The Agenda

● Introduction

● Gen3 Roadmap
  ○ Open source support
  ○ Frontend framework
  ○ Deployment improvements
  ○ Workflow execution in workspaces (Nextflow)
  ○ Large language models for enhanced search
  ○ Data lakehouse improvements
  ○ Other improvements

● Steering Committee Discussion

● Q&A
Gen3 Product Roadmap
1. Gen3 Open-Source Support

- Dedicated support for external contributions to Gen3 codebase including timely review of PRs
- Allow for external contributors to test PRs using our integration test suite (to run on your infrastructure)
- Allow issue reporting and tracking
- **Will provide more details in next Gen3 Community Forum**
2. Gen3 Frontend Framework & Analysis Tool Framework

- Replacement for Gen3 Data Portal
- Moving to “App Store” like framework for analysis and visualization tools
- Updated technology stack
- Improved development and user experience
- Simplifies project specific content and customizations
- Beta launch at the end of January
Analysis Tool Framework (ATF)

- Supports the development of custom analysis tools
- Connects to:
  - Gen3 services
  - 3rd-party APIs
  - Other data sources
- Uses context to filter tools that can be applied to current data selection
- New Gen3 Frontend:
  - Redesigned standard pages: Explorer, Discovery…
  - Analysis tool center for accessing applications
  - Commons specific functionality/components (for use in tables, charts)
  - More integration/data sharing between applications
3. Deployment Improvements

- In 2023, Helm was rolled out and can now be used to deploy Gen3.
- In 2024, CTDS will begin to use Helm for production deployments. Also planned for 2024:
  - Full test suite w/ helm - Getting a full testing suite running against Helm deployments, which will be available to external users.
  - Observability - Incorporating observability tools in the Gen3 Helm charts – better monitoring and logging, and general visibility into the health of the deployments.
- Gen3 Lite - Run a lightweight Gen3 on a single instance. Cheaper and great for non-production environments.
4. Workflow execution in workspaces (Nextflow)

- A general purpose workflow execution system in Gen3 that researchers can use to run *containers* on the cloud for various applications in a secure and isolated manner
- Phase 1
  - Develop and test infrastructure on containers developed by users and run by Gen3 operator - Complete
- Phase 2
  - Develop and test infrastructure with CLI push credentials for containers developed by users and run by users - Currently being tested.
- Phase 3
  - Develop a friendly portal for users to submit containers, track jobs
5. Large Language Models for Enhanced Search

- Natural language queries and responses of data on the Discovery page
- Will use a RAG approach initially (Retrieval Augmented Generation) and in the future may involve training a new model or fine-tuning an existing model
- Planning to include an API for vectorbases to support RAG
6. Gen3 Data Lakehouse Improvements

- Allow data files to be shared through Gen3 either without or before creation of a data dictionary and population of the data model
- Can associate searchable metadata with files or groups of files and projects (via Gen3 Metadata API)
- Adding a per-user data library and updating pages in Gen3 to improve ability to access files before metadata harmonization
- Groups of files and/or a data dictionary could also be distributed in the data lake
  - Will support a packaged file format like Portable Format for Bioinformatics (PFB)
    - PFB includes the structured data (i.e. graph model) and pointers to data files all in one
  - Will also support a .zip or other combination of file formats
7. Other improvements

- Frictionless
  - Export Gen3 data into Frictionless format
  - Import metadata from frictionless data packages
  - Adding tools in workspaces for working with frictionless data packages
- We plan to release an annual Gen3 product roadmap
Gen3 Panel Discussion
Gen3 Steering Committee

- Robert Grossman - University of Chicago
- Claire Rye - New Zealand eScience Infrastructure
- Steven Manos - Australian BioCommons
- Plamen Martinov - Open Commons Consortium
- Kyle Ellrott - Oregon Health and Science University
Questions