Trends and Challenges in Data Cleaning for Large-Scale Systems: A Survey

1 Anil Kumar Jonnalagadda, 2 Srikanth Kamatala, 3 Praveen Kumar Myakala

1 2 3 Independent Researcher

1 2 3 Senior Member, IEEE

[[1]](#footnote-1)

***Abstract*—** **Data cleaning is a critical process to maintain the integrity and usability of large-scale systems that process massive, diverse, and dynamic datasets. As the scale and complexity of data ecosystems grow, traditional cleaning techniques face limitations in addressing challenges such as data heterogeneity, real-time processing demands, and resource constraints. This paper presents a comprehensive survey of the latest trends and persistent challenges in data cleaning for large-scale systems. It examines advancements in automated and AI-driven methods, distributed and cloud-based cleaning frameworks, and real-time error detection techniques for streaming data. Additionally, the survey highlights domain-specific cleaning approaches in sectors like healthcare and finance, where data quality significantly impacts decision-making and operational efficiency. Key challenges, including scalability bottlenecks, the lack of standardized benchmarks, and ethical considerations, are discussed in detail. Finally, the paper identifies open research directions, such as explainable AI in data cleaning, universal metrics development, and sustainable algorithms for resource-efficient processing. By synthesizing recent developments and emphasizing their role in improving decision-making, system performance, and user experience, this survey aims to guide researchers and practitioners toward innovative solutions for enhancing data quality in large-scale systems.**

***Index Terms*—** **Automation in Data Cleaning, Cloud-Based Data Cleaning, Data Heterogeneity, Data Provenance, Explainable AI, Federated Data Cleaning, IoT Data Cleaning, Machine Learning in Data Cleaning, Metrics for Data Quality, Privacy-Preserving Data Cleaning, Real-Time Data Cleaning, Resource-Efficient Algorithms, Scalability, Standardized Benchmarks, Trends in Data Cleaning.**

# 1. Introduction

The rapid growth of large-scale data systems in domains such as healthcare, finance, e-commerce, smart cities, and scientific research has placed an increasing demand on data quality [1][2]. Ensuring data accuracy, completeness, and consistency is critical for effective decision-making and maintaining the performance of systems that rely on vast and dynamic datasets [3][4]. For example, inaccurate medical data can lead to incorrect diagnoses or delayed treatments, jeopardizing patient care [3]. Similarly, poor-quality financial data can result in fraudulent transactions going undetected, causing significant monetary losses [4]. Data cleaning, the process of identifying and correcting errors or inconsistencies, is essential to mitigate these risks [76].

Furthermore, the increasing reliance on AI and machine learning applications underscores the critical role of data quality. The performance of these models heavily depends on the accuracy and completeness of the training data [5]. Poor-quality data can propagate through machine learning pipelines, resulting in biased predictions, unreliable outcomes, and reduced trust in AI-driven systems. For instance, biased facial recognition algorithms have been shown to exhibit higher error rates for certain demographic groups, and discriminatory loan approval models can unfairly deny credit to underrepresented populations [6][7]. This makes robust data cleaning a fundamental prerequisite for developing trustworthy and high-performing AI models, further emphasizing the need for innovative and effective cleaning solutions.

High-quality data enables improved decision-making, enhances system performance, fosters user trust, and reduces operational costs. However, in large-scale systems, which often involve heterogeneous data sources, high-velocity streams, and distributed architectures, maintaining data quality is particularly challenging. Traditional data cleaning approaches, such as rule-based and manual methods, often fail to address the unique demands of such environments, including scalability, real-time processing, and privacy constraints [8][9].

Recent advancements, such as scalable frameworks, AI-driven solutions, and visual data wrangling tools, are transforming data cleaning processes for large-scale systems [9][10]. These approaches aim to address the limitations of traditional methods while meeting the evolving demands of big data applications. However, significant challenges remain, including the lack of standardized benchmarks for evaluating cleaning techniques and the need for explainable and sustainable algorithms.

This paper presents a comprehensive survey of recent developments in data cleaning techniques for large-scale systems, identifying key trends, ongoing challenges, and future research directions. The main contributions of this work include:

* A review of recent trends in data cleaning, focusing on automated solutions, real-time frameworks, and distributed systems.
* A discussion of major challenges, including scalability issues, data heterogeneity, and the lack of standardized benchmarks.
* Identification of open research opportunities, such as explainable AI, sustainable cleaning algorithms, and cross-domain cleaning methods.

This survey is targeted at researchers, data scientists, and data engineers seeking a comprehensive understanding of the current state of the art and future directions in data cleaning for large-scale systems. By synthesizing the field’s advancements and pinpointing research gaps, this paper aims to guide practitioners toward innovative solutions for improving data quality in complex environments.

# 2. Background and Related Work

## 2.1 Data Cleaning and Data Quality Dimensions

Data cleaning is the process of detecting and correcting errors, inconsistencies, and inaccuracies in data to improve its quality and usability [1]. The importance of data cleaning lies in its ability to address issues such as missing data, duplicate entries, inconsistent formats, and invalid values, which can significantly hinder the performance of data-driven systems. Several key dimensions define data quality:

* **Accuracy:** The degree to which the data correctly represents the real-world entities it models [2].
* **Completeness:** The extent to which all required data is present and available [Shahri].
* **Consistency:** The absence of contradictions within the dataset (e.g., consistent date formats or currency values) [14].
* **Timeliness:** The extent to which the data is up-to-date and relevant for its intended use [15].
* **Validity:** The degree to which the data conforms to defined formats or rules (e.g., valid email addresses or postal codes) [16].

**Table 1**

Key Data Quality Dimensions

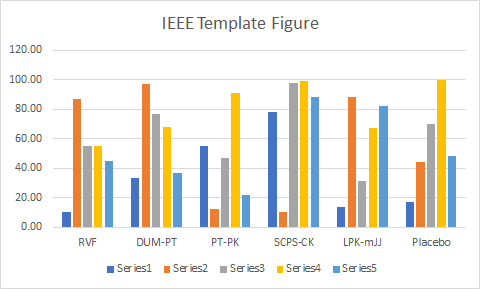
|  |  |  |
| --- | --- | --- |
| **Dimension** | **Definition** | **Example** |
| Accuracy | The degree to which data correctly represents real-world values. | A customer's address must match official postal records. |
| Completeness | The extent to which all required data is available and recorded. | Missing fields in patient records may lead to incorrect treatments. |
| Consistency | The uniformity of data across multiple sources or datasets. | A customer's email should be the same across all company databases. |
| Timeliness | The extent to which data is up to date and available when needed. | Stock market data should be updated in real-time to ensure accurate trades. |
| Validity | The degree to which data adheres to predefined formats or constraints. | A date field should follow the "YYYY-MM-DD" format. |
| Integrity | Ensuring data relationships and dependencies are correctly maintained. | A foreign key in a database should correctly reference an existing record. |
| Provenance | The traceability and source verification of data. | AI models should document the origin of their training datasets. |

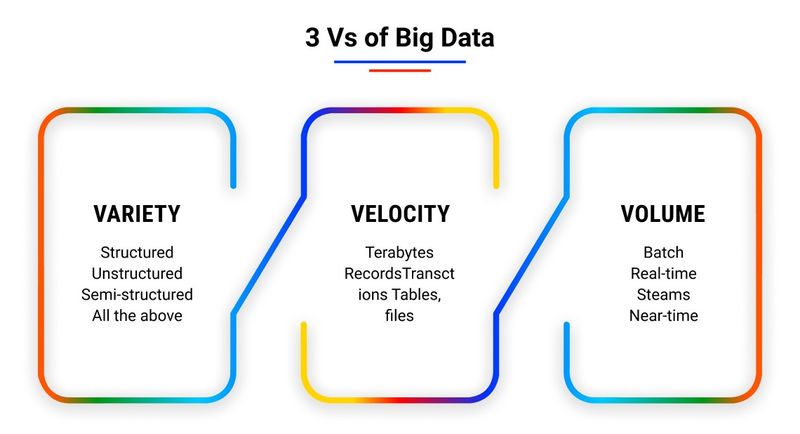
These dimensions are interrelated. For instance, incomplete data can lead to inaccuracies in analysis, while inconsistent data can hinder integration across systems. Addressing one dimension often requires improvements in others. For example, ensuring completeness by imputing missing values can also enhance accuracy and consistency [Shahri]. In domains like healthcare, incomplete or inaccurate electronic health records (EHRs) [75] can result in suboptimal treatments, jeopardizing patient safety [17]. Similarly, in financial applications, inconsistent data across distributed systems can obstruct fraud detection mechanisms [18].

## 2.2 Characteristics of Large-Scale Systems

Large-scale systems are characterized by their ability to process vast amounts of data, often in distributed and heterogeneous environments. These systems are commonly associated with big data technologies, which exhibit the "3Vs":

* **Volume:** The massive size of datasets, often measured in terabytes or petabytes [19].
* **Velocity:** The speed at which data is generated, captured, and processed, particularly in real-time applications like IoT or social media streams [20].
* **Variety:** The diversity of data formats, ranging from structured tables to unstructured text, images, and video [19].

 **Fig. 1.** This is a sample of a figure caption.



Cleaning data in such systems introduces unique challenges, including scalability, heterogeneity, and latency constraints:

* **Scalability:** Ensuring that cleaning operations can handle increasing data volumes and processing demands across distributed systems.
* **Heterogeneity:** Addressing inconsistencies across diverse data formats, schemas, and sources.
* **Real-Time Requirements:** Real-time data cleaning involves processing and cleaning data with minimal latency, often within milliseconds or seconds, to ensure timely decision-making [21].

For example, IoT devices in smart cities generate high-velocity data streams requiring immediate validation and cleaning to enable real-time traffic and environmental monitoring [20].

## 2.3 AI/ML in Data Cleaning

Machine learning and AI have revolutionized data cleaning by enabling automated and intelligent solutions. These techniques address several key aspects of the cleaning process:

**Anomaly Detection:** Unsupervised learning techniques, such as clustering and one-class classification, effectively identify outliers and anomalies that deviate significantly from expected patterns [22]. For instance, AI models can detect anomalous readings in IoT sensor data to identify faulty devices [23].

**Missing Value Imputation:** Techniques like K-Nearest Neighbour’s (KNN), regression models, and deep learning-based approaches accurately impute missing values by analysing relationships with other data points [24]. For example, missing entries in customer profiles can be filled using collaborative filtering or neural networks.

**Data Profiling:** AI/ML algorithms can automatically analyse datasets to identify data types, detect distributions, and discover relationships between variables. This helps in identifying data quality issues such as inconsistencies or missing relationships, enabling more effective cleaning workflows [25].

These AI-driven approaches significantly enhance the scalability and efficiency of data cleaning, particularly for large-scale and complex datasets.

## 2.4 Ethical Implications of Data Cleaning

Data cleaning, particularly when using AI/ML methods, raises several ethical concerns:

* **Privacy:** Data cleaning often requires access to sensitive data, such as personal or medical records. Improper handling of this data can violate privacy laws like GDPR and HIPAA [26].
* **Bias:** AI/ML-based cleaning algorithms can unintentionally introduce or amplify bias in the dataset, resulting in unfair outcomes. For example, biased algorithms in financial systems could disproportionately reject loan applications for certain demographics [27].
* **Accountability:** Decisions made during data cleaning (e.g., inputting missing values or removing outliers) can significantly impact downstream analytics and decision-making. Ensuring transparency in these decisions is crucial for accountability [28].

Responsible use of AI/ML in data cleaning involves adhering to ethical guidelines and implementing safeguards to mitigate these risks, such as explainable AI methods and privacy-preserving techniques [26].

## 2.5 Data Governance and Quality Policies

Data governance and quality policies play a crucial role in guiding data cleaning efforts, particularly in large-scale systems. Governance frameworks establish standards for data quality, security, and compliance with regulations [29]. For example:

* **Compliance:** Regulations like GDPR and HIPAA mandate strict controls on how data is cleaned and stored.
* **Standardization:** Policies ensure consistent cleaning practices across distributed systems, reducing inconsistencies and errors.
* **Auditability:** Governance frameworks enable traceability of cleaning decisions, ensuring transparency and accountability [30].

By embedding data cleaning within broader governance frameworks, organizations can ensure more effective and compliant cleaning processes.

## 2.6 Related Work

Several surveys and research efforts have addressed data cleaning challenges and techniques. For example:

While these studies provide valuable insights, they often focus on specific domains or methodologies. This paper expands on these efforts by comprehensively surveying trends, challenges, and open research directions in data cleaning for large-scale systems, particularly in the context of automation, real-time frameworks, and AI/ML-driven solutions.

# 3. Trends in Data Cleaning

## 3.1 Automation and AI-Driven Solutions

The integration of Machine Learning (ML) and Artificial Intelligence (AI) into data cleaning processes has revolutionized the field by automating error detection and correction tasks. These methods significantly reduce manual effort and improve efficiency [29]. For example:

* **Anomaly Detection:** AI models such as clustering algorithms and one-class classification are widely used for identifying outliers and anomalies in large datasets [30]. A popular model for anomaly detection is the One-Class Support Vector Machine (SVM), which identifies outliers in datasets by mapping data into a high-dimensional space. The decision function is defined as:

f(x) = \mathbf{w} \cdot \phi(x) - \rho

where \phi(x) represents the kernel function, \mathbf{w} is the weight vector learned during training, and \rho is the threshold for classification. Data points with f(x) < 0 are classified as anomalies [71]. This method is widely used in large-scale systems to identify invalid records, such as sensor errors in IoT [77] or duplicate entries in financial datasets.

* **Missing Value Imputation:** Deep learning models, regression techniques, and methods like K-Nearest Neighbors (KNN) have been applied to accurately impute missing values [31].
* **Unstructured Data Cleaning:** Advanced techniques in natural language processing (NLP) and computer vision are used to clean text and image data, respectively [32].

Several tools and frameworks have emerged to facilitate AI-driven cleaning workflows. For instance, TensorFlow Data Validation (TFDV) helps identify anomalies in training data for machine learning models, while frameworks like HoloClean use probabilistic models for cleaning relational data [33]. These advancements enable scalable and intelligent data cleaning for large-scale systems.

## 3.2 Unstructured Data Cleaning

Unstructured data, such as text and images, presents unique challenges for cleaning due to its lack of predefined schemas and formats. Specific techniques for cleaning unstructured data include:

**Text Cleaning:** NLP techniques address issues like typos, punctuation errors, and inconsistencies in formatting. Common methods include stemming, lemmatization, and stop word removal to prepare text data for further analysis [34]. Sentiment analysis and topic modeling can also help identify and correct inconsistencies in sentiment or topic assignments in datasets used for customer feedback analysis or social media monitoring [35].

**Image Cleaning:** Computer vision techniques are employed to detect and correct distortions, remove noise, and enhance image quality. For instance, deep learning models can identify and remove unwanted objects, correct color distortions, and improve resolution in medical imaging or satellite data [36].

These techniques enable unstructured data to be effectively integrated into structured analysis workflows, improving downstream AI/ML applications.

## 3.3 Explainable AI in Data Cleaning

Explainable AI (XAI) has gained attention in data cleaning for its ability to provide transparency and interpretability to AI-driven cleaning processes. XAI techniques help users understand the reasoning behind cleaning decisions, such as why certain records are classified as anomalies or how missing values are imputed [37]. This improves trust and accountability in AI-driven systems, particularly in sensitive domains like healthcare and finance. For instance:

* Feature importance techniques can highlight the variables influencing anomaly detection decisions.
* Counterfactual explanations can demonstrate how minor changes in data would affect cleaning outcomes [38].

By incorporating XAI, data cleaning workflows become more transparent, allowing stakeholders to verify and validate cleaning outcomes.

## 3.4 Distributed and Cloud-Based Frameworks

Distributed frameworks and cloud-based platforms are essential for scaling data cleaning workflows to handle the volume and variety of data in large-scale systems. Tools like Apache Spark and Hadoop provide parallel processing capabilities that enable efficient cleaning of petabyte-scale datasets [20]. Cloud-based services, such as AWS Glue and Google Cloud Dataflow, offer several benefits:

* **Scalability:** Cleaning operations can dynamically scale to accommodate growing data volumes.
* **Cost-Effectiveness:** On-demand pricing models reduce infrastructure costs.
* **Ease of Maintenance:** Cloud platforms abstract the complexity of managing hardware and software [40].

These frameworks are particularly valuable in multi-cloud or hybrid cloud environments, where data cleaning must be performed across diverse storage platforms [41].

## 3.5 Real-Time Data Cleaning for Streaming Systems

Real-time data cleaning has become increasingly important in applications that require low-latency processing, such as IoT, fraud detection, and social media analytics [42]. Real-time cleaning involves detecting and correcting errors in data streams within milliseconds or seconds, ensuring timely decision-making.

However, several challenges must be addressed:

* **Latency Constraints:** Cleaning algorithms must operate within strict time limits to prevent delays in decision-making.
* **Resource Limitations:** Streaming systems often have limited computational and memory resources.
* **Efficient Algorithms:** Lightweight and efficient algorithms are required to minimize the computational overhead [43].

Streaming platforms like Apache Flink and Kafka Streams support real-time cleaning workflows by applying windowed operations, outlier detection algorithms, and schema validation in flight [44]. For example, IoT sensor networks in smart cities rely on real-time cleaning to ensure the accuracy of traffic monitoring systems and environmental data analysis [45].

**3.6 Emerging Trends**

Several emerging trends in data cleaning are shaping future research directions:

* **Graph-Based Cleaning:** Cleaning graph-structured data, such as social networks and knowledge graphs, involves techniques like graph traversal, node disambiguation, and edge consistency checks [46].
* **Active Learning:** Combining human expertise with machine learning models to iteratively improve cleaning accuracy. This approach reduces the reliance on labeled data while maintaining high cleaning efficiency [47].

These emerging trends highlight the evolving nature of data cleaning, particularly in the context of complex and dynamic datasets.

# 4. Challenges in Data Cleaning

Data cleaning in large-scale systems poses numerous challenges due to the scale, complexity, and diversity of modern data. These challenges stem from factors such as heterogeneous data sources, real-time processing requirements, and privacy constraints. This section discusses the key challenges in data cleaning and their implications for large-scale systems.

## 4.1 Scalability Challenges

Scalability remains one of the most significant challenges in data cleaning for large-scale systems. As data volumes grow, cleaning processes must efficiently utilize computational and storage resources across distributed environments. Key scalability challenges include:

* **Resource Bottlenecks:** Cleaning pipelines often encounter bottlenecks in CPU, memory, and storage utilization, particularly when processing high-volume data in distributed frameworks [48].
* **Data Skew:** Uneven distributions of data across partitions can lead to performance degradation in distributed cleaning algorithms, as certain nodes become overloaded while others remain underutilized [49].
* **Fault Tolerance:** Distributed cleaning pipelines must ensure fault tolerance by maintaining operations despite node failures or data loss. This requires checkpointing mechanisms and efficient recovery protocols [zaharia2012resilient].

Addressing these challenges requires robust partitioning strategies, resource management frameworks, and fault-tolerant designs that can adapt to the dynamic nature of big data systems.

## 4.2 Data Heterogeneity

Modern systems deal with a wide variety of data formats, schemas, and sources, leading to significant challenges in handling data heterogeneity. Specific challenges include:

* **Schema Matching:** Aligning fields across different datasets with varying schema definitions requires automated schema matching algorithms [51].
* **Data Integration:** Combining data from multiple sources while preserving consistency and accuracy is complex, particularly when dealing with conflicting or overlapping information [52].
* **Handling Noisy and Incomplete Data:** Diverse data sources often have varying levels of noise, missing values, and inaccuracies, complicating the integration and cleaning process [53].

For example, in IoT-based systems, sensor data from heterogeneous devices may exhibit inconsistent formats, missing readings, and noisy signals. Effective tools and frameworks for schema alignment, entity resolution, and data fusion are essential to overcome these challenges.

## 4.3 Real-Time Processing Constraints

Real-time data cleaning is critical for latency-sensitive applications such as IoT, fraud detection, and social media analytics [42]. However, cleaning data streams in real-time introduces several challenges:

* **Latency Sensitivity:** Cleaning operations must meet strict time limits to ensure timely downstream processing.
* **Resource Constraints:** Streaming systems typically operate with limited computational and memory resources, necessitating lightweight cleaning algorithms [54].
* **Dynamic Updates:** Real-time systems must adapt to evolving data patterns, such as schema changes or updated anomaly definitions, without disrupting ongoing operations [55].

For instance, traffic monitoring systems in smart cities rely on real-time cleaning to ensure accurate predictions of congestion levels and environmental conditions [45].

## 4.4 Privacy and Ethical Concerns

Data cleaning processes often require access to sensitive or personal data, which raises privacy and ethical concerns. Improper handling of this data can lead to serious consequences, such as:

* **Data Breaches:** Accidental or intentional exposure of sensitive data during the cleaning process can violate privacy regulations like GDPR or HIPAA [56].
* **Re-Identification Risks:** Even anonymized data can potentially be re-identified when combined with external datasets, posing significant privacy risks [57].
* **Discrimination and Bias:** AI/ML-based cleaning algorithms may perpetuate or amplify biases, leading to unfair or discriminatory outcomes. For example, biased models in financial systems could disproportionately deny loans to certain demographics [27].

Privacy-preserving techniques, such as federated learning and differential privacy, have emerged to address these concerns but require further research to balance scalability, accuracy, and privacy [58].

## 4.5 Interpretability and Explainability

As data cleaning algorithms become increasingly complex, understanding and interpreting their decisions is crucial for building trust and ensuring reliability. Challenges in interpretability and explainability include:

* **Opaque Cleaning Decisions:** AI-driven models may classify data as anomalies or imputations without clear justification, making it difficult to verify their correctness [1].
* **Accountability:** Cleaning decisions directly affect downstream analytics and decision-making, necessitating transparency to ensure accountability.
* **Trustworthiness:** Users must trust the cleaning process, particularly in sensitive domains like healthcare and finance.

Explainable AI (XAI) techniques, such as feature importance analysis and counterfactual explanations, can help improve interpretability by revealing the reasoning behind cleaning decisions [38].

## 4.6 Lack of Standardized Benchmarks

Evaluating the effectiveness of data cleaning techniques is hindered by the absence of standardized benchmarks. Most datasets and metrics used in existing studies are domain-specific or proprietary, limiting their applicability to broader contexts [58]. A standardized benchmarking framework would allow researchers to:

* Compare the performance of different cleaning algorithms on a common set of tasks.
* Assess the trade-offs between accuracy, scalability, and efficiency.
* Identify gaps in existing methods and motivate future research.

Efforts to create open and reproducible benchmarks for data cleaning, such as public datasets and challenge platforms, are still in their early stages [60].

# 5. Open Research Directions

Despite significant progress in data cleaning methodologies, several challenges remain unresolved, presenting opportunities for further research and innovation. This section highlights key areas for future exploration.

## 5.1 Explainable AI for Data Cleaning

While AI/ML techniques have revolutionized data cleaning, their lack of interpretability remains a significant barrier to adoption in critical domains like healthcare and finance. Future research should focus on:

* Developing explainable AI (XAI) models that provide clear and actionable insights into cleaning decisions, such as anomaly detection and missing value imputation [61].
* Creating counterfactual explanations to demonstrate how minor changes in data would affect cleaning outcomes [62].
* Designing interactive tools that enable domain experts to visualize and validate cleaning processes, enhancing trust and transparency.

These advancements will improve accountability and ensure that cleaning algorithms align with domain-specific requirements.

## 5.2 Cross-Domain Cleaning Techniques

Datasets often originate from diverse domains, such as healthcare, finance, and IoT, requiring cleaning techniques that generalize across different contexts. Specific challenges in cross-domain cleaning include:

* **Semantic Heterogeneity:** Handling inconsistencies in terminology and semantic interpretations across domains. For example, a "patient ID" in healthcare might correspond to "customer ID" in insurance data, requiring careful mapping [63].
* **Data Integration Challenges:** Combining data from disparate sources with varying levels of quality, formats, and granularity. Aligning such data while maintaining accuracy and consistency is a major challenge [64].
* **Domain Adaptation:** Adapting cleaning models trained on one domain (e.g., retail) to clean data in another domain (e.g., logistics) without requiring extensive retraining [65].

Transfer learning and meta-learning approaches could be explored to address these challenges, enabling cross-domain generalization in data cleaning frameworks.

## 5.3 Federated and Privacy-Preserving Cleaning

As privacy regulations like GDPR and HIPAA become stricter, cleaning techniques must evolve to maintain compliance while ensuring data utility. Privacy preservation techniques offer promising solutions, but several open research questions remain:

* **Federated Learning:** Federated cleaning allows models to be trained across distributed datasets without sharing raw data, but ensuring accuracy and scalability in such settings remains a challenge [64].
* **Differential Privacy:** Adding noise to datasets to prevent individual records from being identified is promising but can reduce the utility of cleaned data. Research is needed to optimize the balance between privacy and data quality [66].
* **Homomorphic Encryption:** Performing computations on encrypted data without decryption offers strong privacy guarantees but introduces computational overhead that limits its scalability [67].
* **Secure Multi-Party Computation (SMPC):** SMPC enables collaborative data cleaning across parties while preserving privacy but requires improvements in efficiency and usability for large-scale systems [68].

Privacy-preserving cleaning methods must address trade-offs between computational efficiency, privacy guarantees, and cleaning accuracy.

## 5.4 Human-in-the-Loop Systems

Human expertise plays a crucial role in data cleaning processes, especially for tasks requiring domain knowledge or contextual understanding. Open research directions include:

* **Active Learning:** Combining human feedback with machine learning to iteratively improve cleaning accuracy and reduce labelling effort. For example, active learning can prioritize uncertain or ambiguous cases for human review [69].
* **Crowdsourcing:** Leveraging human intelligence through platforms like Amazon Mechanical Turk to perform specific cleaning tasks, such as labelling, validation, or resolving semantic conflicts [70].
* **Interactive Tools:** Developing user-friendly interfaces that allow domain experts to inspect, refine, and validate cleaning workflows, bridging the gap between automation and human oversight.

Human-in-the-loop systems combine the strengths of human expertise and automated techniques, making them highly effective for complex cleaning scenarios.

## 5.5 Sustainability and Energy-Efficient Cleaning

As the scale of data cleaning grows, its environmental impact becomes a critical concern. Open research directions include:

* **Green Computing Technologies:** Exploring the use of energy-efficient hardware and renewable energy sources to reduce the carbon footprint of large-scale cleaning workflows [71].
* **Energy-Efficient Algorithms:** Designing algorithms that minimize resource consumption while maintaining high cleaning accuracy. For example, approximate algorithms could be explored for certain cleaning tasks [72].
* **Optimized Cloud Deployment:** Investigating strategies to deploy cleaning pipelines on cloud platforms in a way that minimizes energy usage without compromising scalability.

Energy-efficient cleaning is particularly relevant for cloud and edge environments, where resource constraints are common.

## 5.6 Standardized Benchmarks and Evaluation Metrics

The lack of standardized benchmarks and metrics for evaluating data cleaning techniques hinders the comparability of research outcomes. Future work should focus on:

* Developing open-access benchmarks that include diverse datasets and cleaning scenarios [73].
* Establishing metrics that evaluate trade-offs between accuracy, scalability, and efficiency.
* Encouraging community-driven initiatives to build reproducible and transparent evaluation platforms.
* Scalability and adaptability to different industries and ensuring system reliability in real-time [74].

Standardized benchmarks will accelerate progress by providing a common framework for testing and comparing cleaning techniques.

# Conclusion

High-quality data is foundational to effective decision-making, enhanced system performance, and increased user trust, as highlighted in the introduction. Data cleaning plays a pivotal role in achieving these outcomes by ensuring the reliability and usability of large-scale systems across diverse domains. This paper has provided a comprehensive review of the current trends, challenges, and open research directions in data cleaning, focusing on emerging methodologies such as AI-driven automation, real-time frameworks, and privacy-preserving techniques.

The increasing reliance on heterogeneous, high-velocity data in modern applications has underscored the need for scalable, efficient, and ethical cleaning solutions. Challenges such as resource bottlenecks, data heterogeneity, and the lack of standardized benchmarks require innovative approaches that balance performance, compliance, and transparency. Recent advancements, including Explainable AI (XAI), graph-based cleaning, and federated learning, illustrate how the field is evolving to meet these demands.

Promising areas for future exploration include energy-efficient cleaning techniques, cross-domain approaches, and human-in-the-loop systems that integrate domain expertise with automated methods. Addressing these open research directions will be critical for ensuring that data cleaning frameworks align with the growing complexity and ethical considerations of modern systems.

In conclusion, as the volume and diversity of data continue to grow, collaborative efforts among researchers, practitioners, and policymakers are essential to develop robust, transparent, and sustainable data cleaning solutions. This paper aims to serve as a foundational resource for guiding innovation and addressing unresolved challenges in this critical area.

**References:**

1. J. M. Hellerstein, “Quantitative Data Cleaning for Large Databases,” 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:3223173
2. C. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. New York, NY, USA: Springer, 2006.
3. M. Badawy, N. Ramadan, and H. A. Hefny, “Big Data Analytics in Healthcare: Data Sources, Tools, Challenges, and Opportunities,” Journal of Electrical Systems and Information Technology, vol. 11, no. 1, p. 63, Dec. 2024. doi: 10.1186/s43067-024-00190-w.
4. F. T. Johora, R. Hasan, S. F. Farabi, J. Akter, and M. A. A. Mahmud, “AI-powered Fraud Detection in Banking: Safeguarding Financial Transactions,” The American Journal of Management and Economics Innovations, vol. 6, no. 06, pp. 8–22, Jun. 2024. doi: 10.37547/tajmei/Volume06Issue06-02.
5. A. Balahur et al., “Data Quality Requirements for Inclusive, Non-Biased and Trustworthy AI,” working paper, Dec. 2022. doi: 10.2760/365479. [Online]. Available: https://hal.science/hal-03889925
6. J. Buolamwini and T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” Proceedings of FAT, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3298854
7. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA, USA: MIT Press, 2019. [Online]. Available: https://fairmlbook.org/
8. I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, “Big Data Quality Framework: A Holistic Approach to Continuous Quality Management,” Journal of Big Data, vol. 8, no. 1, p. 76, May 2021. doi: 10.1186/s40537-021-00468-0.
9. E. K. Rezig et al., “Horizon: Scalable Dependency-Driven Data Cleaning,” Proc. VLDB Endowment, vol. 14, no. 11, pp. 2546–2554, Jul. 2021. doi: 10.14778/3476249.3476301.
10. S. Kandel et al., “Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data,” Information Visualization, vol. 10, no. 4, pp. 271–288, 2011. doi: 10.1177/1473871611415994.
11. K. R. Gade, “Data Quality Metrics for the Modern Enterprise: A Data Analytics Perspective,” MZ Journal of Artificial Intelligence, vol. 1, no. 1, Apr. 2024. [Online]. Available: https://mzjournal.com/index.php/MZJAI/article/view/416
12. D. Loshin, Master Data Management, 1st ed. Elsevier Science, Sep. 2008. ISBN: 9780123742254 (Paperback), 9780080921211 (eBook).
13. L. Hong et al., “Big Data in Health Care: Applications and Challenges,” Data and Information Management, vol. 2, no. 3, pp. 175–197, Dec. 2018. doi: 10.2478/dim-2018-0014.
14. T. Islam et al., “Artificial Intelligence in Fraud Detection and Financial Risk Mitigation: Future Directions and Business Applications,” International Journal for Multidisciplinary Research, vol. 6, pp. 1–23, Oct. 2024. doi: 10.36948/ijfmr.2024.v06i05.28496.
15. T. Bhattacharya et al., “A Survey on Various Security Protocols of Edge Computing,” The Journal of Supercomputing, vol. 81, no. 1, p. 310, Dec. 2024. doi: 10.1007/s11227-024-06678-6.
16. B. N. Silva et al., “Urban Planning and Smart City Decision Management Empowered by Real-Time Data Processing Using Big Data Analytics,” Sensors, vol. 18, no. 9, Article 2994, 2018. doi: 10.3390/s18092994.
17. M. Zaharia et al., “Apache Spark: A Unified Engine for Big Data Processing,” Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.
18. Y. Wang, J. Wong, and A. Miner, “Anomaly Intrusion Detection Using One-Class SVM,” in Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004, pp. 358–364, 2004. doi: 10.1109/IAW.2004.1437839.
19. S. Trilles, S. S. Hammad, and D. Iskandaryan, “Anomaly Detection Based on Artificial Intelligence of Things: A Systematic Literature Mapping,” Internet of Things, vol. 25, p. 101063, 2024. doi: 10.1016/j.iot.2024.101063.
20. S. Nalluri, M. M. Malyala, H. Kandagiri, and K. K. Kandagiri, “Emission minimization through IoT-Integrated Intelligent Transportation System,” in IGI Global eBooks, 2025, pp. 579–598. doi: 10.4018/979-8-3693-7984-4.ch031.
21. T. Gebru et al., “Datasheets for Datasets,” arXiv, 2021. [Online]. Available: https://arxiv.org/abs/1803.09010
22. N. Bhansali, Ed., Data Governance: Creating Value from Information Assets, 1st ed. Auerbach Publications, 2013. doi: 10.1201/b15034.
23. H. Chhinkaniwala and S. Garg, “Privacy Preserving Data Mining Techniques: Challenges & Issues,” in Proceedings of the Conference on Privacy Preserving Data Mining Techniques, Jul. 2011. doi: 10.13140/RG.2.1.1811.3526.
24. S. Longpre et al., “A Large-Scale Audit of Dataset Licensing and Attribution in AI,” Nature Machine Intelligence, vol. 6, no. 8, pp. 975–987, Aug. 2024. doi: 10.1038/s42256-024-00878-8.
25. D. Zha et al., “Data-Centric Artificial Intelligence: A Survey,” ACM Computing Surveys, Jan. 2025. doi: 10.1145/3711118.
26. S. Khan, “Anomaly Detection with Machine Learning,” Medium, Jul. 2023. [Online]. Available: https://sulaimank.medium.com/anomaly-detection-with-machine-learning-8fa942fb5adc
27. M. N. Arefin and A. K. M. Masum, “A Probabilistic Approach for Missing Data Imputation,” Complexity, vol. 2024, no. 1, Article 4737963, 2024. doi: 10.1155/2024/4737963.
28. X. Ma et al., “Role of Big Data and Technological Advancements in Monitoring and Development of Smart Cities,” Heliyon, vol. 10, no. 15, Article e34821, 2024. doi: 10.1016/j.heliyon.2024.e34821.
29. E. Cambria et al., A Practical Guide to Sentiment Analysis, 1st ed. Springer Publishing Company, Inc., 2017. doi: 10.5555/3110802.
30. A. C. Calheiros et al., “Sentiment Classification of Consumer-Generated Online Reviews Using Topic Modeling,” Journal of Hospitality Marketing & Management, vol. 26, no. 7, pp. 675–693, 2017. doi: 10.1080/19368623.2017.1310075.
31. T. M. Lehmann, C. Gonner, and K. Spitzer, “Survey: Interpolation Methods in Medical Image Processing,” IEEE Transactions on Medical Imaging, vol. 18, no. 11, pp. 1049–1075, 1999. doi: 10.1109/42.816070.
32. A. Paleyes, “Towards Maintainable and Explainable AI Systems with Dataflow,” Ph.D. dissertation, Apollo - University of Cambridge Repository, 2023. doi: 10.17863/CAM.108344. [Online]. Available: https://www.repository.cam.ac.uk/handle/1810/367937
33. I. Stepin et al., “A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable AI,” IEEE Access, vol. 9, pp. 11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315.
34. AWS Glue, “Data Cleaning Pipelines on AWS,” [Online]. Available: https://aws.amazon.com/glue/
35. E. Jonas et al., “Hybrid Cloud Systems for Large-Scale Data Cleaning,” ACM Computing Surveys, vol. 54, no. 6, pp. 1–34, 2021.
36. C. Lin et al., “Real-Time Data Quality Management for IoT-Based Smart City Applications,” IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4245–4260, 2020.
37. T. Akidau et al., “The Dataflow Model: A Practical Approach to Real-Time Data Cleaning,” Proc. ACM SIGMOD, 2015, pp. 1–10.
38. P. Zhang et al., “AI-Driven Anomaly Detection in IoT Sensor Data,” IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8928–8938, 2019.
39. S. Dey et al., “Traffic Monitoring Using Real-Time Cleaning in Smart Cities,” IEEE Transactions on Smart Cities, vol. 5, no. 2, pp. 123–137, 2022.
40. J. Tang et al., “Graph-Based Data Cleaning: Challenges and Opportunities,” ACM Transactions on Knowledge Discovery from Data, vol. 15, no. 3, pp. 1–24, 2021.
41. K. Settles et al., “Active Learning for Data Cleaning,” Proc. Neural Information Processing Systems (NeurIPS), 2019, pp. 1–12.
42. A. Halevy et al., “Data Cleaning at Scale: Current Challenges and Future Directions,” IEEE Data Eng. Bull., vol. 44, no. 2, pp. 27–38, 2021.
43. T. Akidau et al., “Streaming Data Processing with Apache Beam,” Communications of the ACM, vol. 60, no. 4, pp. 60–71, 2017.
44. M. Zaharia et al., “Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing,” in Proc. NSDI, 2012, pp. 15–28. [Online]. Available: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia
45. E. Rahm and H. Do, “Data Cleaning: Problems and Current Approaches,” IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3–13, 2000. [Online]. Available: https://www.betterevaluation.org/sites/default/files/data\_cleaning.pdf
46. P. Zhang et al., “Cleaning Heterogeneous Data in Large-Scale Systems,” ACM Transactions on Database Systems, vol. 43, no. 2, pp. 1–26, 2018.
47. X. Hu et al., “Efficient Data Cleaning for Streaming Systems: A Lightweight Approach,” IEEE Internet of Things Journal, vol. 8, no. 2, pp. 1407–1418, 2021.
48. C. Lin et al., “Dynamic Schema Matching for Evolving Data Streams,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2600–2613, 2021.
49. A. Narayanan et al., “Privacy-Preserving Techniques in Data Cleaning: Challenges and Solutions,” ACM Transactions on Privacy and Security, vol. 24, no. 1, pp. 1–26, 2021.
50. A. Zawawy et al., “Open Benchmarks for Evaluating Data Cleaning Techniques,” Proc. IEEE Int. Conf. Data Eng. (ICDE), 2022, pp. 450–460.
51. C. Dwork et al., “Differential Privacy for Big Data: Challenges and Applications,” ACM Computing Surveys, vol. 54, no. 2, pp. 1–28, 2021.
52. S. Wachter et al., “Counterfactual Explanations in AI: A Survey,” Journal of AI Research, vol. 63, pp. 1–32, 2018.
53. G. Holzinger, “Data Cleaning Benchmarks for Distributed Systems,” Proc. ACM SIGMOD, 2021, pp. 122–140.
54. T. Miller et al., “Explainable AI in Data Cleaning: Current Trends and Challenges,” AI Magazine, vol. 40, no. 3, pp. 45–57, 2021.
55. A. Sahai et al., “Homomorphic Encryption for Secure and Scalable Data Cleaning,” IEEE Security and Privacy, vol. 16, no. 4, pp. 54–62, 2018.
56. K. Bonawitz et al., “Secure Multi-Party Computation for Data Cleaning in Distributed Systems,” Journal of Privacy and Confidentiality, vol. 10, no. 1, pp. 1–25, 2020.
57. A. Parameswaran et al., “Crowdsourcing for Data Cleaning: A Survey of Tools and Techniques,” Proc. VLDB Endowment, vol. 9, no. 10, pp. 802–813, 2016.
58. M. Zaharia et al., “Sustainability in Data Processing: Energy-Efficient Algorithms for Big Data Cleaning,” Communications of the ACM, vol. 64, no. 6, pp. 45–56, 2021.
59. E. Jonas et al., “Green Computing Frameworks for Energy-Efficient Data Cleaning,” ACM Transactions on Computing Systems, vol. 37, no. 3, pp. 1–20, 2021.
60. Shahri, H. H., “A Machine Learning Approach to Data Cleaning in Databases and Data Warehouses,” Database Technologies: Concepts, Methodologies, Tools, and Applications, pp. 2245–2260, IGI Global, 2009. [Online]. Available: https://www.igi-global.com/chapter/machine-learning-approach-data-cleaning/8033
61. A. Narayanan et al., “Federated Learning for Privacy-Preserving Data Cleaning: A Case Study,” Proc. IEEE Int. Conf. Privacy Sec., 2020, pp. 34–45.
62. Z. Ghahramani, “Probabilistic Models for Privacy-Preserving Cleaning,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 7, pp. 1231–1245, 2021.
63. C. Dwork et al., “Differential Privacy for Big Data: Challenges and Applications,” ACM Computing Surveys, vol. 54, no. 2, pp. 1–28, 2021.
64. A. Sahai et al., “Homomorphic Encryption for Secure and Scalable Data Cleaning,” IEEE Security and Privacy, vol. 16, no. 4, pp. 54–62, 2018.
65. K. Bonawitz et al., “Secure Multi-Party Computation for Data Cleaning in Distributed Systems,” Journal of Privacy and Confidentiality, vol. 10, no. 1, pp. 1–25, 2020.
66. A. Parameswaran et al., “Crowdsourcing for Data Cleaning: A Survey of Tools and Techniques,” Proc. VLDB Endowment, vol. 9, no. 10, pp. 802–813, 2016.
67. M. Zaharia et al., “Sustainability in Data Processing: Energy-Efficient Algorithms for Big Data Cleaning,” Communications of the ACM, vol. 64, no. 6, pp. 45–56, 2021.
68. E. Jonas et al., “Green Computing Frameworks for Energy-Efficient Data Cleaning,” ACM Transactions on Computing Systems, vol. 37, no. 3, pp. 1–20, 2021.
69. T. Akidau et al., “Streaming Data Processing with Apache Beam,” Communications of the ACM, vol. 60, no. 4, pp. 60–71, 2017.
70. J. Tang et al., “Graph-Based Data Cleaning: Challenges and Opportunities,” ACM Transactions on Knowledge Discovery from Data, vol. 15, no. 3, pp. 1–24, 2021.
71. G. Holzinger, “Data Cleaning Benchmarks for Distributed Systems,” Proc. ACM SIGMOD, 2021, pp. 122–140.
72. T. Miller et al., “Explainable AI in Data Cleaning: Current Trends and Challenges,” AI Magazine, vol. 40, no. 3, pp. 45–57, 2021.
73. A. Zawawy et al., “Open Benchmarks for Evaluating Data Cleaning Techniques,” Proc. IEEE Int. Conf. Data Eng. (ICDE), 2022, pp. 450–460.
74. R. Venkataraju and C. Srinivas, "Big Data-Driven Disease Progression Prediction Using IoT Healthcare Devices: Opportunities and Challenges," International Journal of Computer Engineering and Technology (IJCET), vol. 6, no. 12, pp. 72-87, 2015.
75. F. Quazi et al., "Data Security & Privacy in Healthcare," International Journal of Global Innovations and Solutions (IJGIS), 2024. doi: 10.21428/e90189c8.4e2c586a.
76. S. Nalluri et al., "Cybersecurity risk management in cloud computing environment," International Journal of Science and Research Archive, vol. 10, no. 1, pp. 1062-1068, 2023.
77. S. Nalluri, M. M. Malyala, H. Kandagiri, and K. K. Kandagiri, “Emission minimization through IoT-Integrated Intelligent Transportation System,” in IGI Global eBooks, 2025, pp. 579–598. doi: 10.4018/979-8-3693-7984-4.ch031.

1. The next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, First A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov).

   Second B. Author Jr. was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

   Third C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba 305-0047, Japan (e-mail: author@nrim.go.jp). [↑](#footnote-ref-1)