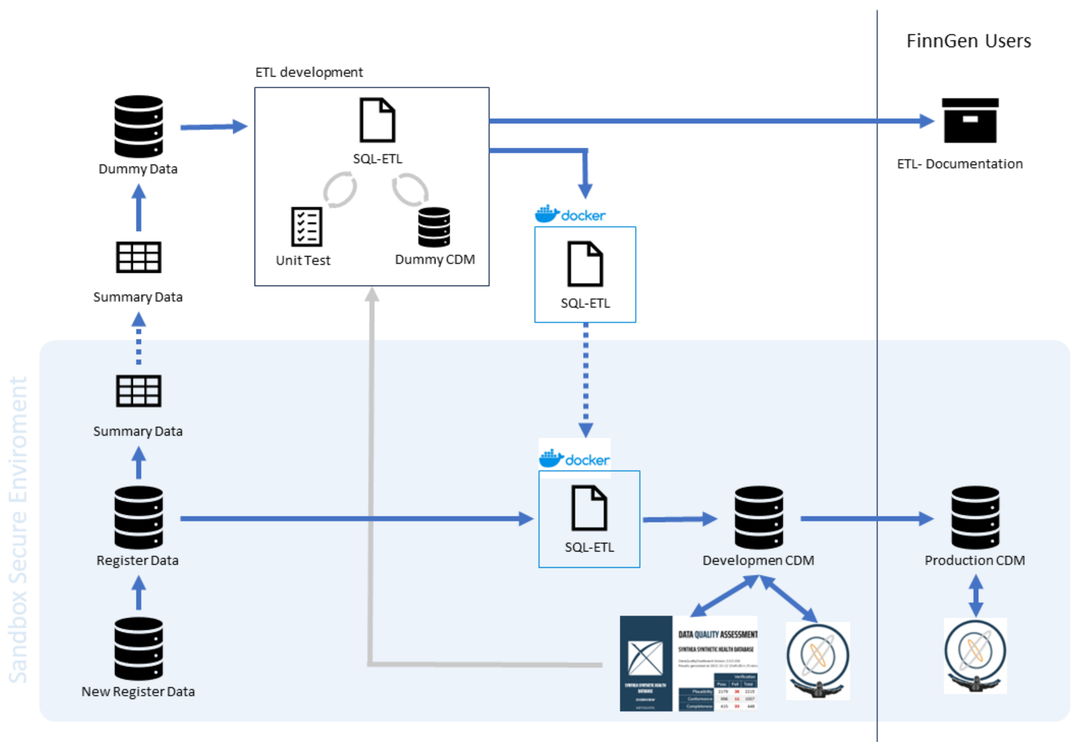
Continuous delivery ETL: Register data to CDM in FinnGen

FinnGen ETL using BigQuery SQL

Background: In addition to 500,000 samples of genotype data, FinnGen continuously expanding its register base phenotype data. We developed a process to transform current and incoming registers to the OMOP-CDM.

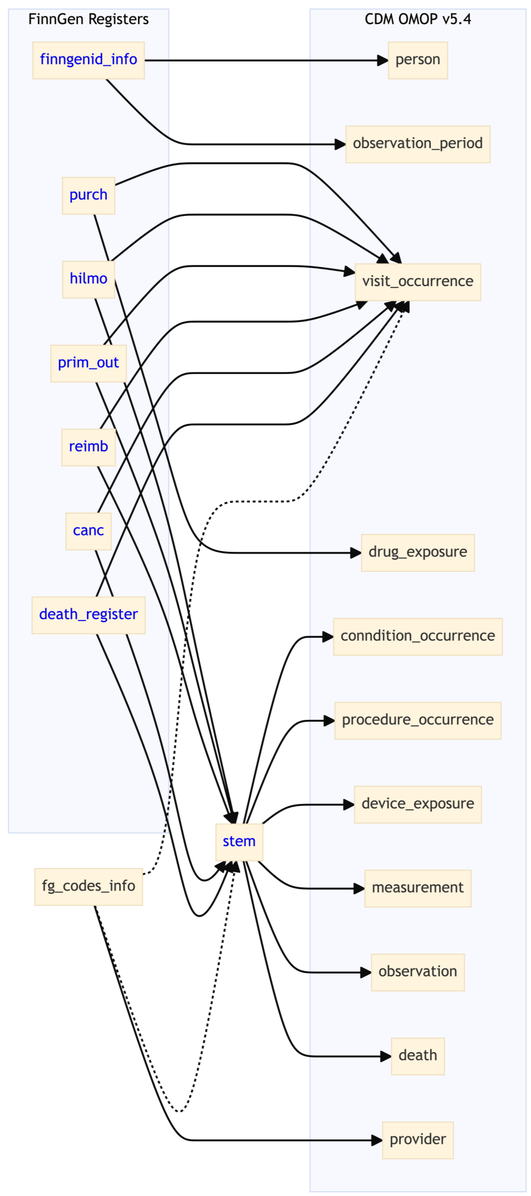
Methods



1. Custom R scripts to summarize the source data to be download from the safe environment.
2. Summary data is used to generate dummy data.
3. ETL is develop on the dummy data and unit tested.
4. Up to date documentation is generated and published.
5. SQL-ETL and support scripts are dockerized and uploaded into the safe environment.
6. SQL-ETL is applied to the source data to generate a development CDM instance.
7. Development CDM instance is tested on DQD and new features tested in Atlas.
8. When ready the development CDM is copied into production to be utilized by the users.

Results

- Six out of 17 national registries converted.
- 73 unit tests passed covering 10 CDM tables.
- Used stem and mapping tables.
- ETL ran within an hour.
- Utilized DQD to fix most of Plausibility and Conformance problems.
- Scan the QR code to explore the ETL documentation.



ETL: source and custom stem tables

Limitation: There are quite good number of unmapped codes in the vocabularies. Dummy data generator does not consider sex variable while assigning codes.

Shanmukha Sampath Padmanabhuni
Javier Gracia-Tabuenca, Mary Pat Reeve