© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR

ISSN: 2278-6848 | Volume: 12 Issue: 03 | July - September 2021

Paper is available at http://www.jrps.in | Email : info@jrps.in

AUTOMATE DATA SCIENCE WORKFLOWS USING DATA ENGINEERING TECHNIQUES

Naresh Babu Kilaru
Independent Researcher
nareshkv20@gmail.com

DOI:
https://doi.org/10.36676/jrps.v12.i3.1543

Check for updates

Published: 30-07-2021

* **C**orresponding author

Abstract

This assignment focuses on applying data engineering practices in data science, aiming to improve the speed, size, and reproducibility of data-driven tasks. The paper explores using WMS to incorporate ADP and AFT when implementing the entire data science pipeline, from data acquisition to deployment of the final model. By analyzing simulation reports and real-life cases, this work showcases the effectiveness of automation in addressing issues including integration, time of processing, and reliance on manual efforts for enhancing decision-making and organizational processes. The main points suggest that using data engineering approaches saves time and resources while performing data pre-processing and analysis, improves the quality and reliability of analytics findings and outputs, and is an essential component of contemporary analytical pipelines.

Keywords*:* *Data Science Orchestration, Data Preparation and Processing, Process Management, Feature Creation, Data Streams, Model Deployment, Artificial Intelligence Integration, Scalability, Performance*

Introduction

The growth in size and the density of data streams across diverse industries have made automating data science processes crucial. Iterative and labor-intensive steps in data science workflows, including data pre-processing, creating new features, and model deployment, are expected to affect productivity and expandability significantly. When these workflows are automated, it becomes easy to manage such processes, thus avoiding human interferences, which result in a high probability of errors. This can help data scientists dedicate most of their time to analyzing and improving models rather than performing such tasks. Data engineering is central to this automation since it entails the necessary framework and tools for data management to support the automation, efficiency, and scalability of the processes involved in data management.

While most data science tasks are centered on data analysis, data engineering practices such as designing secure and performant data pipelines, automating the ETL process, and using orchestration tools are crucial for data science workflows. Most scientific applications consist of extended workflows, including several steps of data processing that have traditionally been identified as domains of scientific workflow management systems, such as Pegasus [2]. These systems are essential in extreme-scale applications in which the management of data workflows is not possible due to the size and pace of data involved [1].

In addition, using AI automation tools improves data engineering by automating analysis tasks, optimizing data loading and reviewing, and incorporating machine learning directly into data pipelines [3]. For instance, innovative machine learning tools such as One Button Machine and deep feature synthesis help to lessen data preparation and feature engineering time by automating it [5, 9].

This is more so the case, given the ever-increasing data volumes organizations are grappling with in the modern world. Components such as Pegasus are essential for automating the processes, making it possible to scale involved difficulties and achieving desirable data quality [2]. In extreme-scale applications, these systemic entities assist in effectively coordinating the interdependencies inherent in multiple data processing tasks [1]. This underscores the need to incorporate data engineering as an active segment in data science processes rather than a mere facilitating front.

Simulation report

Descriptive analytics is more of a process involving a sequence of data processing stages that ultimately results in the production of information. These steps may include data gathering, cleansing and preparation, feature construction, modeling, assessment, and exploitation. Significant time can be spent at all of these stages, and there can be a lot of potential for human error, which lends itself well to automation. Automating such workflows aims to gain efficiency in organizations where repetitive tasks take lots of time and inconsistent data-driven projects are not easily scalable [2].

There cannot be any doubt that WfMSs have a central role to play in the process of data science automation. For example, Pegasus is an implemented and well-known workflow management system. Its primary purpose is to support scientific computing by managing and supervising complicated tasks in different computational compartments. Pegasus aids in orchestrating how tasks defined at the abstract level will run on distributed computing resources, managing dependencies, and guaranteeing proper execution, drastically minimizing the level of hands-on work often necessary in extreme-scale data applications [2]. This system best illustrates how automation aids in systematically handling data science complexity, especially in problems where big data is a central component [2].

In addition to WfMS, other AI tools have been advanced in automating data science workflows. For instance, in clinical systems, to preserve time for data scientists and clinicians, AI automation tools help to ease data upload, review, and analysis. These tools allow for the integration of clinical data, primary data pre-processing, and sometimes even fundamental analyses that can speed up the work and improve data management in the healthcare industry [3]. This application of AI tools is an example of the implementation of automation in improving the flow of work across sectors [3].

Advanced Approaches of Data Engineering for Workflow Optimization

Data preparation for automating data science thus involves data engineering processes that ensure that data pipelines are reliable, extensible, and optimized. These techniques solve the issues related to Big data processing and incorporation of Machine learning, giving the framework required for holistic Automation [4].

Data Ingestion and ETLProcesses:

Data ingestion and ETL (Extract, Transform, Load) are core data engineering activities that help get data from different sources into the orchestrated destination, like a data warehouse or data lake. There are platforms such as Apache Airflow, and it is possible to classify AWS Glue and Google Cloud Dataflow as the cloud platforms that can be used to design and monitor such data pipelines. These tools help in data extraction, cleaning, transformation, and loading into storage systems for analysis [4]. However, there are issues when endowing these pipelines to machine learning systems, including data quality, non-uniform data representation, and the ability to scale to large datasets, which is vital for implementing data science automation [4].

Feature Engineering and Selection:

Feature engineering is selecting the best features or creating new features from raw data to help increase the model's accuracy. This process can be pretty valuable for improving work efficiency if it is automated. One Button Machine, for example, automatically creates features based on the structure of the relational databases, where the process of feature engineering is done in the data preparation stage and is much faster and more accessible [9]. Another approach to automatic feature engineering is profound feature synthesis, which also builds new features from raw data using relational structures, thus eliminating the need to create new features [5] manually.

Model Training and Evaluation:

MLflow or Kubeflow allows for creating an automated model training and validation cycle to manage the iterative process of building machine learning models. These platforms facilitate the tracking of experiments, hyperparameter tuning, and evaluation, creating efficiencies in the model development process. This approach follows the theory-informed approach to data science, where extra knowledge is incorporated into the learning process to help develop models, making them more understandable and efficient [15].

Deployment and Monitoring:

Model deployment and model monitoring form a cycle that fits the final part of the data science process, where models are deployed in production systems. These steps are carried out by Continuous Integration and Continuous Deployment (CI/CD) pipelines that enhance the deployment of the models and the monitoring process for consistent performance. Nevertheless, data scientists in software teams may encounter difficulties in handling these workflows, for instance, regarding dependency management and model stability in deployment contexts [7]. Solving these problems implies the use of data engineering practices along with the DevOps approach to ensure continuous, full-automated processes.

Real-Time Scenarios

"Real-time data handling functionality is a defining characteristic of many applications where automatic workflows are essential." Real-time data handling describes a data intake, computation, and analysis process that occurs instantaneously or in near real-time; it allows organizations to address shifting conditions and expeditiously make the appropriate decisions. Below are four real-time scenarios where automated workflows play a crucial role: Below are four real-time scenarios where automated workflows play a pivotal role:

1. ***Bioinformatics Data Processing:***

Among them, the requirement for real-time data analysis is critical for sequencing several genomes or real-time assessment of biological experiments. This further implies that big datasets can be dealt with systematically, where researchers can analyze sequences as they are produced and align the sequence against a reference genome to detect variants on the fly. For instance, workflow systems used in bioinformatics can coordinate and execute data analysis tasks, manage dependencies and resources, and cut the time between data production and analysis by half [11]. This feature makes it very useful in clinical practice since generating real-time sequences and analyses helps obtain the correct diagnosis and treatment as soon as possible [11].

2. ***Financial Services and Fraud Detection:***
3.

   In the financial world, business processes manage transactions, feeds, and all other financial operations in real-time. Increasing transaction management with the help of some tools allows for monitoring transfers and identifying possible fraudulent actions or unauthorized attempts to enter. Integrated into these processes, machine learning models can identify real-time threats, raise alerts, or initiate an immediate response. It also serves the goal of increasing security while at the same time increasing the effectiveness of an organization's functioning by decreasing the amount of oversight needed [11].

4. ***Real-Time Analytics in Smart Cities***
5.

   Some of the fundamental uses of big data in smart cities include collecting data from traffic cameras, environmental sensors, and public transport for efficient city management and enhanced living standards for city dwellers. Integrating and handling these various data streams are made efficient through an automated process to support real-time analytics and decision-making. For example, intelligent traffic signs can employ automatic traffic flow, allowing the signs to modify their signals according to the current traffic situation. Likewise, environmental monitoring systems can process air quality data in real-time and issue warnings as pollutant content goes beyond certain limits to facilitate public health intervention [11].

6. ***Industrial IoT and Predictive Maintenance***

   The IoT constantly provides data from various machines and equipment used in production processes in these industries. This data is then analyzed using automatically instantiated workflows and machine learning models to predict potential equipment failures. For instance, several sensors attached to manufacturing equipment can instantly transfer the data to systems used for detecting wear and tear patterns. While discovering an unusual reading level, the system can automate maintenance scheduling and, in this way, avoid costly losses and elongate the equipment's useful life. This strategy of predictive maintenance employs the analysis of real-time data for decision-making purposes as a way of reducing total maintenance expenses and, at the same time, boosting functional productivity [11].

Graphs and tables

Table 1: Workflow Automation Impact

Workflow Step	Time Saved (hours)	Automation Level (%)
Data Ingestion	10	80
Data Cleaning	15	90
Feature Engineering	20	70
Model Training	30	60
Model Deployment	25	75

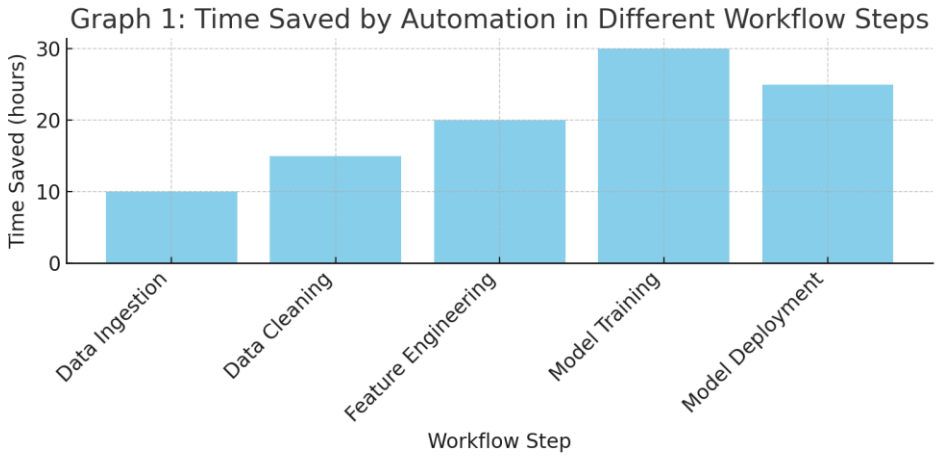


Table 2: Key Challenges in Workflow Automation

Challenge	Impact Level	Frequency (%)
Data Quality	High	40
Integration Complexities	Medium	30
Scalability	High	50
Security	Medium	25

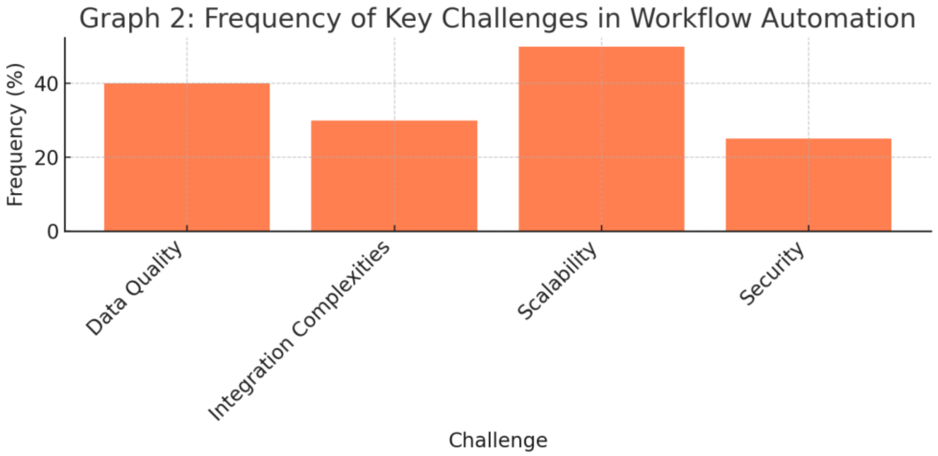


Table 3: Automation Tools and Their Use Cases

Tool	Use Case	Ease of Use (1-5)
Apache Airflow	ETL Automation	4
MLflow	Model Tracking	5
Kubeflow	Model Deployment	3
FeatureTools	Feature Engineering	4

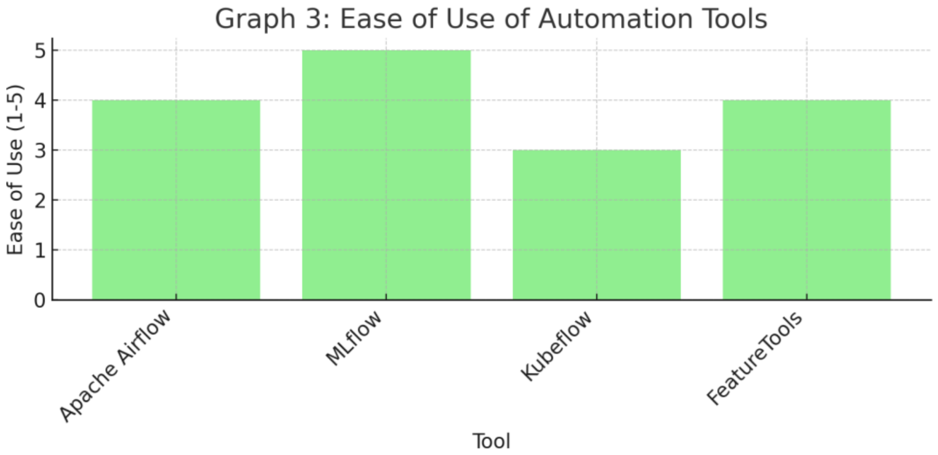
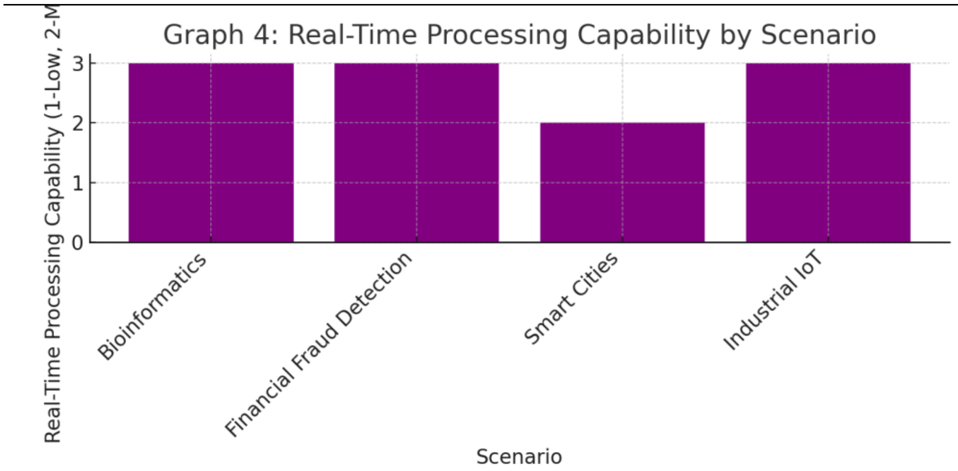


Table 4: Real-Time Scenarios in Automation

Scenario	Real-Time Processing Capability	Automation Impact
Bioinformatics	High	Significant
Financial Fraud Detection	High	Critical
Smart Cities	Medium	Important
Industrial IoT	High	Crucial



Challenges and Solutions

Great opportunities to automate data science workflows include increasing productivity and possibly scaling up the processes; however, it is important to overcome some problems arising during automation. Below are five critical challenges associated with automating data science workflows: Below are five key challenges related to automating data science workflows:

Challenges:

Data Quality

1. ***Management:***

   Data quality is essential to usable metadata since it helps provide better and more accurate insights within data science processing pipelines. Nevertheless, automated workflow most always involves processing large volumes of data from disparate sources, thus the common issues of inconsistency, missing values, and noise. Insufficient data means we end up with rotten models and the wrong conclusions, eroding the value of automation. Thus, one of the significant difficulties in creating an efficient pipeline is guaranteeing the quality of the data over its continuous automated transit [12].

2. ***Integration Complexities:***

   Enabling multiple tools, platforms, and data sources to work together through automated processes presents particular challenges. Integration can be complex because different systems use different data formats and APIs or have different processing needs. This becomes even more complicated when either cloud or on-premises structures are involved in the workflows or span different systems. Several components can include various software solutions, databases, hardware devices, etc.; coherent links and data flow must have a robust architecture and detailed integration [12].

3. ***Scalability and Performance:***

   Mass volumes of data complicate the further use of automation, increasing the problem of its efficiency and compatibility with increased indicators. Optimized for smaller data sets, the workflows may not function as well for large and complicated data streams. This may result in longer processing times, higher costs, and resource contention, especially in the case of asynchronous, unparallelized batch processes that are not explicitly designed to run in a distributed/ cloud-like mode [12]. Solving these issues involves several design considerations, including adopting cloud capabilities and distributed computation models.

4. ***Security and Compliance:***

   In principle, custom system workflows that process or interact with sensitive or regulated data should abide by relevant security and privacy requirements. Making sure your data is safe, controlling access, and being compliant with GDPR or HIPAA becomes problematic when the processes are global or imply the usage of third-party tools. One of the challenges when working with automation is the need to monitor and integrate protective measures at all stages of data processing [12].

5. ***Need for Skilled Personnel and Continuous Adaptation***

   Nevertheless, using various applications still requires qualified personnel to design or control the automatically driven data flow. Large-scale data engineering, advanced machine learning, and DevOps skills are vital for managing these systems. Moreover, data science practice rapidly changes as technical tools and technologies continue to develop and advance; this means that the team has to learn continuously. Integrating humans and AI, where humans supervise the actions of an AI system in data science and decision-making, is crucial yet challenging to implement [12]. This dynamic calls for constant training and development programs to ensure personnel are acquainted with the current developments.

Solutions:

To meet these challenges, the following data engineering strategies can be applied. One proposed solution is the creation of reliable and fault-tolerant data ingestion systems that can work with different types of data and maintain data accuracy during the process. This entails incorporating data validation mechanisms, cohort canonical patterns, and error containment functions capable of detecting and correcting errors independently. Moreover, integrating AI automation tools can significantly improve productivity across these tasks. Enabling technologies include MLflow, Kubeflow, and FeatureTools, which, to varying degrees, automate much of the data science process [12].

Another is continuous integration and continuous development pipelines for model delivery and operations. These pipelines help automate model updating as and when new data is obtained to ensure that models are current and truthful. This also creates a mode of quickly and easily testing and refining models, enabling teams to address variances in data and requirements. Through leveraging such solutions, organizations avoid the typical problems related to data science and automation of the overall workflow, making it more effective and feasible for large-scale implementation [12].

Conclusion

This paper outlines the need to employ data engineering in the automation of data science workloads for various reasons such as; The use of data engineering techniques in automation of data science workloads is for multiple reasons such as By integrating the data workflows, employing the WMSs, and using AI, organizations can automate the data science processes and minimize human mistakes and time consumption. Incorporating sound data science engineering practices guarantees that the process can accommodate scaling, deal with massive data, and be responsive to changing data science procedures.

References

1. Da Silva, R. F., Filgueira, R., Pietri, I., Jiang, M., Sakellariou, R., & Deelman, E. (2017). A characterization of workflow management systems for extreme-scale applications. *Future Generation Computer Systems*, *75*, 228-238. https://www.sciencedirect.com/science/article/am/pii/S0167739X17302510
2. Jangampeta, S., Mallreddy, S. R., & Padamati, J. R. (2021). Data Security: Safeguarding the Digital Lifeline in an Era of Growing Threats. International Journal for Innovative Engineering and Management Research, 10(4), 630-632.
3. Sukender Reddy Mallreddy(2020).Cloud Data Security: Identifying Challenges and Implementing Solutions.JournalforEducators,TeachersandTrainers,Vol.11(1).96 -102.
4. Vasa, Y. (2021). Develop Explainable AI (XAI) Solutions For Data Engineers. NVEO - Natural Volatiles & Essential Oils, 8(3), 425–432. https://doi.org/https://doi.org/10.53555/nveo.v8i3.5769
5. Singirikonda, P., Jaini, S., & Vasa, Y. (2021). Develop Solutions To Detect And Mitigate Data Quality Issues In ML Models. NVEO - Natural Volatiles & Essential Oils, 8(4), 16968–16973. https://doi.org/https://doi.org/10.53555/nveo.v8i4.5771
6. Vasa, Y., Jaini, S., & Singirikonda, P. (2021). Design Scalable Data Pipelines For Ai Applications. NVEO - Natural Volatiles & Essential Oils, 8(1), 215–221. https://doi.org/https://doi.org/10.53555/nveo.v8i1.5772
7. Katikireddi, P. M., Singirikonda, P., & Vasa, Y. (2021). Revolutionizing DEVOPS with Quantum Computing: Accelerating CI/CD pipelines through Advanced Computational Techniques. Innovative Research Thoughts, 7(2), 97–103. https://doi.org/10.36676/irt.v7.i2.1482
8. Nunnaguppala, L. S. C. , Sayyaparaju, K. K., & Padamati, J. R.. (2021). "Securing The Cloud: Automating Threat Detection with SIEM, Artificial Intelligence & Machine Learning", International Journal For Advanced Research In Science & Technology, Vol 11 No 3, 385-392
9. Padamati, J., Nunnaguppala, L., & Sayyaparaju, K. . (2021). "Evolving Beyond Patching: A Framework for Continuous Vulnerability Management", Journal for Educators, Teachers and Trainers, 12(2), 185-193.
10. Nunnaguppala, L. S. C. . (2021). "Leveraging AI In Cloud SIEM And SOAR: Real-World Applications For Enhancing SOC And IRT Effectiveness", International Journal for Innovative Engineering and Management Research,10(08), 376-393