Stanford MEDICINE | Technology & Digital Solutions
STANFORD HEALTH CARE & SCHOOL OF MEDICINE

TDS Connection

THE LATEST FOR TDS PARTNERS | *May 2024*

FEATURED STORY

Streamlining Adoption of Research Results with Databricks

*By **Lisa Tsering**, Internal Communications Specialist, TDS*

Stanford has always been at the cutting edge of research and of building innovative technologies. One of the mandates for the newly formed Stanford Health Care Data Science team is to build a framework to assess the feasibility of machine learning–guided workflows in the hospital. Among the Data Science team's many goals, one is to fast-track AI innovations that happen at the School of Medicine.

Toward this specific goal, the Data Science team has developed a collaboration with the TDS Research Technology, or RT, team. In a recent study titled "Enabling Innovation at the Bedside Using STARR-OMOP," the two teams describe research studies originally done using STARR-OMOP on Google Cloud Platform that have led to applications that are providing real benefit at the point of clinical care.

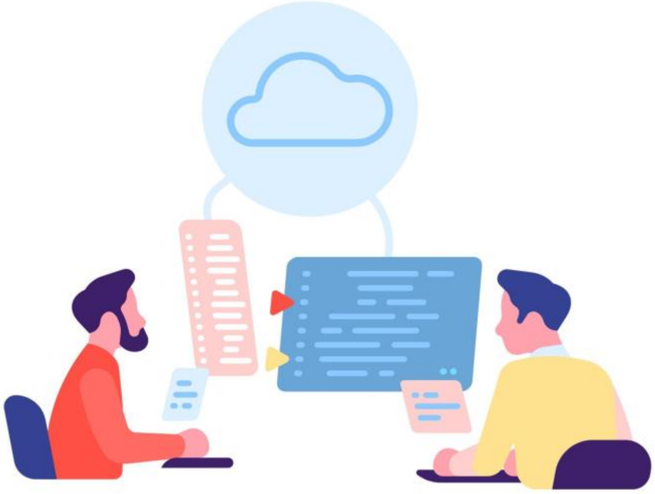
The primary motivation for the OMOP normalized and harmonized data model is to have a common "vocabulary" and analysis methods to enable and encourage global cross-institutional clinical research. The study, one of the earliest such studies in the field, demonstrates that the OMOP data model is well suited to hospital bedside applications. However, **Priya Desai**, lead author of the above study and RT product lead for STARR-OMOP, laments that while access to good quality research data has meant more scientific publications and an advancement in our understanding of disease and methods of prevention, the pace of adopting clinical applications of basic science discoveries has continued to be slower than what the Stanford Medicine community desires.

OMOP: Observational Medical Outcomes Partnership

STARR: STanford Medicine Research data Repository

To accelerate the path between the bench and the bedside, like the study described above, the two TDS teams, RT and Data Science, in collaboration with University IT, have implemented their first Databricks environment within the Stanford University Google Cloud platform. SHC has adopted Databricks for some of its data warehousing, AI/ML, and data science workloads. Now, with an integrative approach and rapid translation in mind, RT brings Databricks capabilities to the School of Medicine. **Garrick Olson**, TDS Infrastructure and Architecture Lead, describes how we've streamlined the process of sharing data across Stanford Medicine.

Q: What is Databricks and what benefits does it bring to Stanford?



A: Databricks is a commercial product designed to store, transform, and analyze large amounts of data efficiently. It offers a user-friendly notebook experience for working with the data, and powerful underlying capabilities based on Spark/SQL, Delta Lake, PyTorch, TensorFlow, MLflow, etc. It has a strong governance and security model, and good multi-cloud support to facilitate working with data in multiple locations and organizations.

The SHC Databricks instance is hosted in our Microsoft Azure infrastructure, while the School of Medicine Databricks instance is hosted in our Google Cloud infrastructure. As a result, we rely on the multi-cloud architecture, sharing features, and governance model for our solution.

The data science team needs considerable amounts of hospital data for their work in operationalizing data science and AI models. In addition to creating new algorithms and models, they shepherd algorithms and models Stanford researchers develop into the hospital context. This requires being able to reproduce the research results on data similar to that originally used, for understanding, validating, tuning, and further development.

The researchers at the school often train algorithms using the STARR-OMOP which contains data from the adult hospital along with data from Stanford Medicine Children's Hospital and other sources. We don’t currently have approvals for the adult hospital to work with all STARR data, so we created a version of the STARR data warehouse that has only SHC data in it. This data goes into Databricks so it can be used with other SHC data needed for the data science projects.

Q: What are a couple of examples of the types of data within Databricks?

A: The STARR data we made available in Databricks is a translation of patient clinical records from Epic into the OMOP Common Data Model. This is a community standard, normalized model for patient data that is designed to make it easier to use for research. It is now widely used at many institutions, facilitating collaboration and replicability. The research technology team has invested a tremendous amount of effort over the years to create and continuously improve the quality and completeness of how we transform our clinical data into this format. In fact, in an earlier newsletter, Priya presents this multi-year journey.

It is important to make the OMOP asset available for our data science team because it makes them more productive and effective as they work with researcher-developed algorithms in the same way as the researchers. Doing so in the Databricks environment reduces the technology friction.

> “We solved for the immediate needs, but we also built out a resource I expect will become a valuable tool more broadly for researchers at the school.”
> *Garrick Olson, TDS Infrastructure and Architecture Lead*

Q: Is it related to Cosmos?

A: No, Cosmos data is held by Epic and you can only access it from within the environment they provide. You can’t pull data out of Cosmos and put it into Databricks or other systems. Cosmos is obviously of great interest for data science and other work, but I expect it will remain separate from our Databricks infrastructure.

Q: What kind of challenges did you and your team face when putting this together?

A: Purchasing and licensing Databricks, and getting all the legal agreements, reviews, and approvals in place. The way it works within Google Cloud is through the third-party marketplace, so the relationships and agreements got a bit complicated. We took care as we went through these processes to make sure we approached it as platform infrastructure so we can use more generally beyond the immediate needs of our teams.

There were a lot of unexpected technical challenges in terms of getting our data into this environment, setting up the sharing mechanisms between hospital and school environments, and working through subtleties in the data formats and conversions, so we spent significant time working on the technical integration.

Now that we have overcome the challenges and are operational, I am happy with the solution. We solved for the immediate needs, but we also built out a resource I expect will become a valuable tool more broadly for researchers at the school.