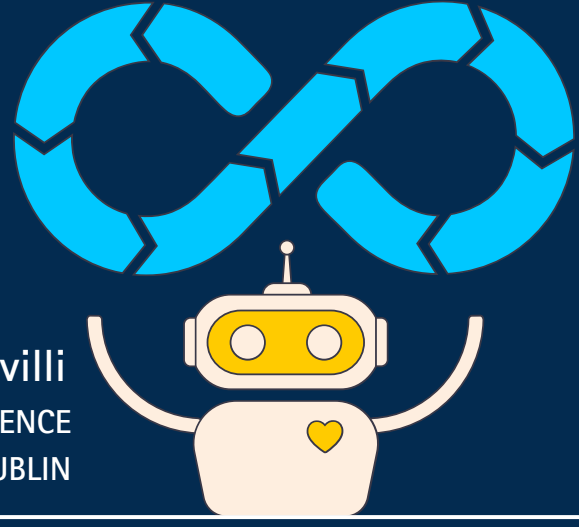
Automating MLOPs Data Ingestion Pipeline – A novel approach to democratize data collection



Surya Teja Gowd Ayinavilli
MSC. HUMAN CENTERED ARTIFICIAL INTELLIGENCE
TECHNOLOGICAL UNIVERSITY DUBLIN

Abstract

The exponential growth of AI applications has intensified the need for efficient data ingestion pipelines that seamlessly integrate data from diverse sources. This research proposes an automated MLOps data ingestion framework that streamlines data transport, transformation, and preparation for AI-ready storage. By optimizing ingestion strategies for batch and streaming architectures, this framework enhances reproducibility, sustainability, and accessibility for analytics teams. The Proof of Concept (PoC) demonstrates how automation can reduce manual intervention, ensuring a scalable and reproducible data pipeline. This novel approach democratizes data collection, enabling faster AI development and decision-making.

Introduction

Data fuels AI, but its collection and preparation remain major bottlenecks in MLOps workflows. Traditional ingestion pipelines are often manual, error-prone, and non-reproducible, limiting scalability. This research presents an automated data ingestion pipeline that accelerates data flow from diverse sources while ensuring traceability, sustainability, and reproducibility. By optimizing ingestion for batch and streaming AI architectures, this solution enhances accessibility for data science teams. The proposed framework minimizes setup overhead and democratizes AI-ready data collection for enterprises.

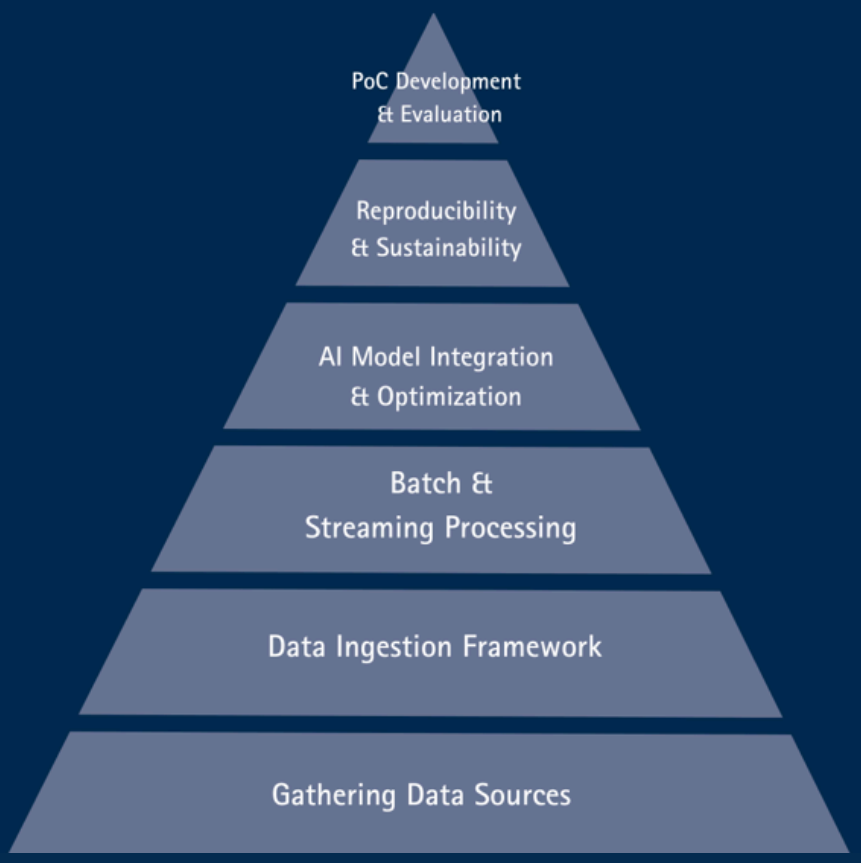
Literature Review

Machine learning (ML) has significantly improved automated data pipelines, yet challenges persist in ensuring scalability, adaptability, and efficiency. Vajpayee (2023) highlights ML-driven anomaly detection and self-healing mechanisms, enhancing data quality but primarily focusing on structured datasets [1]. Thopalle (2017) explores schema adaptation and deduplication using ML techniques, though real-time integration with heterogeneous data sources remains underdeveloped [2]. Zhao (2021) propose a disaggregated ingestion architecture to improve scalability in recommendation systems, yet its applicability across diverse AI workflows is limited [3].

While these advancements optimize data ingestion and processing, the need for a flexible, reproducible, and sustainable framework that seamlessly integrates diverse data sources and dynamically adapts to evolving AI models remains a critical challenge.

Research Question

How can machine learning-driven automation enhance the efficiency, scalability, and data quality of an MLOps data ingestion pipeline while minimizing manual intervention ?



Expected Output

- A fully automated data ingestion pipeline capable of seamlessly transporting data from diverse sources to AI-ready storage.
- Enhanced scalability and reproducibility of data pipelines for both batch and streaming data ingestion.
- A validated Proof of Concept (PoC) demonstrating a significant reduction in manual intervention and improved data quality for AI model training.

Further Work

- Explore the integration of additional data sources and formats, including real-time data streams from IoT devices and unstructured data such as text and images.
- Enhance the framework's adaptability by incorporating advanced machine learning techniques for dynamic schema evolution and real-time data validation.
- Investigate the deployment of the automated ingestion pipeline in different industry contexts, such as healthcare or finance, to assess its scalability and impact on domain-specific AI applications.

References.

1. A. Vajpayee, "The role of machine learning in automated data pipelines and warehousing: enhancing data integration, transformation, and analytics," Sep. 23, 2023. https://www.espjeta.org/jeta-v3i7p111
2. P. K. Thopalle, "REVOLUTIONIZING DATA INGESTION PIPELINES THROUGH MACHINE LEARNING: a PARADIGM SHIFT IN AUTOMATED DATA PROCESSING AND INTEGRATION," INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY (IJARET), Nov. 30, 2017. https://iaeme.com/Home/article_id/IJARET_08_06_017
3. Zhao, Mark et al. "Understanding and Co-designing the Data Ingestion Pipeline for Industry-Scale RecSys Training." *ArXiv* abs/2108.09373 (2021)

Acknowledgements
The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organized under the Erasmus + Programme of the European Union