International Journal of Multidisciplinary Research and Growth Evaluation.

# Developing an Automated ETL Pipeline Model for Enhanced Data Quality and Governance in Analytics

**Kolade Olusola Ogunsola [1*], Emmanuel Damilare Balogun [2], Adebanji Samuel Ogunmokun [3]**
[1, 2] Independent Researcher; USA
[3] United Utilities, UK

* Corresponding Author: **Kolade Olusola Ogunsola**

## Article Info

**ISSN (online):** 2582-7138
**Volume:** 03
**Issue:** 01
**January-February 2022**
**Received:** 20-12-2021
**Accepted:** 15-01-2022
**Page No:** 791-796

**Abstract**
This paper presents the development of an automated Extract, Transform, Load (ETL) pipeline model aimed at enhancing data quality and governance for improved business intelligence and decision-making. The model integrates advanced data processing technologies, ensuring high-quality data extraction, consistent transformation, and seamless loading into data warehouses. The pipeline model enables organizations to achieve reliable analytics outcomes and optimized decision-making by embedding automated quality checks and governance measures. The study explores key components of the ETL pipeline, discusses the technologies utilized, and provides case studies illustrating the model's application in real-world scenarios. Furthermore, the paper addresses the challenges associated with adopting automated ETL systems, such as data integration complexity and resistance to change, offering recommendations for organizations seeking to leverage this technology for improved business performance. Finally, the study identifies avenues for future research, including the integration of artificial intelligence into ETL processes and the broader application of the model across various industries.

**DOI:** https://doi.org/10.54660/.IJMRGE.2022.3.1.791-796

**Keywords:** Automated ETL Pipeline, Data Quality, Data Governance, Business Intelligence, Data Transformation, Cloud-based Analytics

## 1. Introduction

In the digital age, organizations across industries are relying heavily on data-driven insights to inform decision-making and drive business success. One of the most critical components of data management in modern analytics is the Extract, Transform, and Load (ETL) pipeline (Nylén & Holmström, 2015). This process enables businesses to collect raw data from various sources, transform it into a usable format, and load it into a central data repository for analysis. ETL pipelines are indispensable for facilitating real-time and accurate analytics, making them a core part of any data strategy (Thumburu, 2020). The significance of ETL pipelines has grown with the exponential increase in data generation and the need for businesses to stay agile in a competitive market. The ability to quickly analyze high-quality, integrated data allows organizations to make informed decisions that can optimize operations and improve outcomes (Raj, Bosch, Olsson, & Wang, 2020).

However, as organizations scale their operations, managing vast amounts of data becomes increasingly complex. Data governance—ensuring that data is accurate, complete, and consistent—becomes more difficult when traditional ETL processes are used. These processes are often manual, which increases the potential for human error, reduces efficiency, and leads to delays in data integration (Thumburu, 2020). Furthermore, traditional ETL models struggle to keep pace with the increasing volumes and complexity of modern data flows, which impacts the reliability of business insights. This is where automated ETL pipelines come into play. By leveraging automation, organizations can streamline data processing, improve accuracy, and enhance governance. Automated ETL pipelines reduce manual intervention, minimize errors, and can scale seamlessly to handle large datasets. This transition is crucial for businesses seeking to enhance their analytics capabilities and ensure timely, data-driven making (Singu, 2022).

### 1.1 Problem Statement

Traditional ETL processes, although foundational in data integration, have several inherent challenges that limit their effectiveness in modern data-driven environments. One significant issue is the reliance on manual interventions, which introduce the potential for human error during data extraction, transformation, or loading. Human errors can lead to inaccurate, inconsistent, or incomplete data, which undermines the reliability of subsequent analytics and decision-making (Ikegwu, Nweke, Anikwe, Alo, & Okonkwo, 2022). Moreover, traditional ETL processes often struggle to handle the increasing scale and complexity of modern data sources, resulting in delays in data processing, inefficiencies, and bottlenecks. As businesses continue to expand, these limitations in traditional ETL models hinder their ability to make timely, data-driven decisions (Russom, 2011).

Another key challenge is the impact of poor data quality on analytics outcomes. Inadequate data governance can lead to inconsistencies, inaccuracies, and incomplete datasets that prevent organizations from deriving actionable insights (Rachakatla, Ravichandran, & Machireddy, 2021). Without robust data governance policies and automated checks, organizations risk making decisions based on flawed or unreliable data, which can have serious consequences for operational efficiency and business performance. The need for better data quality and governance has never been more pressing. Organizations need ETL systems that ensure data integrity, security, and accessibility, enabling real-time analysis and reliable insights (Gade, 2021).

The adoption of automated ETL pipelines offers a solution to these challenges. Automation can reduce human error, improve data processing speeds, and ensure data consistency and accuracy. Organizations can implement standardized, scalable, and more efficient data workflows by automating the extraction, transformation, and loading of data. Automated ETL pipelines optimize data quality and support data governance by incorporating automated checks, balances, and validation protocols, ensuring that only high-quality data is fed into the analytics process (Dhayne, Haque, Kilany, & Taher, 2019).

### 1.2 Objectives of the study

The primary objective of this study is to design and develop an automated ETL pipeline model that enhances data quality and governance in analytics. This model will aim to address the limitations of traditional ETL processes by introducing automation that improves data accuracy, scalability, and efficiency. The study will focus on how automation can streamline the ETL process while maintaining the integrity of the data, ultimately leading to more reliable business insights.

Secondary objectives of this study include:

- To assess the key components and technologies involved in an automated ETL pipeline:
- To evaluate the impact of this model on data governance and analytics performance:
- To identify challenges in the adoption and implementation of automated ETL pipelines:

By achieving these objectives, this study aims to provide businesses with a comprehensive framework for developing and implementing automated ETL pipelines that optimize data quality, governance, and analytics capabilities.

## 2. Literature Review

### 2.1 ETL pipelines in data analytics

The ETL (Extract, Transform, Load) process is the backbone of data integration in modern analytics. Traditionally, ETL pipelines were designed to extract data from various sources, transform it into a format suitable for analysis, and load it into data warehouses or databases for further analysis (Ajayi & Akerele, 2021). The process plays a crucial role in ensuring that businesses can make sense of large, disparate datasets. Early-stage ETL systems were often manual and rigid, requiring human intervention to handle data extraction, transformation, and integration. Over time, this led to inefficiencies, errors, and scalability challenges, particularly as data volumes and complexity grew exponentially (Adewoyin, 2021; Olbert & Spengel, 2017).

In the context of data warehousing and analytics, ETL pipelines are essential for consolidating data from various sources, ensuring that the data is clean, consistent, and ready for analysis. They serve as the critical link between raw data and actionable insights, feeding business intelligence tools and analytical systems with the required datasets. ETL processes have evolved from manual batch processes to more automated, real-time integrations, particularly with the rise of cloud computing and big data technologies. Modern ETL practices incorporate incremental updates, enabling more timely and accurate data flows (Elumilade, Ogundeji, Achumie, Omokhoa, & Omowole, 2021; Hassan, Collins, Babatunde, Alabi, & Mustapha, 2021).

However, traditional ETL processes have inherent challenges, including the need for constant manual intervention, slow data integration speeds, and difficulties in maintaining data consistency across multiple sources. These manual processes are also prone to human errors, which compromise data quality and analytics outcomes. As businesses scale and require faster data processing, these limitations make traditional ETL systems insufficient for the growing demands of modern analytics and business decision-making. Therefore, there has been a significant push toward automating ETL pipelines to overcome these issues and optimize data management (Odio *et al*, 2021).

### 2.2 Automating ETL pipelines

Automating ETL pipelines has gained considerable attention as a solution to the challenges posed by traditional manual systems. A variety of technologies and tools have been developed to automate the process of extracting, transforming, and loading data in real time, making it possible to streamline workflows, enhance data processing speeds, and reduce human errors. Notable tools include Apache Nifi, Talend, and Apache Airflow, all of which provide a range of functionalities to manage and automate complex data integration tasks. These platforms enable organizations to design, schedule, and monitor ETL workflows, significantly improving operational efficiency (Otokiti, Igwe, Ewim, & Ibeh, 2021; Paul, Abbey, Onukwulu, Agho, & Louis, 2021).

Apache Nifi, for example, offers a user-friendly interface for building data pipelines that support real-time data flow, enabling automated data ingestion and transformation without manual intervention. Talend provides a suite of tools for data integration, which includes cloud-based and on-premise solutions for automating ETL processes. Apache Airflow, an open-source orchestration tool, helps automate ETL workflows by scheduling tasks and managing dependencies between different parts of the data pipeline.

These tools also ensure scalability and flexibility, crucial for businesses dealing with vast amounts of constantly changing data (Abisoye & Akerele, 2022a; Paul *et al*, 2021).

Automating ETL pipelines brings numerous benefits, such as enhanced scalability, real-time data processing, and the ability to handle complex workflows. Automated systems can process large datasets in real-time, ensuring that the data is always up to date and ready for analysis. Moreover, automation helps reduce human errors, improve data consistency, and increase the overall efficiency of data management (Singu, 2022). However, the adoption of automated ETL pipelines is not without challenges. Scalability can become an issue when dealing with vast datasets, especially in industries like healthcare or finance, with enormous data volumes. Additionally, organizations may face difficulties in integrating these automated pipelines with existing systems, which can hinder the seamless flow of data across various platforms (Abisoye & Akerele, 2022b; Achumie, Oyegbade, Igwe, Ofodile, & Azubuike, 2022).

### 2.3 Data quality and governance

Data quality and governance are fundamental to ensuring business analytics' accuracy, consistency, and reliability. Data quality refers to the attributes that make data useful for analysis, such as accuracy, completeness, consistency, timeliness, and reliability (Elumilade, Ogundeji, Achumie, Omokhoa, & Omowole, 2022). High-quality data ensures that the insights generated from analytics are dependable and actionable. Poor data quality can lead to misleading results, which could negatively impact business decisions. Data quality dimensions are critical for businesses, as even small errors can have far-reaching consequences in strategic planning, marketing, and operational efficiency (Mustapha & Ibitoye, 2022).

On the other hand, data governance involves the processes, policies, and standards put in place to manage data assets within an organization. It includes ensuring data security, privacy, compliance with regulations, and data integrity. Effective data governance also establishes accountability and clear ownership, ensuring that data is managed consistently and securely across different departments. Governance structures often involve data stewards, roles responsible for maintaining data standards and ensuring data quality (Adaralegbe *et al*, 2022).

The role of automated ETL pipelines in improving data governance and quality is significant. Automation ensures that data is consistently processed, validated, and transformed according to predefined rules and standards, reducing the risk of errors and inconsistencies. Automated ETL pipelines can also integrate validation steps, such as data cleansing and format standardization, to ensure that only high-quality data is loaded into the data warehouse or analytical platforms (Ajayi & Akerele, 2022b). This reduces the need for manual interventions, allowing organizations to maintain strict data governance protocols while improving operational efficiency. Furthermore, automation ensures that the data flows continuously and in real-time, which is crucial for maintaining timely access to accurate information for decision-making. By improving data quality and governance, automated ETL pipelines play a critical role in enhancing the overall effectiveness of business analytics (Adewoyin, 2022; Ajayi & Akerele, 2022a).

## 3. Methodology

### 3.1 Research Design

This study's research design will be qualitative, primarily based on the analysis of existing reports related to the implementation and performance of automated ETL pipelines in various industries. This qualitative approach is appropriate for understanding the broader implications of automated ETL pipelines on data quality, governance, and analytics processes by synthesizing the findings and experiences reported in industry and academic studies. Reports will be sourced from a range of relevant sources, such as technology whitepapers, case studies, industry reports, and best practices documentation. This approach will allow the researcher to explore how organizations are utilizing automated ETL pipelines, the benefits they have realized, and the challenges they have encountered.

The research will involve a detailed review of reports published by ETL tool vendors, consulting firms, industry analysts, and academic publications on data management systems. These reports will serve as secondary data sources, offering insights into the technologies used, the outcomes of automation, and the adoption challenges associated with ETL pipelines. This research design will not involve the development of a new prototype but will instead focus on gathering evidence from established models and practices to provide recommendations for best practices.

### 3.2 Data collection and analysis

Data collection will primarily involve the gathering of existing reports that detail the use of automated ETL pipelines in various industries. These reports will include documentation from organizations that have adopted automated ETL tools, insights from case studies published by vendors or independent analysts, and other relevant publications. Key data sources will include industry whitepapers from organizations like Apache Nifi and Talend, case studies published by companies specializing in ETL solutions, and academic articles focusing on the integration of automation in ETL processes.

The analysis of these reports will focus on extracting relevant information regarding the benefits and limitations of automated ETL pipelines. The analysis will emphasize data quality improvements, scalability, efficiency, and governance enhancements reported by organizations. Data from case studies will be compared to assess common patterns, benefits, and challenges, such as time savings, data consistency, and the cost-effectiveness of automated versus manual ETL processes.

Quantitative data such as performance metrics (e.g., processing speed, error rates) will also be extracted from the reports where available. This data will be analyzed to determine the effectiveness of automation in improving data workflows, reducing manual intervention, and enhancing data governance. The analysis will be comparative, contrasting traditional ETL processes with automated processes where the data is available.

### 3.3 Limitations

The primary limitation of this methodology lies in the dependence on secondary data from reports. The analysis is constrained by the accuracy and relevance of the reports available. Since this research will not involve direct engagement with companies, it will rely heavily on the reported experiences and performance metrics of others. This could lead to biases or selective reporting, where organizations may highlight the successes of their automation projects while minimizing or omitting challenges. Moreover, the study will be limited by the scope of available reports, which may not cover all industries or offer a comprehensive view of ETL automation in practice. Additionally, data quality issues within the reports, such as discrepancies or gaps in performance data, may affect the robustness of the analysis. The study will aim to mitigate these limitations by selecting a diverse set of reports from credible sources, such as vendors, consultants, and academic journals, to ensure the reliability of the findings.

Another limitation is the lack of direct data collection from organizations or actual implementation cases, which may hinder the ability to fully understand the practical nuances and challenges involved in the adoption of automated ETL pipelines. However, by synthesizing findings from multiple credible sources, the research will provide a well-rounded understanding of the benefits and challenges associated with the automated ETL approach. The scope of this study will focus on reports from industries that heavily rely on data, such as finance, healthcare, and retail, and may expand to other sectors based on the availability of reports. By narrowing the scope to these industries, the research will provide specific insights into the impact of automated ETL pipelines in data-intensive environments.

## 4. Development of the automated ETL pipeline model

### 4.1 Key components of the model

The development of an automated ETL pipeline involves several core components, each playing a crucial role in ensuring efficient data processing, transformation, and integration into the analytics system. The first key component is data extraction, which involves gathering data from various sources, such as databases, APIs, cloud storage, or flat files. The extraction process must be automated to regularly pull data in real-time or near-real-time, ensuring the pipeline remains up-to-date with the latest information.

The second critical component is transformation rules. This step involves cleaning, standardizing, and enriching raw data to make it suitable for analysis. Transformation typically includes processes like data validation, handling missing values, and applying business logic to convert data into structured formats. Automated transformation ensures that data is consistently processed according to predefined rules, reducing human error and enhancing the reliability of the data (Oladosu *et al*, 2022; Onukwulu, Fiemotongha, Igwe, & Ewim, 2022).

Data loading mechanisms represent the next core element, which involves loading the transformed data into a target data warehouse or data lake. Automated loading ensures that data is transferred smoothly and without delay, which is critical for enabling real-time analytics. This step can be configured to load data at regular intervals or triggered by specific events.

Finally, monitoring plays a significant role in the success of an automated ETL pipeline. Real-time monitoring allows for the detection of errors, performance bottlenecks, and irregularities, ensuring data quality and pipeline efficiency are maintained. Additionally, data quality checks and governance measures are integrated throughout the pipeline (Onukwulu *et al*, 2022). These checks can include validation rules, completeness checks, and consistency audits to ensure that the data is accurate, complete, and compliant with governance policies. By embedding data quality and governance checks at each stage of the ETL pipeline, businesses can ensure high-quality data flows, which, in turn, enhances analytics outcomes. Regular monitoring and auditing also ensure that the pipeline adheres to data governance standards, such as compliance with privacy regulations and industry standards (Otokiti, Igwe, Ewim, Ibeh, & Sikhakhane-Nwokediegwu, 2022).

### 4.2 Technology Stack

The technology stack used to develop an automated ETL pipeline is fundamental to its scalability, flexibility, and overall performance. Cloud platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) offer the necessary infrastructure for building and deploying scalable ETL pipelines. These platforms provide on-demand computing power, storage, and data processing capabilities, making managing large volumes of data easier while minimizing the need for expensive on-premises infrastructure (Zhu, 2017). For data integration, a variety of tools can be employed, such as Apache Nifi, Talend, Informatica, and Fivetran, which provide seamless integration with diverse data sources and destinations. These tools offer pre-built connectors and robust transformation capabilities that automate much of the data preparation process. The automation of integration tasks improves the efficiency of the pipeline by reducing manual data entry and minimizing errors (Dinh & Nguyen, 2022).

Orchestration software is another vital component of the technology stack. Apache Airflow and AWS Step Functions are widely used for managing complex ETL workflows. These tools allow for the automation of task scheduling, monitoring, and dependency management, ensuring that each step of the ETL process is executed at the right time and in the correct order (Sikeridis, Papapanagiotou, Rimal, & Devetsikiotis, 2017).

The cloud-based approach to building the ETL pipeline is particularly advantageous due to its scalability and flexibility. Cloud platforms allow businesses to scale the infrastructure up or down based on demand, enabling them to handle large data volumes without the need to invest in costly hardware. Additionally, cloud-based platforms provide the flexibility to integrate various data sources, systems, and tools, enabling more efficient data management and processing. This scalability and flexibility are essential for modern businesses that require agile and responsive data architectures to support decision-making (Laszewski, Arora, Farr, & Zonooz, 2018).

## 5. Conclusion and Recommendations

This study highlights the critical role of automated ETL (Extract, Transform, Load) pipeline models in enhancing data quality and governance in data analytics. The development of an automated ETL pipeline model provides numerous benefits, including streamlined data extraction, efficient transformation processes, and reliable loading mechanisms. By embedding data quality checks and governance measures at every stage of the pipeline, businesses can ensure that analytics data is accurate and compliant with regulatory standards. The study further demonstrates how automated pipelines improve decision-making by providing timely and high-quality data, enabling organizations to make more informed, data-driven decisions. Ultimately, the model helps businesses achieve optimized analytics outcomes, such as enhanced forecasting, customer insights, and operational efficiency.

Organizations considering the adoption of automated ETL pipelines must address several practical considerations to ensure successful implementation. One of the key recommendations is to invest in robust integration tools and cloud-based infrastructure to facilitate data flow automation. Additionally, businesses should prioritize employee training to overcome resistance to change and enhance the skill set of staff responsible for managing the pipeline. It is also crucial for organizations to adopt clear data governance policies and regularly monitor the performance of the ETL pipeline to ensure compliance with data quality standards. Overcoming challenges related to integration complexity and long-term maintenance is essential for realizing the full potential of automated ETL pipelines, which will ultimately drive operational efficiency and provide a competitive edge.

Future research could explore the integration of artificial intelligence (AI) and machine learning algorithms in the ETL pipeline model, which would further enhance the data transformation process and enable predictive analytics. The use of AI can automate more complex decision-making aspects of data transformation and quality assurance, improving both efficiency and accuracy. Additionally, the application of the automated ETL pipeline in various industries, such as manufacturing, healthcare, or finance, could provide valuable insights into its adaptability and effectiveness in different contexts. Future studies should also investigate the long-term impact of automated ETL pipelines on business performance, particularly regarding the optimization of data-driven decision-making and the enhancement of data governance across diverse organizational environments.

## 6. References


1. Abisoye A, Akerele JI. A practical framework for advancing cybersecurity, artificial intelligence and technological ecosystems to support regional economic development and innovation. 2022.
2. Abisoye A, Akerele JI. A scalable and impactful model for harnessing artificial intelligence and cybersecurity to revolutionize workforce development and empower marginalized youth. 2022.
3. Achumie GO, Oyegbade IK, Igwe AN, Ofodile OC, Azubuike C. AI-driven predictive analytics model for strategic business development and market growth in competitive industries. J Bus Innov Technol Res. 2022.
4. Adaralegbe AA, Egbuchiem H, Adeoti O, Abbasi K, Ezeani E, Adaralegbe NJ-F, *et al* Do personality traits influence the association between depression and dementia in old age? Gerontology Geriatr Med. 2022; 8:23337214211068257.
5. Adewoyin MA. Developing frameworks for managing low-carbon energy transitions: overcoming barriers to implementation in the oil and gas industry. 2021.
6. Adewoyin MA. Advances in risk-based inspection technologies: mitigating asset integrity challenges in aging oil and gas infrastructure. 2022.
7. Ajayi A, Akerele JI. A high-impact data-driven decision-making model for integrating cutting-edge cybersecurity strategies into public policy, governance, and organizational frameworks. Int J Multidiscip Res Growth Eval. 2021;2(1):623–37. doi:10.54660/IJMRGE.2021.2.1.623-637.
8. Ajayi A, Akerele JI. A practical framework for advancing cybersecurity, artificial intelligence and technological ecosystems to support regional economic development and innovation. Int J Multidiscip Res Growth Eval. 2022;3(1):700–13. doi:10.54660/IJMRGE.2022.3.1.700-713.
9. Ajayi A, Akerele JI. A scalable and impactful model for harnessing artificial intelligence and cybersecurity to revolutionize workforce development and empower marginalized youth. Int J Multidiscip Res Growth Eval. 2022;3(1):714–19. doi:10.54660/IJMRGE.2022.3.1.714-719.
10. Dhayne H, Haque R, Kilany R, Taher Y. In search of big medical data integration solutions-a comprehensive survey. IEEE Access. 2019;7:91265–90.
11. Dinh V-N, Nguyen H-A. Data process approach by traditional and cloud services methodologies. 2022.
12. Elumilade OO, Ogundeji IA, Achumie GO, Omokhoa HE, Omowole BM. Enhancing fraud detection and forensic auditing through data-driven techniques for financial integrity and security. J Adv Educ Sci. 2021;1(2):55–63.
13. Elumilade OO, Ogundeji IA, Achumie GO, Omokhoa HE, Omowole BM. Optimizing corporate tax strategies and transfer pricing policies to improve financial efficiency and compliance. J Adv Multidiscip Res. 2022;1(2):28–38.
14. Gade KR. Overcoming the data silo divide: a holistic approach to ELT integration in hybrid cloud environments. J Innov Technol. 2021;4(1).
15. Hassan YG, Collins A, Babatunde GO, Alabi AA, Mustapha SD. AI-driven intrusion detection and threat modeling to prevent unauthorized access in smart manufacturing networks. Artif Intell. 2021;16.
16. Ikegwu AC, Nweke HF, Anikwe CV, Alo UR, Okonkwo OR. Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. Clust Comput. 2022;25(5):3343–87.
17. Laszewski T, Arora K, Farr E, Zonooz P. Cloud native architectures: Design high-availability and cost-effective applications for the cloud. Packt Publishing Ltd; 2018.
18. Mustapha SD, Ibitoye B. Comprehension analysis of traffic signs by drivers on urban roads in Ilorin, Kwara State. J Eng Res Rep. 2022;23(6):53–63.
19. Nylén D, Holmström J. Digital innovation strategy: a framework for diagnosing and improving digital product and service innovation. Bus Horiz. 2015;58(1):57–67.
20. Odio PE, Kokogho E, Olorunfemi TA, Nwaozomudoh MO, Adeniji IE, Sobowale A. Innovative financial solutions: a conceptual framework for expanding SME portfolios in Nigeria's banking sector. Int J Multidiscip Res Growth Eval. 2021;2(1):495–507.
21. Oladosu SA, Ige AB, Ike CC, Adepoju PA, Amoo OO, Afolabi AI. Revolutionizing data center security: Conceptualizing a unified security framework for hybrid and multi-cloud data centers. Open Access Res J Sci Technol. 2022;5(2):086–076.
22. Olbert M, Spengel C. International taxation in the digital economy: Challenge accepted? World Tax J. 2017;3.
23. Onukwulu EC, Fiemotongha JE, Igwe AN, Ewim CP-M. Int J Manag Organ Res. 2022.
24. Otokiti BO, Igwe AN, Ewim C, Ibeh AI, Sikhakhane-Nwokediegwu Z. A framework for developing resilient business models for Nigerian SMEs in response to economic disruptions. Int J Multidiscip Res Growth Eval. 2022;3(1):647–59.
25. Otokiti BO, Igwe AN, Ewim CP-M, Ibeh AI. Developing a framework for leveraging social media as a strategic tool for growth in Nigerian women entrepreneurs. Int J Multidiscip Res Growth Eval. 2021;2(1):597–607.
26. Paul PO, Abbey ABN, Onukwulu EC, Agho MO, Louis N. Integrating procurement strategies for infectious disease control: Best practices from global programs. Prevention. 2021;7:9.
27. Rachakatla SK, Ravichandran P, Machireddy JR. The role of machine learning in data warehousing: Enhancing data integration and query optimization. J Bioinform Artif Intell. 2021;1(1):82–103.
28. Raj A, Bosch J, Olsson HH, Wang TJ. Modelling data pipelines. Presented at: 2020 46th Euromicro Conference on Software Engineering and Advanced

Applications (SEAA); 2020.
29. Russom P. Next generation data integration. TDWI Best Practices Report. 2011;Q2.
30. Sikeridis D, Papapanagiotou I, Rimal BP, Devetsikiotis M. A comparative taxonomy and survey of public cloud infrastructure vendors. arXiv preprint. 2017;arXiv:1710.01476.
31. Singu SK. ETL process automation: Tools and techniques. ESP J Eng Technol Advancements. 2022;2(1):74–85.
32. Thumburu SKR. A comparative analysis of ETL tools for large-scale EDI data integration. J Innov Technol. 2020;3(1).
33. Zhu D. Large scale ETL design, optimization and implementation based on Spark and AWS platform. 2017.