FuncX in NIH’s Common Fund Data Ecosystem (CFDE) Portal

Lee Liming
Director, Professional Services
Globus @ University of Chicago
lliming@uchicago.edu
NIH Common Fund Data Ecosystem (CFDE)

- The Common Fund Data Ecosystem (CFDE) is an attempt to standardize, simplify, and enhance researchers’ experience with NIH’s Common Fund data, much like...
  - NASA EOS portal (Earth observational data)
  - USGS ScienceBase (geological data)
  - ACCESS (NSF supercomputing centers)
- NIH’s Common Fund sponsors several dozen big programs (each with dozens of individual awards) that produce valuable medical data.
  - Each program has one (or more) Data Coordinating Centers (DCCs) that collect and manage the data from the all the projects in the program.
  - Each DCC is essentially a data silo.
CFDE Portal - Discover data from the NIH Common Fund

...but how does CFDE know what each DCC has?

https://app.nih-cfde.org/
Updating the CFDE Portal’s copy of a Data Coordinating Center’s holdings: Beginning to End

A DCC bioinformatician creates a set of metadata files representing the DCC’s complete data holdings. This is a heavily CLI-based process.

Getting the submission files from the DCC’s computer to CFDE’s portal in AWS is a fully-automated process, managed by a Globus Flow, initiated by a command on the DCC computer.

After the submission is present in the CFDE Portal, curation by authorized DCC and CFDE personnel is performed using the portal’s web interface.
**FuncX in the CFDE submission flow**

The CFDE Submit function executes Python code provided by the Deriva team that reads the TSV files in the submission and ingests them into Deriva’s submission registry.

The FAIR Assess function assesses the metadata provided in the submission on FAIR principles and adds the resulting metric scores to the submission in the registry for use by reviewers.

Both funcX functions are memory-intensive, and submission traffic is bursty around submission deadlines, so we use an AWS ParallelCluster to handle multiple submissions in parallel if needed and avoid paying for idle resources.