Datasets

Contents

Datasets

A dataset is a specific view of data in PHEMI Central.

Datasets are virtual constructs, created by selecting data elements from across all of the digital assets data collections stored in PHEMI Central. You can contsruct a dataset using any available data elements, across multiple and disparate data collections. They can be created in advance or on demand, and they can be altered whenever circumstances dictate. You don't have to predefine the data and you don't have to navigate a complex database schema.

To be able to define a dataset, a user must have a role of PHEMI Administrator. To be able to execute or export a dataset, a user must have a role of Data Analyst.

Datasets are instantiated only when they are executed. Datasets can be be executed directly by users and downloaded or exported to spreadsheets, applications, or analytics tools. Because all data is indexed and cataloged in advance, dataset execution is fast.

Defining a dataset does not modify any configuration for data collections or impact any protections to data, because the PHEMI Central Policy Enforcement Engine mediates all requests for data. Users can only access and export data to which they have rightful access, based on data visibility tags, user authorizations, and access policies, so access to data remains governance-compliant at all times.

Datasets are often created for research or exploratory data use. As such, the intended use for a given dataset may be different from the intended use of the original data collections. In particular, the governance and controls intended for the data in a dataset may differ from the governance of the original data collections. In PHEMI Central, you can attach an access policy to a dataset that is different from, and independent of, the access rules applied to the consistuent data collections.

Similarly, a single dataset can be executed by multiple users—even by users with different authorizations. A given user's actual view into the data in an executed dataset is always mediated by the Policy Enforcement Engine, and is therefore restricted by data visibility and user authorization, as experssed in an access policy. Thus, two users with different authorizations might execute the same dataset but end up with different views of the data.
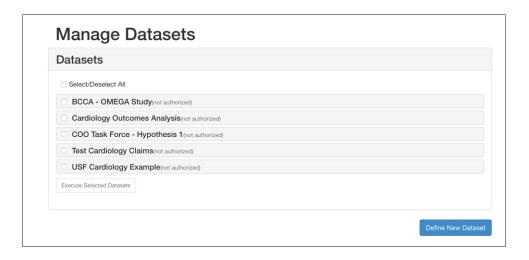
View Defined Datasets

View the list of defined datasets on the **Manage Datasets** page.

To see what datasets have already been defined:

Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.

The **Manage Datasets** page opens, listing the datasets that have already been defined.

Define a Dataset

Define a new dataset on the **Manage Datasets** page.

> Note: Only users with a role of PHEMI Administrator can define datasets.

To define a dataset:

1.
Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.
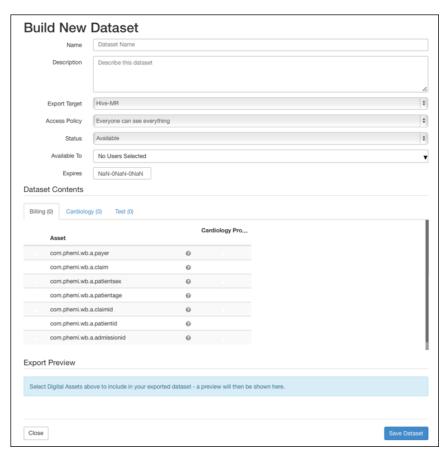
The **Manage Datasets** page opens, listing the datasets that have already been defined.

2. Click the **Define New Dataset** button.
The **Build New Dataset** screen opens.

Build New Dataset

Name	Dataset Name
Description	Describe this dataset
Export Target	Hive-MR
Access Policy	Everyone can see everything
Status	Available
Available To	No Users Selected
Expires	NaN-0NaN-0NaN

Dataset Contents

Billing (0)   Cardiology (0)   Test (0)

Asset	Cardiology Pro...	
com.phemi.wb.a.payer	❓	
com.phemi.wb.a.claim	❓	
com.phemi.wb.a.patientsex	❓	
com.phemi.wb.a.patientage	❓	
com.phemi.wb.a.claimid	❓	
com.phemi.wb.a.patientid	❓	
com.phemi.wb.a.admissionid	❓	

Export Preview

Select Digital Assets above to include in your exported dataset - a preview will then be shown here.

Close   Save Dataset

3. Describe the dataset.

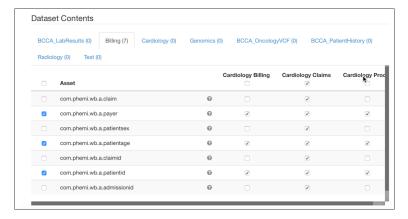| Option	Description
Name	Mandatory. A name for the dataset.
Description	Mandatory. A brief description of the dataset.
Export Target	Optional. The format required by the system to receive the dataset. Supported values are as follows: • Hive-MR. Apache Hive MapReduce format. • JDBC. Java Database Connectivity format. • CSV. Comma-separated values format. • TSV. Tab-separated values format. The default is Hive-MR. If you choose JDBC as the format, you have the option of selecting one of the dataset destinations that has been configured for the system. *Tell me about dataset destinations.*
Access Policy	Optional. The access policy to protect the dataset. The access policy selected for the dataset may match or may be different from any applied to the constituent data collection(s). By default, the first configured access policy is applied.
Status	Optional. The availability of the dataset. The only supported value is Available.
Available To	Mandatory. The users who are allowed to access this dataset. Select one or more users from the drop-down list; only users with a role of Data Analyst appear on the list. By default, no users are allowed to access the dataset.

Option	Description
Expires	Mandatory. An expiry date for the dataset. When this date is reached, PHEMI Central deletes the dataset from the system.

4. Choose the fields you want in the dataset.

You select the fields for the dataset in the **Dataset Contents** pane. This pane will show a number of tabs; these tabs reflect the data categories configured for your data. Clicking a tab will list all the fields extracted from all the data collections within the category, where each data collection is represented by a column. Each data source and each field has a checkbox.

- Select an entire data collection by checking the checkbox for the column.
- Select any number of individual fields.

Continue for each data category you want to include in the dataset.

As you select fields, a preview of your selections appears in **Export Preview** pane.

5. When your dataset is complete, click the **Save Dataset** button. The system confirms when the dataset has been successfully saved. Click **Close** to dismiss the **Build New Dataset** screen.

View Dataset Information

View information for a dataset on the **Manage Datasets** page.

📝 **Note:** Only users with a role of PHEMI Administrator can view dataset information.

To see information for a dataset:

1.

Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.
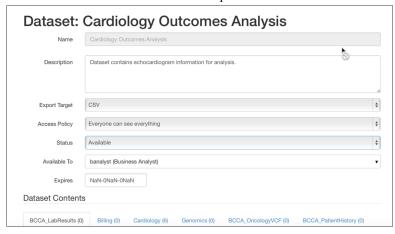
The **Manage Datasets** page opens, listing the datasets that have already been defined.

2. Click the name of the dataset you want to view.

   The information screen for the dataset opens.

3. Click **Close** to dismiss the screen.

Modify a Dataset

Modify information for a dataset on the **Manage Datasets** page.

> **Note:** Only users with a role of PHEMI Administrator can modify a dataset definition.
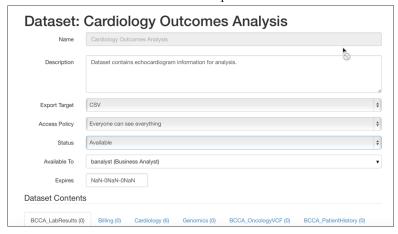
To see information for a dataset:

1.
   Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.

   The **Manage Datasets** page opens, listing the datasets that have already been defined.

2. Click the name of the dataset you want to modify.

The information screen for the dataset opens.

3. Make your changes. *What do these fields mean? How do I change the fields in the dataset?*

Click **Save Dataset** to save your changes. The system confirms when the dataset has been successfully saved. Click **Close** to dismiss the dataset information screen.

Execute a Dataset

View the list of defined datasets on the **Manage Datasets** page.

> **Note:** Only users with a role of Data Analyst can execute a dataset.

To see what datasets are available:

1.

Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.

The **Manage Datasets** page opens, listing the defined datasets.

2. Check the checkbox next to the dataset you want to execute, or check **Select/Deselect All** to select all the datasets.

You will only be able to select datasets that the PHEMI Administrator has designated as available to you in the dataset definition.

3. Click **Execute Selected Datasets**.

The dataset executes.

Delete a Dataset

Delete a dataset from the **Manage Datasets** page.

> **Note:** Only users with a role of PHEMI Administrator can delete a dataset.
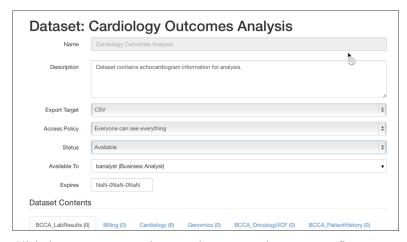
To delete a dataset:

1.

Open the **Manage Datasets** page, by clicking the **Datasets** icon in the left navigation bar.

The **Manage Datasets** page opens, listing the datasets that have already been defined.

2. Click the name of the dataset you want to delete.

The information screen for the dataset opens.

3. Click the **Delete Dataset** button. The system asks you to confirm permanent deletion. Click **Delete**.