Cureus Journal of Computer Science

Open Access Review Article

An Exhaustive Survey of Big Data Storage Reduction Techniques

Samita A. Patil [1] , Savita Sangam [1]

1. Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, IND

Corresponding authors: Samita A. Patil, patilsamita85@gmail.com, Savita Sangam, savita.sangam@gmail.com

Received 03/24/2025
Review began 03/31/2025
Review ended 04/16/2025
Published 04/16/2025

DOI: https://doi.org/10.7759/s44389-025-03518-3

Abstract

The exponential growth of data in the era of big data has introduced significant challenges in storage, processing, and analysis, necessitating the development of effective data reduction methods. This survey provides a concise overview of key big data reduction techniques, systematically reviewing methodologies such as data compression, data deduplication , dimensionality reduction, and feature selection. By analyzing data reduction techniques, algorithms, and frameworks, the study highlights the critical need for data reduction to manage storage demands, improve computational efficiency, and enable meaningful data analysis in large-scale systems. Key findings reveal that hybrid approaches combining multiple reduction strategies excel in maintaining data integrity while significantly reducing storage and computational overhead, with advanced techniques like deep learning-based compression and adaptive sampling proving highly effective. The study concludes that innovative data reduction methods are essential for efficient big data management and emphasizes the importance of aligning these strategies with specific application requirements, paving the way for future advancements in scalable and optimized storage frameworks.

Categories: Big Data and Analytics, Data Mining, Storage Systems
Keywords: data compression, data preprocessing, data deduplication, dimensionality reduction, data reduction

Introduction And Background

The rapid increase in Big Data has created significant challenges in data storage, processing, and management. As organizations and systems increasingly rely on large datasets, efficient storage solutions are essential for optimizing resource utilization and ensuring data accessibility. To address these challenges, several Big Data storage reduction strategies have been developed. These strategies aim to reduce the amount of storage space required while maintaining the integrity and usability of the data. Key strategies explored in this survey encompass pre-processing, deduplication, dimensionality reduction, and compression. Pre-processing methods refine data for efficient storage by removing noise and irrelevant information. Deduplication eliminates duplicate data copies to optimize storage space. Dimensionality reduction techniques reduce the range of capabilities in data without significant data loss, while compression strategies reduce storage size by encoding data more efficiently. This survey provides a comprehensive evaluation of these techniques, comparing their strengths, weaknesses, and application areas in the context of Big Data systems [1]. Despite significant progress in Big Data storage reduction, several research gaps remain. Scalability is still a mission, as many reduction strategies, particularly dimensionality reduction and compression, struggle to scale effectively with the increasing volume of Big Data, especially in distributed environments. Real-time processing is another hurdle; while pre-processing and deduplication are commonly used strategies, their application in real-time systems remains difficult, as they may introduce latency or inefficiencies in time-sensitive packages. Moreover, data heterogeneity presents a project, as there is a loss of comprehensive studies addressing the integration of storage reduction techniques across heterogeneous data sorts, together with structured semi-structured, and unstructured data, and throughout various systems like cloud computing. Loss of data is another subject, as techniques like dimensionality reduction and compression can lead to the loss of vital data, requiring in addition research to expand strategies that strike a balance between reduction efficiency and data preservation [2]. Privacy and security also remain underexplored, in particular with regard to the effect of storage reduction strategies on data protection and confidentiality in sensitive applications. Finally, there may be a want for adaptive techniques which could adjust robotically based on the form of data and specific requirements of a given application, taking into account more green and tailor-made storage solutions [3]. Closing these gaps will pave the way for creating more sustainable data storage solutions.

Background of the study

Big data is defined by the "5Vs": volume, variety, velocity, veracity, and value. Volume refers to the massive scale of data, while variety encompasses different data types such as text, images, and videos. Velocity highlights the rapid generation of data, veracity ensures data reliability, and value represents the insights gained. Managing big data poses challenges due to storage costs, bandwidth limitations, and

computational constraints. To address these issues, data reduction techniques like compression, sampling, and feature selection optimize storage and processing efficiency [4]. These methods help retain essential information while reducing redundancy, making them vital for applications in healthcare, finance, and research. The study focuses on overcoming inefficiencies in traditional storage and processing frameworks as data volumes continue to grow. The cost of storage infrastructure increases with the need for larger capacities to accommodate massive datasets. Moreover, data centers consume significant energy, raising environmental concerns. Storage limitations present significant challenges in handling large datasets, particularly in big data environments. Processing bottlenecks arise as analytical and machine learning models require substantial computational resources, leading to delays that can compromise the timeliness of insights, especially in real-time applications [5]. Additionally, bandwidth constraints become a concern when transferring massive volumes of data over networks, resulting in latency issues and increased operational costs. Scalability challenges further exacerbate the problem, as traditional storage and processing systems struggle to expand efficiently, leading to decreased performance and reliability. Another critical issue is data redundancy and noise, where large datasets often contain duplicate or irrelevant information. This not only inflates storage requirements but also degrades the quality of analytical results. Effectively filtering out redundant data is essential to optimizing storage, improving processing efficiency, and enhancing the overall accuracy of insights derived from big data [6]. These challenges underscore the urgent need for effective data reduction methods that can mitigate the limitations of existing systems. By reducing the size of datasets while preserving their essential information, data reduction techniques enable organizations to achieve scalable and cost-effective solutions for big data management.

Objectives of the study

This study provides a comprehensive survey of big data reduction techniques, focusing on their principles, methodologies, and applications. It categorizes key methods like data compression, sampling, dimensionality reduction, and feature selection, analyzing their strengths and limitations. Recent advancements, including deep learning-based compression and hybrid frameworks, are reviewed for their effectiveness in modern data challenges. A comparative analysis of techniques offers insights into performance, efficiency, and trade-offs for practical applications. The study connects theoretical research with practical application, showcasing effective realizations. Additionally, it identifies research gaps and explores future directions, such as integrating data reduction with emerging technologies like edge computing and federated learning. This study aims to contribute to efficient and scalable storage frameworks for big data, helping organizations optimize data management strategies. Data reduction techniques are crucial across various domains, including healthcare for personalized medicine, finance for fraud detection, and scientific research for analyzing large datasets. These methods enable better storage, processing, and analysis of massive data volumes. The survey serves as a valuable resource for researchers and practitioners, addressing current challenges while paving the way for future innovations in big data management.

Review

Categorization of big data storage reduction

Data Preprocessing

Data pre-processing is a crucial step in data mining that enhances data quality by removing errors, inconsistencies, and redundancies. Since raw data is often noisy and incomplete, pre-processing ensures it meets the necessary standards for analysis. This stage is essential in the Big Data Lifecycle, as it refines data before it is used for decision-making. Data pre-processing involves several key steps to improve data quality and ensure its suitability for analysis. The first step, noise identification, detects and eliminates meaningless fluctuations in data using techniques like the binning method, which smooths sorted data by dividing it into equal-sized segments, and regression, which fits data to a regression function to reduce noise [7]. Next, data cleaning corrects corrupt or faulty records, such as misspellings, missing values, and typographical errors, improving the accuracy of analytical models. A pre-processing framework has been developed to upgrade data quality in weather tracking applications. The framework addresses challenges including data cleaning, noise elimination, integration, and transformation to ensure accurate analysis. The study highlights the significance of pre-processing in enhancing the reliability of climate forecasting, especially when considering global warming parameters [8]. It discusses the Big Data Lifecycle, emphasizing the role of high-quality data management. The proposed framework aims to enhance data consistency, integrity, and accuracy. The study concludes that pre-processing is critical in reducing errors and improving decision-making in weather-related data analysis. It presents a complete evaluation of pre-processing techniques for big data in smart grids. It discusses the function of pre-processing in improving power demand forecasting and grid performance. The study categorizes pre-processing into four key steps: data cleansing, reduction, integration, and transformation. It compares various pre-processing techniques used in smart grid packages and evaluates their effectiveness. The study highlights challenges such as managing large-scale, real-time data streams from smart meters and sensors. Future research directions focus on optimizing pre-processing techniques for better grid performance [9]. This study also explores novel data reduction techniques for processing bathymetric (seafloor mapping) information,

introducing a fusion of reduction strategies: True Bathymetric Data Reduction and Optimum Dataset. These methods aim to preserve accuracy while significantly reducing the volume of hydrographic data. The research validates the proposed fusion method through real-world tests in marine environments. The results demonstrate that this approach improves data processing efficiency while retaining essential spatial information, contributing to advances in hydrographic data analysis and navigation safety [10].

The study presents opinions about data pre-processing techniques used in data evaluation, categorizing pre-processing into three main steps: data cleaning, data transformation, and data reduction. It investigates different algorithms and their effectiveness in enhancing data quality, highlighting the importance of pre-processing in handling missing values, filtering noise, and normalizing data. The research finds that flawed pre-processing leads to poor model performance in data mining and machine learning. It concludes that an effective pre-processing approach is crucial for accurate and reliable data evaluation [11]. It focuses on pre-processing big data collected from tunnel boring machine (TBM) operations, discussing challenges such as noise filtering, anomaly classification, and handling missing data. The study introduces a data-processing framework that identifies and corrects errors caused by mechanical faults and human intervention. A machine learning-based approach is proposed to improve TBM performance predictions, and it is concluded that proper pre-processing significantly enhances the reliability of machine learning models in tunnel excavation tasks. The findings underscore the importance of data quality in predictive analytics for engineering packages [12]. Data normalization then scales all variables to a standard range, such as 0 to 1, eliminating skewness and improving processing efficiency. Data transformation converts raw data into a format suitable for integration and analysis through methods like attribute selection, which creates new attributes; discretization, which categorizes numerical values; and concept hierarchy generation, which aggregates attributes into higher-level categories (e.g., changing "City" to "Country"). Finally, data integration merges data from multiple sources into a unified format, a critical step in managing growing data volumes, especially in machine learning applications. These pre-processing steps ensure that data is clean, structured, and optimized for further analysis. Without proper pre-processing, raw data remains unstructured and unreliable, making it unsuitable for analysis [13]. Overall, data pre-processing ensures accuracy, completeness, and consistency, enabling organizations to derive valuable insights and make informed decisions.

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible. This helps simplify models, improve performance, and facilitate data visualization. It projects high-dimensional data into a lower-dimensional space, preserving essential patterns. Machine learning often requires extensive resources to process high-dimensional data, making dimensionality reduction a crucial step. Dimensionality reduction is categorized into two main approaches as feature selection and feature extraction. Feature selection retains only the most relevant variables, discarding irrelevant ones to improve model accuracy and efficiency. Feature selection can be supervised, relying on labeled data, or unsupervised, operating without labels [14]. Feature extraction, also known as feature projection, transforms high-dimensional data into a low-dimensional form. Key techniques include Principal Component Analysis (PCA), which transforms correlated features into uncorrelated principal components; Linear Discriminant Analysis (LDA), a supervised method that maximizes class separation; and t-Distributed Stochastic Neighbor Embedding (t-SNE), used for visualizing high-dimensional data. These techniques effectively handle complex datasets while preserving essential information. These techniques are applied to datasets such as cardiotocography, diabetic retinopathy, and intrusion detection systems [15]. PCA consistently outperforms LDA in maintaining classification performance. For large datasets, PCA improves the accuracy, specificity, and sensitivity of classifiers. PCA and LDA significantly influence the performance of machine learning classifiers. Random forest and support vector machine (SVM), when combined with PCA, produce the best results. For small datasets, dimensionality reduction can adversely affect model performance [16]. The authors propose a supervised dimensionality reduction method that uses class-conditional moments, introducing Linear Optimal Low-rank Projection (LOL) for scalable classification, which outperforms PCA, LDA, and other established techniques in classification accuracy. The study tests LOL on brain imaging and genomics datasets with millions of features, showing that LOL delivers higher classification accuracy and computational efficiency, confirming its effectiveness for large-scale classification tasks [17].

The review explores feature selection and extraction strategies to tackle the "curse of dimensionality," analyzing key methods like PCA, LDA, and deep learning-based approaches. It assesses the strengths and weaknesses of different dimensionality reduction techniques, emphasizing the challenges of maintaining essential data properties during dimension reduction. It underscores the need for scalable solutions to manage growing data complexity. It proposes hybrid approaches that integrate feature selection and extraction to enhance performance [18]. The Word Pattern Similarity Function (WP-SC) is designed for clustering high-dimensional text data. It employs dimensionality reduction techniques to maintain classification accuracy while lowering computational costs. WP-SC shows superior performance compared to LDA-SVM and adaptive firefly optimization with multi-kernel SVM. It achieves a higher data compression ratio through dimension reduction, enhancing classification accuracy. WP-SC attains 99.95% accuracy on high-dimensional datasets, surpassing existing methods. The results confirm WP-SC's

effectiveness for large-scale text classification [19]. It introduced a polar projection-based algorithm for dimensionality reduction in smart grid data. It uses K-means clustering with a novel distance metric that accounts for peak consumption. It introduces a stochastic approach to speed up cluster center identification. Results demonstrate a significant reduction in computational complexity compared to traditional methods. It achieves low estimation errors while enhancing the efficiency of real-time data analysis. The method is particularly effective for processing smart meter data and optimizing tariffs [20].

Data Compression

The research introduces a selective data compression scheme that leverages compressibility prediction to optimize storage and processing efficiency. By assessing whether data segments are compressible, the approach minimizes unnecessary computational overhead, offering a more efficient alternative to traditional uniform compression methods. Instead of compressing all information, the method determines whether a bit is compressible, lowering needless computational overhead. It employs Huffman coding and LZ77 encoding in mixture with field-programmable gate array-based acceleration to improve compression efficiency. The experimental results display a 34.15% boom in throughput with a minor 0.09% decrease in compression ratio [21]. This study presents an innovative graph-based data compression technique employing graph wedgelets, which extend geometric wavelet methods to graph structures, enabling efficient data representation and processing. It applies adaptive partitioning to optimize compression efficiency, specializing in piecewise steady representations. The proposed technique allows for green sparse representations, making it in particular beneficial for photograph and sign compression tasks [22]. The paper additionally provides theoretical foundations and experimental validation showing aggressive overall performance. This study proposes an enhanced security framework for large-scale data storage by integrating blockchain technology with a path compression algorithm, optimizing both data protection and storage efficiency. The proposed version ensures confidentiality, integrity, and availability by using managing get admission to through a special key-primarily based structure [23]. Path compression improves performance with the aid of decreasing the variety of nodes in blockchain transactions, optimizing garage utilization. The study shows that the method improves security while also increasing data retrieval speed. The Hadoop Data Reduction Framework (HDRF) combines lossless compression and deduplication within the Hadoop Distributed File System layer. By minimizing redundant facts garage and optimizing compression strategies, HDRF appreciably reduces storage requirements whilst keeping high throughput. Experimental results indicate up to 48.96% better throughput with a storage reduction exceeding 50% in positive datasets [24]. Various data compression techniques, including lossless and lossy methods, demonstrate their applications across diverse fields. The paper classifies compression algorithms by their redundancy elimination and encoding methods, highlighting their importance in optimizing big data storage. It also examines recent advancements in compression and addresses emerging challenges in reducing storage needs [25].

Data Deduplication

Data deduplication is a crucial technique for reducing redundant copies of data, optimizing storage usage, and improving efficiency. In today's data-driven world, where organizations manage vast amounts of information, deduplication helps control storage costs and enhances system performance. It ensures that duplicate files or data blocks are removed while maintaining system integrity [26]. Deduplication is particularly beneficial in environments with frequent backups, snapshots, and virtual desktop infrastructure images, where repeated data storage is common. The primary types of deduplication include file-level, block-level, and byte-level deduplication, each offering varying levels of granularity and efficiency [27]. An analytical model to predict the performance of deduplication systems evaluates the effect of various chunking parameters. It is revealed that the anticipated variable chunk size and minimum bite size appreciably affect deduplication efficiency. The proposed model predicts greatest deduplication parameters a lot faster than machine searches, attaining comparable consequences with as much as a 19.8× speedup. The results display that optimizing deduplication parameters can reduce storage utilization appreciably as compared to default configurations [28]. The research validates its findings via empirical experiments across different datasets. Some introduces Metadata Region Adaptive Container Structure (MACS), a metadata scheme designed to enhance garage performance and repair overall performance in deduplication systems. MACS complements container utilization via optimizing metadata storage, leading to a higher container filling ratio of over 95%. Experiments display that MACS improves repair overall performance by means of minimizing disk I/O operations, reducing the variety of box reads during data restoration. Caching efficiency has also been examined, demonstrating that MACS efficaciously reduces redundant disk accesses by leveraging locality-conscious indexing [29]. An inline deduplication approach, InDe, dynamically adjusts the variety of vintage container references in keeping with section to optimize restore performance. The gadget introduces a metric, Valid Container Referenced Count, to prioritize boxes with high utilization. The proposed F-greedy and F-greedy+ algorithms enhance restore velocity by 1.3× to 2.4× while maintaining high backup performance. Experimental results verify that InDe outperforms existing deduplication methods in terms of repair performance, making it a valuable answer for garage optimization [30]. The study examines various deduplication strategies and their effect on garage optimization. It categorizes deduplication into area-based, time-based, and bite-based methods, studying their blessings and barriers. They have a look at highlights that variable-sized

deduplication gives higher overall performance compared to fixed-sized strategies. Additionally, it discusses the trade-offs among processing time and deduplication efficiency [31]. The studies conclude that an optimized deduplication method can shop up to 60% of storage space, enhancing cloud storage usage. Statistics-theoretic analysis of deduplication performance when repeated data segments are approximate in place of specific [32]. It introduces a probabilistic model where repeated strings go through random substitutions, affecting chunking performance. They have a look at evaluates each fixed-length and variable-length deduplication schemes, displaying that fixed-period tactics suffer from boundary-shift problems. The findings recommend that variable period deduplication achieves better compression ratios, specifically as supply entropy decreases. The research contributes to growing more strong deduplication strategies for real-world data [33].

File-level deduplication removes duplicate files by storing a single instance and replacing other copies with pointers. Block-level deduplication divides files into unique data blocks, achieving higher deduplication ratios. Byte-level deduplication provides the most granular approach by comparing individual bytes but is computationally intensive [34]. Deduplication can be performed at different stages, such as source-based (before transfer) and target-based (after transfer to storage). Additionally, inline deduplication occurs in real-time, while post-process deduplication happens after data is written. Advanced methods like content-aware and application-aware deduplication optimize storage for specific data types and applications [35]. Although deduplication improves storage efficiency and reduces bandwidth usage, it requires significant processing power, memory, and careful implementation. As data volumes grow, effective deduplication strategies become essential for managing storage infrastructure and ensuring optimal data management.

Table *1* presents a comprehensive literature summary of big data reduction methods, highlighting the reduction techniques, algorithms, datasets used, results, strengths, limitations, and research gaps. It covers various methods like dimensionality reduction and feature selection, detailing how each algorithm performs on specific datasets. The table also discusses the outcomes of applying these methods, their advantages, such as computational efficiency, and limitations, such as challenges in preserving data interpretability. Additionally, it identifies gaps in current research, particularly around improving scalability and integrating diverse reduction techniques.

Title	Reduction Method Used	Algorithm Used	Dataset	Results	Strengths	Limitations	Research Gaps
A Survey of Preprocessing Methods Used for Analysis of Big Data Originated From Smart Grids [6].	Data Cleansing, Data Reduction, Data Integration	Clustering, Binning, Regression	Smart Grid Data	Enhanced classification accuracy with feature selection	Improves forecasting accuracy, reduces redundant data	Processing large real-time datasets is challenging	Need for automated and scalable preprocessing techniques
Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application [7].	Moving Average, Noise Reduction Filter	Processing software	Excavation datasets from Yinsong and projects	Varies by dataset and method	Improves machine learning accuracy, cleans irrelevant data	Requires manual tuning for best results	Needs further automation validation in real-world scenarios
Significance and methodology: Preprocessing the big data for machine learning on TBM performance [26].	Data Cleaning & Reduction	Moving Average Method, Noise Reduction Filter	TBM Performance Data (29.52 km, 17,051 cycles)	Improved signal reliability by filtering irrelevant data	Automated anomaly detection	High computational cost for real-time processing	Need for lightweight AI-driven optimization techniques
A comprehensive review on data preprocessing techniques in data analysis [27].	Data Filtering, Noise Reduction	Machine Learning-based Anomaly Detection	IoT Sensor Data	96-99% data reduction while maintaining accuracy	Lightweight, efficient for IoT devices	not generalize across different IoT environments	Need for adaptive models for diverse IoT scenarios
Analysis of Dimensionality Reduction Techniques on Big Data [10].	Principal Component Analysis (PCA), Linear Optimal Low-Rank Projection	LDA, Canonical Correlation Analysis	Brain imaging datasets (150M+ features), Genomics datasets (500k+ features)	Significant reduction while maintaining classification accuracy	Scales to millions of dimensions, strong statistical guarantees, computationally efficient	May lose non-linear relationships in data	Need for better generalization across diverse datasets

Supervised dimensionality reduction for big data [11].	PCA, Linear Discriminant Analysis (LDA)	Decision Tree, SVM, Naive Bayes, Random Forest	Cardiotocography, Diabetic Retinopathy, Intrusion Detection System	PCA outperforms LDA in high-dimensional settings	Reduces computation load, improves model performance	Limited to linear methods, does not always enhance performance for low-dimensional data	Hybrid approaches integrating multiple techniques
A Novel Clustering and Matrix Based Computation for Big Data Dimensionality Reduction and Classification [12].	Feature Selection, Feature Extraction	Genetic Algorithm, Ant Colony Algorithm, Neural Networks	Various high-dimensional datasets	Varies by dataset	Improves pattern recognition, enhances machine learning efficiency	Loss of interpretability in complex transformations	Need for adaptive reduction techniques based on dataset characteristics
Fast Big Data Analytics for Smart Meter Data [14].	Feature Extraction, PCA-based methods	Machine Learning-based dimensionality reduction	Various machine learning datasets	Improved efficiency while retaining key features	Enhances classification performance, reduces overfitting	Dependent on quality of extracted features	Investigation of non-linear feature extraction methods
Low-Overhead Compressibility Prediction for High-Performance Lossless Data Compression [15].	Selective Data Compression	DEFLATE (LZ77 + Huffman Coding)	FPGA-based benchmark datasets	34.15% throughput improvement, 0.09% compression ratio decrease	Reduces computational overhead, improves compression efficiency	Minor decrease in compression ratio (0.09%)	Further optimization of compressibility prediction methods
New Blockchain Based Special Keys Security Model With Path Compression Algorithm for Big Data [17].	Lossy Compression	SZ, ZFP, Gzip, LZ4, Zstd	ML/AI datasets (CANDLE, Superconductor, Ptychonn)	50-100× compression with <1% loss in accuracy	High compression ratio with minimal quality loss	Selection of optimal compression parameters is complex	Further exploration of adaptive lossy compression for ML models
Hadoop Data Reduction Framework: Applying Data Reduction at the DFS Layer [18].	Hybrid (Lossy + Lossless) Compression	SAX Quantization, LZW Compression	Intel Lab Sensor Data, Simulated IoT Data	Improved network efficiency, extended sensor lifetime	Efficient for IoT sensor networks, reduces transmission overhead	Accuracy loss due to lossy quantization	Optimization for diverse IoT environments
Data Deduplication With Random Substitutions [23].	Cloud Storage Deduplication	Bit-Level Representation for Deduplication	Cloud storage environments (Amazon S3, Microsoft Azure)	90-95% storage reduction	Efficient cloud storage utilization, reduced network bandwidth	Security concerns in encrypted deduplication	Enhancing deduplication with privacy-preserving techniques
Predicting Deduplication Performance: An Analytical Model and Empirical Evaluation [19].	Deduplication, Compression	LZ4, Gzip, Deduplication schemes	Wikipedia dump (400GB), Image dataset (35GB)	Up to 48.96% higher throughput. Significant storage savings	Modular design, adaptable to multiple datasets	Limited evaluation on real-world large-scale deployments	Integration with emerging storage technologies
A Cost-Efficient Metadata Scheme for High-Performance Deduplication Systems [20].	Metadata-Adaptive Deduplication	Metadata Region Adaptive Container Structure	Real-world backup datasets	95%+ average container filling ratio	High space utilization, improved restore performance	Limited impact on highly fragmented datasets	Exploring integration with existing rewriting methods

TABLE 1: Summary of Various Big Data Reduction Techniques

DFS, Distributed File System; FPGA, Field-Programmable Gate Array; IoT, Internet of Things; LZW, Lempel-Ziv-Welch; ML/AI, Machine Learning/Artificial Intelligence; SAX, Symbolic Aggregate approXimation; SVM, Support Vector Machine; TBM, Tunnel Boring Machine

Analysis of research gaps and potential enhancements in research

The analysis of research gaps in data reduction methods highlights critical areas for improvement, emphasizing the need for adaptive, automated preprocessing techniques and hybrid approaches that balance data retention and reduction. Real-time processing remains a challenge, particularly in spatial and TBM performance data, necessitating lightweight AI-driven optimization methods [36]. Standardization issues across fields call for unified benchmarking frameworks, while multi-source data fusion in smart grid and weather monitoring requires robust integration algorithms. In data deduplication, existing techniques struggle with diverse data types, demanding optimized strategies for text, audio, image, and video, as well as adaptive chunking models [37]. Additionally, balancing storage efficiency and retrieval speed remains underexplored, requiring advanced probabilistic frameworks. Dimensionality reduction methods, including PCA, LDA, and feature selection techniques, face scalability and adaptability issues, necessitating hybrid deep learning-based models and privacy-preserving approaches [38]. Furthermore, computational efficiency and interpretability need enhancement for broader AI-driven applications. Reduction techniques such as low-overhead compressibility prediction, graph wedges adaptive compression, block chain-based path compression, and the Hadoop data reduction framework require predictive model refinements, computational optimizations, scalability improvements, and real-world testing to enhance their effectiveness and applicability across diverse domains [39].

Evaluation parameters of data reduction methods

1. Compression Ratio: Measures the effectiveness of compression by means of comparing the unique information size to the compressed size. A better compression ratio suggests better compression efficiency.

2. Data Reduction Ratio: Evaluates how much data is eliminated during preprocessing. Higher data reduction ratio values signify effective data reduction, improving storage and processing efficiency.

3. Information Loss: Assesses the loss of data first-class due to reduction strategies using metrics like root mean squared error and peak signal-to-noise ratio. Lower values make certain higher retention of data integrity.

4. Feature Retention Rate: Determines the percentage of essential features retained post dimensionality reduction. A higher feature retention rate ensures minimal loss of critical information.

5. Computational Efficiency: Measures the time and resources required for data reduction strategies. Optimized techniques have to offer great reduction with minimal processing overhead.

Evaluating data reduction methods based on these parameters can optimize techniques to enhance storage efficiency, processing speed, and data integrity while minimizing losses. The studies included in the review were chosen based on their direct contribution to the research gaps identified. These gaps focused on areas where existing methods were lacking or where further enhancement could lead to significant improvements in practical applications. For instance, studies proposing hybrid approaches that combine traditional techniques with newer AI-driven models were prioritized [40]. Research addressing real-time processing challenges in domains like spatial data and TBM performance was highlighted. Advanced methods for fusing data from multiple sources (e.g., weather monitoring and smart grids) were particularly relevant. Proposals for deep learning-based hybrid models for scalability and privacy preservation in dimensionality reduction were prioritized. This approach ensured that the most relevant and innovative studies in data reduction methods were included, contributing to a comprehensive analysis of research gaps and potential enhancements in the field.

Conclusions

The comprehensive survey of big data reduction methods reveals a dynamic landscape of techniques aimed at managing the exponential growth of data across industries. Feature selection and extraction methods, such as PCA, LDA, and deep learning-based approaches, effectively address the curse of dimensionality while preserving critical information. Advanced clustering techniques, like WP-SC and polar projection-based algorithms, demonstrate superior performance in high-dimensional text and smart grid data processing. Data deduplication strategies, leveraging lossless compression and probabilistic models, optimize storage efficiency without compromising retrieval speeds. Emerging methods, including LOL and blockchain-based path compression, show promise in scalable classification and computational efficiency. The survey highlights the pivotal role of these techniques in domains like healthcare, where AI-driven, privacy-preserving solutions enhance medical imaging, genomic analysis, and patient monitoring. However, challenges remain, including the need for adaptive preprocessing, real-time optimization, and standardized frameworks to ensure cross-domain applicability. Future research should focus on integrating hybrid approaches, improving interpretability, and developing energy-efficient solutions to unlock the full potential of big data reduction, enabling smarter decision-making and innovation across sectors.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Samita A. Patil, Savita Sangam

Drafting of the manuscript: Samita A. Patil, Savita Sangam

Critical review of the manuscript for important intellectual content: Samita A. Patil, Savita Sangam

Acquisition, analysis, or interpretation of data: Savita Sangam

Supervision: Savita Sangam

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

We would like to express our gratitude to Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai, for providing the necessary resources and support for this study. We also extend our appreciation to our Principal, Dr. Bhavesh Patel, for his valuable insights and guidance throughout the research process. Additionally, we acknowledge the contributions of Dr. Pramod Bhavarthe for his support in facilitating this work. Lastly, we thank the open-source community and researchers whose work has laid the foundation for advancements in Big Data Storage Techniques.

References

1. Fernandes V, Carvalho G, Pereira V, Bernardino J: Analyzing data reduction techniques: an experimental perspective. Applied Sciences. 2024, 14:3436. 10.3390/app14083436
2. Haeri H, Kathiriya N, Chen C, Jerath K: Adaptive granulation: data reduction at the database level. Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. 2023, 3:29-39. 10.5220/0012190700003598
3. Hancock JT, Wang H, Khoshgoftaar TM, Liang Q: Data reduction techniques for highly imbalanced medicare Big Data. Journal of Big Data. 2024, 11:8. 10.1186/s40537-023-00869-3
4. Gyamerah S, Tour Soori G, Redeemer Korda D, Kwame Tawiah J, Ayintareba Akolgo E, Oteng Dapaah E: Comparative analysis of feature extraction of high dimensional data reduction using machine learning techniques. American Journal of Electrical and Computer Engineering. 2023, 7:27-39. 10.11648/j.ajece.20230702.12
5. Khan S, Alam M: Preprocessing framework for scholarly big data management. Multimedia Tools and Applications. 2022, 82:39719-39743. 10.1007/s11042-022-13513-8
6. Alghamdi TA, Javaid N: A survey of preprocessing methods used for analysis of big data originated from smart grids. IEEE Access. 2022, 10:29149-29171. 10.1109/ACCESS.2022.3157941
7. Juneja A, Das NN: Big data quality framework: pre-processing data in weather monitoring application. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). 2019, 559-563. 10.1109/COMITCon.2019.8862267
8. Wlodarczyk-Sielicka M, Blaszczak-Bak W: Processing of bathymetric data: the fusion of new reduction methods for spatial big data. Sensors. 2020, 20:6207. 10.3390/s20216207
9. Jahin MA, Shovon MSH, Shin J, Ridoy IA, Mridha MF: Big data—supply chain management framework for forecasting: data preprocessing and machine learning techniques. Archives of Computational Methods in Engineering. 2024, 31:3619-3645. 10.1007/s11831-024-10092-9
10. Reddy GT, Lakshmanna K, Kaluri R, Singh Rajput D, Srivastava G: Analysis of dimensionality reduction techniques on big data. IEEE Access. 2020, 8:54776-54788. 10.1109/access.2020.2980942
11. Vogelstein JT, Bridgeford EW, Tang M, Zheng D, Douville C, Burns R, Maggioni M: Supervised dimensionality reduction for big data. Nature Communications. 2021, 12:2872. 10.1038/s41467-021-23102-2
12. Varghese J, Tamil Selvan P: A novel clustering and matrix based computation for big data dimensionality reduction and classification. Journal of Advanced Research in Applied Sciences and Engineering Technology. 2023, 32:238-251. 10.37934/araset.32.1.238251
13. Jia W, Sun M, Lian J, Hou S: Feature dimensionality reduction: a review. Complex & Intelligent Systems. 2022, 8:2663-2693. 10.1007/s40747-021-00637-x
14. Mohajeri M, Ghassemi A, Gulliver TA: Fast big data analytics for smart meter data. IEEE Open Journal of the Communications Society. 2020, 1:1864-1871. 10.1109/OJCOMS.2020.3038590
15. Kim Y, Choi S, Lee D, Jeong J, Kwak J, Lee J: Low-overhead compressibility prediction for high-performance

lossless data compression. IEEE Access. 2020, 8:37105-37123. 10.1109/ACCESS.2020.2975929
16. Erb W: Graph wedgelets: Adaptive data compression on graphs based on binary wedge partitioning trees and geometric wavelets. IEEE Transactions on Signal and Information Processing over Networks. 2023, 9:24-34. 10.1109/tsipn.2023.3240899
17. Bakir C: New blockchain based special keys security model with path compression algorithm for big data. IEEE Access. 2022, 10:94738-94753. 10.1109/access.2022.3204289
18. Widodo RNS, Abe H, Kato K: Hadoop data reduction framework: Applying data reduction at the DFS layer. IEEE Access. 2021, 9:152704-152717. 10.1109/access.2021.3127499
19. Randall O, Lu P: Predicting deduplication performance: An analytical model and empirical evaluation. IEEE International Conference on Big Data (Big Data). 2022, 319-328. 10.1109/BigData55660.2022.10020871
20. Mo Y, Hua Y, Li P, Cao Q, Liu X: A cost-efficient metadata scheme for high-performance deduplication systems. 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). 2021, 49-56. 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00034
21. Lin L, Deng Y, Zhou Y, Zhu Y: InDe: An inline data deduplication approach via adaptive detection of valid container utilization. ACM Transactions on Storage. 2023, 19:1-27. 10.1145/3568426
22. Sujatha G, Raj JR: A comprehensive study of different types of deduplication technique in various dimensions. International Journal of Advanced Computer Science and Applications. 2022, 13:10.14569/ijacsa.2022.0130339
23. Lou H, Farnoud F: Data deduplication with random substitutions. IEEE Transactions on Information Theory. 2022, 68:6941-6963. 10.1109/TIT.2022.3176778
24. Vijayalakshmi K, Jayalakshmi V: Analysis on data deduplication techniques of storage of big data in cloud. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). 2021, 976-983. 10.1109/ICCMC51019.2021.9418445
25. Kumar N, Shobha, Jain SC: Efficient data deduplication for big data storage systems. Progress in Advanced Computing and Intelligent Engineering. Panigrahi C, Pujari A, Misra S, Pati B, Li KC (ed): Springer, Singapore; 2019. 714:351-371. 10.1007/978-981-13-0224-4_32
26. Xiao HH, Yang WK, Hu J, Zhang YP, Jing LJ, Chen ZY: Significance and methodology: Preprocessing the big data for machine learning on TBM performance. Underground Space. 2022, 7:680-701. 10.1016/j.undsp.2021.12.003
27. Çetin V, Yýldýz O: A comprehensive review on data preprocessing techniques in data analysis. Pamukkale University Journal of Engineering Sciences. 2022, 28:299-312. 10.5505/pajes.2021.62687
28. Jeong H, Seo G, Hwang E: Lossless data compression with bit-back coding on massive smart meter data. 2022 IEEE International Conference on Big Data (Big Data). 2022, 6667-6669. 10.1109/BigData55660.2022.10020726
29. Al-Tammie HK, Al-Anber NJ: A comparative dimensionality reduction study in diabetes patients using LDA and PCA. 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA). 2022, 170-175. 10.1109/IT-ELA57378.2022.10107946
30. Song L, Ma H, Wu M, Zhou Z, Fu M: A brief survey of dimension reduction. Intelligence Science and Big Data Engineering. Peng Y, Yu K, Lu J, Jiang X (ed): Springer, Cham; 2018. 11266:189-200. 10.1007/978-3-030-02698-1_17
31. Albattah W, Khan RU, Khan K: Attributes reduction in big data. Applied Sciences. 2020, 10:4901-4901. 10.3390/app10144901
32. Yang H: The impact of the double reduction policy on the educational anxiety of parents under big data. International Journal of Sociotechnology and Knowledge Development. 2024, 16:1-16. 10.4018/ijskd.342131
33. Gandor T, Nalepa J: First gradually, then suddenly: Understanding the impact of image compression on object detection using deep learning. Sensors. 2022, 22:1104. 10.3390/s22031104
34. Jayasankar U, Thirumal V, Ponnurangam D: A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. Journal of King Saud University - Computer and Information Sciences. 2021, 33:119-140. 10.1016/j.jksuci.2018.05.006
35. Peng M, Southern DA, Ocampo W, et al.: Exploring data reduction strategies in the analysis of continuous pressure imaging technology. BMC Medical Research Methodology. 2023, 23:56. 10.1186/s12874-023-01875-y
36. Ragavan N, Yesubai Rubavathi C: A novel big data storage reduction model for drill down search. Computer Systems Science and Engineering. 2022, 41:373-387. 10.32604/csse.2022.020452
37. García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F: Big data preprocessing: methods and prospects. Big Data Analytics. 2016, 1:9. 10.1186/s41044-016-0014-0
38. Poletaev AY, Spiridonova EM: Hierarchical clustering as a dimension reduction technique in the Markowitz portfolio optimization problem. Automatic Control and Computer Sciences. 2022, 55:809-815. 10.3103/s0146411621070270
39. Gahar RM, Arfaoui O, Hidri MS, Hadj-Alouane NB: A distributed approach for high-dimensionality heterogeneous data reduction. IEEE Access. 2019, 7:151006-151022. 10.1109/access.2019.2945889
40. Velliangiri S, Alagumuthukrishnan S, Iwin Thankumar Joseph S: A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science. 2019, 165:104-111. 10.1016/j.procs.2020.01.079