Title: A Modular Framework for Data Harmonization: Enhancing Quality and Efficiency in Healthcare ETL Pipelines

PRESENTER: **Freija Descamps**

BACKGROUND:

- Harmonizing healthcare data is key for research and interoperability, but the development process is often slowed down by heterogeneous source data, project-specific setup needs and data quality issues.

METHODS

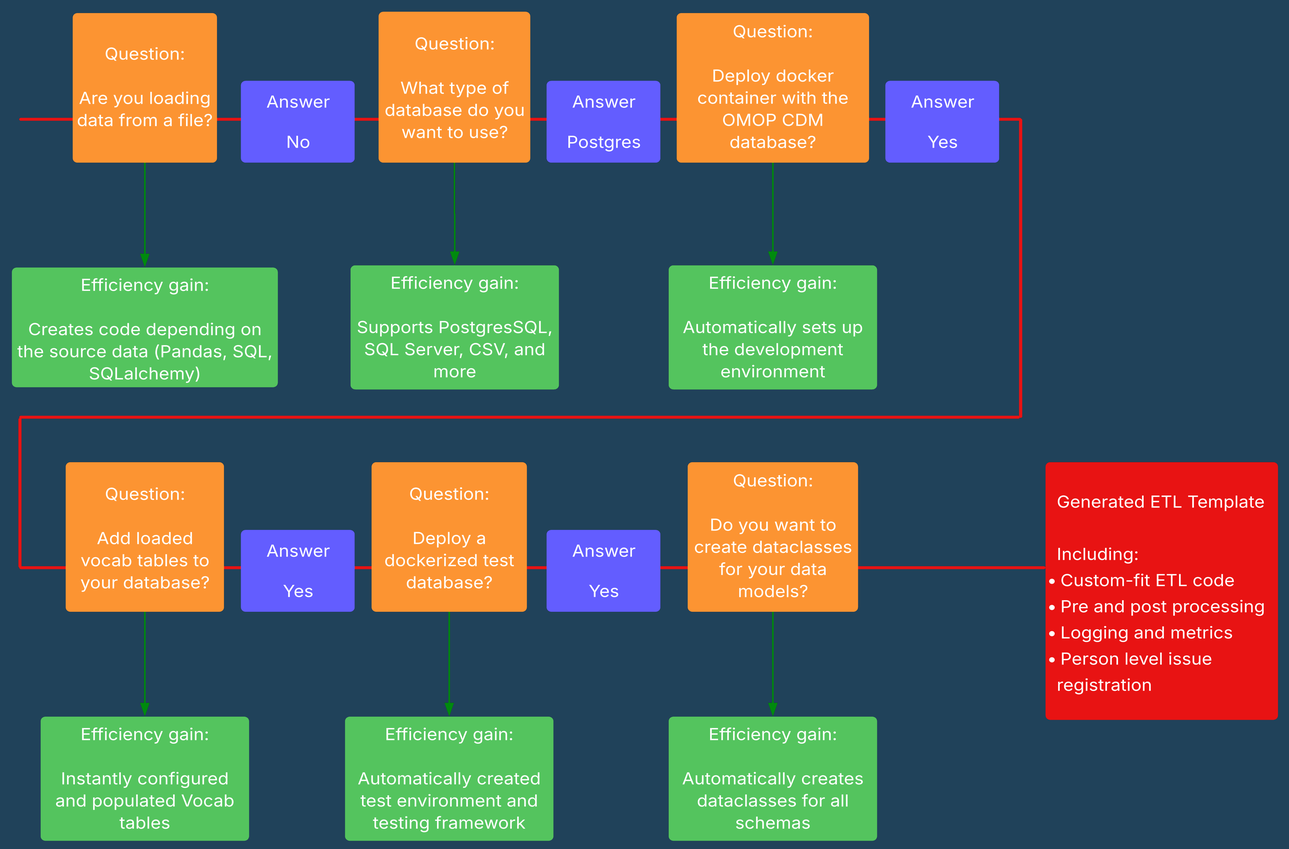
- Our framework automates ETL development by generating project templates based on user input. It supports multiple databases (e.g., PostgreSQL, SQL Server) and input types (e.g., CSV, other databases), and allows users to choose their preferred development approach (Pandas, SQLAlchemy, or raw SQL).

It includes:

1. **Pre- and post-processing modules** for standardizing formats (e.g., dates, nulls)
2. **Automated scaffolding** for customized, consistent project code
3. **Dockerized local environments** with source, target, and lookup schemas
4. **Ready-to-use local test setup** with a dedicated Docker database and preconfigured test code
5. **Auto-generated data classes** for all tables
6. **Comprehensive logging and metrics** to monitor data quality
7. **Issue tables** after each ETL run to detect person-level inconsistencies while preserving privacy

A Modular Framework That Reduces ETL Development Time and Boosts Data Quality in Healthcare Pipelines

HOW IT WORKS:



Take a picture to **download** the **full paper**

CONCLUSION

Our framework streamlines healthcare data harmonization by automating key ETL tasks. It reduces development time, improves data quality, and ensures consistency—making ETL pipelines easier to build, maintain, and share across teams.

Isaac Claessen[1], Silvia Jimenez Navarro[1], Shirah Cashriel[1], Panagiotis Gialernios[1]

[1]*edenceHealth NV*

edence Health

OHDSI