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

TDS Connection

THE LATEST FOR TDS PARTNERS | *July 2024*

FEATURED STORY

STARR Data Lake Expansion Enables Access to Research Data

*By **Deepa Balraj**, Senior Software Engineer; **Joseph Mesterhazy**, Manager, Big Data Engineering; and **Garrick Olson**, Infrastructure and Architecture Lead (TDS Research Technology Engineering Services); **Donald Mitchell**, Director, Academic Application Services; **John Maul**, Senior Systems Analyst; and **Amanda Unrue**, Business/Applications/Systems Analyst (OnCore); and **A. Solomon Henry**, Senior Software Developer and **Douglas J. Wood**, Senior Software Developer (Stanford Cancer Institute Research Database)*

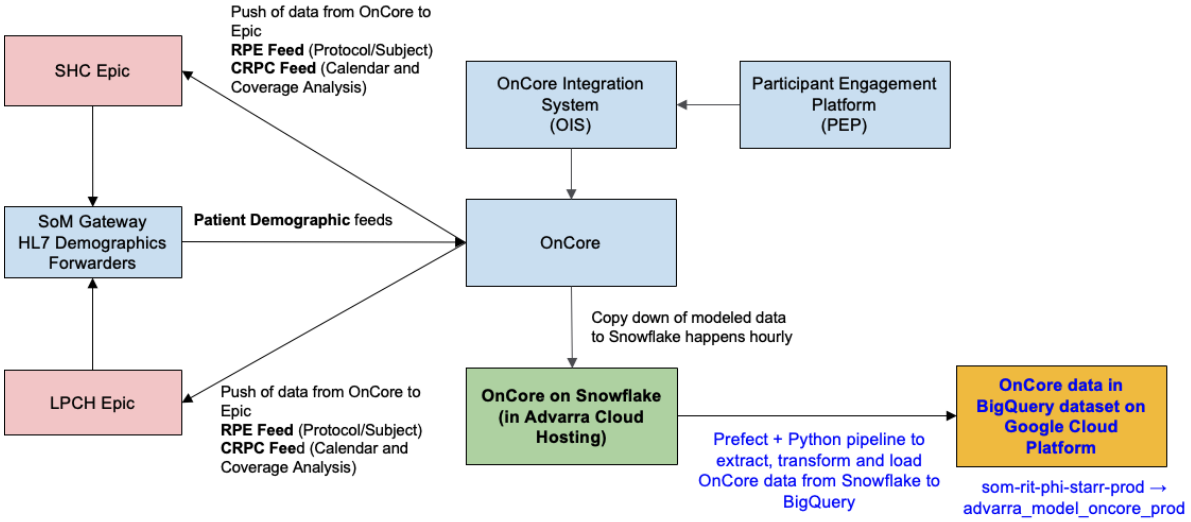
A recent collaborative effort among three teams at Stanford Health Care aims to enhance the efficiency of drawing research conclusions, promising to expedite responses to research inquiries and accelerate dissemination of medical knowledge to the public.

The STARR (STAnford Research Repository) data lake in BigQuery on the Google Cloud Platform is the data resource developed by the Research Technology team that is designed to improve researchers' access to Electronic Health Record data from our hospitals and clinics. This includes raw data, analysis ready data, linked data across different data modalities, support for different data models, multiple clinical data warehouses, data search and access tools, data de-identification pipelines, concierge services, training, and documentation.

While Stanford’s oncology data is extensive, it is currently dispersed across various sources, so the oncology data lake in STARR will consolidate Stanford’s oncology and cancer data from these diverse sources into a single repository in BigQuery/GCP.

A pipeline to bring in OnCore data into the STARR oncology data lake has been completed; this pipeline runs on the first day of every month.

WORKFLOW DIAGRAM


SHC Epic
Push of data from OnCore to Epic
RPE Feed (Protocol/Subject)
CRPC Feed (Calendar and Coverage Analysis)
OnCore Integration System (OIS)
Participant Engagement Platform (PEP)
SoM Gateway HL7 Demographics Forwarders
Patient Demographic feeds
OnCore
Copy down of modeled data to Snowflake happens hourly
LPCH Epic
Push of data from OnCore to Epic
RPE Feed (Protocol/Subject)
CRPC Feed (Calendar and Coverage Analysis)
OnCore on Snowflake (in Advarra Cloud Hosting)
Prefect + Python pipeline to extract, transform and load OnCore data from Snowflake to BigQuery
OnCore data in BigQuery dataset on Google Cloud Platform
som-rit-phi-starr-prod → advarra_model_oncore_prod

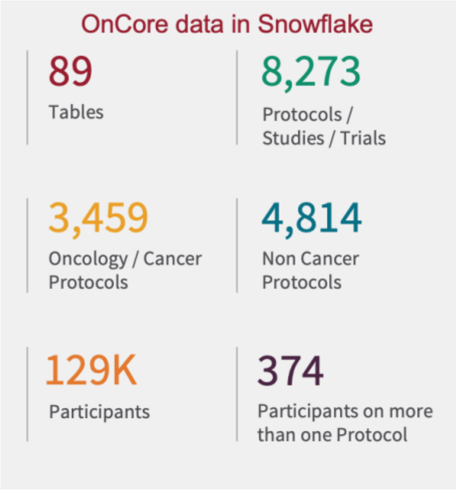
OnCore is the Clinical Trial and Research Management System used by Stanford Medicine, and is our secure centralized system for tracking clinical research. OnCore data is now being loaded hourly into a reporting database on a cloud-based Snowflake server. This data is a transformed version of the transactional data modeled for reporting purposes. OnCore has cancer and non-cancer clinical trial data which includes protocols/study/trial, participants, patients on studies, biospecimen data, and other information.

As part of the initial use case for the oncology data lake in STARR, we brought in OnCore data from Snowflake into a BigQuery dataset in GCP.

Ahead of bringing this data into STARR, there was another significant project where the OnCore was migrated to cloud by the Academic Application Services team. This migration itself is part of a larger project to enable the financial modules in OnCore to streamline the billing processes and integrate the Clinical Trial management system with Epic to make billing research participants more streamlined and efficient. To enable all the financial features, we needed to migrate to the cloud version of the OnCore product. This required us to rebuild all existing integrations with Stanford systems and point them to the new instance of OnCore in the cloud.

Benefits:

- Accelerate oncology research by compiling cancer data from different data sources into one place in BigQuery within the GCP.
- Bring OnCore data into the STARR data lake to allow us to easily combine this data with other research and Epic data for analysis.



OnCore data in Snowflake

89 Tables	8,273 Protocols / Studies / Trials
3,459 Oncology / Cancer Protocols	4,814 Non Cancer Protocols
129K Participants	374 Participants on more than one Protocol

By integrating OnCore data into the STARR data lake, researchers will have enhanced access to clinical trial information for both cancer and non-cancer studies. This data can be seamlessly combined with information from diverse sources like Epic, empowering researchers to address their inquiries more effectively.

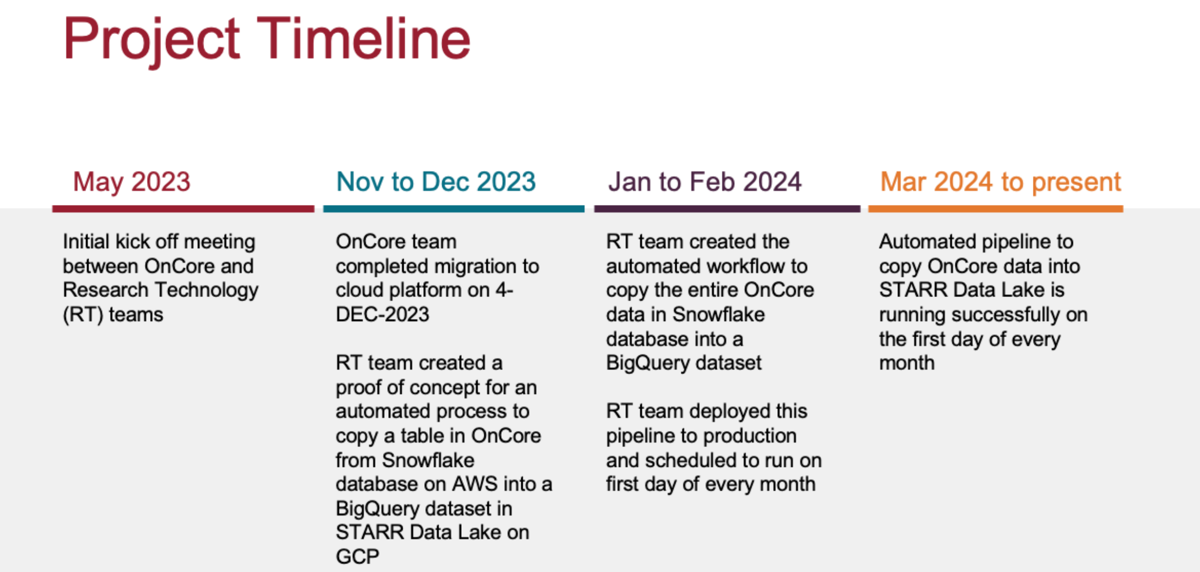
The data can also enhance our internal tools. For example, SCIRDB (the Stanford Cancer Institute Research Database, managed by the Data Coordinating Center, now part of Research Technology) aims to leverage data from the OnCore system. The DCC team produces patient lists for clinical trial recruitment, which are then used by clinical trial coordinators. A subset of these patients, who meet the trial's inclusion and exclusion criteria, are ultimately recruited. Feedback from the clinical trial coordinator on why a patient is excluded is essential but time-consuming — so instead, OnCore data is utilized for this feedback. The DCC uses the recruited patients from the clinical trial study as a validation dataset to refine the decision logic for creating patient recruitment lists.

The project promotes Stanford’s mission of research, education, and clinical care by improving accessibility to data — which enhances our efficiency in addressing research queries, consequently enhancing our research output.

Overcoming Challenges

Since OnCore Snowflake is a new data source, we needed to become familiar with ways we could export and transform the data. Some of the data was not handled consistently in the OnCore system, which made mapping the data with consistency into STARR challenging and an extra, unanticipated step. We also needed to transform some data types to different types as we migrated the data, which sometimes created some complexity.

> “As someone accustomed to working with Epic Clarity data, delving into a new realm of medical data was both enjoyable and enlightening.”
> *Deepa Balraj, Senior Software Engineer, TDS Research Technology Engineering Services*



Project Timeline

May 2023	Nov to Dec 2023	Jan to Feb 2024	Mar 2024 to present
Initial kick off meeting between OnCore and Research Technology (RT) teams	OnCore team completed migration to cloud platform on 4-DEC-2023 RT team created a proof of concept for an automated process to copy a table in OnCore from Snowflake database on AWS into a BigQuery dataset in STARR Data Lake on GCP	RT team created the automated workflow to copy the entire OnCore data in Snowflake database into a BigQuery dataset RT team deployed this pipeline to production and scheduled to run on first day of every month	Automated pipeline to copy OnCore data into STARR Data Lake is running successfully on the first day of every month

TDS Collaborations

The project to bring OnCore data into the STARR oncology data lake required close collaborations between many individuals from the Research Technology team led by Somalee Datta; the Academic Application Services team led by Donald Mitchell; the Advarra (vendor) team; and the Data Coordinating Center team, who specializes in oncology/cancer research.

The culmination of collaborative efforts and technological advancements underscores Stanford Health Care's ongoing dedication to advancing medical research and improving patient outcomes.