

Single Day Event

***Open-Source Frameworks for Scalable Data
Quality Orchestration in Clinical Data
Pipelines***

Presenters :

Pavan Kumar Reddy Sakam

Vinodh Kumar Jestadi

AstraZeneca

Disclaimer : The views and opinions expressed in this presentation are those of the speakers and cannot be considered as a statement from AstraZeneca.

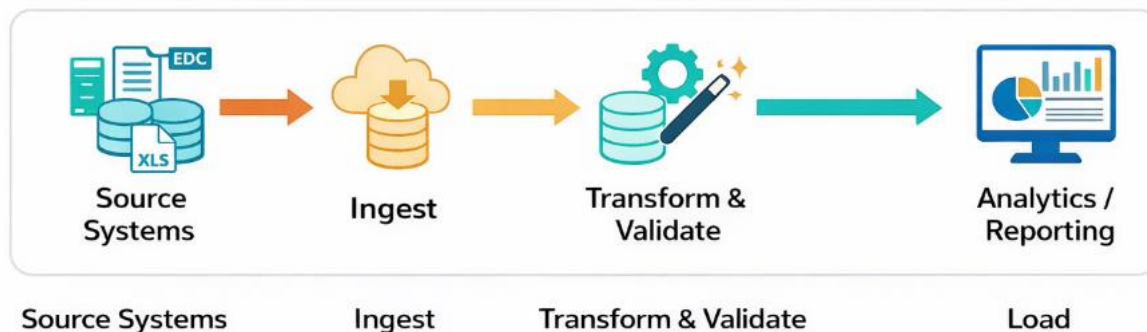
All materials have been prepared for informational purposes only and not for the purpose of providing legal advice.



Data Pipeline ?

A data pipeline is a set of automated steps that collect, move, transform, validate, and deliver data from source systems to target platforms, ensuring the data is reliable, consistent, and ready for analysis or reporting.

Example of a Data Pipeline



Importance in Clinical Domain:

- Regulatory-grade quality: Consistent, validated, audit-ready data.
- Faster insights: Near real-time ingestion from EDC, labs, ePRO/wearables, imaging, RWD.
- Efficiency & reuse: Standardized processes, less manual work, improved interoperability.
- Traceability: End-to-end lineage for reproducibility and transparency.

Problem Statement

Why Data Quality Is Still a Bottleneck ?



**Tightly Coupled
to legacy
systems**



**Hard to re-use
Validation rules**

Hard-coded, inconsistent,
and difficult to reuse



**Scaling issues
across studies &
platforms**



**Change require
manual effort and
rework**

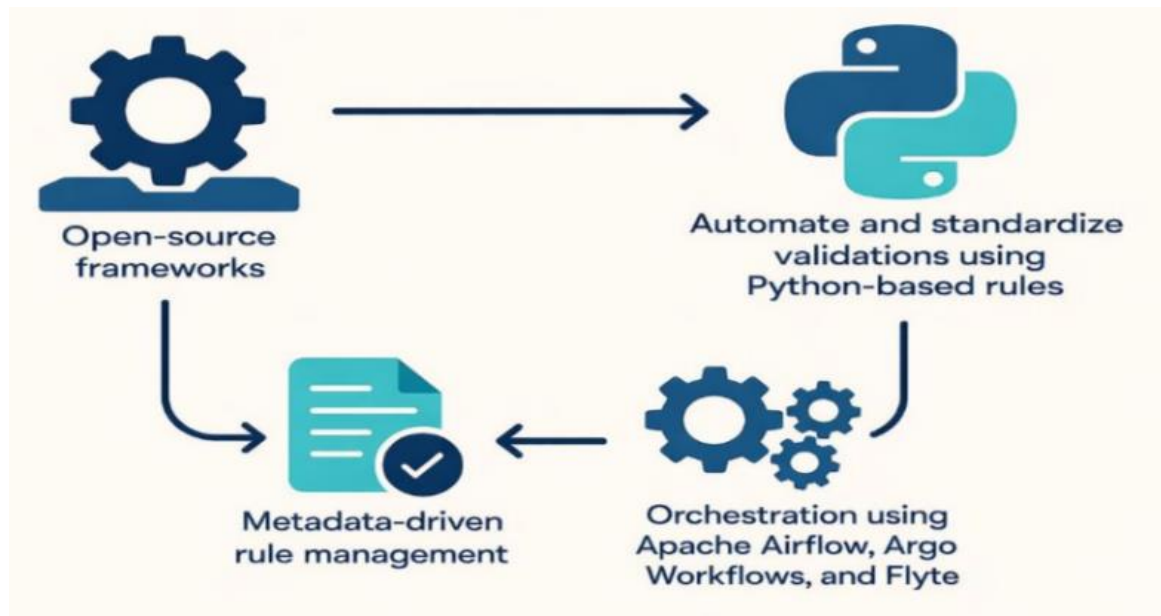
Re-work & long turn
arounds



**Limited Visibility
& Traceability**

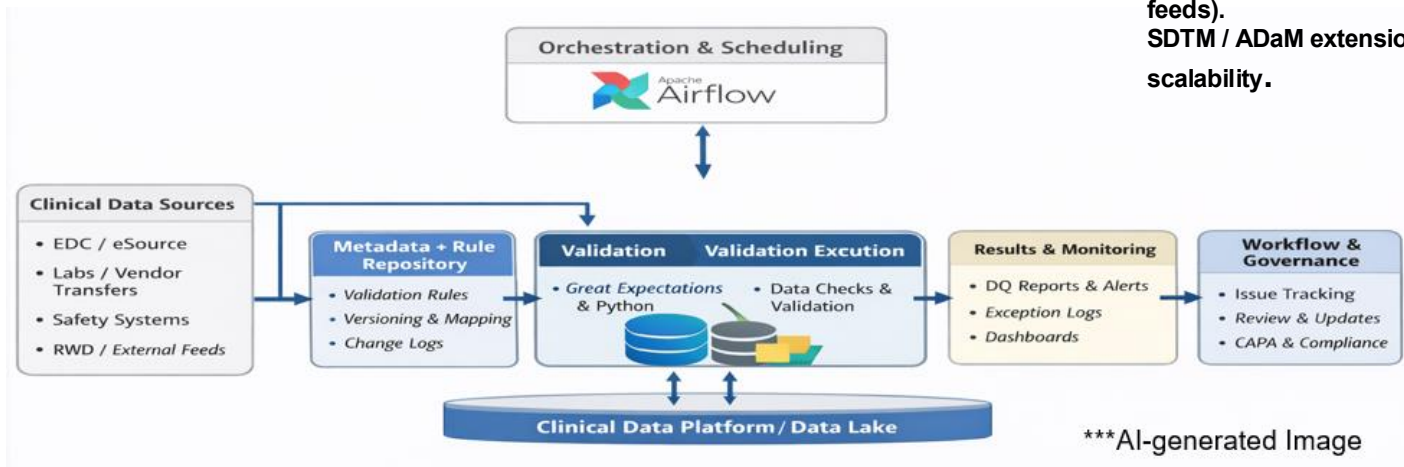


Solution : Open-Source–Driven Data Quality Orchestration





Conceptual Architecture



Current implementation focuses on source clinical data (EDC, labs, vendor feeds).
SDTM / ADaM extension is part of future scalability.

GXP Compliance Considerations :

Ensure pipelines / python programs / rules are validated, version-controlled, and fully auditable, with documented requirements, tested changes, access controls, and secure, traceable data processing aligned to 21 CFR Part 11 and Annex 11.



Apache Airflow

Enterprise-grade orchestration for controlled, scalable data processing

- Apache Airflow is an open-source orchestration platform that enables controlled, automated, and traceable execution of data quality workflows across clinical data pipelines.
- Widely adopted across enterprise and regulated environments

As a Data Quality Orchestration Layer :

- **Risk Reduction** : Minimizes manual intervention, missed checks, and inconsistent execution in critical data flows.
- **Operational Scale** : Standardizes data quality execution across studies, domains, and platforms without custom rebuilds.
- **Control & Compliance Enablement** : Supports disciplined execution, traceability, and inspection readiness aligned to GxP expectations.
- **Automate data quality execution** : Automation at the right points in the data lifecycle, ensuring consistent and repeatable validations across studies and domains



Case Study: Metadata-Driven Dynamic Validations using Airflow,Python

What was implemented :

- Dummy Data of DM,AE in CSV format
- Design Metadata validation rules required to check Quality of the Data
- Create Python scrips, DAG code and Docker rules
- Automated scheduling of Data Quality checks via Airflow

Key Outcomes :

- Single output file with PASS/FAIL status for the created validation rules
- Rule updates enabled without code redeployment
- Automated execution across multiple domains
- Reduced manual checks and dependency on individual programmers
- Easy to maintain the rule file and stronger data governance and traceability



Dummy_data_dm



Python_DQ_Samp
le_Code_DAG



Data_Quality_Vali
dator_Rules



Docker_Compose
r_Script



Dummy_Output



Stakeholder Utilization & Value

Data Management : Early detection of issues, reduced downstream rework

Programming Teams: Reduced manual reruns, reusable governed rules

Medical Review: Increased confidence in source data used for safety review

Quality & Compliance: Traceable, audit-ready execution evidence



Limitations & Future Scalability

Current Scope and Limitations

- Current scope limited to source clinical data
- SDTM / ADaM validation not included in current implementation

Future Direction

- Extend rule libraries for SDTM / ADaM
- Expand cross-domain and cross-dataset checks
- Progressive adoption across study lifecycle



Impact: Business, Technical & Governance Value

Measurable benefits

-  Reduced manual effort and rework
-  Full transparency and audit readiness
-  Lower cost compared to monolithic legacy solutions
-  Faster innovation using open-source ecosystems

Strategic alignment

- Supports **Resilience and Innovation**
- Builds **future-ready, trustworthy clinical data pipelines**

Contact Info:

- Pavan Kumar Reddy Sakam – [*Senior Clinical Programmer, AstraZeneca*]
- Email: pavankumar.sakam@astrazeneca.com
- LinkedIn: <https://www.linkedin.com/in/pavan-kumar-reddy-sakam-122634b4/>



- Vinodh Kumar Jestadi – [*Associate Director - Clinical Programming, AstraZeneca*]
- Email: vinodh.kumarjestadi@astrazeneca.com
- LinkedIn: <https://www.linkedin.com/in/vinodh-kumar-jestadi-081489a1/>



**The Global Healthcare
Data Science Community**

Contact Channels

- 📞 UK +44 1843 609600
- ✉ office@phuse.global
- 🌐 phuse.global

Social Media

- 🐦 [@phusetwitta](https://twitter.com/phusetwitta)
- 📘 [/phusebook](https://facebook.com/phusebook)
- 📺 [/phusetube](https://youtube.com/phusetube)
- 🌐 [/company/phuse](https://company.phuse.com)