IADIS International Journal on Computer Science and Information Systems
Vol. 19, No. 2, pp. 19-36
ISSN: 1646-3692

A SYSTEMATIC MAPPING REVIEW ON DATA CLEANING METHODS IN BIG DATA ENVIRONMENTS

Cláudio Keiji Iwata, Napoleão Verardi Galegale, Márcia Ito,
Marília Macorin de Azevedo, Marcelo Duduchi Feitosa and Carlos Hideo Arima
CEETEPS – Centro Estadual de Educação Tecnológica Paula Souza, Brazil

ABSTRACT

The evolution of information technology combined with artificial intelligence, IoT (Internet of Things) and robotics has made processes integrated and intelligent. The increased use of technology and the need for evidence-based decisions have contributed to the rapid expansion of a large volume of data in recent years. The quality of data generated mainly by humans must be given special attention, as errors can occur more frequently, making the pre-processing phase, such as data cleaning, a determining factor for better results in data analysis. The aim of this article is therefore to analyze data cleaning methods applied in Big Data environments by conducting a systematic review. The review method was based on the Kitchenham protocol, and the search databases were Scopus, Web of Science and CAPES. After searching and selecting the articles according to the protocol, 69 articles were analyzed, revealing the use of a wide variety of techniques, such as machine learning, data mining, natural language processing and others. The review also emphasized the various publication formats and the wide dissemination and discussion of research on data cleaning in Big Data in the academic community. Finally, this study provides the state of the art of data cleansing techniques that have been used in a Big Data context, offering insights and directions for future research.

KEYWORDS

Data Cleaning, Big Data, Data Quality, Mapping Review

1. INTRODUCTION

Big Data is a heterogeneous concept that encompasses various types and volumes of digital information, along with specific analytical tools tailored to different industrial contexts (Manyika et al., 2011). The definition of Big Data hinges on organizations' ability to capture, manage, and process vast amounts of data that conventional computers cannot handle within an acceptable timeframe. Characterized by its novelty, Big Data technologies and architectures have yet to be fully integrated into commercial systems and management software (Chen et al., 2012). It is crucial to highlight the role of information security policies in processing large volumes of data (Galegale et al., 2017) and conducting risk analyses to adequately protect this data (Souza et al., 2020). Big Data analysis provides valuable insights across various domains and industries, including sales forecasting methods (Martins et al., 2022, 2023), doctor-patient relationship analysis (Ito et al., 2017), and studies aimed at developing solutions for children's learning in the health sector (Ito et al., 2013), among others.

Data refers to any set of observations or measurements that can be analyzed to generate insights or support decision-making. It can be structured or unstructured, and its quality is crucial for obtaining accurate and reliable insights. To ensure data quality, organizations must perform data preparation activities such as cleaning, transforming, and integrating data. Data quality is a critical factor for the success of data-driven decision-making, as inaccurate or incomplete data can lead to misleading insights (Wamba et al., 2017). Therefore, it is essential for organizations to adopt rigorous data management practices to ensure data quality.

Data cleaning is a fundamental process in preparing data for analysis, involving the identification and correction of errors, inconsistencies, and anomalies (Y.-Y. Zhang et al., 2021). It is a critical step in any data mining project, as "dirty" data can lead to inaccurate results and erroneous conclusions. The goal of data cleaning is to ensure that the data is correct, complete, and consistent, allowing it to be used confidently in analyses and decision-making (Hellerstein, 2008). Data cleaning can involve several steps, such as removing missing values, correcting typographical errors, standardizing formats, and detecting and removing outliers (Zhang et al., 2021).

Thus, data cleaning is a critical step in the Big Data analysis process and must be performed with care and attention to detail. It can be conducted manually or through automated algorithms, depending on the size and complexity of the data. Additionally, it can be an iterative process, as identifying errors may lead to new discoveries and the need for further data collection (Manyika et al., 2011).

The volume and complexity of data present significant challenges in data cleaning within Big Data contexts. Traditional methods may not be scalable or efficient enough to handle large and complex datasets. Furthermore, the heterogeneity and variability of data sources can pose substantial challenges, as different sources may have varying formats, structures, and quality levels. Another challenge is the need for real-time or near-real-time data cleaning, especially in applications that rely on streaming data. This necessitates fast and efficient algorithms capable of detecting and repairing anomalies in real-time, thereby minimizing the impact on overall system performance. Addressing these challenges requires the development of new methods and techniques, as well as the integration of data cleaning into the overall data management and analysis pipeline (Y.-Y. Zhang et al., 2021).

Organizations continue to be adversely affected by poor data quality as they struggle to extract value from their data. Recent studies estimate that up to 80% of the data analysis pipeline is consumed by data preparation tasks, such as data cleaning. A wide variety of data cleaning solutions have been proposed to reduce this effort, including constraint-based cleaning, statistical cleaning, and leveraging master data as a source of truth (Huang et al., 2020).

This study aims to identify the main data cleaning techniques used in Big Data, evaluate their effectiveness, and explore their applicability across different industrial sectors. Additionally, the study seeks to synthesize and validate previous research through the categorization of articles and the exploration of the relationships between Big Data and the practice of data cleaning.

2. RESEARCH METHODOLOGY

To consolidate evidence and deepen the understanding of the implementation of data cleaning practices in Big Data environments, a Systematic Mapping Review (SMR) was conducted with the aim of identifying the main sources, authors, techniques, and industrial sectors addressed by the articles exploring this topic. The conduct of this systematic review aimed to evaluate the scientific production related to data cleaning in Big Data contexts. Specific metrics were employed for the analysis and selection of articles, covering aspects such as production, dissemination, impact, and interrelationships among academic works.

This SMR was based on a Systematic Literature Review (SLR) because the research is broader and aims to map the challenges and impacts of the area by analyzing the articles. The Systematic Literature Review is a meticulous, clear, comprehensive, and replicable process aimed at identifying, evaluating, and summarizing the complete works produced by academic and professional researchers (Gallardo-Gallardo, 2016). The SMR must be comprehensive and independent, following a clear and explicit methodology to describe the procedures used. It is essential that the review covers all relevant material and is replicable by other researchers who wish to adopt the same approach in reviewing the topic in question (Okoli, 2019).

The online platform Parsifal - Performing a Systematic Literature Review (https://parsif.al/) was adopted for conducting the SLR. This platform enables the selection of pertinent articles directly from the Web of Science, Scopus, and CAPES databases. The choice of these databases was based on their wide coverage of academic publications and frequent content updates. No specific time restriction was established during the search period, except for the exclusion of the year 2023, the current year of the research, as the annual publication analysis could be hampered by incomplete data. The keywords were selected in accordance with the research objectives and the questions sought to be addressed. In this context, the research adopted the following keywords: "data cleaning" AND "big data" AND (technique OR process OR method).

The systematic review was based on the PRISMA-P protocol (Moher et al., 2015), which consists of four stages: identification, screening, eligibility, and selection of documents for critical analysis.

As an example, Iwata, Cláudio, and Ito (2023) provided a comprehensive analysis of data cleaning techniques, highlighting the need to consider innovative approaches in the review process. This framework contributed to the careful delineation of the inclusion and exclusion criteria, as well as guiding the selection of primary sources that form the knowledge base of this review.

In the review phase, three initial data points were entered into the tool: the title of the work, which coincides with the title of this article; the description, which encompasses the abstract of the article; and the names of the authors. In the planning stage, the first sub-step pertains to the definition of the protocol, in which the overall objective was formalized as mapping the publications related to data cleaning in Big Data, addressing techniques, processes, and procedures up to the year 2022. The specific objectives were formalized as follows:

- Demonstration of the publication landscape on the grouped themes: Data Cleaning, Big Data, and techniques, processes, and methods.
- Demonstration of the main groups of contributions resulting from a Systematic Literature Review.

The design of the inclusion and exclusion criteria for articles to be considered for analysis is an important step to be executed after the collection of articles by search engines.

Table 1. Formalization of PICOC Data

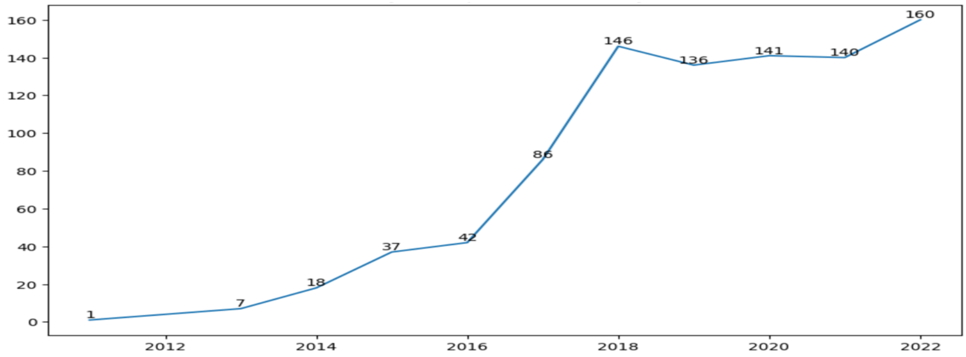
PICOC Method	Formalized Data
Population:	Scientific articles published in journals and at academic events.
Intervention:	Collection of data on machine learning applications in various industries.
Comparison:	Different applications of machine learning.
Outcome:	A report on the main applications observed.
Context:	Scientific articles published in journals and at academic events regarding studies conducted in various industries.

Source: Authors, 2023

The PICOC method—Population, Intervention, Comparison, Outcome, and Context - was employed to structure the data within the tool. Table 1 presents the formalized data for each field, providing a solid foundation for the systematic and rigorous conduct of the review.

The inclusion criteria were established as follows: (i) articles that discuss technology evaluation; (ii) articles that address data cleaning; and (iii) articles that respond to the research questions.

The exclusion criteria were defined as follows: (i) articles that are not in English; (ii) articles that do not meet the research criteria; and (iii) articles that are not primary studies.



Source: Iwata, Cláudio e Ito

Figure 1. Temporal Evolution of Publications

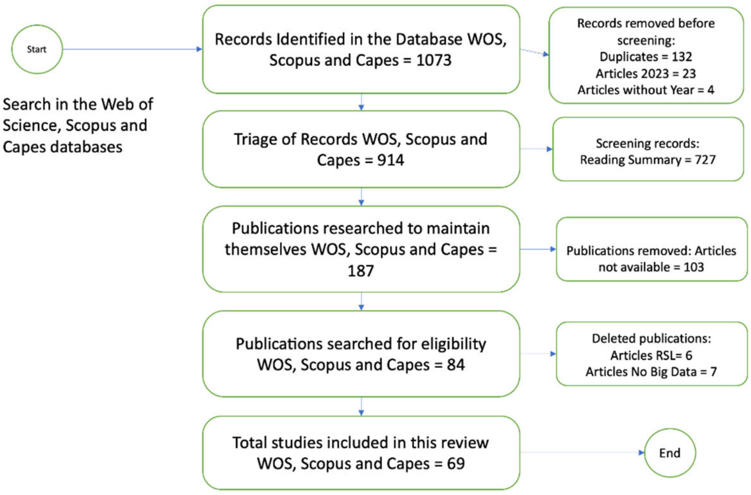
The graphic in Figure 1, extracted from Iwata and Ito (2023), illustrates the number of articles selected for the literature review. A total of 914 academic works were analyzed.

After applying the criteria, a total of 1,073 potential academic works were identified in the Scopus, Web of Science, and CAPES databases. Of this total, 27.56% had duplications across the databases, with 27.34% originating from Web of Science, 36.54% from Scopus, and 8.56% from CAPES. Subsequently, 132 duplicated works were removed, 23 articles published in 2023 were excluded, and 4 articles without date information were invalidated, resulting in a total of 914 academic works available for analysis.

Through the analysis of titles and abstracts, it was found that some works did not have a direct relationship with the topic in question or did not qualify as case studies. Consequently, 727 works were not selected as they did not contribute to the research objectives.

Additionally, 103 publications were excluded due to restrictions on free access, redirection to pages without a download link, or excessive delays in the download process. The remaining 84 articles were then distributed among the authors for full-text reading. At this stage, 6 articles that consisted of systematic literature reviews and 7 articles that were not clearly related to Big Data were identified and eliminated.

Figure 2 illustrates the flowchart based on the PRISMA-P protocol, detailing the step-by-step screening process that resulted in the selection of 69 articles from the initial 1,073.


Start
Search in the Web of Science, Scopus and Capes databases
Records Identified in the Database WOS, Scopus and Capes = 1073
Records removed before screening: Duplicates = 132 Articles 2023 = 23 Articles without Year = 4
Triage of Records WOS, Scopus and Capes = 914
Screening records: Reading Summary = 727
Publications researched to maintain themselves WOS, Scopus and Capes = 187
Publications removed: Articles not available = 103
Publications searched for eligibility WOS, Scopus and Capes = 84
Deleted publications: Articles RSL= 6 Articles No Big Data = 7
Total studies included in this review WOS, Scopus and Capes = 69
End

Source: Iwata, Cláudio e Ito

Figure 2. Flowchart based on the PRISMA-P protocol

3. RESULTS AND DISCUSSION

Table 2 presents a synthesis of the results of data cleaning practices in Big Data based on the 69 selected articles. The analysis of each article aimed to identify the following aspects:

- **The technical approach used**: This refers to the strategies and methodologies employed for data cleaning (e.g., benchmark, framework, method).
- **The specific technique**: This pertains to the specific techniques utilized for deduplication, handling missing values, error correction, and normalization, among others (e.g., algorithms, neural networks, machine learning).
- **Type of industry**: This refers to the specific data treatment requirements of each sector (e.g., health, finance, construction, manufacturing, transportation).
- **Database used**: This pertains to the origin and volume of the data utilized in the data cleaning process.

Table 2. Analysis of Articles

Author (Article)	Approach	Technique	Industry	Database
(Jäger et al. 2021)	Benchmark	The article benchmarks 69 diverse datasets, demonstrating that the proposed imputation method outperforms five alternative methods in various tasks	General	General
(Liu et al. 2020)	Framework	The article employs an enhanced approach combining the Local Outlier Factor (LOF) algorithm and Random Forest for outlier detection and missing data imputation in large electricity datasets	Electrical	Electrical
(Du et al. 2017)	Framework	The article utilizes the Conditionally Combined Functional Dependency (CCFD) technique, an innovative approach for identifying and correcting inconsistencies in related data.	General	Health/NBA
(Jin et al. 2018)	Model	The methodology in the article is graph-based, employing community detection algorithms and deep CNN models to clean an extremely large facial dataset. Additionally, the "TensorFlow" framework is used to train the models.	General	Celebrities
(Oni et al. 2019)	Method	The methodology outlines the datasets used, the four data cleaning tools evaluated, and the data cleaning tasks performed.	General	Climate
(X. Yang et al. 2018)	Method	The article involves a data cleaning method for large GPS trajectory datasets using movement consistency.	General	GPS
(Shen et al. 2021)	Method	The methodology in the article is based on a data cleaning algorithm for PMUs using an ensemble model and a soft voting approach. The algorithm is implemented in a big data environment utilizing the Apache Spark platform.	Electrical	Electrical
(Zhou et al. 2021)	Framework	The proposed methodology in the article involves using algorithms, frameworks, and big data tools to map suitable methodologies for health big data analysis based on user needs.	Health	Health
(J. Wang et al. 2014)	Framework	The article introduces a framework called Sample-and-Clean, which enables the application of any data cleaning technique on data samples for processing aggregated numerical queries in large dirty datasets.	General	Citation Index
(Feric et al. 2021)	Framework	The article describes a custom software architecture developed using the Django web framework for harmonizing biomedical data. It outlines the techniques, processes, and methods employed for cleaning, transforming, visualizing, and analyzing large biomedical datasets.	Biomedical	Biomedical
(Mok et al. 2017)	Framework	The methodology used in the article is the Reynolds Decomposition approach for cleaning financial security price data, based on data analysis and big data processing techniques.	Finance	Finance

Author (Article)	Approach	Technique	Industry	Database
(Abu Ahmad & Wang 2018)	Model	The methodology in the article is based on a multi-attribute weighted rule generation algorithm (MAWRG) and an entity resolution algorithm (MAWR-ER) that uses generated rules to identify entities in large bibliographic datasets.	General	Bibliographic
(Dong et al. 2018)	Framework	The article proposes an efficient data management framework to support an ongoing research project at the PROTECT Center in Puerto Rico, encompassing workflows for data import, cleaning, and secure transmission of privacy-sensitive data.	General	Health
(Elouataoui et al. 2022)	Model	The methodology in the article is based on a machine learning algorithm for real-time detection and removal of duplicate records in large datasets, along with a set of transformations required to prepare the data for deduplication.	General	None
(Rahul et al. 2020)	Benchmark	The article does not focus on a specific methodology but rather on various techniques, processes, and methods related to data cleaning in big data applications.	General	None
(Mavrogiorgos et al. 2022)	Method	The article proposes an automated rule-based data cleaning technique that utilizes Natural Language Processing (NLP) technology to ensure data reliability, detailing the technique and its components.	General	Health
(Kenda & Mladenić 2018)	Method	The article is based on a data cleaning technique using the Kalman filter, adaptable to conceptual drift and effective for detecting random additive outliers in sensor data streams, along with an unsupervised machine learning approach for automatic parameter tuning and an evaluation procedure based on indirect modeling.	General	Water Sensors
(De et al. 2014)	Framework	The methodology in the article is based on developing a data cleaning and query response system for structured big data using probabilistic models and query rewriting techniques, empirically evaluated with controlled and real datasets.	General	None
(Ribeiro et al. 2022)	Framework	The methodology in the article involves the description and presentation of a new tool, the bdc package, developed to standardize, integrate, and clean biodiversity data.	Biological	Biodiversity
(Fayyad et al. 2017)	Benchmark	The article does not employ a specific methodology but addresses issues related to benchmarks, process management, and data management tools and models in large-scale environments.	General	None
(T. Wang et al. 2020)	Method	The methodology in the article is based on experimental research, involving tests and analyses to evaluate the effectiveness of the proposed data cleaning method.	Manufacturing	Networks
(Prakash et al. 2019)	Benchmark	The article does not focus on a specific methodology but discusses techniques and	General	None

Author (Article)	Approach	Technique	Industry	Database
		tools for data preprocessing in big data environments.		
(Tang 2015)	Model	The methodology in the article involves presenting a proposal to address data quality issues in large RDF datasets, along with discussing specific techniques and tools for data cleaning.	General	RDF Data
(Tian et al. 2017)	Framework	The article presents a distributed stream data cleaning system called Bleach, which uses an incremental equivalence class algorithm to detect and repair data violations in real-time, detailing the system's architecture and performance optimizations.	General	None
(Yang et al. 2019)	Framework	The methodology in the article combines data cleaning techniques, deep recurrent neurais networks, and cross-validation to predict energy consumption in buildings, detailing a step-by-step approach for preprocessing data, training, and evaluating the prediction model.	Construction	Energy
(Staegemann et al. 2021)	Method	The article's methodology combines data cleaning techniques, deep recurrent neurais networks, and cross-validation in a step-by-step approach to preprocess data, train, and evaluate an energy consumption prediction model for buildings.	Construction	Energy
(Farid et al. 2016)	Framework	The article presents a system called CLAMS for cleaning data in large unstructured and semi-structured datasets, detailing the system's architecture, cleaning techniques, and the challenges faced by companies managing large volumes of data in "data lakes."	General	None
(Nugroho et al. 2021)	Benchmark	The article discusses techniques and algorithms for handling missing data in predictive systems, but does not specifically mention a methodology.	General	None
(P.-I. D. Lin et al. 2022)	Model	The article's methodology involves applying the growthcleanr algorithm to identify and correct errors in anthropometric data from electronic health records.	Health	Health
(Shende et al. 2022)	Framework	The article's methodology proposes an R package called cleanTS, which provides tools and a benchmarking system to compare different data cleaning techniques for univariate time series.	General	None
(F. Li et al. 2021)	Framework	The article's methodology includes applying the 5S data management approach to enhance data cleaning efficiency in a big data environment, along with a task allocation model considering indicators like time, production cost, product quality, and service quality for collaborative manufacturing.	Manufacturing	None
(W. Zhang & Tan 2019)	Benchmark	The article's methodology employs a machine learning technique for data cleaning in large datasets for supervised learning, combining robust deep autoencoder-based outlier detection with reconstruction error	General	None

Author (Article)	Approach	Technique	Industry	Database
		minimization to filter and label misclassified data.		
(Yousef 2015)	Framework	It proposes an enhanced generic framework for detecting duplicate records in large datasets based on rules and dictionaries, comparing current available tools and their detailed components, including similarity functions, classification algorithms, dictionary construction components, and blocking techniques.	General	None
(Z. Chen et al. 2019)	Framework	The article's methodology presents a scalable and customizable fuzzy join technique for large datasets, utilizing locality-sensitive hashing (LSH) to generate signatures for each record and create an index on the signatures in the reference table R.	General	None
(Y.-Y. Zhang et al. 2021)	Framework	The article's methodology proposes a data cleaning method that combines threshold-based techniques and clustering to enhance the quality of energy consumption data in buildings, along with an indicator to assess the quality of the cleaned data, relying on data processing algorithms and techniques.	Construction	Energy
(Xu et al. 2015)	Benchmark	The article does not present a specific methodology but discusses various data cleaning techniques and methods that can be applied in industrial processes.	General	None
(Bramantoro 2018)	Method	The article employs an experimental methodology, conducting comparative tests between data deduplication services in a data warehouse environment, based on performance metrics to determine the most effective service for data deduplication.	General	None
(Martinez-Mosquera et al. 2017)	Benchmark	The article proposes a comparative technique based on Fellegi-Sunter theory for data cleaning in security logs within a Big Data environment, emphasizing the importance of data preprocessing to enhance the efficiency of data mining processes.	General	None
(Y. Lin et al. 2019)	Method	The article's methodology is based on proposing new methods and algorithms for data source selection for information integration in the Big Data era, featuring a novel probabilistic approach, a scalable index-based algorithm, and two pruning strategies.	General	None
(Kim et al. 2019)	Model	The article proposes a new automatic standardization algorithm for categorical laboratory test results in electronic health records, called SALT-C (Standardization Algorithm for Laboratory Test - Categorical Results).	Health	Health
(Chiang et al. 2021)	Method	The article's methodology is based on electronic health records and involves deep data cleaning and phenotyping using ICD codes.	Health	Health

Author (Article)	Approach	Technique	Industry	Database
(S. Zhang et al. 2022)	Model	The article's methodology involves a Bayesian Networks-based algorithm to enhance data maintenance efficiency and quality in Big Data, utilizing a combination of data cleaning and mining technologies along with a manual Hadoop-based data cleaning architecture.	General	None
(Sainju et al. 2021)	Benchmark	The article does not specify the methodology used but mentions various techniques and algorithms for handling incomplete or missing data in flood mapping-based observations.	General	None
(Sheoran & Parmar 2022)	Framework	The article presents a data cleaning tool called GeoWebCln, designed to clean geospatial metadata, detailing the data cleaning process and its applicability to any geographic area.	General	Geospatial
(Chu et al. 2013)	Framework	The article employs a unified data cleaning approach that combines heterogeneous rule evidence encoded in a conflict hypergraph, presenting repair algorithms and system optimizations to enhance data cleaning quality and efficiency.	General	None
(Xie & Cheng 2020)	Model	The article proposes a new data cleaning algorithm designed to handle large, imbalanced datasets, thus basing its methodology on algorithmic approaches.	General	None
(Ding et al. 2018)	Framework	The article proposes a framework called Improve3C, consisting of four steps to detect and enhance data quality, along with algorithmic solutions for repairing incomplete and inconsistent data, and a metric for currency difference and consistency to address inconsistent attributes.	General	None
(Koehler et al. 2021)	Method	The article's methodology is based on a scalable approach to automate data preparation, considering data context, and includes defining a data schema, data profiling, data matching, value format transformation, data repair, and mapping generation and selection.	General	None
(Fan et al. 2013)	Method	The article's methodology is based on heuristic algorithms that integrate data consistency and currency in a single process to resolve conflict resolution issues, evaluating the accuracy and efficiency of their method using real and synthetic data, without employing bibliometrics, frameworks, benchmarks, or specific tools.	General	None
(G. Chen et al. 2017)	Framework	The article's methodology is a framework that sequentially maps datasets using filtering techniques to accelerate searches in generic metric spaces, implemented with MapReduce for large-scale data processing.	General	None
(Maccio et al. 2014)	Method	The article's methodology involves queue modeling to analyze the behavior of distributed data cleaning systems under various configurations and parameters, applying queue	General	None

Author (Article)	Approach	Technique	Industry	Database
		theory principles to assess the trade-off between system performance and data quality.		
(Hossen et al. 2018)	Method	The article's methodology proposes a new data cleaning method for big data analytics that involves feature selection based on Random Forest and the application of two classification algorithms (Random Forest and linear SVM) to train and develop an intelligent model.	General	None
(Huang et al. 2018)	Framework	The article's methodology presents a framework called PACAS, which implements a pricing scheme for sensitive data and proposes new extensions to define generalized data repairs that obfuscate sensitive information.	General	None
(B. Li et al. 2021)	Method	The article's methodology is based on k-means clustering algorithms, a BERT model for transforming text data into vectors, and a parallel implementation scheme to enhance the efficiency of the data cleaning process.	General	Competitions
(Drakopoulos & Megalooikonomou 2016)	Method	The article's methodology is based on data regularization techniques implemented through the conjugate gradient method using a finite difference matrix, discussing the connection between finite differences and the discrete Laplacian operator, and evaluating the proposed regularization techniques using heart rate time series data from the MIT-BIH dataset.	Biomedical	Cardiology
(Roy et al. 2022)	Method	The article presents a technique for removing impulse noise in images from low-cost sensors using a hybrid detection and fuzzy filter algorithm, and discusses the efficiency of this technique compared to other impulse noise removal methods in the literature.	Multimedia	M-IOT
(Wu et al. 2018)	Method	The article does not specifically mention a methodology; it discusses data mining and the algorithms used, but does not provide a specific methodology for the article in question.	General	None
(Lyu 2021)	Method	The article employs data cleaning technology to enhance dataset quality, thereby improving the performance of pedestrian detection models in subway stations.	Transportation	None
(Van Keulen et al. 2018)	Method	The article utilizes an iterative approach to incorporate user evidence into probabilistically integrated data, presenting a technique for remapping random variables in a probabilistic database and extending query languages to specify evidence as hard and soft rules, along with methods for updating the database with this evidence.	General	None
(Adolfo et al. 2021)	Method	The article presents an iterative method for integrating user evidence into probabilistic data, including techniques for remapping random variables and extending query languages to handle hard and soft rules.	General	Physiological

Author (Article)	Approach	Technique	Industry	Database
(Rollo et al. 2022)	Framework	The methodology combines various data cleaning techniques for anomaly detection and classification in large traffic sensor datasets, including flow-speed correlation filtering, replacing filtered observations with nearby reliable averages, applying the FFIDCAD model for anomaly classification and using the ARIMA model for anomaly detection.	Traffic	Traffic
(A. Zhang et al. 2016)	Method	The methodology employs a statistical approach for data cleaning in sequences, presenting exact algorithms and heuristic methods to optimize the cleaning process.	General	GPS
(H. Yang et al. 2019)	Method	The methodology involves creating a data cleaning system based on a big data platform, allowing for customizable rules and algorithms.	General	None
(Shimizu et al. 2019)	Method	The methodology proposes a data cleaning architecture based on visualizations to enable data managers to efficiently navigate and correct data portions, without relying on bibliometrics, algorithms, frameworks, benchmarks, or specific tools.	General	Scientific
(O'Shea et al. 2020)	Model	he methodology utilizes Model-Driven Development (MDD), a software development approach focused on creating high-level models to specify and automate the software development process.	General	None
(Ilyas & Chu 2015)	Method	The methodology does not specifically involve bibliometrics, algorithms, frameworks, benchmarks, or tools, but discusses various data cleaning techniques, including rule-based methods, clustering methods, and machine learning methods.	General	None
(Cao et al. 2017)	Model	The methodology employs a video data cleaning algorithm based on the Bradley-Terry model, detailing its implementation and specifics.	General	None
(Ganibardi & Ali 2018)	Method	The methodology is based on experiments and analyses of web log data, processed with various data cleaning algorithms and evaluated using a confusion matrix, implemented in R within the Apache Spark API.	General	Web Records
(Li et al. 2019)	Framework	The methodology proposes a new framework based on attribute correlation under blocking for unsupervised data cleaning.	General	None

The technical approaches utilized in the articles were categorized by similarity, with decreasing frequency as follows: (i) Method: 36 – 52%; (ii) Framework: 24 – 35%; and (iii) Benchmark: 9 – 13%.

Regarding the specific techniques cited in the articles, it was initially anticipated that traditional data cleaning methods would be the most prevalent. However, the review revealed a diverse array of techniques, including approaches based on machine learning, data mining, natural language processing, and other advanced methodologies.

In terms of the industries referenced, approximately 74% of the articles do not specify any particular sector. The remaining articles highlight health, manufacturing, and construction to a lesser extent.

Databases were utilized or specified in about half of the articles, with a notable emphasis on health data.

This review has some limitations that should be acknowledged. First, the exclusion of non-English articles may have resulted in the omission of relevant studies conducted in other languages. Second, reliance on specific databases (Scopus, Web of Science, CAPES) may have introduced selection bias, potentially overlooking significant works indexed in other databases. Additionally, the exclusion of articles published in 2023 due to incomplete data may have omitted recent developments in the field. Future research should consider these limitations and strive to include a broader range of sources to mitigate these biases.

To further illustrate the applicability of data cleaning techniques across different industrial sectors, we provide specific examples:

- **Healthcare**: Techniques such as real-time machine learning (Elouataoui et al., 2022) and deep data cleaning (Chiang et al., 2021) have been employed to enhance diagnostic validity and mortality assessment in healthcare settings.
- **Finance**: The Reynolds Decomposition approach (Mok et al., 2017) has been utilized to clean security price data, thereby improving the accuracy of financial models.
- **Manufacturing**: Methods such as the 5S data management methodology (F. Li et al., 2021) have been implemented to enhance data quality in manufacturing processes, leading to more efficient production workflows.
- **Transportation**: Data cleaning aimed at improving data quality (Lyu, 2021) has been applied in transportation systems to enhance pedestrian detection algorithms, contributing to safer and more efficient transit systems.

4. CONCLUSIONS

The systematic mapping review (SMR) on data cleaning in Big Data has revealed the growing significance of this research field, highlighting the approaches and techniques employed to manage large volumes of data. The methodology adopted in this study, based on the Parsifal online platform and the specific PRISMA-P protocol, facilitated a comprehensive and independent approach, following a clear and explicit methodology to describe the procedures utilized.

The review demonstrated the use of a wide variety of advanced techniques grounded in artificial intelligence, including machine learning, data mining, and natural language processing, among others. This diversity of techniques reflects the complexity and heterogeneity of data in Big Data environments, as well as the necessity for innovative approaches to address these challenges.

The predominant approach adopted in the articles was the use of methods, as opposed to frameworks and benchmarks. There was a lack of emphasis on the types of industries and data sources in the context of data cleaning.

It is evident that data cleaning in Big Data is a multifaceted process that necessitates a well-planned technical approach, effective specific techniques, and an understanding of the unique characteristics of the databases utilized. Effective data cleaning ensures that data is reliable and useful, providing a solid foundation for accurate analyses and informed decision-making.

This study is expected to serve as a starting point for researchers interested in the topic, offering perspectives and guiding future investigations in the field of data cleaning in Big Data. Future research should focus on developing new data cleaning techniques capable of handling real-time data processing and integrating these techniques into the broader data management and analysis pipeline to enhance the reliability and usability of Big Data.

REFERENCES

Abu Ahmad, H. and Wang, H., 2018. An effective weighted rule-based method for entity resolution. *Distributed and Parallel Databases*, Vol. 36, No. 3, pp. 593-612.

Adolfo, C. M. S., Chizari, H., Win, T. Y. and Al-Majeed, S., 2021. Sample Reduction for Physiological Data Analysis Using Principal Component Analysis in Artificial Neural Network. *Applied Sciences*, Vol. 11, No. 17, 8240.

Bramantoro, A., 2018. Data Cleaning Service for Data Warehouse: An Experimental Comparative Study on Local Data. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, Vol. 16, No. 2, 834-842.

Cao, B., Chen, L. and Lie, H., 2017. Vast Amounts of Video Data Clean Algorithm Base on Bradley-Terry Model. *2016 7th International Conference on Education, Management, Computer and Medicine (EMCM 2016)*, Shenyang, China.

Chen, H., Chiang, R. and Storey, V., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, Vol. 36, No. 4, 1165-1188.

Chen, G., Yang, K., Chen, L., Gao, Y., Zheng, B. and Chen, C., 2017. Metric Similarity Joins Using MapReduce. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 3, pp. 656-669.

Chen, Z., Wang, Y., Narasayya, V. and Chaudhuri, S., 2019. Customizable and scalable fuzzy join for big data. *Proceedings of the VLDB Endowment*, Vol. 12, No. 12, pp. 2106-2117.

Chiang, H.-Y., Liang, L.-Y., Lin, C.-C., Chen, Y.-J., Wu, M.-Y., Chen, S.-H., Wu, P.-H., Kuo, C.-C. and Chi, C.-Y., 2021. Electronic Medical Record-Based Deep Data Cleaning and Phenotyping Improve the Diagnostic Validity and Mortality Assessment of Infective Endocarditis: Medical Big Data Initiative of CMUH. *BioMedicine*, Vol. 11, No. 3, pp. 59-67.

Chu, X., Ilyas, I. F. and Papotti, P., 2013. Holistic data cleaning: Putting violations into context. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 458-469.

De, S., Hu, Y., Chen, Y. and Kambhampati, S., 2014. BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata. *Proceedings of the 2014 IEEE International Conference on Big Data (Big Data)*, pp. 15-24.

Ding, X., Wang, H., Su, J., Li, J. and Gao, H., 2018. *Improve3C: Data Cleaning on Consistency and Completeness with Currency*. https://doi.org/10.48550/arXiv.1808.00024

Dong, S., Feric, Z., Yu, L., Kaeli, D., Meeker, J., Padilla, I. Y., Cordero, J., Vega, C. V., Rosario, Z. and Alshawabkeh, A., 2018. An Efficient Data Management Framework for Puerto Rico Testsite for Exploring Contamination Threats (PROTECT). *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*, pp. 5316-5318.

Drakopoulos, G. and Megalooikonomou, V. (2016). Regularizing large biosignals with finite differences. *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA).*

Du, Y.-F., Shen, D.-R., Nie, T.-Z., Kou, Y. and Yu, G., 2017. A cleaning method for consistency and currency in related data. *Chinese Journal of Computers*, Vol. 40, No. 1, pp. 92-106.

Elouataoui, W., Alaoui, I. E., Mendili, S. E. and Gahi, Y., 2022. An End-to-End Big Data Deduplication Framework based on Online Continuous Learning. *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 9.

Fan, W., Geerts, F., Tang, N. and Yu, W., 2013. Inferring data currency and consistency for conflict resolution. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp. 470-481.

Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F. and Chu, X., 2016. CLAMS: Bringing Quality to Data Lakes. *Proceedings of the 2016 International Conference on Management of Data*, pp. 2089-2092.

Fayyad, U. M., Candel, A., Ariño De La Rubia, E., Pafka, S., Chong, A. and Lee, J.-Y., 2017. Benchmarks and Process Management in Data Science: Will We Ever Get Over the Mess? *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 31-32.

Feric, Z., Agostini, N. B., Beene, D., Signes-Pastor, A. J., Halchenko, Y., Watkins, D., MacKenzie, D., Karagas, M. R., Manjourides, J., Alshawabkeh, A. N. and Kaeli, D. R., 2021. A Secure and Reusable Software Architecture for Supporting Online Data Harmonization. *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2801-2812.

Galegale, N. V., Fontes, E. L. G. and Galegale, B. P., 2017. Uma contribuição para a segurança da informação: um estudo de casos múltiplos com organizações brasileiras. *Perspectivas em Ciência da Informação*, Vol. 22, No. 3, pp. 75-97.

Gallardo-Gallardo, E., 2016. *Systematic Literature Reviews*. [Unpublished.]

Ganibardi, A. and Ali, C. A., 2018. Web Usage Data Cleaning: A Rule-Based Approach for Weblog Data Cleaning. In C. Ordonez and L. Bellatreche (orgs.), *Big Data Analytics and Knowledge Discovery* (Vol. 11031). Springer International Publishing, pp. 193-203.

Hellerstein, J. M., 2008. *Quantitative Data Cleaning for Large Databases*. Available at: https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf

Hossen, J., Jesmeen H, M. Z. and Sayeed, S., 2018. Modifying Cleaning Method in Big Data Analytics Process using Random Forest Classifier. *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, pp. 208-213.

Huang, Y., Milani, M. and Chiang, F., 2018. PACAS: Privacy-Aware, Data Cleaning-as-a-Service. *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1023-1030.

Huang, Y., Milani, M. and Chiang, F., 2020. Privacy-Aware Data Cleaning-as-a-Service. *Information Systems*, Vol. 94, 101608.

Ilyas, I. F. and Chu, X., 2015. Trends in Cleaning Relational Data: Consistency and Deduplication. *Foundations and Trends® in Databases*, Vol. 5, No. 4, pp. 281-393.

Ito, C., Marinho Filho, A., Ito, M, Azevedo, M. M. and Almeida, M. A., 2013. Preliminary evaluation of a serious game for the dissemination and public awareness on preschool children's oral health. In C. U. Lehmann et al. (eds.) *MEDINFO 2013*. IOS Press, pp. 1034-1034.

Ito M., Appel A. P., de Santana V. F. and Moyano L. G., 2017. Analysis of the Existence of Patient Care Team Using Social Network Methods in Physician Communities from Healthcare Insurance Companies. *Studies in Health Technology and Informatics*, Vol. 245, pp. 412-416.

Iwata, C. and Ito, M., 2023. Limpeza de Dados em Big data: Uma Revisão Bibliométrica. *XVIII Simpósio dos Programas de Mestrado Profissional Unidade de Pós-Graduação, Extensão e Pesquisa*, pp. 1259-1275. Available at: http://www.pos.cps.sp.gov.br/files/artigo/file/1335/86b356802ddf22c51f191115ccff47ba.pdf

Jäger, S., Allhorn, A. and Bießmann, F., 2021. A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, Vol. 4, 693674.

Jin, C., Jin, R., Chen, K. and Dou, Y., 2018. A Community Detection Approach to Cleaning Extremely Large Face Database. *Computational Intelligence and Neuroscience,* Vol. 2018. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1155/2018/4512473

Kenda, K. and Mladenić, D., 2018. Autonomous Sensor Data Cleaning in Stream Mining Setting. *Business Systems Research Journal*, Vol. 9, No. 2, pp. 69-79.

Kim, M., Shin, S.-Y., Kang, M., Yi, B.-K. and Chang, D. K., 2019. Developing a Standardization Algorithm for Categorical Laboratory Tests for Clinical Big Data Research: Retrospective Study. *JMIR Medical Informatics*, Vol. 7, No. 3, e14083.

Koehler, M., Abel, E., Bogatu, A., Civili, C., Mazilu, L., Konstantinou, N., Fernandes, A. A. A., Keane, J., Libkin, L. and Paton, N. W., 2021. Incorporating Data Context to Cost-Effectively Automate End-to-End Data Wrangling. *IEEE Transactions on Big Data*, Vol. 7, No. 1, pp. 169-186.

Li, B., Wang, J. and Liu, X., 2021. Parallel Cleaning Algorithm for Similar Duplicate Chinese Data Based on BERT. *Scientific Programming*, Vol. 2021. https://doi.org/10.1155/2021/5916748

Li, F., Li, X., Yang, Y., Xu, Y. and Zhang, Y., 2021. Collaborative Production Task Decomposition and Allocation among Multiple Manufacturing Enterprises in a Big Data Environment. *Symmetry*, Vol. 13, No. 12, 2268.

Li, P., Dai, C. and Wang, W., 2019. When Considering More Elements: Attribute Correlation in Unsupervised Data Cleaning under Blocking. *Symmetry*, Vol. 11, No. 4, 575.

Lin, P.-I. D., Rifas-Shiman, S. L., Aris, I. M., Daley, M. F., Janicke, D. M., Heerman, W. J., Chudnov, D. L., Freedman, D. S. and Block, J. P., 2022. Cleaning of anthropometric data from PCORnet electronic health records using automated algorithms. *JAMIA Open*, Vol. 5, No. 4, ooac089.

Lin, Y., Wang, H., Li, J. and Gao, H., 2019. Data source selection for information integration in big data era. *Information Sciences*, Vol. 479, pp. 197-213.

Liu, J., Cao, Y., Li, Y., Guo, Y. and Deng, W., 2020. A big data cleaning method based on improved CLOF and Random Forest for distribution network. *CSEE Journal of Power and Energy Systems*. doi: 10.17775/CSEEJPES.2020.04080

Lyu, Z., 2021. Research on Subway Pedestrian Detection Algorithm Based on Big Data Cleaning Technology. *Wireless Communications and Mobile Computing*, Vol. 2021, pp. 1-10.

Maccio, V. J., Chiang, F. and Down, D. G., 2014. Models for Distributed, Large Scale Data Cleaning. In W.-C. Peng, H. Wang, J. Bailey, V. S. Tseng, T. B. Ho, Z.-H. Zhou and A. L. P. Chen (orgs.) *Trends and Applications in Knowledge Discovery and Data Mining* (Vol. 8643). Springer International Publishing, pp. 369-380.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H., 2011. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Available at:https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_exec_summary.pdf

Martinez-Mosquera, D., Luján-Mora, S., López, G. and Santos, L., 2017. Data Cleaning Technique for Security Logs Based on Fellegi-Sunter Theory. In S. Wrycza and J. Maślankowski (orgs.) *Information Systems: Research, Development, Applications, Education* (Vol. 300). Springer International Publishing, pp. 3-12.

Martins, E. and Galegale, N., 2022. Retail sales forecasting information systems: comparison between traditional methods and machine learning algorithms. *Proceedings of the 2022 International Conference Information Systems (IADIS)*, pp. 30-38.

Martins, E. and Galegale, N. V., 2023. Machine learning: A bibliometric analysis. *International Journal of Innovation*, Vol. 11, No. 3, pp. 1-37.

Mavrogiorgos, K., Mavrogiorgou, A., Kiourtis, A., Zafeiropoulos, N., Kleftakis, S. and Kyriazis, D., 2022. Automated Rule-Based Data Cleaning Using NLP. *2022 32nd Conference of Open Innovations Association (FRUCT),* pp. 162-168.

Moher, D. et al., 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, Vol. 4, No. 1.

Mok, R. V., Mok, W. Y. and Cheung, K. Y., 2017. A Security Price Data Cleaning Technique: Reynold's Decomposition Approach. In J. A. Lossio-Ventura and H. Alatrista-Salas (orgs.) *Information Management and Big Data* (Vol. 656). Springer International Publishing, pp. 108-119.

Nugroho, H., Utama, N. P. and Surendro, K., 2021. Class center-based firefly algorithm for handling missing data. *Journal of Big Data*, Vol. 8, No. 1, 37.

Okoli, C., 2019. Guia Para Realizar uma Revisão Sistemática de Literatura*. EaD em Foco*, Vol. 9, No. 1.

Oni, S., Chen, Z., Hoban, S. and Jademi, O., 2019. A Comparative Study of Data Cleaning Tools. *International Journal of Data Warehousing and Mining*, Vol. 15, No. 4, pp. 48-65.

O'Shea, E., Khan, R., Breathnach, C. and Margaria, T., 2020. Towards Automatic Data Cleansing and Classification of Valid Historical Data an Incremental Approach Based on MDD. *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1914-1923.

Prakash, A., Navya, N. and Natarajan, J., 2019. Big Data Preprocessing for Modern World: Opportunities and Challenges. In J. Hemanth, X. Fernando, P. Lafata and Z. Baig (orgs.) *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018* (Vol. 26). Springer International Publishing, pp. 335-343.

Rahul, K., Banyal, R. K. and Goswami, P., 2020. Analysis and processing aspects of data in big data applications. *Journal of Discrete Mathematical Sciences and Cryptography*, Vol. 23, No. 2, pp. 385-393.

Ribeiro, B. R., Velazco, S. J. E., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S. P. and Loyola, R., 2022. bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, Vol. 13, No. 7, pp. 1421-1428.

Rollo, F., Bachechi, C. and Po, L., 2022. Semi Real-time Data Cleaning of Spatially Correlated Data in Traffic Sensor Networks. *Proceedings of the 18th International Conference on Web Information Systems and Technologies*, pp. 83-94.

Roy, A., Bandopadhaya, S., Chandra, S. and Suhag, A., 2022. Removal of impulse noise for multimedia-IoT applications at gateway level. *Multimedia Tools and Applications*, Vol. 81, No. 24, pp. 34463-34480.

Sainju, A. M., He, W., Jiang, Z., Yan, D. and Chen, H., 2021. Flood Inundation Mapping with Limited Observations Based on Physics-Aware Topography Constraint. *Frontiers in Big Data*, Vol. 4, 707951.

Shen, L., He, X., Liu, M., Qin, R., Guo, C., Meng, X. and Duan, R., 2021. A Flexible Ensemble Algorithm for Big Data Cleaning of PMUs. *Frontiers in Energy Research*, Vol. 9, 695057.

Shende, M. K., Feijóo-Lorenzo, A. E. and Bokde, N. D., 2022. cleanTS: Automated (AutoML) tool to clean univariate time series at microscales. *Neurocomputing*, Vol. 500, pp. 155-176.

Sheoran, S. K. and Parmar, V., 2022. GeoWebCln: An Intensive Cleaning Architecture for Geospatial Metadata. *Quaestiones Geographicae*, Vol. 41, No. 1, pp. 51-62.

Shimizu, T., Omori, H. and Yoshikawa, M., 2019. Toward a view-based data cleaning architecture. *ArXiv, abs/1910.11040*. Available at: https://api.semanticscholar.org/CorpusID:204852143

Staegemann, D., Volk, M., Saxena, A., Pohl, M., Nahhas, A., Häusler, R., Abdallah, M., Bosse, S., Jamous, N. and Turowski, K., 2021. Challenges in Data Acquisition and Management in Big Data Environments *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security*, pp. 193-204.

Tang, N., 2015. Big RDF data cleaning. *2015 31st IEEE International Conference on Data Engineering Workshops*, pp. 77-79.

Tian, Y., Michiardi, P. and Vukolic, M., 2017. Bleach: A Distributed Stream Data Cleaning System. *2017 IEEE International Congress on Big Data (BigData Congress)*, pp. 113-120.

Van Keulen, M., Kaminski, B. L., Matheja, C. and Katoen, J.-P., 2018. Rule-Based Conditioning of Probabilistic Data. In D. Ciucci, G. Pasi and B. Vantaggi (orgs.) *Scalable Uncertainty Management* (Vol. 11142). Springer International Publishing, pp. 290-305).

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R. and Childe, S. J., 2017. Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, Vol. 70, pp. 356-365.

Wang, J., Krishnan, S., Franklin, M. J., Goldberg, K., Kraska, T. and Milo, T., 2014. A sample-and-clean framework for fast and accurate query processing on dirty data. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 469-480.

Wang, T., Ke, H., Zheng, X., Wang, K., Sangaiah, A. K. and Liu, A., 2020. Big Data Cleaning Based on Mobile Edge Computing in Industrial Sensor-Cloud. *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 2, pp. 1321-1329.

Wu, L., Li, W., Wang, C., Yang, Y., An, Z. and Zhang, N., 2018. Research on Acquisition of Clean Governance Evaluation Techniques for Big Data. *Proceedings of the 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017)*, Xi'an, China.

Xie, W. and Cheng, X., 2020. Imbalanced big data classification based on virtual reality in cloud computing. *Multimedia Tools and Applications*, Vol. 79, No. 23-24, pp. 16403-16420.

Xu, S., Lu, B., Baldea, M., Edgar, T. F., Wojsznis, W., Blevins, T. and Nixon, M., 2015. Data cleaning in the process industries. *Reviews in Chemical Engineering*, Vol. 31, No. 5.

Yang, H., Liu, W., Wang, X., Liu, H., Yu, B. and Zhou, H., 2019. The Design and Implementation of a Cleaning System Prototype. *IOP Conference Series: Earth and Environmental Science*, Vol. 252, 032218.

Yang, J., Tan, K. K., Santamouris, M. and Lee, S. E., 2019. Building Energy Consumption Raw Data Forecasting Using Data Cleaning and Deep Recurrent Neural Networks. *Buildings*, Vol. 9, No, 9, 204.

Yang, X., Tang, L., Zhang, X. and Li, Q., 2018. A Data Cleaning Method for Big Trace Data Using Movement Consistency. *Sensors* (Basel, Switzerland), Vol. 18. Available at: https://api.semanticscholar.org/CorpusID:4717373

Yousef, A. H., 2015. Cross Language Duplicate Record Detection in Big Data. In A. E. Hassanien, A. T. Azar, V. Snasael, J. Kacprzyk and J. H. Abawajy (orgs.) Big Data in Complex Systems (Vol. 9). Springer International Publishing, pp. 147-171.

Zhang, A., Song, S. and Wang, J., 2016. Sequential Data Cleaning: A Statistical Approach*. Proceedings of the 2016 International Conference on Management of Data*, pp. 909-924.

Zhang, S., Wang, Y. and Lv, Q., 2022. Exploring Artificial Intelligence Architecture in Data Cleaning Based on Bayesian Networks. *Advances in Multimedia*, Vol. 2022, 1-11.

Zhang, W. and Tan, X., 2019. Combining Outlier Detection and Reconstruction Error Minimization for Label Noise Reduction. *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1-4.

Zhang, Y.-Y., Hu, Z.-Z., Lin, J.-R. and Zhang, J.-P., 2021. Data Cleaning for Prediction and its Evaluation of Building Energy Consumption. *38th International Symposium on Automation and Robotics in Construction*, Dubai, UAE.

Zhou, S., Zhang, R., Chen, D. and Zhu, X., 2021. A novel framework for bringing smart big data to proactive decision making in healthcare. *Health Informatics Journal*, Vol. 27, No. 2, 146045822110246.