ISSN 2278-3091
Volume 8, No.1.4, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5181.42019.pdf
https://doi.org/10.30534/ijatcse/2019/5181.42019

Systematic Literature Review: Leveraging Data Deduplication Strategies & Hashing Techniques to Eliminate Data Redundancy in Cloud Environments

M.A. Fazlina[1], Rohaya Latip[1], [2], Hamidah Ibrahim[1], Azizol Abdullah[1],
[1]Faculty of Computer Science and Information Technology, University Putra Malaysia.
[2] Institute for Mathematical Research (INSPEM),
University Putra Malaysia.
rohayalt@upm.edu.my

ABSTRACT

Technology advancements are core contributors to massive data evolutions across the globe. As data is blooming universally the storage and network bandwidth are on high demands. Many survey evidences that data for in storage seems to have multiple same or similar copies of data and to be exact very minimal of the total data are unique where else, the remaining amount are identical data shared between numerous users. This scenario resulting high bandwidth and storage consumptions, especially for cloud providers. Variety of solutions has anticipated enhancing cloud storage performance. This Systematic Literature Review (SLR) is focusing on Deduplication in Cloud Replication Environment and the results are alleviated to show holistic deduplication adoption and adaption in cloud environments. The study successfully addressed four (4) research questions and analysis was done for future research.

Key words: Cloud Environment, Deduplication, Hashing Techniques, Replication.

1. INTRODUCTION

Digital data attained tremendous growth at current global world [1]. International Data Corporation (IDC) in year 2011, testified that data volume generated and copied across the globe will be 35ZB by 2020 [2]. This exponential data counts are flooding in storages causing huge space been occupied mostly by redundant data [2]. Currently, the main challenge faced by the cloud as storage platform service provider is to reduce the storage consumption without degrading the performance service. Hence, the best state-of-art strategy available is data deduplication (dedup) techniques [3]–[5].

Numerous researchers depicted 'Data Deduplication' is the most efficient strategy to resolve the multiple data copies in storage[6] and [7].Data deduplication (dedup) techniques are extensively used to eliminate duplicate copy of data in cloud storage [8], [9]. Rather than only reducing the storage overheads, deduplication similarly capable to optimized the bandwidth usage [10]. There are two types of deduplication techniques are widely employed in cloud environments; server-side dedup and client-side dedup [11]. Client-side dedup is always deploy data elimination approaches before data sent to storage and satisfies users in saving bandwidth and storage as well [12]. On the other hand, the server-side dedup performed when data are placed in storage and achieve to secure storage space and enhance system performance on client side. [4]

There are countless studies embarked in data deduplication and researchers produced various dedup strategies with different techniques. Each of proposed approaches is proven to enhanced and improved various system performances in cloud environment

The rest of the paper is organized as follows; Section 2 explains thoroughly on systematic review methodology. Discussions on research question are in section 3 and Section 4 is conclusion.

2. METHODOLOGY

This study was conducted referring to guidelines by[13]–[15]. There are three (3) main phases in the review process; Planning the Review, Conducting the Review and Reporting the Review. Figure 1 below displays SLR guideline model:

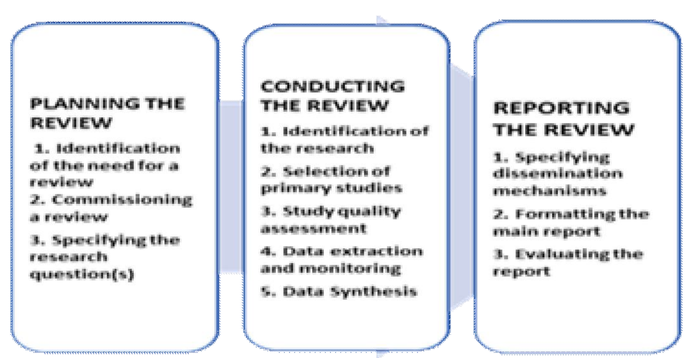

Figure 1: Systematic Literature Review (SLR) Guideline [16]

2.1 Research Questions

There are Four (4) research questions was constructed for this study purposes. The entire literature review process for this study was guided by these questions. In order to produce precise and well-organized research questions, a model called 'Population, Intervention, Comparison and Outcomes' (PICO) approach by [17] was adopted in this study. Table 1 below presents in detailed about criteria, research questions and purpose of study in PICO structure.

Table 1: PICO - Criteria and Research Questions

Criteria	Research Question	Purpose
Population	R1: How many researches have been done within the year of 2015 till 2019, related to data deduplication on cloud environment?	To identify how many articles had been published in this research area between the year of 2015 till 2019 with inclusion of deduplication in cloud replication cloud replication environment.
Intervention	R2: What is the redundancy elimination strategies used for Data Deduplication?	To understand deduplication strategies developed and used by researchers for their study.
Comparison	R3: What are the most used hashing techniques in data deduplication?	To define the technique used in the experiment in order to attain best data duplicate detector.
Outcomes	R4: What are the advantages & limitations of each research?	To identify limitation/gaps of related studies.

2.2 Data Sources

The data gathered for this study was retrieved from five (5) online databases; IEEEXplore Digital Library, ScienceDirect, Scopus, SpringerLink and Web of Science.

2.3 Search Strategies

In order to obtain all related research articles for this work, specific identification steps were used to produce efficient search string [13]. The string identification processes are as below:

1. Find similar or synonym for the search term using thesaurus.
2. A phrase searching was done to get overview of search results.
3. Truncation and Wildcard (symbol *) was added for wider search results to include singular/plural or any spelling variances in search phrases.
4. Boolean AND & OR are used to combine few words to complete better sentences for accurate search results.

After merging all necessary functions in search terms, the final search string was formed as follow: ((" Data Deduplication " OR " Data Reduplication " OR " Data Replicating" OR " Data Duplication) AND " Replication " OR "Data Replication " OR " Cloud Replication Environment " OR " Cloud Environment " OR " Cloud Replication "))

Through-out this search process, there was some modification made in the search string for IEEEXplore and Scopus to meet better search results.

2.4 Study Selection Procedure

As the study selection, field code function was not limited only to Abstract/Keywords/Title, on the other hand, search was performed in entire document to avoid any related research article unnoticed in search results. This determination was made due to some papers are not specifically mentioning the search string in particular field code function.

2.5 Inclusion and Exclusion Criteria

Throughout this study, screening by computer function was completed through few inclusion and exclusion conditions. This process is crucial to obtain significant search article results which literally addressing the research questions [16]. The criteria itemized in Table 2 as below:

Table 2: Inclusion and Exclusion Criteria

Inclusion	Exclusion
i. Language: English.	i. Papers with duplicate study.
ii. Publication Year: 2015 to 2019.	ii. Not Relevant: Title/Domain/Abstract/ Introduction/Conclusion.
iii. Empirical Data with Results: Research Article/ Journal/ Conference Paper (Special Issue)/Book Chapters.	iii. Paper which does not contain any empirical study/data.

2.6 Results

A thorough process was done to retrieve most relevant papers for this study. As first, the total search result for all five online databases was 3247 papers. Initial screening was started by looking into not valid article (article content not readable). Therefore, 6 papers were eliminated as rubbish record due to no paper found with unrelated title. Additionally, another 25 papers were identified as duplicate papers across those five databases. Subsequently, researcher found 2851 papers need to be discarded from the list due to inappropriate title and domain of study. The manual checking continues by reading the abstract for the remaining papers and that resulting 421 papers removed due to irrelevant study material indeed. Next, total of 71 papers were classified as unrelated study background as the introduction and conclusion was not addressing the research questions. Eventually, researcher read through the rest of the papers and finalized 21 papers to be removed from the available list since the content of papers does not produce novel contribution and it's the study leads to different direction which is insignificant for this research purposes. Ultimately, 27 papers were identified as anchor papers for this study

The overall process to acquire the most eligible articles for this study is shown in the Figure 2 below:

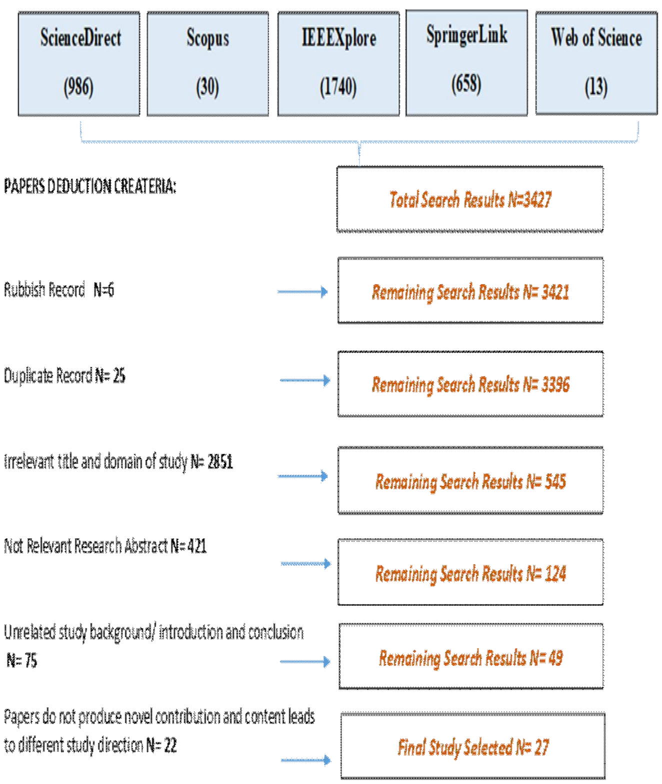

Figure 2: Final Studies Selection Process

The final anchor papers are further analyzed to discover regarding the most trending publishers in this research area. The information is presented as in Figure 3 below:

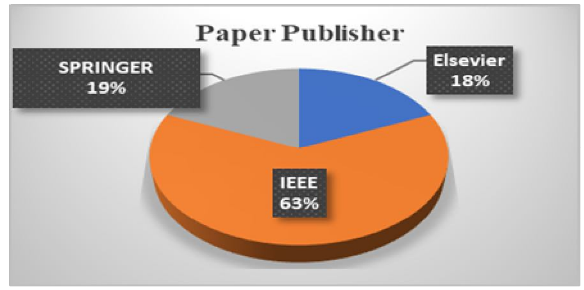

Figure 3: Paper Publisher

Based on analysis, pie chart above apparently demonstrates the research area for this study is quite popular and mostly published in IEEE (63%), followed by Springer (19%) and Elsevier (18%).

3. DISCUSSION

The constructed Research Questions (RQ) for this systematic literature review will be addressed and discussed in this section thoroughly. All RQ answered based on the final 27 selected as described in previous section in Figure 2.

RQ1: How many researches have been done within the year of 2015 till 2019, related to data deduplication on cloud environment?

Based on this SLR study within the year of 2015 till early 2019, thousands of studies were found embarked related to topic "Deduplication in Cloud Environment" inclusive cloud replication environment. As mentioned in [18], replication is among the best strategy to provide data availability for users. Hence, one of the best approaches to improve the replication strategy is using deduplication methods which optimize the overall performance for system especially in storage consumptions [7], [19]. Thus, this SLR study visibly proves that this study area is vital and many researchers are proactively trying hard to explore similarly improve various extent in this matter.

There are many perspectives where Deduplication (dedup) is implemented in cloud environment. Based on the final selected studies for this SLR, researcher identified 5 main areas where dedup was adopted and adapted. The areas are Primary Cloud Storage, Cloud Storage, Replication Storage, Backup/Recovery and Cache Memory. Table 1 below, shows number of studies implemented dedup in various areas.

Table 3: Areas of Deduplication (Dedup) Deployment

Dedup Implementation Areas	Cloud Storage	Primary Cloud Storage	Replication Storage	Backup/ Recovery Storage	Cache Memory
Number of Studies	14	5	3	4	1

RQ2: What are the redundancy elimination strategies used for Data Deduplication?

Deduplication (dedup) method used to identify same or similar data in cloud environments [20], [21]. Researchers has exploits dedup methods in variety type of data such as; images, videos, structured data, semi-structured data and unstructured data [22], [23]. Therefore, through this SLR study researcher managed to gather significant data on proposed strategies by respective researchers in final selected studies. Every individual study is given Identification Number (ID) for easier analysis. Table 4 later, presents the summary for 27 final selected papers which contributed in producing novel deduplication strategies in cloud environments. Literally, Table 4 answering the RQ2 and RQ4 for this study.

RQ3: What are the most used hashing techniques in data deduplication?

Identifying and removing duplicates data copies is promising by using deduplication methods. Nevertheless, choosing and using best techniques in dedup are the challenge for researchers. Research questions 3 specifically to address the most used hashing techniques in deduplication. Hashing is technique used in deduplication to produce certain value in order to identify data similarity [23]. There are numerous hashing algorithms available and the most common and well-known are Message Digest (MD5), Secure Hashing Algorithms (SHA-1, SHA-2, SHA-256, SHA-3) and more various techniques. Different hashing generates dissimilar value with certain bytes depends on the chosen hashing type. This study addressing the RQ3 by providing informative data and state-of-art on hashing techniques most used in recent studies as in Figure 4.

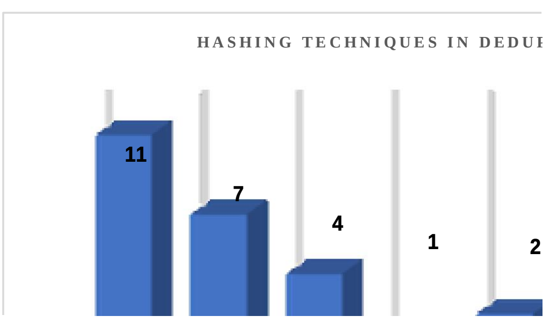

Figure 4: Hashing Techniques in Deduplication

Graph is produced based on final selected studies for this SLR. Based on the graph SHA-1 apparently used most compared to others. After thorough reading and analysis, researcher found the popularity of SHA-1 not only because its good performance version of hashing algorithm available but on the other hand, it has very low collision rate compared to other hashing algorithm [24]. Therefore, this study proves that, although SHA-1 is older version hashing algorithm the stability and capability to identify and eliminate duplicate data is not questionable.

RQ4: What are the limitations of each related research?

Last discussion is for RQ4, but most important research question to be answered in this study is to find research limitation in related studies. As discussed in RQ1, there are numerous researchers focusing in this same area "Deduplication in Cloud Environments". All of them have their own research goals and aims. They successfully achieved their research objectives with different perceptions, nevertheless the research gaps or limitations are still existed as trade-off for every study.

The gaps in selected studies are varies, but then analysis depicted similar major and crucial issue in researchers work. The key issues are categorized as in Figure 5 below:

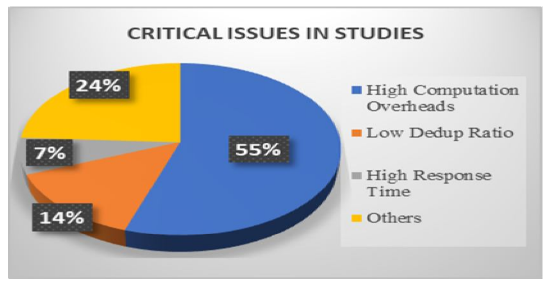

Figure 5: Critical Issues in Studies

Figure 5 shows critical issues in most studies in pie chart where 'high computation' visibly the critical issue in most of studies. After exhaustive reading and analysis, researcher perceived the possible reason for 'computation overhead' and 'high response time' is when particular studies intended to achieve 'best deduplication ratio'. Meticulous and extensive process is required to get 'high deduplication ratio', which consequently impacts 'response time' and 'process time' to be much longer. On the other hand, usually the trade-off for 'faster response time' is 'low deduplication ratio' caused by the computation process in algorithms short with simple formula but deduplication rate is less efficient.

Regardless of researchers' novel contributions and improvements, the respective studies have trade-offs and limitation which are considered tolerable. Essentially, all gaps are definitely an opportunity for researchers' future enhancements.

Table 4 is the detailed summary for each study consist of answers for RQ1, RQ2, RQ3 and RQ4 that derived from final 27 selected papers for this SLR.

DEDUP AREA: CLOUD STORAGE					
AUTHOR	**CONTRIBUTION (DEDUP STRATEGIES)**	**DEDUP LEVEL**	**HASHING METHOD**	**MEASUREMENT METRIC**	
				ADVANTAGES	**LIMITATION**
[25]	Enhanced Cuckoo Hashing	Fixed-Sized	Cuckoo	• Low Storage Usage • Low Bandwidth Usage	• High Computation
[26]	Secure Block-Level Message-Locked Encryption (MLE)	Block-Level	Short Hashing (SH)	For Large Block-Size: • Low Bandwidth Usage	For Small Block-Size: • High Computation
[27]	Secure Data Dedup based Rabin CDC & MD5	Content Defined Chunking (CDC)	MD5	• High Security • High Dedup Ratio • Low Bandwidth Usage	• High Computation
[28]	Based Similarity & Delta Encoding (DABSD) Algo	Byte-Level	Bloom Filter	• High deduplication ratio • Low Storage Usage	• High Computation
[29]	Secure Dedup with Proof-of-Ownership (PoW)	File-Level & Block Level	SHA-256	• Low Bandwidth usage • High Security	• High Computation
[30]	Primitive Randomize MLE2 (μR-MLE2)	File-Level	SHA-1	• High Security • High Performance	• High Computation for Dynamic Dedup • Low Dedup Ratio
[31]	Genetic Evolution (GE) based Two Thresholds Two Divisors (TTTD-P) Algo	Content Defined Chunking (CDC)	SHA-1	• High Dedup Ratio • High Server Through-put • Low Dedup Computation	• High Bandwidth Usage
[11]	Enhanced Chord Algo with Client & Server-Side Dedup	Block-Level	MD5	• High Dedup Ratio • Low Storage Usage	• Low Response Time

DEDUP AREA: CLOUD STORAGE					
AUTHOR	**CONTRIBUTION (DEDUP STRATEGIES)**	**DEDUP LEVEL**	**HASHING METHOD**	**MEASUREMENT METRIC**	
				ADVANTAGES	**LIMITATION**
[32]	Secure Authorized Deduplication with Token mechanism	File-Level	SHA-256	• High Data Security	• High Computation (Big file size)
[33]	Secure Comprehen-sive Sensing (CS) Scheme	Content Defined Chunking (CDC)	SHA-3	• Low Storage Usage • High Dedup • High Security	• High Computation
[34]	Cloud Data Management Interface (CDMI)	File-Level	SHA-2	• Small file • Low Response Time (Upload =<1kb) • High Data Transmission	• Big Files • High Response Time (Up/Download >1kb) • High Dedup Computation
[20]	Bucket-based Deduplication Mechanism	Fixed-Sized Level	MD5	• Low Storage Consumption • High Dedup Ratio	• High Computation
[1]	BOAFFT	Super-Chunk	MinHash	• Low Network Usage • Improve Data Dedup	• High Storage Resource (Need Large Cluster)
[9]	Secure Distributed Deduplication System	Fixed-Size Block-level & File- level	SHA-256	• High Reliability • Low Storage • Low Network Usage • High Consistency	• Encoding/ Decoding Overhead
DEDUP AREA: PRIMARY STORAGE					
[35]	Stream Locality Aware Deduplication (SLADE) Algo	Fixed-Size level	MD5	• High Dedup Ratio • Low Storage Usage	• High Computation
[19]	DC- Dedup	Fixed-Sized	SHA-1	• Low Storage Usage • Low Bandwidth Usage	• Compression Time Overhead
[36]	Cluster-based Incremental Sorted Neighbourhood Method(ciSNM)	Large-Block(Pair Comparison)	ciSNM	• High Dedup Ratio (for Large Size data) • Reduce Num. of data for Dedup	• Low Dedup Ratio (Big size data)

DEDUP AREA: PRIMARY STORAGE					
AUTHOR	CONTRIBU-TION (DEDUP STRATEGIES)	DEDUP LEVEL	HASHING METHOD	MEASUREMENT METRIC	
				ADVANTAGES	LIMITATION
[37]	Dedup-lication Strategy using HDFS	File-Level	MD5 & SHA-1	• Efficient Memory • Low Storage Usage • Low Computa-tion	• Low Dedup Ratio
[24]	Dedup-based non-volatile Phase-Change Memory (PCM)	Content Defined Chunking (CDC)/ Var-Sized	SHA-1	• Low Storage Usage	• Performance Degrade
DEDUP AREA: PRIMARY CACHE					
[38]	Pre-Cache	Fixed-Sized & Var-Sized (CDC)	SHA-1	• High Dedup Ratio	• High Computa- tion
DEDUP AREA: CLOUD REPLICATION STORAGE					
[39]	Autho-rize Dedup Technique using DARE	Block-Level	SHA-1/SHA-256 & Delta Compres-sion (XDelta)	• High Dedup Ratio • High Server Through-put • Low Storage Usage	• High Computa- tion
[40]	Deduplica tion-Assisted primary storage in Cloud-of-Clouds (DAC)	Fixed-Sized	MD5 & SHA-1	• Low Storage Usage • Efficient Memory • Low Response Time	• Low Data Availability (1 replica only)
[41]	Exten-ded Data Dedup with Replica-tion Control	Block-Level	MD5/ SHa-1	• Low Storage Usage	• High Computa-tional (async. & ack.)

4. CONCLUSION

Objectives of this SLR are to obtain state-of-art research trends, focus and gaps for deduplication in cloud replication environment. Through methodology guide all four (4) research questions successfully addressed through-out this study. As the replication is the well-known key to address business continuity, more research are expressively recomended especially via integrating deduplication techniques. Comprehensive studies in many prespectives must be completed prior to produce suitable dedup strategies for replication and that is the core challenge faced by scholars. Hence, this SLR proves not many studies did specifically delivers dedup strategy in replication. Consequently, deduplication for replication environement is vibrant zone for research and there still rooms for new contribution and improvement.The trends in this SLR show this research topic is significant as continuous research was progressing annually. Existing deduplication method and techniques are advancing contributed by numerous researchers who are proactively enhancing deduplication capabilities. Unfortunately, research gaps are still persisted and the limitation finding in this SLR useful for future explore and improvements.

DEDUP AREA: BACKUP/RECOVERY STORAGE					
AUTHOR	CONTRIBU-TION (DEDUP STRATEGIES)	DEDUP LEVEL	HASHING METHOD	MEASUREMENT METRIC	
				ADVANTAGES	LIMITATION
[42]	Multi-Level Pattern Matching Algo-(MLPMA)	File-Level	Bloom Filters	• Low Computa-tion • High Server Through-put • Efficient Memory	• Low Dedup Ratio
[43]	Asymmet-ric Extremum (AE) Algo	Content Defined Chunk-ing (CDC)	SHA-1	• High System Throug-hput • Low Computa-tion • High deduplicat-ion ratio	• Small Data Sizes (<3GB)
[44]	Resemble & Mergence based Dedup (RMD) scheme	Content Defined Chunk-ing (CDC)	RMD	• High Performan-ce • Low Response Time	• High Memory
[45]	Deduplica-tion Aware Resemble (DARE)	Fixed-Sized	SHA-1	• High Dedup ratio • Fast Backup/ Recovery	• High Computation

ACKNOWLEDGEMENT

This work is supported by Putra Grant, University Putra Malaysia (Grant No: 95960000). Utmost appreciation and thanks to provide sufficient facilities and funding through-out this research.

REFERENCES

1. K. Li, C. Wu, G. Zhang, S. Khan, and S. Luo, **Boafft: Distributed Deduplication for Big Data Storage in the Cloud**, *IEEE Trans. Cloud Comput.*, vol. 61, no. 11, pp. 1–1, 2015. https://doi.org/10.1109/TCC.2015.2511752
2. R. Kaur, I. Chana, and J. Bhattacharya, **Data Deduplication Techniques For Efficient Cloud Storage Management: A Systematic Review**, *J. Supercomput.*, vol. 74, no. 5, pp. 2035–2085, 2018.
3. W. Chen, Y. Hu, S. Yin, and W. Xia, **EEC-Dedup: Efficient erasure-coded deduplicated backup storage systems,** *Proc. - 15th IEEE Int. Symp. Parallel Distrib. Process. with Appl. 16th IEEE Int. Conf. Ubiquitous Comput. Commun. ISPA/IUCC 2017*, vol. 2, no. Goal 1, pp. 251–258, 2018.
4. H. Hovhannisyan, K. Lu, R. Yang, W. Qi, J. Wang, and M. Wen, **A novel deduplication-based covert channel in cloud storage service**," *2015 IEEE Glob. Commun. Conf. GLOBECOM 2015*, pp. 1–6, 2015.
5. J. Patil and S. S. Barve, **DDFP: Duplicate detection and fragment placement in deduplication system for security and storage space,** *Proc. - 1st Int. Conf. Intell. Syst. Inf. Manag. ICISIM 2017*, vol. 2017–Janua, pp. 225–229, 2017.
6. B. Mao, H. Jiang, S. Wu, and L. Tian, **Leveraging Data Deduplication to Improve the Performance of Primary Storage Systems in the Cloud**, *IEEE Trans. Comput.*, vol. 65, no. 6, pp. 1775–1788, 2016. https://doi.org/10.1109/TC.2015.2455979
7. D. Sureshpatil, R. V. Mane, and V. R. Ghorpade, **Improving the Availability and Reducing Redundancy using Deduplication of Cloud Storage System**, *2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017*, pp. 1–5, 2018.
8. A. Khan, C. G. Lee, P. Hamandawana, S. Park, and Y. Kim, **A robust fault-tolerant and scalable cluster-wide deduplication for shared-nothing storage systems**, *Proc. - 26th IEEE Int. Symp. Model. Anal. Simul. Comput. Telecommun. Syst. MASCOTS 2018*, pp. 87–93, 2018.
9. J. Li *et al.*, **Secure Distributed Deduplication Systems with Improved Reliability**, *IEEE Trans. Comput.*, vol. 64, no. 12, pp. 3569–3578, 2015.
10. H. Cui, C. Wang, Y. Hua, Y. Du, and X. Yuan, **A Bandwidth-Efficient Middleware for Encrypted Deduplication**, *2018 IEEE Conf. Dependable Secur. Comput.*, pp. 1–8, 2018.
11. X. Xu, N. Hu, and Q. Tu, **Two-side data deduplication mechanism for non-center cloud storage systems**, *2016 IEEE Int. Conf. Ubiquitous Wirel. Broadband, ICUWB 2016*, 2016.
12. Z. Pooranian, K. C. Chen, C. M. Yu, and M. Conti, **RARE: Defeating side channels based on data-deduplication in cloud storage**, *INFOCOM 2018 - IEEE Conf. Comput. Commun. Work.*, pp. 444–449, 2018. https://doi.org/10.1109/INFCOMW.2018.8406888
13. H. A. M. Shaffril, S. E. Krauss, and S. F. Samsuddin, **A Systematic Review On Asian's Farmers' Adaptation Practices Towards Climate Change**, *Sci. Total Environ.*, vol. 644, pp. 683–695, 2018. https://doi.org/10.1016/j.scitotenv.2018.06.349
14. B. Kitchenham and S. Charters, **Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3**, *Engineering*, vol. 45, no. 4ve, p. 1051, 2007.
15. M. P. and H. Roberts, **Systematic Reviews in the Social Sciences: A Practical Guide.**, *Couns. Psychother. Res.*, 2006.
16. [16] B. Kitchenham and S. Charters, **Procedures for Performing Systematic Literature Reviews in Software Engineering**, *Keele Univ. Durham Univ. UK*, 2007.
17. M. Petticrew, " M. Petticrew and H. Roberts. **Systematic Reviews in the Social Sciences: A Practical Guide** . Oxford: Blackwell 2006. 352 pp. ISBN 1 4051 2110 6. £29.99 ," *Couns. Psychother. Res.*, vol. 6, no. 4, pp. 304–305, 2006.
18. M. A. Fazlina, R. Latip, H. Ibrahim, and A. Abdullah, **A Review : Replication Strategies for Big Data in Cloud Environment**, vol. 7, pp. 357–362, 2018.
19. B. Zhang, C. Wang, B. B. Zhou, D. Yuan, and A. Y. Zomaya, **DCDedupe: Selective Deduplication and Delta Compression with Effective Routing for Distributed Storage**, *J. Grid Comput.*, vol. 16, no. 2, pp. 195–209, 2018.
20. N. Kumar, R. Rawat, and S. C. Jain, **Bucket based data deduplication technique for big data storage system**, *2016 5th Int. Conf. Reliab. Infocom Technol. Optim. ICRITO 2016 Trends Futur. Dir.*, pp. 267–271, 2016. https://doi.org/10.1109/ICRITO.2016.7784963
21. L. T. Yang, R. Yang, Y. Zhou, L. Si, and Y. Deng, **LDFS: A Low Latency In-Line Data Deduplication File System**, *IEEE Access*, vol. 6, pp. 15743–15753, 2018.
22. M. Parekh, S. Bidani, and V. Santhi, *Proof of* **Retrieval and Ownership for Secure Fuzzy Deduplication of Multimedia Data**, vol. 710. Springer Singapore, 2018.
23. G. Walsh, **Guide to Big Data: Deduplication Practices for Multimedia Data in the Cloud**, *Rashid F., Miri A. Deduplication Pract. Multimed. Data Cloud. Srinivasan S. Guid. to Big Data Appl. Stud. Big Data, vol 26. Springer, Cham*, vol. 5, no. 2, pp. 48–56, 2018.
24. C. T. Lin, Y. H. Chang, T. W. Kuo, H. S. Chang, and H. P. Li, **How to improve the space utilization of dedup-based PCM storage devices**?, *2015 Int. Conf. Hardware/Software Codesign Syst. Synth. CODES+ISSS 2015*, pp. 11–20, 2015.
25. R. N. K. and L. N. K. J. Sridharan, C. Valliyammai, **Data De-duplication Using Cuckoo Hashing in Cloud Storage,** vol. 758. Springer Singapore, 2019.
26. H. Shin, D. Koo, Y. Shin, and J. Hur, **Privacy-Preserving and Updatable Block-Level Data Deduplication in Cloud Storage Services**, *IEEE Int. Conf. Cloud Comput. CLOUD*, vol. 2018–July, no. 1, pp. 392–400, 2018.
27. H. Kambo and B. Sinha, **Secure data deduplication**

mechanism based on Rabin CDC and MD5 in cloud computing environment, *RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.*, vol. 2018–Janua, pp. 400–404, 2018.

28. Q. S. Song B., Xiao L., Qin G., Ruan L., **A Deduplication Algorithm Based on Data Similarity and Delta Encoding**, *Yuan H., Geng J., Bian F. Geo-Spatial Knowl. Intell. GRMSE 2016. Commun. Comput. Inf. Sci. vol 699. Springer, Singapore*, vol. 698, pp. 154–163, 2017.
29. S. Jiang, T. Jiang, and L. Wang, **Secure and Efficient Cloud Data Deduplication with Ownership Management**, *IEEE Trans. Serv. Comput.*, vol. 1374, no. c, pp. 1–14, 2017. https://doi.org/10.1109/TSC.2017.2771280
30. [30] T. Jiang, X. Chen, Q. Wu, J. Ma, W. Susilo, and W. Lou, **Secure and Efficient Cloud Data Deduplication with Randomized Ta**g, *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 3, pp. 532–543, 2017.
31. N. Kumar, S. Antwal, G. Samarthyam, and S. C. Jain, **Genetic optimized data deduplication for distributed big data storage systems**, *4th IEEE Int. Conf. Signal Process. Comput. Control. ISPCC 2017*, vol. 2017–Janua, pp. 7–15, 2017.
32. V. Waghmare and S. Kapse, **Authorized Deduplication: An Approach for Secure Cloud Environment,** *Phys. Procedia*, vol. 78, pp. 815–823, 2016.
33. F. Rashid and A. Miri, **Secure image data deduplication through compressive sensing**, *2016 14th Annu. Conf. Privacy, Secur. Trust. PST 2016*, pp. 569–572, 2016.
34. X. L. Liu, R. K. Sheu, S. M. Yuan, and Y. N. Wang, **A file-deduplicated private cloud storage service with CDMI standard**, *Comput. Stand. Interfaces*, vol. 44, pp. 18–27, 2016.
35. H. Wu, C. Wang, Y. Fu, S. Sakr, K. Lu, and L. Zhu, "**A differentiated caching mechanism to enable primary storage deduplication in clouds**," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 6, pp. 1202–1216, 2018.
36. A. Samiei and F. Naumann, **Cluster-Based Sorted Neighborhood for Efficient Duplicate Detection**, *IEEE Int. Conf. Data Min. Work. ICDMW*, pp. 202–209, 2017. https://doi.org/10.1109/ICDMW.2016.0036
37. S. Ranjitha, P. Sudhakar, and K. S. Seetharaman, **A Novel and Efficient De-duplication System for HDFS**, *Procedia Comput. Sci.*, vol. 92, pp. 498–505, 2016.
38. M. Li, H. Zhang, Y. Wu, and C. Zhao, **Prefetch-Aware Fingerprint Cache Management For Data Deduplication Systems**, *Front. Comput. Sci.*, pp. 1–16, 2018.
39. R. V Gode and R. Dalvi, **An Effective Storage Management In A Twin Cloud Architecture Using An Authorized Deduplication Technique**, *2017 IEEE Int. Conf. Power, Control. Signals Instrum. Eng.*, pp. 796–801, 2017.
40. S. Wu, K. C. Li, B. Mao, and M. Liao, **DAC: Improving storage availability with Deduplication-Assisted Cloud-of-Clouds,** *Futur. Gener. Comput. Syst.*, vol. 74, pp. 190–198, 2017.
41. [41] P. Sobe, **Combination of Data Deduplication and Redundancy Techniques in Distributed Systems**, pp. 4–7, 2016.
42. A. Sahaya Jenitha and V. Sinthu Janita Prakash, **An Effective Content-Based Strategy Analysis for Large-Scale Deduplication Using a Multi-level Pattern-Matching Algorithm**, vol. 2. Springer Singapore, 2017. https://doi.org/10.1007/978-981-13-1747-7_23
43. Y. Zhang *et al.*, **A Fast Asymmetric Extremum Content Defined Chunking Algorithm for Data Deduplication in Backup Storage Systems**, *IEEE Trans. Comput.*, vol. 66, no. 2, pp. 199–211, 2017.
44. X. He, H. Wang, K. Zhou, P. Huang, and P. Zhang, **Resemblance and mergence based indexing for high performance data deduplication**, *J. Syst. Softw.*, vol. 128, pp. 11–24, 2017. https://doi.org/10.1016/j.jss.2017.02.039
45. W. Xia, H. Jiang, D. Feng, and L. Tian, **DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads,** *IEEE Trans. Comput.*, vol. 65, no. 6, pp. 1692–1705, 2016. https://doi.org/10.1109/TC.2015.2456015