



National
COVID
Cohort
Collaborative

Introduction to Analyzing Real-World Data Using the National COVID Cohort Collaborative (N3C) Spring 2024 - N3C Education and Training Domain Team

Session 1, Assignment 1

Enclave Navigation, Submitting a Data Use Request

DUE: Jan. 25, 2024

Pre-requisites

To complete this assignment, you will need to:

- Be able to log into the N3C Enclave at <https://unite.nih.gov>
- Have completed your Human Subjects Research Protection Training within the past 3 years, and have the date of your completion on hand. More information [here](#).

Part 1 - Workspaces, Folders, Favorites

After logging into the enclave, you'll see the enclave homepage with quick links to various resources. Start by browsing the **Files** link in the left menu bar:

National COVID Cohort Collaborative Data Enclave
Discover, collaborate, and analyze data from the N3C.

Welcome to N3C, Shawn

Educational Resources

- Training material
- N3C Community Notes
- Results Download
- N3C Assist

21,704,702
TOTAL N3C PATIENTS

8,463,370
CONFIRMED COVID-19 (+)

220,651
POSSIBLE COVID-19 (+)

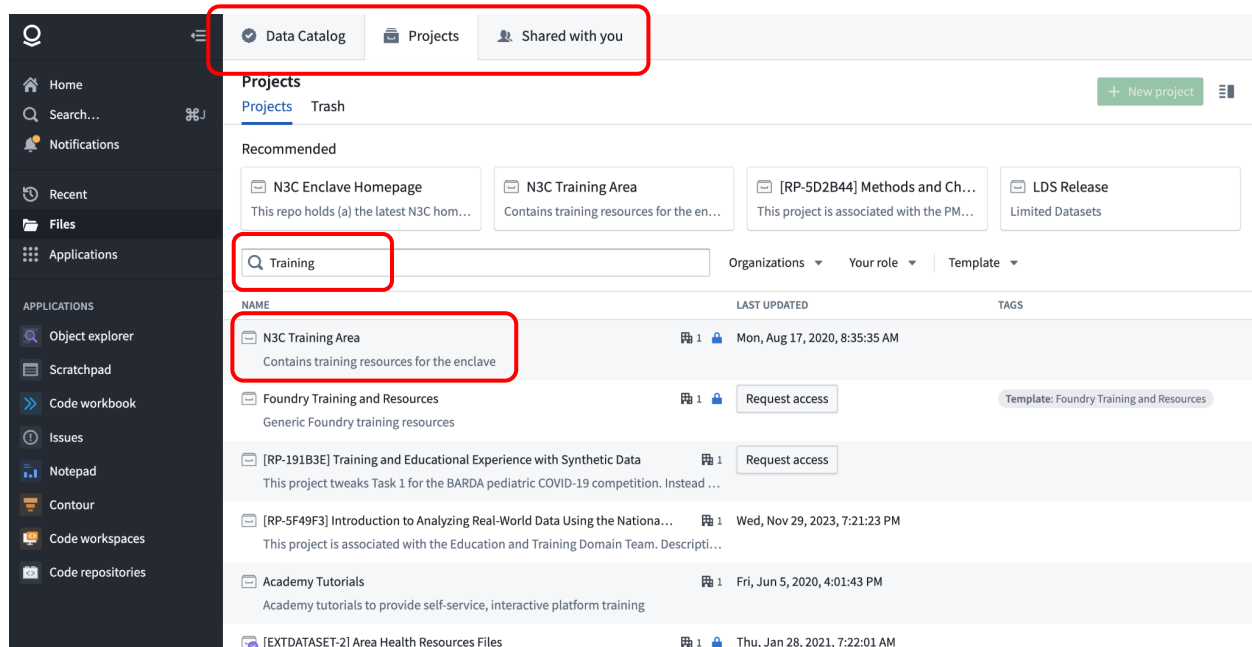
83
SITES

30.7b
TOTAL ROWS

Files (highlighted in red box)

This will open a dashboard where you can browse **Projects** - also known as project workspaces. These act like folders with strict access permissions, and use a filing-drawer icon. This dashboard also provides access to the **Data Catalog**, which we'll cover later and provides collections of shortcuts to frequently-accessed files (box 1).

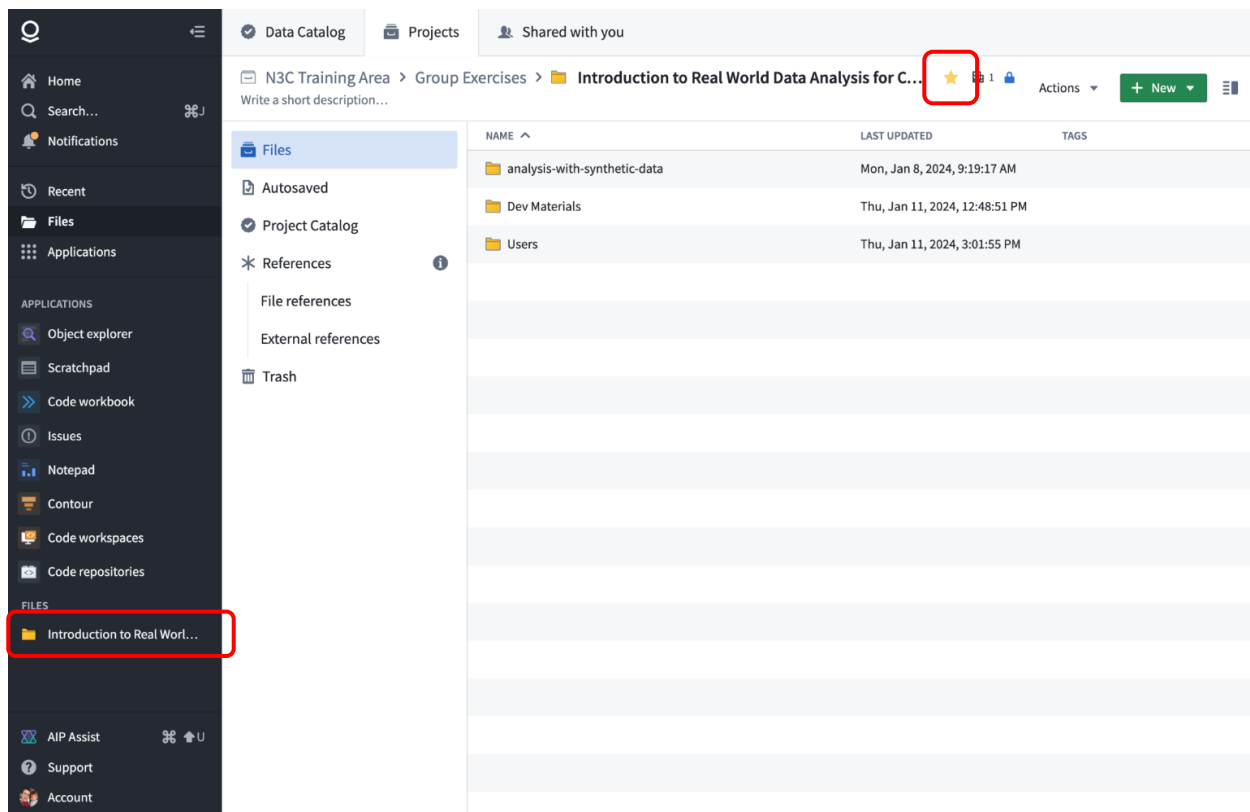
There are many project workspaces listed - search for "Training" to find the N3C Training Area workspace and open it (boxes 2 and 3).



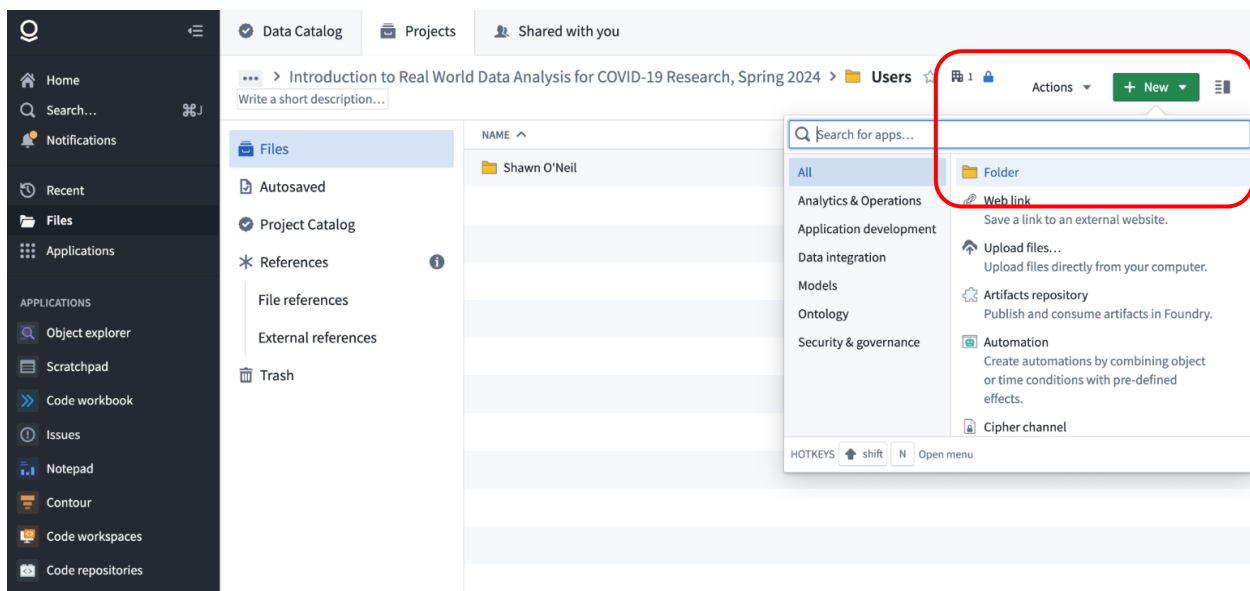
In the workspace, open the folders **Group Exercises** → **Introduction to Real World Data**

Analysis for COVID-19 Research, Spring 2024. Once you have this folder open, click the "star" icon that is next to the "breadcrumb trail" for the folder (box 1). This will put the folder in your "favorites", which are listed in the left navigation bar under **Files** (box 2). We highly recommend using these star-favorites and the breadcrumb trails for faster navigation in the enclave!

Another feature of the Enclave is the use of persistent URLs. For example, the URL displayed in the browser for the N3C Training Area workspace is <https://unite.nih.gov/workspace/compass/view/ri.compass.main.folder.86a7020f-db30-4fd1-b735-bbaf53512365> - this URL will always take you directly to the N3C Training Area workspace, even if it is moved to a different location or renamed in the future. You can thus bookmark specific enclave resources in your browser, in addition to starring favorites in the interface.



Next, open the Users folder, and create a **New Folder** using the green **New** button, naming the folder your first and last name. *If you are unable to access the workspace or create a new folder, contact Shawn O'Neil on the CD2H/N3C Slack or at shawn@tislabs.org to have permissions added for you.*



Next, let's take a look at some notional synthetic data in the enclave. Rather than find the specific location it lives in, we'll use the **Data Catalog** from the file browser to view a collection of shortcuts to commonly used resources. Open the Data Catalog tab, and then open **Patient Data** → **Synthea Data (Open Source Synthetic)**.

The screenshot shows the 'Data Catalog' interface. On the left is a sidebar with navigation options: Home, Search..., Notifications, Recent, Files, Applications, and a list of applications including Object explorer, Scratchpad, Code workbook, Issues, Notepad, Contour, Code workspaces, and Code repositories. The main panel is titled 'Data Catalog' and has tabs for 'Collections' and 'Files'. A 'Request data' button is in the top right. Below the tabs is a table listing datasets:

NAME	FILES
Clinical Data Model Resources Contains resources for OMOP Code Sets, OMOP Concepts, and Unit Harmonization Materials	11
Level 2 De-identified (DE-ID) Harmonized De-Identified Data	25
Level 3 Limited Dataset (LDS) Limited EHR Dataset	25
N3C Knowledge Store N3C Shared Logic and Template Derived Datasets. All logic is shared and open	13
Patient Data Contains links to Synthetic Data, Level 2 Data, and Level 3 Data	4
PPRL Linked Datasets Contains restricted external datasets that have been linked to N3C Data using Privacy Preserving Record Linkage, and may be requested via Data...	7
Publicly Available Datasets Publicly Available Datasets for Use in Research Questions	67

This will open a folder listing OMOP-formatted data tables, known as **datasets** in Enclave parlance. This dataset is notional (fake), but generated based on a statistical model of patient histories developed early in the pandemic. You can optionally learn more in the **A README** file. Next, open the **condition_occurrence** file.

The screenshot shows the 'Synthea Data (Open Source Synthetic)' folder view. The breadcrumb path is 'Synthea > N3C Processing > Versioning > workbook-output > Synthea Data (Open S...'. There is a 'Write a short description...' field and a '+ New' button. On the left is a sidebar with navigation options: Home, Search..., Notifications, Recent, Files, Applications, and a list of applications including Object explorer, Scratchpad, Code workbook, Issues, Notepad, Contour, Code workspaces, and Code repositories. The main panel shows a list of files:

NAME	LAST UPDATED	TAGS
A README	Thu, Jan 11, 2024, 3:17:29 PM	
condition_era	Tue, Nov 9, 2021, 3:46:19 PM	
condition_occurrence	Tue, Nov 9, 2021, 3:46:16 PM	
conditions_to_macrovisit	Thu, Jun 24, 2021, 3:57:57 PM	
death	Thu, Jun 24, 2021, 3:58:36 PM	
device_exposure	Thu, Jun 24, 2021, 3:58:21 PM	
drug_era	Thu, Jun 24, 2021, 3:58:52 PM	
drug_exposure	Thu, Jun 24, 2021, 3:58:39 PM	
measurement	Thu, Jun 24, 2021, 3:58:55 PM	
measurements_to_macrovisits	Thu, Jun 24, 2021, 3:59:01 PM	

Opening a dataset this way opens the Dataset Preview application, a powerful interface for quickly browsing and summarizing the table.

Synthea > N3C Processing > *** > condition_occurrence ☆

File Help 1 master

Preview History **Details** Health Compare

Analyze in Contour Explore pipeline All actions Build

Showing 300 of 971k rows 21 columns Search columns...

condition_occurrence

person_id Integer condition_occurrence_id Integer condition_concept_id Integer condition_start_date Date condition_end_date Date

1 20346 161036 80502 1999-09-30 1999-09-30

2 20371 161234 80502 2001-09-06 2001-09-06

3 20417 161566 80502 1991-11-02 1991-11-02

4 20422 161598 80502 2003-03-05 2003-03-05

5 20426 161631 80502 2010-02-14 2010-02-14

6 20467 161925 80502 1982-12-09 1982-12-09

7 20468 161932 80502 2015-12-20 2015-12-20

8 20473 161969 80502 2017-12-02 2017-12-02

9 20543 162558 80502 2018-02-05 2018-02-05

10 20550 162592 80502 2002-06-16 2002-06-16

11 20582 162837 80502 2011-04-02 2011-04-02

12 20585 162875 80502 1991-09-17 1991-09-17

13 20594 162942 80502 2012-11-15 2012-11-15

14 20609 163048 80502 2014-04-19 2014-04-19

15 20624 163145 80502 1993-02-18 1993-02-18

16 20628 163183 80502 1998-12-29 1998-12-29

17 20647 163346 80502 2003-12-24 2003-12-24

18 20660 163450 80502 2013-10-08 2013-10-08

19 20712 163928 80502 2006-08-04 2006-08-04

20 20715 163952 80502 2003-08-25 2003-08-25

21 20717 163977 80502 2012-08-24 2012-08-24

22 20726 164040 80502 1990-09-18 1990-09-18

23 20735 164097 80502 1977-07-25 1977-07-25

24 20748 164191 80502 1995-05-07 1995-05-07

25 20750 164211 80502 2000-10-13 2000-10-13

26 20752 164227 80502 2018-10-04 2018-10-04

27 20809 164643 80502 1982-08-26 1982-08-26

28 20821 164755 80502 2018-09-10 2018-09-10

Updated Nov 9, 2021, 3:46 PM by Shawn O'Neil

Created Apr 12, 2021, 4:25 PM by Shawn O'Neil

Location /UNITE/Synthea/N3C Processing/Versionin...

Type Dataset

Size 21 columns • 971k rows • 1 file • 22.8MB

Updated via versioning

Show more

Tags

Add tags

Health Checks

View details

There are no health checks on this dataset.

Inputs

Explore data lineage

condition_occurrence

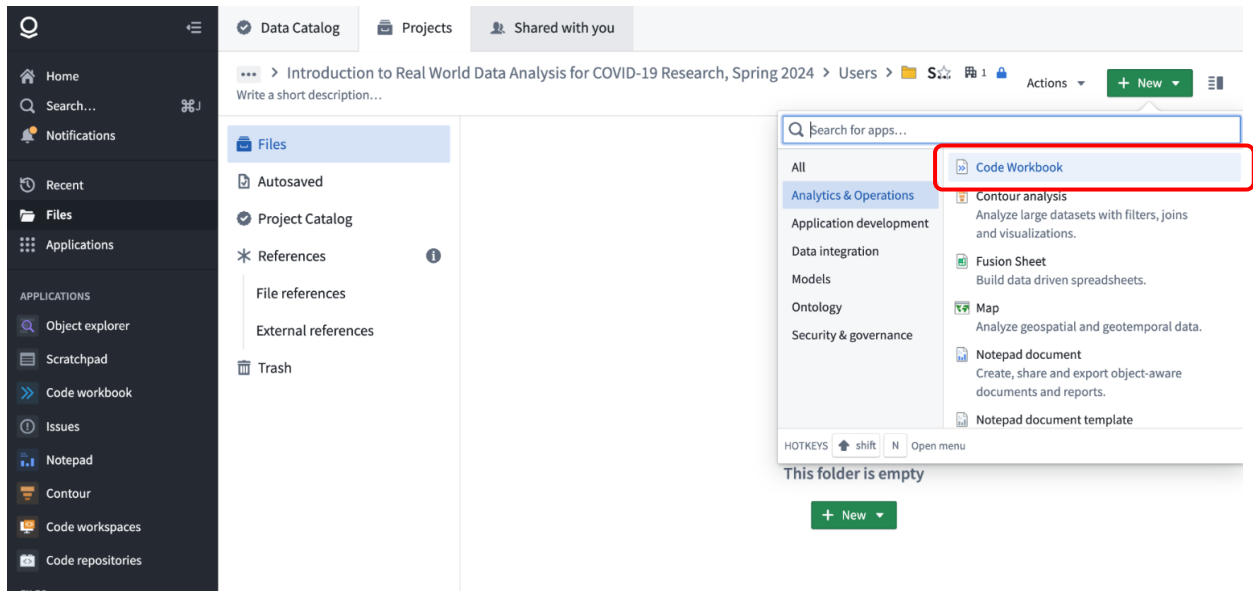
/UNITE/Synthea/N3C Processing/Fuzzing Labs and Conditions

This table is in OMOP format, which we'll cover more in later sessions. Here are some simple data manipulation exercises you can try in the data browser interface:

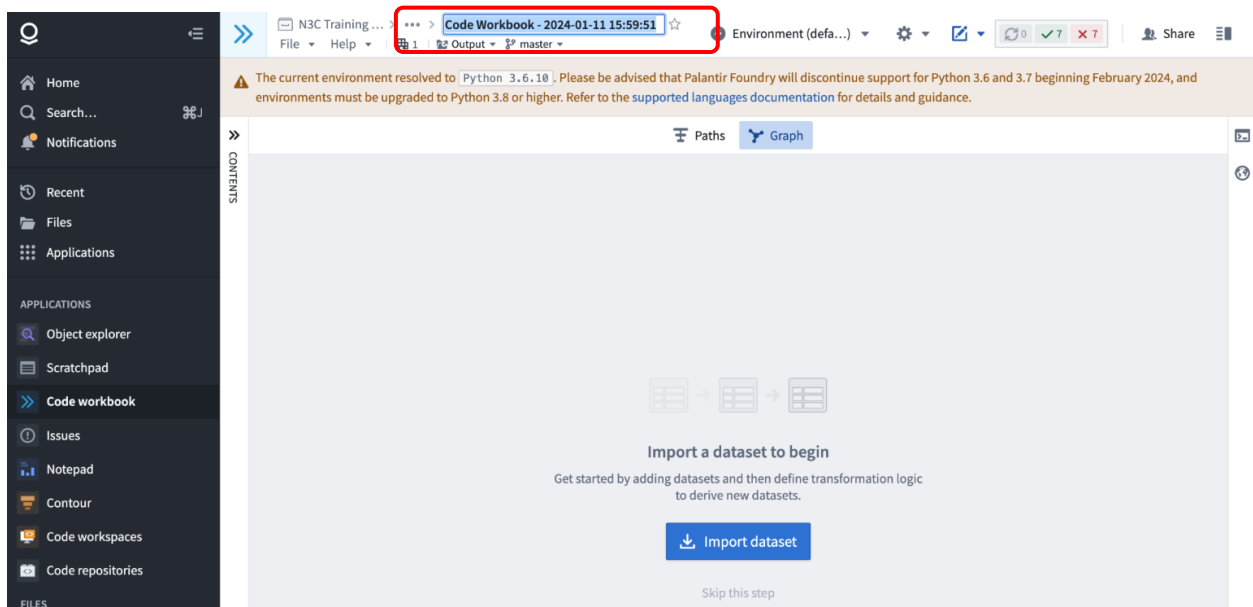
1. Use the dropdowns next to one of the columns (box 1) to **Filter** values (though note that filtering is limited to exact matches only in this basic interface), **Sort** them, or **View Stats**. The View Stats feature is particularly useful for data browsing, and can be done after filtering by one or more other columns.
2. Check out the Details tab (box 2) for more information about the dataset, noting especially the **Files** information. Datasets in the enclave are handled by Apache Spark, a high-performance distributed computing technology, which breaks large datasets over many files for efficiency (this dataset consists of just one file because it is small). For the most part we won't need to worry about this, but Spark is a powerful tool for advanced enclave usage, particularly for analyzing N3C's very large datasets!
3. Visit the **Explore data lineage** tool (box 3), to see the input datasets and code that were used to create this dataset. This is another advanced tool you won't need to worry about now, but highlights that the enclave has very powerful features for data provenance.

Part 2 - Code Workbooks and Datasets

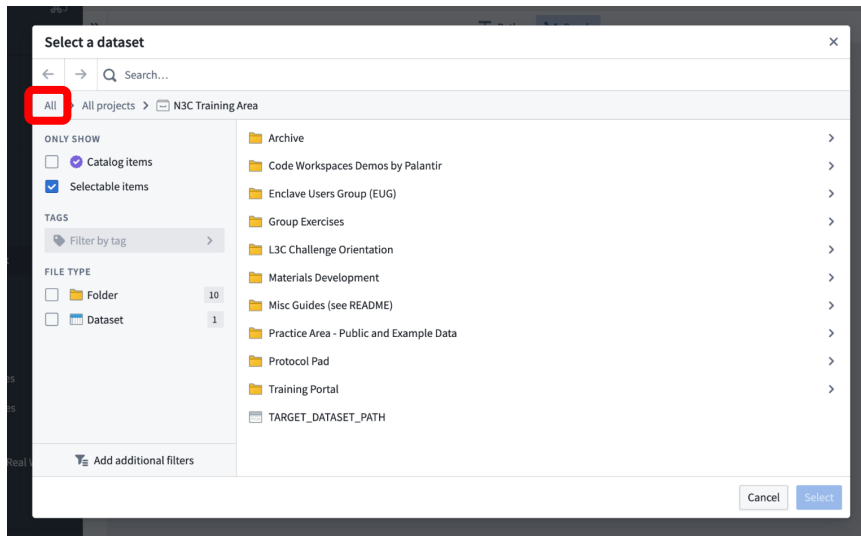
In this part we're going to create a very basic data manipulation pipeline with the Code Workbook tool, largely to practice working with workbooks, folders, and datasets together. First, navigate to your student folder from part 1 above (via your favorites list in the navigation menu!), and use the green **New +** button to create a **New Code Workbook**. It will be listed under "Analytics & Operations":



The resulting Code Workbook will be named according to current date; click the filename in the breadcrumb menu to rename it **Assignment 1 Workbook**. When prompted to rename the output folder, select **Rename**.



Now, you can click the **Import dataset** button to open the file browser and choose a dataset to import. We are going to import the **condition_occurrence** dataset we saw above, by finding it in the Data Catalog. To do that, navigate and select it using **All → Data Catalog → Patient Data → Synthea Data (Open Source Synthetic) → condition_occurrence**.



The next view is of the Code Workbook main interface, which shows the dataset we've imported as a "node" in the workspace. When a node is selected, a panel will show information about it on the bottom, and when it is highlighted a blue "Add transform" button will appear.

The screenshot shows the Palantir Foundry interface. On the left is a sidebar with navigation options like Home, Search, Recent, Files, Applications, and a list of code repositories. The main workspace displays a 'DATASET' node with a table of data. A red box highlights the 'Add Transform' button (a blue plus sign) next to the dataset. Another red box highlights the 'Selected node panel' below the dataset, which shows a table of data with columns like person_id, condition_occurrence_id, condition_concept_id, condition_start_date, condition_end_date, and condition_end_date. A red arrow points from the text 'Add Transform button' to the plus sign, and another red arrow points from the text 'Selected node panel' to the table below.

Dataset Table:

person_id	condition_occurrence_id	condition_concept_id	condition_start_date	condition_end_date	condition_end_date
20346	161036	80502	1999-09-30	1999-09-30	1999-09-30
20371	161234	80502	2001-09-06	2001-09-06	2001-09-06
20417	161566	80502	1991-11-02	1991-11-02	1991-11-02
20422	161598	80502	2003-03-05	2003-03-05	2003-03-05
20426	161631	80502	2010-02-14	2010-02-14	2010-02-14
20467	161925	80502	1982-12-09	1982-12-09	1982-12-09
20468	161932	80502	2015-12-20	2015-12-20	2015-12-20
20473	161969	80502	2017-12-02	2017-12-02	2017-12-02
20543	162558	80502	2018-02-05	2018-02-05	2018-02-05

Selected node panel Table:

person_id	condition_occurrence_id	condition_concept_id	condition_start_date	condition_end_date	condition_end_date
1	20346	161036	1999-09-30	1999-09-30	1999-09-30
2	20371	161234	2001-09-06	2001-09-06	2001-09-06
3	20417	161566	1991-11-02	1991-11-02	1991-11-02
4	20422	161598	2003-03-05	2003-03-05	2003-03-05
5	20426	161631	2010-02-14	2010-02-14	2010-02-14
6	20467	161925	1982-12-09	1982-12-09	1982-12-09
7	20468	161932	2015-12-20	2015-12-20	2015-12-20
8	20473	161969	2017-12-02	2017-12-02	2017-12-02
9	20543	162558	2018-02-05	2018-02-05	2018-02-05

NOTE: There is generally no need to “save” files as they are edited in the enclave - they are continually saved as you work.

Next, add a transform to the dataset by clicking the blue + “Add Transform” button, and selecting **SQL code**. This will create a new SQL “transform” node, with an arrow connecting the dataset and the transform. The lower panel will show a code editor, and some tabs you may want to browse.

Edit the code to read `SELECT * FROM condition_occurrence WHERE data_partner_id = 1346` as shown, and click **Preview** to run the code. This will result in the code being run, and the result being shown in the Preview tab.

(Side note: data_partner_id is not a standard column of OMOP data, but is an N3C addition indicating which data partner each row of data comes from. Although these data are notional, we’ve added data_partner_id columns for realism.)

The screenshot displays the Palantir Foundry interface. On the left is a dark sidebar with navigation options: Home, Search..., Notifications, Recent, Files, Applications, Object explorer, Scratchpad, Code workbook, Issues, Notepad, Contour, Code workspaces, Code repositories, and a FILES section with 'Introduction to Real Worl...'. The main workspace has a top bar with 'N3C Training ... > *** > Assignment 1 Workbook', environment details 'Environment (default-r3.5)', and a share button. A warning banner states: 'The current environment resolved to: Python 3.6.10. Please be advised that Palantir Foundry will discontinue support for Python 3.6 and 3.7 beginning February 2024, and environments must be upgraded to Python 3.8 or higher. Refer to the supported languages documentation for details and guidance.'

The workspace contains two main components:

- DATASET:** A table with columns: person_id, condition_occurrence_id, condition_concept_id, condition_occurrence_start_date, condition_occurrence_end_date. It shows 21 columns and 971,227 rows. A red box highlights the 'condition_occurrence' column.
- SQL Transform Node:** A node labeled 'SQL' with a plus icon. It contains the text 'No preview available' and 'unnamed'. A red box highlights this node, with a red arrow pointing to it from the text 'Transform Node'.

Below the dataset, a red box highlights the 'Code editor (aka Logic)' section. It shows a SQL query:

```
1 SELECT *
2 FROM condition_occurrence
3 WHERE data_partner_id = 1346
```

At the bottom of the red box, a red arrow points to the 'Useful tabs' section, which includes tabs for Logic, Inputs, Preview, Visualizations, Logs, and Description.

Now, we've only previewed this result - it is not saved out as a dataset, and if we were to re-open this workbook we would lose the preview. It is much more common to use the **Save as dataset feature** (box 1), giving the output reasonable filename a name like conditions_site_1346 (box 2; you'll also be asked if you want to rename the node in the workbook, say yes), and clicking Run (box 3).

The screenshot shows the Databricks workspace interface. On the left is a sidebar with navigation options: Recent, Files, Applications, Object explorer, Scratchpad, Code workbook, Issues, Notepad, Contour, Code workspaces, Code repositories, FILES, Introduction to Real Worl..., AIP Assist, Support, and Account. The main area displays a SQL node. The node is highlighted with a red box, and its 'Run' button is circled in red. The 'Save as dataset' toggle is also circled in red. The SQL query is: `SELECT * FROM condition_occurrence WHERE data_partner_id = 1346`. The result is a dataset named 'conditions_site_1346' with 21 columns and 398,059 rows. The dataset view shows a table with columns: person_id, condition_occurrence_id, condition_concept_id, condition_occurrence_id, and condition_concept_id. The dataset is named 'conditions_site_1346' and has 21 columns and 398,059 rows.

Running the node executes the code, and this time the result is saved to a dataset in the enclave. **NOTE: if the code or upstream input datasets change, the result dataset will *not* be updated unless it is explicitly re-run.**

Finally, you should be able to navigate back to your personal folder (where you created the Assignment 1 Workbook), and see a new **workbook-output** folder. In here you can navigate to the **Assignment 1 Workbook** folder (named after the code workbook by default), and find your **conditions_site_1346** dataset for review!

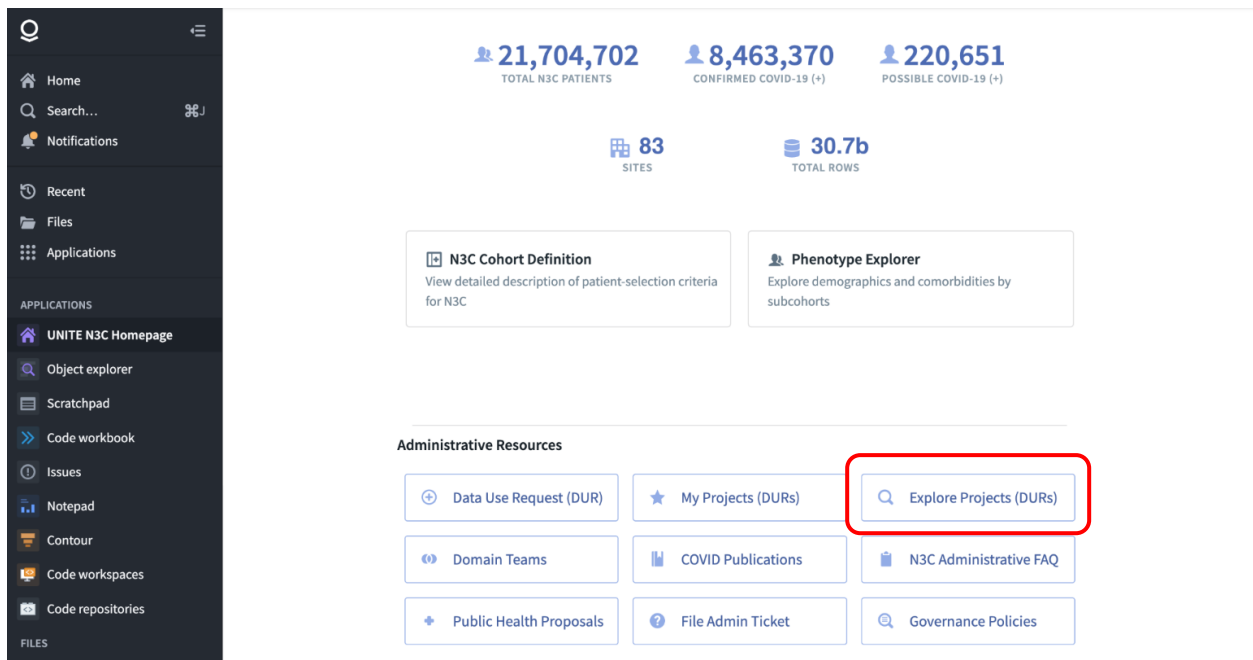
The screenshot shows the Databricks Data Catalog interface. On the left is a sidebar with navigation options: Home, Search..., Notifications, Recent, Files, Applications, Object explorer, Scratchpad, Code workbook, Issues, and Notepad. The main area displays a table with columns: NAME, LAST UPDATED, and TAGS. The 'workbook-output' folder is highlighted with a red box. The table lists 'Assignment 1 Workbook' and 'workbook-output'.

Part 3 - Joining a DUR

In a later part of this course we will be working with real Level 2 data in a research project workspace. To access the workspace, we need to join the DUR and get approved. We're going to find the DUR via the N3C dashboard, but here's a direct link just in case you have trouble: [Malnutrition and COVID-19 Outcomes](#).

If you have any questions or issues, please reach out to your course instructors! For more information on N3C DURs in general, check out [this chapter](#) in the Guide to N3C, and [this tutorial video](#).

First, from the Enclave homepage, open the **Explore Projects** dashboard:



This will open a dashboard showing all N3C DURs, which come in three kinds: those configured as joinable by the project lead (box 1), those not listed as joinable (box 2), and operational DURs used by N3C staff (box 3). Use the search box to look for “Malnutrition”, and in the entry for **Malnutrition and COVID-19 Outcomes** (box 4), scroll to the right to find the corresponding **Request to Join** link (box 5), and open it.

N3C Projects All Projects Projects by N3C Domain Team Closed Projects

SEARCH PROJECTS 4

Malnutrition

Projects to Join 1

Scroll to the right to find the Join link

Title	Statement	Lead Investigator	Email Lead Investigator	Lead Investigator Accessing Institution	Review Status	Allow Joining	Request to Join
Impact of Preoperative Dysphagia on	condition in middle-aged and older f malnutrition, and can adversely after surgery. Yet, preoperative	Seth Cohen	Email	Duke University	Accepted	Allow	Request to Join
Malnutrition and COVID-19 Outcomes	health crisis that effects up to one out results in an estimated annual cost of the United States. Malnutrition is linked	Alfred Anzalone	Email	University of Nebraska Medical Center	Accepted	Allow	Request to Join 5
Mortality and Complications Among Patients	exposed the existing healthcare d added on to the current burden in ealth disparities. Several previous	Karthik Raghunathan	Email	Duke University	Pending	Allow	

Projects to Explore 2

Operational Projects (N3C Technical Development) 3

Scroll to the right to find the Join link

This will open the DUR form for the project, in which the project lead has specified a Title, Abstract, and Data Level for the DUR.

Request to join as a Collaborator

Join an Existing Data Use Request

DUR Information

Project Title
Malnutrition and COVID-19 Outcomes

Research Project Abstract
Malnutrition is a global health crisis that effects up to one out of two older adults and results in an estimated annual cost of \$51.3 billion per year in the United States. Malnutrition is linked to weaker immune systems, reduced cardiac output, poor diaphragmatic and respiratory muscle function and impaired gastrointestinal function. Early studies have explored the prevalence and severity of malnutrition in COVID-19 patients. However, small sample sizes and single-site studies limit the generalizability of results. Currently, a gap in knowledge exists regarding the relationship of malnutrition in hospital admissions in COVID-19 patients and the impact of malnutrition on clinical outcomes.

Data Security Tier
De-identified Data (Level 2)

PPRL External Dataset
Choose PPRL External Dataset from displayed values

De-Identified Data (Level 2)
Patient-level records scrubbed of identifying information.

New sheet

Collaborator Attestations

To join the DUR and eventually get access to the project workspace, scroll down and fill out the required information:

- Read and attest to the institutional DUA.
- Verify that you have completed the required NIH IT Security training in the past year (the (i) button provides a link - the 2023 refresher is the needed one).
- Verify that you have completed appropriate Human Subjects Research Protection Training in the past 3 years, and provide the date of completion. (More information [here](#).)
- Read and attest to the User Code of Conduct.
- Answer the question “Does your Institution policy require IRB review for use of Level 2 data?”
 - Note that most institutions do not require IRB review for Level 2 (De-Identified) data, which is commonly not considered human subjects research. However, we cannot guarantee this for your institution, and you should check with your IRB office if unsure.
- Read and attest to the N3C Download policy.
 - Importantly, you are not allowed to screenshot or record video of row-level patient data (notional data such as we have worked with in this assignment is ok, but be careful!) All results for publication, such as figures and tables, must be approved by a submission-and-review process before they can be exported.
- Acknowledge that if you plan to use publicly available datasets alongside patient data you will need to ensure it is consistent with your institution’s policies.
 - Some institutions regard linking De-Identified patient data with public data (e.g. US Census or other data) an increased risk of identification, potentially triggering the need for IRB review otherwise not required. Again, this is not typical, but we cannot guarantee this for your institution, and you should check with your IRB office if unsure.
- Finally, click Submit!

What happens next: Your request to join the DUR will be sent to the project lead (Dr. Anzalone) for approval; once approved by the lead, it will head to the N3C Data Access Committee (DAC) for approval; once approved, your account will be given permissions to access the corresponding workspace and you will receive an email as well.

To see the status of your DUR request, you can navigate to the **My Projects** dashboard of the enclave homepage.