Research Article

EXAMINING THE IMPORTANCE OF ARTIFICIAL INTELLIGENCE IN THE SINGULARIZATION OF BIG DATA WITH THE DEVELOPMENT OF CLOUD COMPUTING

Authors: Serkan KESKİN, Ali Hakan IŞIK

To cite to this article: Keskin, S. & Isık, A. H. (2023). EXAMINING THE IMPORTANCE OF ARTIFICIAL INTELLIGENCE IN THE SINGULARIZATION OF BIG DATA WITH THE DEVELOPMENT OF CLOUD COMPUTING . International Journal of Engineering and Innovative Research ,5(2),p170-180 . DOI: 10.47933/ijeir.1261330

DOI: 10.47933/ijeir.1261330

To link to this artice: https://dergipark.org.tr/tr/pub/ijeir/archive

International Journal of Engineering and Innovative Research

http://dergipark.gov.tr/ijeir

EXAMINING THE IMPORTANCE OF ARTIFICIAL INTELLIGENCE IN THE SINGULARIZATION OF BIG DATA WITH THE DEVELOPMENT OF CLOUD COMPUTING

Serkan KESKİN[*], Ali Hakan IŞIK [2]

[1]Burdur Mehmet Akif Ersoy University, Institute of Science and Technology, Department of Computer Engineering, Burdur, Turkey.
[2] Burdur Mehmet Akif Ersoy University, Faculty of Architecture-Engineering, Department of Computer Engineering,
Burdur, Turkey

[*]Corresponding Author: serkankeskin@isparta.edu.tr
(**Received:** 07.03.2023; **Accepted:** 22.05.2023)

https://doi.org/10.47933/ijeir.1261330

ABSTRACT: Big data is a huge amount of structured or unstructured data that cannot be processed, managed and analyzed by traditional data methods. This data often comes from multiple sources and of different types. With the emergence of big data, it has become difficult to process data with the algorithms used to process data. Therefore, new algorithms and technologies have been developed. One of the most important of these technologies is data deduplication. Data deduplication is a process in which data from different data sources are grouped according to their similarities in order to reduce repetitive data and prevent inconsistencies. In this way, it aims to save storage space by storing only one copy of many repeated data. The most commonly used deduplication architectures are inline deduplication, post-process deduplication and hybrid deduplication. Using these architectures, 90% data savings can be achieved in deduplication. Thus, data occupies less space in storage units. Today, artificial intelligence technologies are advancing very rapidly and their application areas are expanding. Therefore, artificial intelligence will continue to be a very important technology for the industry and our lives in the future. The aim of this study is to give an idea about the relationship between deduplication technology and artificial intelligence by examining various deduplication systems and algorithms. In addition, in the light of detailed information about deduplication, the importance of the contributions of today's deduplication systems to information systems is also explained.

Keywords: Cloud computing, artificial intelligence, big data, deduplication.

1. INTRODUCTION

Today, with the introduction of big data, technologies have started to evolve and transform. On the way from desktop computers to a mobile life, our devices have shrunk. In the opposite direction of this shrinkage, there has been a serious growth in our data. Although we use the term "big data", what is actually happening is 'unstructured' data. The vast majority of the source of unstructured data is the internet. This big data has caused the storage areas to fill up over time and new storage areas are needed. This situation has also led to storage costs. The increase in storage space and costs has led to the development of new technologies. Technologies such as Thin Porivisioning, Automated Tiering, Snapshot, Virtual Storage Management, Archiving, Deduplication, Compression help to close the gaps of storage systems that have difficulty in

catching up with the growth rate of data. Thanks to these technologies, our data does not take up as much space as before. Deduplication, one of the most important of these technologies, aims to save storage space by storing a single copy of repeated data on a file or block basis.

Deduplication is also known as single instance storage, intelligent compression, data reduction or common factorization. Deduplication is a technology that storage device manufacturers rely on to use storage space more efficiently. Another storage technology is data compression. This feature is often lumped into a larger category called data reduction. All these systems help achieve the same goal - data efficiency for increased storage space. With proper deduplication techniques, businesses can efficiently store more data than their overall storage capacity can offer. Deduplication will always have a huge positive impact on overall storage utilization and reduce costs. But it is important to know which type of deduplication method is required to maximize efficiency in the right way.

Artificial intelligence systems, especially those using machine learning algorithms, require large amounts of high quality data for training and prediction. By using deduplication, the AI system is able to process a dataset that is refreshed and free of redundant data. This aims to provide a cleaner and more accurate dataset for training. The relationship between deduplication and AI is that deduplication is often used as a pre-processing of data before it is fed into an AI system to improve the quality of the data and make working with it more manageable. At the same time, AI models can use deduplication within their tasks to make the process more efficient.

2. RELATED WORKS

In this section, we discuss deduplication technology and studies that examine big data issues. Studies in the literature examining deduplication algorithms, data security, deduplication in the cloud system and artificial intelligence are analyzed. As a result of the analysis, it shows that data deduplication provides significant savings in storage space, the importance of data security, artificial intelligence and deduplication are used as a whole. The prominent studies in the literature review are presented.

In the study by Keles and Nevcihan, the issue of deduplication of rapidly and uncontrollably increasing data in a hotel database was addressed. Various text similarity algorithms were used for text similarity. Distance regularization based similarity, token based similarity, sequence based similarity algorithms and fuzzy string matching algorithms were examined. At the end of the study, a hybrid model using fuzzy string matching algorithm and Jaro Winkler distance is proposed. In the comparison on the test dataset, the model achieved 94% accuracy on the easy dataset and 83% accuracy on the difficult dataset [1].

Pg et al. present an approach to solve the problem of wasted storage space due to the large amount of data. It develops a new method that uses Convergent and Modified Elliptic Curve Cryptography algorithms on a cloud environment to create secure deduplication systems. The performance analysis confirms that the proposed system has 96% security. This result is higher than other existing methods. The evaluation result reveals that the proposed system is highly secure and effective for deduplication for an integrated cloud environment [2].

The study by Jiang et al. investigated data privacy and property management issues for deduplication. In this work, a scheme that resists the poison attack and the recursive forgery attack of cross-user file-level deduplication is used. For this, a new proof of ownership scheme

using the Bloom filter is designed. As a result, the security analysis conserved bandwidth for deduplication. The proof of ownership verification process was used to achieve mutual proof of ownership verification. It has been shown that the proposed schemes can guarantee security requirements other than side channel attack [3].

In the study by Yang et al. existing deduplication schemes were used to resist brute force attacks. It is understood that schemes are designed to ensure efficiency and data availability. However, it is explained that it is not valid for both conditions. In this study, a three-tier cross-domain architecture is investigated. A cloud-efficient and privacy-preserving big data deduplication is envisioned. Decision tree algorithms are utilized. Improved privacy protection, data availability, and accountability capability are shown to outperform the current state of the art while resisting brute force attacks [4].

Barik et al. describe that with the rapid expansion of the Internet of Things (IoT), IoT devices are generating enormous amounts of big data. It is explained that traditional cloud computing systems are not efficient enough to process large volumes of data simultaneously. In order to build a big data deduplication scheme, this paper designs a new GeoBD 2 scheme that defines a geographic deduplication structure. According to the experimental results, it is found that the proposed scheme has minimum overhead cost compared to the existing big data deduplication scheme [5].

Vijayalakshmi and Jayalakshmi, in a research paper, discuss the deduplication techniques used to deduplicate big data generated by connecting many devices to the internet. The main purpose of this paper is to give a good idea about deduplication techniques by examining sufficient information and available data [6].

In the study by Manogar and Abirami, it was seen that the increase in data such as text, images, audio, video, data centers and backup data in recent years has caused many problems in both storage and retrieval process. For this reason, location-based deduplication, time-based deduplication, and chunk-based deduplication techniques are examined and it is explained that it is aimed to determine the best result. Among these techniques, variable dimensional deduplication within the scope of stack-based deduplication technique was found to give better results compared to other deduplication techniques. As a result, it was found that variable dimensional deduplication technique increases storage efficiency and increases the performance of storage resources by allowing more data to be transferred and processed [7].

Aiming to improve storage efficiency, Leesakul et al. propose a dynamic deduplication scheme in cloud storage systems that aims to preserve redundancy for fault tolerance. Deduplication mechanisms in cloud systems imply a static scheme that limits their applicability to the dynamic nature of the data in this system. Therefore, it emphasizes the need to strike a balance between changing storage efficiency and fault tolerance requirements. Furthermore, a dynamic deduplication scheme for cloud storage is proposed to improve performance in cloud storage systems. The number of copies of files is dynamically changed according to the changing Quality of Service (QoS) level. Experimental investigations show that the performance of the proposed system is significantly improved and can cope with the scalability problem [8].

Fan et al. consider an encrypted deduplication mechanism that enables the cloud storage server to eliminate duplicate ciphertexts and improve privacy protection. A feasible encrypted deduplication mechanism is proposed where all data is stored in the form of a cipher structure consisting of a control block, a transformation block, an activation block and a cipher block.

The cloud storage server can define duplicate cipher structures with control blocks. It is then mentioned that it can transform duplicate activation blocks. Each data owner of duplicate cipher structures corresponding to the same plaintext can share the same copy of the cipher blocks. As a result, the cloud storage server can avoid wasting storage space by storing only one copy of the duplicate cipher block [9].

In a study by Park et al. on neural networks and deduplication, a new reference search technique, DeepSketch, was used. In this way, an average of 21% more success was achieved compared to the delta compression technique [10].

In a study by Tarun et al. on machine learning and deduplication in distributed systems, it was reported that repeated data entry of stored data poses a major problem for query processing and data analysis. To minimize this problem, an advanced machine learning based algorithm is proposed to detect repeated data entries and improve efficiency [11].

3. METHODS

3.1. Cloud computing, Data Storage, Data Security, Multi-Device Access and Cost Effectiveness

Cloud computing refers to the provision of services over the internet. Depending on a cloud computing service provider, users can access many different services over the internet for a fee. These services include software as a service (SAAS), infrastructure as a service (IAAS) and platform as a service (PAAS). IaaS allows users to create their own virtual machines and storage. This allows users to install their own operating systems and applications. PaaS provides a platform for users to develop and deploy their own applications. Thus, users can use the infrastructure provided by the cloud computing service provider while developing their own applications. SaaS aims to enable users to use applications over the internet. These applications are usually accessible through a browser and do not need to be downloaded or installed by users [12]. Examples include applications such as Google Docs or Microsoft Office 365.

Data storage is the general name given to the process of storing and retaining information. Data storage systems are physical devices and software used to store users' data. Data storage systems have features such as keeping data secure, fast access and management. Cloud computing services also include data storage services. Users can store their data for a certain fee, connected to cloud computing service providers over the internet [13].

The use of cloud computing services brings some risks in terms of data security. These risks include the ability to transmit and store data over the internet, the inability to control access to data, the possibility of damaging data, and the loss of data [4]. Therefore, it is necessary to take precautions regarding data security when using cloud computing services. Measures that can be taken for data security include encryption, authentication, authorization, backup and update [9]. It is also important to investigate the security measures and data protection policies of the cloud computing service provider and to choose a reliable service provider in this regard [14].

Cloud computing services allow users to access their data from any device over the internet thanks to data storage systems. In this way, users can access their data from any device with an internet connection from their home, workplace or even while traveling [15]. This is referred to as data multi-device access. Data multi-device access aims to make it easier and faster for users to access their data. Data multi-device access is made possible through the use of cloud

computing services. Users can benefit from data storage services over the internet for a certain fee depending on the cloud computing service provider [16]. Thus, users can access their data from any device over the internet. However, data multi-device access brings some risks in terms of data security. Therefore, it is necessary to take precautions for data security when using data multi-device access.

Cloud computing allows a software or service to be used over the internet. Instead of purchasing their own software and hardware, this method enables organizations to outsource services by paying a fee for use when needed. With this method, organizations can use software and hardware without having to invest upfront. The cost-effectiveness of cloud computing is a major advantage for organizations. For example, if an organization purchases its own software and hardware, it will only use a portion of that investment and the rest may be idle. However, when using cloud computing, the organization can use the software and hardware by paying only when needed. In this way, organizations can benefit from services in a less costly way instead of making upfront investments. Cloud computing also reduces additional costs for organizations such as maintenance and updates. Cloud providers maintain and update the software and hardware themselves and offer these services to organizations for a fee [5]. Thus, organizations can use resources such as time and money allocated to the maintenance and updating of their own software and hardware in other areas.

3.2. Artificial Intelligence and Data Analytics

Artificial Intelligence aims for a machine or program to have the ability to think and learn like a human. Machine learning is the ability of a machine to improve itself using data. It enables machine learning systems to become more effective and deliver better results throughout the learning process. This goal involves automating the processes of data collection, learning and decision making. Data analysis is the process of collecting, organizing and making sense of data and producing usable information. Data analysis is used to understand data collected from many different data sources and to generate meaningful information. Data analytics is often based on technologies such as artificial intelligence and machine learning. There are important differences between these two concepts [17]. This data can take various forms. However, the most used data types are tables and graphs. Data analysis is usually done using advanced data analysis techniques. These techniques include regression, clustering and classification. Artificial intelligence and data analysis are used in many different industries and play an important role in these fields. For example, in a healthcare company, data analytics can be used to predict the spread of diseases and take preventive measures. Likewise, in a retail company, data analytics can be used to predict customer behavior and take preventive measures.

3.3. Integration of Artificial Intelligence and Cloud Computing

Artificial intelligence and cloud computing are technologies that match each other very well. They can be used by integrating with each other. If an AI application needs data processing, it is possible to store and process that data with cloud computing services. It can enable the application to meet its data processing requirements and continue to operate without the need for more data processing power as it grows. Furthermore, AI applications often work with large data sets and require large data centers to store and process this data [18]. Cloud computing is ideal for meeting these data processing needs. This is because data processing services can be accessed without the need to own data centers. Cloud computing aims to meet the data processing needs of AI applications and enable them to continue to operate without the need for more data processing power as they grow.

3.4. Big Data and Types

Big data refers to large, complex data sets that are difficult to process with traditional data management tools and methods. This is because these tools and methods are inadequate due to factors such as the volume, velocity or variety of data. [19]. These datasets often come from a variety of sources, including social media, transactional data, and sensors [20]. The data can be structured or unstructured. The term big data is often used to describe the challenges and opportunities of working with large datasets, as well as the tools and technologies used to store, process and analyze them [12].

There are several types of big data, each with its own characteristics and challenges. Some common types of big data are; Structured data: Data that is organized in tabular form and can be easily processed. Unstructured data: Data that does not have a specific order. Semi-structured data: Data that is a combination of structured and unstructured data. Examples include emails or web pages. Multimedia data: Refers to data in the form of pictures, videos or audio recordings. Sensor data: Refers to data generated by sensors or devices such as IoT devices, wearables [5]. Transaction data: Data generated as a result of transactions, such as financial or retail data.

3.5. Deduplication of Big Data and Deduplication Architectures

Deduplication is the process of identifying and removing duplicate records from a dataset. This is often an important step in the process of working with big data, as large datasets can contain a significant number of duplicate records. Deduplication can be useful for a variety of purposes, including reducing the storage space required for a dataset, improving the accuracy of analytics and machine learning models, and simplifying the process of analyzing data [21].

Deduplication architectures can be divided into three classes. Inline Deduplication: In this architecture, data is deduplicated as it is written to the storage system. The system compares incoming data with existing data in real time and eliminates duplicates before the data is stored. This approach requires more processing power and can impact system performance. However, it ensures the most efficient use of storage space. Post-Process Deduplication: In this architecture, data is stored in its original form and deduplication is performed later in a separate process. This approach requires fewer resources than inline deduplication but may require more storage. It can result in longer backup and restore times. Hybrid Deduplication: This architecture combines inline and post-process deduplication. Inline deduplication is used for data that is accessed or written frequently, while post-process deduplication is used for data that is accessed less frequently.

Deduplication works by identifying and removing duplicate records from a dataset. There are several different approaches to deduplication depending on the characteristics of the data and the goals of the deduplication process. A common approach to deduplication is record linkage, which involves identifying records that refer to the same real-world entity, even if the records do not match exactly. It can be done using various techniques such as matching records based on common attributes such as name and address, or using more advanced techniques such as machine learning to identify patterns in the data that indicate that two records refer to the same entity [22]. Another approach to deduplication is hash-based deduplication, which involves creating a unique hash value for each record and comparing the hash values to identify duplicates. It can be an effective way to identify duplicates. However, it is important to ensure

that the hash function used is sufficient to distinguish all unique records in the dataset. Ordered list deduplication involves sorting the records in a dataset and then identifying and removing duplicates based on the sorted order. This method can be an effective way to identify duplicates and requires sorting the entire dataset. This can be time-consuming for large datasets. Probabilistic deduplication involves using statistical techniques to identify and remove duplicates based on the probability that two records refer to the same real-world entity [22], [23]. As such, it can be an effective way to identify duplicates. However, it may not be as accurate as other approaches as it relies on statistical estimates rather than exact matches. In general, the most effective approach to deduplication will depend on the characteristics of the dataset and the goals of the deduplication process.

3.6. How Data Deduplication Works

If a user stores the same file in multiple locations, deduplication stores only one copy of the file and all other copies are saved as references to this copy. This means that when backing up or storing data, much less data is sent and storage space is used more efficiently. Technically, deduplication eliminates duplicate blocks of data. It stores unique data blocks at the 4KB block level within a volume and across all volumes in the aggregate. It then relies on unique digital signatures for all 4 KB data blocks. The deduplication engine examines the newly incoming blocks, develops a digital signature, and stores it in a hash store of the digital signature when data is written to the system. After the digital signature is calculated, a search is performed on the hash store. The block of data that matches the repeated digital signature is examined in the cache in case it matches a digital signature in the hash store [24]. If a match is detected, a byte-by-byte comparison between the valid data block and the digital signature is performed as verification. The receiving block is shared with the matching digital signature without writing the receiving block to disk during verification. Only the metadata is updated to keep track of the sharing details. If the transmitting block is not found in the cache, it is pre-cached from disk and compared byte-by-byte to ensure an exact match. Without actually writing to disk, the receiving block is marked as a duplicate during verification. Metadata is updated to keep track of sharing details. Likewise, the background deduplication engine runs. It searches all data blocks in bulk and removes duplicates by comparing block digital signatures and performing byte-by-byte comparison to eliminate false positives [25]. This method also ensures that no data is lost during the deduplication process. Figure 1 visually illustrates how deduplication works.

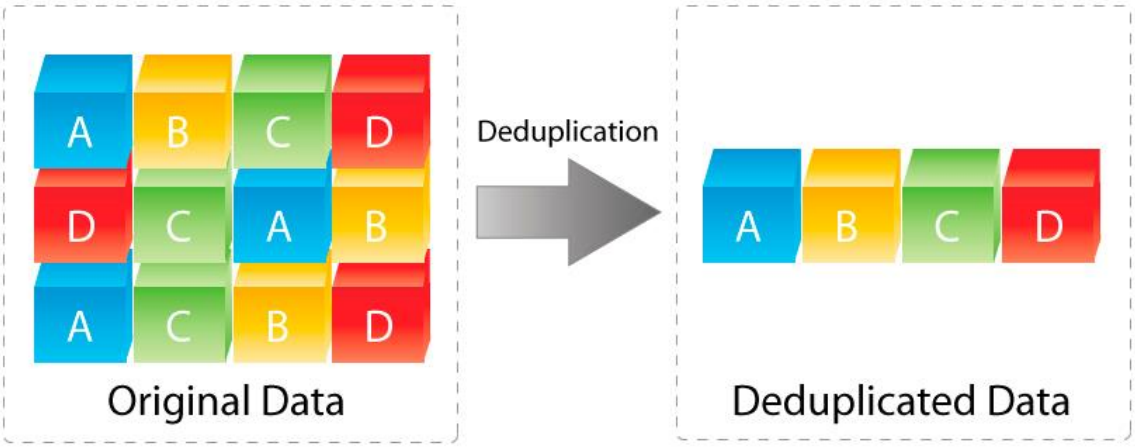

Figure 1. Example of data deduplication [26]

4. RESULTS AND DISCUSSION

Today, deduplication is very important. Because it reduces your storage requirements, saves you money and reduces the amount of bandwidth used to move data to and from remote storage sites. While deduplication can reduce storage requirements by up to 95% in some cases, it can be affected by factors such as the type of data you are trying to deduplicate. Even if your storage requirements are less before deduplication than after deduplication, it can reduce your cost of data storage and significantly increase your bandwidth availability.

Once the deduplication process is complete, it is possible to use the deduplicated dataset to build predictive models. The most appropriate type of model will depend on the nature of the data and the objectives of the analysis. Some common types of predictive models that can be built from de-duplicated data can be categorized into 4 groups. If we briefly discuss these models; Regression models: Can be used to predict continuous outcomes based on a set of input variables. Classification models: Can be used to predict categorical outcomes based on a set of input variables. Clustering models: These models can be used to identify groups of similar records in the data. Anomaly detection models: Can be used to identify unusual or unexpected patterns in data.

Prediction models created by combining artificial intelligence and big data are models that aim to predict future events or outcomes by analyzing large data sets. These prediction models are created using artificial intelligence methods and big data technologies. For example, a prediction model can analyze a large dataset of customer buying behaviors to predict when and what a customer will buy in the future. These prediction models can be used in a variety of different fields. Finance: A finance company can use AI and big data to predict the future payment behavior of customers. Health: A healthcare company can use AI and big data to predict future health issues. Marketing: A marketing company can use AI and big data to predict customer buying behavior. Energy: An energy company can use AI and big data to predict future energy demand. Forecasting models built using AI and big data are often very comprehensive and powerful and can analyze large data sets collected from a wide range of different data sources. As a result, these models can often predict future events and outcomes with very high accuracy.

As can be seen from the comparison in Table 1, a comparison of the performance of machine learning algorithms based on artificial intelligence is given. The reasons for the different performances are due to the fact that the data sets used in the study were not used with the correct machine learning techniques. It is thought that the success rate will increase with the development of machine learning algorithms and artificial intelligence.

Table 1. Success comparison table on different studies.

Author	Methodology	Performance
Keleş ve Nevcihan [1]	Deduplication by text similarity	Medium
Pg ve ark. [2]	Deduplication with convergent and modified elliptic curve cryptography algorithms	Medium
Jiang ve ark. [3]	Creating a new proof of ownership scheme using the Bloom filter	Medium
Yang ve ark. [4]	Deduplication with decision tree algorithm	High
Barik ve ark. [5]	Unification with GeoBD 2 schema	High
Vijayalakshmi ve Jayalakshmi [6]	Deduplication techniques	Low
Manogar ve Abirami [7]	Stack-based deduplication	Medium
Leesakul ve ark. [8]	Dynamic deduplication	Medium
Fan ve ark. [9]	Deduplication of duplicate passwords	Medium
Park ve ark. [10]	Deduplication with DeepSketch technique	High
Tarun ve ark. [11]	Deduplication using machine learning algorithms	High

5. CONCLUSIONS

Deduplication and artificial intelligence are two increasingly important technologies used in almost all industries today. Deduplication and AI complement each other in a number of ways, notably in improving data quality, reducing storage costs and enhancing data analytics. Deduplication is the process of processing and analyzing large amounts of data and extracting meaningful information. Artificial intelligence is the ability of computer systems to think and make decisions in a human-like manner. An AI system continuously scans the dataset and can detect and automatically eliminate multiple copies of the same data. This is a crucial step in the deduplication process. In the AI deduplication process, various methods can be proposed that can be used to process and understand the dataset. An AI system can identify which data is important in the dataset and by highlighting them, it can optimize the deduplication process. By transforming the data obtained in the deduplication process into meaningful information, AI can facilitate the understanding and use of the data. Artificial intelligence can quickly access large amounts of data using deduplication algorithms. Thus, it can make the process more efficient and effective. From another perspective, deduplication helps AI models find similar or identical information in the data sets used during training. This method helps to organize and clean the data sets. This leads to better results when training AI models. Deduplication is especially important for very large data sets. Duplication of information can slow down the training process of AI models and reduce the likelihood of accurate results. Deduplication helps to eliminate such repetitive information. In this way, the training of AI models can be faster and more accurate. In addition to improving the accuracy and efficiency of deduplication, AI can

also improve data analysis by identifying relationships and correlations between data points that human analysts may not immediately recognize. This can help organizations extract more value from their data, leading to more informed decision-making and better business outcomes. As you can see, artificial intelligence and data deduplication feed off each other and are inseparable. If we use a program to analyze a data set, it must have deduplication and artificial intelligence in the background. Today, these two are among the constants of data analysis. Today, as the size of our data increases, we need new storage spaces. It becomes important to reduce the volume of our data in storage areas. It is possible to achieve this with data deduplication. This study aims to provide a different perspective to researchers who will study the importance of deduplication and artificial intelligence against each other. In future studies, it is aimed to experimentally study how important deduplication is.

REFERENCES

[1] Keleş, Ü., & Nevcihan, D. U. R. U. (2021). Metin Benzerliği Algoritmaları ile Veri Tekilleştirme: Oteller Veri Tabanında Bir Uygulama. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 14(2), 86-98.

[2] PG, S., RK, N., Menon, V. G., Abbasi, M., & Khosravi, M. R. (2020). A secure data deduplication system for integrated cloud-edge networks. Journal of Cloud Computing, 9(1), 1-12.

[3] Jiang, S., Jiang, T., & Wang, L. (2017). Secure and efficient cloud data deduplication with ownership management. IEEE Transactions on Services Computing, 13(6), 1152-1165.

[4] Yang, X., Lu, R., Choo, K. K. R., Yin, F., & Tang, X. (2017). Achieving efficient and privacy-preserving cross-domain big data deduplication in cloud. IEEE transactions on big data, 8(1), 73-84.

[5] Barik, R. K., Patra, S. S., Patro, R., Mohanty, S. N., & Hamad, A. A. (2021, March). GeoBD2: Geospatial big data deduplication scheme in fog assisted cloud computing environment. In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 35-41). IEEE.

[6] Vijayalakshmi, K., & Jayalakshmi, V. (2021, April). Analysis on data deduplication techniques of storage of big data in cloud. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 976-983). IEEE.

[7] Manogar, E., & Abirami, S. (2014, December). A study on data deduplication techniques for optimized storage. In 2014 Sixth International Conference on Advanced Computing (ICoAC) (pp. 161-166). IEEE.

[8] Leesakul, W., Townend, P., & Xu, J. (2014, April). Dynamic data deduplication in cloud storage. In 2014 IEEE 8th International Symposium on Service Oriented System Engineering (pp. 320-325). IEEE.

[9] Fan, C. I., Huang, S. Y., & Hsu, W. C. (2015, May). Encrypted data deduplication in cloud storage. In 2015 10th Asia Joint Conference on Information Security (pp. 18-25). IEEE.

[10] Park, J., Kim, J., Kim, Y., Lee, S., & Mutlu, O. (2022). {DeepSketch}: A New Machine {Learning-Based} Reference Search Technique for {Post-Deduplication} Delta Compression. In 20th USENIX Conference on File and Storage Technologies (FAST 22) (pp. 247-264).

[11] Tarun, S., Batth, R. S., & Kaur, S. (2021, December). A Scheme for Data Deduplication Using Advance Machine Learning Architecture in Distributed Systems. In 2021 International Conference on Computing Sciences (ICCS) (pp. 53-60). IEEE.

[12] Dokuz, A. Ş., & Çelik, M. (2017). Bulut Bilişim Sistemlerinde Verinin Farklı Boyutları Üzerine Derleme.Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, 6(2), 316-338.

[13] Çelik, K. (2021). Bulut Bilişim Teknolojileri. Bartın Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 12(24), 436-450.

[14] Jamsa, K. (2022). Cloud computing. Jones & Bartlett Learning.

[15] Hurwitz, J. S., & Kirsch, D. (2020). Cloud computing for dummies. John Wiley & Sons.

[16] Singhal, S., Sharma, P., Aggarwal, R. K., & Passricha, V. (2018). A global survey on data deduplication. International Journal of Grid and High Performance Computing (IJGHPC), 10(4), 43-66.

[17] Alonso, J. M., & Casalino, G. (2019, June). Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In International workshop on higher education learning methodologies and technologies online (pp. 125-138). Springer, Cham.

[18] Gill, S. S., Tuli, S., Xu, M., Singh, I., Singh, K. V., Lindsay, D., ... & Garraghan, P. (2019). Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. Internet of Things, 8, 100118.

[19] Demirol, D., Das, R., & Hanbay, D. (2019, September). Büyük veri üzerine perspektif bir bakış. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-9). IEEE.

[20] Süzen, A. A., & Kayaalp, K. (2019). Büyük Verilerde Gizlilik Tabanlı Yaklaşım: Federe Öğrenme. International Journal of 3d Printing Technologies and Digital Industry, 3(3), 297-304.

[21] Xu, L. J., Hao, R., Yu, J., & Vijayakumar, P. (2021). Secure deduplication for big data with efficient dynamic ownership updates. Computers & Electrical Engineering, 96, 107531.

[22] Premkamal, P. K., Pasupuleti, S. K., Singh, A. K., & Alphonse, P. J. A. (2021). Enhanced attribute based access control with secure deduplication for big data storage in cloud. Peer-to-Peer Networking and Applications, 14(1), 102-120.

[23] Kumar, N., & Jain, S. C. (2019). Efficient data deduplication for big data storage systems. In Progress in Advanced Computing and Intelligent Engineering (pp. 351-371). Springer, Singapore.

[24] Savić, I., & Lin, X. (2021, November). The Analysis and Implication of Data Deduplication in Digital Forensics. In International Symposium on Cyberspace Safety and Security (pp. 198-215). Springer, Cham.

[25] Wu, H., Wang, C., Fu, Y., Sakr, S., Zhu, L., & Lu, K. (2017). Hpdedup: A hybrid prioritized data deduplication mechanism for primary storage in the cloud. arXiv preprint arXiv:1702.08153.

[26] URL-1 (2022). https://www.hsb.nl/the-importance-of-deduplication-and-adjudication-in-identity-management-solutions/ (Erişim tarihi 12.02.2022)