Archives of Materials Science and Engineering
Volume 104
Issue 1
July 2020
Pages 5-18
International Scientific Journal
published monthly by the
World Academy of Materials
and Manufacturing Engineering

DOI: 10.5604/01.3001.0014.3864

Collection of hospital wastewater data using deduplication approaches

N.A. Khan [a,*], G.R. Sinha [b], S. Ahmed [a], A. Feshchenko [c], F. Changani [d], A. Qureshi [e], M.A. Mazhar [a], I. Neklonskyi [c]

[a] Civil Engineering Department, Jamia Millia Islamia, New Delhi, India
[b] Adjunct Professor, IIIT Bangalore & Professor, MIIT Mandalay, Myanmar, India
[c] National University of Civil Defence of Ukraine, Kharkiv, Ukraine
[d] Department of Environmental Health Engineering, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[e] Environmental Engineering Department, College of Engineering A13, Imam Abdulrahman Bin Faisal University, Dammam 34212, Saudi Arabia
* Corresponding e-mail address: er.nadimcivil@gmail.com
ORCID identifier: https://orcid.org/0000-0003-4366-9639 (N.A.K.)

ABSTRACT

Purpose: This investigation aims to study the various approaches currently used to reduce the load on computer servers in order to better manage data on hospital wastewater treatment and solid waste generation.

Design/methodology/approach: This manuscript investigates the taxonomies of deduplication procedures based on literature and other data sources, thereby presenting its classification and its challenges in detection.

Findings: Based on the literature survey of deduplication techniques, the method of deduplication dispensed on cloud gadget devices has been found to be a promising research challenge. The gaps discussed include a reduction in storage space, bandwidth, type of disks used, and expenditure on energy usage and heat emissions when implementing these strategies. The art work on a scalable, robust, green and allocated approach to deduplication for a cloud gadget will remain of interest in destiny.

Research limitations/implications: Considerable attention is focused on the deduplication due to efficient, extensive storage system.

Practical implications: This research paper will be useful to identify deduplication techniques which are nowadays used in different hospital wastewater data collection systems and put significant proposals for further improvements in deduplication.

Originality/value: This manuscript portrays a broader assessment of the available literature for data duplication along with the classification of different methods for the data storage used in the different level of storage of hospital wastewater data collection.

Keywords: Statistics, Hospital wastewater, Medical waste, Deduplication, Approaches, Servers

Reference to this paper should be given in the following way:

N.A. Khan, G.R. Sinha, S. Ahmed, A. Feshchenko, F. Changani, A. Qureshi, M.A. Mazhar, I. Neklonskyi, Collection of hospital wastewater data using deduplication approaches, Archives of Materials Science and Engineering 104/1 (2020) 5-18.

DOI: https://doi.org/10.5604/01.3001.0014.3864

METHODOLOGY OF RESEARCH, ANALYSIS AND MODELLING

1. Introduction

The waste' amount generated and their quality, as well as the mechanisms for their treatment, especially in medical institutions, have become a matter of concern [1-5]. As know, the biomedical waste management system in many countries is not perfect [3,6,7]. Concerns about potential exposure to pharmaceutical contaminants, especially for water consumption and bathing, have prompted numerous studies to clarify the relationship between specific health effects and the chemical constituents of drinking water from water supplies [8-10], as well as the possibility of reducing disease risk by timely solution of this problem [10,11].

For economic and environmental reasons, wastewater should be post-treated for reuse; Waste quantities should be reduced by implementing a comprehensive program to oversee environmentally sound separation methods such as source separation, storage, transport, handling and disposal; solid organic waste should be composted through an organic waste converter and reused as manure; rainwater harvesting systems are required; Hybrid hot water generators such as solar water heaters are recommended; etc. – this is a small list of recommendations that can help reduce human exposure to health care waste, improve the environment and sustainably manage the environment. However, all this is extremely difficult to achieve without automated systems.

Today, the automated control systems' development for technological processes follows the path of improving monitoring and data transmission, as well as the introduction of microprocessor control systems at local facilities. Dispatching of technological processes is mainly developing along the way of collecting and presenting data on the operating modes of systems and signalling the occurrence of off-design values of indicators and characteristics. According to many authors, improving the efficiency and economy of waste management systems based on the development of information monitoring and management systems is an urgent direction in the modernization of facilities, such as industrial enterprises, waste dumps, organizations, institutions, including medical facilities, etc. [12-14]. At the same time, the development of geoinformatics and geoinformation technologies has provided a powerful tool for managing territories [13-16]. However, for the effective use of geographic information systems (GIS), it is necessary to develop modern methods for analyzing geoinformation data using the latest software products in order to solve specific environmental problems, including for the collection, systematization and analysis of multi-factor information in the aspect of providing information support for managerial decision-making, conducting environmental monitoring and information basis for predictive modelling of the development of the territory [13].

Thus, one of the most important components of the waste management system is the data archiving and storage subsystem. This subsystem stores and archives information necessary for solving various problems, such as obtaining statistical information about pollution of rivers and lakes, about the characteristics of medical wastewater, predicting emergency situations, etc [13].

In the recent years with the development of cloud computing there is an increase in internet usage, acceptance of smartphones and social media platforms which are considerably increasing the information saved in clouds. Global statistics organization (IDC) reported in 2011 that 35ZB size data might be generated and copied for the whole world [17].

That is why this investigation aims to study the various approaches currently used to reduce the load on computer servers in order to better manage data on hospital wastewater treatment and solid waste generation.

2. Material and methods

This manuscript investigates the taxonomies of deduplication procedures based on literature and other data sources, thereby presenting its classification and its challenges in detection.

Subsequent to making an analysis on the cutting-edge studies in deduplication strategies, there was a need for a thorough evaluation of the text, therefore:

(i) The functions and need of deduplication strategy to enhance the overall implementation of a large memory system has been explained along with its advantages and disadvantages.

(ii) The present deduplication strategies have been arranged on the basis of gadget, point of utility and stage completely. In addition, deduplication approach has been characterized dependent on literature substance, photo and videocassette. The consideration of content deduplication procedure and their criticalness have been clarified. An assessment of content, picture and video-based deduplication and their different scientific categorizations are introduced. Consequently, this literature evaluates writing and gives an overall view on deduplication strategies.

(iii) Upcoming research guidelines within the area of deduplication were featured for analysts of the scholarly world and enterprise.

This paper alludes outstanding journals and legal disputes of different gatherings and information of numerous research offices. This manuscript has several segments:

(i) Segment that portrays foundation, development of deduplication, excess information markdown procedures and its examination, benefits and bad marks of deduplication strategies;
(ii) Segment that bears a survey approach, inquire about inquiries and an investigations strategy used to choose and evaluate the past research material. It additionally manages system meant for assessment and conversation of research texture and statistics deduplication;
(iii) Segment that makes a specialty to fame of introducing a common deduplication way, scientific categorization of deduplication methods. So, also, the procedures are arranged fundamentally dependent on literary substance, picture and video;
(iv) Segment that conveys an exchange on open demanding circumstances and upcoming research rules inside the territory of deduplication strategies;
(v) Segment that ends the study and carries conclusions and tips for potential research which can be classified as lossless statistics.

3. Results and discussion

Corporations are dealing with issues in putting away and handling a lot of statistics volumes in order to ensure that the improvement of unshakable quality and availability, as well as data recovery after loss, is ensured. The information is usually copied on different gadgets; the greater part of those copied realities applies an additional heap at the capacity machine in expressions of additional zone and transfer speed to switch the copied information at the system. Management of a green information gadget is dangerous and realities deduplication approach is considered as an empowering period for green gadget of gigantic records. Deduplication strategy is a unique reality compacting technique to eradicate excess statistics and drop off grid transference cost and storage volume within the cloud gadget frameworks. The procedures situate out the copy records, spare handiest one imitation of the information and deliberately utilize legitimate tips for copied realities. Deduplication tends to the developing requirement for the capacity ability [18]. Several cloud storage organizations similar to Amazon S3 and backing solutions, for example, drop box and Memo pal are using data deduplication strategies to enhance gadget productivity [19].

The deduplication methods are facts kind of distinctive, and special techniques are utilized on exceptional forms of statistics consisting of content, picture and video statistics. Every one of the three sorts of records has excellent capacity positions and certain characteristics. Basically, dependent on sort of data, deduplication systems have explicit techniques to find and put off copy records [20,21]. Along these lines, kind of facts is significant for the improvement of deduplication procedures. The layout of records is imperative for considering, obtaining and coordinating the facts. Bit-stage coordinating is needed to discover duplication in workable records. The strategies to examine copies in textual content, picture and video have distinct techniques because of various configurations of statistics. To get high records accessibility a minimum figure of facts duplicates known as replication aspect are maintained in a massive dispersed gadget device. To reduce gadget necessity, storing cost, calculation and power any reproduction statistics above replication factor is evacuated. Because of these widespread advantages to enterprise, deduplication techniques for a massive dispensed storage structures received thrust in scholarly world and enterprise. Nonetheless, those strategies are going through difficulties because of performance and adequacy of facts coordinating methods. The analysts in the scholarly community and enterprise are operating to expand effective dispensed deduplication strategies. Discern correspond to facts deduplication in which replica portions of same information are diminished to specific sections of records.

The any document is part into constant-or variable size portions. With deduplication system, just a single duplicate of every section is saved [22] and guidelines are utilized for multiplication portions. If deduplication engine encounters a bit of facts that this is as of now stocked up someplace inside the capacity device, it spares pointer inside the facts duplicate region that leads lower back to bona fide duplicate. It enables in releasing up the obstructions inside the gadget device, hence releasing the reminiscence space. Figure 1 affords the deduplication method [23-25].

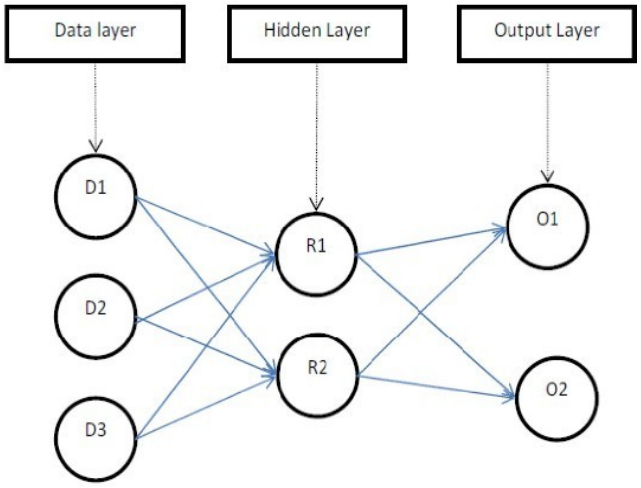

Fig. 1. Deduplication method

3.1. Development of deduplication, excess information markdown procedures and its examination

Development of deduplication

Space-green method, i.e. smart delta compression techniques, was introduced later in the 1990s to concentrate on compressing completely similar records or nearly identical lumps. In the early 2000s, the term deduplication of facts was coined to support the big storage systems at a stage of towering granularity [17]. It expels each between document and intra-record-degree excess over gigantic datasets over a few dispensed gadget servers dislikes conventional data compression strategies that discard repetition over small organization of records primarily based on excess intra-document. Distinguishing the copies is calculated by crypto-graphic hashes of each text or lump. These strategies have also been updated on mixed media substance in 2008, and the use of highlight extraction and hashing methods looks at the replica multimedia content by estimating the resemblance of pix or frames in recordings. Such methods for deduplication of records came to life to explain the problem found with the creation of data length within storage structures [26,27-29].

Hufman coding and dictionary coding are conventional ways to compress bytes or strings while deduplication techniques decline log or bit redundancy. Redundant fact-reduction methods originated in the 1950s, and were increasingly encountered in the 1990s via delta compression and loss-related compression strategies. Earlier, in 2000, computational deduplication strategies accompanied by optical deduplication came here.

Statistics warehouses which are archives of facts accumulated from several statistics assets comprises the inspiration of mostly current decision assisting packages and CRM (client dating control). Accuracy of choice guide evaluation on statistics warehouses is vital due to the fact important commercial enterprise choices are inspired by such evaluation. Impartial and nearly incompatible requirements can be observed via facts sources as they are unbiased. In maximum circumstances, doubt consequences are a combination of pages containing diverse things that percentage the similar nomenclature. In a best recovery gadget, a client could genuinely enter an entity or idea call and receive seek outcomes clustered in keeping with the extraordinary entities/standards that percentage that call. One strategy to improve such gadget is to encompass extra records in the listed files [30]. In general, businesses befall privy to rational specific disparities or inconsistencies whilst gathering information from distinct assets to appoint an information warehouse. Such troubles fit in to the group known as facts heterogeneity. Foolish replication of the facts occurs while data from various statistical resources which keep the information overlapping is incorporated. However, the facts obtained in the facts warehouse from outside assets have errors such as spelling errors, inconsistent conventions across statistical sources, neglected fields, and so on.

Arriving statistics tuples outdoor assets want rationale and change for the provision of excessive statistical quality. The 'errors-unfastened' manner in the facts warehouse is suggested through high-quality statistics. Record cleaning techniques are crucial to the improvement of the first class of records. Records mining techniques and techniques using software or algorithms called facts mining equipment, successfully mined and offered artistic and beneficial evaluation. Extrapolative and descriptive are two kinds of obtaining images. Descriptive representations such as Clustering, Summarization, Affiliation Rule, Series Discovery and many others determine the homes of the analysed records by means of tracing patterns or associations' records. Clustering categorizes a set of substance at a subsets' quantity known as clusters, so that analogous cluster' gadgets are rather comparable, and gadgets of different clusters vary by cluster [31-34]. Clustering is renowned as a basic way to consolidate and sum up the data, since it may give an abstract of the saved records [22,23]. Predictive versions, such as categorization, regression, time series evaluation, prediction and so on, decide on the use of recognized values for unknown information type technique which comprises many choice-

hypothesis strategies for spotting records is a broad variety studies subject capable of handling an expansive records' group than regression. The classification set of rules constructs a version by way of getting to know from the training set.

New items be categorized by means of model described above, Strict classification algorithms include the k Nearest Neighbours algorithm (kNN), the Naive Bayes algorithm, the neural network [18,19,35-37], and the support vector mechanism (SVM). The particularly accepted type algorithm referred to as kNN reveals desirable performance uniqueness and is widely used in a number of different applications such as three-dimensional item interpretation, text material primarily based on picture recovery [23,27,29, 38,39], facts (evaluation of entropies and deviations) [18-20,22,37,40]. Information cleansing, also known as records cleaning or scrubbing, decorate the quality of statistics by using figuring out and removing mistakes and contradiction from the records [12,26,41,42]. It ambitiously enhances the overall statistics compatibility by focusing on abolition of alteration in data sets and reducing facts replication. Record replica, unnoticed value, record and field resemblances and replica elimination are identified with the aid of present-day information removing strategies. Recognition of other or numerous facts that characterize one wonderful real global entity or item is executed with the aid of the duplicate document detection technique.

Repetitive data minimisation methods

These repetitive statistics minimizing Procedures have been developed to manage the growing quantity of virtual records and to select excess as of byte to string stage and from chew to report degree. The company and their development of redundant statistics reduction techniques are similarly. Compression of information is a bit-price discount technique that represents data in a compact form. This decreases the storage area required and attempts to locate redundant information. Compression of statistics is generally defined by lossless compression procedure [18,19,35,43]. In lossless compression method, the specific unique information is recreated from the compacted information. Lossy compression decreased the details by pointing out meaningless statistics such as compression of jpeg images. This recreates a first act prediction. Data on films and sounds allow using of lossy compression techniques [20]. In this section, critical heritage associated with redundant techniques of information reduction is provided, showing the advancement of each traditional approach to compression of lossless data, delta compression strategies and information deduplication strategy. It shows a scientific categorization of the considerable number of strategies and their growth [31-33,44].

The techniques for lossless data compression. The compression of information became coined with the use. The strategies such as entropy encoding, run duration encoding and word reference encoding are strategies for lossless compression of information. The character series was symbolized through a small amount of sequences in bit. In these strings enormous quantities of redundant statistics are created and expelled in such designs of facts:

(i) Byte stage: The premature records compression techniques utilize entropy encoding to turn into aware of redundancy at byte degree. Huffman encoding and mathematics coding are the entropy encoding' styles used to signify often happening sample with fewer bits. Huffman coding advanced with the aid of D.A. Huffman applies frequency-looked after binary tree to produce the best prefix code. It provides replacement of constant length codes with variable-length codes [20,21,45,46]. Usually used images are presented as short coding. Elias in 1960 developed arithmetic coding [25]. It encrypts the entire message into a fixed range of floating-point values.

(ii) String Level: String Level approach has been intended to review & eradicate recurring strings. Among the byte level approaches, there are two key approaches, namely LZ77/LZ78 and LZW/LZO [20,21,26,41,45, 46], introduced by is a dictionary level solution to help sliding window identifying and eradicating duplicate string sets. LZW/LZO, planned by Terry Welch in the 1980s, is the LZ compression variant designed to pace up or boost up the firmness cycle. The redundant strategies' organization for data decline [31-34,44].

Mechanisms of the delta compression. In the 1990s, in practice, delta compression approach had here aimed at compressing identical files or chunks. Far ranging synchronization and backup storage device are the main widely used applications. It uses sliding window that is byte-sensitive to find matched strings among similar chunks. Variations are stored in "delts" or "diffs" form between sequential file and complete record [30]. The Delta Compression Strategies include the string step. The X delta and Z delta string stage are delta compression techniques which utilize a byte-clever sliding window [41] to detect redundant strings between the target chunk and the source chunk used for the delta computer.

Data deduplication techniques. There is no single correct way to deduplicate data. There are many different variables by which the best approach can be chosen, for example:

(i) Records deduplication approach became first of all proposed in 2000 to aid worldwide compression at coarse granularity. The preceding techniques are extremely long in order to be informed of comparable chunks and are not feasible, while statistical deduplication methods be able to apply at the reporting or sub-reporting level. It compresses statistics by utilizing chunks of constant or variable size. Using cryptographic hash capabilities generates the hash values of certain chunks, and duplicates are recognized by identical hash values. Both techniques are applied in record-level deduplication at file degree, and reporting is regarded as unmarried entity. This checks the index of the backup document to check the attributes saved within the document. If the equal record available, it provides a indicator to the existing document, or else it update and stores the index price. Therefore, the best example of the study is saved, and it is also known as the unmarried-instance gadget, where it is straightforward to apply the whole document hashing method. Because report hash numbers are clean to make, and it requires considerably fewer energy to process. A few or one-byte exchange in the report results in a one-of-a-kind or new hash price technology that needs exclusive data. The problem of deduplication at document level ends with the advent of block deduplication techniques.

(ii) Sub-report-level (block-degree) deduplication: In these methods, a file is taken apart at numerous small blocks, consisting of constant or variable-length blocks. MD5, SHAI, Rabin fingerprinting and comparable hash algorithms are used to identify comparable blocks. As the result, a block that is unlike the others blocks, is written to the disk and its index is up to date. For any other case, the pointer shall be taken to the original position of the block of equivalent truth. Since the identifiers' number will be significantly increased, additional processing power is required for processing. Block-stage deduplication is similarly classified as constant duration or variable-length deduplication [10]. Deduplication of the fixed-length block period techniques analyze a fixed length of into blocks of fixed-length computer does not back up the same statistics block twice. Important this process' advantage is its unfussiness. One byte, and consequently all blocks of records are subsidized. This limitation of fixed-length block strategies ends in variable length innovation. There blocks of variable duration statistic. Variable-block algorithms use one sort of strategies to decide the block time. This allows their cord block boundaries to "drift" with in. So that modifications to one part of the structure have no effect on the boundaries of various block locations [12]. In a content-based way and within a selection the phase consists of any bytes in the period. It gives greater manage granularity and flexibility to a block.

Merits and demerits of deduplication

The following deduplication features deserve significant attention:

(i) Lessen gadget area: Deduplication helps to reduce the gadget area needed for backups, documents, or bundles of various documents. A completely unique reproduction of statistics is stored as the most accurate, and copy copies are removed. So, to store larger data, it creates more unfastened space [20].

(ii) Provides community bandwidth' improved: Since the exact copies are archived on the disk, therefore, therefore, it would be logical to suggest data duplication to eliminate the need to send replication copies over the network. That is, by deduplication, requirements can be reduced for community bandwidth.

(iii) Reducing energy consumption: Deduplication is a capacity optimization strategy that reduces the requirements for capacity and power. Less power and coolants are needed in decreased storage. It saves power and lowers the load on the gadget tools.

(iv) Lessen usual gadget value: Deduplication allows for widespread time, area, network bandwidth, human assets and finance savings. It ends in greater storage system efficiency and efficacy.

At the same time, there are some demerits of deduplication:

(i) Effect on gadget performance: Fixed-size method in chief gadget machine leads to more than one chunks stored at one-of-a-kind memory locations. It leads to fragmentation problems which have an unfavourable effect on results. The deduplication process calls for additional sources such as memory, reminiscence and bandwidth for its execution. Any inefficient approach to deduplication is affecting the efficiency of a huge storage network.

(ii) Loss of integrity of information. The blocks of information are listed for better lookup through hash

values. For extraordinary blocks of facts, the equal hashes can be generated due to a hash collision that may basis lack of integrity of records. Therefore, hash collisions have to be addressed carefully to keep away from any loss of information and its integrity.

(iii) Backup device problems: Facts deduplication can also need a different hardware tool to relocate and technique information. Such support appliance may additionally cause extra price and effect storage performance [23].

(iv) Compliance with the security and privacy: The deduplication methods are always open to complete.

(v) Archiving (storage): It can be used to get from the entry to the repository. The deductibility strategies' security must be vigilantly designed to protect gadgets against such violations of safety and the loss of secret statistics.

ANN method

A synthetic neural network is a gadget focused mainly on the biological neural networks' function, is an emulation of organic neural gadget in other phrases. ANN is a tool that is used for particular approaches such as categorizing, optimizing, etc. A neural network can fulfil responsibilities not feasible for a linear program. When a detail of the neural community fails, the use of their parallel nature can continue without any hassle. A neural network implements a learning strategy so it does not need to be reprogrammed any more. A synthetic neural culture is of the following forms in particular:

1. If the data must be specifically feed forward from input to output computers, then such a neural network is called a Feed Forward Neural Network. The processing of records may expand over numerous (layers of) gadgets, however no remarks are present, i.e. connections extending from gadget outputs to device inputs in the similar layer or preceding layers.
2. Recurring neural networks that do have references to statements. The community's diverse houses are important, in contrast to feed-ahead networks. In certain cases, the devices' activation values go through a rest process so that the neural population will develop into a stable country in which such activations no longer alternate. The activation values' changes of the output neurons are important in other packages, so that the complex action constitutes the neural network output. For deduplication purpose we make use of feed ahead neural network in the current paintings. A multi-layered neural network was used by us. The neural network' fundamental structure consists of the facts layer, a hidden and an output layer. The structure may be demonstrated by the subsequent discernment, which determines the simple form of a multilayered neural community. Here the statistics layer includes records since the values, which are received from the calculation of incoming data and random weights generated are contained in the input and the hidden layer. The technique performed through the neural community is the segment of education and the segment of experimentation. The information input is fed to the nodes in the training section, in order to find the weights between each node. It will include the cost of weight measured during the schooling process for the production fee during the test.

3.2. Statistics deduplication

As facts exponentially develop in cloud gadget offers, these records are duplicated on dispensed storing machine used for excessive trustworthiness, accessibility, and restoration of calamity [20,21,25,26,41,42,45-47]. The minimum range of data replications called replication factor is critical to protecting the device against screw-ups and high availability. Any wide variety beyond the element of replication must be removed from gadget system. In every other scenario duplicate fact in the gadget computer causes extra stress in conditions of extra room plus bandwidth. To decrease or manage this duplication of information, deduplication strategies are applied to make the storage machine more capable in fee and usage phrases. The utility of deduplication strategies relies on the facts' type, for instance based, not having the structure (unstructured) and semi structure. Also, statistics can be similarly categorized as textual content, picture and video. The knowledge about replication affects overall storage capacity, storage device efficiency, and maximum allowable network traffic processing speed [48-51]. This made it possible for researchers to become aware of how advanced deduplication techniques for gadget structures are being developed. This consists of deleting duplicated records and delivering green information to a gadget unit. Deduplication can be defined as a method which robotically eliminates the statistics on reproduction within gadget systems. The deduplication reduction of facts is reported with the aid of Microsoft and NetApp Microsoft has conducted a record gadget experiment to estimate the stability of area savings between full-document and sub-record deduplication over a four-week period of 857 desktop windows machines [23]. Totally based on observation, complete record deduplication has

a gap reserve equal to 75%, and block-level deduplication 32% of the original needs. Records deduplication was also applied to a virtual library that recognized the use of similarity functions on two actual datasets by duplicated bibliographic metadata reports [39,27,28,52].

Datasets are meta-data statistics for two genuine virtual libraries (BDB Comp and DBLP) & article quotation information for the Cora collection. Consider the effects of the first rate of metadata deduplication in the digital library data set, which improves from 2 to 62% and from 7 to 188% in the item dataset. NetApp considers that 95% replicate information inside storage structures can be reduced by deduplication [41]. Experimental findings show that regular financial savings for backups account for 95%, VMware accounts for 72%, email accounts for 30% and records for 35%.

3.3. Deduplication techniques' classification

Modern deduplication techniques were classified according to the type of storage, i.e. primary or secondary deduplication, source/goal, handling time that is online and post-process deductions, which are based on local and worldwide deductions and cloud-based. A deduplication classification taxonomy based on 4 criteria: length, form, timing, and degree [44]:

(i) *Deduplication based on the storage type (storage-based).* Deduplication classification based on storage type was done. Deduplication is used for primary [36] or secondary storage [28]. Primary storage: The primary storage-based deduplication runs on main memory or active storages that are directly accessible to CPU. CPU continuously reads and executes the servers in memory data are an example of primary storage. Secondary storage it is an auxiliary or external storage system that has not directed access to CPU. It back up primary storage data. These systems are accessed only for retrieving old data and data retention. The examples are storage archives, snapshots and back up storage.

(ii) *Type-based deduplication*. The deduplication procedure is executed either on supply side or heading in the right direction facet. Based on these sorts, deduplication is characterized into deduplication based primarily on source and target. Deduplication is performed on the statistics on the supply aspect prior to it smiles being transferred to the backup goal [23]. The software established on the server's CPU and reminiscence of the moving data to backup server. So, it additionally reduces the bandwidth necessities, gadget and time required to backup information. On the duplicates Deduplication is usually performed on a dedicated gadget device on backup servers. Dedicated hardware deduplication appliances address all deduplication functionalities in this regard. There may be no overhead on the statistics source used for big gadget systems. It calls for extra resources and is more classified as or publish-manner like mentioned underneath.

(iii) *Timing-based deduplication*. Timing-based deduplication applies entirely to the moment of execution of the deduplication algorithm. It imposes a time limit for performing deduplication. The main solution for timing-based deduplication is deduplication operations such as duplicate-searching. As a synchronous/in-band operation or as an asynchronous/out-of-band operation it can be achieved. In addition, the timing-based deduplication was labelled as inline deduplication and post-process deduplication. The deduplicated information is executed on the source facet or earlier than it is written to the disk. So, there's no need for extra disk space to keep the facts to be backed up and defend them. It increases efficiency since the information are exceeded and processed best as soon as Inline deduplication needed extra computation. Finished after backup data is written temporary to gadget device i.e., to a disk. It's also referred to as offline deduplication [17]. It is also faster than inline deduplication it allows in reducing the backup time.

(iv) *Level-based deduplication.* Facts deduplication can be classified as nearby level-primarily based and international-stage primarily based deduplication as discussed underneath Local deduplication is possible in a single VM only, and replicas are detected in a single node. It has a terrible effect on over-all performance, because it cannot get rid of all duplicates completely [41]. It has a little of nodes and of facts. Deduplications are referred to as noneducation of the report is performed in a distributed environment, i.e. across more than one dataset.

It is also called multi-node deduplication, and has a node cluster that paints as a unit together. Records sent to at least one cluster node are compared with previous records sent to that appliance and the information sent to another cluster node. The main aim is to apply deduplication to scattered storage that makes use of more than one storage server.

It eliminates redundant disk accesses and eliminates all viable replicas inside or throughout VMs. similarly, it has extra hashing overhead.

Cloud-based data deduplication on storage systems. Data deduplication technique is commonly used in the Cloud Storage. Backup environment and Data Storage System as it reduces gadget requirements and storage costs. Deduplication technique allows decreasing internet bandwidth more than the community or the sum of facts uploaded to the cloud as only one body replica is saved preferably to duplicate copies of facts. It encourages the speed enhancement of cloud backup, resulting in quicker and green security operations for information.

Cloud storage deduplication may be installed with straight cloud deduplication, secondary gadget copies, and cloud gateway deduplication. Similarly, deduplication can be used in one-of-a-kind gadget systems from primary to secondary virtual machines to cloud storage systems. Personal, civic and mixture cloud storage devices benefit from the deduplication method. It will help researchers identify deduplication strategies based entirely on primary and secondary gadget structures, digital machine systems, network systems, SSD-primarily based multimedia and cloud gadget devices [31-33,44].

3.4. Upcoming research in the field of deduplication strategies

This survey paper provides paintings of cutting-edge studies relevant to deduplication techniques and is in addition to earlier surveys. Deduplication methods were discussed in depth and primarily focused on different taxonomies. An attempt has been made to address the problems of the experiments, which still remain unresolved in deduplication strategies in large distributed gadget systems. Many deduplication strategies for gadget systems have attracted interest in the latest beyond and new strategies are being developed. It is predicted that the artwork on optimizing bite duration or granularity, green detection of reproduction information in a cloud gadget, safety, protection, overall performance enhancement by indexing in deduplication techniques will stay the focal point nearby in the coming years. The paintings on allocated deduplication scalability, breakup, disk bottleneck, I/O latency and average output enhancement in multimedia record reproduction detection are increasing fields to look at. For a cloud gadget system, a scalable, solid, green and assigned deduplication approach is required. This survey provides a scientific and categorized evaluation of 128 research articles. This survey article also offers studies paintings on textual content-based and multimedia-based deduplication techniques. There was also mention for researchers in industry and academia of the annoying conditions and future guidelines for the next generation of green deduplication techniques for big cloud facts.

Open challenges and future research directions based on the numerous problems discussed inside the presented literature, a number of demanding situations in deduplication strategies were recognized and are mentioned under:

(i) Actual or Near-Genuine image deduplication for large-scale disbursed storage device: Most of the social media pictures are either slightly modified or identical replica of a unique picture. The massive quantity of reproduction photos or close to-duplicate pictures calls for big gadget, affects the show and price of a gadget systems. The precise photo or near-actual picture reproduction detection in a huge allotted storage gadget is an open studies task. Such photograph detections require extra CPU, reminiscence and bandwidth. Consequently, a capable and real-time deduplication method for actual and close to-exact picture is a first-rate undertaking in allotted photo storage device.

(ii) Store the transformation of near-genuine photos in a large-scale storage device: The close to-precise images are the tailored version of a novel picture; consequently, it isn't really helpful to save the near-precise photos. Most effective the transformation of near actual photos want to be saved, and those are reconstructed on software online calls. At gift, it's far a huge undertaking in itself to save the alteration of close to-specific pictures in a massive disbursed storage machine.

Overall performance problem for distributed gadget deduplication: Deduplication is achieved in dispensed deduction techniques on an allocated storage gadget using either online or offline approaches. Research for duplicate chunks in huge disbursed gadget is a useful resource-extensive project. Those responsibilities raise the right latency of chunks. Making use of deduplication approach on the time of mark lesser the write overall performance of chunks in phrases of chunks in step with second. Offline deduplication has been widespread, running as a heritage carrier, but requiring extra temporary gadget and increasing the I/O bandwidth. The open task is inline dispensed method of deduplication and most desirable of offline use of assets.

This hassle is still extra difficult in a large scalable gadget machine.

Optimization method for chunk size: A record is separated into tiny chunks of length varying as of 4 KB to 256 MB. This little information is listed and cached for higher performance. Still minor chunks shop the gap, but those generate large hash tables. However, the selection of huge chew length reduces the hash entries, increases the wastage of gadget and takes a big time to compare. So, it makes the deduplication process in depth even more aid. The chunking variable size degrades the overall performance because it generates large index structures.

The selection of chew dimension or granularity is an unlock hassle. There should be a proficient technique to calculate the accurate length of chunks for performance so as to improve the general performance of the gadget. Disk bottleneck problem: The approach to data deduction is usually applied to the disk based on the structure of a secondary gadget. Even though the storage structures are expanding, there are still performance problems with disk I/O operations [31]. The record is divided into chunks to increase the data streaming rate, and is allocated across multiple nodes in a disbursed environment. Chunks are stored on disbursed nodes to conquer the bottleneck disk problem. The radical information distribution strategies are growing with the intention to cause a demand of a newbie deduplication technique.

Throughput and latency: The file is damaged in smaller bit sizes. The metadata of each chew is indicated for the best possible performance and is kept in memory. All new incoming chunks are checked in opposition to a massive bit-in-wheel list. So, the number of I/O disk operations is huge. Hence, fingerprinting indexing has come to be a bottleneck to efficient deduplication systems, which has a negative effect throughout [31] and increases the latency of right operations. The device must use parallel and multiple streams on distributed gadget nodes to meet the necessities of increasing dataset length and deduplication scalability. The gadget throughput is booming. Thus, job can be performed to optimize deduplication throughput and latency in distributed gadget systems.

Fragmentation trouble records deduplication reasons: breakup on disk that minimizes the overall show of read operations [53]. It's going to boom studies time for chronological reads from the equal data and additionally extra disk I/O is required to get admission to on-disk metadata. Deduplication outcomes in statistics breakup and wishes to be addressed cautiously. Fragmentation is each different principal difficulty of withholding for a prolonged duration and can lessen the zone of reference. Higher dealing with inherent fragmentation in disk records requires deduplication.

Scalability and average overall performance of deduplication: The major challenges in deduplication techniques are its tailoring and performance. Every chunk is in comparison with each exceptional chunk in a massive scale storage gadget, and if a match is positioned, the same can be deleted as in step with replication coverage. However, due to the fact of device growth, it becomes difficult for chunks to match entirely. To get throughput, scalability and availability, the centralized index has its non-public troubles and bottleneck. Since the storage necessities are developing suddenly, it poses a terrific task to observe green allotted deduplication strategies on huge-scale allotted gadget structures.

Privateness and protection: In big distributed gadget structures, each information and metadata are disbursed to acquire scalability and availability in this sort of allotted tool, a safety skeleton must be necessary to employ distributed deduplication techniques to defend it in the direction of robbery, attacks and to stick the authoritarian compliances for privateness and protection.

4. Conclusions

In the state of affairs of recent times, deduplication of statistics and cloud computing seemed to be most up-to date practice. The demand for digital information green gadget in cloud computing in a vast storage network has caused the demand for record deduplication to increase. Huge load of data on storage systems underlines the point of interest in developing inexperienced techniques to get rid of redundant data. Cloud computing has the ability to dispose of redundant duplicates, with the development of statistics deduplication and its numerous techniques. Deduplication is a crucial strategy for lowering storage costs of statistics, which is actively accumulated and used to identify, penetrate and solve problems with medical waste (sewage and solid waste), as well as for lowering storage costs of bandwidth and energy use. This studies article presented a methodical survey on statistics deduplication techniques. Contemporary studies in this field have focused exclusively on storage-based deduplication techniques. The first understanding of this survey is to discover deduplication techniques focused on text and multimedia primarily. Based on the evaluation, deduplication poses many challenging circumstances that

could be tackled in complete deduplication based on textual material and multimedia.

For enhancing the gadget in cloud gadget structures, the subsequent studies demanding situations required solutions:

1. There is a requirement to expand a green online information deduplication approach with top-quality use of resources in a Cloud storage tool.
2. To make facts deduplication powerful and power efficient in phrases of area, there may be a need to expand a green deduplication technique with top of the line use of CPU, memory and community resources.
3. To resolve the difficulty of fingerprint indexing in reminiscence, the gadget should parallelize backup streams to multiple nodes for green deduplication.
4. Disk I/O bottleneck is one of the critical problems in gadget systems which have an effect on the overall performance. To resolve this issue; allotted multi-node deduplication approach wishes to be superior.
5. In order to fulfil the requirements, protection and privacy are also very relevant in cloud setting. Based on the literature survey of deduplication techniques, the method of deduplication dispensed on cloud gadget devices has been found to be a promising research challenge. The gaps discussed above include a reduction in storage space, bandwidth, type of disks used, and expenditure on energy usage and heat emissions when implementing these strategies. The art work on a scalable, robust, green and allocated approach to deduplication for a cloud gadget will remain of interest in destiny.

Acknowledgements

The authors would like to acknowledge the administration of Jamia Millia Islamia, Mewat Engineering College, Nuh and administration of other institutions, which authors represent.

Conflicts of interest

None of the authors have not potential conflicts of interest associated with the present study.

References

[1] N.A. Khan, S. Ahmed, S. Vambol, V. Vambol, I.H. Farooqi, Field hospital wastewater treatment scenario, Ecological Questions 30/3 (2019) 57-69. DOI: https://doi.org/10.12775/EQ.2019.022

[2] S. Vambol, V. Vambol, V. Sobyna, V. Koloskov, L. Poberezhna, Investigation of the energy efficiency of waste utilization technology, with considering the use of low-temperature separation of the resulting gas mixtures, Energetika 64/4 (2018) 186-195. DOI: https://doi.org/10.6001/energetika.v64i4.3893

[3] E. Walkinshaw, Medical waste-management practices vary across Canada, Canadian Medical Association Journal 183/18 (2011) 1307-E1308. DOI: https://doi.org/10.1503/cmaj.109-4032

[4] A. Dhingra, N.A. Khan, S. Ahmed, S. Gautam, S. Vambol, V. Vambol, S. Kovalenko, Investigation of medical institutions in India as a source of surface water pollution, World Review of Science, Technology and Sustainable Development (2020) (in press).

[5] N.A. Khan, S.U. Khan, S. Ahmed, I.H. Farooqi, A. Hussain, S. Vambol, V. Vambol, Smart ways of hospital wastewater management, regulatory standards and conventional treatment techniques, Smart and Sustainable Built Environment (2019) (ahead of print). DOI: https://doi.org/10.1108/SASBE-06-2019-0079

[6] V. Hegde, R.D. Kulkarni, G.S. Ajantha, Biomedical waste management, Journal of Oral and Maxillofacial Pathology 11/1 (2007) 5-9. DOI: https://doi.org/10.4103/0973-029X.33955

[7] G.V. Patil, K. Pokhrel, Biomedical solid waste management in an Indian hospital: a case study, Waste Management 25/6 (2005) 592-599. DOI: https://doi.org/10.1016/j.wasman.2004.07.011

[8] F. Bove, Y. Shim, P. Zeitz, Drinking water contaminants and adverse pregnancy outcomes: a review, Environmental Health Perspectives 110 (2002) 61-74. DOI: https://doi.org/10.1289/ehp.02110s161

[9] N.A. Khan, S. Ahmed, I.H. Farooqi, I. Ali, V. Vambol, F. Changani, M. Yousefi, S. Vambol, S. U. Khan, A.H. Khan, Occurrence, sources and conventional treatment techniques for various antibiotics present in hospital wastewaters: a critical review, Trends in Analytical Chemistry 129 (2020) 115921. DOI: https://doi.org/10.1016/j.trac.2020.115921

[10] A. Maazouzi, A. Kettab, A. Badri, B. Zahraoui, A. Kabour, L. Chebbah, Contribution to the study of the effect of urban wastewater on the degradation of ground water quality and to the treatment by filtration on dune sand of the city of Bechar (Algeria),

Desalination and Water Treatment 30/1-3 (2011) 58-68. DOI: https://doi.org/10.5004/dwt.2011.1637

[11] N.A. Khan, S. Ahmed, I.H. Farooqi, S. Vambol, V. Vambol, V. Koloskov, Operational parameters optimisation for simultaneous removal of two drugs in hospital wastewater – RSM approach, International Journal of Environment and Waste Management (2021) (in press).

[12] S.M. Romanchuk, Algoritmy upravleniya tekhnologicheskimi rezhimami vodosnabzheniya gorodov, Problemi Yekologíĭ 1 (2013) 98-108.

[13] N. Badalov, KH. Mamedov, S.D. Tsybulya, Yntehral′naya ynformatsyonnaya systema ydentyfykatsyy zahryaznenyya morskoy poverkhnosty Kaspyyskoho morya. Visnyk Chernihivs′koho Derzhavnoho Tekhnolohichnoho Universytetu. Seriya: Tekhnichni Nauky 3 (2013) 244-249.

[14] S. Vambol, V. Vambol, M. Sundararajan, I. Ansari, The nature and detection of unauthorized waste dump sites using remote sensing, Ecological Questions 30/3 (2019) 43-55. DOI: https://doi.org/10.12775/EQ.2019.018

[15] C.H. Swartz, R.A. Rudel, J.R. Kachajian, J.G. Brody, Historical reconstruction of wastewater and land use impacts to groundwater used for public drinking water: exposure assessment using chemical data and GIS, Journal of Exposure Science & Environmental Epidemiology 13/5 (2003) 403-416. DOI: https://doi.org/10.1038/sj.jea.7500291

[16] S. Karuppasamy, S. Kaliappan, R. Karthiga, C. Divya, Surface area estimation, volume change detection in lime stone quarry, tirunelveli district using cartosat-1 generated digital elevation model (dem), Circuits and Systems 7/06 (2016) 849.

[17] M. Gu, X. Li, Y. Cao, Optical storage arrays: a perspective for future big data storage. Light: Science & Applications 3/5 (2014) e177-e177. DOI: https://doi.org/10.1038/lsa.2014.58

[18] B. Mao, H. Jiang, S. Wu, Y. Fu, L. Tian, Read-performance optimization for deduplication-based storage systems in the cloud, ACM Transactions on Storage 10/2 (2014) 6. DOI: https://doi.org/10.1145/2512348

[19] J. Wang, X. Chen, Efficient and secure storage for outsourced data: a survey, Data Science and Engineering 1/3 (2016) 178-188. DOI: https://doi.org/10.1007/s41019-016-0018-9

[20] A.J. Maan, Analysis and comparison of algorithms for lossless data compression, International Journal of Information and Computation Technology 3/3 (2013) 139-146.

[21] W. Xia, H. Jiang, D. Feng, L. Tian, M. Fu, Y. Zhou, Ddelta: a deduplication-inspired fast delta compression approach, Performance Evaluation 79 (2014) 258-272. DOI: https://doi.org/10.1016/j.peva.2014.07.016

[22] A. Venish, K.S. Sankar, Framework of data deduplication: a survey, Indian Journal of Science and Technology 8/26 (2015) 1-7. DOI: https://doi.org/10.17485/ijst/2015/v8i26/80754

[23] D.T. Meyer, W.J. Bolosky, A study of practical deduplication, ACM Transactions on Storage 7/4 (2012) 1-20. DOI: https://doi.org/10.1145/2078861.2078864

[24] J. Barreto, P. Ferreira, Efficient locally trackable deduplication in replicated systems, in: J.M. Bacon, B.F. Cooper (eds), Middleware 2009. Middleware 2009. Lecture Notes in Computer Science, vol. 5896, Springer, Berlin, Heidelberg, 2009, 103-122. DOI: https://doi.org/10.1007/978-3-642-10445-9_6

[25] I.H. Witten, R.M. Neal, J.G. Cleary, Arithmetic coding for data compression, Communications of the ACM 30/6 (1987) 520-540. DOI: https://doi.org/10.1145/214762.214771

[26] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering—a systematic literature review, Information and Software Technology 51/1 (2009) 7-15. DOI: https://doi.org/10.1016/j.infsof.2008.09.009

[27] J. Paulo, J. Pereira, Distributed exact deduplication for primary storage infrastructures. in: K. Magoutis, P. Pietzuch (eds), Distributed Applications and Interoperable Systems. DAIS 2014. Lecture Notes in Computer Science, vol. 8460, Springer, Berlin, Heidelberg, 2014, 52-66. DOI: https://doi.org/10.1007/978-3-662-43352-2_5

[28] A.F. Banu, C. Chandrasekar, A survey on deduplication methods, International Journal of Computer Trends and Technology 3/3 (2012) 364-368.

[29] C. Alvarez, NetApp deduplication for FAS and V-Series deployment and implementation guide, Technical ReportTR-3505, 2011. Available from: https://aptiris.com/support/netapp_faq/Dedup.pdf

[30] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, Y. Zhou, A comprehensive study of the past, present, and future of data deduplication, Proceedings of the IEEE 104/9 (2016) 1681-1710. DOI: https://doi.org/10.1109/JPROC.2016.2571298

[31] C. Kim, K.-W. Park, K.H. Park, GHOST: GPGPU-offloaded high performance storage I/O deduplication for primary storage system, Proceedings of the 2012 International Workshop on Programming Models and Applications for Multicores and Manycores, PMAM'12, 2012, 17-26. DOI: https://doi.org/10.1145/2141702.2141705

[32] M. Lillibridge, K. Eshghi, D. Bhagwat, V. Deolalikar, G. Trezis, P. Camble, Sparse Indexing: Large Scale, Inline Deduplication Using Sampling and Locality, Proceedings of the Fast'09, 7th USENIX Conference on File and Storage Technologies, San Francisco, Vol. 9, 2009, 111-123.

[33] B. Zhu, K. Li, R.H. Patterson, Avoiding the Disk Bottleneck in the Data Domain Deduplication File System, Proceedings of the Fast'08, 6th USENIX Conference on File and Storage Technologies, San Jose, Vol. 8, 2008, 269-282.

[34] C. Dubnicki, L. Gryz, L. Heldt, M. Kaczmarczyk, W. Kilian, P. Strzelczak, J. Szczepkowski, C. Ungureanu, M. Welnicki, HYDRAstor: A scalable secondary storage, Proceedings of the Fast'09, 7th USENIX Conference on File and Storage Technologies, San Francisco, Vol. 9, 2009, 197-210.

[35] N. Mandagere, P. Zhou, M.A. Smith, S. Uttamchandani, Demystifying data deduplication, Proceedings of the ACM/IFIP/USENIX Middleware'08 Conference Companion, 2008, 12-17. DOI: https://doi.org/10.1145/1462735.1462739

[36] J. Paulo, J. Pereira, A survey and classification of storage deduplication systems, ACM Computing Surveys 47/1 (2014) 1-30. DOI: https://doi.org/10.1145/2611778

[37] C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275 (2014) 314-347. DOI: https://doi.org/10.1016/j.ins.2014.01.015

[38] E.N. Borges, M.G. de Carvalho, R. Galante, M.A. Gonçalves, A.H. Laender, An unsupervised heuristic-based approach for bibliographic metadata deduplication, Information Processing and Management 47/5 (2011) 706-718. DOI: https://doi.org/10.1016/j.ipm.2011.01.009

[39] J. Xu, W. Zhang, Z. Zhang, T. Wang, T. Huang, Clustering-based acceleration for virtual machine image deduplication in the cloud environment, Journal of Systems and Software 121 (2016) 144-156. DOI: https://doi.org/10.1016/j.jss.2016.02.021

[40] R. Di Pietro, A. Sorniotti, Proof of ownership for deduplication systems: A secure, scalable, and efficient solution, Computer Communications 82 (2016) 71-82. DOI: https://doi.org/10.1016/j.comcom.2016.01.011

[41] P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80/4 (2007) 571-583. DOI: https://doi.org/10.1016/j.jss.2006.07.009

[42] J. Gantz, D. Reinsel, The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East, IDC Analyze the Future (2012) 1-16. Available from: https://www.speicherguide.de/download/dokus/IDC-Digital-Universe-Studie-iView-11.12.pdf

[43] Y. Tian, S.M. Khan, D.A. Jiménez, G.H. Loh, Last-level cache deduplication, Proceedings of the 28th ACM International Conference on Supercomputing, 2014, 53-62. DOI: https://doi.org/10.1145/2597652.2597655

[44] B. Mao, H. Jiang, S. Wu, L. Tian, Leveraging data deduplication to improve the performance of primary storage systems in the cloud, IEEE Transactions on Computers 65/6 (2015) 1775-1788. DOI: https://doi.org/10.1109/TC.2015.2455979

[45] S. Shanmugasundaram, R. Lourdusamy, A comparative study of text compression algorithms, International Journal of Wisdom Based Computing 1/3 (2011) 68-76.

[46] U.S. Bhadade, A.I. Trivedi, Lossless text compression using dictionaries, International Journal of Computer Applications 13/8 (2011) 27-34.

[47] D.A. Reed, J. Dongarra, Exascale computing and big data, Communications of the ACM 58/7 (2015) 56-68. DOI: https://doi.org/10.1145/2699414

[48] C.H. Ng, M. Ma, T.Y. Wong, P.P. Lee, J.C. Lui, Live deduplication storage of virtual machine images in an open-source cloud, in: F. Kon, A.M. Kermarrec (eds), Middleware 2011, Lecture Notes in Computer Science, vol. 7049, Springer, Berlin, Heidelberg, 81-100. DOI: https://doi.org/10.1007/978-3-642-25821-3_5

[49] X. Zhao, Y. Zhang, Y. Wu, K. Chen, J. Jiang, K. Li, Liquid: A scalable deduplication file system for virtual machine images, IEEE Transactions on Parallel and Distributed Systems 25/5 (2013) 1257-1266. DOI: https://doi.org/10.1109/TPDS.2013.173

[50] C.A. Waldspurger, Memory resource management in VMware ES X server, ACM SIGOPS Operating

Systems Review 36/SI (2002) 181-194. DOI: https://doi.org/10.1145/844128.844146

[51] A.T. Clements, I. Ahmad, M. Vilayannur, J. Li, Decentralized Deduplication in SAN C luster File Systems, Proceedings of the USENIX Annual Technical Conference, 2009, 101-114.

[52] Q. He, Z. Li, X. Zhang,. Data deduplication techniques, Proceedings of the 2010 International Conference on Future Information Technology and Management Engineering, Changzhou, 2010, 430-433. DOI: https://doi.org/10.1109/FITME.2010.5656539

[53] K. Srinivasan, T. Bisson, G.R. Goodson, K. Voruganti,. iDedup: latency-aware, inline data deduplication for primary storage, Proceedings of the Fast'12, 10th USENIX Conference on File and Storage Technologies, San Jose, Vol. 12, 2012, 1-14.