Framework and Implementation of an OMOP-Oriented Clinical Data Warehouse Using Databricks

Jared Houghtaling[a], Kyrylo Simonov[a], Kyle Zollo-Venecek[a], Elina Hadelia[a], Manlik Kwong[a], Polina Talapova[a], Clark Evans[a], Robert Miller[a], Andrew E. Williams[a]

[a] Tufts Medicine, Clinical and Translational Science Institute (CTSI)

Background:

The Clinical and Translational Science Institute (CTSI) at Tufts Medical Center was founded in 2008 and has since received four consecutive Clinical and Translational Science Awards (CTSA) from the National Institutes of Health (NIH) in support of efforts to innovate across the health data space. One core component of these efforts is the creation and maintenance of a Tufts Research Data Warehouse (TRDW) that captures and consolidates rich observational health data from different clinical data acquisition systems across Tufts Medicine's three hospitals, 40-practice physician network, and home health care organization. Once fully implemented, the TRDW will facilitate the following: (1) efficient and accurate handling of medical data requests in support of clinical research across Tufts Medicine, Tufts University, and Tufts CTSI's large set of affiliate and partner institutions (2) participation in broad consortia – such as the OHDSI Data Network, Bridge2AI **[1]**, ACT, TriNetX, N3C **[2, 3]**, and CRITICAL **[4]** – on the cutting edge of clinical informatics and medical machine learning (ML), (3) advancement and utilization of powerful open-source tooling pioneered by the Observational Health and Data Sciences and Informatics (OHDSI) collaborative, and (4) novel approaches to integrate unstructured and semi-structured data (e.g. waveforms, free text, images, flowsheets, genomic tumor profiles) with a structured relational framework like the Observational Medical Outcomes Partnership (OMOP) common data model (CDM). This latter integration will enable broad and deep phenotyping and linking to knowledge graphs for translational research on biological causes of disease **[5]** and alignment of health systems' data definitions and support infrastructure through FHIR. In this work, we outline efforts to achieve each of these functionalities and give particular attention to the architectural design and implementation of the various data transformations, as well as to the tooling for interacting with the resulting data output. We place our efforts in a broader context, describing both the highlights and the pitfalls we have encountered thus far, and providing feedback to those with similar ambitions.

Methods:

The TRDW is hosted on Azure cloud and incorporates a range of containerized services. All services are managed via Terraform **[6]** and ephemeral processes - like rendering notebooks, executing transformations, or producing extracts - are orchestrated using Prefect **[7]**, which ultimately converts Python functions into parallelized, executable jobs that can be scheduled at

regular intervals or launched manually from a web interface; an overview of these services is shown below in **Figure 1**. We use GitHub as the first step in our continuous integration/continuous delivery (CI/CD) process, whereby any updates to configuration files, orchestration scripts, transformation logic, or analytic workflows within a given repository can trigger a cascade of deployment steps that result in a relatively seamless update for end users.

We designed the Extract-Transform-Load (ETL) processes using a combination of languages (python, R, julia) with a fundamental goal of producing a thoroughly documented, highly transparent, and easily maintainable codebase. Transformations are stratified by OMOP table, and further sub stratified by data source; each transformation includes associated SQL operations that accompany an interactive notebook with logic documentation, analytics, and unit tests. The core methods in the processing layer of the pipeline (**Figure 2**) are scheduled to run nightly, executing a full reload of the key relational source data in less than an hour. Note that processing pipelines for waveform signals, images, and free-text notes can require significant computational time, and the associated data enter the TRDW through different routes and at different frequencies than the bulk electronic health records. We currently transform these data *ad hoc* and independently from the other transformations, but we will eventually schedule them in parallel with the core ETL in a Delta-load format after evaluating general requirements for data refresh frequency. Once transformed, we organize the various sources in an instance architecture, in which independent OMOP instances represent specific sources (e.g. devices, pre-hospital care data, veterinary info) or combinations thereof, and with a single merged OMOP dataset that integrates, matches, and deduplicates data from all other instances.

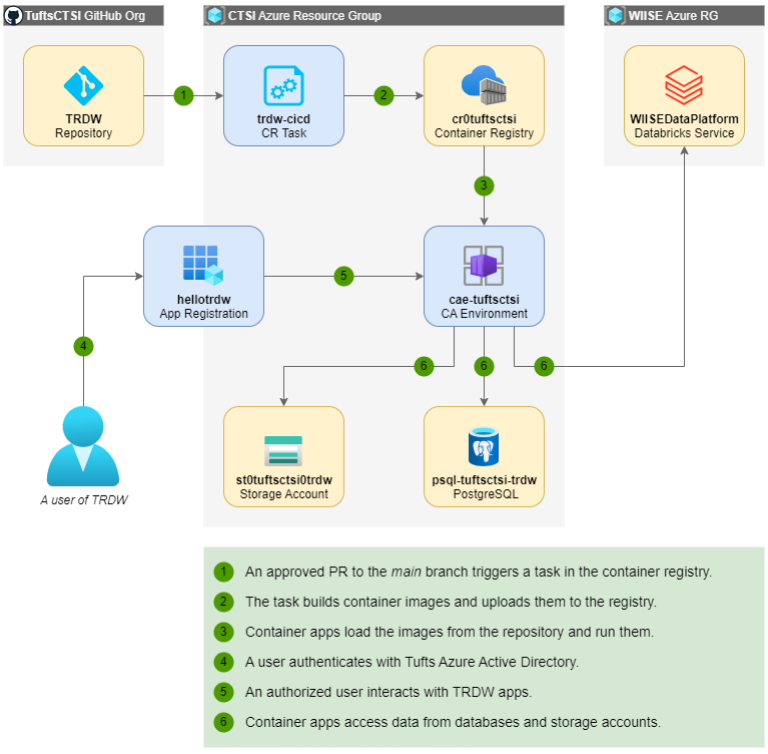

Figure 1: Overview of Terraform-based infrastructure deployment that handles user access, clinical data transformations, and wide-ranging analytics. We are currently collaborating with the WIISE data team at Tufts Medical Center on multiple aspects of the Databricks and EPIC integrations.

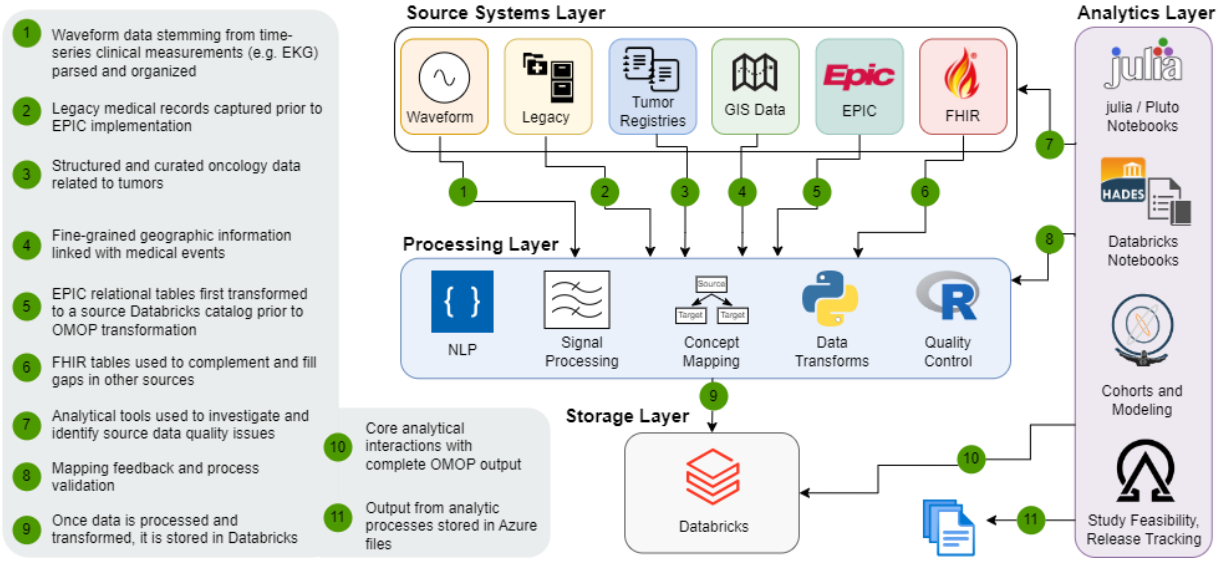

Figure 2: Overview of data pipeline and associated services to analyze and extract insights from the resulting output.

Results:

We are still in the proof-of-concept (POC) stage of defining and implementing a fully operational clinical data warehouse that supports each of the functionalities described above. One major challenge in any large-scale effort involving sensitive clinical data is governing authentication and authorization; here, we are implementing newly offered features from Databricks (e.g. Delta Share) that enable highly granular permission structures for data access. We are using these features, together with a modular notebook-based approach, to provide researchers and clinicians with configurable tooling that allows them to interact with precisely the data they request and are permitted to view.

The framework itself is highly scalable in terms of data (e.g. quantity, diversity), computational performance, and services provided. New tools continue to be released from the OHDSI community that enable research collaboration in a federated manner, and we are taking an active role in contributing to those tools to ensure they are compatible with Databricks/Spark data structures as their adoption continues to grow. The TRDW supports approximately 100 research requests from more than 50 distinct users annually and produces project-specific data extracts for six research platforms; we maintain the infrastructure with a team of four engineers and expect to extend services to more than 20 concurrent users – most of whom are clinical researchers - in the coming months.

We are currently evaluating approaches to integrate high-density data forms into our processing pipeline and subsequent informatics work. One promising method for parsing and storing tremendous volumes of physiological time-series data, proposed by Goodwin *et al.* and now packaged into a product called Atrium DB **[8]**, may address some of our needs, although practical challenges remain.

Conclusion:

The TRDW currently serves as a sandbox for securely and efficiently interacting with multiple rich OMOP datasets as well as with diverse data types. We expect that in the months to come it will enable the construction and implementation of sophisticated statistical models based on multimodal data. Such modeling approaches have considerable potential when paired with federated, real-world evidence studies like those recently initiated between the Darwin-EU project consortium and European Medicines Agency **[9]**; they also have potential to support translational engineering and design of clinical decision support (CDS) tools with bed-side impact **[10]**. Lastly, much of the work presented above builds on the effort and dedication of so many others in the OHDSI community; we will continue to contribute to - and advocate for - open-source development of these powerful tools, and we plan to share our efforts and experiences along the way.

References:

[1] Xu, Huimin, et al. "Cross-Team Collaboration and Diversity in the Bridge2AI Project." *Companion Proceedings of the ACM Web Conference 2023*. 2023.

[2] Haendel, Melissa A., et al. "The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment." *Journal of the American Medical Informatics Association* 28.3 (2021): 427-443.

[3] "National COVID Cohort Collaborative (N3C)", June 2023, https://covid.cd2h.org

[4] "CRITICAL Consortium", June 2023, https://critical-consortium.github.io/

[5] Callahan, Tiffany J., et al. "Ontologizing health systems data at scale: making translational discovery a reality." *NPJ Digital Medicine* 6.1 (2023): 89.

[6] Brikman, Yevgeniy. *Terraform: Up and Running*. " O'Reilly Media, Inc.", 2022.

[7] "Prefect Orchestration Software", June 2023, https://www.prefect.io/

[8] Goodwin, Andrew J., et al. "A practical approach to storage and retrieval of high-frequency physiological signals." Physiological measurement 41.3 (2020): 035008.

[9] Arlett, P., Kjær, J., Broich, K., & Cooke, E. (2022). Real-world evidence in EU medicines regulation: enabling use and establishing value. Clinical Pharmacology and Therapeutics, 111(1), 21.

[10] Lin, V., et al. "Training prediction models for individual risk assessment of postoperative complications after surgery for colorectal cancer." Techniques in Coloproctology 26.8 (2022): 665-675.