dbt for OMOP Phase I: dbt-synthea

Katy Sadowski[1], Vishnu Chandrabalan[2], Lawrence Adams[3], Adam Bouras[4], Evanette K Burrows[5], Roger Carlson[6]

[1]Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, [2]Lancashire Teaching Hospitals NHS Foundation Trust, UK, [3]London Secure Data Environment (SDE), UK, [4]Tritonis, Inc., Boundbrook, NJ, [5]Janssen Research & Development, LLC, Raritan, NJ, [6]Corewell Health, Grand Rapids, MI

Background

Data build tool (dbt) is an open-source data transformation tool. It leverages the power of modern data warehouse technology to enable a SQL-first approach to data modeling, and provides simple interfaces for building modular, testable, and well-documented data transformation workflows. dbt has become extraordinarily popular since its launch in 2016 and is used by over 25,000 companies worldwide.[1]

The OHDSI community currently lacks a go-to tool for transforming source data into the OMOP CDM. While ETL "helpers" like White Rabbit and Rabbit-in-a-hat are widely adopted, the actual transformation of data into OMOP tends to be accomplished via closed-source vendor solutions or bespoke in-house ETL code. A handful of open-source ETL tools have emerged over the years, such as CaRRoT[2], Rabbit-in-a-blender[3], and Perseus[4]; however, no tool has yet achieved wide use as an OHDSI standard.

We believe that dbt has the potential to become such a standard. Successful implementations of dbt for site-specific OMOP ETLs have been demonstrated at recent OHDSI Symposia by Lancashire Teaching Hospitals (UK)[5], SiData+ (Thailand)[6], and Corewell Health (US)[7]. These projects cited superior performance; simplicity of development workflow; ease of integration with other technologies; and robust built-in documentation features as reasons for selecting dbt as their ETL technology.

We aim to build upon the learnings of these projects to develop a generalized dbt project template for OMOP ETL. The project will be data-source-agnostic and will leverage dbt's highly flexible configuration features and macros library to provide source-specific templates and modules for common transformations.

The first phase of this project, which we present in this paper, is a simple proof-of-concept which re-writes the ETL-Synthea[8] project in dbt - "dbt-synthea". dbt-synthea aims to provide a demonstration of how dbt can be used to develop an OMOP ETL and to promote a set of ETL development principles we believe will make OMOP ETL easier and more robust.

Methods

Building dbt-synthea has been a collaborative process undertaken by a small team of experts in dbt, OMOP ETL, and Synthea.[9] The basis of our collaboration is a series of monthly knowledge sharing meetings. All coding work occurs asynchronously via GitHub in the dbt-synthea repository, which is hosted in the OHDSI GitHub organization: https://github.com/OHDSI/dbt-synthea.

In Phase I of this project we produced a working replication of ETL-Synthea using dbt. We leveraged dbt's **modular, SQL-first architecture** to produce a set of parameterized SQL files called "models". As part of a dbt ETL run, dbt auto-compiles and executes the directed acyclic graph (DAG) of these models, the output of which is a "data mart" (an OMOP CDM instance, in this case). Following dbt best practices, dbt-synthea includes:[10]

- **Staging** models - SQL queries that are 1:1 with the source Synthea and vocabulary tables and are used to adjust data types and column names as needed

- **Intermediate** models - SQL queries used to perform the bulk of the joining and transformation needed to move from source tables into the OMOP CDM, with a focus on joins and transformations which may be reused several times in the final models; for example, source-to-standard concept mappings or modeling of different visit types
- **Mart** models - SQL queries that are 1:1 with the final OMOP tables

In a production dbt run, staging and intermediate models are generally not materialized in the database, but materialization can be toggled on for debugging and performance purposes.

dbt-synthea also leverages a feature of dbt called **seeds**. Seeds are csv files stored alongside the code in a dbt project's git repository which can be loaded into the database as part of a dbt run. We generated a small, 27-person Synthea dataset and corresponding OMOP vocabulary "shard"[11] comprising only concepts found in the Synthea data. These datasets are available in dbt-synthea as seeds for use by project developers as well as users looking to quickly spin up a working ETL.

dbt-synthea includes **built-in documentation and testing** for both the staging and mart (OMOP) models, defined using dbt's yaml-based configuration syntax. Documentation includes table- and column-level descriptions and DDL information, and tests can range from simple non-NULL and key constraint verifications to complex custom SQL checks.

Finally, we explored dbt's support for **cross-database usage** by adding support for both Postgres and duckdb in dbt-synthea. dbt supports 16+ database platforms and has features that enable cross-database SQL translation.

To test our implementation, we compared dbt-synthea's final OMOP CDM tables to those produced by an ETL-Synthea run on the seed datasets described above. We ran scripts to compare table structure, row count, and contents of the tables. The comparison was performed only in Postgres, as ETL-Synthea does not support duckdb.

Results

dbt-synthea successfully replicated the output of ETL-Synthea when run on the same source data. Row counts and table contents largely matched between all output OMOP CDM tables. The only differences were expected due to truncation of timestamps in ETL-Synthea and the auto-generation of primary keys in both projects.

We were also able to demonstrate dbt's powerful documentation capabilities by generating a local documentation website for dbt-synthea. With the run of a single command, dbt generates a site that includes detailed model-level descriptions; code previews; and an interactive lineage diagram of the entire ETL DAG.

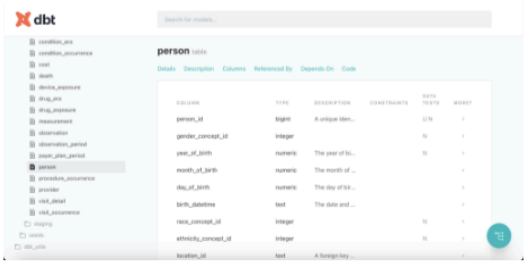

Figure 1. Documentation page for PERSON

Figure 2. Lineage graph for OBSERVATION_PERIOD

Conclusion

dbt is a powerful data transformation tool well-suited to the OMOP ETL use case. Its support for transparency, accessibility, and open-source collaboration are aligned with OHDSI principles. With dbt-synthea, we have demonstrated that it is technically feasible to perform an OMOP ETL using

dbt and have done so in a way that allows users to easily replicate and expand upon our work.

Our work on dbt-synthea is not done. Before moving onto Phase II of the project, we aim to better assess dbt's cross-database support features; our initial work revealed dbt's cross-database macros library to be quite immature. We will also evaluate newer tools like SQLMesh, a dbt derivative which natively supports cross-dialect SQL transpilation.[12]

We additionally aim to explore performance and scalability; expansion of built-in tests to include CDM/THEMIS conventions and DataQualityDashboard checks; and incorporation of source-specific vocabularies. We welcome collaborators to join us in our journey to build this essential resource for the OHDSI community!

References

1. dbt Labs Builds Momentum as the Industry Standard for Data Transformation [Internet]. dbt Labs. [cited 2024 Jun 12]. Available from: https://www.getdbt.com/blog/dbt-labs-builds-momentum-as-the-industry-standard-for-data-transformation
2. CaRROT-CDM [Internet]. University of Nottingham Health Informatics; Available from: https://carrot4omop.ac.uk/CaRROT-CDM/ETL/
3. Rabbit-in-a-blender [Internet]. RADar, AZ Delta; Available from: https://github.com/RADar-AZDelta/Rabbit-in-a-Blender
4. Perseus [Internet]. OHDSI; Available from: https://github.com/OHDSI/Perseus
5. Ashcroft Q, Kirkwood D, Howcroft T, Knight J, Dobson S, Chandrabalan V. Implementing the OMOP common data model using dbt. In: OHDSI Global Symposium Collaborator Showcase [Internet]. 2023. Available from: https://www.ohdsi.org/wp-content/uploads/2023/10/20-Ashcroft-BriefReport.pdf
6. Pitchayarat T, Pinyo G, Tanchotsrinon W, Khamsrimuang S, Issarasittiphap C, Bootnumpech C, Siangchin N, Promma K, Bovornmongkolsak N, Suriyaphol P, Adulyanukosol N. Using dbt—a free and open-source software framework— to transform data into OMOP CDM in the ETL process. In: OHDSI Global Symposium Collaborator Showcase [Internet]. 2022. Available from: https://www.ohdsi.org/wp-content/uploads/2022/10/2-Pitchayarat-abstract.pdf
7. Carlson R, Phad M, Martin S. Moving OMOP to the cloud with DBT and Snowflake (All of Us Research Program). In: OHDSI Global Symposium Collaborator Showcase [Internet]. 2022. Available from: https://www.ohdsi.org/2022showcase-47/
8. ETL-Synthea [Internet]. OHDSI; Available from: https://github.com/OHDSI/ETL-Synthea
9. SYNTHEA [Internet]. MITRE; Available from: https://synthetichealth.github.io/synthea/
10. How we structure our dbt projects [Internet]. dbt Labs. [cited 2024 Jun 19]. Available from: https://docs.getdbt.com/best-practices/how-we-structure/1-guide-overview
11. Scheumie M. FilterVocabulary.R [Internet]. OHDSI. Available from: https://github.com/OHDSI/Tutorial-Hades/blob/main/extras/FilterVocabulary.R
12. SqlMesh [Internet]. Tobiko Data. Available from: https://sqlmesh.readthedocs.io/en/stable/