© 2018 IJSRST | Volume 4 | Issue 5 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Survey On Various Data Deduplication Techniques Applied In Cloud

Thilagavathi N[1], Praveena C[2], Niranjani S[2]
[1]Associate professor, Department of Information Technology, SMVEC,Puducherry, Tamil Nadu, India
[2]UG student, Department of Information Technology, SMVEC, Puducherry, Tamil Nadu, India

ABSTRACT

Data deduplication is useful for organizations dealing with highly redundant operations that requires constant copying and storing of data for future reference or recovery purpose. The technique is a part of backup and disaster recovery solution as it allows enterprises to save data repeatedly and promotes fast, reliable and cost-effective data recovery. Deduplication run an analysis and eliminates these sets of duplicate data and keeps only what is unique and essential, thus significantly clearing storage space. Here are some benefits of data deduplication for organizations

Keywords : Deduplication, Progressive, Convergent Encryption, Proxy Re-Encryption

I. INTRODUCTION

Cloud Computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the internet. It is a general term used to describe a new class of network-based computing that takes place over the Internet. Cloud computing is easy for data owners to outsource their data to public cloud servers and it allows data users to retrieve the data. It is the use of resources that are delivered as a service over a network. The cloud providers manage the infrastructure and platforms on which the applications run. End users access cloud-based applications through web browser or a lightweight desktop or mobile app while the business software and user's data are stored on servers at a remote location. Cloud computing enables companies and applications, which are system infrastructure dependent, to be infrastructure-less. By using the Cloud infrastructure on “pay as used and on demand”, all of us can save in capital and operational investment.

Big Data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them.

Duplication of data refers to the process of creating the exact copy of already existing data. This duplication of data in cloud has adverse effects. Some of the adverse effects of data duplications are

- ✓ Wasted cost and low income
- ✓ Waste of space in cloud storage
- ✓ Inefficiency and lack of productivity
- ✓ Poor customer service
- ✓ Poor business processes

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. This technique is used to improve storage utilization and can also be applied to network data transfer to reduce the number of bytes that must be sent. In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. There are various data deduplication techniques available. Hence in this paper we present the various deduplication techniques and their benefits.

II. DATA DEDUPLICATION

Data deduplication is data compression technique for eliminating duplicate copies of repeating data. It is used to improve storage utilization.

Risk Factors in Deduplication Detection

1) A user has only limited, maybe unknown time for data cleansing and wants to make best possible use of it. Then, simply start the algorithm and terminate it when needed. The result size will be maximized.
2) A user has little knowledge about the given data but still needs to configure the cleansing process. Then, let the progressive algorithm choose window/block sizes and keys automatically.
3) A user needs to do the cleaning interactively to, for instance, find good sorting keys by trial and error. Then, run the progressive algorithm repeatedly; each run quickly reports possibly large results.
4) A user must achieve a certain recall. Then, use the result curves of progressive algorithms to estimate how many more duplicates can be found further; in general, the curves asymptotically converge against the real number of duplicates in the dataset.

INFLUENCE IN DEDUPLICATION

Improved early quality Let t be an arbitrary target time at which results are needed. Then the progressive algorithm discovers more duplicate pairs at t than the corresponding traditional algorithm. Typically, termination is smaller than the overall runtime of the traditional algorithm.
Same eventual quality If both a traditional algorithm and its progressive version finish execution, without early termination, they produce the same results

REVIEW OF LITERATURE

Most research on duplicate detection also known as entity resolution and by many other names, focuses on pair selection algorithms that try to maximize recall on the one hand and efficiency on the other hand. The most prominent algorithms in this area are Blocking and the sorted neighbourhood method (SNM).

Adaptive techniques. Previous publications on duplicate detection often focus on reducing the overall runtime. Thereby, some of the proposed algorithms are already capable of estimating the quality of comparison candidates. The algorithms use this information to choose the comparison candidates more carefully. For the same reason, other approaches utilize adaptive windowing techniques, which dynamically adjust the window size depending on the amount of recently found duplicates. These adaptive techniques dynamically improve the efficiency of duplicate detection, but in contrast to our progressive techniques, they need to run for certain periods of time and cannot maximize the efficiency for any given time slot.

Progressive techniques. In the last few years, the economic need for progressive algorithms also initiated some concrete studies in this domain. For instance, pay-as-you-go algorithms for information integration on large scale datasets have been presented. Other works introduced progressive data cleansing algorithms for the analysis of sensor data streams. However, these approaches cannot be applied to duplicate detection.

In this paper they explained building a deduplication storage system over cloud computing [1]. This is an efficient deduplication system, which is not able to handle encrypted data.

In this paper, they explained a verifiable [2] data deduplication scheme in cloud computing. This technique adopts two servers to achieve verifiability of deduplication.This scheme is an image deduplication scheme.

In this paper, they explained a secure two-phase data deduplication scheme [3], Meye et al. proposed to

adapt two servers for intra user deduplication and inter deduplication.

This paper deals with; CloudDedup [4]: Secure deduplication with encrypted data for cloud storage. This paper aims to cope with the inherent security exposures of convergent encryption, but it cannot solve the issues caused by data deletion.

In this paper, Bellare et al. proposed DupLESS: Server aided encryption for deduplicated storage that provides secure deduplicated storage to resist brute-force attacks [5]. In DupLESS, a group of affiliated clients (company employees) encrypt their data with the help of Key server(KS) that is separate from a Storage Service(SS).

In this paper, they explained the policy-based deduplication in secured cloud storage [6]. This proxy scheme was proposed but it did not consider duplicated data management.eg deletion and owner management and did not evaluate scheme performance.

In this paper, they explained the efficient hybrid inline and out of line deduplication for back up storage [7]. The strict latency requirements of primary storage lead to the focus on offline deduplication systems.

In this paper, Fu et al. explained the accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information [8], they proposed history aware rewriting algorithm to accurately identify and rewrite fragmented chunks, which improved the restore performance.

In this paper, Kaczmarczyk et al. [9] explained reducing the impact of data fragmentation caused by inline deduplication, focused on inter-version duplication and proposed context-based rewriting to improve the restore performance by shifting fragmentation to older backups for latest backups.

In this paper, they explained how to improve restore speed for backup systems that use inline chunk-based [10] deduplication. This work proposed to forfeit the deduplication to reduce the chunk fragmentation by container capping.

In this paper, they introduced a scheme to manage encrypted data storage with deduplication in cloud. It proposed using Proxy Re-encryption [11] for cloud data deduplication. This scheme applies the hash code of data. This hash code of data helps us to check ownership with signature verification. This is insecure because H(M) is disclosed to a malicious user. In this paper, we aim to support big data deduplication in an efficient way by surveying the existing techniques to ideate a new technique.

In this paper, they have discussed about the attacks on convergent Encryption [12]. Ng et al. adapted the PoW to manage the deduplication of encrypted data. This scheme also generates verification information based on Merkle trees for deduplication check. Based on several data blocks each leaf value is generated, while each interactive proof protocol can challenge one leaf to the Merkle tree. Higher security is achieved by executing the protocol multiple times and by checking more data.

In this paper, they explained the provable ownership of file [13] in deduplication cloud storage. Yang et al. also proposed a cryptographically secure and efficient scheme using which the ownership of the file can be verified, in which a client must prove to the server that it indeed possesses the entire file without uploading the file.

This paper deals with secure and constant cost public cloud storage [14] auditing with deduplication. Yuan and Yu attempted to solve the issue of supporting

efficient and secure data integrity auditing with storage deduplication for cloud storage . They proposed a novel scheme based on techniques including polynomial-based authentication tags and homomorphic linear authenticators. Their design allows deduplication of both files and their corresponding authentication tags.

Table 1. Overview Of The Deduplication Techniques

Paper	Feature	Result
DupLESS: Server-Aided Encryption for Deduplicated Storage	Space saving High performance	Simple Storage
A Secure Client-Side Deduplication Scheme in Cloud Storage Environments	Access control Privacy	Ensure better confidentiality
Hybrid Data Deduplication in Cloud Environment	Easy to implement Provide transparency	Data can be store as per requirement either in encrypted or un-encrypted area
Dynamic Data Deduplication in Cloud Storage	Improve Storage efficiency	Maintaining redundancy for fault tolerance
A Verifiable Data Deduplication Scheme in Cloud Computing	Storage saving	Verifiable data deduplication
Twin Clouds: An Architecture for Secure Cloud Computing (Extended Abstract)	Secure computation Store large amount of data Secure execution environment	Client uses the trusted Cloud as a proxy that provides a clearly defined interface to manage the outsourced data, programs, and queries.
A Hybrid Cloud approach for secure Authorized Deduplication	Differential authorization Authorized duplicate check Unforgeability of token	Reduce storage space and save network bandwidth

Data integrity auditing and storage deduplication are achieved simultaneously. Public auditing and batch auditing are both supported. But feasibility of supporting deduplication big data was not discussed in this work.

This paper deals with improving accessing efficiency of cloud storage using de-duplication and feedback schemes. To reduce workloads due to duplicate files, Wu et al. proposed [15] Index Name Servers (INS) to manage not only file storage, data deduplication, optimized node selection, and server load balancing, but also file compression, chunk matching, real-time feedback control, IP information, and busy level index monitoring. To manage and optimize storage nodes based on a client-side transmission status by the proposed INS, all nodes must elicit optimal performance and offer suitable resources to clients. In this way, not only can the performance of a storage system be improved, but the files can also be reasonably distributed, decreasing the workload of the storage nodes. However, this work cannot deduplicate encrypted data.

In this paper, they discuss about hybrid data deduplication in cloud environment [16]. Fan et al. proposed a hybrid data deduplication mechanism that provides a practical solution with partial semantic security. This solution supports deduplication on plaintext and ciphertext. But this mechanism cannot support encrypted data deduplication very well. It works based on the assumption that CSP knows the encryption key of data. Thus, it cannot be used in the situation that the CSP cannot be fully trusted by the data holders or owners.

III. CONCLUSION

In this paper we have analysed concerning deduplication concept[table 1], where we tend to confer many new duplicate check in cloud architecture and access privileges given to retrieve or store the data in cloud. We have done a review on data deduplication challenges in chunking process, scalability, throughput, metadata processing, parallelizing dedupe process and deploying data deduplication on cluster to achieve good deduplication performance. For the comparison of chunking algorithms and deduplication ratio with them, first we divided data into fixed and variable size chunks using chunking algorithms with maximum efficiency and less time. In future we would like to explore the deduplication on images, videos and deploy the same on cloud-based storage with proposed storage policies.

IV. REFERENCES

[1]. Z. Sun, J. Shen, and J. M. Yong, "DeDu: Building a deduplication storage system over cloud computing," in Proc. IEEE Int. Conf. Comput. Supported Cooperative Work Des., 2011, pp. 348–355, doi:10.1109/CSCWD.2011.5960097.

[2]. Z. C. Wen, J. M. Luo, H. J. Chen, J. X. Meng, X. Li, and J. Li, "A verifiable data deduplication scheme in cloud computing," in Proc. Int. Conf. Intell. Netw. Collaborative Syst., 2014, pp. 85–90, doi:10.1109/INCoS.2014.111.

[3]. P. Meye, P. Raipin, F. Tronel, and E. Anceaume, "A secure twophase data deduplication scheme," in Proc. HPCC/CSS/ICESS, 2014, pp. 802–809, doi:10.1109/HPCC.2014.134.

[4]. P. Puzio, R. Molva, M. Onen, and S. Loureiro, "ClouDedup: Secure deduplication with encrypted data for cloud storage," in Proc. IEEE Int. Cof. Cloud Comput. Technol. Sci., 2013, pp. 363–370, doi:10.1109/CloudCom.2013.54.

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

[6]. C. Y. Liu, X. J. Liu, and L. Wan, "Policy-based deduplication in secure cloud storage," in Proc. Trustworthy Comput. Serv., 2013, pp. 250–262, doi:10.1007/978-3-642-35795-4_32.

[7]. Y.-K. Li, M. Xu, C.-H. Ng, and P. P. C. Lee, "Efficient hybrid inline and out-of-line deduplication for backup storage," ACM Trans. Storage, vol. 11, no. 1, pp. 2:1-2:21, 2014, doi:10.1145/2641572.

[8]. M.Fu, et al., "Accelerating restore and garbage collection in deduplication-based backup systems via exploiting historical information," inProc.USENIXAnnu.Tech.Conf., 2014, pp.181–192.

[9]. M. Kaczmarczyk, M. Barczynski, W. Kilian, and C. Dubnicki, "Reducing impact of data fragmentation caused by in-line deduplication," in Proc. 5th Annu. Int. Syst. Storage Conf., 2012, pp. 15:1–15:12, doi:10.1145/2367589.2367600.

[10]. M. Lillibridge, K.Eshghi, and D.Bhagwat,"Improving restore speed for backup systems that use inline chunk-based deduplication," in Proc.USENIXConf.FileStorageTechnol.,2013,pp. 183–198.

[11]. Z. Yan, W. X. Ding, and H. Q. Zhu, "A scheme to manage encrypted data storage with

deduplication in cloud," in Proc. ICA3PP2015, Zhangjiajie, China, Nov. 2015, pp. 547–561.

[12]. D. Perttula, B. Warner, and Z. Wilcox-O'Hearn, "Attacks on convergent encryption." (2016). [Online]. Available: http://bit.ly/ yQxyvl

[13]. C. Yang, J. Ren, and J. F. Ma, "Provable ownership of file in deduplication cloud storage," in Proc. IEEE Global Commun. Conf., 2013, pp. 695–700, doi:10.1109/GLOCOM.2013.6831153.

[14]. J.W.Yuan and S.C.Yu, "Secure and constant cost public cloud storage auditing with deduplication," in Proc. IEEE Int.Conf. Communic. Netw.Secur.,2013, pp.145–153, doi:10.1109/CNS.2013.6682702.

[15]. T. Y. Wu, J. S. Pan, and C. F. Lin, "Improving accessing efficiency of cloud storage using de-duplication and feedback schemes," IEEE Syst. J., vol. 8, no. 1, pp. 208–218, Mar. 2014, doi:10.1109/ JSYST.2013.2256715.

[16]. C. Fan, S. Y. Huang, and W. C. Hsu, "Hybrid data deduplication in cloud environment," in Proc. Int. Conf. Inf. Secur. Intell. Control, 2012, pp. 174–177, doi:10.1109/ISIC.2012.6449734.