ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

REVIEW ON DATA DE-DUPLICATION IN CLOUD COMPUTING

[1] Mrs. K. Geetha, [2] Dr. A. Vijaya,
[1] Guest Lecturer, [2] Assistant professor and Head,
[1, 2] Department of Computer Applications,
[1, 2] Government Arts College (Autonomous),
[1, 2] Salem-7.

ABSTRACT - Cloud computing is a rising innovation for giving foundation as an administrations to cloud clients. The foundation as an assistance depends on virtualization where it dispenses the virtual machine to client through web. Virtual machine is a visitor machine runs in the earth of host machine. VMI are utilized to buy VM examples to run on virtual machine in cloud stages. The capacity of huge number of VMI and provisioning stays testing issue. Data De-duplication assumes a significant part in disposing of this repetitive data and diminishing the capacity utilization. Its primary expects is the manner by which to decrease more copies productively, eliminating them at rapid and to accomplish great copy expulsion proportion. Numerous components have been proposed to meet these goals. In this paper we review on different data de-duplication in cloud computing.

Keywords: [cloud computing, de-duplication, virtual machine, data compression.]

1. INTRODUCTION

Cloud Computing administrations has gigantic measure of computational resources on demand by utilizing pay-per-use. It provides computational resources with the assistance virtualization innovation. It has the capability to store data and run applications. It empowers us to get to all the archives and run applications from anyplace on the planet through the Internet. Cloud Computing empowers arranges admittance to a mutual pool of configurable Computing resources. Under Cloud Computing, numerous clients have option to use on its own worker to recover and refresh their data.

Data de-duplication is a procedure data de-duplication is utilized to store single case of repetitive data and dispenses with the copy data in datacenter. It is utilized to decrease the size of datacenter and lessen the replications of data that were copied on cloud. The de-duplication measure assists with eliminating any square or document that are not remarkable and store in littler gathering of squares. The essential strides for data de-duplication measure are

The documents are changed over into little portions

Then new and existing data are checked for excess

Metadata are refreshed and sections are packed

Duplicate data are deleted and check the data uprightness.

They are two strategies used to break the record into fragments, called, fixed size chunking and variable size chunking. The fixed size chunking will parts the first record into hinders in same size. The variable size chunking is finished with Rabin unique mark on document substance and it likewise detects the limits inside the record. Contrasting both VMI designs generally utilize fixed size chunking which is acceptable in de-duplication. Preferences are decreased capacity, effective volume replication, versatility and IO execution.

2. LITERATURE REVIEWS

Chi Yang and Jinjun Chen [1] proposed a novel scalable data compression based on similarity calculation among the partitioned data chunks with Cloud computing. A similarity model was developed to generate the standard data chunks for compressing big data sets. Instead of compression over basic data units, the compression was conducted over partitioned data chunks. The MapReduce programming model was adopted for the algorithms implementation to achieve some extra scalability on Cloud. With the real meteorological big sensing data experiments on this U-Cloud platform, it was demonstrated that this proposed scalable compression based on data chunk similarity significantly improved data compression performance gains with affordable data accuracy loss. The significant compression ratio brought dramatic space and time cost savings. With the popularity of Spark and its specialty in processing streaming big data set [1].

Youjip Won, Kyeongyeol Lim, and Jaehong Min [2] proposed a novel multicore chunking algorithm, MUCH, which parallelizes the variable size chunking. To date, most of the existing works on deduplication focus on expediting the redundancy detection process, while less attention has been paid on how to make the file chunking faster. That proposed a multicore chunking algorithm, MUCH, which guarantees Chunking Invariability. They developed a performance model to compute the segment size that maximizes the chunking bandwidth while minimizing the memory requirement [2]. Through extensive physical experiments, it showed that the performance of MUCH scales linearly with the number of cores. In quad-core CPUs, MUCH brings a 400 percent performance increase when the storage device is sufficiently fast. The benefits of MUCH are evident when it chunks large files, e.g., tar images of file system snapshot, at high performance storage. MUCH successfully increases the chunking performance with the factor being as high as the number of available CPU cores without any additional hardware assistance.

Xu Zhang and Yue Cao [3] propose a fully distributed ICN-based caching scheme for content objects in Radio Access Network (RAN) at eNodeBs. Such caching scheme operates in a cooperative way within neighborhoods, aiming to reduce cache redundancy to improve the diversity of content distribution [3]. The caching decision logic at individual eNodeBs allows for adaptive caching, by considering dynamic context information, such as content popularity and availability. The efficiency of the proposed distributed caching scheme is evaluated via extensive simulations, which show great performance gains, in terms of a substantial reduction of backhaul content traffic as well as great improvement on the diversity of content distribution, etc.

Chuanshuai Yu, Chengwei Zhang, Yiping Mao, Fulu Li [4] presented the leap-based CDC algorithm and added a secondary condition to it to reduce the computing overhead and maintain the same deduplication ratio. This algorithm satisfies both the content defined condition and the equal probability condition [4]. The leap-based CDC algorithm with or without a secondary condition can significantly reduce the computing overhead while maintaining the same deduplication ratio. To resolve the technique issue of not being able to use the rolling hash in the new algorithm, they introduced the pseudo-random transformation to replace the role of rolling hash.

Daniel Posch, Hermann Hellwagner and Peter Schartner [5] proposed a framework for multimedia delivery in VoD use cases. The concepts of CCN, DASH and BE to create dynamic adaptive encrypted chunks of data, which can be inherently cached in the network [5]. The evaluation results show that network inherent caching can increase the efficiency of multimedia delivery. However, the usage of adaptive concepts leads to the question of how to synchronize clients to exploit the advantage of cached data perfectly. Finding a solution to this issue would enhance the framework greatly.

Chi Yang and Jinjun Chen [6] proposed a novel scalable data compression based on similarity calculation among the partitioned data chunks with Cloud computing. A similarity model was developed to generate the standard data chunks for compressing big data sets. Instead of compression over basic data units, the compression was conducted over partitioned data chunks [6]. The MapReduce programming model was adopted for the algorithms implementation to achieve some extra scalability on Cloud. With the real meteorological big sensing data experiments on this U-Cloud platform, it was demonstrated that this proposed scalable compression based on data chunk similarity significantly improved data compression performance gains with affordable data accuracy loss. The significant compression ratio brought dramatic space and time cost savings.

C. Goktug Gurler , S. Sedef Savas , and A. Murat Tekalp [7] proposes two modifications to the Torrent protocol, variable chunk size and adaptive scheduling window, for efficient, error-resilient, adaptive P2P streaming of scalable video. The proposed modifications yield superior results in terms of number of decoded frames, hence superior quality of experience, in P2P video streaming [7]. The proposed modifications to BitTorrent for video streaming yield superior results both in terms of chunks exchanged between leechers (P2P activity) and the number of decoded frames (superior quality of experience). In the variable size chunk tests show that the number of decodable frames has significantly increased, improving the PSNR and the QoE. In the variable size chunk tests show that the proposed adaptive windowing allows better scalability against increasing number of leechers. Therefore, with the proposed modifications, the peers would receive video at a higher quality and the content providers (seeders) have lower cost of bandwidth.

Haiying Shen and Jin Li [8] propose a DHT-aided chunk-driven overlay for P2P live streaming that targets higher scalability, better availability, and low latency. The design has three main components: a two-layer hierarchical DHTbased infrastructure, a chunk sharing algorithm, and a video provider selection algorithm. The hierarchical DHTbased infrastructure offers high scalability. The chunk sharing algorithm provides service for chunk index collection and discovery, which guarantees high availability. The provider selection algorithm enables full utilization of system bandwidth. As a result, the overlay can provide high-quality video streaming. They also propose a centralized and simplified decentralized provider selection algorithm. DCO is superior to tree-based systems in dealing with churn and mesh-based systems in bandwidth consumption and latency. More importantly, it can flexibly take full advantage of system bandwidth by dynamically matching chunk requesters and providers [8]. The experimental results show that DCO improves the performance of the mesh-based systems (pull and push) and tree-based systems, in term of scalability, availability, latency, and overhead. The experimental results also confirm the importance of providing incentives to encourage nodes to serve as coordinators in the DHT-based infrastructure and the importance of selecting chunk providers with enough bandwidth in chunk delivery.

Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long and Mark Lillibridge [9] introduced a new method, Extreme Binning, for scalable and parallel deduplication, which is especially suited for workloads consisting of individual files with low locality. Existing approaches which require locality to ensure reasonable throughput perform poorly with such a workload. Extreme Binning exploits file similarity instead of locality to make only one disk access for chunk lookup per file instead of per chunk, thus alleviating the disk bottleneck problem. It splits the chunk index into two tiers resulting in a low RAM footprint that allows the system to maintain throughput for a larger data set than a flat index scheme. Partitioning the two-tier chunk index and the data chunks is easy and clean. In a distributed setting, with multiple backup nodes, there is no sharing of data or index between nodes. Files are allocated to a single node for deduplication and storage using a stateless routing algorithm – meaning it is not necessary to know the contents of the backup nodes while making this decision. Maximum parallelization can be achieved due to the one file-one backup node distribution. Backup nodes can be added to boost throughput and the redistribution of indices and chunks is a clean operation because there are no dependencies between the bins or between chunks attached to different bins [9]. The autonomy of backup nodes makes data management tasks such as garbage collection, integrity checks, and data restore requests efficient. The loss of deduplication is small and is easily compensated by the gains in RAM usage and scalability.

Chu-Hsing Lin, Chen-Yu Lee, Yi-Shiung Yeh, Hung-Sheng Chien and Shih-Pei Chien [10] generalized the SHA family as SHA-mn that takes arbitrary length message as input to generate a message digest with required length. They modify each of the steps of SHA-mn as generalized version that contains padding and parsing; setting the initial hash values, constants, Boolean expressions and functions and message schedule; initializing the eight working variables and for-loop operation; and, computing the ith intermediate hash values. Further, the LHV problem that does not exist in the original SHA standard is solved. Owing to security considerations, SHA-mn is generalized based on the rules of SHA family design [10]. Although many may not agree the method for calculating complexity according to the birthday paradox as the collision of full SHA-1 has been found in 2005, the design of SHA is improved. Efficient ways of finding collisions of SHA-256 remain the focus of many researchers to date.

Some data de-duplication in cloud computing methods are tabulated below

Author's Name	Proposed Method	Merits	Demerits
Amdewar Godavari , Chapram Sudhakar, and T. Ramesh (2020)	Hybrid deduplication system (HDS), a block-based partial deduplication system with similarity-based indexing	The framework has performed reliably better in diminishing the metadata overhead and expanding the normal portion length for every one of the three arrangements of I/O follow data.	The HDS is designed to be utilized as a deduplication framework for a solitary stockpiling node just, it cann't to help different capacity nodes or conveyed stockpiling

			frameworks.
Weimin Lang, Weiguo Ma, Yin Zhang, Shengyun Wei , Han Zhang (2020)	Edge-IoT encrypted data deduplication scheme	It improves computational effectiveness and data security with the assistance of secure data deduplication and edge computing.	Security difficulties in edge computing are high because of enormous measure of data.
Yinjin Fu, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen (2017)	AppDedupe, an application-aware scalable inline distributed deduplication framework in cloud environment	It beats the amazingly exorbitant and ineffectively adaptable stateful tight coupling plan in the cluster wide de-duplication proportion however just at a somewhat higher sys-tem overhead than the exceptionally versatile loose coupling schemes. It essentially improves the stateless free coupling plans in the group wide viable de-duplication proportion while holding the last's high framework adaptability with low overhead.	The primary disadvantages of inline de-duplication is the overhead presented in the inactivity of compose demands, as the greater part of the handling is done in the compose way
Jia, G., Han, G., Rodrigues, J., Lloret, J., & Li, W. (2015)	Proposed a coordinate memory deduplication and partition approach named CMDP	CMDP neither adjusting equipment nor including extra data, it is truly light plan. CMDP can proficiently improve execution then oblige more virtual machines simultaneously.	Performing page examinations is confined into a similar grouping, never surpassing to various arrangements, which will do numerous pointless correlations
Helei Cui, Huayi Duan, Zhan Qin, Cong Wang, and Yajin Zhou**(2019)**	Proposed SPEED, a secure and generic computation deduplication system in the context of	SPEED improves execution by up to multiple times. The source code is accessible on GitHub for open use.	This procedure can't change dynamic investigating the underlying calculations during its

	Intel SGX		runtime.
Zhichao Yan, Hong Jiang, Yujuan Tan, Stan Skelton and Hao Luo (2019)	Proposed Z-Dedup, a novel deduplication system	The design and usage of Z-Dedup additionally address a potential security weakness because of customer site compression, just as bundles created by both non-strong and strong compression techniques. Z-Dedup model decreases repetitive data in compacted bundles than conventional deduplication framework.	Z-Dedup model turns into an unpredictable and less-proficient cycle if reinforcements have been running for some time.

Amdewar Godavari , Chapram Sudhakar, and T. Ramesh (2020) [11] proposed and implement a hybrid deduplication system (HDS), a block-based partial deduplication system with similarity-based indexing. The proposed framework applies deduplication out of sight to decrease the idleness and furthermore targets decreasing the data fracture. It applies likeness based indexing to decrease high number of metadata queries emerging out of irregular access examples of the solicitations. HDS for essential outstanding burdens is recreated in the Linux condition utilizing three distinct sorts of FIU follows, and the adequacy of the framework is contrasted and full deduplication dependent on the boundaries—metadata access overhead, normal portion length, and reaction time.

Weimin Lang, Weiguo Ma, Yin Zhang, Shengyun Wei , Han Zhang (2020) [12] designs an edge-IoT encrypted data deduplication scheme supporting dynamic ownership management and security assurance, which accomplishes fine-grained admittance control by client level key and update system and decreases the correspondence overhead significantly on the grounds that the possession update is performed by the cloud worker. In IoT condition, mass data delivered by the IoT devices are excess. These repeat data devour additional transmission transfer speed and extra room. A direct way to deal with transfer speed and capacity sparing is to receive data deduplication.

Yinjin Fu, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen (2017) [13] proposed AppDedupe, an application-aware scalable inline distributed deduplication framework in cloud environment, to address this difficulty by misusing application mindfulness, data similitude and region to enhance circulated deduplication with between node two-layered data directing and intra-node application-mindful deduplication. It initially apportions application data at record level with an application-mindful directing to keep application region, at that point doles out comparable application data to a similar stockpiling node at the super-lump granularity utilizing a handprinting-based stateful data steering plan to keep up high worldwide deduplication effectiveness, in the interim adjusts the outstanding burden across nodes.

AppDedupe manufactures application-mindful similitude lists with super-lump impressions to speedup the intra-node deduplication measure with high productivity.

Jia, G., Han, G., Rodrigues, J., Lloret, J., & Li, W. (2015) [14] proposed a coordinate memory deduplication and partition approach named CMDP to diminish memory necessity and impedance all the while for improving execution in virtualization. Besides, CMDP embraces a lightweight page conduct based memory deduplication approach named BMD to diminish purposeless page correlation overhead in the interim to detect page sharing open doors productively. Furthermore, a virtual machine based memory segment called VMMP is added into CMDP to diminish impedance among virtual machines. As per page shading, VMMP dispenses novel page hues to applications, virtual machines and hypervisor.

Helei Cui, Huayi Duan, Zhan Qin, Cong Wang, and Yajin Zhou (2019) [15] proposed SPEED, a secure and generic computation deduplication system in the context of Intel SGX. It permits SGX-empowered applications to identify excess calculations and reuse calculation results, while securing the confidentiality and uprightness of code, information sources, and results. To augment the advantage of calculation deduplication, we design a cross-application deduplication plot, engaging various applications to safely use the mutual outcomes as long as they perform identical calculations. To facilitate the utilization of SPEED, we execute a completely practical model and provide a compact and expressive API for developers to deduplicate rich calculations with negligible exertion, as not many as 2 lines of code for each capacity call.

Zhichao Yan, Hong Jiang, Yujuan Tan, Stan Skelton and Hao Luo (2019) [16] proposed Z-Dedup, a novel deduplication system that is able to detect and remove redundant data in compressed packages, by misusing some key invariant data embedded in the metadata of packed bundles, for example, record based checksum and unique document length data. Assessments demonstrates that Z-Dedup can essentially improve both space and transmission capacity effectiveness over conventional methodologies by killing 1.61% to 98.75% excess data of a compacted bundle dependent on our gathered datasets, and significantly more extra room and transfer speed are relied upon to be spared after the capacity workers have amassed more packed substance.

CONCLUSION

In this paper, the survey on de-duplication with cloud computing work with different algorithms tabulated them on the basis of algorithm, target standards, condition to which the works being performed. From the writing review obviously, part of work had been done as of now in de-duplication yet at the same time it needs further development. (i.e) Deduplication need to build up with elevated level security and least space wastage.

REFERENCES

[1] Chi Yang and Jinjun Chen, “A Scalable Data Chunk Similarity based Compression Approach for Efficient Big Sensing Data Processing on Cloud”, IEEE Transactions on Knowledge and Data Engineering; Print ISSN: 1041-4347; Electronic ISSN: 1558-2191; CD-ROM ISSN: 2326-3865; June 1 2017

[2] Youjip Won, Kyeongyeol Lim, and Jaehong Min, “MUCH: Multithreaded Content-Based File Chunking”,IEEE Transactions on Computers (Volume: 64, Issue: 5, May 1 2015), 14 May 2014; 14 May 2014

[3] Xu Zhang and Yue Cao, “A Cooperation-Driven ICN-based Caching Scheme for Mobile Content Chunk Delivery at RAN”, 13th International Wireless Communications and Mobile Computing Conference (IWCMC); IEEE; **ISSN:** 2376-6506,2017

[4] Chuanshuai Yu, Chengwei Zhang, Yiping Mao, Fulu Li, "Leap-based Content Defined Chunking --- Theory and Implementation", IEEE; Print ISSN: 2160-195X; Electronic ISSN: 2160-1968,31st 2015

[5] Daniel Posch, Hermann Hellwagner and Peter Schartner, "On-Demand Video Streaming based on Dynamic Adaptive Encrypted Content Chunks", IEEE International Conference on Network Protocols (ICNP); ISBN: 978-1-4799-1270-4; ISSN: 1092-1648,21st2013

[6] Chi Yang and Jinjun Chen, "A Scalable Data Chunk Similarity based Compression Approach for Efficient Big Sensing Data Processing on Cloud", IEEE Transactions on Knowledge and Data Engineering (Volume: 29, Issue: 6, June 1 2017);ISSN: 1041-4347; 2017

[7] C. Goktug Gurler1 , S. Sedef Savas2 , and A. Murat Tekalp3, "VARIABLE CHUNK SIZE AND ADAPTIVE SCHEDULING WINDOW FOR P2P STREAMING OF SCALABLE VIDEO", IEEE International Conference on Image Processing; Print ISSN: 1522-4880; Online ISSN: 1522-4880; Electronic ISSN: 2381-8549; 19th2012

[8] Haiying Shen and Jin Li , "A DHT-Aided Chunk-Driven Overlay for Scalable and Efficient Peer-to-Peer Live Streaming", IEEE Transactions on Parallel and Distributed Systems (Volume: 24, Issue: 11, Nov. 2013); Print ISSN: 1045-9219; Electronic ISSN: 1558-2183; CD-ROM ISSN: 2161-9883; 22 October 2012

[9] Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long and Mark Lillibridge, "Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup", IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems; Print ISBN: 978-1-4244-4927-9; CD-ROM ISBN: 978-1-4244-4928-6; 2009

[10] Chu-Hsing Lin, Chen-Yu Lee, Yi-Shiung Yeh, Hung-Sheng Chien and Shih-Pei Chien, "Generalized Secure Hash Algorithm: SHA-X",IEEE EUROCON - International Conference on Computer as a Tool; Print ISBN: 978-1-4244-7486-8;CD-ROM ISBN: 978-1-4244-7485-1; 2011

[11]. Amdewar Godavari , Chapram Sudhakar, and T. Ramesh (2020), "Hybrid Deduplication System—A Block-Level Similarity-Based Approach", **DOI:** 10.1109/JSYST.2020.3012702, **Print ISSN:** 1932-8184, IEEE.

[12]. Weimin Lang, Weiguo Ma, Yin Zhang, Shengyun Wei , Han Zhang (2020), "EdgeDeup: an Edge-IoT Data Deduplication Scheme with Dynamic Ownership Management and Privacypreserving", **DOI:** 10.1109/ITNEC48623.2020.9085119, **Electronic ISBN:** 978-1-7281-4390-3, IEEE.

[13]. Yinjin Fu, Nong Xiao, Hong Jiang, Guyu Hu, and Weiwei Chen (2017), "Application-Aware Big Data Deduplication in Cloud Environment", **DOI:** 10.1109/TCC.2017.2710043, **Electronic ISSN:** 2168-7161, IEEE.

[14]. Jia, G., Han, G., Rodrigues, J., Lloret, J., & Li, W. (2015), "Coordinate Memory Deduplication and Partition for Improving Performance in Cloud Computing", **DOI:** 10.1109/TCC.2015.2511738, **Electronic ISSN:** 2168-7161, IEEE.

[15]. Helei Cui, Huayi Duan, Zhan Qin, Cong Wang, and Yajin Zhou (2019), "SPEED: Accelerating Enclave Applications via Secure Deduplication", **DOI:** 10.1109/ICDCS.2019.00110, **Electronic ISBN:** 978-1-7281-2519-0, IEEE.

[16]. Zhichao Yan, Hong Jiang, Yujuan Tan, Stan Skelton and Hao Luo (2019), "Z-Dedup:A Case for Deduplicating Compressed Contents in Cloud", **DOI:** 10.1109/IPDPS.2019.00049, **Electronic ISBN:** 978-1-7281-1246-6, IEEE.