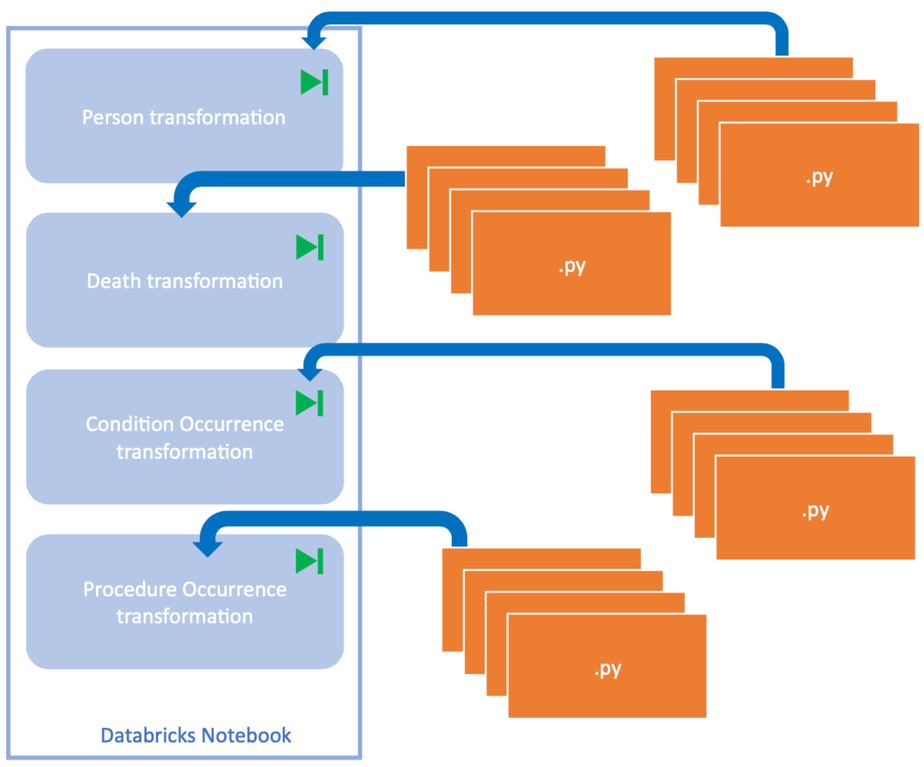
Key lessons from mapping population-wide electronic health records, in the NHS England Secure Data Environment, onto the OMOP Common Data Model v5.4, using Databricks and Apache Spark

Background: The Extract-Transform-Load (ETL) process for the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v5.4, provides a conversion for data resources within the National Health System (NHS) England's Secure Data Environment (NHSE SDE), accessible via the British Heart Foundation (BHF) Data Science Centre's CVD-COVID-UK/COVID-IMPACT Consortium. The NHSE SDE is a secure data and research analysis platform with population-wide person-level electronic health records from over 57 million live people in England. The platform offers various technologies, including Databricks, Spark, R Studio, R Studio Server and an internal GitLab. The ETL was implemented using Apache Spark, an open-source and distributed computing system that provides a fast and general-purpose cluster computing framework for big data processing and analytics.

Objectives and Challenges

The objective of the ETL is to map person-level data from primary and secondary care records, death registrations and cardiovascular disease audit datasets into OMOP CDM v5.4. The main challenges of the project included:

- restricted access to the NHSE SDE for approved researchers only;
- utilisation of Apache Spark as the primary Python library for ETL development.



Architecture Overview

The NHSE SDE is provisioned with high standards of system security and permissions for Delta and Hive tables. Embracing a modular design, specific transformations were implemented in distinct Python modules, which were subsequently imported into a Databricks notebook, orchestrating the entire ETL pipeline.

Workflow to reduce Databricks costs

Databricks operates on a subscription-based model where organisations pay for the resources consumed on the platform. To reduce costs and leverage user-friendly development environments, we opted to use two distinct environments: a Python virtual environment and Databricks. Primary development took place in the virtual environment.

Best practices for Apache Spark

- Optimise the Directed Acyclic Graph (DAG) of Spark transformations to improve ETL performance and overall cluster efficiency.
- Use Spark-friendly approaches in implementation, such as zipWithIndex() to set the primary key of the OMOP tables.
- Avoid repeated scanning of the entire DataFrame to optimise performance.
- Instead of relying on dictionaries for mapping, the recommended approach when working with Apache Spark is to perform a join transformation on two DataFrames. This helps efficient data mapping without the performance drawbacks associated with using dictionaries.
- Choose the correct mode of joining depending on the dimensions of Spark DataFrame and potential use of broadcasting to handle the data skew.

Conclusion: This project has addressed some of the challenges of conducting an ETL to map population-wide electronic health records on over 57 million people within a Secure Data Environment onto the OMOP CDM v5.4, using Databricks.

We recommend making essential workflow adjustments, including the use of multiple development environments and the use of best practices for Apache Spark, and highlight the 'team-science' approach that was integral to this project's success.

Silvia Jimenez[a], Mehrdad A. Mizani[b], Shirah Cashriel[a], Emma Gesquiere[a], Jadene Lewis[b], Angela Wood[b], Rouven Priedon[b], Anne Li[a]

[a]edenceHealth NV (BE)
[b]British Heart Foundation Data Science Centre, Health Data Research UK (UK)

edence Health
British Heart Foundation Data Science Centre
Led by Health Data Research UK
HDRUK Health Data Research UK
OHDSI