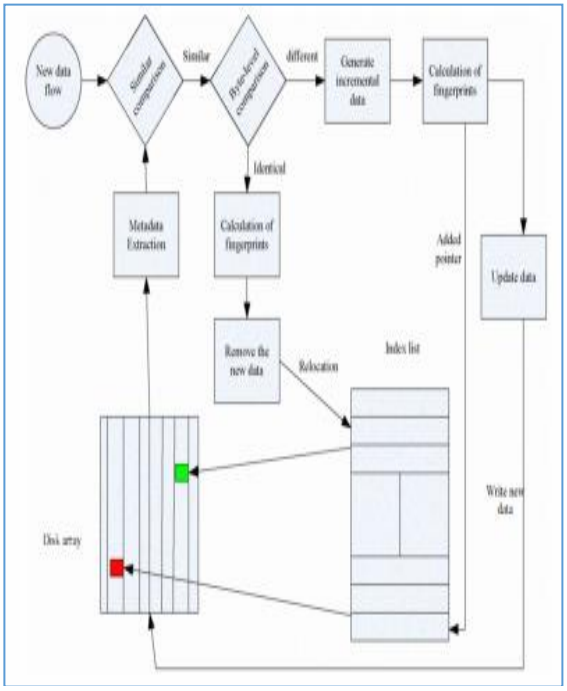

Figure 4: Logical flow of byte-level de-duplication strategy

As can be seen in Figure 4, the byte level de-duplication strategy is illustrated. When the de-duplication process is employed at byte level, it can reduce more disk space and recover data as well as and when needed. Thus it plays crucial role in backup and recovery procedures used in the real applications. Consistency checking is an important activity that is part of de-duplication process. It needs to consider data dynamics and see that the duplicates are eliminated but the accessibility to original copy of data is not lost. In the recovery process, there is need for reconstruction of data.

5. DATA DEDUPLICATION PROCESS

There are many researchers' contribution to data de-duplication process as found in [3], [8], [12] and [15. The first phase is known as data collection phase. It is achieved by comparing new and old backup and reducing the scope of duplications. The second phase is the process of identification of data in terms of bytes. When data uniqueness is considered an algorithm is used to know the real duplicates. In case of backup and restore procedures, sufficient care is needed to remove duplicates and then give access to original data by keeping pointers. In the process meta-data is also used for controlling the procedure and distinguish from the previous backup and recognize the duplicate data effectively. Data re-assembling is another phase of data de-duplication. When new data is saved, the previous duplicate copy may be marked as duplicate and keep a pointer to it rather than leaving physical copy there. The result of this process is to gain a copy of data that will not have duplicates. In the fourth phase, all the marked data for deletion is actually removed and at the same time integrity is verified. Redundant storage is removed and thus disk space is reclaimed for other potential users.

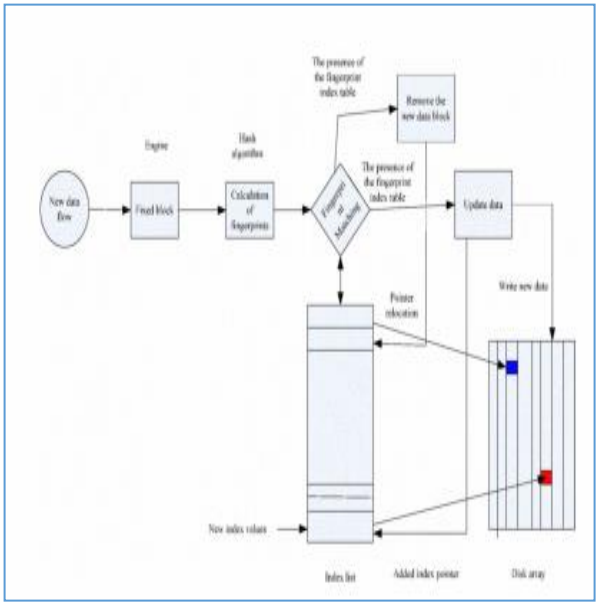

Figure 5: Logical flow of data de-duplication process

When data is stored different flat files, it is essential to use de-duplication process to see that the file content is not lost from user point of view but internal duplicates re eliminated. File level de-duplication strategy takes care of file duplications in terms of content. A file may have number of versions. However, there are duplicates in the files and thus it can be eliminated with de-duplication techniques. Backup performance will not be affected by duplicate removal. Instead, its performance will be improved when duplicates are avoided. There is less impact on recovery time as well. The pointer used to point original copy of data is essential to ensure that the data is not duplicated but the pointer.

5. DISASTER RECOVERY AND EVIDENCE RECOVERY

Secure de-duplication plays an important role in disaster recovery and evidence recovery applications. Such applications can be realized as illustrated in Figure 6.

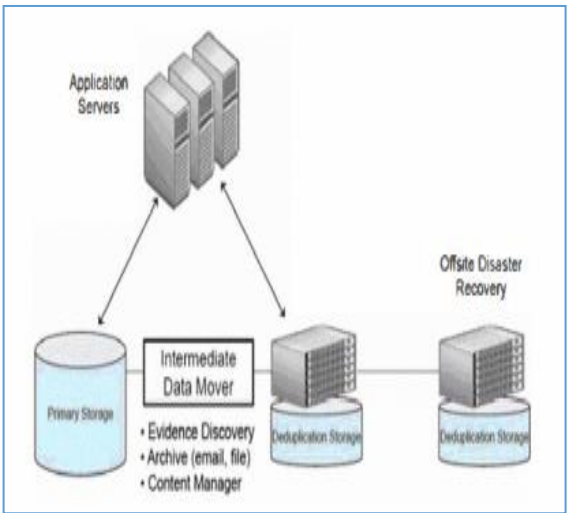

Figure 6: Realization of disaster recovery and evidence recovery

Business systems may have both background and foreground processing. Thus the business may maintain in-premise digital infrastructure and offsite disaster recovery procedures. The foreground processing is made using software implementations. The software may be stand alone or it may support client-server architecture. Servers are used to take backup at given time and allow clients to use the backup up data. In the same process, there might be hash options to ensure data integrity. A duplicate block gets removed but pointer is maintained to the original copy of data. There is provision for data recovery and restoration of deleted data. The secure-deduplication techniques also reduce network traffic and amount of storage in storage media. There needs to be an integrated mechanism to realize the system with de-duplication procedures. When business applications do not follow secure de-duplications, they severely affect performance of the system. For secure de-duplication, there is an inline approach or the remote approach. Both need to use hash algorithms in order to ensure data integrity. It has mechanisms like de-duplication, recovery of lost data and also backup. Hash algorithms ensure data integrity in presence of different operations like data deletion as explored in [12], [14] and [15]. There are post-processing techniques to have hash based differential algorithms or technologies as studied in [12] and [13].

6. SECURE DEDUPLICATION FOR CLOUD BASED HEALTHCARE SYSTEMS

Healthcare is the industry which has high impact on the lives of people. It produces large volumes of data. Such data needs help of computing resources available in cloud. Electronic Health Record (EHR) or Electronic Medical Record (EMR) is the data related to one patient permanently stored in storage media. When there is cloud-assisted healthcare system, it can have benefits of scalability and availability. Such system needs both security and deduplication of data for protection of data and efficiency. Many existing systems such as strived to provide security. Less research is found on the deduplication of medical records [1]. The security also needs to be improved with lightweight approaches.

7. CONCLUSION AND FUTURE WORK

In this paper, we reviewed different techniques available for deduplication. Different levels of data deduplication strategies are covered. They include block level techniques, byte level techniques and file level techniques. It also covers general deduplication process and how it works. The data deduplication technology is presented. From the review, it is understood that there are many approaches for deduplication. Some of the approaches are able to provide security while performing deduplication. The strategies are categorized based on granularity, time of application and point of application. As mentioned above, file level and block level are based on granularity. Inline deduplication and post process deduplication are techniques based on time of application. Source based and target based deduplication techniques are based on point of application. Secure deduplication techniques threw light on both aspects like removal of duplicates and security. In future we intend to research on medical records for more effective and secure deduplication.

References

[1] Zhang, Y., Xu, C., Li, H., Yang, K., Zhou, J., & Lin, X. (2018). HealthDep: An Efficient and Secure Deduplication Scheme for Cloud-Assisted eHealth Systems. IEEE Transactions on Industrial Informatics, 1–11.

[2] He, X., Jin, R., & Dai, H. (2018). Deep PDS-Learning for Privacy-Aware Offloading in MEC-Enabled IoT. IEEE Internet of Things Journal, 1–8.

[3] Xu, G., Li, H., Tan, C., Liu, D., Dai, Y., & Yang, K. (2017). Achieving efficient and privacy-preserving truth discovery in crowd sensing systems. Computers & Security, 69, 114–126.

[4] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicaterecord detection: A survey. Knowledge and Data Engineering,IEEE Transactions on, 19:1-16, 2007.

[5] Y. Wang and S. Madnick. The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In Proceedings of the Fifth International Conference on Data Engineering, pages 46-55, Washington, DC, USA, 1989. IEEE Computer Society.

[6] http://www.linux-mag.comlidl7535

[7] H. Li, Y. Yang, Y. Dai, J. Bai, S. Yu, and Y. Xiang, “Achieving secure and efficient dynamic searchable symmetric encryption over medical cloud data,” IEEE Transactions on Cloud Computing, 2017, to appear.

[8] M.Bellare, S. Keelveedhi, and T. Ristenpart, “Message-locked encryption and secure deduplication,” in Proceedings of EUROCRYPT. Springer, 2013, pp. 296–312.

[9] M.Bellare, S. Keelveedhi, and T. Ristenpart, “Dupless: Server-aided encryption for deduplicated storage,” in Proceedings of USENIX Security Symposium. USENIX, 2013, pp. 179–194.

[10] “Smartphones as practical and secure location verification tokens for payments,” in Proceedings of NDSS. Internet Society, 2014, pp. 1–15.

[11] Y. Zhang, C. Xu, H. Li, and X. Liang, “Cryptographic public verification of data integrity for cloud storage systems,” IEEE Cloud Computing, vol. 3, no. 5, pp. 44–52, 2016.

[12] http://www.snia.orglsearch?cx=001200299847728093177%3A3rwmjfdm8ae&cof=FORID%3AII&q=data+deduplication&sa=Go#994

[13] http://bbs.chinabyte.comlthread-393434-1-I.htrnl

[14] http://storage.chinaunix.netlstor/c/

[15] Austin Clements, Irfan Ahmad, MuraliVilayannur, and Jinyuan Li. Decentralized deduplication in san cluster file systems. In Proc. of the USENIX Annual Technical Conference, June 2009.

[16] J. Li, C. Qin, P. P. Lee, and X. Zhang, “Information leakage in encrypted deduplication via frequency analysis,” in Proceedings of DSN. IEEE, 2017, pp. 1–12.

[17] D. T. Meyer and W. J. Bolosky, “A study of practical deduplication,” ACM Transactions on Storage, vol. 7, no. 4, 2012.
[18] J. Xu, E. Chang, and J. Zhou, “Weak leakage-resilient client-side deduplication of encrypted data in cloud storage,” in Proceedings of ASIACCS. ACM, 2013, pp. 195–206.
[19] Miao, M., Wang, J., Li, H., & Chen, X. (2015). Secure multi-server-aided data deduplication in cloud computing. Pervasive and Mobile Computing, 24, 129–137.
[20] Shin, Y., Koo, D., &Hur, J. (2017). A Survey of Secure Data Deduplication Schemes for Cloud Storage Systems. ACM Computing Surveys, 49(4), 1–38.