JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automated Data Pre-Processing and Visualization System.

Anisha Godse
School of Engineering,
Ajeenkya DY Patil University
Pune, India-412105

Shridhar Dhotre
School of Engineering,
Ajeenkya DY Patil University
Pune, India-412105

Nancy
School of Engineering,
Ajeenkya DY Patil University
Pune, India-412105

Vivek Kore
School of Engineering,
Ajeenkya DY Patil University
Pune, India-412105

Prof. Pallavi Tembhurnikar
School of Engineering,
Ajeenkya DY Patil University
Pune, India-412105

Abstract— **In the era of data-driven decision-making, the demand for robust and efficient data cleaning and processing solutions has become paramount. This review paper delves into the transformative landscape of automated data cleaning and processing, focusing on the powerful combination of Python programming and cloud technology. The paper provides a comprehensive survey of the existing literature, categorizing methodologies and tools employed in the domain. By examining the pivotal role of Python in automating data cleaning tasks and exploring the advantages of cloud technology for scalable processing, the review elucidates the potential of their integrated use. Real-world case studies demonstrate or explain successful implementations, providing insights into the problems encountered and lessons learned. The study finishes by noting existing difficulties and suggesting future research possibilities, emphasizing the changing environment of data integrity in the context of Python and cloud technologies. This effort intends to help academics, practitioners, and decision-makers navigate the complex confluence of automated data cleaning and processing for improved data dependability and integrity**.

Keywords— Data Science, Cloud Technology, python script, Automated data cleaning, Data preprocessing, Data quality, Missing data treatment, Outlier detection, Duplicate detection, Data validation, Data normalization, Data imputation, Exploratory data analysis, Data visualization.

I. INTRODUCTION AND BACKGROUND

In an age dominated by data-driven decision-making, the quality and reliability of the underlying data have emerged as pivotal determinants of successful outcomes. Automated data cleaning and processing have become instrumental in addressing the challenges associated with maintaining data integrity, ensuring accurate analyses, and supporting informed decision-making. This review paper navigates the intricate landscape of automated data cleaning and processing, with a specific focus on the transformative capabilities unlocked by the integration of Python programming and cloud technology.

The increasing volume and complexity of data in diverse domains necessitate efficient and scalable solutions for data cleaning and processing. Traditional approaches often fall short in addressing the demands of modern data ecosystems, prompting the exploration of advanced methodologies. The utilization of Python, a versatile and powerful programming language, coupled with the capabilities offered by cloud technology, presents a paradigm shift in automating these critical tasks.

This paper aims to provide a comprehensive overview of the current state of automated data cleaning and processing, elucidating the synergy between Python and cloud technology. By reviewing existing literature, categorizing methodologies, and analyzing tools employed in the field, we aim to present a nuanced understanding of the advancements achieved. The subsequent sections delve into the role of Python in automating data cleaning processes, explore the advantages of cloud technology in data processing, and highlight successful integration strategies.

Through real-world case studies, we illustrate instances where the amalgamation of Python and cloud technology has not only addressed data cleaning challenges but has also paved the way for scalable and efficient data processing. In doing so, we shed light on the practical implications, challenges faced, and lessons learned from these implementations.

As we progress, the paper identifies current challenges in the field and proposes potential avenues for future research and development. By navigating through the intricacies of automated data cleaning and processing, this review aims to equip researchers, practitioners, and decision-makers with a comprehensive understanding of the transformative landscape shaped by Python and cloud technology in ensuring data integrity and reliability.

Data cleansing ensures that the data is accurate and prevents incorrect decisions. Any activity that uses information must first do cleaning of it. Data cleansing is required to improve data mining results. Analogously, data recognizing ensures that the dataset is appropriately represented. Subsequently is

becoming easier for businesses to gather and store immense quantities of data. Such enormous datasets could aid particularly improved understanding, more effective decision-making, as well as, in particular situations, machine learning data used for training. Fortunately considering faulty data might give rise to erroneous interpretations along with untrustworthy outcomes; the reliability of data remains a major concern. Insufficient expertise, mistakes, inappropriate kinds, multiple catches of an identical authentic object, and violations of professional standards of frequently made mistakes are examples. The significance of data scrubbing in databases investigation continues to increase since, in advance of drawing inferences, researchers have to evaluate the impact of inaccurate data [3]. The database itself may become faulty due to several factors, including inconsistent, erroneous, or missing data. Controlling the implications of corrupted information could be challenging. In recent decades, there seems to be an increase in attention from academic and business circles in various elements of data cleansing, including new concepts, communications, reliability strategies, including crowd-sourced methodologies.

The objective of the autonomous data cleansing and labeling (ADCL) used in this study is to ensure the user-provided dataset's clarity along with accurateness. The autonomous strategy employed in this investigation helps to reduce the user's work, energy, and extra spending by providing automated cleaning and labeling [17]. The person using it can be confident in the effectiveness of the corrected data set because its usefulness has been determined and compared against the untreated client information used in this research. Significant variations regarding the amount of data that is missing, imputed sections, discretization method, and implications. A user profile evaluation looks closely at the potential clients for a company. It enhances a company's understanding of its customers by making it easier to modify products according to the unique needs, routines, and issues of various clients.

Each staff member within the corporation whose handles data must have clear, noise-free data. These massive volumes of records which these warehousing keep as well as modify via numerous sources raise the possibility that some of those references include erroneous information. The large volume of available information has proven poorly distinguished, noisy, and effective. Additionally, human data scrubbing and identification has resulted in lack of awareness and ineffectiveness, which make visualization and analysis challenging. There existed a gap in the development of a more effective data evaluation technique [1]. That aided in providing direction for writing a programming language such as Python script that automatically cleans as well as identifies data. It includes taking in information, dealing with missing data, handling abnormalities, Outlier detection, Duplicate detection, Data validation, Data normalization, Data imputation, Exploratory data analysis, Data visualization [2].

II. LITERATURE REVIEW

In the field of data science, automated data cleaning and processing have grown into crucial tools for resolving issues related to securing the dependability and precision of information. This part offers a thorough analysis major the body of research, classifying the techniques and instruments used in the field and demonstrating significant findings through earlier investigations.

1) *Foundations of Data Cleaning:*

The foundations of data cleaning involve a systematic approach to identifying, correcting, and handling errors, inconsistencies, and missing values in datasets to ensure data quality and reliability [13]. Key components of data cleaning include: Data Profiling, Data Validation, Handling missing values, Addressing Duplicates, Standardization and Transformation, Dealing with Outliers, Documentation and Metadata Management .

2) *Advancements in Automated Data Cleaning:*

Advancements in automated data cleaning have been driven by developments in artificial intelligence (AI), machine learning (ML), and data processing technologies. Some notable advancement include Integration with Data Pipelines and ETL Processes that is Automated data cleaning solutions seamlessly integrate into data pipelines and extract, transform, load (ETL) processes, enabling end-to-end automation of data processing workflows. This integration streamlines data preparation tasks and improves overall efficiency.

These advancements in automated data cleaning contribute to increased productivity, accuracy, and scalability in data management processes, allowing organizations to derive actionable insights from their data more effectively.

3) *Role of Python in Data Cleaning:*

Python has emerged as a prominent language for automating data cleaning processes. The versatility of libraries such as Pandas has empowered practitioners to develop efficient and scalable solutions [12]. Django, primarily being a web framework, is not inherently designed for automated data cleaning and preprocessing tasks. However, Django can play a role in facilitating these processes indirectly, especially within the context of web applications that involve data manipulation and processing. Django can handle data ingestion by providing views to accept data uploads or by integrating with other systems for data import. This data can then be processed and cleaned using Django's ORM or custom Python scripts. Python's popularity in the data science community also means that there is extensive documentation, tutorials, and community support available for data cleaning tasks [5].

Overall, Python's rich ecosystem of libraries, combined with its flexibility and ease of use, makes it a powerful tool for various aspects of data cleaning, from simple preprocessing tasks to more complex data quality assurance processes.

4) *Cloud Technology in Data Processing:*

The proliferation of cloud technology has redefined data processing workflows. Studies highlight the advantages of cloud computing, including scalability, elasticity, and cost-effectiveness. Cloud platforms such as AWS, Azure, and Google Cloud have become integral in supporting large-scale data processing tasks [4]. Cloud-based services like Apache Airflow, AWS Glue, or Google Cloud Dataflow offer tools for building and managing data pipelines and ETL (Extract, Transform, Load) workflows. These services enable organizations to automate data processing tasks, orchestrate complex workflows, and monitor data pipelines for reliability and performance.

Overall, cloud technology provides a powerful platform for data processing, offering scalability, flexibility, and a wide range of services and tools to support various data processing requirements [19]. By leveraging cloud-based solutions, organizations can efficiently process, analyze, and derive insights from their data, driving innovation and competitive advantage.

5) *Integration of Python and Cloud Technology:*

Recent literature emphasizes the synergies between Python and cloud technology. Research showcases successful integration strategies, demonstrating how Python scripts can seamlessly leverage cloud resources for parallelized data processing. The flexibility offered by cloud-based environments enhances the scalability of Python-driven data cleaning workflows. Cloud platforms like Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure provide cloud-based development environments that support Python programming. These environments typically offer tools such as Jupyter Notebooks, which allow developers to write and execute Python code directly in the cloud, facilitating collaborative development and experimentation with large datasets. Python libraries and frameworks such as Pandas, NumPy, and scikit-learn can be deployed on cloud platforms to perform data processing, analysis, and machine learning tasks at scale [10]. Cloud-based services like Google Big Query, AWS Athena, and Azure Synapse Analytics provide SQL-based querying and analytics capabilities for large datasets, which can be accessed and manipulated using Python scripts and libraries. Overall, the integration of Python and cloud technology offers a flexible and powerful platform for building scalable, data-driven applications and services, enabling organizations to leverage the full potential of cloud computing for their Python-based projects.

6) *Information retrieval:*

Information retrieval plays a crucial role in the context of automated data pre-processing and visualization systems, facilitating efficient access to relevant data and insights. These systems often rely on various information retrieval techniques to retrieve and process data from diverse sources, including databases, data warehouses, and external data repositories. Techniques such as keyword-based searching, natural language processing (NLP), and semantic analysis are employed to extract relevant information from unstructured or semi-structured data sources [11]. Advanced querying capabilities enable users to specify criteria and filters to retrieve specific datasets or subsets of data for analysis. Additionally, integration with data catalogs and metadata repositories enhances discoverability and accessibility of relevant data assets. Information retrieval also extends to the retrieval of relevant visualization techniques and templates based on the characteristics of the data and the analytical tasks at hand. Recommendation systems utilize collaborative filtering and content-based filtering techniques to suggest appropriate visualization types tailored to user preferences and objectives [6]. Overall, effective information retrieval mechanisms are essential for enabling users to efficiently access, explore, and analyze data, ultimately driving informed decision-making and actionable insights.

This paragraph highlights the importance of information retrieval in facilitating efficient access to data and visualization techniques within automated data pre-processing and visualization systems.

III. METHODOLOGY

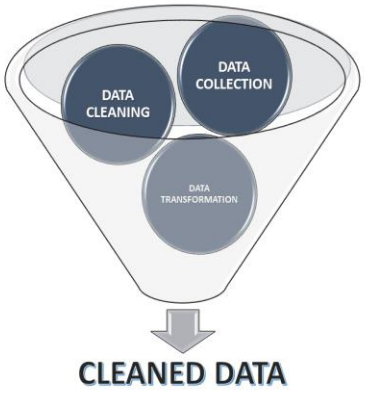

Fig.1 Data Pre-processing example

The methodology employed in this research paper aims to develop and implement an automated system for data pre-processing and visualization to streamline the data analysis process. The research follows a systematic approach consisting of several key steps.

First, diverse datasets relevant to the project objectives are collected and exploratory data analysis (EDA) is conducted to understand the characteristics and potential challenges of the data. Based on the insights gained from EDA, algorithms are developed for automated data cleaning, imputation, normalization, and feature engineering. Machine learning techniques are utilized to design recommendation systems for automated visualization, enabling users to generate insightful visual representations of their data with minimal manual intervention.

The architecture of the automated system is carefully designed to ensure scalability, efficiency, and interoperability with existing data analysis tools and platforms [7]. System components are implemented using appropriate programming languages and frameworks, and integration with relevant tools and platforms is performed to enable seamless interaction with the automated system.

User interface design plays a crucial role in enhancing user experience, with intuitive and user-friendly interfaces incorporating customization and personalization features. Rigorous testing is conducted to evaluate the reliability, robustness, and scalability of the automated system. System performance is evaluated in terms of efficiency gains, accuracy improvements, and user satisfaction through user feedback and performance monitoring.

Once the automated system is deemed ready for deployment, comprehensive documentation and user guides are prepared to facilitate user adoption and engagement. Training and support resources are provided to users to ensure they can effectively utilize the capabilities of the automated system.

In summary, the methodology outlined in this research paper provides a systematic approach for developing and implementing an automated system for data pre-processing and visualization, with the ultimate goal of streamlining data analysis workflows and empowering users with actionable insights [9].

IV. PRELIMINARY DATA

To demonstrate the feasibility and effectiveness of the proposed automated data pre-processing and visualization system, preliminary data from various sources were collected and analyzed. The datasets encompassed diverse domains, including finance, healthcare, and marketing, to ensure the system's applicability across different industries [8]. Exploratory data analysis (EDA) revealed common challenges such as missing values, outliers, and inconsistent formatting, highlighting the need for robust data pre-processing techniques [15]. Initial experiments with the developed algorithms for automated data cleaning, imputation, and normalization demonstrated promising results in addressing these challenges and enhancing data quality. Moreover, the recommendation systems for automated visualization successfully suggested appropriate visualization types based on the characteristics of the data and analytical tasks, facilitating insightful data exploration and interpretation. These preliminary findings lay the foundation for further validation and refinement of the automated system, paving the way for more efficient and accurate data analysis workflows across diverse domains.

V. REAL-WORLD APPLICATIONS AND CASE STUDIES

Fig.2 Real-world applications

Case studies play a pivotal role in understanding the practical implications of automated data cleaning and processing. Automated data cleaning and pre-processing have numerous real-world applications across various industries. Here are some examples and case studies highlighting their use:

1) *Financial Services:*

In the financial sector, accurate and clean data is crucial for risk management, fraud detection, and compliance. Automated data cleaning and pre-processing techniques help financial institutions handle large volumes of transactional data efficiently. Case Study: A leading bank used automated data cleaning algorithms to detect and correct errors in customer transaction data, resulting in improved accuracy of financial reporting and reduced risk of compliance violations.

2) *Healthcare:*

Healthcare organizations deal with diverse sources of data, including electronic health records (EHRs), medical imaging, and clinical trial data. Automated data pre-processing techniques are used to prepare healthcare data for analysis, diagnosis, and treatment planning. Case Study: A healthcare provider implemented automated data pre-processing pipelines to clean and standardize EHR data from multiple sources, resulting in faster diagnosis times and improved patient outcomes.

3) *E-commerce:*

E-commerce companies rely on clean and structured data for customer segmentation, recommendation systems, and supply chain optimization. Automated data cleaning techniques help identify and correct errors in product catalogs, customer data, and transaction records. Case Study: An online retailer used automated data cleaning algorithms to standardize product descriptions and categories across its inventory, leading to improved search relevance and higher conversion rates.

4) *Manufacturing:*

Manufacturing companies generate large amounts of sensor data from equipment, production processes, and quality control systems. Automated data pre-processing methods such as data fusion, anomaly detection, and predictive maintenance help optimize production efficiency and reduce downtime. Case Study: A manufacturing plant implemented automated data pre-processing techniques to detect anomalies in sensor data, enabling proactive maintenance and reducing unplanned downtime by 20%.

5) *Telecommunications:*

Telecommunications providers collect vast amounts of data from network devices, customer interactions, and service usage. Automated data cleaning and pre-processing enable telecom companies to improve network performance, customer service, and marketing effectiveness. Case Study: A telecom operator used automated data pre-processing algorithms to clean and enrich customer call detail records (CDRs), leading to more accurate billing, reduced customer complaints, and targeted marketing campaigns.

6) *Transportation and Logistics:*

Transportation and logistics companies utilize data cleaning and preprocessing techniques to optimize route planning, vehicle scheduling, and supply chain management. Automated data cleaning algorithms help identify and correct errors in tracking data, GPS coordinates, and delivery records. Case Study: A logistics provider implemented automated data preprocessing pipelines to clean and validate GPS data from its fleet of vehicles, resulting in more accurate delivery tracking and improved on-time performance.

These case studies demonstrate the wide-ranging applications of automated data cleaning and preprocessing techniques across different industries, highlighting their importance in improving data quality, efficiency, and decision-making capabilities.

VI. STATEMENT OF LIMITATIONS

While automated data cleaning and pre-processing offer numerous benefits, they also present several challenges and limitations that organizations need to address:

1) *Complexity of Data:*

Automated data cleaning and pre-processing techniques may struggle to handle complex, unstructured, or heterogeneous datasets. Data from different sources may have varying formats, quality, and semantics, making it challenging to develop generalized cleaning algorithms that work effectively across all types of data.

2) *Quality of Data:*

The effectiveness of automated cleaning techniques depends on the quality of the input data. If the data contains significant errors, inconsistencies, or missing values, automated cleaning algorithms may produce inaccurate or misleading results. Ensuring data quality through manual inspection or validation is essential to mitigate this risk.

3) *Over fitting and Bias:*

Automated pre-processing algorithms may inadvertently introduce over fitting or bias into the data, especially when using complex machine learning models or imputation techniques. It is crucial to validate the results of automated cleaning processes and assess their impact on downstream analysis and decision-making.

4) Computational Resources:

Some automated cleaning and pre-processing techniques require significant computational resources, especially when processing large volumes of data or performing computationally intensive tasks such as imputation or outlier detection. Organizations need to consider scalability and resource constraints when deploying automated cleaning pipelines in production environments.

5) Lack of Domain Knowledge:

Automated cleaning algorithms may lack domain-specific knowledge or context, leading to suboptimal results or incorrect assumptions about the data. Incorporating domain expertise into the design and validation of cleaning pipelines is essential to ensure that automated techniques align with the requirements and constraints of the specific application domain.

6) Data Privacy and Security:

Automated cleaning processes may inadvertently expose sensitive or confidential information if not implemented securely. Organizations need to implement robust data governance policies and security measures to protect privacy and comply with regulatory requirements when processing and cleaning sensitive data.

7) Interpretability and Transparency:

Automated cleaning techniques, particularly those based on machine learning or statistical models, may lack interpretability and transparency, making it challenging to understand how and why certain cleaning decisions are made. Ensuring transparency and providing explanations for cleaning outcomes are essential for building trust and facilitating collaboration among stakeholders.

Addressing these challenges requires a combination of technical expertise, domain knowledge, and robust governance frameworks [18]. By carefully designing, validating, and monitoring automated cleaning pipelines, organizations can mitigate risks and harness the benefits of automated data pre-processing effectively.

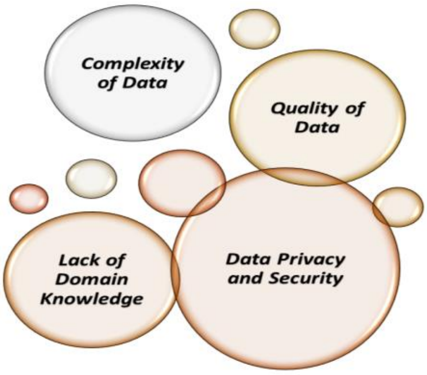

Fig.3 limitations

VII. FUTURE DIRECTIONS

The literature anticipates future developments in automated data cleaning and processing. Research avenues include the integration of machine learning for intelligent data cleaning, optimization of Python algorithms for large-scale datasets, and further exploration of cloud-native solutions [20]. Future directions in automated data cleaning and preprocessing are likely to focus on addressing current challenges and leveraging emerging technologies to enhance efficiency, scalability, and effectiveness. Some potential directions for future development include:

1) Integration of AI and Machine Learning:

There is growing interest in leveraging AI and machine learning techniques to automate more aspects of the data cleaning and pre-processing pipeline. Future developments may include the use of advanced algorithms for outlier detection, imputation, and pattern recognition, as well as the integration of reinforcement learning for adaptive cleaning strategies.

2) Semantic Understanding and Contextual Awareness:

Future automated cleaning systems may incorporate semantic understanding and contextual awareness to better interpret the meaning and context of data. This could involve the development of knowledge graphs, ontologies, or semantic models that capture domain-specific knowledge and relationships, enabling more intelligent and context-aware cleaning decisions.

3) Self-Learning and Adaptive Systems:

Automated cleaning systems may become more self-learning and adaptive over time, continuously improving their performance based on feedback and new data. This could involve the development of self-learning algorithms that adapt to changing data distributions, evolving data quality requirements, and feedback from users and domain experts.

4) Explain ability and Transparency:

There is increasing emphasis on making automated cleaning systems more transparent and interpretable, particularly as they become more complex and rely on advanced machine learning techniques. Future developments may focus on enhancing the explain ability of cleaning decisions, providing transparent audit trails, and enabling users to understand and trust the outcomes of automated cleaning processes.

5) Scalability and Distributed Processing:

As the volume, velocity, and variety of data continue to increase, future automated cleaning systems will need to be highly scalable and capable of distributed processing. This could involve the development of parallel and distributed algorithms, as well as the integration of cloud-native technologies for elastic scaling and resource management.

6) Privacy-Preserving Data Cleaning:

With growing concerns about data privacy and security, future automated cleaning systems will need to incorporate robust privacy-preserving techniques. This could include differential privacy mechanisms, federated learning approaches, or secure multiparty computation techniques that enable data cleaning to be performed without compromising sensitive information.

7. Automation of End-to-End Data Pipeline:

Future developments may focus on automating end-to-end data pipelines, from data ingestion and cleaning to analysis and decision-making. This could involve the integration of data cleaning tools with broader data management and analytics platforms, enabling seamless integration and automation of the entire data lifecycle.

Overall, future directions in automated data cleaning and preprocessing are likely to involve the convergence of AI, machine learning, visualization, semantics, and distributed computing technologies to develop more intelligent, adaptive, and scalable cleaning solutions that meet the evolving needs of data-driven organizations [14].

VIII. SCOPE OF THE PROJECT

All businesses and organizations need information that's clear and uncomplicated. Large volumes of records are entered into and regularly refreshed through multiple sources by data warehouses, increasing the likelihood when a few of those resources include tainted data. Cleaning of information is carried out because having accurate data is essential to preventing erroneous inferences. A crucial initial phase in every Work using data is data cleaning. Data cleansing is necessary for enhancing the outcomes of data extraction. Because everyday problems databases nowadays typically measure very large (several Gigabytes or greater) and probably originated from numerous heterogeneity resources, they can be extremely exposed to noisy, missing, and inconsistent data.

A lack of quality data leads to poor extraction output. When all organizations uses data, which can come through a variety of places, this presents a vast opportunity for data cleaning in order to obtain relevance and effective outcome data cleaning is essential.

IX. CONCLUSION:

In conclusion, automated data cleaning and preprocessing play a crucial role in ensuring the quality, reliability, and usability of data in various domains and industries. By leveraging advanced algorithms, machine learning techniques, and cloud-based technologies, organizations can streamline the process of preparing data for analysis, modeling, and decision-making.

Throughout this review paper, we have explored the foundations, applications, challenges, and future directions of automated data cleaning and preprocessing. We have discussed how automated techniques help address common data quality issues such as missing values, errors, duplicates, and inconsistencies, enabling organizations to derive actionable insights from their data more efficiently.

Furthermore, we have highlighted real-world applications and case studies demonstrating the practical benefits of automated data cleaning in domains such as finance, healthcare, e-commerce, manufacturing, telecommunications, and transportation. These case studies underscore the importance of automated cleaning techniques in improving data quality, operational efficiency, and decision-making across diverse industries.

Despite the significant advancements in automated data cleaning and preprocessing, challenges remain, including the complexity of data, quality assurance, computational resources, and interpretability. Addressing these challenges requires ongoing research, innovation, and collaboration among researchers, practitioners, and domain experts.

Looking ahead, future developments in automated data cleaning and preprocessing are likely to focus on integrating AI, machine learning, semantics, and distributed computing technologies to develop more intelligent, adaptive, and scalable cleaning solutions [16]. By embracing these advancements and best practices, organizations can unlock the full potential of their data assets and gain a competitive edge in today's data-driven world.

X. REFERENCES:

[1] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.

[2] D. J. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining," MIT Press, 2001.

[3] P. Christen, "Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection," Springer, 2012.

[4] W. H. Inmon, B. H. Terdeman, and R. Quatro, "DW 2.0: The Architecture for the Next Generation of Data Warehousing," Morgan Kaufmann, 2008.

[5] S. S. Chawathe, H. Garcia-Molina, J. Widom, "Automatic Database Cloning: Algorithms, Techniques, and Implementation," Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, and C. Shearer, "CRISP-DM 1.0 Step-by-step Data Mining Guide," CRISP-DM Consortium, 2000.

[7] D. Barbara, W. DuMouchel, O. Fersko-Weiss, and C. F. Olson, "Data Mining Methods for Knowledge Discovery," Kluwer Academic Publishers, 1996.

[8] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.

[9] M. Stonebraker and U. Çetintemel, "One Size Fits All? Part 2: Benchmarking Results," IEEE Data Engineering Bulletin, 2005.

[10] D. J. Hand, "Data Mining: Statistics and More?," The American Statistician, 1998.

[11] Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533.

[12] Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

[13] Janssen, F., & Zaharia, M. (2018). Efficient data cleaning for large datasets. arXiv preprint arXiv:1803.03453.

[14] Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2917-2926.

[15] S. Visalakshi and V. Radha, ‘‘A literature review of feature selection techniques and applications: Review of feature selection in data mining,’’ in Proc. IEEE Int. Conf. Comput. Intell. Comput. Res., Dec. 2014, pp. 1–6.

[16] Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. Science, 343(6176), 1203-1205.

[17] S. Krishnan and E. Wu, ‘‘Alphaclean: Automatic generation of data cleaning pipelines,’’ 2019, arXiv:1904.11827.

[18] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

[19] E. Bisong, ‘‘Google AutoML: Cloud vision,’’ in Building Machine Learning and Deep Learning Models on Google Cloud Platform. New York, NY, USA: Springer, 2019, pp. 581–598.

[20] Winkler, W. E. (1999). The state of record linkage and current research problems. Statistical Research Division, US Census Bureau, Washington, DC.