

# Cloud-based, Automated Solution for Transforming Clinical Datasets to the OMOP CDM

Simon Thompson & Abigail Carter  
Genomics England

## BACKGROUND

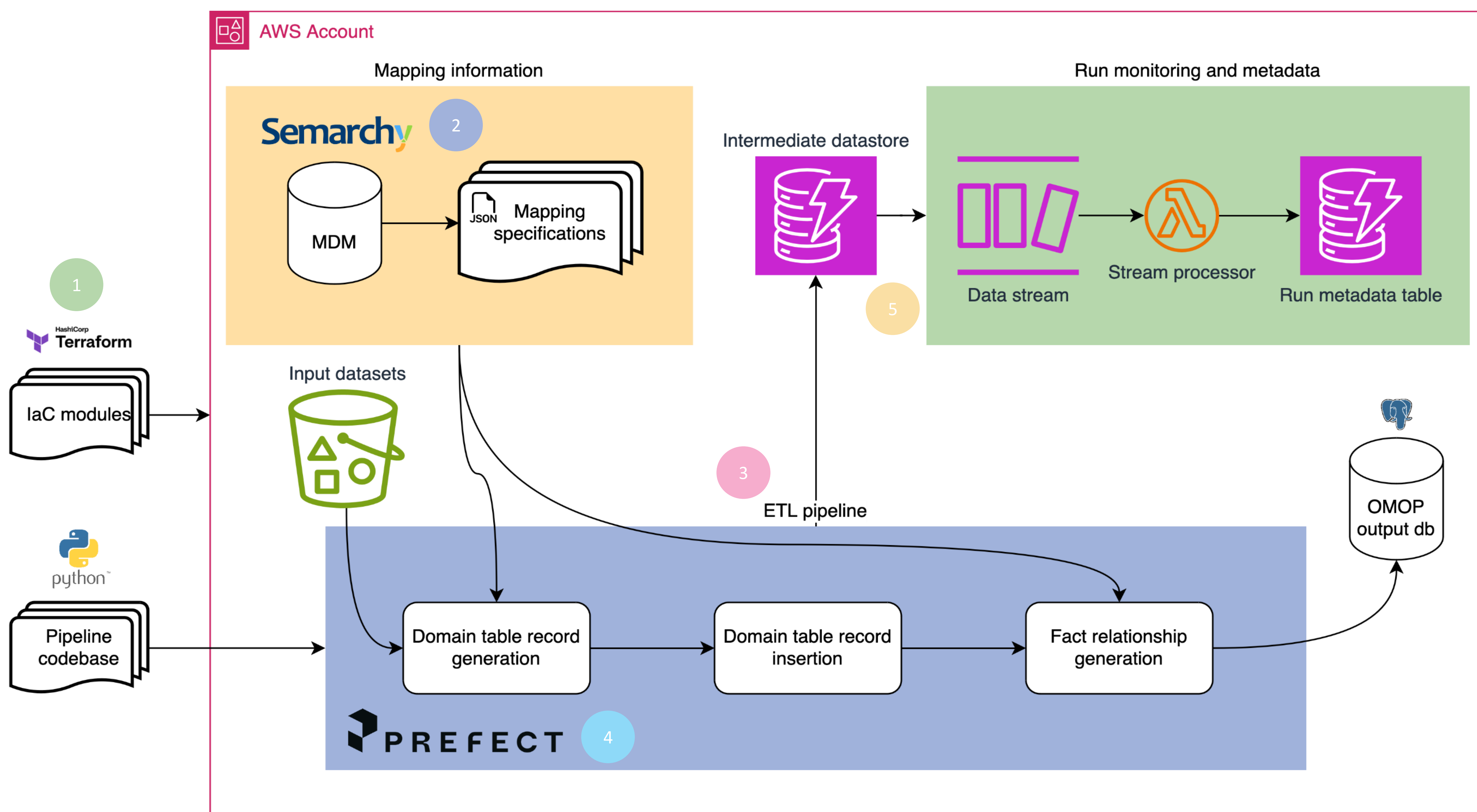
- Genomics England is a public-funded body that partners with the NHS to provide whole genome sequencing diagnostics and equip researchers with a large genomic database, the National Genomic Research Library (NGRL).
- The NGRL holds clinical data on over 100,000 patients. The information is currently held in separate data models without standardised concepts or enumerations. There is a desire to transform as much of this diverse dataset as possible to the OMOP CDM.

## PIPELINE REQUIREMENTS

The pipeline should:

- be cloud-based (in AWS) in its entirety and use serverless technology where appropriate,
- be high-throughput and process data in parallel,
- store mapping information in a format that is easily interrogated and updated,
- be secure, have a high level of testing and monitoring, and capture failing records at runtime.

## SOLUTION



- 1 Infrastructure is deployed using Terraform (an infrastructure-as-code tool).
- 2 Mapping information for all datasets and concepts is stored in a Master Data Management (MDM) platform, Semarchy (see poster by Carter et al. for further information).
- 3 Individual pipeline components (written in Python) for generating OMOP domain table records, generating fact relationship records, and inserting records into a database, are generic and parametrised by the MDM platform.
- 4 Prefect orchestrates the pipeline and manages the parallel execution of processes
- 5 Dynamodb, a serverless NoSQL database, is used as an intermediate datastore and feeds processes that monitor each pipeline run. Data is encrypted at rest and in transit throughout.