Data Sources

Contents

Data Sources

In PHEMI Central, a data source is the set of management and governance rules and policies that will be applied to a collection of data. A data source configuration should be defined for each collection of data to be stored and managed according to the same retention, legal, and governance rules.

Data source configuration includes:

- A data policy. *Tell me about data policies.*
- The Data Processing Function associated with the data source. *Tell me about DPFs.*

The **Data Sources** page also includes a facility for manually ingesting data. *How do I manually ingest data?*

Connection information to specific systems is not a part of data source configuration. Data is "pushed" the site submitting the data. The submitting system initiates a session on a designated port, on which PHEMI Central listens. PHEMI Central extracts any necessary connection information and login credentials from the session information. *How do I submit data to PHEMI Central?*

Data Policies

The data policy provides several key items of information about a data source.

The data policy describes what type and format of data is expected from the data source and classifies it into a data category. It also lists the users who are responsible for various aspects of the information.

The data policy also specifies the privacy settings for the data source. It specifies one of the configured data visibilities, which describe the sensitivity of the data. The data policy also specifies which access policy (if any) will be used to enforce rightful access to the information. All raw data ingested from this data source, as well as every item of data derived from processing this data source, will be tagged with the specified data visibility and access policy to protect privacy.

The data policy is also where the organization's data sharing agreement is recorded. In this context, the data sharing agreement is a document that records permission for data from this data source to reside on the PHEMI system. You upload the agreement document and specify the time period for which the agreement is valid.

The data policy also records the retention rules to be applied to information from this data source. The system uses the retention rules to calculate a time to live for each data item. When the time to live expires, PHEMI Central deletes the raw data and all associated derived data items from the data store.

If you want your information to be version-controlled, the data policy is the place where you enable version control.

Data Processing Functions

Data Processing Functions are associated with a data source during data source configuration.

A Data Processing Function, or DPF, is an executable piece of code that supplies the instructions for processing raw data (for example, a log message or medical report) to extract from heterogeneous data sources meaningful, context-specific information (such as a temperature reading or blood glucose measurement) that can be queried or exported for analysis. The code is executed by the PHEMI Central DPF Engine, which uses it to direct curation of the data. The input to a DPF is the raw binary data ingested into the system. The output of a DPF is a set of structured elements, each of which includes a type property (for example, INT or STRING) and can specify data visibilities (for example, SECRET or IDENTIFIABLE) on a per-field basis. The data elements output by a DPF are called derived data. The collection of derived data produced by a DPF is automatically indexed in PHEMI Central.

DPFs converted ingested raw data into analytics-ready digital assets. The DPF instructs the system how to parse the raw data, extract key content and perform data cleansing, ontology matching, enhanced indexing and cataloging, and structuring data according to the organization’s needs. Standard PHEMI DPFs are included to index and describe

structured data, such as Microsoft Excel spreadsheets, comma-separated value (CSV) files, database records, or strongly typed XML/JSON, for consumption of the data through a REST or SQL interface. Ingested genomic Variant Call Format (VCF) files are converted into a series of variants, with each variant governed by the parent file's privacy specifications.

A DPF is associated with a data source as part of data source configuration. Association consists of uploading the DPF archive.

A DPF archive is the set of code that makes up a DPF. A DPF archive is delivered as a ZIP file archive. It consists of two parts: a manifest file and a code library. To associate a DPF with a data source, the DPF archive is ``registered`` with the data source by uploading the archive during data source configuration.

User-defined DPFs can also be developed for PHEMI Central for specific needs, such as analyzing semi-structured data or performing natural language processing on free text. DPFs can catalog data and standardize it into ontologies such as SNOMED or LOINC, so that data analysts to find the right information in the required format. DPFs can also analyze streams of machine data to find patterns and exceptions, calculating aggregates and converting the telemetry into an analytics-ready state for trending and predictive analysis. For parsing unstructured documents, such as scans or X-rays, the DPF can include specialized parsing functions such as OCR or image parsing.

As your organization's needs evolve and as knowledge advances, DPFs can be updated and re-executed, to leverage the value of the organization's historical data in new ways.

PHEMI Central includes a library of helper functions to simplify DPF development. DPFs can be developed in either Java or Python, or PHEMI Central can be extended to support DPF development in any modern programming language that runs on a Linux OS. MapReduce or YARN knowledge is not necessary. Your DPF can be written by PHEMI, by your organization's in-house programmers, or by third-party developers. Training in DPF development is also available from PHEMI.

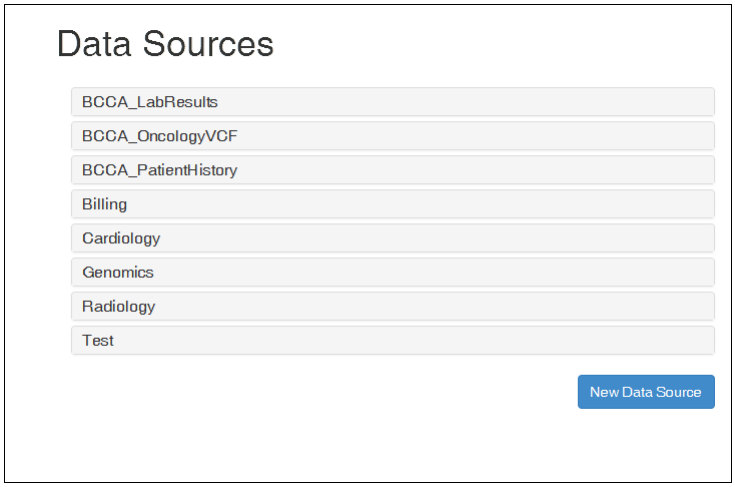
View Data Sources

View data sources on the **Data Sources** page.

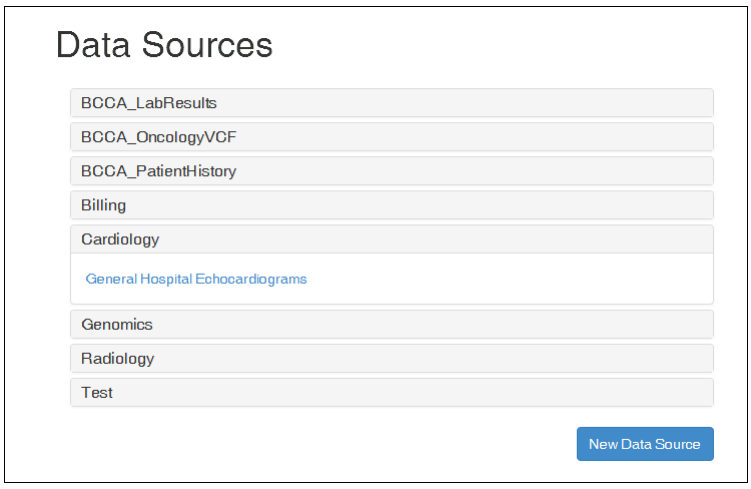
To view defined data sources:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon in the left navigation bar. 

   The **Data Sources** page opens showing defined data categories. *Tell me about data categories.*



2. Click any data category to expand the category and see the data sources included in the category.



Click the data category a second time to collapse the category again.

Define a Data Source

Define a data source on the**Data Sources** page.

Before defining a data source, you must configure the following:

- Data categories
- Data visibilities
- Users

To be able to configure users, you must first configure *user authorizations*. Your users must include at least one user with a role of PHEMI Administrator and at least one user with a role of Privacy Officer.

If you are applying an access policy to the data source, you must first configure the *access policies*.

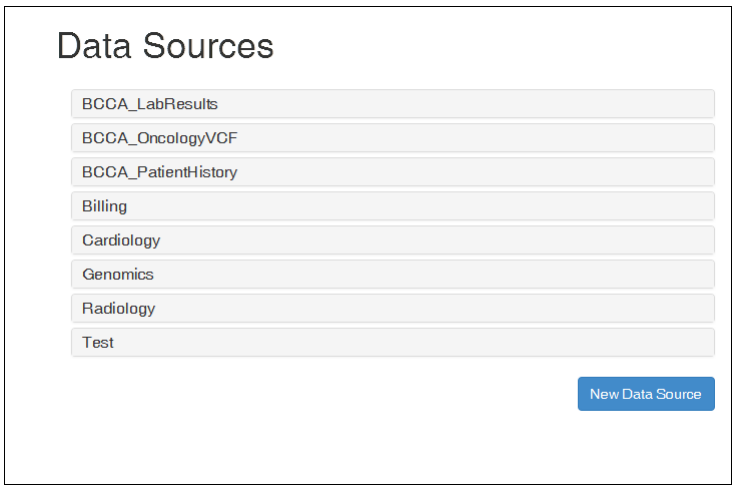
When you first define a data source, only the **Data Policy** screen is available to you. Configure the data policy and save it. After saving the data policy information, the **Data Processing Function** and **Ingest Data** tabs become available to you.

Define the Data Policy

To define the data policy:

1. Open the **Data Sources** page, by clicking the **Data Sources** icon in the left navigation bar. 

   The **Data Sources** page opens showing all defined data categories.



2. Click the **New Data Source** button.

   The **New Data Source** screen opens with only the **Data Policy** tab showing.

3. Describe the data source.

Option	Description
Name	Mandatory. A name for the data source. Numbers, letters, spaces, hyphens, and the underscore character are supported. **Note:** Once you save a data source, the name cannot be edited. To change the name, you must delete the whole data source and reconfigure it with the new name.
Source Category	Mandatory. The data category of the data source. Choose from the drop-down list of data categories. *How do I define data categories?*
Institutional Owner	Mandatory. The individual responsible overall for data stored in PHEMI Central. Only users with a role of PHEMI Administrator are eligible. Choose from the drop-down list of eligible users.
Privacy Officer	Mandatory. The individual responsible for defining the organization's governance policy and for approving access policies. Only users with a role of Privacy Officer are eligible. Choose from the drop-down list of eligible users.

Option	Description
Source Owner	Mandatory. The individual responsible for approving dataset requests involving this data source. Only users with a role of PHEMI Administrator are eligible. Choose from the drop-down list of eligible users.
Data Visibilities	Optional. The data visibility (privacy tag) to be associated with this data source. Select from the list of configured data visibilities. You can select multiple items by holding down the Shift key.
Document Format	Optional. The kinds of document expected from this data source. Examples are Microsoft Word, Excel, or JSON.
Definition	Optional. A brief description of the kinds of documents expected from this data source.
Notes	Optional. Any additional notes for the data source.

4. Upload the data sharing agreement. The data sharing agreement records permission for this data to reside on PHEMI Central.

   Click the **Choose File** button and navigate to the document in your local file system. Double-click the file to select it. The system uploads the document.

   Specify the period during which this data sharing agreement is in effect. The format for the start and end dates is *mm-dd-yyyy*.

5. Specify the retention rules. These rules are used to determine how long each item from this data source can remain in the system.

Option	Description
Please choose...	Mandatory. Specifies when data should be erased from the system. Supported values are as follows: • Retain for a time period. Keeps the data for the period specified. Specify some number of minutes, hours, days, weeks or years. • Delete after time period. Erases the data after the specified time period. Specify some number of minutes, hours, days, weeks or years. • Do not delete. The data is never deleted. • Delete oldest data once capacity is reached. Deletes data items with the oldest timestamp after the specified capacity is reached. Specify the capacity as some number of Kilobytes, Megabytes, or Gigabytes.
Version control	Optional. Maintains version control over data. Check to enable version control; uncheck to disable version control. By default, version control is disabled.

6. Click the **Save Data Source** button to save the data policy. When the data policy has been successfully saved, the **Data Processing Function** and **Ingest Data** tabs appear as available on the screen.

Specify the Data Processing Function

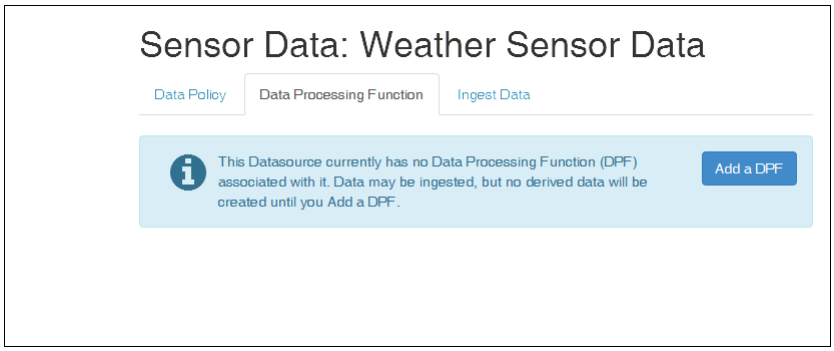
To associate a DPF with a data source, you upload the DPF archive onto the Data Processing Function screen of the Data Sources page.

Tell me about DPFs and DPF archives.

To associate a DPF with a data source:

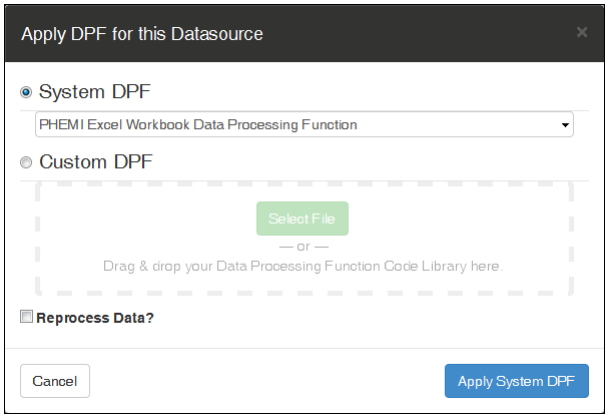
1. From the **Data Policy** screen of the **Data Sources** page, click the **Data Processing Function** tab.

   The **Data Processing Function** screen opens. When you first define a data source, no Data Processing Function (DPF) is associated with it.



2. Click the **Add a DPF** button.

   The **Apply DPF for this Datasource** screen opens.



3. Choose to use either a **System DPF** or a **Custom DPF**.

   - If you choose **System DPF**, select which PHEMI system DPF you want to use from the drop-down menu. Choose between a DPF for processing **Variant Call Format (VCF) data** and a DPF for processing **Microsoft Excel workbook** data. Once the system DPF is selected, click the **Apply System DPF** button to apply the DPF.
   - If you choose **Custom DPF**, click the **Select File** button, navigate to the DPF archive and double-click the ZIP file. Or, drag and drop the DPF archive from your file manager oto the target area on the Management and Governance Console screen.Click the **Apply Custom DPF** button to upload the ZIP file and apply the DPF.

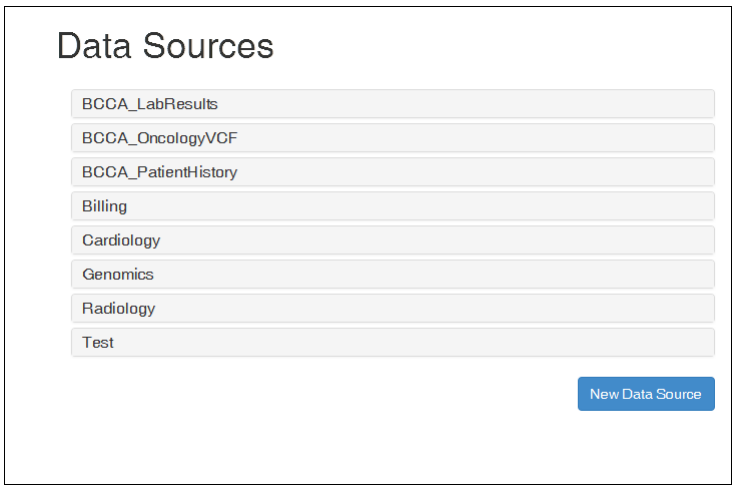
View Data Source Information

View data source information from the **Data Sources** page.

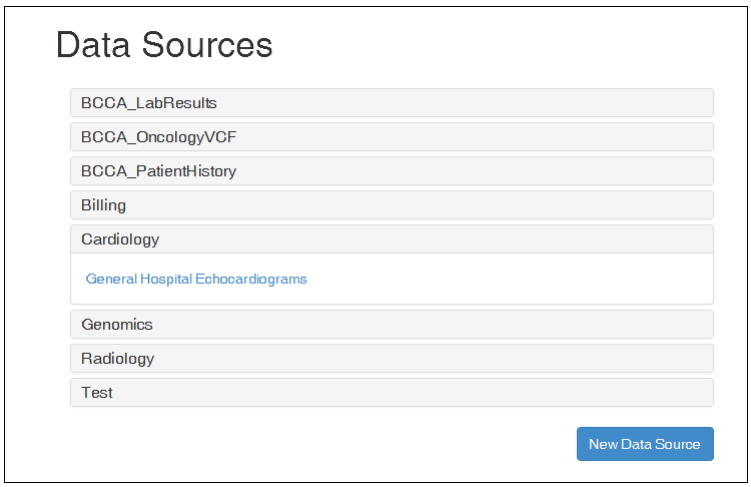
To view information about a particular data source:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon. in the left navigation bar.

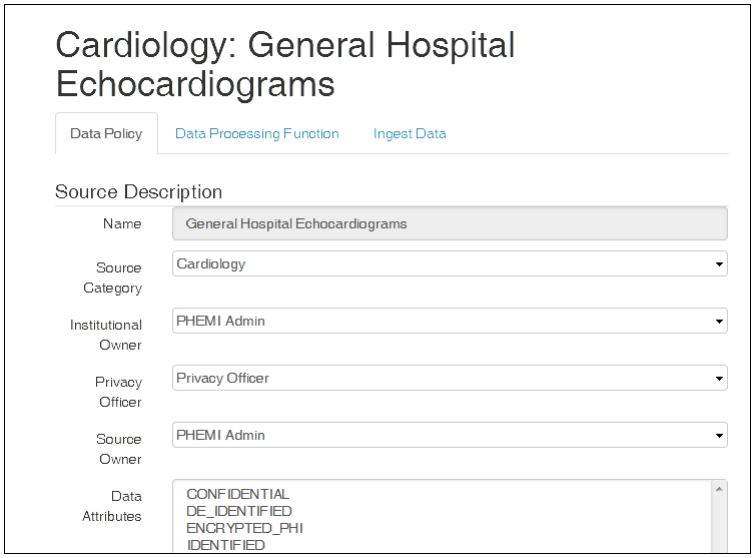
   The **Data Sources** page opens showing defined data categories.



2. Click the data category that contains the data source, so that it expands and shows the data sources in the category.



3. Click the data source name. The page for the data source opens on the **Data Policy** screen.



4. Do any of the following:

- View data policy information on the **Data Policy** tab.
- Click the **Data Processing Function** tab to see the DPF associated with this data source.

- Click the **Ingest Data** tab to access a screen where you can manually ingest files. *How do I manually ingest files?*

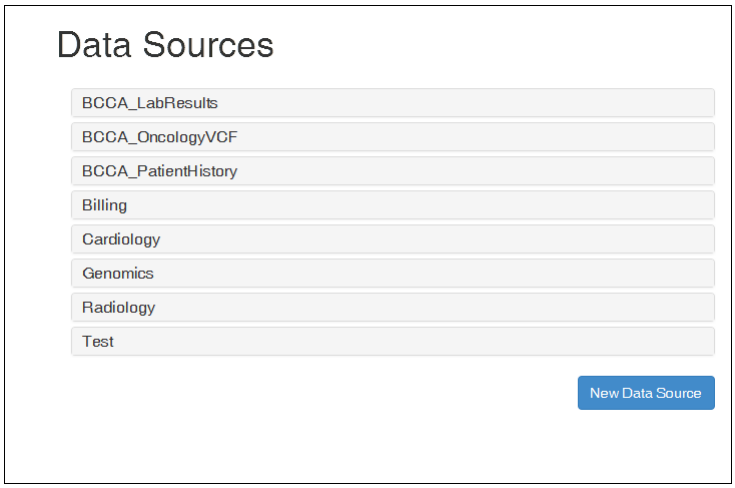
Modify Data Source Information

Modify data source information from the **Data Sources** page.

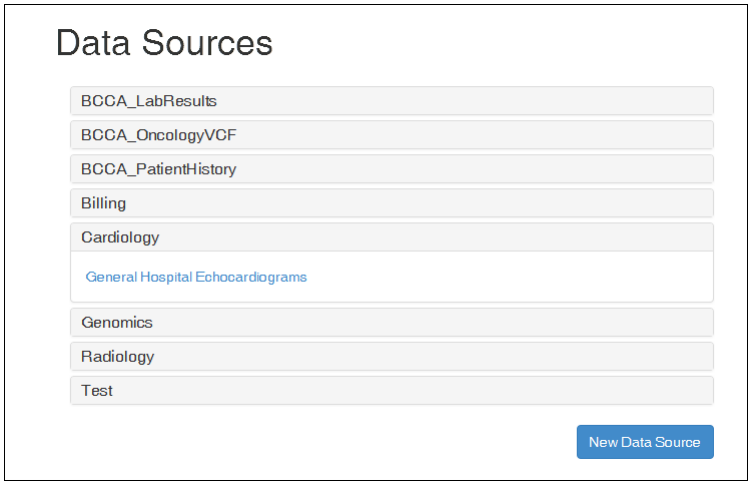
To modify information for a particular data source:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon. in the left navigation bar.

   The **Data Sources** page opens showing defined data categories.

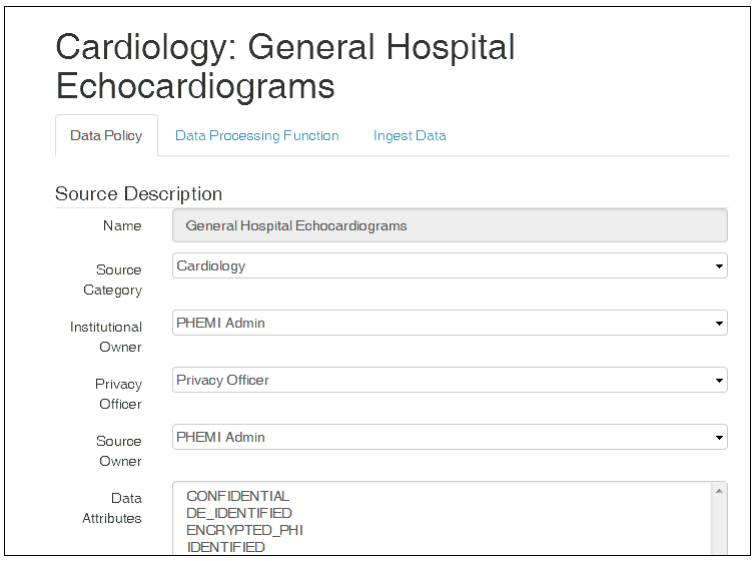


2. Click the data category that contains the data source, so that it expands and shows the data sources in the category.



3. Click the data source name.

   The page for the data source opens on the **Data Policy** screen.



4. Do any of the following:
   - Modify data policy information on the **Data Policy** screen. *How do I define a data policy?*
   - Click the **Data Processing Function** tab to change the DPF associated with this data source. *How do I define the DPF?*
   - Click the **Ingest Data** tab to access a screen where you can manually ingest files. *How do I manually ingest files?*

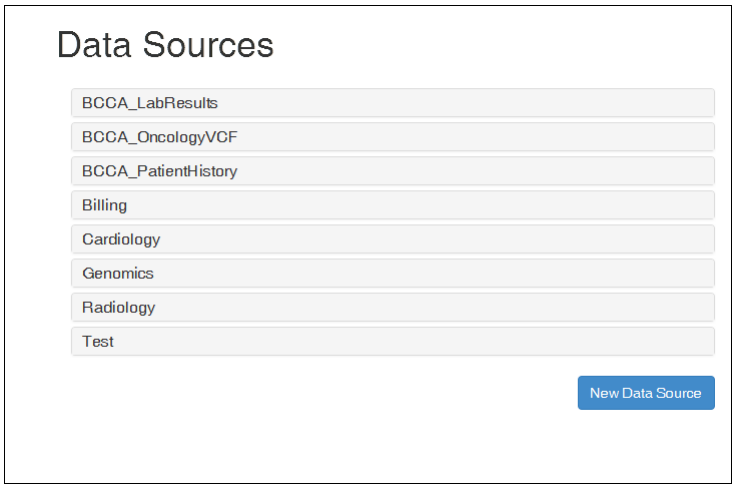
Delete a Data Source

Delete a data source from the **Data Sources** page.

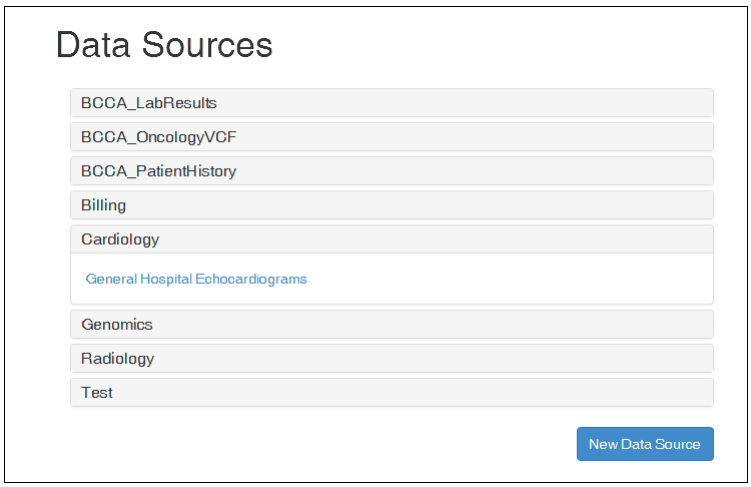
To delete a data source:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon. in the left navigation bar.

   The **Data Sources** page opens showing defined data categories.

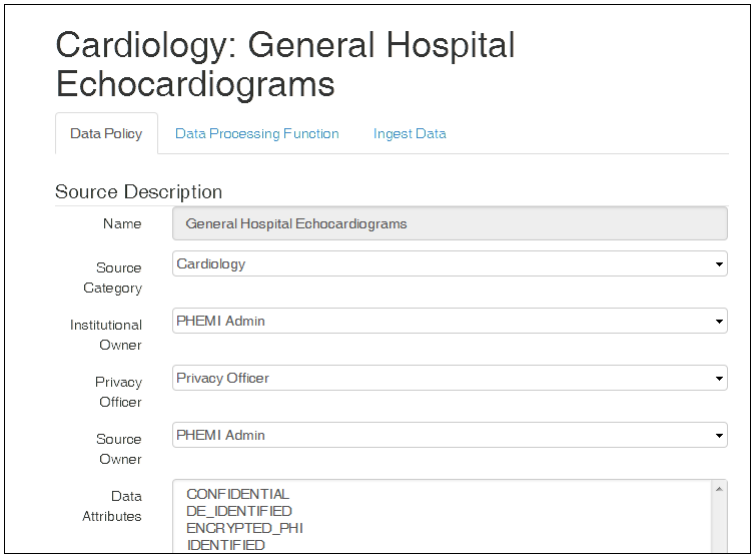


2. Click the data category that contains the data source, so that it expands and shows the data sources in the category.



3. Click the data source name.

   The page for the data source opens on the **Data Policy** screen.



4. Click the **Delete Data Source** button.

   The system asks you to confirm permanent deletion of the data source. Click **Delete**.

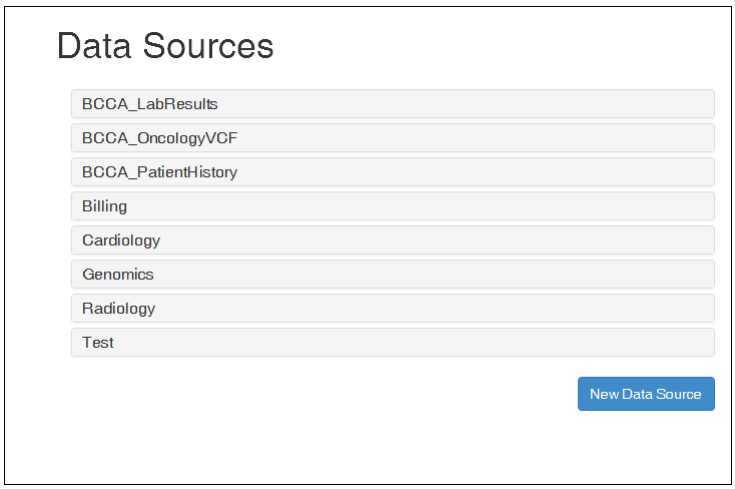
Manually Ingest Files

You can quickly manually ingest data files from a data source into PHEMI Central from the **Data Sources** page.

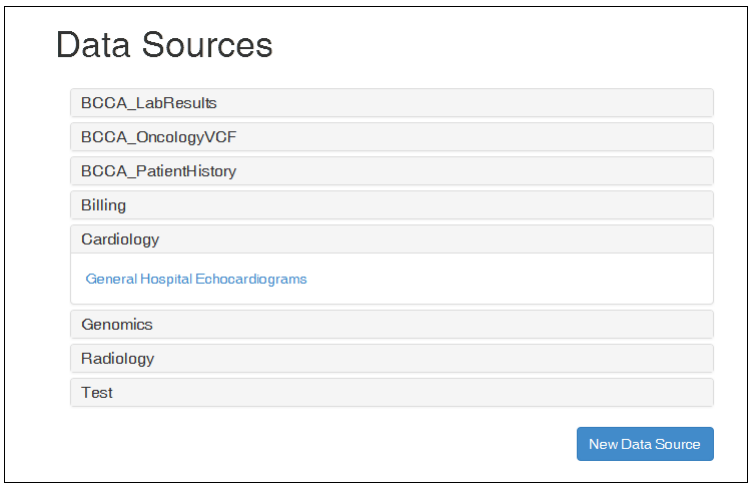
To manually ingest files:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon in the left navigation bar. 

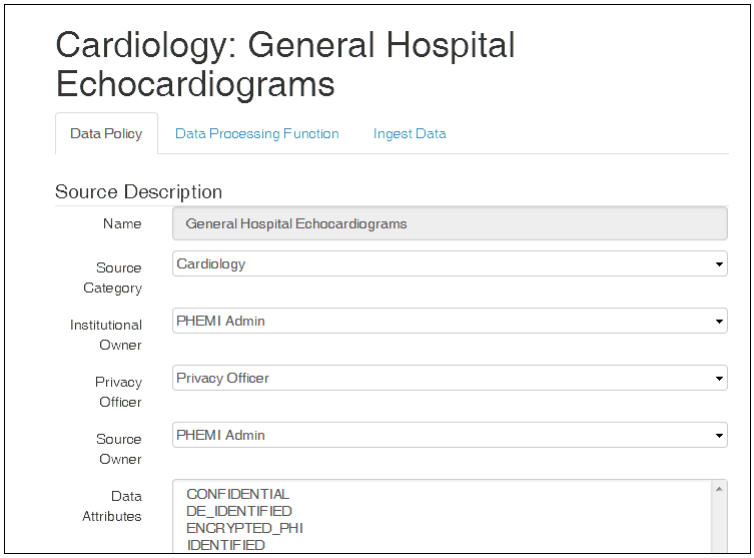
   The **Data Sources** page opens showing defined data categories. *Tell me about data categories.*



2. Click any data category to expand the category and see the data sources included in the category.

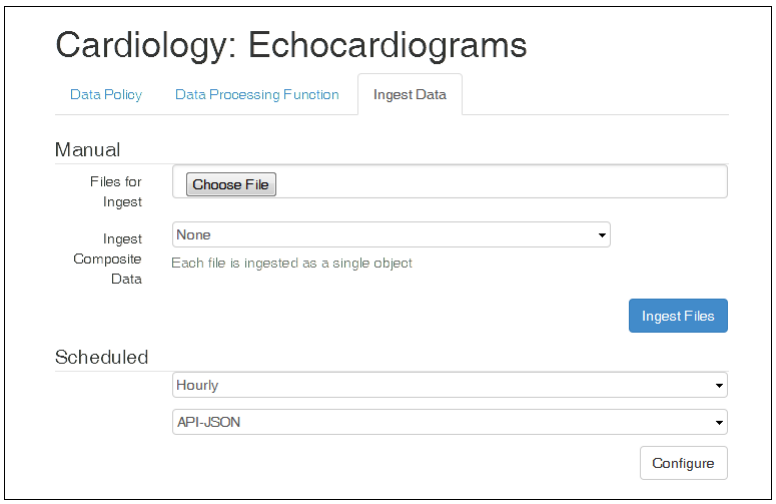


3. Click the data source name. The **Data Policy** screen for the data source opens.



4. Click the **Ingest Data** tab.

   The **Ingest Data** screen opens.



5. Click the **Choose File** button (or **Browse** button, depending on your browser). Navigate to the folder and select the file or files you want to ingest. Click the **Choose** or **Open** button to set your selection.
6. For structured or composite file such as a ZIP files or CSV files, if you want the PHEMI system DPFs to automatically process the file on ingest, click the drop-down arrow in the **Ingest Composite Data** field and select the file type from the list. If you want PHEMI Central to store the original data along with the derived data items, check the **Store Original File** checkbox.
7. Click the **Ingest Files** button.

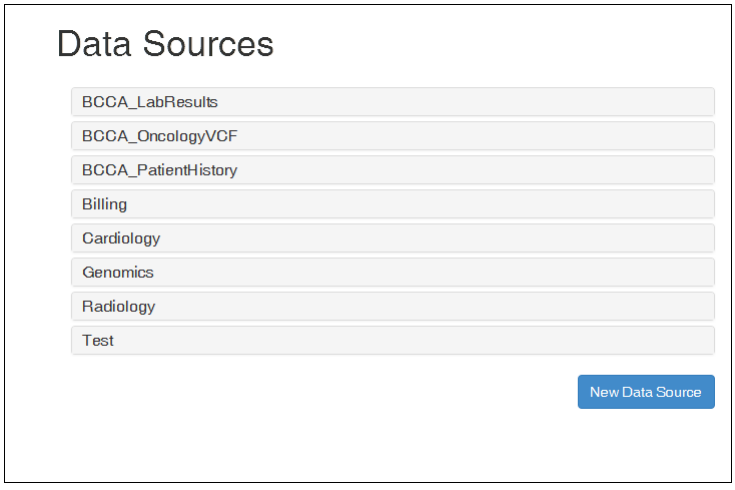
Schedule File Ingest

You can have the Management and Governance Console ingest files on a schedule from the **Data Sources** page.

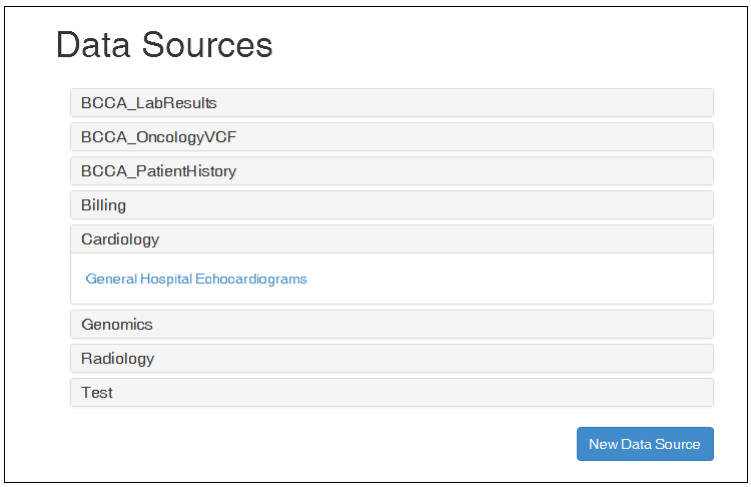
To schedule file ingest:

1. Open the **Data Sources** screen, by clicking the **Data Sources** icon in the left navigation bar. 

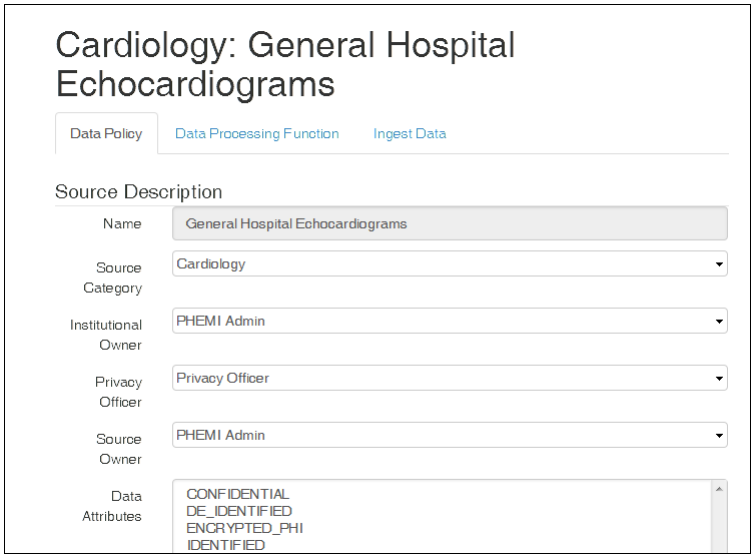
   The **Data Sources** page opens showing defined data categories. *Tell me about data categories.*



2. Click any data category to expand the category and see the data sources included in the category.

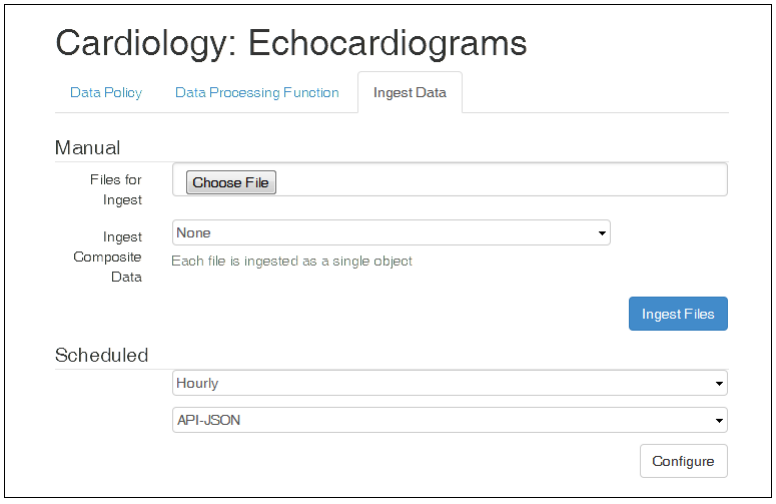


3. Click the data source name. The **Data Policy** screen for the data source opens.



4. Click the **Ingest Data** tab.

   The **Ingest Data** screen opens.



5. Click the **Choose File** button (or **Browse** button, depending on your browser). Navigate to the folder and select the file or files you want to ingest. Click the **Choose** or **Open** button to set your selection.

6. For structured or composite file such as a ZIP files or CSV files, if you want the PHEMI system DPFs to automatically process the file on ingest, click the drop-down arrow in the **Ingest Composite Data** field and select the file type from the list. If you want PHEMI Central to store the original data along with the derived data items, check the **Store Original File** checkbox.
7. In the **Scheduled** pane, choose from **Hourly**, **Daily**, or **Weekly** ingest schedule.
8. Select one of **API-JSON**, **SFTP**, **SMB/Samba**, or **ODBC** as the format of the ingested files.
9. Click the **Configure** button.
10. Click the **Ingest Files** button.

    PHEMI Central ingests the files and continues to ingest files automatically on the specified schedule.