Cloud Workspace Pilot: A Practical Hands-On Introduction to the CFDE Portal and CAVATICA Data Analysis Platform

Purpose: Deliver working knowledge of CFDE Portal and CAVATICA Data Analysis Platform to CFDE all-hands workshop attendees.
Duration: 45 minutes (5 min Accessing Data Sources, 25 min. Platform, 10 min. Questions)
Instructors: Sangeeta Shukla (CHOP/Kids First: Data) and Eric Tobin (Velsera: CAVATICA)

Suggested Prerequisites

The suggested prerequisites below are designed to help attendees optimize their learning during the workshop, however, are not required. All interested attendees are welcome.

CAVATICA

- Distribute the User Resources guide and ask attendees to examine before the conference.
    - Information on Knowledge Center Guides, Onboarding videos, Research examples (webinars), Public Projects, and Support. Even without completing each subtask before the training, reading this document will establish at least an orientation to the platform.
- Gather usernames for registered attendees by March 1, 2024; apply pilot funds by March 15, 2024. This will encourage users to actively participate. Note: attendees are also welcome after this date to submit responses to the Optional Pre- and Post-Workshop Survey.

Overall Learning Objectives

1. Accessing Data Sources in CAVATICA
    a. Highlight the known diversity of CFDE DCC data sources and access modalities.
    b. Understand that CAVATICA enables you to bring in your own data for analysis in diverse ways–including DRS.
2. CAVATICA Platform
    a. Understand the basics of setting up and managing a CAVATICA account
    b. Familiarize with key features of CAVATICA for data analysis and project management
    c. Cloud Workspace Pilot (CWP)
        i. The CWP is a limited engagement initiative to promote the use of cloud-based data and analysis resources on CFDE data.
        ii. There is tiered funding available that is very accessible
            1. Pilot funds ($100): Email support
            2. CWP tier 1 ($1,000): Letter of intent and budget estimate
            3. CWP tier 2 ($5,000): Process initiated by closeout of CWP tier 1 project

Accessing Data Sources (5 minutes):

Setting the Stage: Exploring cloud-based solutions in CAVATICA for managing access to the diversity of CFDE DCC data modalities.

1. CFDE DCCs and the diversity of data and data access modalities.
2. Big Data Sourcing and Management within CAVATICA.

CAVATICA Platform (25 minutes):

1. Account creation eRA Linking (3mins)
    a. Concepts: Create a CAVATICA account, update account + check dataset access, billing group location and breakdowns
        i. Show login page + account creation in incognito, demo spun up
    b. Competencies: Account management
        i. How to investigate billing group, where to find eRA upgrade and dataset access
2. Project creation and management (5mins)
    a. Concepts: Projects as fundamental research unit, Billing and spot instances, collaboration within projects, project navigation
        i. Create a project
            1. Details: allow network access, cloud location, name vs url, spot instances, work reuse, billing group
        ii. Navigate a project
            1. Details: project vs platform toolbar, Project notebook, collaboration, Dash/Files/Apps/Tasks/Data Studio BRIEF mention,
    b. Competencies: Project Setting management, manage project members, navigate within and among Projects
        i. Explore project options
            1. Details: Settings for deletion/billing group change/lock down/etc., show addition and removal of colleagues, compare two projects and refer back to toolbars
3. Data management (7mins)
    a. Concepts: Data ingestion methods, Symbolic links, Cloud data storage RE: interoperability
        i. Demonstrate data upload methods, specifically DRS (link and manifest); FTP/HTTP(S); upload from local; interoperability
            1. Details: Explain that DRS is a file's "true" location address and can be linked with metadata and authorization (FHIR/PFB + GA4GH), import all GMKF bam files via manifest; demonstrate adding real CFDE DCC data to project via HTTP(S); add DCC data from computer by drag and drop; BDC+CGC+GMKF Portal
    b. Competencies: Export and Locate Data from KF Portal, Search and filter data in a project, apply tags to data, access data between projects

        i. Explore KF portal and import all bam files, add tags
            1. Details: Select cohort on KF and use “Export to SB” button, filter by tags in discussion and apply win ingress, point out between project data sharing option for data upload when discussing others
4. Apps + Tasks overview (12mins)
    a. Concepts: CWL, Public Apps Gallery, Apps vs. Tasks
        i. Describe CWL in brief
            1. Details: “get along shirt” for different languages, cloud agnostic, no coding needed (but available), can develop your own app if necessary
        ii. Explore Tools and Workflows
            1. Details: Show public apps gallery and add DESEQ2 and [TBD] workflow to project, discuss Tool description page (key points: benchmarking, expected inputs/outputs, versions, copying, editing)
        iii. Demonstrate a task
            1. Details:
    b. Competencies: Copy App, Run Task, Locate Results, Edit and Rerun
    c. Short demo: Edit and rerun, show editable params and inputs
        i. Bulk RNA-Seq Public Project
            1. Detailed KC Walkthrough
5. Data Studio overview (5mins)
    a. Concept: Virtual computer, R+Python+Custom; platform vs. notebook timeout
    b. Competencies: Create Data Studio, edit settings, input/output/save
    c. Short demo (options):
        i. Data Studio Interactive Analyses Public Project (Multiple options)
            1. Ballgown
            2. Single Cell RNA Seq
        ii. Data Interoperability Public Project (JSON import of data from AnVIL)
        iii. Custom notebook (basic stats and R/Python running environments)

Questions (10 Minutes):

Solicit questions for CFDE Cloud Workspace Pilot (CWP) and CAVATICA