UKLO Proceedings: A Multidisciplinary Approach to Sustainable Development Goals • Vol. 2, No. 1, 2026
Published by the University "St. Kliment Ohridski" – Bitola

(*) Corresponding author
Original research article
DOI: https://doi.org/10.20544/uklop.2026.611
Received: 25.11.2025 • Accepted after revision: 30.01.2026 • Published: 07.05.2026

#UKLO
Annual **International Scientific Conference**

SNAPLOGIC AUTOMATION AS A CATALYST FOR ETL AND MACHINE LEARNING WORKFLOWS

Aneta Trajkovska*
University "St. Kliment Ohridski" - Bitola, Faculty of Information and Communication Technologies, Republic of North Macedonia
aneta.trajkovska@uklo.edu.mk

Kostandina Veljanovska
University "St. Kliment Ohridski" - Bitola, Faculty of Information and Communication Technologies, Republic of North Macedonia
kostandina.veljanovska@uklo.edu.mk

ABSTRACT

Introduction: *This study explores the use of SnapLogic automation to modernize Extract, Transform, and Load (ETL) processes and support scalable machine learning (ML) workflows. Traditional ETL pipelines are slowed by manual intervention, inefficiency, and limited scalability.*
Methods: *Our methodology followed an applied research design, focusing on the practical implementation of SnapLogic automation for ETL and machine learning workflows. Automated pipelines were built to handle data preparation, model deployment, and performance monitoring, and were tested with both synthetic and real-world datasets.*
Results: *SnapLogic automation significantly improved efficiency, reduced development time, minimized manual errors, and enabled real-time processing of large, dynamic workloads. Continuous monitoring and automated lifecycle management enhanced system reliability and resilience.*
Discussion*: These findings demonstrate that SnapLogic directly addresses the limitations of traditional ETL by streamlining workflows and bridging the gap between data engineering and ML. While the platform may involve a higher upfront investment, it optimizes long-term costs for large organizations.*

Keywords: Automation, ETL, SnapLogic, Machine Learning, Workflows, Enterprise, AI

1. INTRODUCTION

Addressing modern sustainability challenges requires organizations to adopt data-driven, automated, and collaborative approaches that connect expertise from multiple disciplines. Building resilient infrastructure, fostering innovation, advancing economic productivity, reducing inequality, and protecting the environment all depend on the ability to gather, integrate, and analyze large volumes of diverse data. As environmental, social, and economic systems become more complex and interdependent, efficient data engineering, advanced analytics, and intelligent automation are becoming essential components of sustainable decision-making.

In this broader sustainability landscape, automation platforms such as SnapLogic provide significant value. By streamlining Extract–Transform–Load (ETL) workflows and simplifying the orchestration of machine learning processes, SnapLogic enables organizations to manage complex datasets with greater accuracy, speed, and transparency, (Gour et al. 2010, 786-789) . These capabilities support sustainability efforts such as real-time monitoring of key indicators, predictive modeling for more efficient use of resources, and automated reporting that strengthens accountability and long-term planning.

Sustainable development also requires strong collaboration across fields including environmental science, engineering, public administration, business analytics, and information technologies, (Chalmers, 2003). SnapLogic and the low-code approach lowers technical barriers and supports more inclusive participation, making it easier for multidisciplinary teams to contribute to data-driven sustainability initiatives. Its cloud-ready and scalable architecture further promotes responsible digital transformation and efficient use of technological resources.

Automation of ETL and machine learning workflows is not only a technical improvement but also a strategic enabler for sustainability-oriented digital ecosystems, (Bomma et. al. 2024). SnapLogic capacity to unify data, accelerate analysis, and support collaboration across diverse domains positions it as a catalyst for multidisciplinary efforts toward more intelligent, resilient, and sustainable development. This paper examines how SnapLogic automation strengthens these efforts by improving data accessibility, enhancing analytical capabilities, and supporting innovative solutions across sectors.

2. LITERATURE REVIEW

Research in this area documents the transition from traditional ETL solutions toward more flexible integration platforms capable of supporting real-time data movement, distributed processing, and cloud-based analytics. Highlighting the key benefits such as reduced operational complexity, faster data availability, and improved monitoring capabilities, (Mishra,2020, 69-78).

In sustainability-focused research, achieving long-term social, environmental, and economic goals increasingly depends on high-quality, integrated datasets and reliable analytics pipelines. Studies point to the importance of interdisciplinary collaboration and accessible digital tools that enable data modeling, prediction, and monitoring across domains such as environmental management, resource optimization, and urban systems (Ravi 2025, 52-55), (Kumar et al. 2022, 1292-1297).

Our previous research (Trajkovska et. al. 2023, 212-219), focused primarily on automation and monitoring of ETL processes while distributing data, addressing operational efficiency and integration reliability. This established the foundation for understanding how automated pipelines support large-scale data flows. The current study builds directly on that foundation by extending the focus beyond automation alone. It now explores how SnapLogic ETL automation intersects with machine learning capabilities and how this combined functionality can contribute to broader sustainability-oriented initiatives. Thus, the present research situates ETL automation not only as a technical process but as part of a multidisciplinary digital ecosystem that supports sustainable development.

3. THE CHALLENGE OF UTILIZING INTEGRATION PROCESSES

The effective utilization of integration processes represents a persistent challenge within modern information systems. As organizations increasingly rely on heterogeneous data sources, distributed architectures, and complex workflows, ensuring seamless interoperability becomes essential yet difficult to achieve. Integration processes must address issues such as data consistency, system compatibility, real-time synchronization, and secure communication across platforms (Zeng, et. al. 2010). Furthermore, the dynamic nature of business requirements demands adaptable and scalable integration strategies capable of evolving alongside technological advancements, (Maréchal, et. al. 1998). These challenges highlight the need for robust methodologies, advanced automation tools, and well-designed governance frameworks to optimize integration efforts and support reliable, end-to-end operational performance.

Within enterprise integration environments, SnapLogic offers a powerful platform for automating data flows, orchestrating services, and enabling seamless connectivity across disparate systems. However, effectively utilizing integration processes in SnapLogic presents several challenges that require careful consideration. The platform's flexibility demands a thorough understanding of pipeline design principles, data transformation patterns, and integration best practices to ensure optimal performance, (Trajkovska et. al. 2023, 212-219). Additionally, managing complex workflows-particularly those involving real-time processing or large-scale data movement-requires robust error handling, monitoring, and governance mechanisms. As organizations scale their digital ecosystems, the ability to design maintainable, efficient, and resilient SnapLogic pipelines becomes crucial. These challenges underscore the importance of strategic planning, technical proficiency, and continuous optimization to fully leverage SnapLogic capabilities in modern integration architectures (Hussain et. al. 2024).

4. METHODS

This study employs a qualitative, exploratory research design aimed at evaluating how SnapLogic automation capabilities support integrated ETL and machine learning workflows within contexts relevant to sustainable development. The design combines conceptual analysis, platform-based workflow demonstration, and comparative evaluation of existing literature.

- **Study Design**

Our methodology was focused on the practical implementation of SnapLogic automation for ETL and machine learning workflows. Automated pipelines were built to handle data preparation, model deployment, and performance monitoring, and were tested with both synthetic and real-world datasets. A multi-method approach was applied to ensure that the study captures both the technical characteristics of SnapLogic and the broader implications for sustainable and collaborative practices, such as sustainability framework mapping – alignment of identified platform functionalities with sustainability-related objectives such as resource optimization, real-time monitoring, data democratization, and multidisciplinary cooperation.

We have worked with dataset that encompassed detailed information related to vehicles as well as associated organizational attributes. The data included variables such as vehicle specifications, pricing structures, discount parameters, model colors, and identification fields, alongside organizational metadata such as company identifiers, employee salaries. This combination of technical and business-level information provided a suitable foundation for constructing and evaluating automated ETL workflows. It also enabled the testing of machine learning capabilities within SnapLogic, as the dataset allowed for diverse transformation tasks, feature engineering processes, and predictive modeling scenarios relevant to real-world operational contexts (Patel,2023,1-5).

- **Data Integration**

Dataset utilized was generated by API trigger for the implementation of the solution. During the preparation of the examples, we have divided the data sets in two parts :

- **Synthetic datasets**, generated to test pipeline behavior under controlled conditions, including volume scaling and error-handling scenarios.
- **Real-world datasets,** selected to validate the practical applicability and performance of the designed workflows in realistic ETL and ML settings.

ETL is a fundamental data integration process used to collect data from multiple sources, convert it into a suitable format, and load it into a target system for analysis or storage, Fig.1. The extraction phase involves retrieving raw data from heterogeneous sources such as databases, APIs, or flat files. During the transformation phase, data is cleaned, normalized, and structured to ensure consistency, quality, and analytical readiness. Finally, in the loading phase, the processed data is delivered to a data warehouse, data lake, or other storage systems for reporting, analytics, or machine learning applications. ETL processes are critical for enabling accurate, timely, and reliable data-driven decision-making (Khan et. al. 2024).

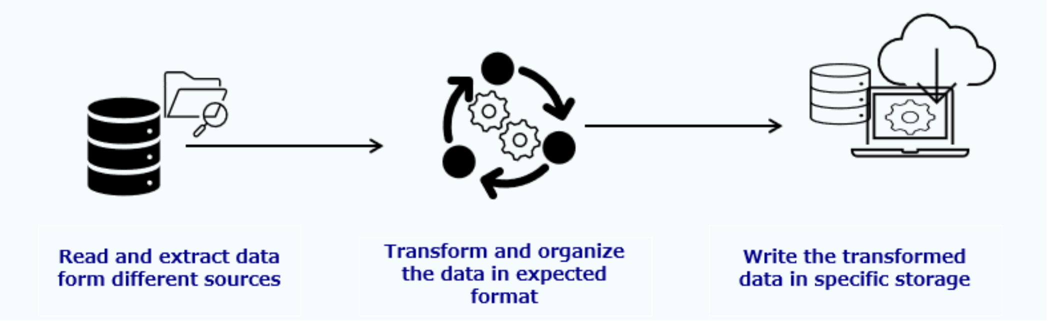

Fig 1. ETL phases overview (Source: Own depiction)

- **Materials and Instruments**

Preparation of the solution relied on the SnapLogic Intelligent Integration Platform, which served as the primary environment for designing, executing, and monitoring automated workflows. The SnapLogic Pipeline Designer was utilized to construct ETL and machine learning pipelines, enabling seamless orchestration of data flows and automation tasks, (Kommera, Reddy,2015). Various data preparation Snaps were integrated to support cleaning, transformation, and feature engineering operations required for model readiness. Machine Learning Snaps were employed for training, scoring, and deploying predictive models within the same unified environment. Additionally, SnapLogic monitoring and dashboard capabilities were used to track execution performance, identify bottlenecks, and evaluate workflow reliability. Complementary tools, including Python-based evaluation scripts and standard performance metrics, were incorporated when needed to validate the outputs of the automated pipelines.

- **Procedure**

The procedures applied in this study were designed to ensure the reliability, accuracy, and consistency of the automated ETL and machine learning workflows developed in SnapLogic. A key component of the procedure involved the implementation of validation and verification rules, which were applied immediately after data ingestion to guarantee that only high-quality and structurally sound data progressed through the pipeline (Clemens-Meyer et. al. 2021). These rules included schema conformity checks, assessments for missing or incomplete values, verification of identifier consistency, and evaluations of numerical fields to ensure they fell within acceptable operational ranges. By enforcing these rules early in the process, the system was able to automatically detect irregularities and either correct them or isolate problematic records for further inspection.

Data cleaning represented a second critical aspect of the procedure, given the mixed structure of the vehicle and organizational dataset. Cleaning operations were essential for improving the accuracy of transformations and predictive models (Dasu et. al. 2003). Tasks such as standardizing formats, handling missing values, resolving duplicate entries, and normalizing numerical fields ensured that the dataset was coherent and analytically robust. These cleaning steps reduced noise in the data and prevented errors in later stages of the workflow, demonstrating the importance of strong preprocessing practices within automated integration environments.

- **Data Analysis**

Analytical methods included evaluation of model performance (accuracy, precision, error rates) and workflow efficiency (execution time, resource consumption, failure rates). Comparative analyses were conducted between synthetic and real-world executions to determine robustness and generalizability of the automation approach. Qualitative assessment of pipeline maintainability and usability was also incorporated.

5. RESULTS

The implementation and testing of automated ETL and machine learning workflows in SnapLogic demonstrated several key outcomes in terms of efficiency, data quality, and workflow reliability. The pipelines successfully ingested, validated, and transformed both synthetic and real-world datasets containing vehicle and organizational information. Validation and verification rules effectively identified inconsistencies, missing values, and outliers, ensuring that only high-quality data progressed to subsequent transformation and modeling stages. This resulted in clean, standardized, and analytically robust datasets that supported accurate machine learning operations (Côté et. al. 2024).

price	discount	company	model	final_price
4704	1.05	Smith-Walter	Getz	4480
1098	1.05	Moore-Hagenes	Fusion	1046
30	1.05	Stiedemann, Jacob...	Note	29
21395	1.05	Feeney Group	Elantra	20376
1098	1.05	Harvey, Parisian an...	Sportage	1046
25089	1.05	Halvorson, Fay and ...	X5 E70	23894
5331	1.05	Mertz, Abshire and ...	C 180	5077
8154	1.05	Kreiger LLC	Vectra B	7766

Fig 2. Dataset examples (Source: Own depiction)

Automated data preparation tasks including normalization, deduplication, and feature engineering, were executed seamlessly within the SnapLogic platform, reducing manual intervention and minimizing the potential for human error, Fig. 2.

The integration of machine learning Snaps enabled end-to-end automation of model training, deployment, and scoring, providing real-time insights from the processed data, Fig. 3. Performance monitoring and logging features allowed continuous observation of pipeline execution, facilitating rapid identification of bottlenecks and system inefficiencies.

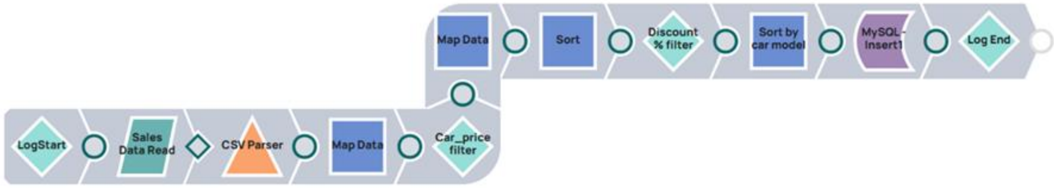

Fig 3. Automation of ETL process (Source: Own depiction)

Comparisons between synthetic and real-world datasets highlighted the scalability and adaptability of SnapLogic workflows. Pipelines maintained stable execution even under varying data volumes and complexity, demonstrating robustness in handling heterogeneous datasets. Furthermore, the low-code environment facilitated workflow adjustments and multidisciplinary collaboration, allowing non-technical users to engage in data preparation and analysis processes effectively.

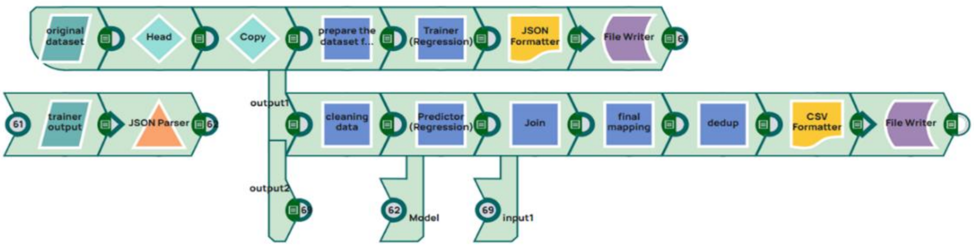

Fig 4. Machine Learning Workflow in SnapLogic (Source: Own depiction)

The machine learning component of the workflow employed a Random Forest Regression model to generate predictive insights based on the numerical and categorical attributes contained in the dataset (Fig. 4). After the data was cleaned, validated, and transformed, the selected features-such as vehicle model, company characteristics, and price-related variables were used to train the regression algorithm. Random Forest was selected due to its ability to capture nonlinear relationships, handle complex feature interactions, and provide higher predictive stability compared to single-model approaches. Once trained, the model produced predicted price values for each record, demonstrating how SnapLogic integrated machine learning capabilities can operationalize advanced regression techniques within automated ETL workflows, Fig.5 .

```
[
  {
    "readable": "RandomForest\n\nBagging with 100 iterations and base learner\n\nweka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -depth 15 -do-not-check-capabilities",
    "metadata": "rO0ABXNyABN3ZWthLmNvcmUuSW5zdGFuY2Vz/7slkzRlqaQCAAZJAAxtX0NsYXNzSW5kZXhJAAdtX0xpbmVzTAAMbV9BdHRyaWJ1dGVzdAAVTGphdmEvdXRpbC9BcnJheUxpc3Q7TAALbV9Jb
    "model": "rO0ABXNyACN3ZWthLmNsYXNzaWZpZXJzLnRyZWVzLlJhbmRvbUZvcmVzdA9/z5TVRjBaAgABWgAcbV9jb21wdXRlQXR0cmlidXRlSW1wb3J0YW5jZXhyAB13ZWthLmNsYXNzaWZpZXJzLm1ldGEuQm
  }
]
```

Fig 5. Model description in SnapLogic (Source: Own depiction)

Additionally, it was applied to the dataset to compute predicted prices for each vehicle entry. The table illustrates a comparison between the actual price values and the predicted_price generated by the regression model, shown below in Fig. 5. These predictions reflect the model's attempt to estimate vehicle pricing patterns based on learned relationships within the data. The output confirms that SnapLogic integrated machine learning capabilities can execute end-to-end prediction tasks directly within automated ETL workflows, enabling real-time scoring and analytical insight generation.

Preview Type
Table

company_name	model	price	predicted_price
Adams, Dibbert and...	Sonata	1792	2755.0464
Beer Inc	Sonata	1792	4061.482933333333
Hoppe, Bins and Go...	Santa FE	46378	40850.373599999...
Leffler-Monahan	FIT	7766	8732.311666666666
Kuphal-Leffler	Avalanche	5227	6928.725199999999
Smitham Group	X5 E70	23894	24059.5454
Cartwright LLC	CX-5	49282	38137.3549999999...
Mertz, Abshire and ...	Caddy	7467	7951.2069999999...
Hintz, Spinka and C...	Optima	597	1478.61979999999...
Grimes-Morissette	Cruze	14336	14904.0538000000...
Leffler-Monahan	Kizashi	7467	7875.0780666666...
Schaden Group	Fusion	821	2055.4871714285714

Fig 6. Predicted price based on the data (Source: Own depiction)

The results indicate that SnapLogic serves as a practical and effective platform for combining ETL automation with machine learning orchestration. The integration of these capabilities within a single platform not only enhances operational efficiency but also provides a scalable foundation for supporting sustainability-oriented analytics, enabling organizations to derive actionable insights and implement data-driven strategies with minimal overhead.

6. DISCUSSION

The findings from the implementation of automated ETL and machine learning workflows in SnapLogic demonstrate the platform's capacity to enhance data processing efficiency, improve workflow reliability, and support multidisciplinary collaboration. The successful ingestion, validation, and transformation of both synthetic and real-world datasets highlight the effectiveness of integrated validation and data cleaning procedures, confirming the importance of these steps in maintaining high-quality data for analytical and predictive tasks.

The results indicate that SnapLogic low-code environment facilitates collaboration across technical and non-technical team members, allowing domain experts to participate in data preparation and model deployment. This feature aligns closely with the goals of multidisciplinary projects, where effective collaboration between data engineers, analysts, and domain specialists is essential for deriving actionable insights. Furthermore, the automated orchestration of machine learning workflows within the same platform demonstrates the potential for end-to-end pipeline efficiency, reducing manual intervention and minimizing the risk of errors while maintaining consistent performance across datasets of varying size and complexity.

From a sustainability perspective, the integration of ETL automation with machine learning capabilities presents opportunities for more informed and responsive decision-making. Automated pipelines enable real-time monitoring, predictive modeling, and resource optimization, which are critical components for organizations pursuing data-driven sustainability initiatives. By reducing time and effort required for data processing and model execution, SnapLogic contributes to operational efficiency while enabling the rapid application of insights in sustainability-focused domains, such as environmental monitoring, energy management, or supply chain optimization, Fig.7. The Monitor interface allows users to track pipeline performance metrics such as execution duration, throughput, memory consumption, and data processing volumes. It also enables real-time observation of task status, including successful executions, warnings, and failure events. Additionally, users can monitor Groundplex health, resource utilization, and node availability to ensure stable operational performance. Detailed logs and execution histories support troubleshooting, auditing, and optimization of workflows, making monitoring a critical component for maintaining reliable and efficient automation within the SnapLogic platform.

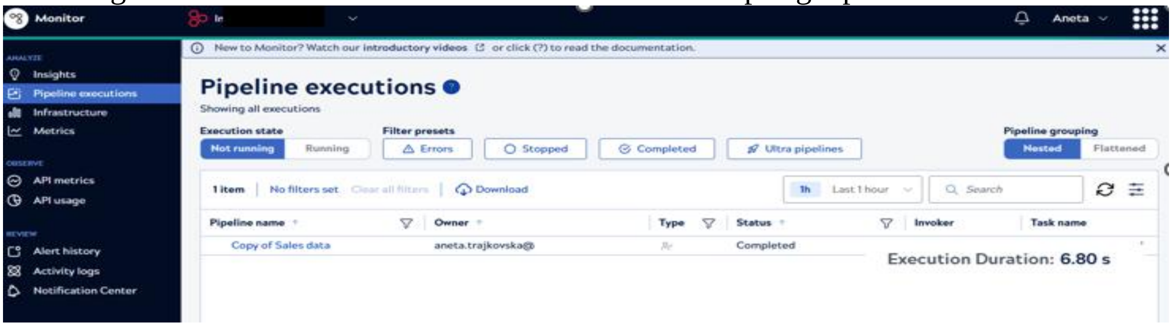

Fig 7. Monitoring workflows execution in SnapLogic (Source: Own depiction)

Despite the demonstrated benefits, several limitations were identified during the study. SnapLogic machine learning capabilities, while efficient for lightweight and moderately sized models, are not ideal for large-scale model training that requires GPU acceleration or distributed cluster compute, which may restrict its applicability in highly complex analytical environments. Pipeline performance was also found to be dependent on the configuration and capacity of the Groundplex infrastructure, meaning that execution speed and stability can vary significantly based on organizational setup, available compute resources, and parallel processing configurations. Furthermore, the full functionality of ETL and machine learning workflows is dependent on specific SnapPack licensing, which may limit access to advanced features if the required components are not purchased. These constraints highlight the need for careful planning, appropriate infrastructure provisioning, and licensing considerations when deploying SnapLogic for enterprise-level automation and analytics. This makes the platform more accessible and cost-effective for large enterprises with long-term integration strategies, while smaller organizations may struggle to justify the upfront expenditure.

Discussion emphasizes that SnapLogic serves not only as a technical tool for ETL and machine learning automation but also as a strategic enabler for multidisciplinary, sustainability-oriented projects. The platform combination of automation, machine learning integration, and user-friendly interface allows organizations to streamline workflows, foster collaboration, and implement data-driven strategies effectively. These findings build directly on prior research focused solely on ETL automation, expanding the scope to include machine learning orchestration and the broader implications for sustainable development initiatives.

7. CONCLUSION

This study examined the application of SnapLogic automation as a catalyst for ETL and machine learning workflows, with a focus on its potential to support multidisciplinary and sustainability-oriented initiatives. The findings demonstrate that SnapLogic enables seamless integration, data preparation, and model orchestration within a single platform, reducing manual effort, enhancing workflow reliability, and improving overall operational efficiency. Validation and data cleaning procedures were shown to be critical in maintaining data quality, while automated ML pipelines provided reproducible, scalable, and actionable insights from both synthetic and real-world datasets.

The research highlights the platform's capacity to facilitate collaboration between technical and non-technical team members, making it a valuable tool for multidisciplinary projects where decision-making is essential. Furthermore, the integration of automation with machine learning offers tangible benefits for sustainability-focused applications, including real-time monitoring, predictive modeling, and optimized resource management. These capabilities position SnapLogic not only as a technical solution but also as a strategic enabler for organizations aiming to implement intelligent, sustainable, and data-driven processes.

Building on prior work focused solely on ETL automation, this study extends the understanding of SnapLogic role in orchestrating machine learning workflows and supporting sustainability initiatives. Future research could explore the performance of such pipelines in large-scale enterprise environments, investigate additional use cases in environmental and social domains, and develop standardized metrics for evaluating the platform's contribution to sustainable development goals. Also, extending the analytical framework to enhance its scalability and integration of advanced ML and predictive analytics.

REFERENCES

1. Gour, Vishal, S. S. Sarangdevot, Govind Singh Tanwar, and Anand Sharma. "Improve performance of extract, transform and load (ETL) in data warehouse." Int. Journal on Comp. Sci. and Eng 2, no. 3 (2010): 786-789.
2. Chalmers, Neil, Stephen Gough, and William Scott. Sustainable development and learning: Framing the issues. Routledge, 2003.
3. Bomma, Hari Prasad. "Revolutionizing ETL with AI Powered Automation." IJLRP-International Journal of Leading Research Publication 5, no. 2 (2024).
4. Mishra, Sarbaree. "Automating the data integration and ETL pipelines through machine learning to handle massive datasets in the enterprise." International Journal of Emerging Research in Engineering and Technology 1, no. 2 (2020): 69-78.
5. Ravi, Chandran. "ETL (Extract, Transform & Load) Automation." International Journal of Emerging Trends in Computer Science and Information Technology 6, no. 1 (2025): 52-55.
6. Kumar, G. Sunil Santhosh, and M. Rudra Kumar. "Dimensions of automated etl management: A contemporary literature review." In 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 1292-1297. IEEE, 2022.
7. Trajkovska, Aneta, Tome Dimovski, Ramona Markoska, and Zoran Kotevski. "Automation and Monitoring on Integration ETL Processes while Distributing Data." (2023): 212-219.
8. Zheng, Mei, Jianhua Ren, Linlin Sun, and Hui Ding. "The Process of Integration." In 2010 International Conference on E-Product E-Service and E-Entertainment, pp. 1-5. IEEE, 2010.
9. Maréchal, François, and Boris Kalitventzeff. "Process integration: Selection of the optimal utility system." Computers & Chemical Engineering 22 (1998): S149-S156.
10. Hussain, Sadath, and Mohan Aradhya. "Big Data Integration & Transformation: A Comparative Analysis of SnapLogic and AWS Glue." In 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), pp. 1-6. IEEE, 2024.
11. Khan, Bilal, Saifullah Jan, Wahab Khan, and Muhammad Imran Chughtai. "An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing." Journal on Big Data 6 (2024).
12. Kommera, Adisheshu Reddy. "Future of enterprise integrations and iPaaS (Integration Platform as a Service) adoption." Neuroquantology 13, no. 1 (2015): 176-186.
13. Clemens-Meyer, Francois HLR, Mathieu Lepot, Frank Blumensaat, Dominik Leutnant, and Guenter Gruber. "Data validation and data quality assessment."

Metrology in Urban Drainage and Stormwater Management: Plug and Pray, edited by: Bertrand-Krajewski, J.-L., Clemens-Meyer, F., and Lepot, M., IWA Publishing (2021): 327-390.

14. Dasu, Tamraparni, and Theodore Johnson. Exploratory data mining and data cleaning. John Wiley & Sons, 2003.
15. Côté, Pierre-Olivier, Amin Nikanjam, Nafisa Ahmed, Dmytro Humeniuk, and Foutse Khomh. "Data cleaning and machine learning: a systematic literature review." Automated Software Engineering 31, no. 2 (2024): 54.
16. Patel, Paraskumar. "Challenges and solutions in cross-platform ML systems integration." Journal of Artificial Intelligence & Cloud Computing 2, no. 4 (2023): 1-5.