# PHEMI Central Management and Governance Console User Guide

# Contents

**Introducing PHEMI Central..... 4**
Data in PHEMI Central..... 4
Collection..... 5
Curation..... 5
Consumption..... 7
Privacy, Security, and Governance..... 8
Privacy..... 9
Security..... 9
Governance..... 10
Data Management..... 10
Metadata Framework..... 10
Lifecycle Management..... 11
Data Immutability..... 11
Version Control..... 11

**Submitting Data to PHEMI Central..... 12**
Using the RESTful API..... 12
Using Manual Ingest..... 12
Using Bulk Ingest..... 12
Using ETL Tools..... 12

**Administrator Quick Start..... 13**
Before You Configure..... 13
System Installation..... 13
Define a Governance Policy..... 13
Initial Configuration Workflow..... 13

**Introducing the Management and Governance Console..... 15**
Logging On and Off..... 15
Quick Tour of the Management and Governance Console..... 15

**Management and Governance Console Reference..... 17**
Quick Tasks..... 17
Change Your Password..... 17
Get System Version Information..... 17
Log Off..... 18
Data Collections..... 18
Data Policies..... 19
Data Processing Functions..... 19
View Data Collections..... 20
Define a Data Collection..... 21
View Data Collection Information..... 25
Modify Data Collection Information..... 26
Delete a Data Collection..... 27
Manually Ingest Files..... 28

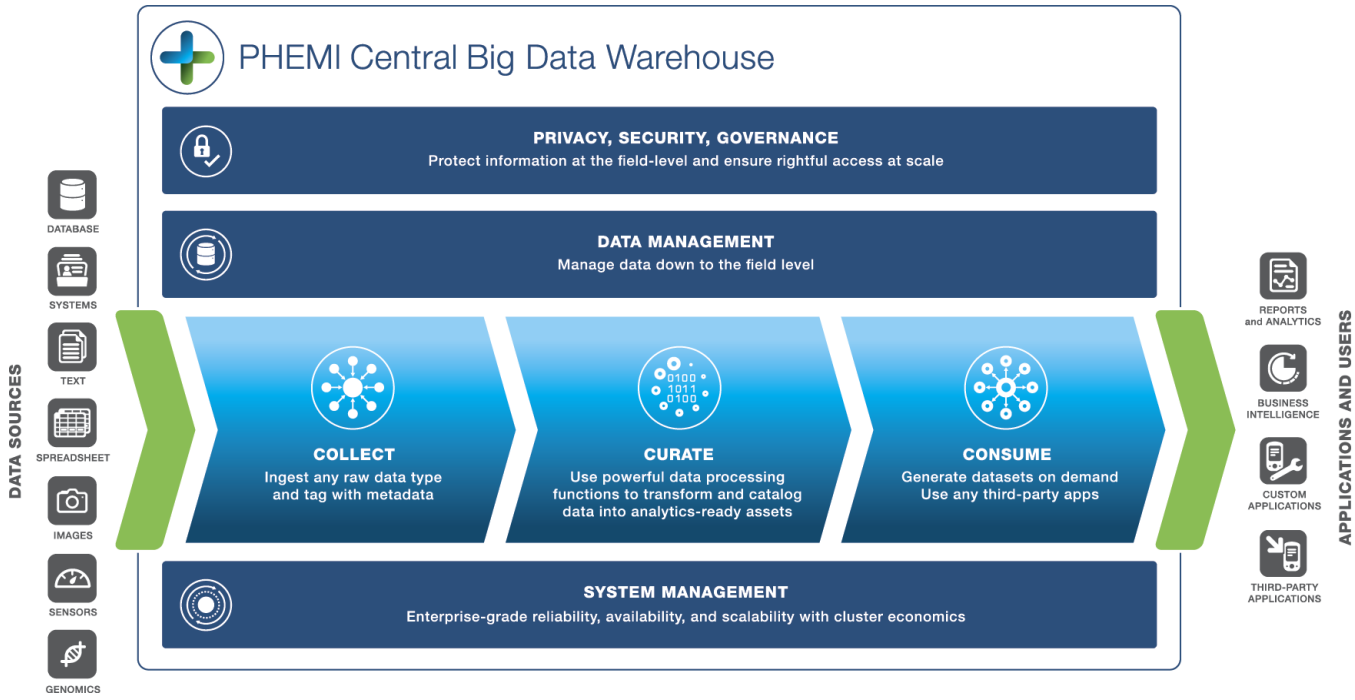
Datasets ... 30
View Defined Datasets ... 30
Define a Dataset ... 31
View Dataset Information ... 33
Modify a Dataset ... 34
Execute a Dataset ... 35
Delete a Dataset ... 36
Access Policies ... 36
View Existing Access Policies ... 37
Create an Access Policy ... 37
View Access Policy Information ... 38
Modify an Access Policy ... 39
Delete an Access Policy ... 40
System Metrics ... 41
View Global Metrics ... 41
View Data Metrics ... 42
View System Performance ... 43
Monitor System Tasks ... 44
Users ... 46
User Roles ... 46
View System Users ... 46
Create a New User ... 47
View User Information ... 48
Modify User Information ... 49
Delete a User ... 50
The Object Browser ... 51
View an Object ... 52
Delete an Object ... 53
Audit Log ... 55
View the Audit Log ... 55
System Configuration ... 56
The Password Policy ... 56
Dataset Destinations ... 58
Data Retention Behavior ... 68
Data Categories ... 71
Data Visibilities ... 76
User Authorizations ... 83

**Glossary of Terms and Concepts ... 89**

# Introducing PHEMI Central

PHEMI Central is a big data warehouse that offers big data capability with fully integrated privacy, security, and governance and advanced data management functionality.

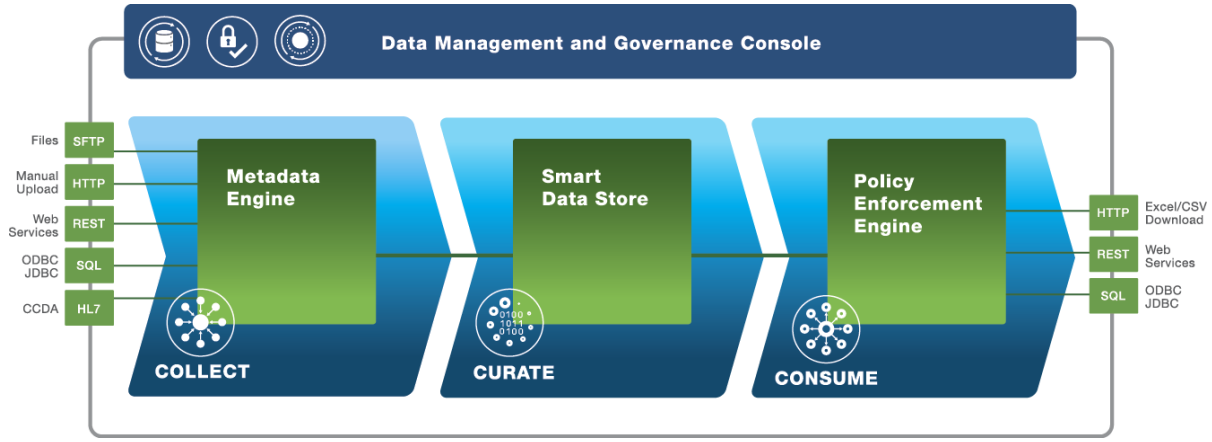
PHEMI Central allows organizations that need to protect and govern the use of their information to take advantage of big data technology to access, catalogue, and analyze their digital assets at speed and scale.




## Data in PHEMI Central

Data in PHEMI Central follows a lifecycle of collect, curate, and consume.

Throughout the data lifecycle, data is managed according to the organization's governance policies.

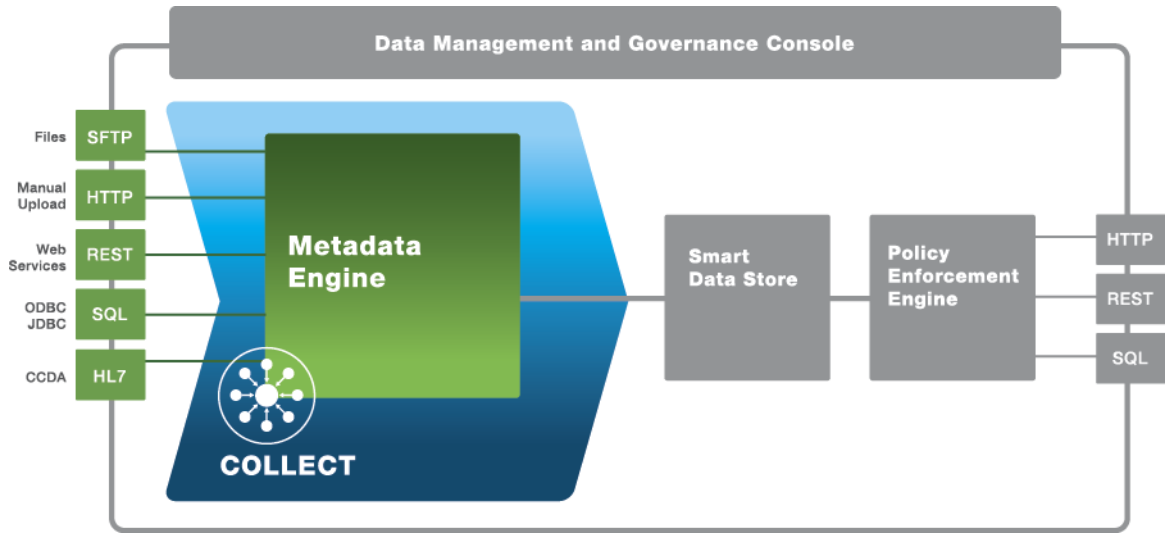


## Collection

PHEMI Central can collect, or "ingest," any type of data.

Data collections can include any data type from small kilobyte messages to large terabyte files.

- **Database records**—Data extracted from information systems, databases, and so on.
- **Structured non-relational data**—Spreadsheets, GIS datasets, genomics, machine data, XML, JSON, HL7, and so on.
- **Semi-structured files**—ECGs, tabular documents, and so on.
- **Unstructured files and datasets**—Images, consult letters, reports, e-mails, customer feedback, social media, and so on.




Data can be ingested and aggregated from multiple disparate sources, bringing together and consolidating data silos. Data can be ingested into in a variety of ways:

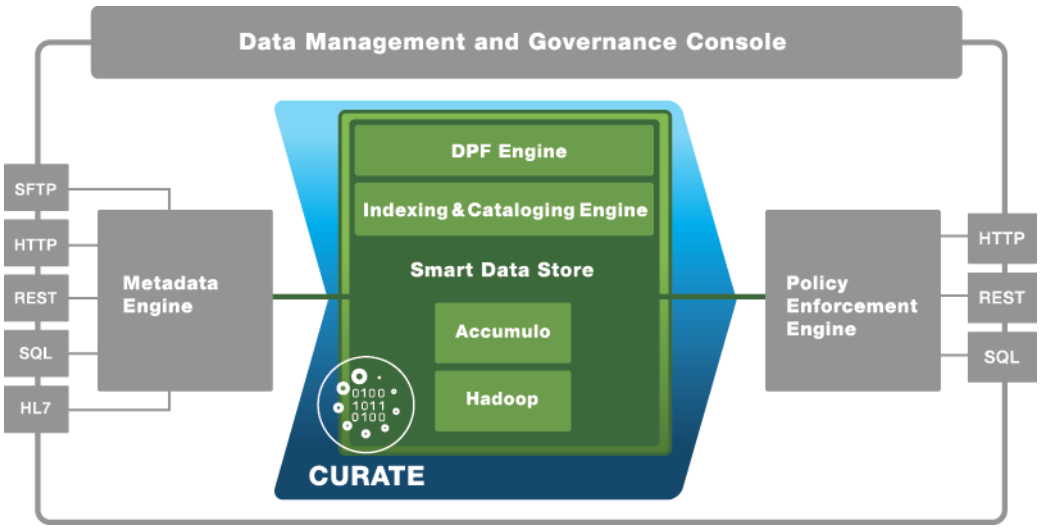
- **Stream**—Machine-to-machine data collections, such as telemetry and hospital bedside monitors, can stream data to PHEMI Central by means of the PHEMI RESTful API.
- **Push**—Data collections and extract, transform, and load (ETL) tools can publish to PHEMI Central using either JDBC or the PHEMI RESTful API.
- **Pull**—Custom connectors based on the PHEMI RESTful API can be deployed to allow PHEMI Central to fetch data from sources.
- **Manually Ingest**—Files can be manually uploaded to PHEMI Central from a standard browser window.
- **Store by Reference and Action**—PHEMI Central can reference remote data or a remote dataset through a URL, stored procedure, SQL query, external table, or the RESTful API. Applications can also be stored and executed, causing external tables or external data to be accessed and pre-processed.

During ingest, PHEMI Central tags each raw data object with metadata that describes the data. Metadata governing digital rights management, retention rules, data sharing agreements, and privacy policies is applied and enforced. Describing information with metadata means that users and applications can query and analyze data based on the data's properties, instead of having to navigate complex directories or schemas to find information. PHEMI Central then places the tagged digital asset into the data store for curation.

## Curation

PHEMI Central's Smart Data Store converts the raw data into analytics-ready digital assets.

PHEMI Central integrates the capabilities of the Hadoop/Accumulo ecosystem with a powerful metadata framework, with indexing and cataloging capabilities, and with an innovative Data Processing Function framework to create a Smart Data Store that transforms your raw data into analytics-ready digital assets.




### Schemaless Storage

PHEMI Central is a key-value store that's graph-based and schemaless.

In a traditional system, data is designed into a file system hierarchy or a database schema. So long as the schema or file system is in force, data must comply with it. If the design does not scale or if the requirements change, migration can be complex and costly.

Unlike schema-based data stores, data in PHEMI Central's store is distributed and based on key-value pairings. Data is stored in a binary format that is unaffected by any schema in source or destination systems. Schemaless storage offers benefits in several situations:

- If the schema of the source or destination system changes
- If the characteristics of your data change
- If the requirements of a user or an application change
- If a new, disparate data source needs to be brought online

### Indexing and Cataloging

PHEMI Central automatically indexes and catalogs all ingested data. The tagged, cataloged, and indexed raw data object is the simplest type of digital asset.

User-defined DPFs enable deeper and more sophisticated indexing and cataloging, while second-order indexes and graph relationships allow data analysts to quickly find and build datasets across digital assets. Linking datasets with common keys makes it possible to build meaningful datasets across disparate data collections, turning the data lake into a set of findable, searchable, easy-to-query, and analytics-ready digital assets.

### DPF Framework

A Data Processing Function (DPF) is an executable piece of code, written in any modern programming language, that transforms the original raw data into analytics-ready digital assets specifically targeted for your organization’s needs.

The DPF supplies the instructions for parsing the raw data (for example, a log message or medical report), extracting key content (for example, a blood glucose measurement) and performing data cleansing and enhanced indexing and cataloging. The DPF also structures data according to the organization’s needs. The result is data description at the element level that embeds the rules and policies governing the data collection, and embeds configured properties such as the data collection ownership, its time to live according to the data collection's retention policy, and what visibility the element should have.

Standard PHEMI system DPFs are included that index and describe structured data, such as spreadsheet files, database records, or XML/ JSON documents. User-defined DPFs can be developed for advanced needs, such as analysing semi-structured data or performing natural language processing on free text. Or, DPFs can catalog and standardize data into ontologies such as SNOMED or LOINC, making it easier for data analysts to find the right information in the right format.

DPFs can also analyze streams of machine data to find patterns and exceptions, calculate aggregates, and convert streaming data for trending and predictive analysis. For extracting information from unstructured documents such as scans or X-rays, the DPF can include specialized parsing functions, like Optical Character Recognition (OCR) or image parsing. As the organization's needs evolve and as knowledge advances, DPFs can be updated or redeveloped and re-executed on existing or historical data to extract new or different information.

PHEMI Central's DPF Framework manages DPF deployment and execution across the entire system. A DPF code library is associated with a data collection by uploading it into PHEMI Central. The code is executed by the PHEMI Central DPF Engine. PHEMI Central manages DPF execution across all datasets and all data elements within the system.

### Data Linking

The indexing, cataloging, and graph relationships PHEMI Central generates allow you to make connections, or links, among data items.

Data linking allows you to connect disparate data and data that might have been isolated in silos. For example, imagine you ingest a patient history from a family doctor, a scan of prescription information from a pharmacy, Medical Resonance Images (MRIs) from a hospital, and X-ray images from a medical laboratory. If data elements are tagged with appropriate metadata, you can link all this disparate data (for example, with a patient ID number) for use in various ways.

Graph-based data linking means that you can query and analyze a more complete picture of your data so that you can see, at scale and efficiently, relationships between objects in the system.

### Data Dictionary

A data dictionary cleanses data by identifying and saving a common interpretation of selected types or fields.

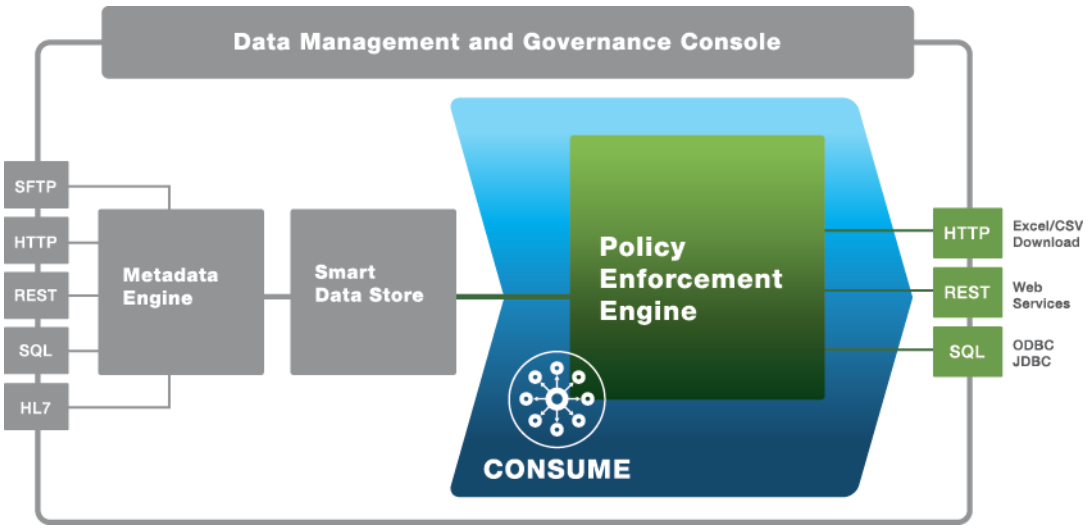
Disparate data collections may have fields that occur in common but are named differently or use different format conventions. For example, one data collection might have a field called "Sex" with values "M" and "F," while another might have a field called "Gender" with values "Male" and "Female." Similarly, different medical imaging systems might use different terminology and conventions for the same concepts and measurements.

You can develop a DPF for your data that acts as a data dictionary, to standardize and cleanse data. Cleansing data with a data dictionary greatly simplifies query and analysis.

## Consumption

The data elements stored in PHEMI Central is accessed by querying the system. Access can be made in a number of ways.

- You can locate and download the original data object using the PHEMI Central Management and Governance Console Object Browser.
- You can query a data collection or dataset using the PHEMI RESTful API.
- You can create a dataset and download it into Excel, CSV, or TSV format.
- You can export a dataset to a portal, tool, or application, using the RESTful API or a JDBC/ODBC connector.
- You can export a dataset to SAP HANA using the SAP Smart Data Access connector.

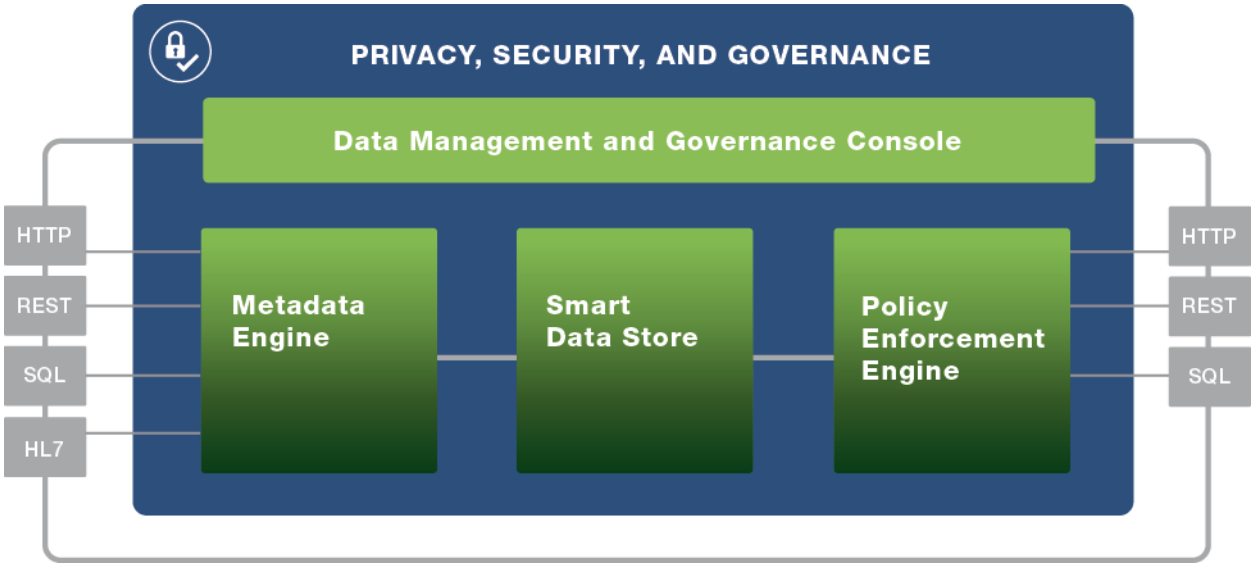



In all cases, PHEMI Central's Policy Enforcement Engine strictly enforces your organization's privacy and security policies to ensure rightful access to data.

## Privacy, Security, and Governance

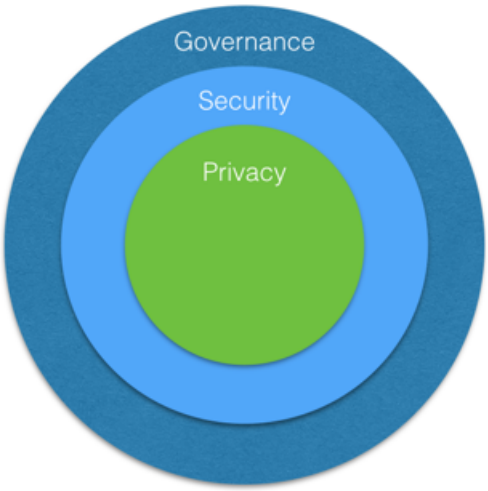
The purpose of privacy, security, and governance is to protect information and ensure rightful access.

PHEMI Central is designed from the ground up to be able to manage crucial aspects of privacy and security, and to be able to accurately reflect your organization's governance policies.




Privacy, security, and governance are not all the same thing.

- Privacy is restricting information access to those who have the right to access it.
- Security is the means by which you maintain privacy and protect information assets.
- Governance is the set of processes, roles, policies, controls, and metrics that an organization develops and implements around information to manage its privacy and security.




## Privacy

Privacy is restricting information access to those who have the right to access it.

PHEMI Central's Privacy by Design framework was designed from the ground up to define, manage, and enforce data sharing agreements and privacy policies. This framework includes the following mechanisms:

- **Attribute based access control (ABAC)**—Users are tagged with attributes that describe their authorizations to access data. Data is tagged with attributes that describe what its visibility should be. These two attributes are used in access policies that are applied to data collections and datasets to enforce rightful access privileges. For example, a data analyst with CONFIDENTIAL authorization might be able to export fully identified data, while an analyst with RESEARCHER authorization might only have access to de-identified data.

  Attribute based access control reduces complexity and reduces the risk of data breach. An attributed based approach to privacy is also especially helpful when not all uses or access requirements for data are understood up-front, or when new types of data are frequently introduced into the system (both common scenarios in health care, for example).
- **Selective data tagging**—The attribute-based access configured in the system can be enriched and expanded with context-specific protections by using the Data Processing Function framework to extract and re-tag information. For example, scans of patient reports can be recognized and extracted by a DPF and fields selectively marked as PII (personally identifying information, as in a Social Security or Social Insurance Number) or NON_IDENTIFYING (as in a blood glucose measurement).
- **Automatic anonymization and de-identification**—PHEMI Central can be set to automatically invoke a Data Processing Function that can de-identify, encrypt, redact, or mask any data element. A DPF can even include sophisticated data dependency algorithms to reduce the risk of re-identification.

  A PHEMI Administrator can also construct datasets that strip out identifying data elements. Centralizing anonymization and de-identification helps reduce data sprawl and reduces the risk of data consistency errors.
- **End-to-end access policy enforcement**—Every query for data to PHEMI Central is mediated by the PHEMI Policy Enforcement Engine, which compares the access request against the privacy protections that have been placed on the data. At no time can users, applications, or external systems bypass the Policy Enforcement Engine to access data directly.

## Security

Security is the means by which you maintain privacy and protect information assets.

PHEMI Central includes a number of security mechanisms:

- **Role Based Access Control (RBAC)**—User roles determine what operations a user can perform. For example, only users with a role of PHEMI Administrator can configure the system and construct datasets, while only users with a role of Data Analyst can query data and execute or export a dataset.

- **Configurable Password Policy**—PHEMI Central allows you to configure the password policy that defines how strong user passwords have to be and how they must be changed.
- **Audit Log**—PHEMI Central maintains complete a audit log of system and user operations. The log includes all create, modify, and delete operations, plus a record of all queries made to the system. The audit log file is completely tamperproof for all users.
- **Encryption in motion**—PHEMI Central assumes your system is deployed on a trusted network. However, you can encrypt links from data sources and to consuming applications and analytics tools using either Secure Sockets Layer (SSL) or Transport Layer Security (TLS).

### Governance

Governance is the set of processes, roles, policies, controls, and metrics that an organization develops and implements around information to manage its security and privacy.

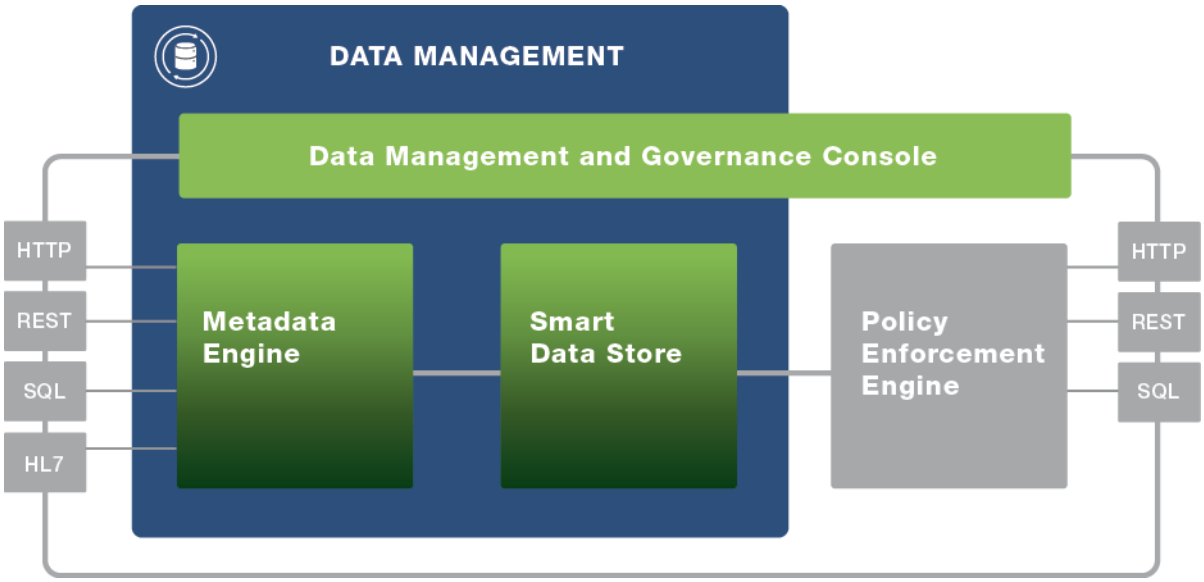
Information governance is about controlling and protecting an organization's data. The data may be sensitive, or perhaps it is important that the data be absolutely accurate, or perhaps the organization must achieve legislative and compliance targets. Data governance includes the process and policies around the protection, curation, and access to data and encompasses all of privacy protection, data security, lifecycle management, and data audit.

A governance policy is a coordinated approach to protecting data and assigning privileges to users. To control and protect your data, your organization should have a clearly defined policy governing data. The governance policy will drive how the PHEMI Administrator configured PHEMI Central.

## Data Management

Proper management of data through its lifecycle is critical as volumes grow and variety increases.

PHEMI Central includes advanced data management features such as version control, rollback, and retention rules. A sophisticated metadata framework allows information to be managed at the field level throughout its lifecycle. This family of features brings the data management capabilities of enterprise-grade traditional data warehouses to big data.




### Metadata Framework

The power and sophistication of PHEMI Central's data management capability arises from its powerful metadata framework, which extends across the system.

Metadata is applied on ingestion and enriched by cataloging, indexing, and invoking Data Processing Functions. The result is data description at the element level that embeds the rules and policies governing the element, as well as configured properties such as the data collection ownership, retention time (time to live), and what visibility the

element should have. This means that de-identification, encryption, and masking, and other privacy restrictions can be enforced per data item, at the cell level.

PHEMI Central's metadata framework, with its flexible distributed key-value store, means that system designers do not have to worry how to structure the system.PHEMI Central structures data automatically, on the fly. Data scales to large volumes while still providing fast access, and changes to requirements do not necessitate changes to design of the data store. Users and integrated applications benefit from the metadata because they can use simple queries based on the properties of the data, rather than having to navigate complex directories or schemas to find the data they seek.

## Lifecycle Management

PHEMI Central uses organization-specific retention rules to manage digital assets throughout their entire life cycle, from data creation through curation, usage, and end of life.

- Expiry dates are set for the data sharing agreement. Each data collection is configured with a data sharing agreement, which can be uploaded into PHEMI Central. Configuring the data sharing agreement includes optionally setting a start and end date. If the end date expires, users trying to access data receive a warning and PHEMI Central prevents them from accessing the system. The data itself remains in the system, but users are unable to access it unless the expiry date for the data sharing agreement is reset.
- Expiry dates are set for datasets. Each dataset is made available for a period of time set by the PHEMI Administrator. When this period expires, the dataset is flagged as expired and users are not able to execute the dataset. The original data remains in the system, but users are unable to access it unless the expiry date for the dataset is reset.
- Retention rules are specified for data collections. Each data collection is configured with retention rules. From the retention rules together with the ingestion timestamp, PHEMI Central calculates a time to live for every raw and derived data element. When the time to live expires, PHEMI Central purges data from the data store.

## Data Immutability

PHEMI Central stores all data in a write-only data system that is never modified.

Data is only deleted when its precalculated time to live expires, as derived from the organization's retention policy. This approach provides assurance of data integrity for audit and compliance requirements.

## Version Control

PHEMI Central has robust version control and rollback capabilities to ensure data is never lost, corrupted, or overwritten.

The system keeps a history of data revisions and allows administrators to trace changes over time, including the ability to audit who made changes and when, plus the ability to roll back changes if necessary. This design provides a complete history for audit and compliance requirements.

# Submitting Data to PHEMI Central

Data can be submitted to PHEMI Central using the PHEMI RESTful API, by manually ingesting it, by using FTP or SSH batch ingestion, or by using extract, transform, and load (ETL) tools.

## Using the RESTful API

If the data source is able to publish data, the system can be programmed to publish to PHEMI Central using the PHEMI RESTful API.

In REST-based ingestion, the client (that is, the data source or submitting system) sends an HTTP or HTTPS POST request. The POST request contains valid user credentials in JSON format in the payload body.

When the credentials are authenticated, PHEMI Central returns the session ID and URI for the session, as well as a session cookie. Once the session is established, the client can POST data to the appropriate data collection by referencing the data collection ID.

PHEMI Central listens for REST queries on port 80 (for HTTP) and port 443 (for HTTPS).

REST-based ingestion is useful in situations where a system submits smaller pieces of data very frequently. Since PHEMI Central always listens on the port, the client system can be set up with a scheduled task to submit the data as often as needed.

## Using Manual Ingest

You can use the Management and Governance Console to manually ingest data objects into PHEMI Central.

Manual upload is a good method when you have very large amounts of data such that HTTP/REST is not suitable (for example, gigabytes or terabytes of data), and/or data that needs to be ingested relatively infrequently.

## Using Bulk Ingest

Batch ingest of data is extremely fast. You configure a secure FTP or an SSH connection to allow a system to write data to a temporary landing space within PHEMI Central. PHEMI Professional Services will help you get this set up.

You can trigger the bulk ingest process remotely or you can use a scheduled task such as a cron job. Triggering the process launches a MapReduce job that inserts the bulk data into PHEMI Central at a very fast rate. The temporary files are then purged from the system.

## Using ETL Tools

Some data collections (for example, some databases) are not able to submit data directly to PHEMI Central. For such systems, extract, transform, and load (ETL) tools can be used to extract data from the source system and then use either REST-based ingest or bulk ingest, depending on the requirements.

# Administrator Quick Start

This section shows you the workflow for a first configuration of the PHEMI Central Management and Governance Console.

## Before You Configure

Before you start configuring PHEMI Central, the system must be installed and you should have your organization's governance policy at hand.

### System Installation

PHEMI Central may be shipped as an appliance, deployed on Amazon Web Services (AWS), or deployed on your organization's VMware environment. Regardless of the deployment type, PHEMI Professional Services will install and set up your base system of cluster nodes and your management node, and will install the PHEMI Central software on the cluster. Installation may include populating your system with some of your organization's pre-existing data.

PHEMI Professional Services will create the first PHEMI Administrator account for you on PHEMI Central and will provide you with the user name and password for the account, together with the URL of the PHEMI Central Management and Governance Console. At that point, you can log on to the Management and Governance Console.

### Define a Governance Policy

Make sure your organization's governance policy is defined before you begin configuration.

Your governance policy will drive how you configure PHEMI Central. Therefore, before configuring PHEMI Central, you should make sure you have your organization's governance policy available to you. If you are uncertain as to the appropriate data visibilities, user authorizations, or access policies to define for your organization, consult with your Information Officer, Privacy Officer, or with someone in a comparable role.

To define your organization's governance policy:

1. Identify the different sensitivity levels of the data you will be storing in PHEMI Central.

   You'll configure these as data visibilities. Data visibilities are attributes used to tag ingested data for purposes of privacy and access control. Examples of data visibilities are "CONFIDENTIAL," "SECRET," and "PII" (Personally Identifiable Information).
2. Identify what authorizations your organization assigns to different users to allow them to access data.

   Your user authorizations should reflect the actual permissions your personnel are granted to interact with data with different visibility.
3. Specify the allowed combinations of user authorizations and data visibilities.

   For example, a user with C_LEVEL authorization (perhaps a CEO, CIO, COO, or CTO) might be permitted to access CONFIDENTIAL data, a user with RESEARCH authorization might only be permitted to access only DE-IDENTIFIED data, and a user with DOCTOR authorization might be allowed to see data of all sensitivity. You'll capture these combinations as access policies.

## Initial Configuration Workflow

Since some configuration tasks must be completed before others can be performed, the PHEMI Administrator can use the workflow in this section to configure PHEMI Central.

In these first steps, order matters. You cannot create users until you have defined user authorizations. Likewise, you cannot create access policies without defining both user authorizations and data visibilities.

1. *Log on to the PHEMI Central Management and Governance Console*.
2. *Define user authorizations.*
3. *Create users.* Create at least one user with a role of PHEMI Administrator, at least one user with a role of Privacy Officer, and at least one user with a role of Data Analyst.
4. *Define data visibilities.*

Data categories must be added before you can define data collections. Access policies are optional for data collection configuration, but if you are using access policies, you must configure the policies before configuring the data collections.

If you are using a Data Processing Function (DPF) to generate derived data from raw data, you must prepare the DPF code archive before configuring the data collection and upload the DPF archive during data collection configuration.

5. *Add data categories.*
6. *Create access policies.*
7. *Set up your data collections.*

It is the Data Analyst, not the PHEMI Administrator, who executes and exports datasets. However, constructing datasets from available data elements and configuring dataset export destinations is part of system configuration, and these tasks are performed by the PHEMI Administrator.

8. *Build datasets.*
9. *Configure dataset destinations.*

# Introducing the Management and Governance Console

The Management and Governance Console is a web-based interface for configuring and monitoring PHEMI Central.

**Note:** Use either Apple Safari, Mozilla Firefox, or Google Chrome to access the PHEMI Central Management and Governance Console. Microsoft Internet Explorer is not supported.

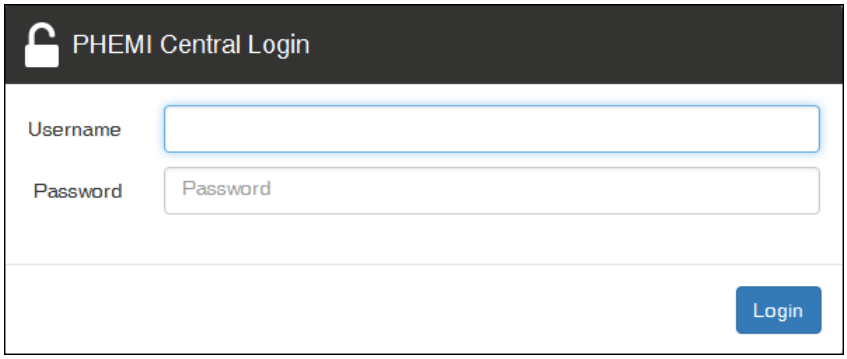
## Logging On and Off

The URL for the PHEMI Central Management and Governance Console is configured during physical installation of PHEMI Central. PHEMI Professional Services will provide you with the URL.

To log on to the PHEMI Central Management and Governance Console:

1. In the address bar of the browser, enter the URL configured for the PHEMI Central Management and Governance Console.

   The PHEMI Central login screen appears.




2. Enter your user name and password. Click **Login**.
   The PHEMI Central Management and Governance Console launches.The Management and Governance Console opens on the **System Metrics** dashboard.
3. Logging off the Management and Governance Console is described *here*.

## Quick Tour of the Management and Governance Console

The PHEMI Central Management and Governance Console consists of a set of main pages. Each main page includes screens where you can perform a set of related functions.

Note that your user role determines what part of the PHEMI Central Management and Governance Console you can access, and therefore what icons appear in the left navigation bar. Different roles have access to different parts of the Management and Governance Console.

Note also that the Privacy Officer does not have a functional role within the Management and Governance Console. In PHEMI Central, the Privacy Officer is the person chiefly responsible for defining the organization's overall governance policies. As such, you must define at least one user with a role of Privacy Officer to "own" that aspect of a data collection.

**Table 1:**

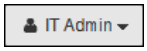
| Main Page | What you can do there: | You must be a: |
|---|---|---|
| Data Collections | Manage data collections; manually ingest files. | PHEMI Administrator. |
| Datasets | Build datasets; execute datasets; export datasets. | PHEMI Administrator to define datasets.<br>Data Analyst to execute and export datasets. |
| Access Policy Builder | Define access policies. | PHEMI Administrator. |
| System Metrics | View metrics about data collections, system performance, and tasks. | PHEMI Administrator. |
| Users | Manage PHEMI Central users. | PHEMI Administrator. |
| Object Browser | View or delete individual objects stored in PHEMI Central. | PHEMI Administrator. |
| Audit log | Monitor what operations have been performed from and what data requests have been made of PHEMI Central. | PHEMI Administrator. |

# Management and Governance Console Reference

This section steps you through all the functions and procedures in the Management and Governance Console.

## Quick Tasks

At the top right corner of the screen, a few tasks are always available from the quick task button. The quick task button is labeled with your username.

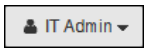



### Change Your Password

Change your password quickly by using the quick task button at the top of any screen.

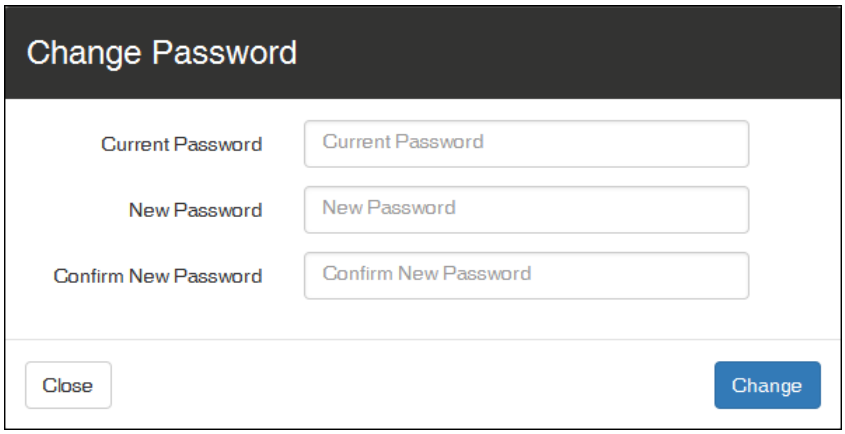
To change your password:

1. Locate the quick task button at the top right of the screen.

2. Click the drop-down arrow and select **Change Password**.

   The **Change Password** dialog opens.

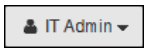
3. Enter your current password in the **Current Password** field.
4. Enter the new password in the **New Password** field. Retype your new password in the **Confirm New Password** field.
5. Click the **Change** button to submit the change.

### Get System Version Information

Get system version information from any screen by using the quick task button at the top of the screen.

To get system version information:

1. Locate the quick task button at the top right of the screen.

2. Click the drop-down arrow and select **About**.

   The **About PHEMI Central** screen opens, showing system version.

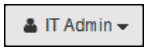
3. Click the **Close** button to dismiss the screen.

### Log Off

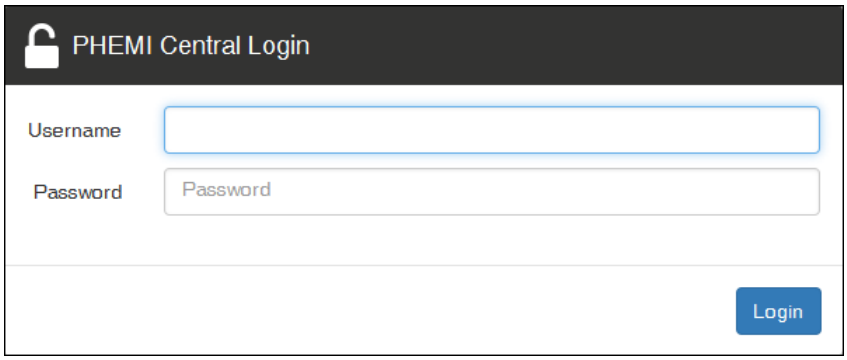
Log off from any screen by using the quick task button at the top of the screen.

To log off the PHEMI Central Management and Governance Console:

1. Locate the quick task button at the top right of the screen.




2. Click the drop-down arrow and select **Logout**. You are logged off the Management and Governance Console and the **Login** screen opens.




## Data Collections

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

Data collection configuration includes:

- Specifying data policy
- Associating the Data Processing Function that will process the data collection

Connection information to specific systems is not a part of data collection configuration. Data is "pushed" from the site submitting the data. PHEMI Central extracts any necessary connection information and login credentials from the session information.

*How do I submit data to PHEMI Central?*

Data collection configuration is set on the **Data Collections** page.

## Data Policies

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

The data policy describes what type and format of data is expected from the data source and classifies the data collection into one of the configured data categories. The data policy also lists the users who "own," or are responsible for, various aspects of the information.

The data policy also specifies the privacy settings for the data collection. It specifies which data visibilities, if any, to attach to data belonging to the collection. 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 the data collection, as well as every item of data derived from processing the data collection, will be tagged on ingestion or processing with the specified data visibility and access policy to protect privacy.

The data policy is also where the data sharing agreement is stored for reference. The data sharing agreement is a document that records permission for the data collection to be stored in PHEMI Central. You upload the agreement document and specify the time period for which the agreement is valid. If the data sharing agreement expires, the data remains in PHEMI Central's data store but users are prevented from accessing it.

The data policy also records the retention rules to be applied to information from this data source. The system uses the retention rules and looks at the ingest timestamp to calculate a time to live for each data element. When the time to live for a data item expires, PHEMI Central purges the raw data and any 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 allow ingested raw data items to be parsed, data elements to be recognized and extracted, and the derived data to be selectively tagged with metadata.

A Data Processing Function, or DPF, is an executable piece of code that supplies the instructions for processing raw data to extract 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 selectively 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.

### DPF Archives

The set of code that makes up a DPF is called a DPF archive. 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 collection, the DPF archive is ``registered`` with the data collection by uploading the DPF archive. Registering the DPF is part of data collection configuration.

### System DPFs

PHEMI Central comes with a library of system DPFs.

### Excel Workbook DPF

The Excel Workbook DPF takes an Excel workbook file and transforms it into derived data.

The Excel workbook is uploaded and transformed into a single JSON document. The JSON document is parsed into JSON objects, each representing one data record (row) of the Excel workbook. The resulting set of JSON objects is stored in PHEMI Central as a set of raw data.

The Excel DPF supports .xlsx and .xlsm file types. The Excel file must be comply with the expected format, where:

- Row 1 specifies the column names.
- Row 2 specifies the data type information for the data in the corresponding column.

- Row 3 specifies the data visibilities for the data in the corresponding column, as comma-separated list. The visibilities must be configured in PHEMI Central or an error will occur and no data will be ingested.
- Rows 4 to *n* contain the data to be ingested.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | PHN | PHN_Hashed | DoB | DoB_Masked | Weight | |
| 2 | STRING | STRING | DATE | DATE | DOUBLE | |
| 3 | PII | PII_HASHED | PII | PII_MASKED | NONPII | |
| 4 | 24568 | 0x293820ec2 | 5/1/1970 | 1/1/1970 | 78.1 | |
| 5 | 54786 | 0x2938fd2b3a | 12/31/1949 | 1/1/1949 | 55.4 | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |

### VCF DPF

The VCF DPF processes a VCF file.

A Variant Call Format (VCF) file is a text file containing tab-separated marker and genotype data. VCF data is used in bioinformatics to store gene sequence variations. As such, VCF files are very large, and a a VCF file can document hundreds, thousands, and even millions of gene sites in a single file.

Because a VCF data file is typically so large (on the order of gigabytes), a VCF file is always parsed into records while the file is being ingested. If you monitor ingestion of a VCF file (using the **Task Monitoring** pane of the **System Metrics**), you will always see the ingest task chained to the derive task for VCF files.

Once the ingestion process finishes, the file that was uploaded to the distributed file system is removed.

### Custom DPFs

A custom DPF is a DPF you develop especially to process your organization's data.

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. No MapReduce or YARN knowledge is 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 Collections

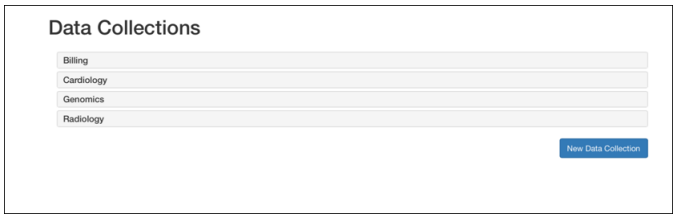
View data collections on the **Data Collections** page.

To view defined data collections:

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

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

   *Tell me about data categories.*




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

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

## Define a Data Collection

Define a data collection on the **Data Collections** page.

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

- Data categories *How?*
- Data visibilities *How?*
- Users *How?*

To be able to configure users, you must first configure user authorizations. *How?*

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 collection, you must first configure the access policy. *How?*

When you first define a data collection, 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:

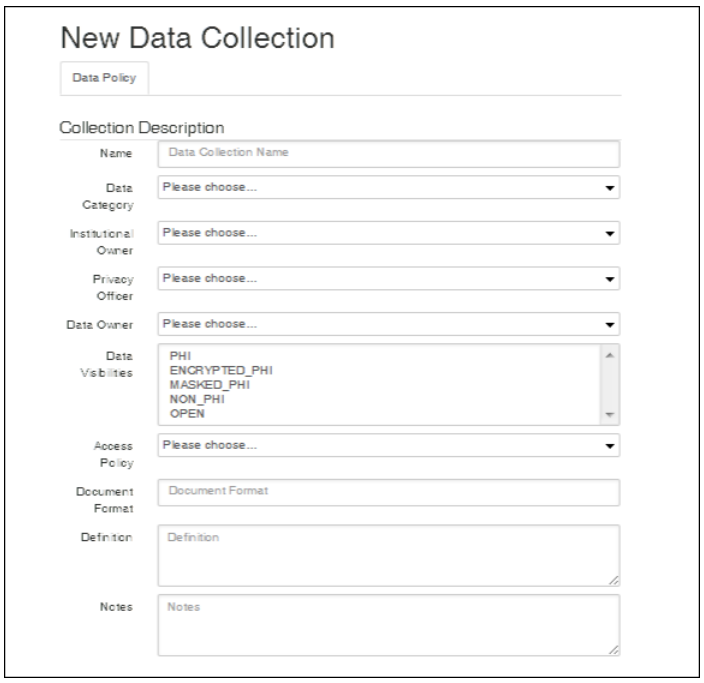
1. Open the **Data Collections** page, by clicking the **Data Collections** icon in the left navigation bar.

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

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

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

3. Describe the data collection.




| Option | Description |
|---|---|
| **Name** | Mandatory. A name for the data collection. Numbers, letters, spaces, hyphens, and the underscore character are supported.<br><br>**Note:** Once you save a data collection, the name cannot be edited. To change the name, you must delete the whole data collection and recreate it with the new name. |
| **Source Category** | Mandatory. The data category of the data collection. Choose from the drop-down list of data categories.<br><br>*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 Data Analyst are eligible. Choose from the drop-down list of eligible users. |
| **Source Owner** | Mandatory. The individual responsible for approving dataset requests involving this data collection. 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 collection. Choose from the list of configured data visibilities. You may apply zero, one, or many data visibilities; select multiple items by holding down the Shift or Control key. By default, no data visibility is applied.<br><br>*How do I define data visibilities?* |
| **Access Policy** | Optional. The access policy for enforcing rightful access to data from this data collection. Choose from the list of configured access policies. By default, no access policy is applied. |

| Option | Description |
|---|---|
| | *How do I create an access policy?* |
| **Document Format** | Optional. The kinds of document expected from this data collection. Examples are Microsoft Word, Excel, or JSON. |
| **Definition** | Optional. A brief description of the kinds of documents expected from this data collection. |
| **Notes** | Optional. Any additional notes for the data collection. |

4. Optionally, 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. When the data sharing agreement period expires, the data remains in the system but users will be prevented from accessing it. The format for the start and end dates is *mm-dd-yyyy*. If you do not specify a time period, the data sharing agreement remains in effect indefinitely.

5. Specify the retention rules. These rules are used to determine how long each item from this data collection are retained in the data store.

| Option | Description |
|---|---|
| **Please choose...** | Mandatory. Specifies when data should be purged from the system. Supported values are as follows:<br>• Retain for a time period. Keeps the data for the period specified. Specify some number of minutes, hours, days, weeks or years.<br>• Delete after time period. Purges the data after the specified time period. Specify some number of minutes, hours, days, weeks or years.<br>• Do not delete. The data is never purged from the data stored.<br>• 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** 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

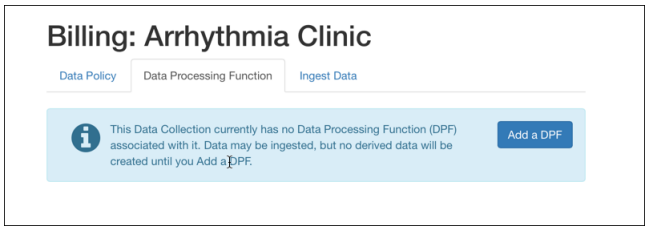
To associate a Data Processing Function (DPF) with a data collection, you upload the DPF archive onto the **Data Processing Function** screen of the **Data Collections** page.

*Tell me about DPFs and DPF archives.*

To associate a DPF with a data collection, start on the **Data Collections** page:

1. Click the **Data Processing Function** tab.

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



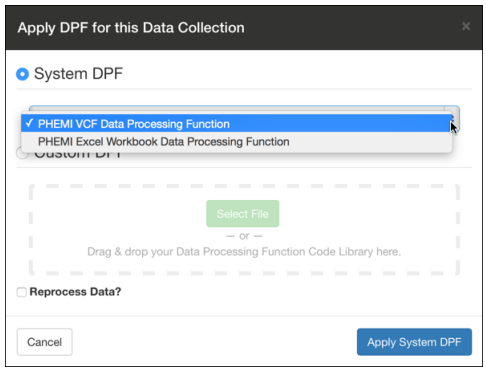

2. Click the **Add a DPF** button.

   The **Apply DPF for this Data Collection** dialog opens, with either the **System DPF** or the **Custom DPF** option selected.

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

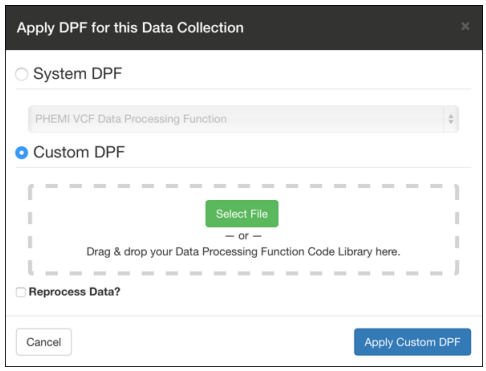
   - If you choose **System DPF**, use the drop-down arrow to select which system DPF you want to use. Choose between **PHEMI VCF Data Processing Function** and **PHEMI Excel Workbook Data Processing Function**.

     *What do these options mean?*



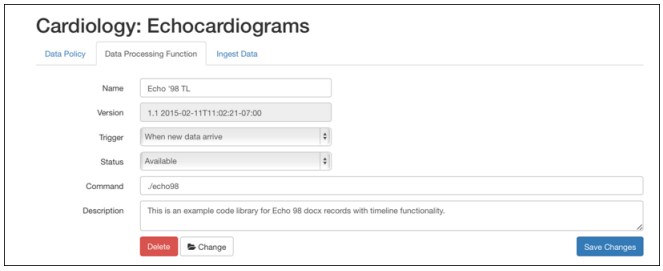

     Choose whether you want data to be reprocessed after the configuration change. Click the **Apply System DPF** button to apply the DPF.

   - 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.




     Select the **Reprocess Data?** checkbox if you want data to be reprocessed after the configuration change. Click the **Apply Custom DPF** button to upload the archive file and apply the DPF.

The system uploads the DPF, validates it, and applies it to the data collection. The **Data Processing Function** screen reappears, showing the DPF name, the DPF version, the default trigger condition, and the DPF description. All information except the trigger is automatically read from the DPF archive or derived by PHEMI Central's DPF Framework.




4. Set the trigger condition. The trigger condition determines at what point the DPF processes data.

| Option | Description |
|---|---|
| **When new data arrives** | The DPF processes data as it is ingested. If the DPF is the **PHEMI VCF Data Processing Function**, this is the only trigger condition available. (This is because VCF files are so large, they are always processed into individual records immediately on ingest.) |
| **Manually** | The DPF must be manually triggered.<br>When you choose this option, an **Execute** button appears next to the Trigger field. Click this button to manually execute a DPF on a data collection. |
| **Periodic** | The DPF executes on ingested data according to the specified schedule. When you choose this option, a scheduling field appears.<br>Set the DPF schedule to execute every **1 minute**, **5 minutes**, **10 minutes**, **30 minutes**, or **60 minutes**. |

The default behavior depends on what is coded into the DPF.

## View Data Collection Information

View data collection information from the **Data Collections** page.

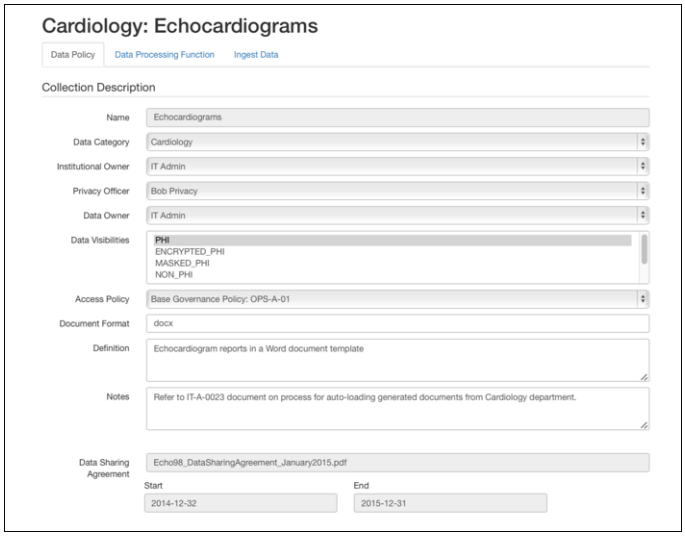
To view information about a particular data collection:

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

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

2. Click the data category that contains the data collection, so that it expands and shows the data collections in the category.

3. Click the data collection name. The page for the data collection opens showing the **Data Policy** screen.




4. Do any of the following:
   - View data policy information on the **Data Policy** tab.

     *What do these fields mean?*
   - See the DPF registered with this data collection by clicking the **Data Processing Function** tab.

     *How do I change the DPF registered with a data collection?*
   - Access a screen where you can manually ingest files by clicking the **Ingest Data** tab.

     *How do I manually ingest files?*

## Modify Data Collection Information

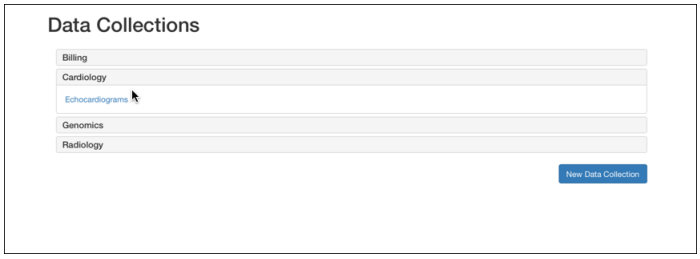
Modify data collection information from the **Data Collections** page.

To modify information for a particular data collection:

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

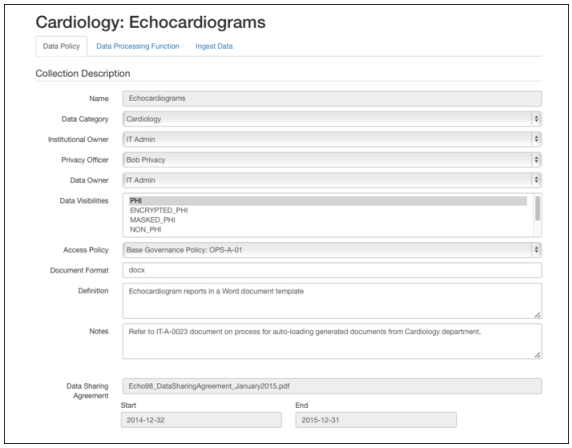
   The **Data Collections** page opens showing defined data categories.

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




3. Click the data collection name.

   The page for the data collection opens showing the **Data Policy** screen.




4. Do any of the following:

   - Modify data policy information on the **Data Policy** screen.

     *What do these fields mean?*
   - Click the **Data Processing Function** tab to change the DPF configuration.

     *How do I change the DPF registered with a data collection?*
   - Click the **Ingest Data** tab to access a screen where you can manually ingest files.

     *How do I manually ingest data?*

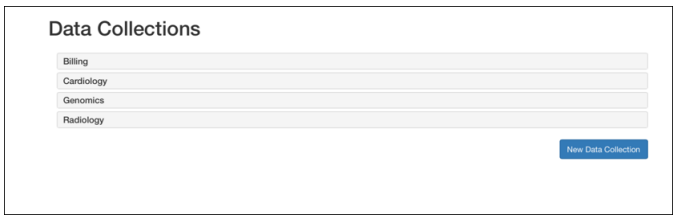
## Delete a Data Collection

Delete a data collection from the **Data Collections** page.

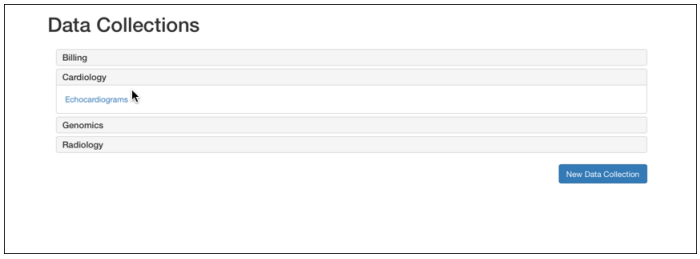
To delete a data collection:

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

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

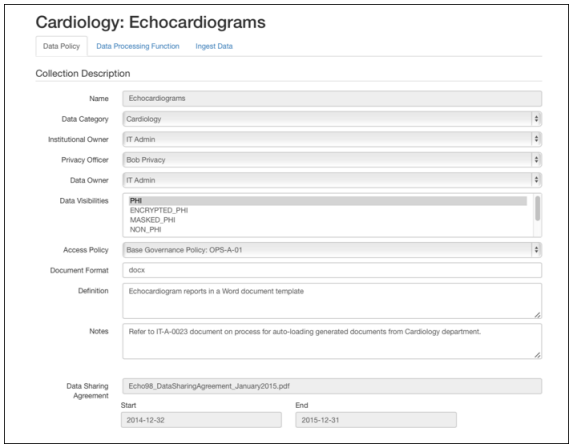



2. Click the data category that contains the data collection, so that it expands and shows the data collection in the category.




3. Click the data collection name.

   The page for the data collection opens showing the **Data Policy** screen.




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

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

## Manually Ingest Files

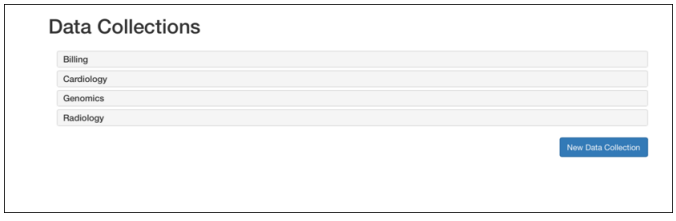
Manually ingest data files from a data collection into PHEMI Central from the **Data Collections** page.

To manually ingest files:

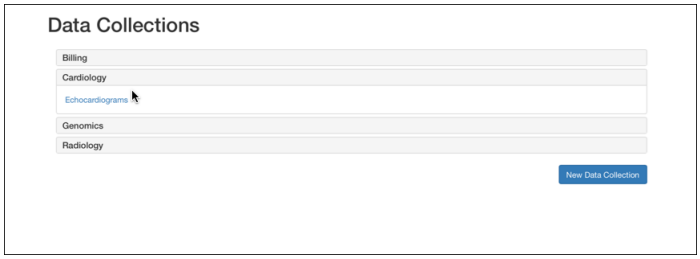
1. Open the **Data Collections** page, by clicking the **Data Collections** icon in the left navigation bar.

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

   *Tell me about data categories.*

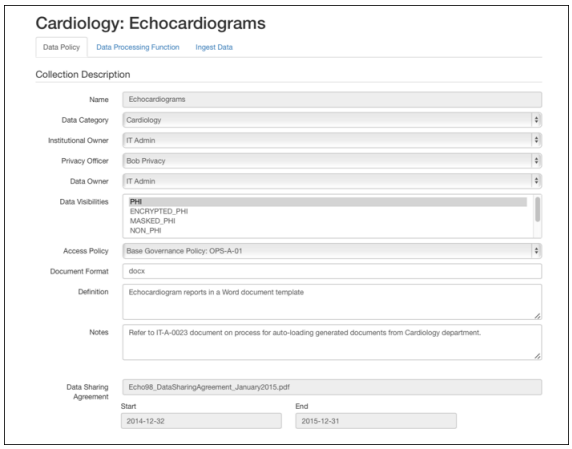



2. Click any data category to expand the category and see the data collection included in the category.




3. Click the data collection name.

The page for the data collection opens showing the **Data Policy** screen.

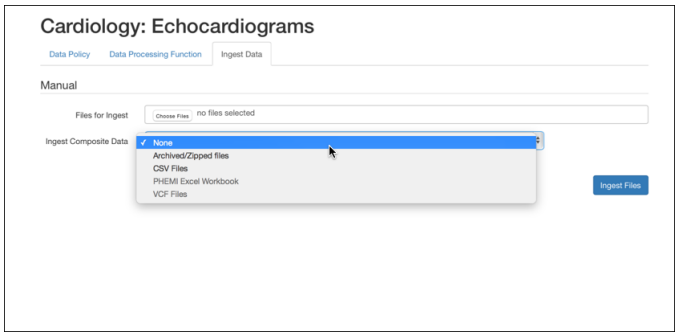



4. Click the **Ingest Data** tab.

   The **Ingest Data** screen opens.

5. In the **Files for Ingest** field, click the **Choose File** button (this may be a **Browse** button, depending on the browser you are using). 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 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.




| Option | Description |
|---|---|
| **Archived/Zipped files** | PHEMI Central extracts the files from the archive and ingests the archive contents. Supported formats as .tar.gz, .tar.bz2 and .zip. |
| **CSV files** | PHEMI Central extracts each row of the CSV file as a single object. Newline characters are interpreted as delimiting rows. Commas are interpreted as delimiting fields within a row. The first row is interpreted as the header row. Rows are ingested as XML documents. |
| **PHEMI Excel Workbook** | PHEMI Central extracts each row of the Excel file as a single object. The first row of the Excel file is expected to contain column names. The second row is expected to contain type information for the data. The third row is expected to specify the data's visibilities. An object is created for each row in the Excel file. |
| **VCF files** | PHEMI Central extracts each row of the VCF file as a single object. The VCF file is expected to contain a header section and at least one header row. The VCF file may be in submitted in text format or as a GZIP archive file. |

7. If you want PHEMI Central to store the original data along with the derived data items, check the **Store Original File** checkbox.

8. Click the **Ingest Files** button.

# 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 stored in PHEMI Central. You can construct 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 executed directly by users and downloaded or exported to spreadsheets, applications, or analytics tools.

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 constituent data collections.

When you attach an access policy to a dataset, the dataset bypasses any access policies protecting the constituent data collections and instead is constrained by its own access policy. For example, the access policy of a data collection may specify that only a user with Doctor authorization may see data with PHI visibility, while the access policy for a dataset including some of the same information may allow users with Nurse authorization to view PHI data.

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 constrained by data visibility and user authorization, as expressed 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.

If the PHEMI Administrator has not made the dataset available to you, you will not be authorized to execute it. In this case, the dataset is marked (not authorized). If you need to see this data, you must ask the PHEMI Administrator add you to the list of users who can access the dataset.

If the dataset has expired, the dataset is marked (expired). If you need to see this data, you must ask the PHEMI Administrator to reset the dataset expiry date.

## Define a Dataset

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

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

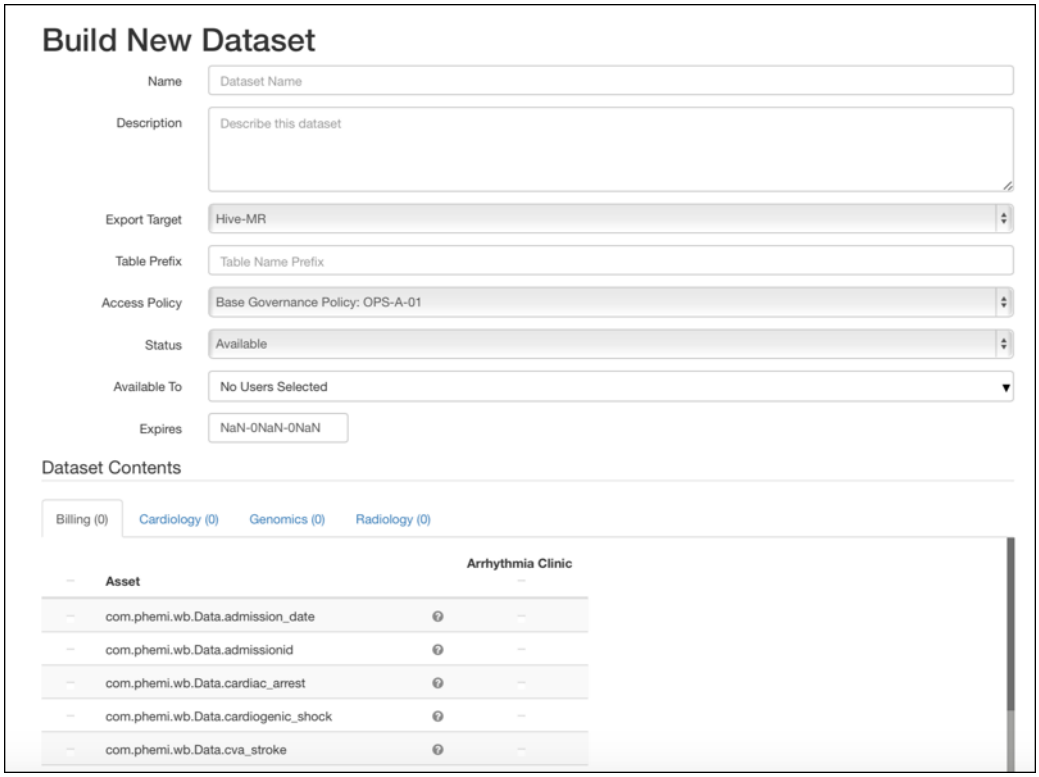
To define a dataset:

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.



3. Describe the dataset.

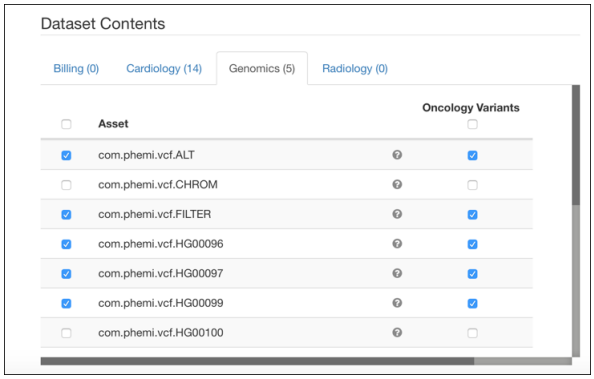
| Option | Description |
|---|---|
| **Name** | Mandatory. A name for the dataset. |
| **Description** | Mandatory. A brief description of the dataset. |
| **Export Target** | Mandatory. The format required by the system to receive the dataset. Supported values are as follows:<br>• Hive-MR. Apache Hive MapReduce format.<br>• JDBC. Java Database Connectivity format.<br>• CSV. Comma-separated values format.<br>• TSV. Tab-separated values format.<br>The default is Hive-MR.<br>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.<br>*Tell me about dataset destinations.* |
| **Table Prefix** | Optional. A database table prefix that can be prepended to data elements exported from this dataset. |
| **Access Policy** | Mandatory. The access policy for protecting the dataset. The access policy selected for the dataset may be different from any applied to the constituent data collection(s).<br>If you apply an access policy to a dataset, the dataset access policy overrides any access policy applied to constituent data collections. The access policies for the collections are bypassed.<br>*Tell me about access policies.* |
| **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.<br>Once a user has been granted access to the dataset, access to individual elements is constrained by the visibility of the element, the user's authorization, and the access policy applied to the dataset. |
| **Expires** | Optional. An expiry date for the dataset. When this date is reached, PHEMI Central disables the dataset and it can no longer be executed. |

4. Choose the fields you want in the dataset.

   All configured data collections and all derived data items are available to include in a dataset.

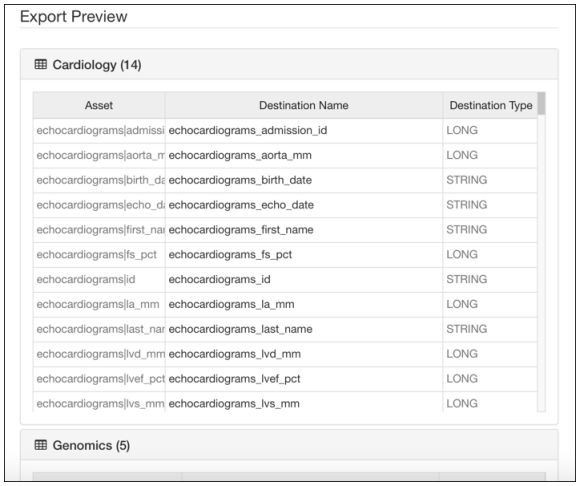
   You select the fields to include in the dataset on the **Dataset Contents** pane. The **Dataset Contents** pane shows a number of tabs; each tab reflects one of the data categories that has been configured for your data.

   Clicking a data category tab lists all the fields extracted from all the data collections within the category. Each data collection is represented by a column. Each data collection and each data element within the collection has a checkbox. Include a data element, or an entire data collection, by checking its checkbox.




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, organized by data category.




5. If you want to, you can change the name of any digital asset in the dataset. In the **Export Preview**, select the **Destination Name** field of the digital asset you want to modify. Double-click to make the field editable and make your changes.
6. 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.

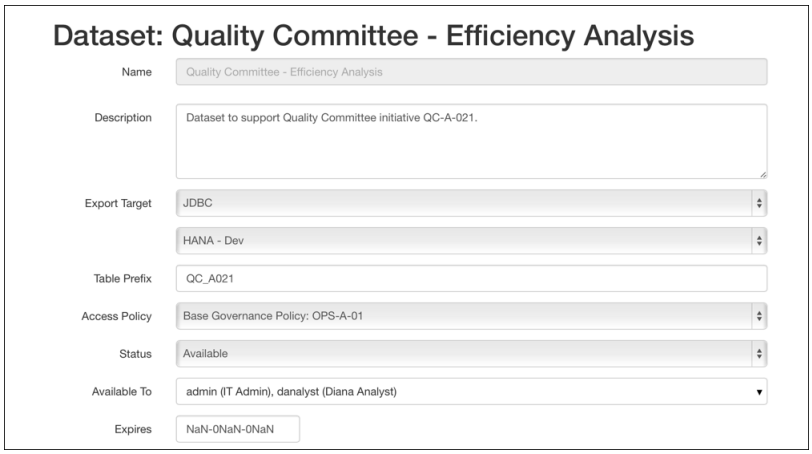
**Note:** Only users with a role of PHEMI Administrator can view dataset information.

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.




   *What do these fields mean?*

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?*

   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. If a dataset has not been made available to you, you will see it in the list but it will be marked not authorized and you will not be able to select it. If you need to see this data, you must ask the PHEMI Administrator add you to the list of users who can access the dataset.

   You will only be able to select datasets that have not expired. If the dataset is expired, you will see it in the list but it will be marked expired and you will not be able to select it. If you need to see this data, you must ask the PHEMI Administrator to reset the dataset expiry date.

If you do have access to a dataset, you may not have access to all the data fields defined for it. The data you are able to see will be constrained by your user authorization and the access policy applied to the dataset.

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**.

# Access Policies

Access policies let you characterize rightful access in terms of user authorization and data visibility.

An access policy is a set of rules that specifies how users can consume data stored in PHEMI Central. The access policy lists what user authorizations are required to interact with data tagged with specified visibility. Access policies can be applied to data collections and datasets.

To create an access policy, you define one or more access rules. Each rule has three parts.

- Subject specifies who may access the data. Access is characterized in terms of user authorizations.

  *Tell me about user authorizations.*
- Action specifies what action authorized users may take to interact with the data.
- Object specifies what kind of data for which the access is being granted. The data is specified in terms of data visibility.

  *Tell me about data visibilities.*

An access policy is matched when the data access request satisfies at least one access rule in the policy. A rule is satisfied if the user making the request has one of the authorizations listed for Subject, and the data being requested one of the visibilities listed for Object. When there is a match, the user may take the specified Action on the data.

The rules in a given access policy implement specific controls over data. Depending on the number and kinds of datasets your organization works with, you may need just one access policy or you may need multiple access policies.

If you have multiple access policies, it is possible for policies to conflict with one another and still represent a consistent governance policy, provided that each access policy is used to control different data collections or datasets. For example, one access policy may allow users with Researcher authority to read CONFIDENTIAL data while another access policy does not. This can be perfectly consistent, as when access policies control different data.

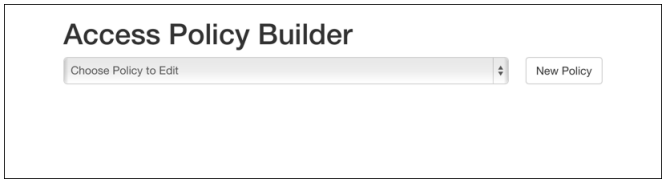
## View Existing Access Policies

See what access policies have been configured on the **Access Policy Builder** page.

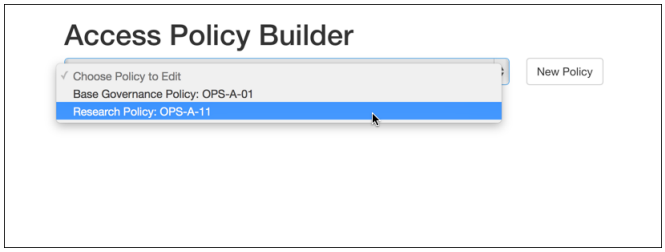
To view defined access policies:

1. Open the **Access Policy Builder** page, by clicking the **Access Policy Builder** icon in the left navigation bar.

   The **Access Policy Builder** page opens.




2. At the right of the **Choose Policy to Edit** field, click the drop-down arrow. The list of configured access policies display.




## Create an Access Policy

Create a new access policy on the **Access Policy Builder** page.

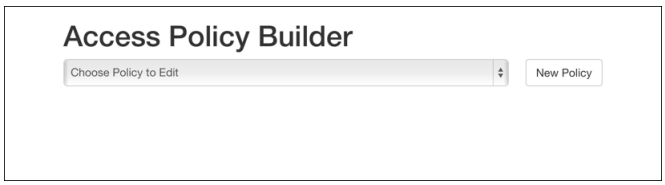
Before you can create an access policy you must configure the following:

- User authorizations *How?*
- Data visibilities *How?*

To create a new access policy, define one or more access rules:

1. Open the **Access Policy Builder** page, by clicking the **Access Policy Builder** icon in the left navigation bar.

   The **Access Policy Builder** page opens.




2. Click the **New Policy** button.

   The form for the new access policy opens, with Rule 1 ready for you to edit.

3. Enter the access rule information.

| Option | Description |
|---|---|
| **Subject** | Mandatory. The user authorization(s) allowed to perform the action on the data. User authorizations are configured for the system by the administrator.<br><br>*Tell me about user authorizations.* |
| **Action** | Mandatory. The action(s) an authorized user may take on the data. Supported actions are as follows:<br><br>• None. The user has no permission to interact with the data.<br>• Read. The user may view the data.<br>• Export. The user may export the data to a destination, such as a local computer for analysis.<br><br>The default value is None. In practice, an action of None would be rarely, if ever, specified. Instead, access rules are generally structured to specify what actions are allowed. Anything not allowed is implicitly disallowed. |
| **Object** | Mandatory. The data visibilities authorized users are allowed to perform the action on.<br><br>*Tell me about data visibilities.* |

4. Add another rule by clicking the **Add Rule** button. Or, save the access policy by clicking the **Save Access Policy** button. The system confirms when the access policy has been successfully saved.

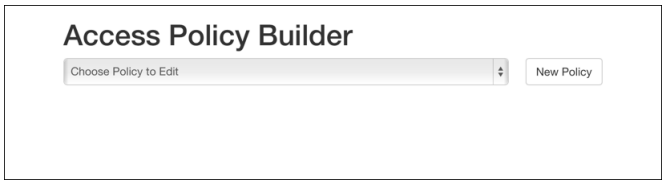
## View Access Policy Information

View the information configured for a given access policy on the **Access Policy Builder** page.

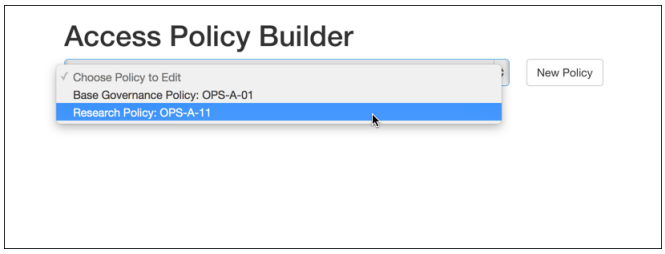
To view information for an access policy:

1. Open the **Access Policy Builder** page, by clicking the **Access Policy Builder** icon in the left navigation bar.

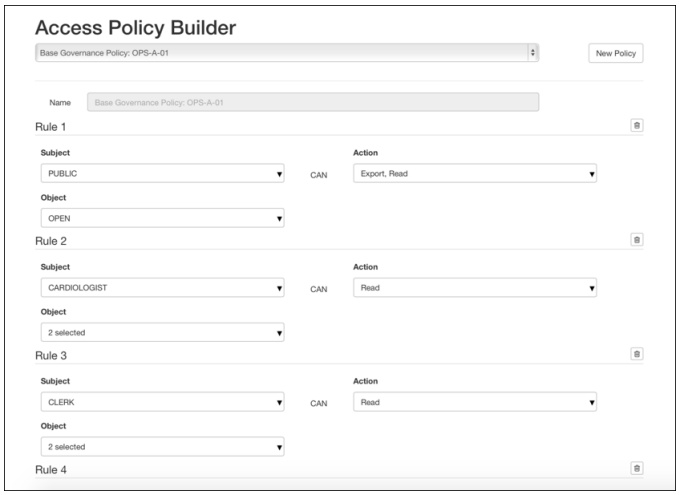
   The **Access Policy Builder** page opens.




2. At the right of the **Choose Policy to Edit** field, click the drop-down arrow to see configured access policies. Select the policy you want to view.




   The screen for the selected access policy opens, showing the information configured for it.




   *What do these fields mean?*

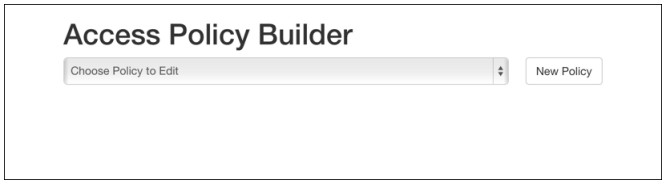
## Modify an Access Policy

Modify an access policy on the **Access Policy Builder** page.

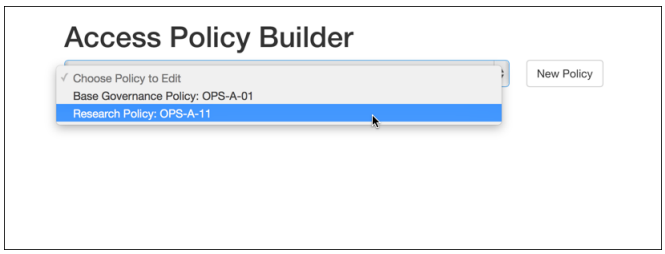
To modify an access policy:

1. Open the **Access Policy Builder** page, by clicking the **Access Policy Builder** icon in the left navigation bar.

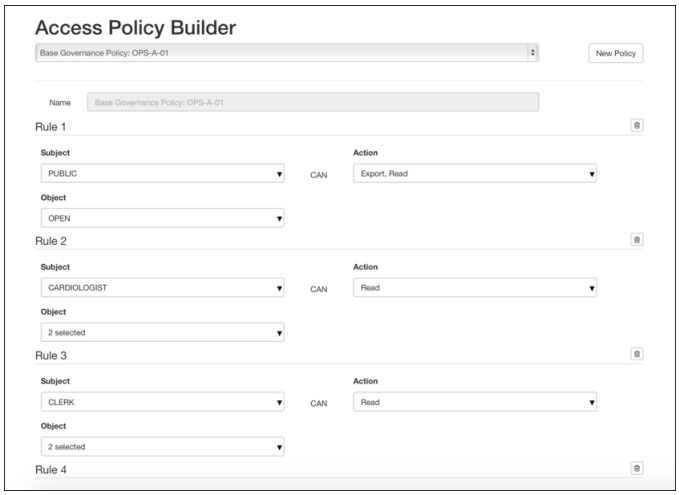
   The **Access Policy Builder** page opens.




2. At the right of the **Choose Policy to Edit** field, click the drop-down arrow to see configured access policies. Select the policy you want to modify.




The screen for the selected access policy opens, showing the information configured for it.




*What do these fields mean?*

3. Do any of the following.
   - Modify an existing rule by editing values for Subject, Action, or Object.

     *What do these fields mean?*
   - Add a new rule by clicking the **Add Rule** button and populating the fields.

     *What do these fields mean?*
   - Delete a rule by clicking the trash can icon to the right of the rule.
4. Click the **Save Access Policy** button to save the changes.

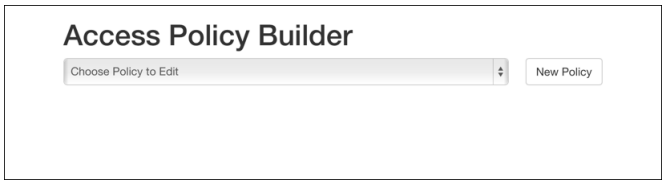
## Delete an Access Policy

Delete an access policy on the **Access Policy Builder** page.

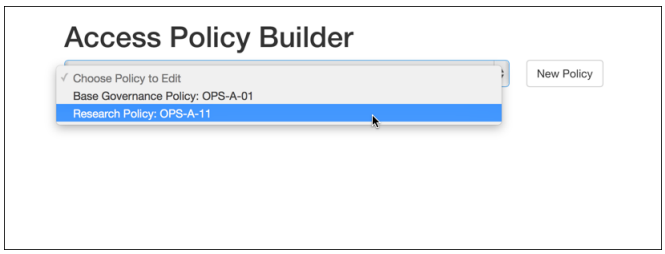
To delete an access policy:

1. Open the **Access Policy Builder** page, by clicking the **Access Policy Builder** icon in the left navigation bar.

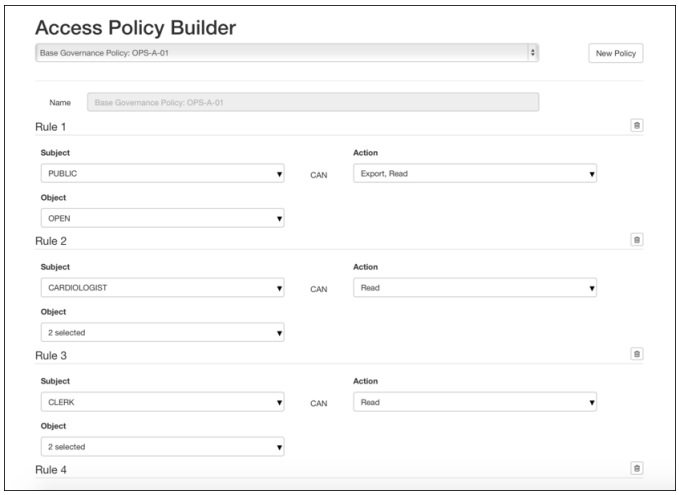
   The **Access Policy Builder** page opens.




2. At the right of the **Choose Policy to Edit** field, click the drop-down arrow to see configured access policies. Select the policy you want to modify.




The screen for the selected access policy opens, showing the information configured for it.




3. Delete a single rule by clicking the trash can icon next to the rule. Delete an entire access policy by clicking the **Delete Access Policy** button. If you try to delete an access policy, the system asks you to confirm deletion; click **Delete**.

# System Metrics

Monitor system status and health on the **System Metrics** page.

The **System Metrics** page of the Management and Governance Console reports information about users and system usage, as well as tracking data collection statistics, system performance, and the status of systems tasks.

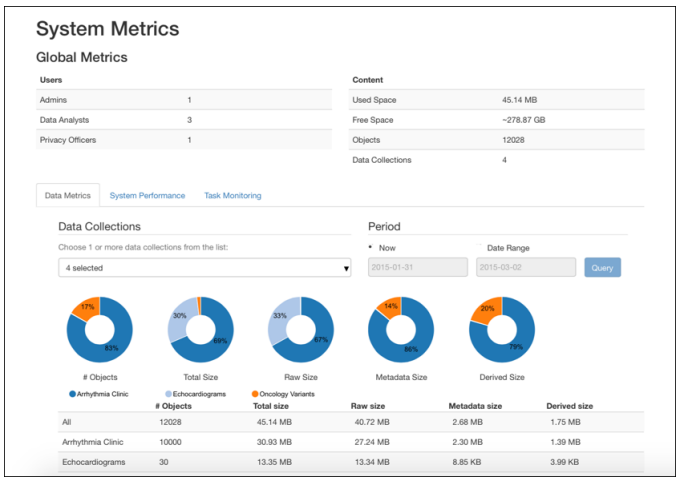
## View Global Metrics

View summary information about your users and the content residing on PHEMI Central in the **Global Metrics** pane of the **System Metrics** page.

To view global metrics:

Open the **System Metrics** page, by clicking the **System Metrics** icon in the left navigation bar.

The **System Metrics** page opens showing the **Global Metrics** pane and the **Data Metrics** pane.




In the **Users** column, you can see how many users of each user role have been created. In the **Content** column, you can see how much used and free space there is on the system, as well as the number of objects the system is currently storing and the number of data collections that have been configured.

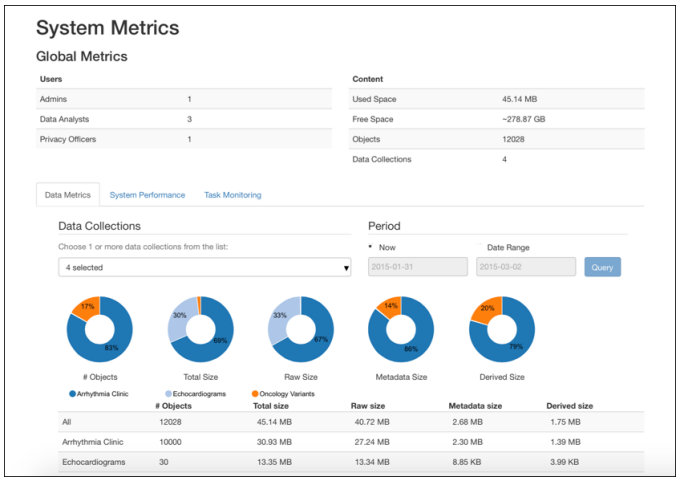
## View Data Metrics

View statistics about data collections on the **Data Collection Metrics** tab of the **System Metrics** page.

To view data collection metrics:

1. Open the **System Metrics** page, by clicking the **System Metrics** icon in the left navigation bar.

   The **System Metrics** page opens showing the **Global Metrics** pane and the **Data Metrics** pane.

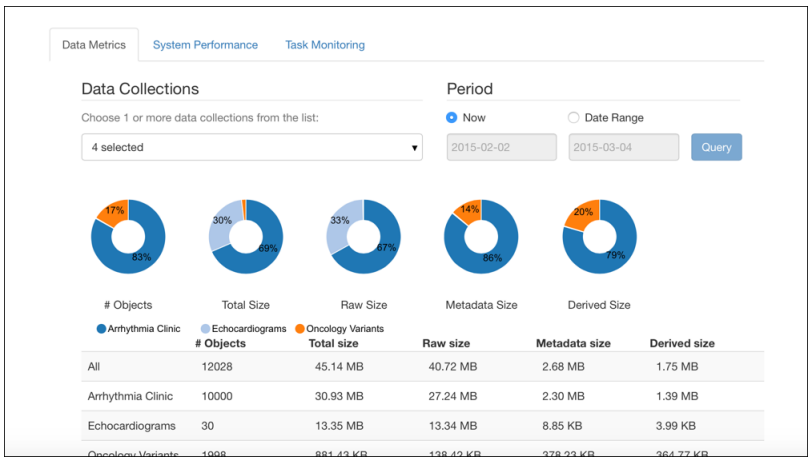



2. Select the data collections to include in the query.

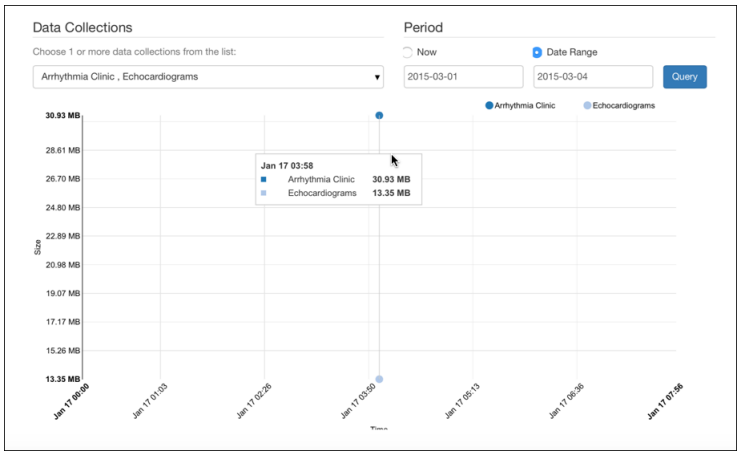
   By default, the **Data Metrics** pane shows a current snapshot of data collection metrics and includes the data from all data collections. To view information for a different set of data collections, click the drop-down arrow in the **Data Collections** field and check each data collection you want included in the metrics. If you want to include all data collections, check **All** to select them all in one operation.
3. Choose the query period. Select **Now** or select **Date Range** and specify the dates in the format *yyyy-mm-dd*.
4. Click **Query**.

   The results of your query are displayed below the selection parameters.

   If the query period is **Now**, the system shows the metrics as a set of ring graphs.




When a date range is specified, the system draws an X-Y graph showing ingestion events for the specified data collections. You can hover your cursor over an ingestion event to view specific data metrics for each data collection.



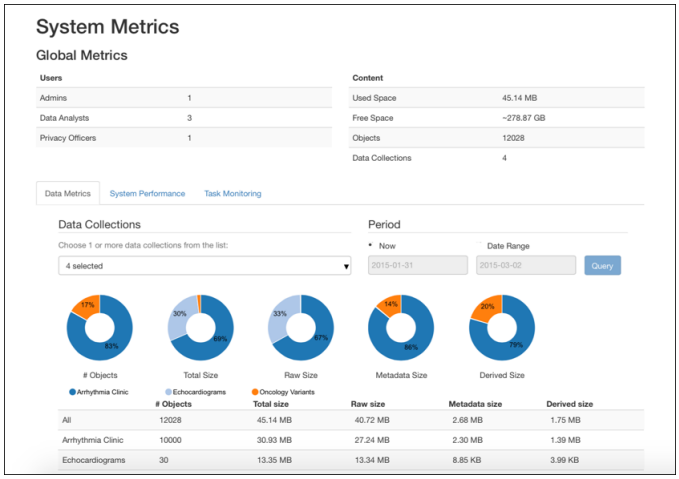

## View System Performance

View performance statistics for data ingest and dataset execution on the **System Performance** tab of the **System Metrics** page.

To view data ingest and dataset execution metrics:

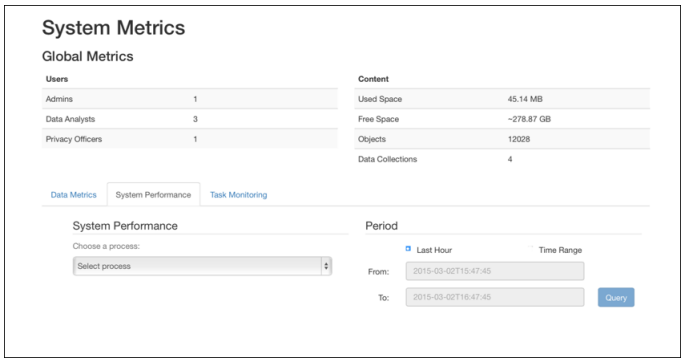
1. Open the **System Metrics** page, by clicking the **System Metrics** icon in the left navigation bar.

   The **System Metrics** page opens showing the **Global Metrics** pane and the **Data Metrics** pane.




2. Click the **System Performance** tab.

   The **System Performance** pane opens.




3. In the **Select process** field, choose **Data Ingest** to view performance information for data ingest, or **Data Execution** to view performance information for dataset execution.
4. Choose the query period: select **Last Hour** to see statistics for the previous hour, or select **Time Range** to specify a range of times. The format for a time is *yyyy*-*mm*-*dd*T*hh*:*mm*:*ss*, where hours are specified according to a 24-hour clock.
5. Click **Query**.

   The results of your query are displayed below the selection parameters.

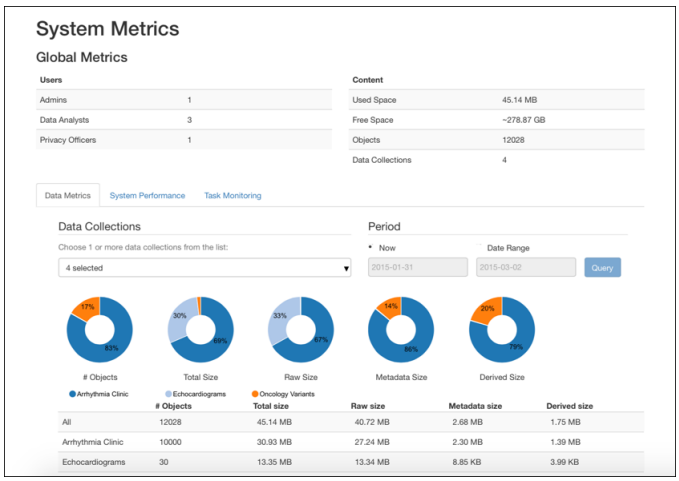
## Monitor System Tasks

Monitor PHEMI Central system tasks on the **Task Monitoring** tab of the **System Metrics** page.

To monitor system tasks:

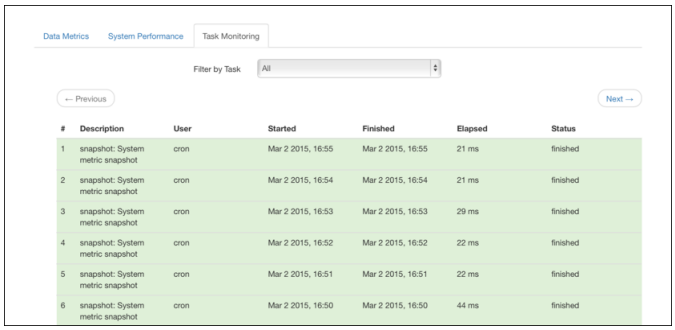
1. Open the **System Metrics** page, by clicking the **System Metrics** icon in the left navigation bar.

   The **System Metrics** page opens showing the **Global Metrics** pane and the **Data Metrics** pane.




2. Click the **Task Monitoring** tab.

   The **Task Monitoring** pane opens, showing system tasks with most recently executed tasks shown first.




3. By default, the **Task Monitoring** pane lists tasks of all task types. You can filter on a single task type by clicking the drop-down arrow in the **Filter by Task** field and checking the task type you want to see. If you want to see all task types, you can check **All**.

| Option | Description |
|---|---|
| **chained** | The system has executed a chained ingest and DPF execution task set, ingesting raw data and executing a DPF as the data was being ingested. This occurs, for example, when Variant Call Format (VCF) files are ingested. |
| **cleanup** | The system has checked retention rules for a data collection and deleted expired data from the data store. |
| **derive** | The system has executed a DPF on data. |
| **download** | The system has exported a dataset. |
| **execute** | The system has executed a dataset. |
| **ingest** | The system has ingested data from the indicated data collection. |
| **snapshot** | The system has taken a snapshot of information for system metrics. |

When you make your selection, the results display from most recent to less recent below the selection parameters.

# Users

User management allows you to say who can use the system to do what, and who can access what data.

PHEMI Central user management is organized around user roles and user authorizations. Your user role defines what parts of the Management and Governance Console you can use, and user authorizations are combined with data visibilities in access policies to specify how users are allowed to interact with data.

Most user information is set and modified on the **Manage Users** page. The exception is user authorizations, which are defined as part of system configuration.

*Tell me about user authorizations*.

## User Roles

User roles define what you are allowed to do within the PHEMI Central Management and Governance Console. Users can be assigned any or all of three roles.

| Role | Purpose | PHEMI Central Management Console Access |
|---|---|---|
| PHEMI Administrator | Configures PHEMI Central and configures access to data. | • System configuration: password policy, dataset destinations, data retention behavior, data categories, data visibilities, and user authorizations<br>• Manage access policies<br>• Manage data collections, deploy DPFs<br>• Construct datasets<br>• Monitor system metrics<br>• Manage users<br>• Monitor audit logs |
| Privacy Officer | Responsible for governance policies that define the organization's approach for safeguarding data and assigning privileges to users. | The privacy officer has no functional ability within the PHEMI Central Management and Governance Console. The privacy officer defines the governance policies that drive system configuration and "owns" this aspect of a data collection. |
| Data Analyst | Submits data for ingestion and consumes data. | • Ingest data<br>• Execute and export datasets |

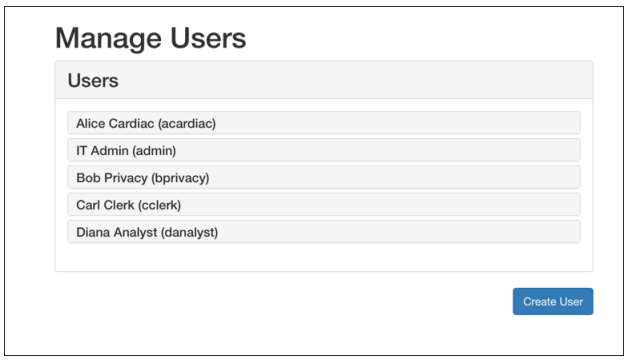
## View System Users

View system users on the **Manage Users** page.

To view system users:

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

The **Manage Users** page lists all the users who have been defined on the system.



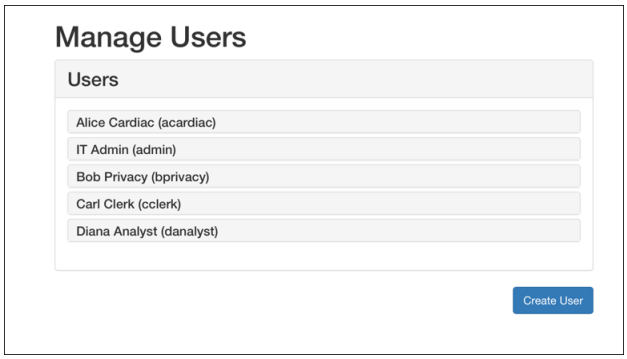

## Create a New User

Create a new user on the **Manage Users** page.

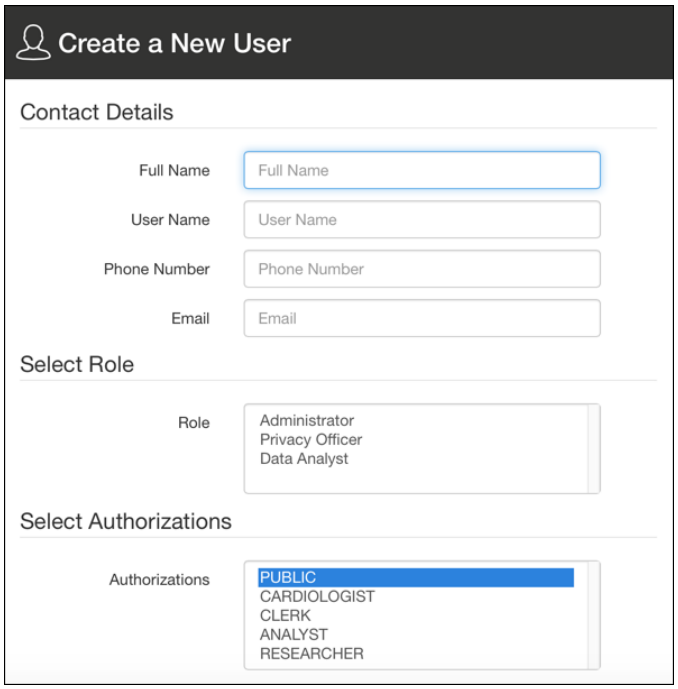
Before you can create users, you must configure user authorizations. *How?*

To create a new user:

1. Open the **Manage Users** page, by clicking the **Users** icon in the left navigation bar.
   The **Manage Users** page lists all users defined in the system so far.




2. Click the **Create User** button.
   The **Create a New User** dialog opens.



3. Enter the user information.

| Option | Description |
|---|---|
| **Full Name** | Mandatory. The user's full name. |
| **User name** | Mandatory. The user ID for this user.<br>The **User Name** field autopopulates with the first initial and last name entered for as the user's full name. For example, if the user's full name is Jane Smith, the **User Name** field autopopulates with jsmith. The user ID can be edited after it autopopulates. IDs can be up to 16 characters long. Alphabetic and numeric characters are permitted, as well as underscore ("_") |
| **Phone Number** | Optional. The user's phone number. You must configure this field if you want the system to send alerts to the user's phone. |
| **Email** | Mandatory. The user's email address. If you configure the system to send email alerts to the user, this is the email address that will be used. |
| **Role** | Mandatory. The user's system role. Together with the user authorization configured and the visibility set for data, the user role determines what access the user will have to data. A user can be assigned more than one role. |
| **Authorizations** | Optional. The types of data the user is authorized to access.<br>The PUBLIC authorization is automatically applied to users by default. You can delete this authorization. Use either the **Shift** key or the **Ctrl** key to make multiple selections. |
| **Password** | Mandatory. Note that the password must comply with your organization's password policy. |

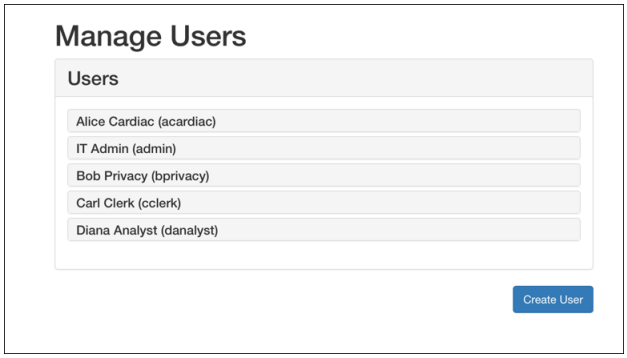
4. Save the user information by clicking the **Save User** button. The system confirms when the user has been successfully saved. Click **Close** to close the screen.

## View User Information

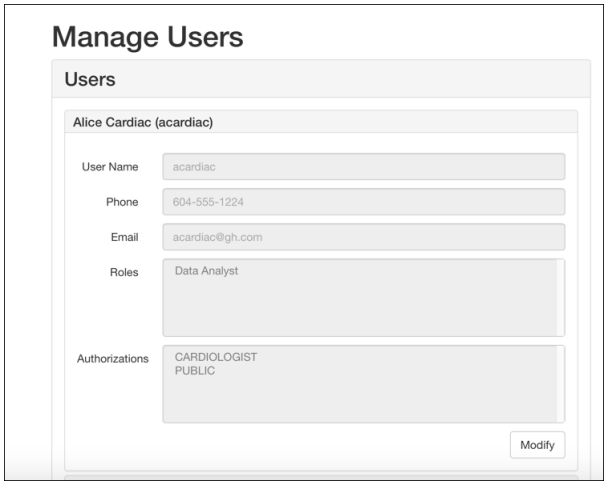
View user information from the **Manage Users** page.

To view user information:

1. Open the **Manage Users** screen, by clicking the **Users** icon in the left navigation bar.




2. Click the user's name to expand the user record.




*What do these fields mean?*

3. Click the user's name a second time to collapse the user record again.

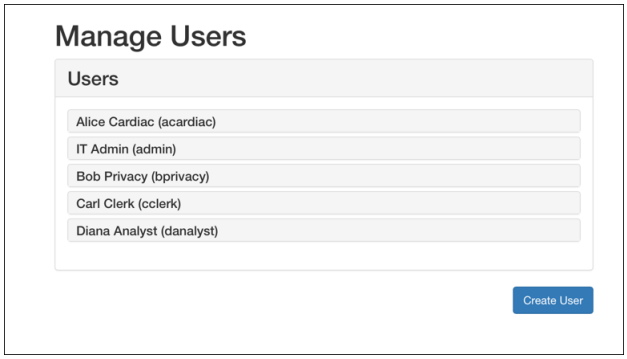
## Modify User Information

Modify user information on the **Manage Users** page.

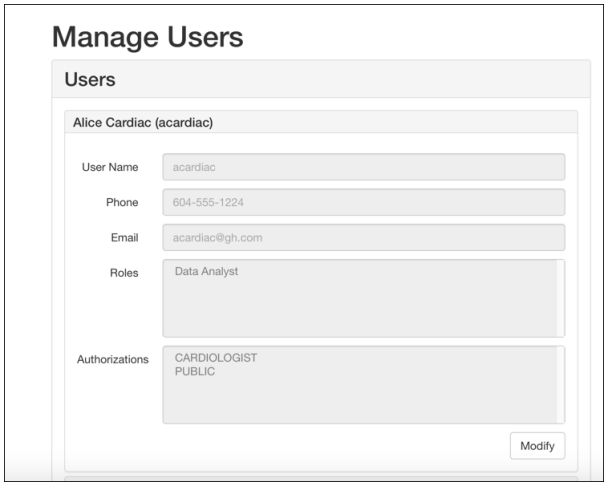
You must know a user's password in order to change their user information.

To modify user information:

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




2. Click the user's name to expand the user record.

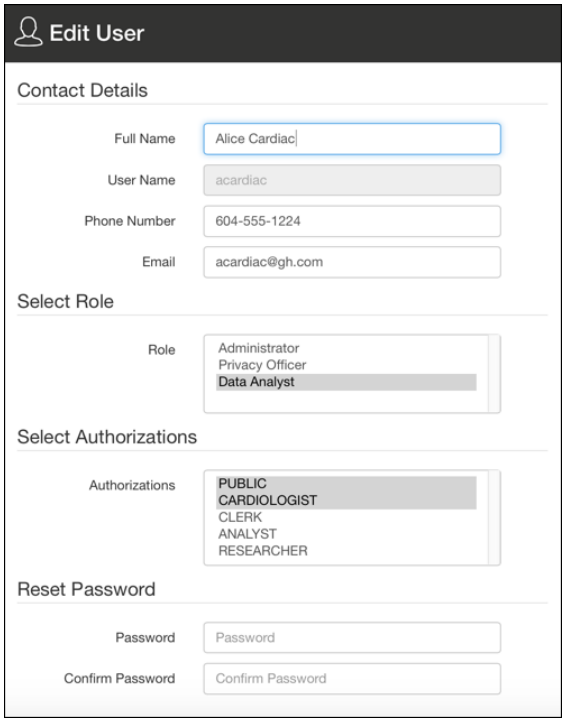


*What do these fields mean?*

3. Click the **Modify** button.

   The **Edit User** dialog opens.

   *What do these fields mean?*



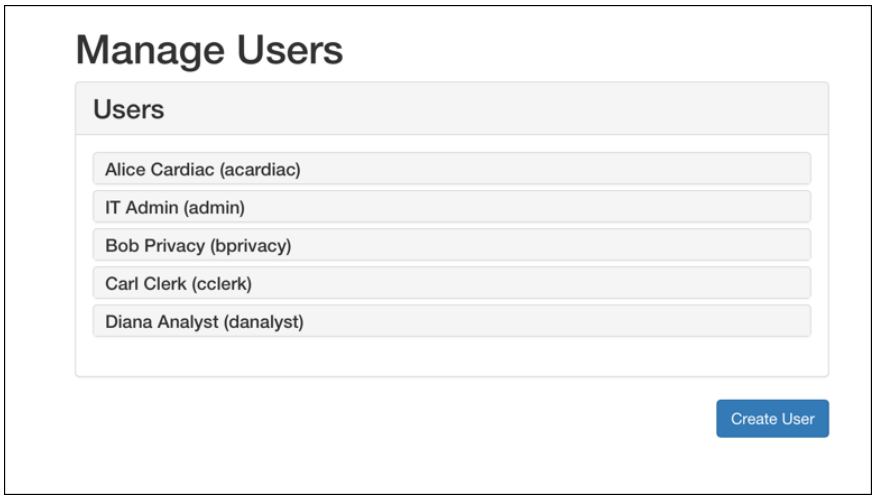

4. Modify the user information as necessary. You must confirm the user's password by typing it in both the **Password** and the **Confirm Password** fields.
5. Save the changes by clicking the **Save User** button. The system confirms when the user has been successfully saved. Click **Close** to close the screen.

## Delete a User

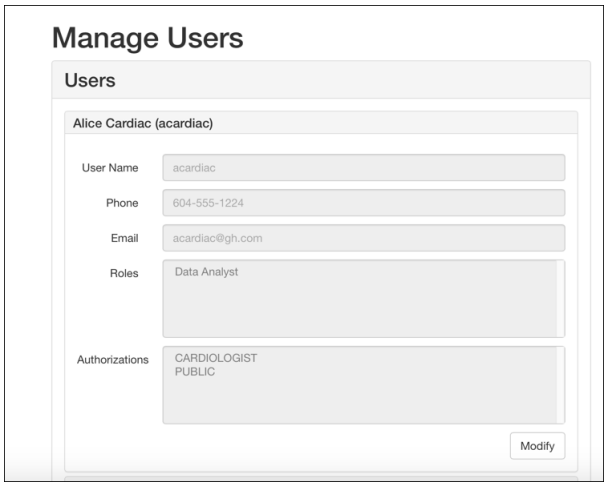
Delete a user from the **Manage Users** page.

To delete a user:

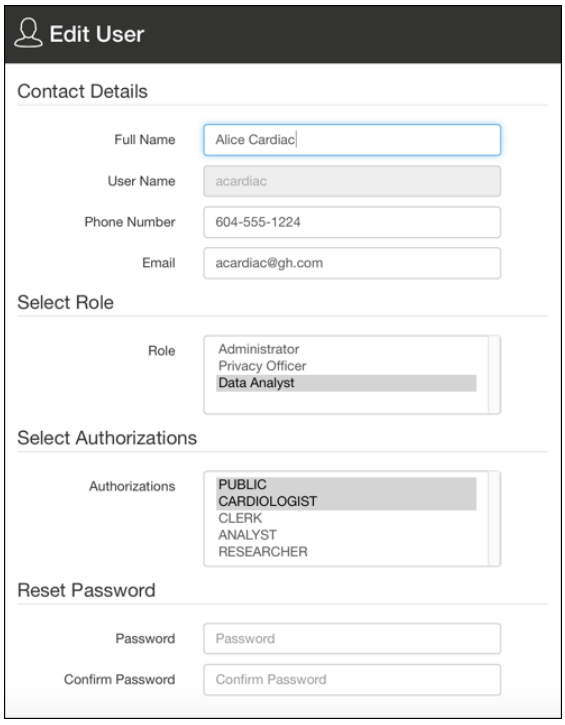
1. Open the **Manage Users** page, by clicking the **Users** icon in the left navigation bar.



2. Click the user's name to expand the user record.




3. Click the **Modify** button.
   The **Edit User**dialog opens.




4. Click the **Delete User** button. The system asks you to confirm. Click the **Confirm Delete** button.
   The system confirms when the user has been successfully deleted. Click **Close** to close the screen.

## The Object Browser

View or delete objects stored in PHEMI Central using the **Object Browser**.

The Object Browser lets you view individual objects that have been stored in PHEMI Central. You can see the metadata that has been applied to the object, the data elements (if any) that have been derived by running a DPF on the data, and the binary data that makes up the object, shown as hexadecimal with an ASCII interpretation.

The Object Browser is the only access to data that a PHEMI Administrator has in PHEMI Central, and only a PHEMI Administrator can access the Object Browser. As with all other kinds of queries to the system, access to data is constrained by the governance placed around the data.

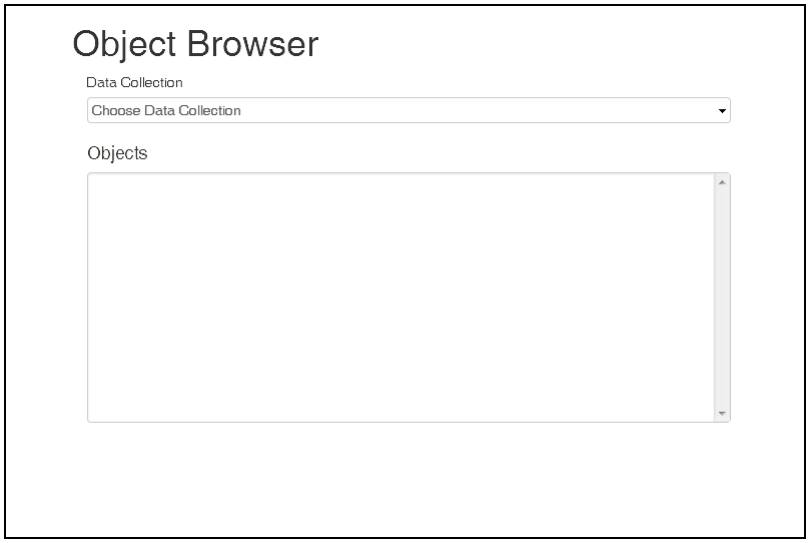
## View an Object

View individual data objects stored in PHEMI Central using the **Object Browser** page.

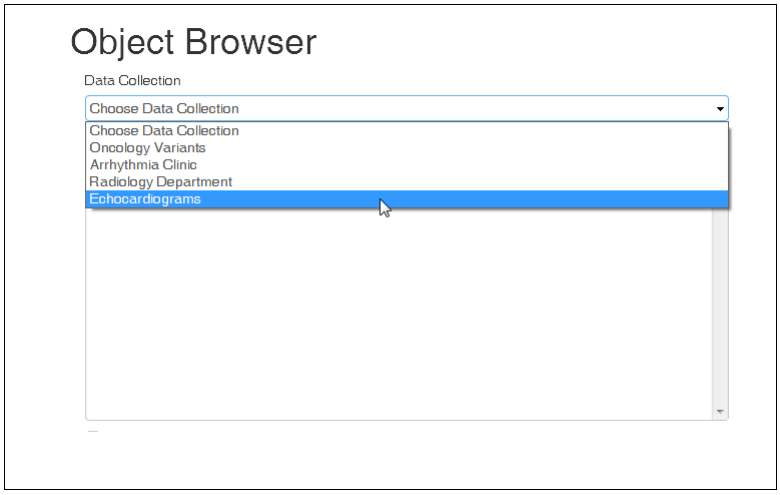
To view an object:

1. Open the **Object Browser** page, by clicking the **Object Browser** icon in the left navigation bar.

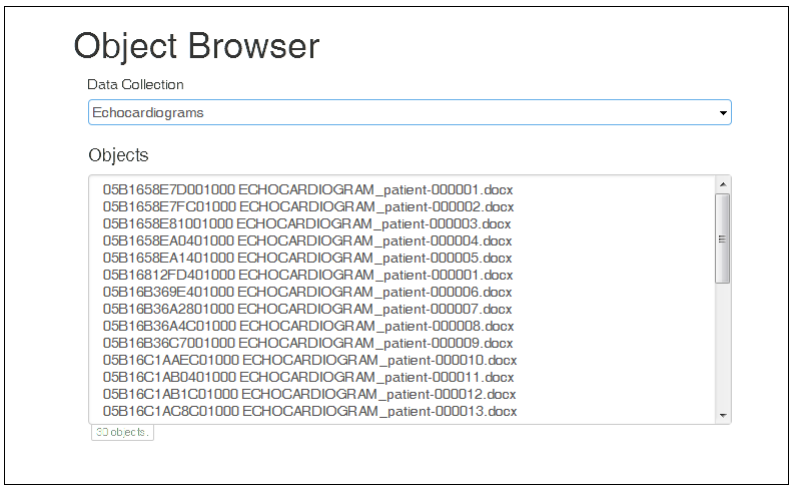
   The **Object Browser** page opens.




2. Click the drop-down arrow on the **Choose Data Collection** field and select the data collection containing the object.

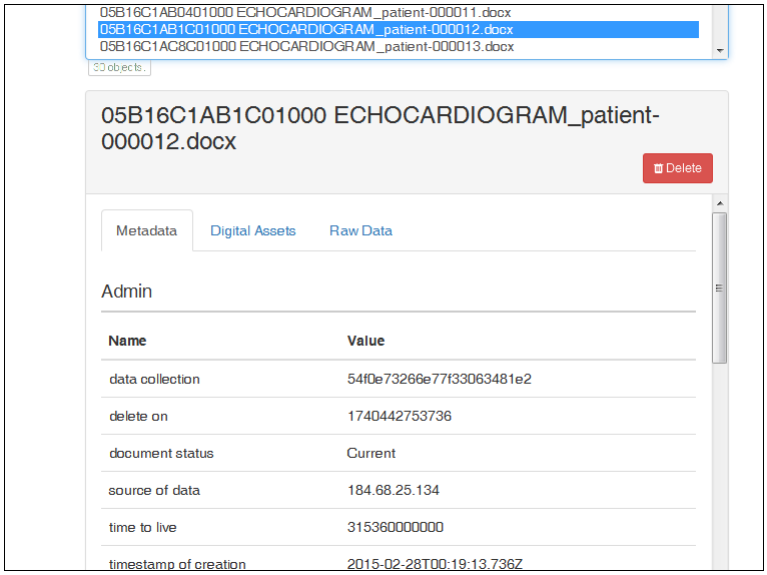



   The available objects are listed in the **Objects** text box. Each object has the original file name it had on ingest, prepended with the system key that PHEMI Central has assigned.




3. Scroll through the list as necessary and select the object you want to view.

   Once selected, the object's information is retrieved and displays with the **Metadata** pane showing.




- View the object's metadata on the **Metadata** pane.
- View any data elements derived by executing a DPF by clicking the **Digital Assets** tab.
- View a hexadecimal representation and ASCII interpretation of a raw binary object by clicking the **Raw Data** tab.

## Delete an Object

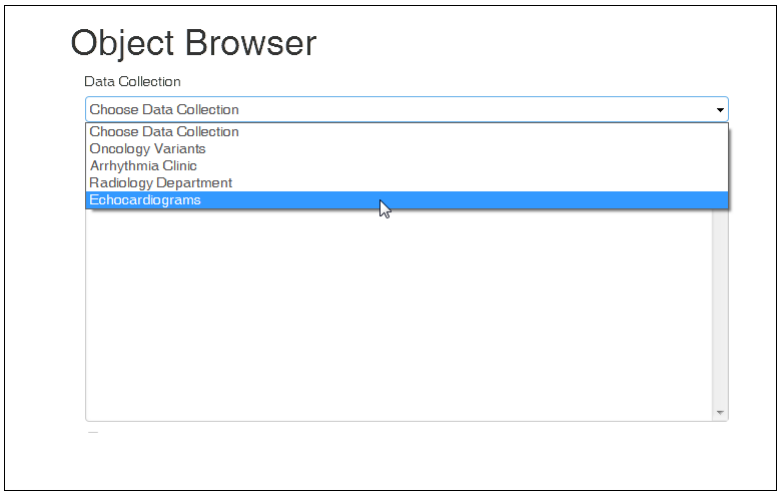
Delete an individual data object from the system using the **Object Browser** page.

To delete an object:

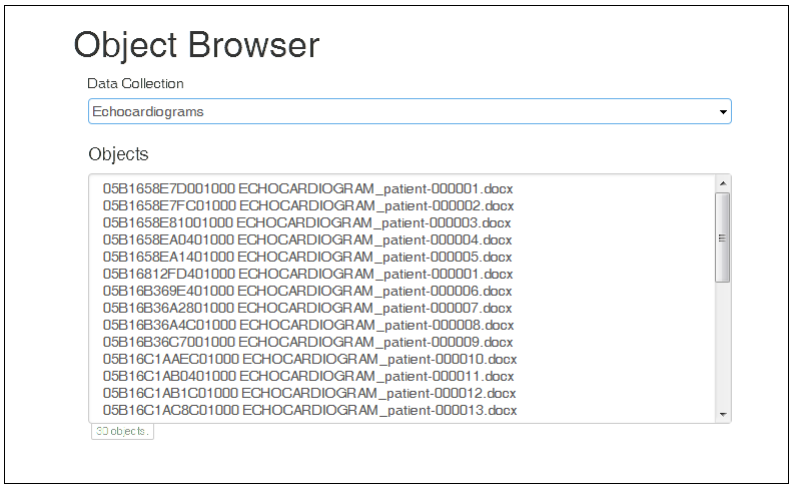
1. Open the **Object Browser** page, by clicking the **Object Browser** icon in the left navigation bar.

   The **Object Browser** page opens.

2. Click the drop-down arrow on the **choose Data Collection** field and select the data collection containing the object.

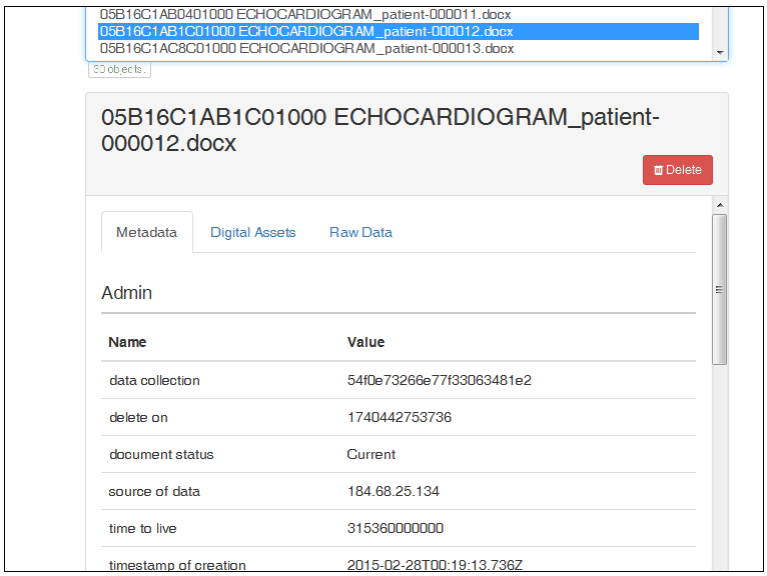



The available objects are listed in the **Objects** text box. Each object has the original file name it had on ingest, prepended with the system key that PHEMI Central has assigned.




3. Scroll through the list as necessary and select the object you want to view.

   Once selected, the object's information is retrieved and displays with the **Metadata** tab showing, and the **Delete** button appears.




4. To delete the original object, click the **Delete** button. The **Delete this object?** dialog opens and the system asks you to confirm permanent deletion of the object. If there are derived data items and you want to delete them as well, check the **Do you also want to delete the associated Derived Data?** checkbox on the **Delete this object?** dialog. Click **Delete**.

# Audit Log

See how PHEMI Central has been accessed and used by viewing the **Audit Log**.

PHEMI Central maintains complete a complete audit log of system and user operations. Log entries include all create, modify, and delete operations performed, as well as all data queries made to the system, regardless of interface. The Audit Log file is completely tamperproof for all users.

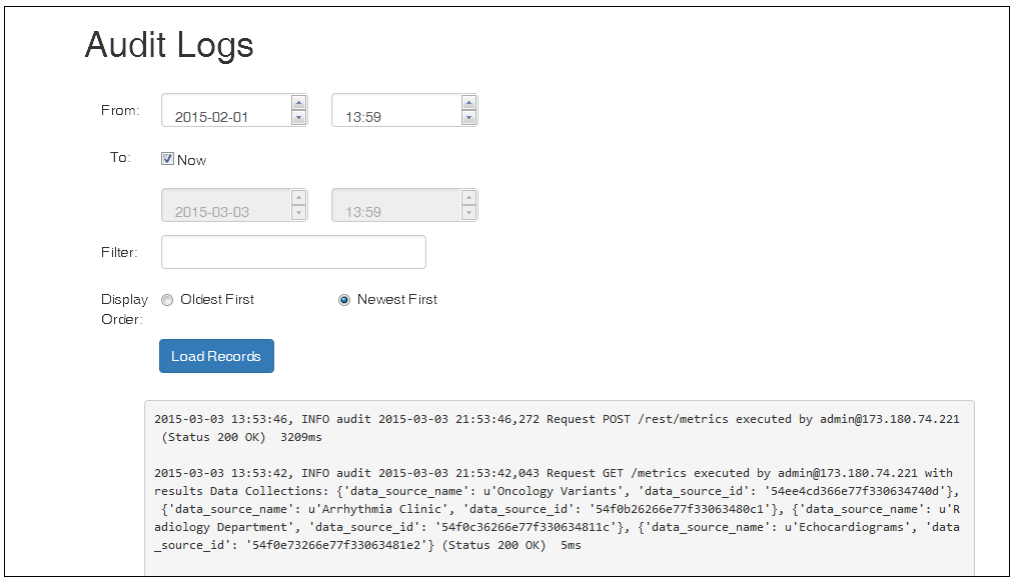
## View the Audit Log

View the Audit Log on the **Audit Logs** page.

To view the Audit Log:

1. Open the **Audit Log** page, by clicking the **Audit Log** icon in the left navigation bar.

   The **Audit Log** page opens.




2. Set your query options.

| Option | Description |
|---|---|
| **From** | Optional. The start time of log entries. The format for the date is *yyyy-mm-dd*. The format for the time is *hh:mm*, in 24-hour format. The default is the current date and time. Leaving the default From and To values shows the complete log file. By default, the From time is one month previous, relative to the current time. |
| **To** | Optional. The stop time of log entries. Choose the current time by checking the Now checkbox, or uncheck the Now checkbox and specify a time using *yyyy-mm-dd* for the date and *hh:mm*, in 24-hour format, for the time. The default is Now. |
| **Display Order** | Optional. The order in which log messages are displayed. Select Oldest First to display from oldest to most recent. Select Newest First to display from most recent to oldest. The default is Newest First. |
| **Filter** | Optional. A space-separated list of keywords, which may include user IDs. The system restricts search results to entries containing at least one of the keywords. |

3. Click **Load Records** to retrieve the specified log messages.

# System Configuration

The **System Configuration** page includes functions the PHEMI Administrator uses to set up and maintain PHEMI Central.

*What is the workflow for initial configuration?*

## The Password Policy

A good password policy is one way your organization can secure your information systems.

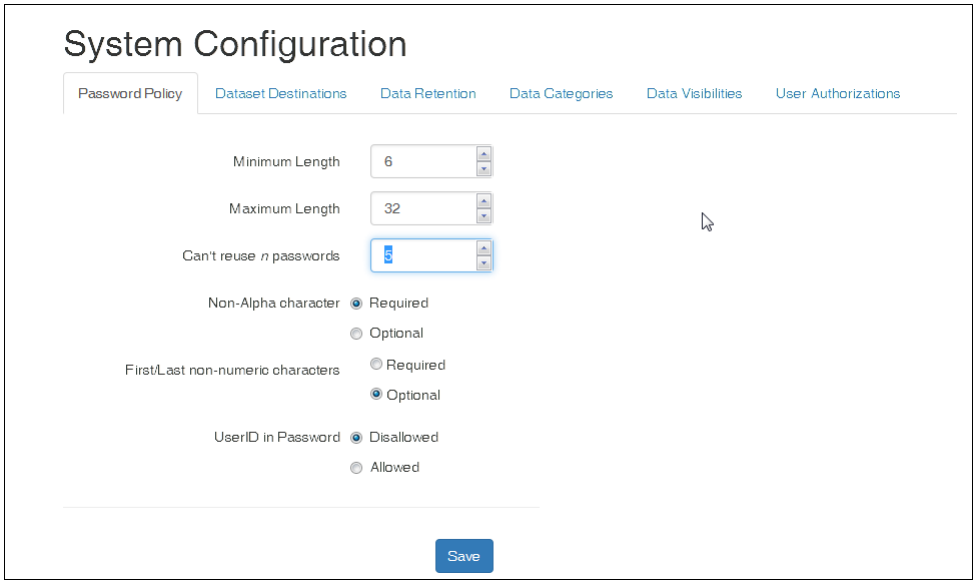
A password policy generally enforces the strength of a password, by requiring passwords to be of a certain length and by stipulating that a password must include a certain mix of letters, numbers, and/or special characters. The password policy may also control how quickly a given password can be reused.

The password policy is generally decided on by the privacy officer, or someone in a similar role. The privacy policy is implemented in PHEMI Central by the PHEMI Administrator, as part of system configuration.

### Configure the Password Policy

Set the password policy on the **Password Policy** screen of the **System Configuration** page.

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.




2. Set the password policy information.

| Option | Description |
|---|---|
| **Minimum Length** | Mandatory. The minimum length for the password. The range is 6 to 15. The default is 6. |
| **Maximum Length** | Mandatory. The maximum length for the password. The maximum starts at the value set for **Minimum Length** (that is, maximum and minimum value can be the same) and ranges to 32 characters. The default is 32. |
| **Can't reuse *n* passwords** | Optional. Specifies the number of times in a row a password can be used. For example, if this value is set to 1, the user cannot reuse the same password twice; at least one more password must intervene before reusing the original password. The range is 0 to 12. The default is 0, which means that the same password can be repeated indefinitely. |
| **Non-Alpha Character** | Optional. Specifies whether the password must include with a non-alphabetical character. Non-alphabetical characters are numbers or special characters. Spaces are not supported. Supported values are as follows:<br>• Required: Passwords must include at least one non-alphabetical character.<br>• Optional: Passwords can consist of only alphabetic characters.<br>The default is Optional. |
| **First/Last non-numeric characters** | Optional. Specifies whether the password must begin and end with a non-numeric character. Non-numeric characters include alphabetical characters and special characters. Spaces are not supported. Supported values are as follows:<br>• Required: Passwords must begin and end with a non-numeric character.<br>• Optional: Passwords may begin and end with alphabetic or special characters.<br>The default is Optional. |

| Option | Description |
|---|---|
| **User ID in Password** | Optional. Specifies whether the string representing the user's ID may appear in the password. Supported values are as follows:<br>• Disallowed: The user ID may not appear in the password.<br>• Allowed: The user ID may appear in the password.<br>The default is Allowed. |

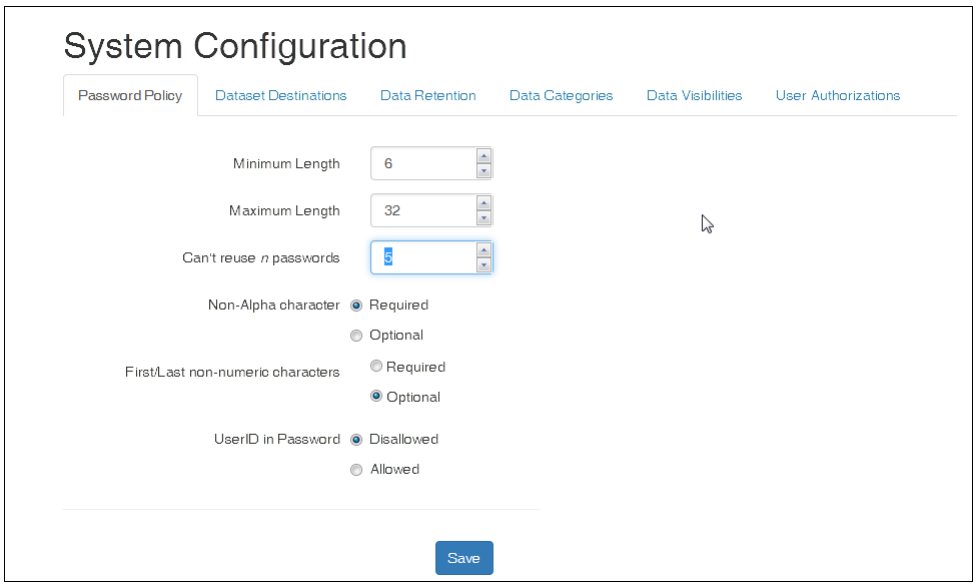
3. Save the password policy by clicking the **Save** button.

### View the Password Policy

View the configured password policy on the **Password Policy** screen of the **System Configuration** page.

Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

The **System Configuration** page opens on the **Password Policy** screen.




*What do these fields mean?*

## Dataset Destinations

Datasets can be exported to consuming applications and tools.

*Tell me about datasets.*

Datasets can be consumed by querying them through the PHEMI RESTful API or by configuring a dataset export target as a "destination" in the Management and Governance Console.

Dataset destinations are configured by the PHEMI Administrator. The destination is characterized in terms of the connection type, the network parameters, the destination database and schema names, and the user credentials for logging on to the destination.

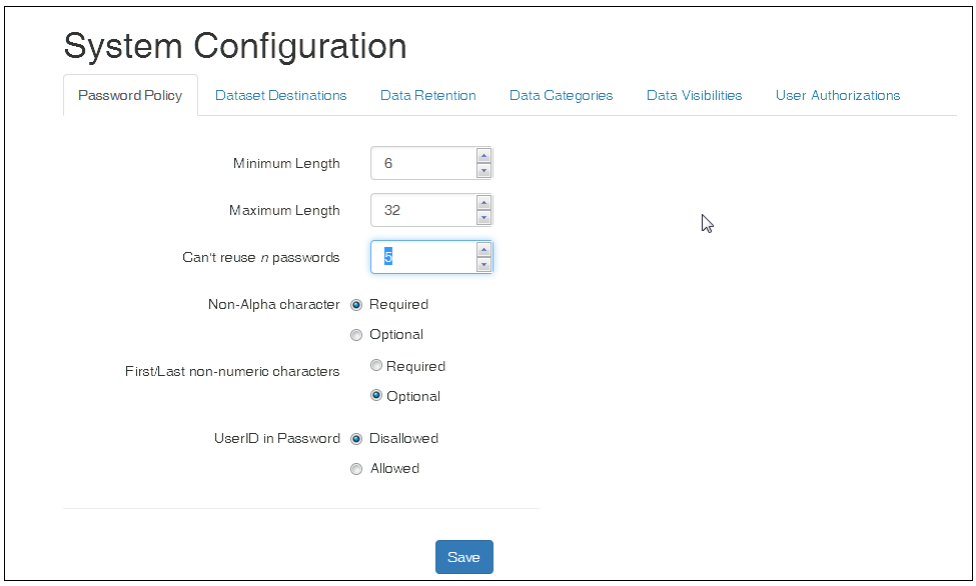
### View Dataset Destinations

View dataset destinations on the **Dataset Destinations** screen of the **System Configuration** page.

To view defined dataset destinations:

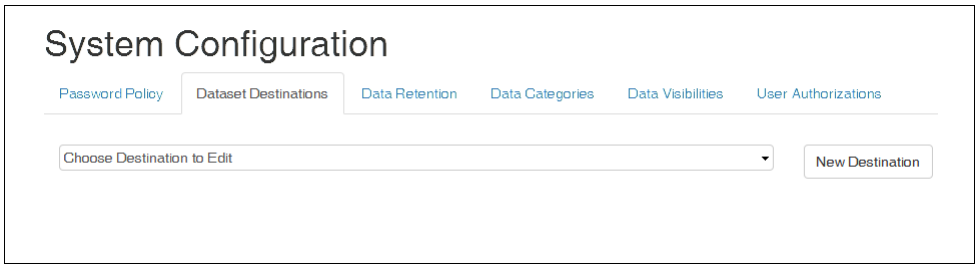
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

   The **System Configuration** page opens on the **Password Policy** screen.

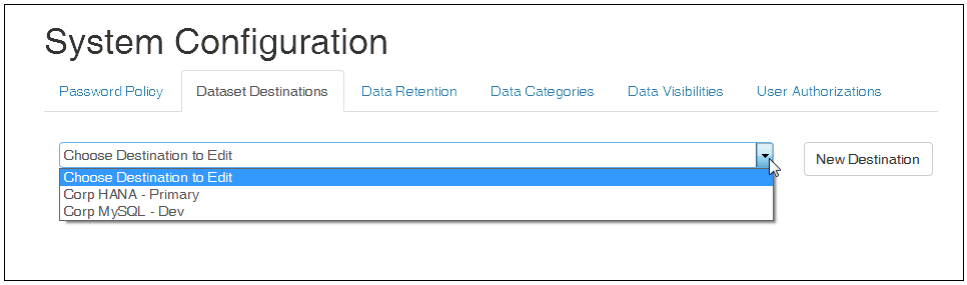



2. Click the **Dataset Destinations** tab.

   The **Dataset Destinations** screen opens.




3. At the right of the **Choose Destination to Edit** field, click the drop-down arrow to see configured dataset destinations.



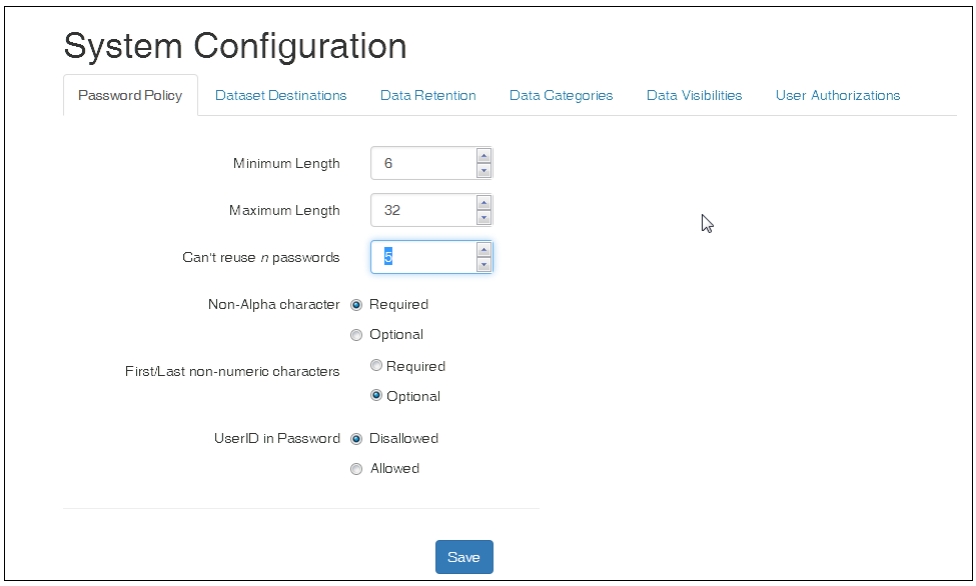

### Create a Dataset Destination

Define a new dataset destination on the **Dataset Destinations** screen of the **System Configuration** page.

To create a new dataset destination:

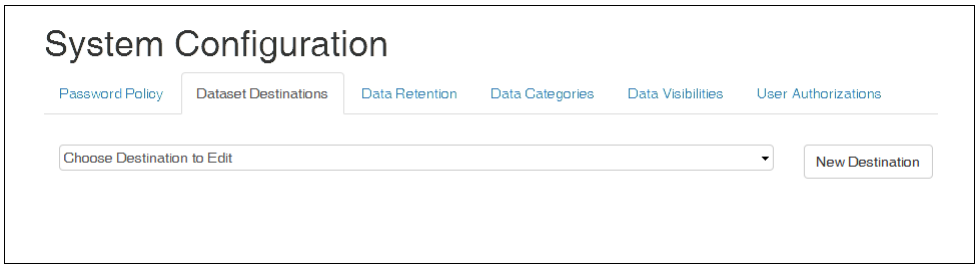
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

   The **System Configuration** page opens on the **Password Policy** screen.



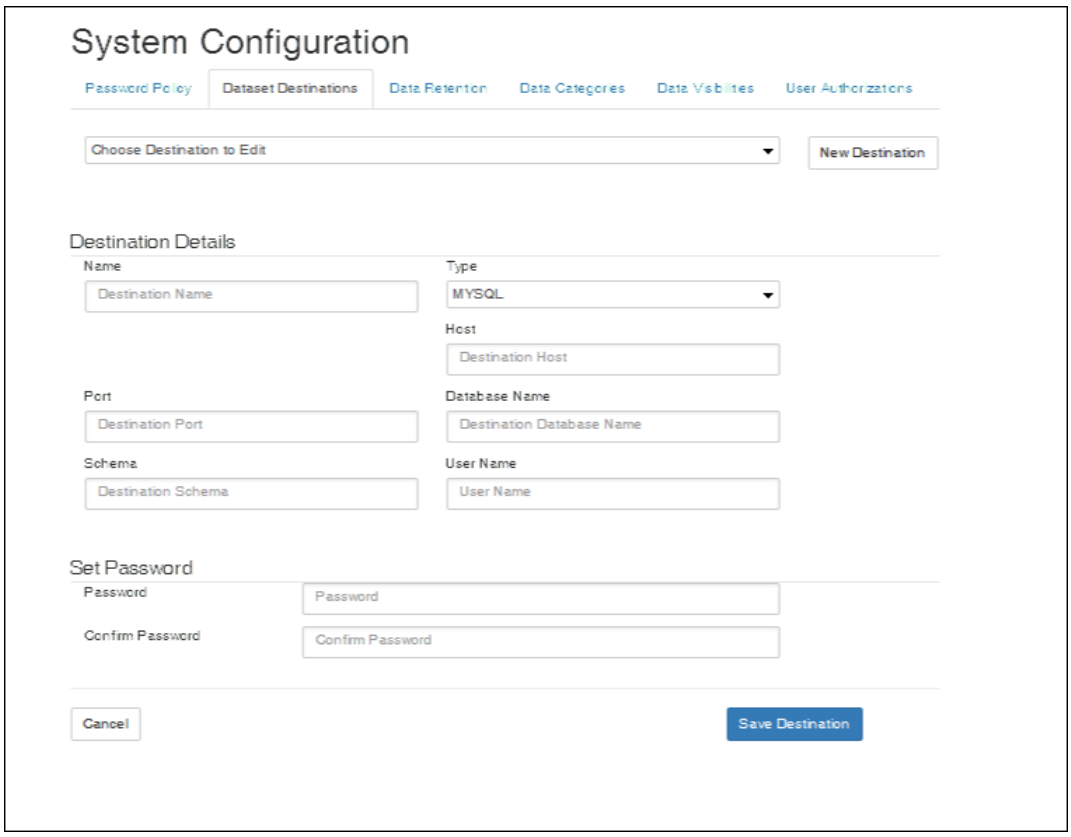

2. Click the **Dataset Destinations** tab.

   The **Dataset Destinations** screen opens.




3. Click the **New Destination** button to the right of the **Choose Destination to Edit** field.

   The **Destination Details** screen opens.




4. Enter the destination details.

| Option | Description |
|---|---|
| **Name** | Mandatory. Provide a name for the destination. |
| **Type** | Mandatory. The destination type. Supported values are as follows:<br>• MySQL. The destination is a MySQL database.<br>• HANA. The destination is a SAP HANA database.<br>Both MySQL and HANA destination types use a Jave Database Connectivity (JDBC) connection. |
| **Host** | Mandatory. The IP address or hostname of the destination system. |
| **Port** | Mandatory. The destination port. |
| **Database Name** | Mandatory. The name of the destination database. |
| **Schema** | Mandatory. The name of the schema being used in the destination database. |
| **User Name** | Mandatory. The user name for logging on to the destination system. |
| **Password** | Mandatory. The password for logging on to the destination system. |
| **Confirm Password** | Mandatory. Re-enter the password to ensure it is correct. |

5. Click the **Save Destination** button to save the new destination.

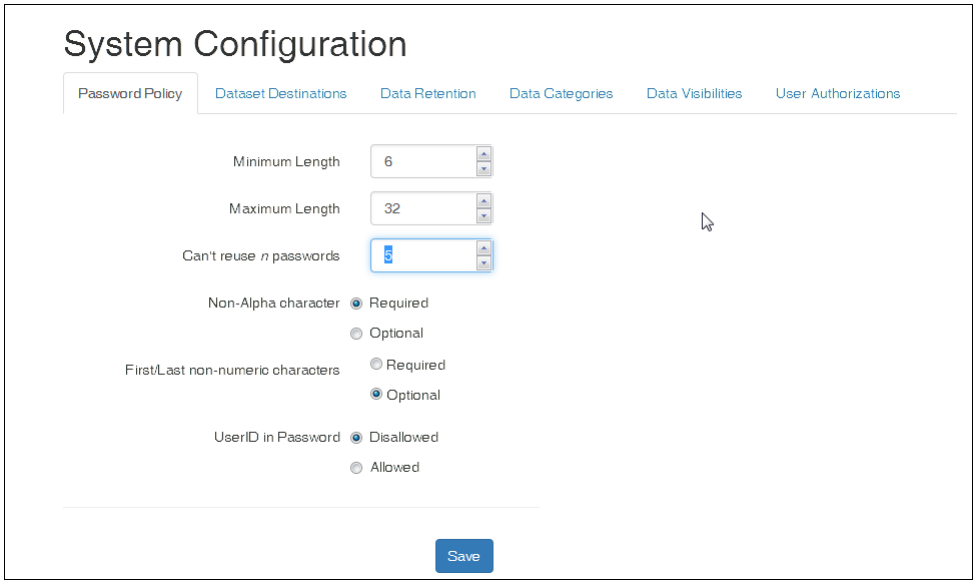
### View Dataset Destination Information

View information about a dataset destination on the **Dataset Destinations** screen of the **System Configuration** page.

To view information for a dataset destination:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

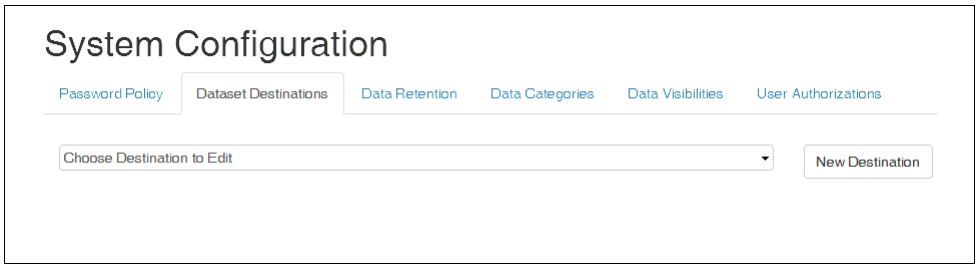
   The **System Configuration** page opens on the **Password Policy** screen.



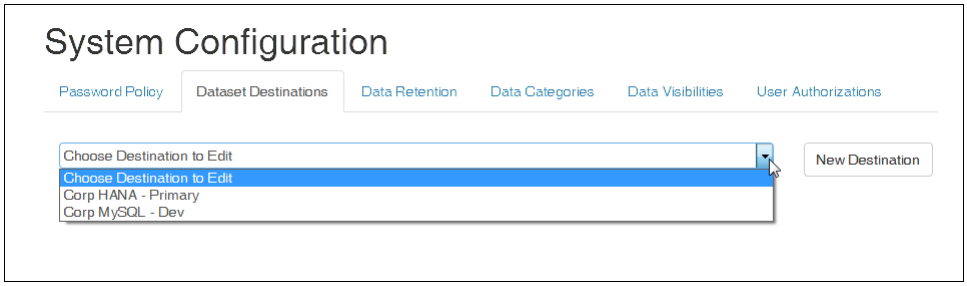

   *What do these fields mean?*

2. Click the **Dataset Destinations** tab.

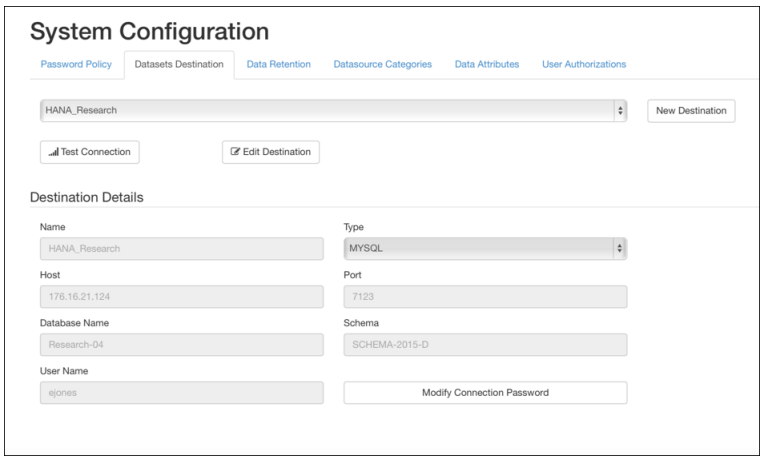
   The **Dataset Destinations** screen opens.




3. Click the drop-down arrow at the right of the **Choose Destination to Edit** field to see configured dataset destinations. Select the destination you want to view.




4. The **Destination Details** screen opens, showing information for the selected dataset.



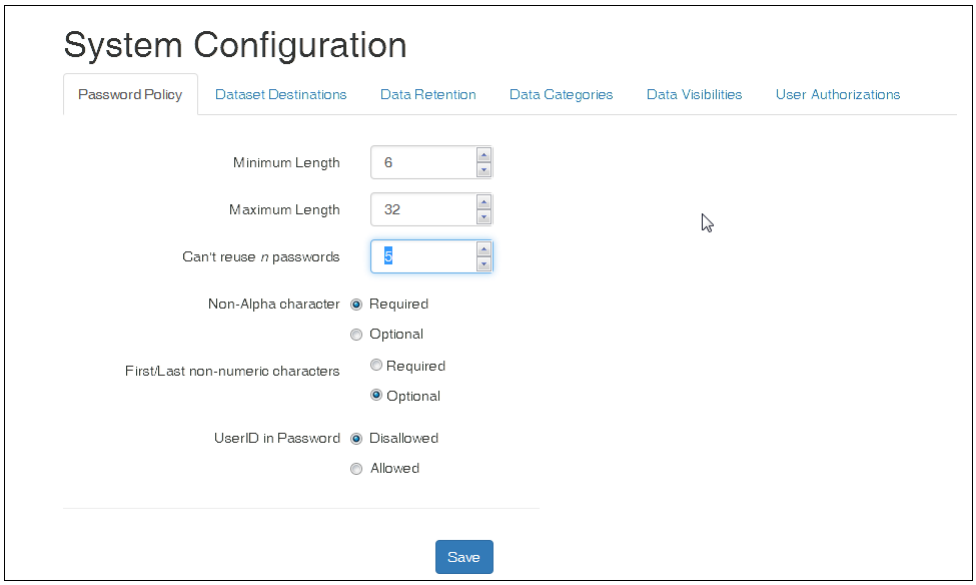

## Test a Dataset Destination Connection

You can test the connection for a dataset destination from the **Dataset Destinations** screen of the **System Configuration** page.

To test a dataset destination connection:

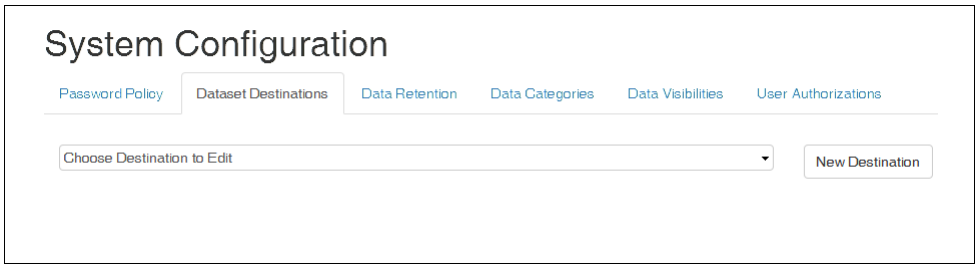
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

   The **System Configuration** page opens on the **Password Policy** screen.

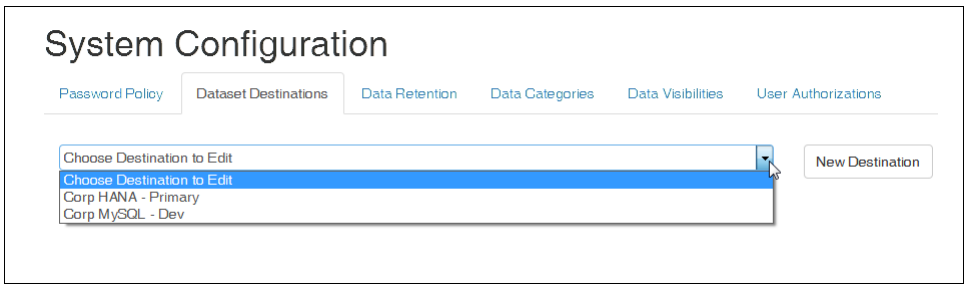



2. Click the **Dataset Destinations** tab.

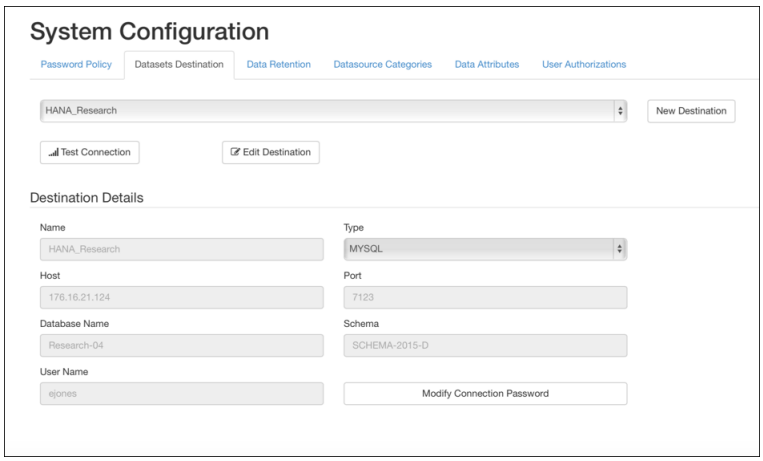
   The **Dataset Destinations** screen opens.




3. Click the drop-down arrow at the right of the **Choose Destination to Edit** field to see configured dataset destinations. Select the destination you want to test.

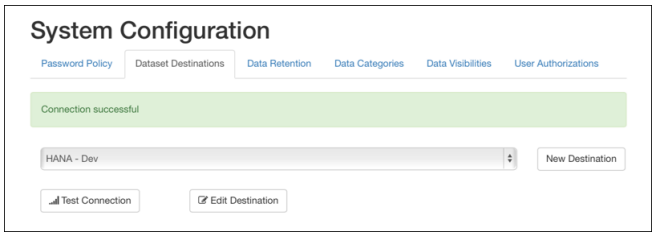



4. The **Destination Details** screen opens, showing information for the selected dataset.




5. Click the **Test Connection** button.

   The system initiates an attempt to connect with the dataset destination system. If it connects the system displays a Connection successful message.




   If the connection does not succeed, the system opens a diagnostic pane that displays status and error messages related to the connection attempt.

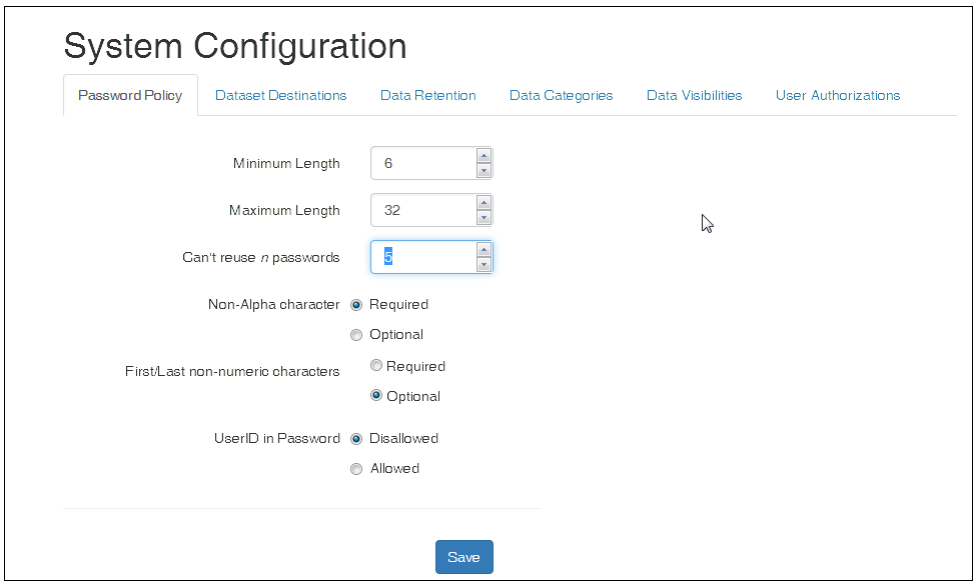
## Modify Dataset Destination Information

Modify information for a dataset destination on the **Dataset Destinations** screen of the **System Configuration** page.

To modify information for a dataset destination:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

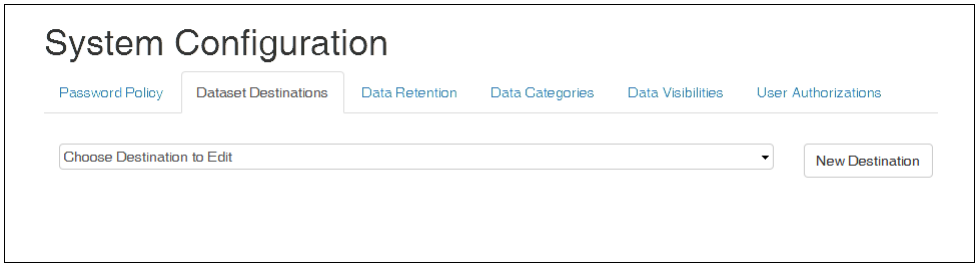
   The **System Configuration** page opens on the **Password Policy** screen.



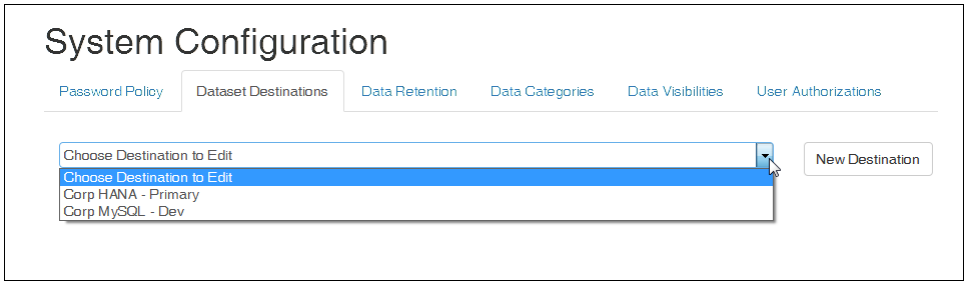

*What do these fields mean?*

2. Click the **Dataset Destinations** tab.

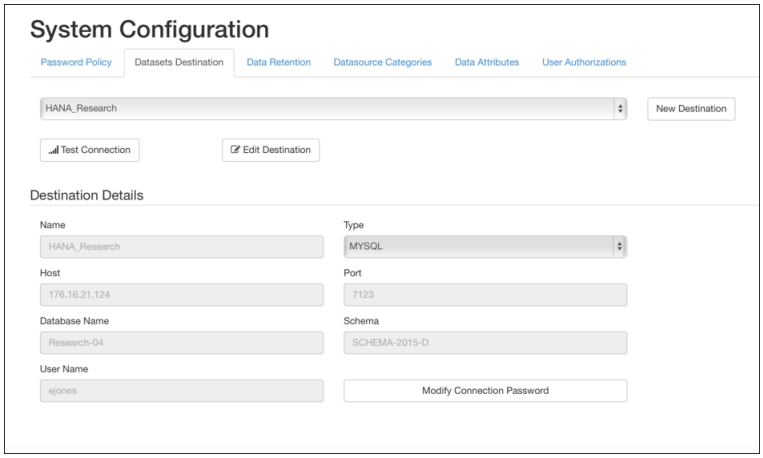
   The **Dataset Destinations** screen opens.




3. Click the drop-down arrow at the right of the **Choose Destination to Edit** field to see configured dataset destinations. Select the destination you want to modify.

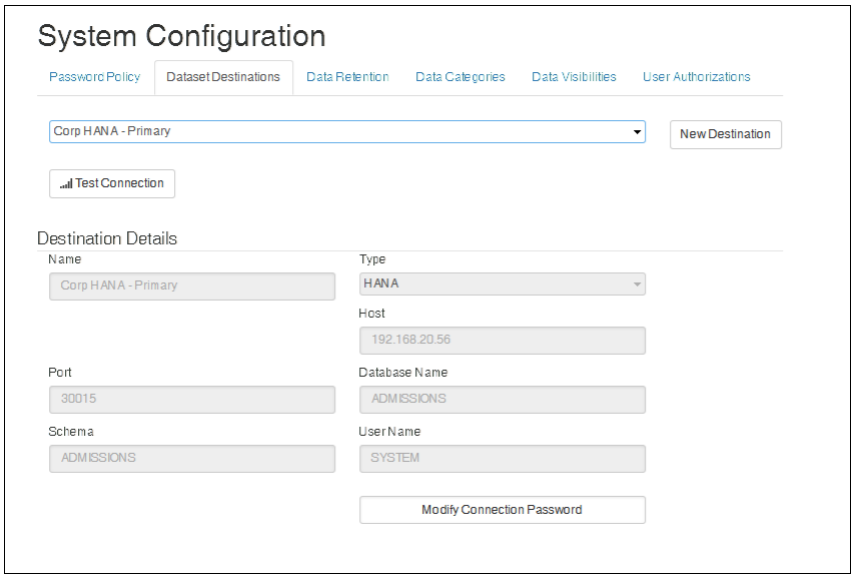



4. The **Destination Details** screen opens, showing information for the selected dataset.



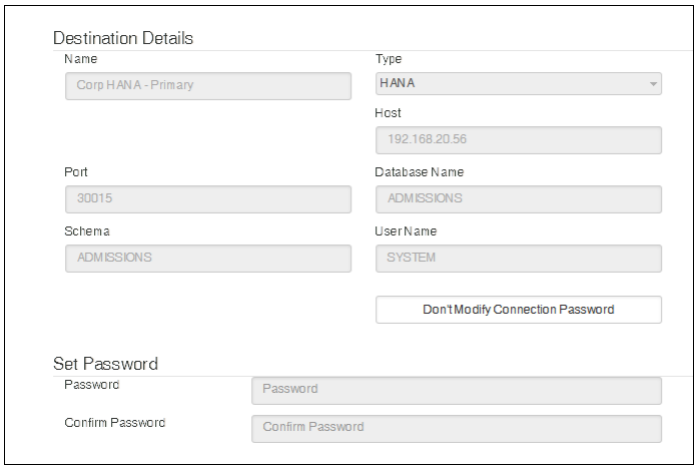

*What do these fields mean?*

5. Click the **Edit Destination** button. The fields on the **Destination Details** screen become editable, and the **Delete Destination** buttons and **Save Destination** buttons appear.




6. Make your changes. If you need to change the destination password information, click the **Modify Connection Password** button.

   The **Set Password** pane opens underneath the other destination information.




7. Complete your changes, then click the **Save Destination** button to save the changed information.

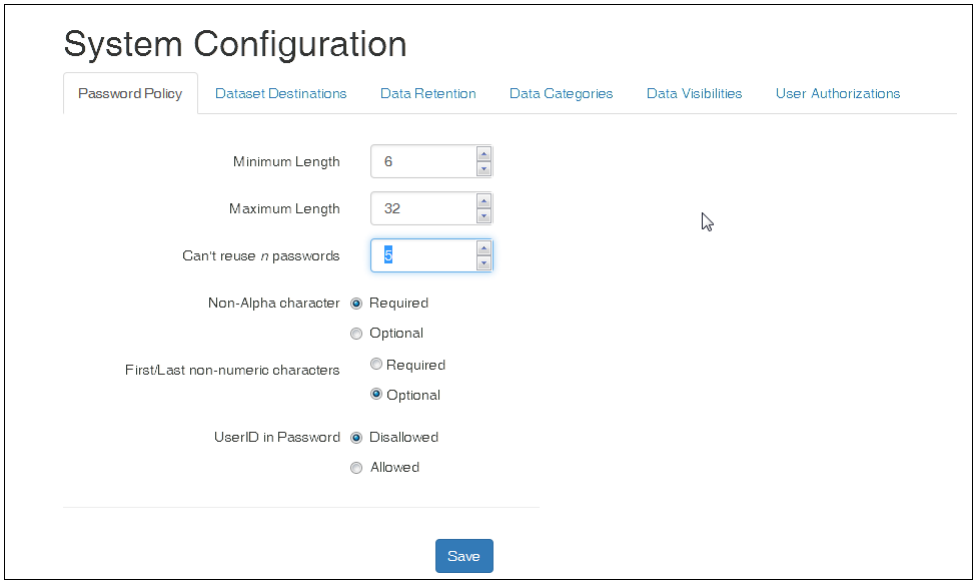
### Delete a Dataset Destination

Delete a dataset destination on the **Dataset Destinations** screen of the **System Configuration** page.

To delete a dataset destination:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

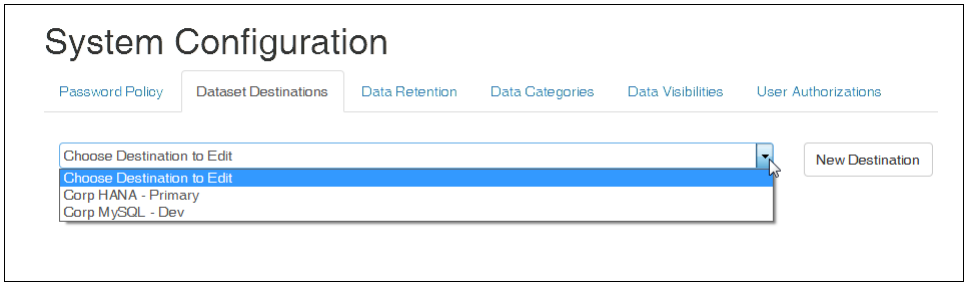
   The **System Configuration** page opens on the **Password Policy** screen.



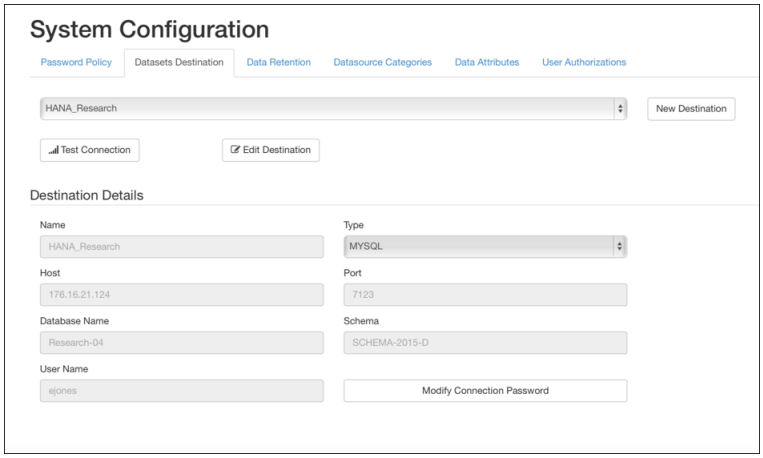

2. Click the **Dataset Destinations** tab.

   The **Dataset Destinations** screen opens.

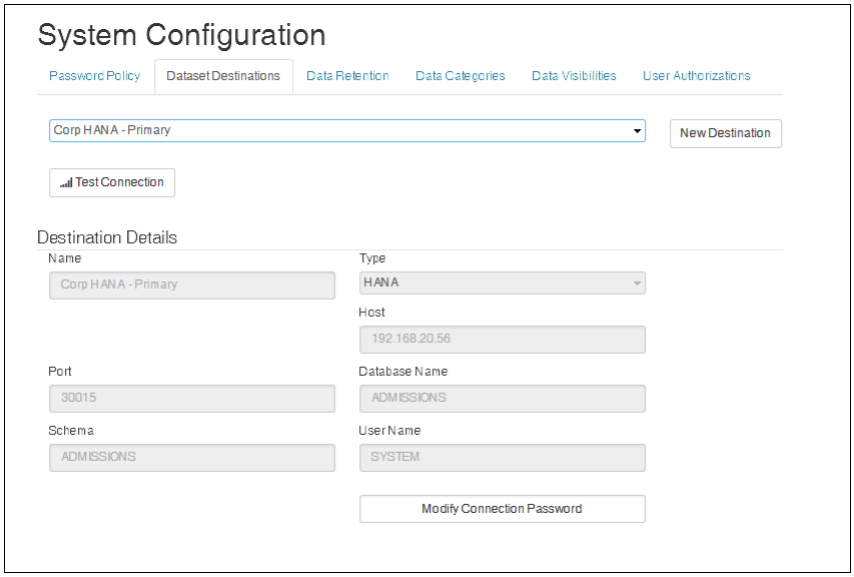
3. Click the drop-down arrow at the right of the **Choose Destination to Edit** field to see configured dataset destinations. Select the destination you want to delete.




4. The **Destination Details** screen opens, showing information for the selected dataset.




5. Click the **Edit Destination** button. The fields on the **Destination Details** screen become editable, and the **Delete Destination** buttons and **Save Destination** buttons appear.




6. Click the **Delete Destination** button. The system asks you to confirm permanent deletion. Click **Delete**.

## Data Retention Behavior

Each data collection has a data policy that specifies, among other things, how long data items should be retained in the system. When the item's "time to live" expires, PHEMI Central deletes the item from the data store. The data retention behavior, which is set in system configuration, specifies when PHEMI Central checks the data store to see what data, if any, is expired and should be deleted from the data store.

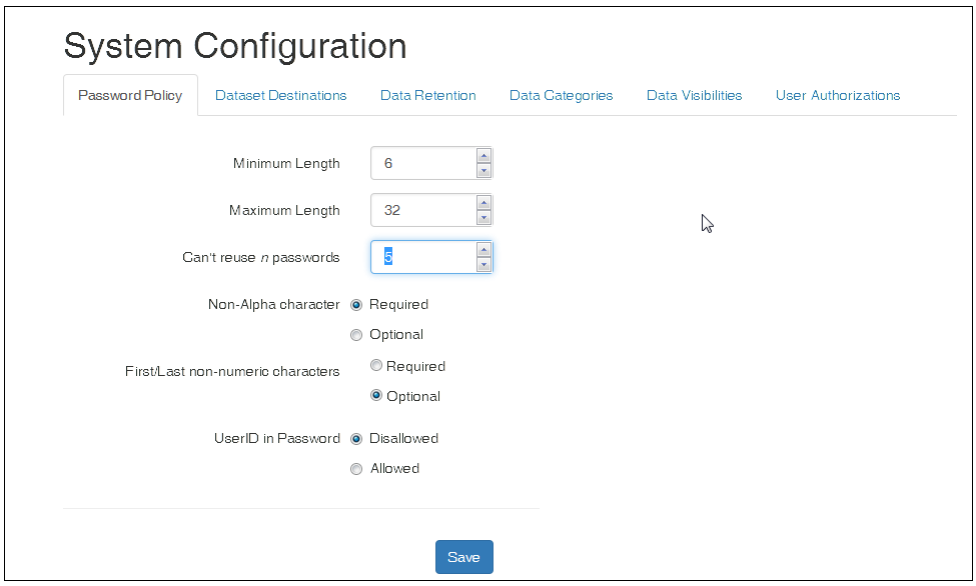
You can configure PHEMI Central to check on a schedule, or you can manually initiate checking.

### View the Data Retention Schedule

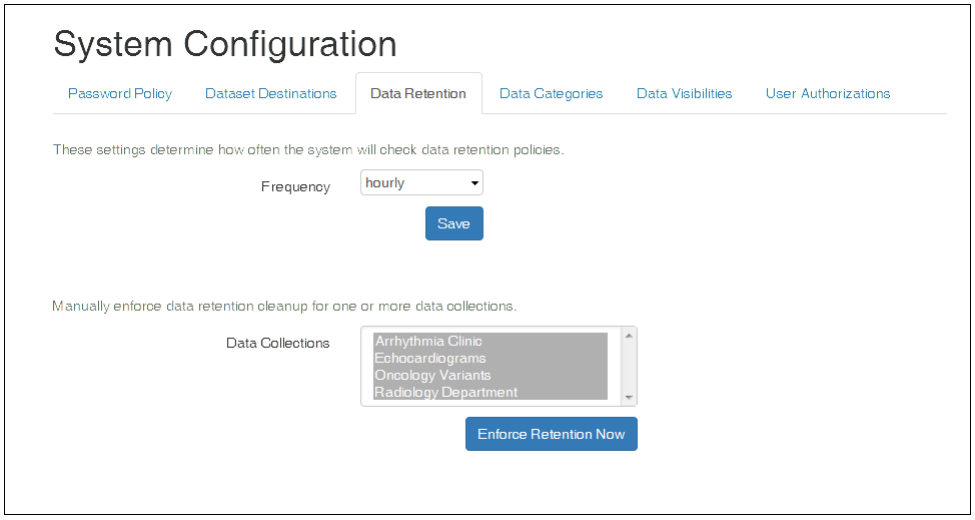
View the schedule for data retention behavior on the **Data Retention** screen of the **System Configuration** page.

To see the configured data retention schedule:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Retention** tab.
   The **Data Retention** screen opens, showing the schedule for automatic checking.

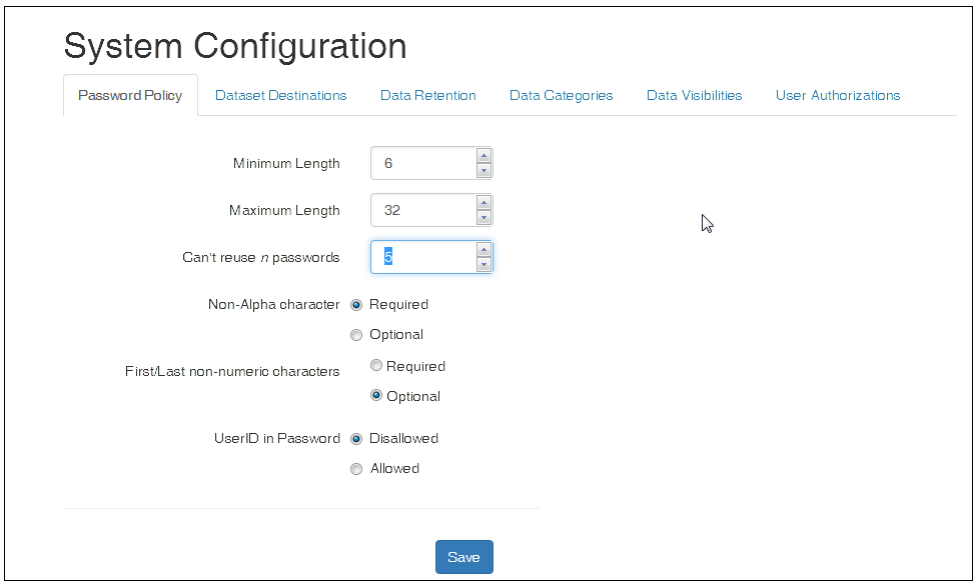



### Set the Data Retention Schedule

Set the schedule for data retention on the **Data Retention** screen of the **System Configuration** page.

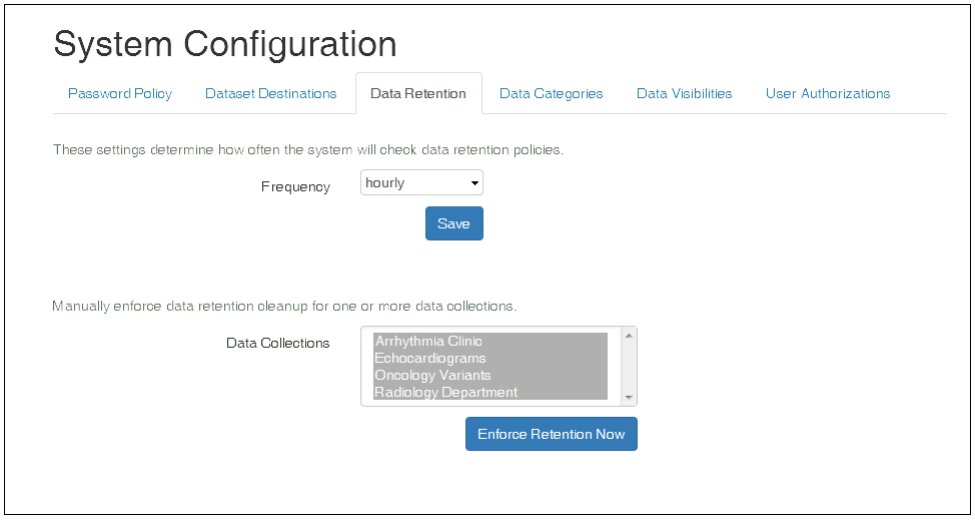
To set how often the system checks for expired data:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Retention** tab.

   The **Data Retention** screen opens.




3. In the **Frequency** field, choose between **hourly**, **daily**, or **weekly**, or else specify the number of minutes between checks.
4. Click **Save** to save the changes.

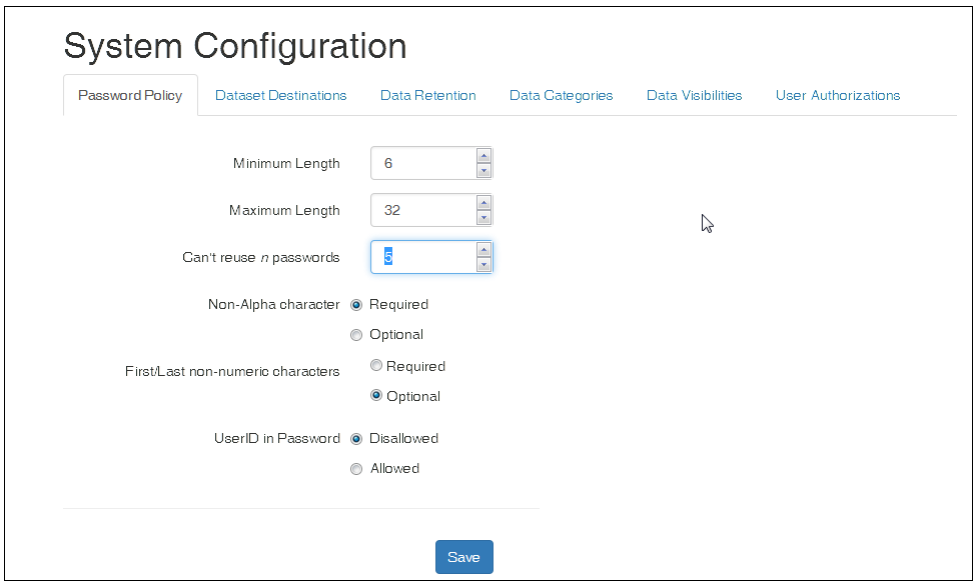
### Manually Trigger Data Retention Check

Trigger data retention checking at any time on the **Data Retention** screen of the **System Configuration** page.

To manually initiate checking for expired data:

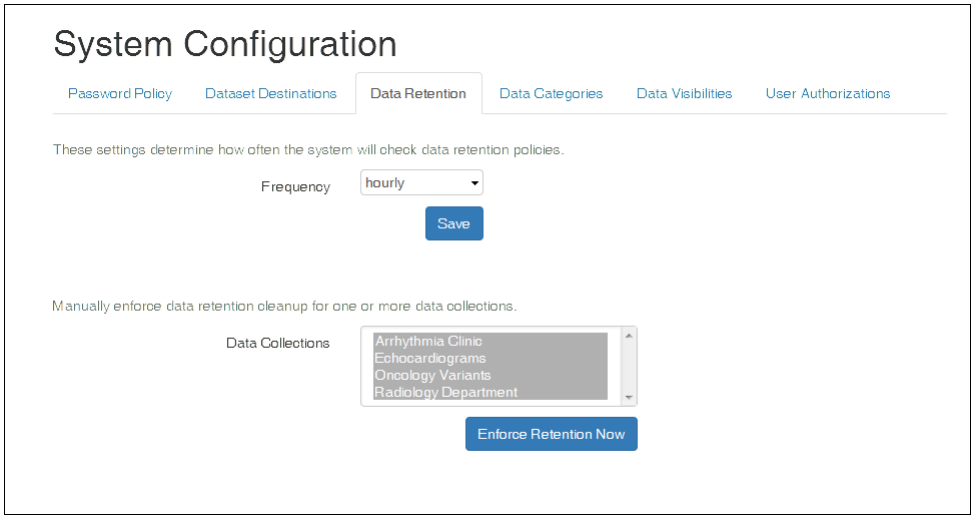
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Retention** tab.

   The **Data Retention** screen opens.




3. In the **Data Collections** field, select all the data collections you want to check.
4. Click **Enforce Retention Now**. PHEMI Central checks the selected data collections for expired data. If expired data is detected, the system deletes it from the data store.

## Data Categories

High-level data categories help you classify the data that will be stored in PHEMI Central.

Each data category can include multiple data sources, systems, or collections. For example, an organization might have a category BILLING, which could include data from several different billing systems in the organization. Another might have a category GENOMICS, which might contain genomic data submitted from different research organizations.

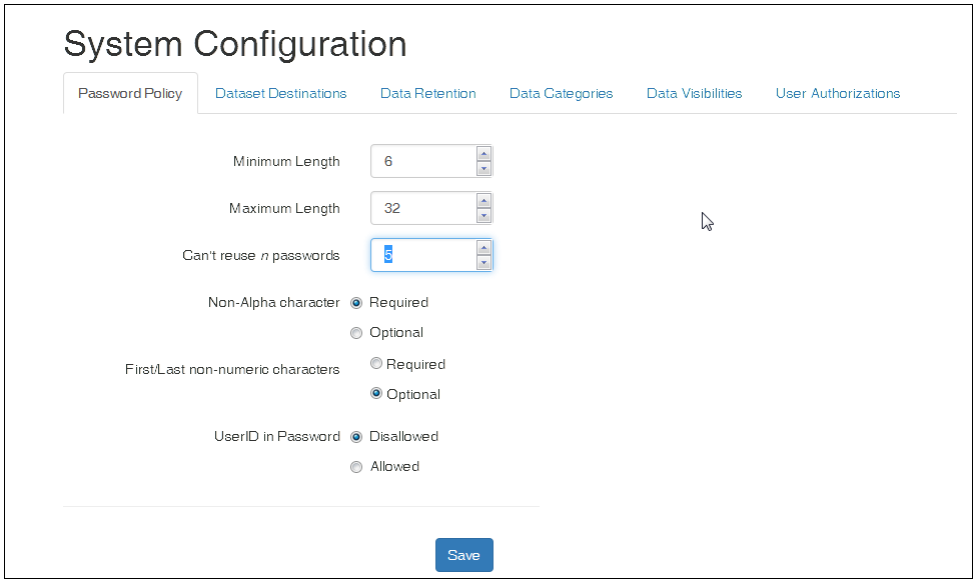
### View Data Categories

View the names of defined data categories on the **Data Categories** screen of the **System Configuration** page, or from the **Data Collections** page.

To view data category names from the **System Configuration** page:

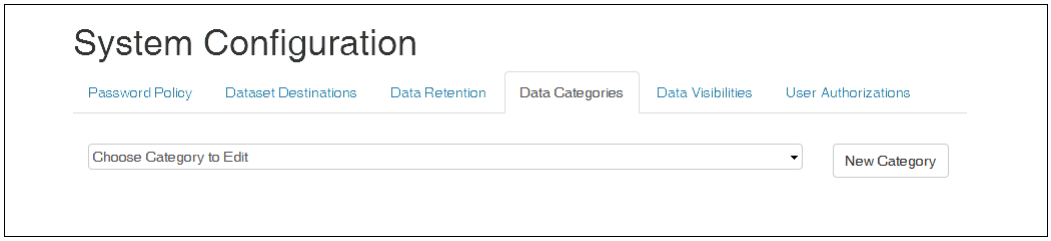
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

The **System Configuration** page opens on the **Password Policy** screen.

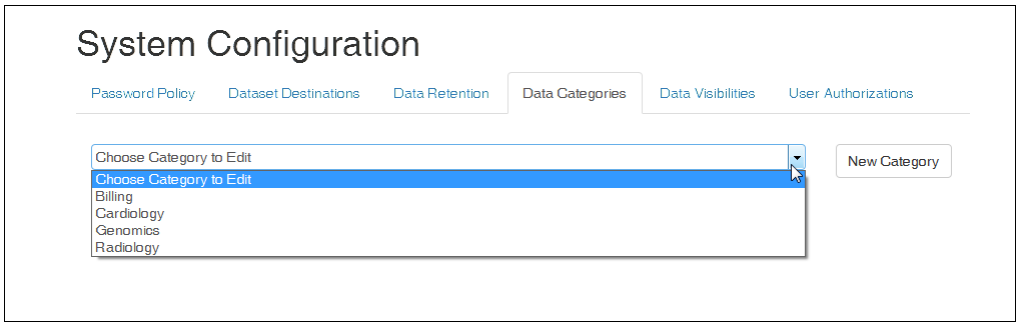



2. Click the **Data Categories** tab.

   The **Data Categories** screen opens.




3. In the **Choose Category to Edit** field, click the drop-down arrow to see the list of defined data categories.




You can also view data categories from the **Data Collections** page.

4. Open the **Data Collections** screen, by clicking the **Data Collections** icon in the left navigation bar.

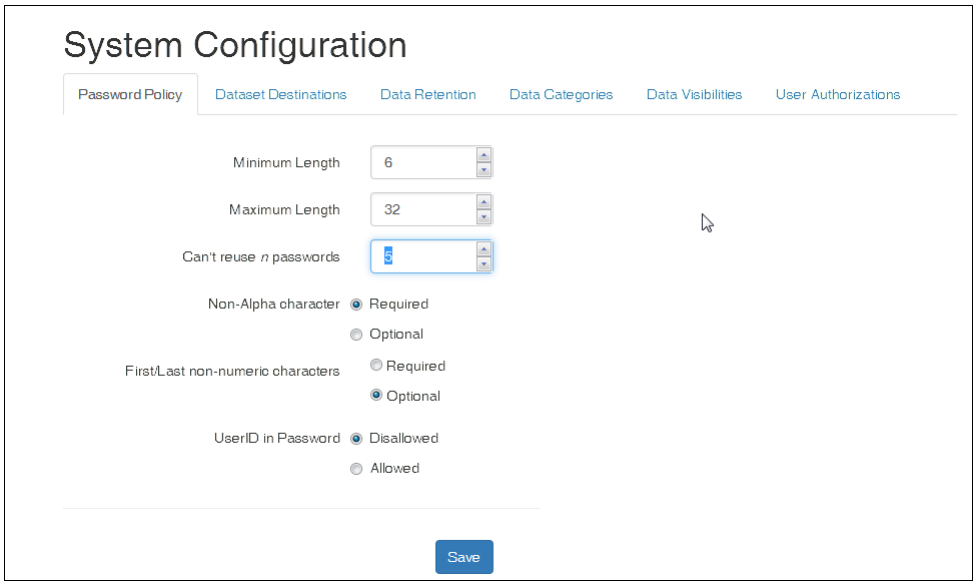
   The **Data Collections** page opens, showing all defined data categories.

### Add a Data Category

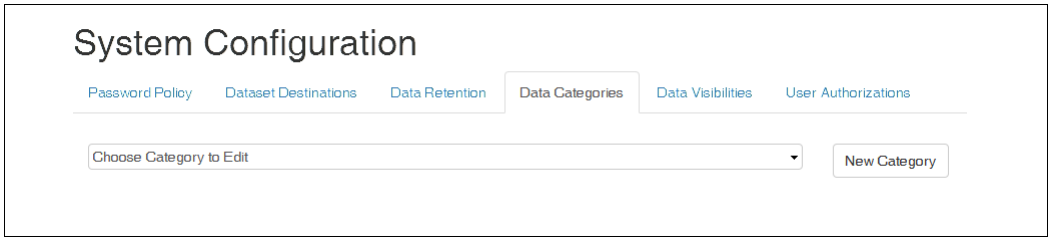
Add a data category on the **Data Categories** screen of the **System Configuration** page.

To add a data category:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.

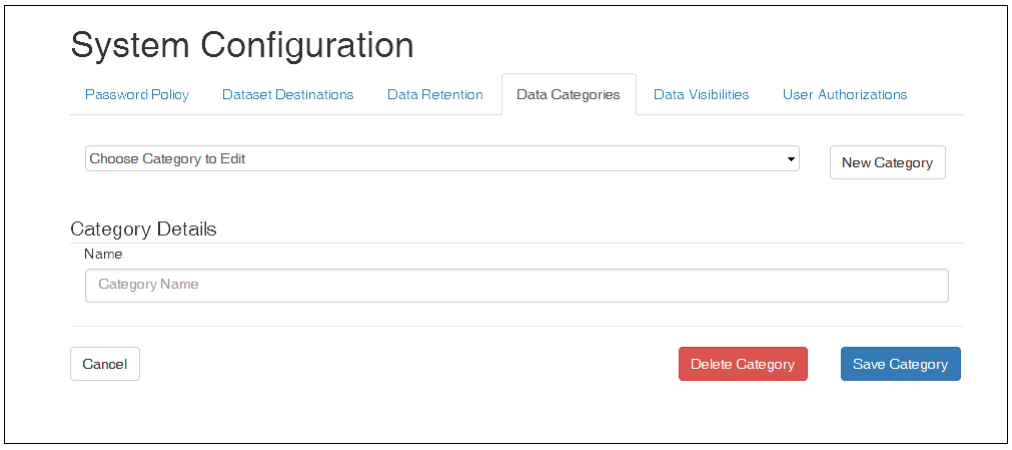



2. Click the **Data Categories** tab.
   The **Data Categories** screen opens.




3. Click the **New Category** button.

   The **Data Categories** screen expands to show the **Category Details** area.



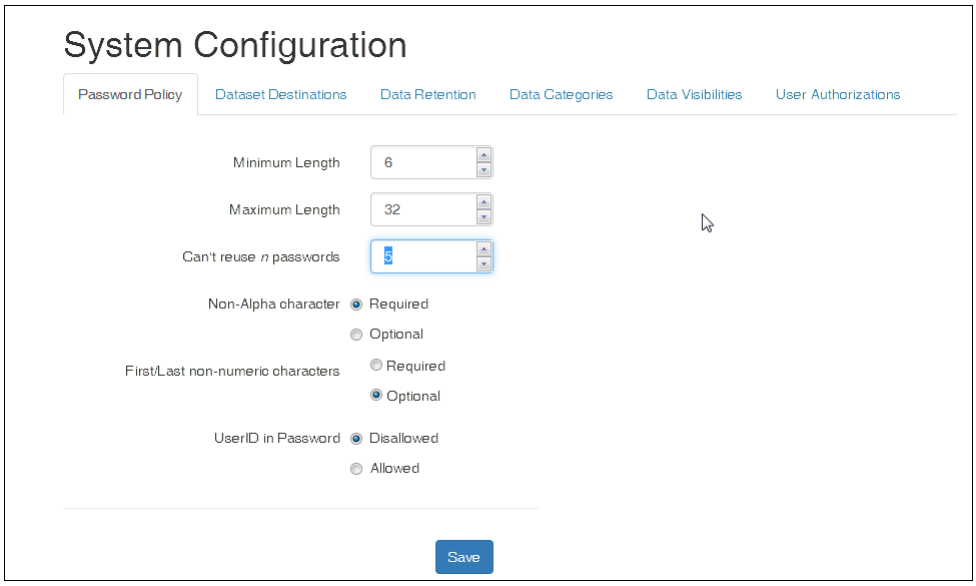

4. Enter a name for the category.
5. Save the data category by clicking the **Save Category** button.

### Edit a Data Category Name

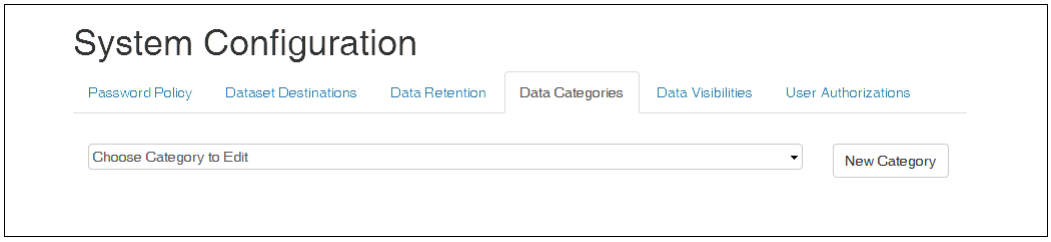
Modify a data category name on the **Data Categories** screen of the **System Configuration** page.

To edit a data category name:

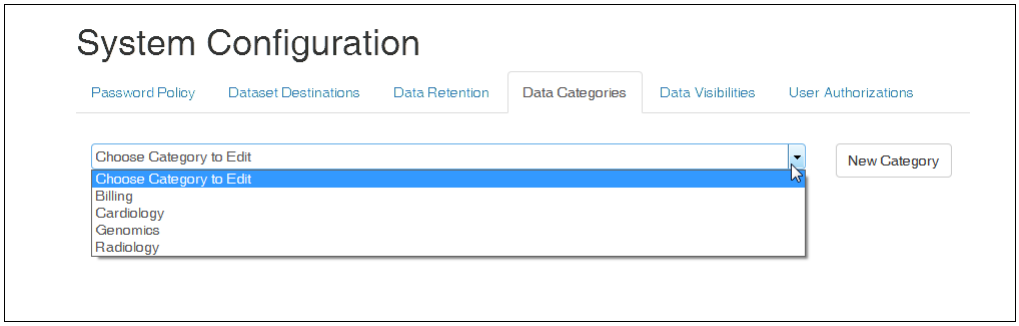
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar. The **System Configuration** page opens on the **Password Policy** screen.



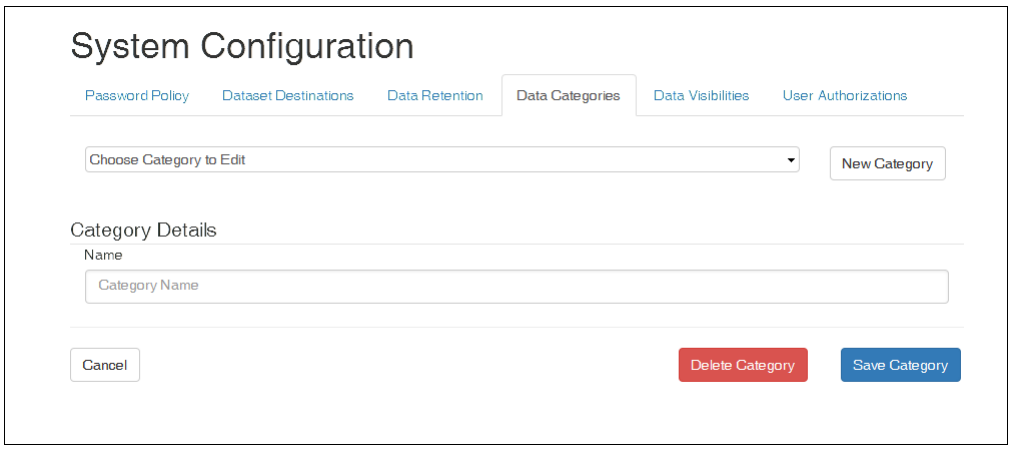

2. Click the **Data Categories** tab.
   The **Data Categories** screen opens.




3. At the right side of the **Choose Category to Edit** field, click the drop-down list and select the category you want to edit.




   The **Category Details** screen opens.



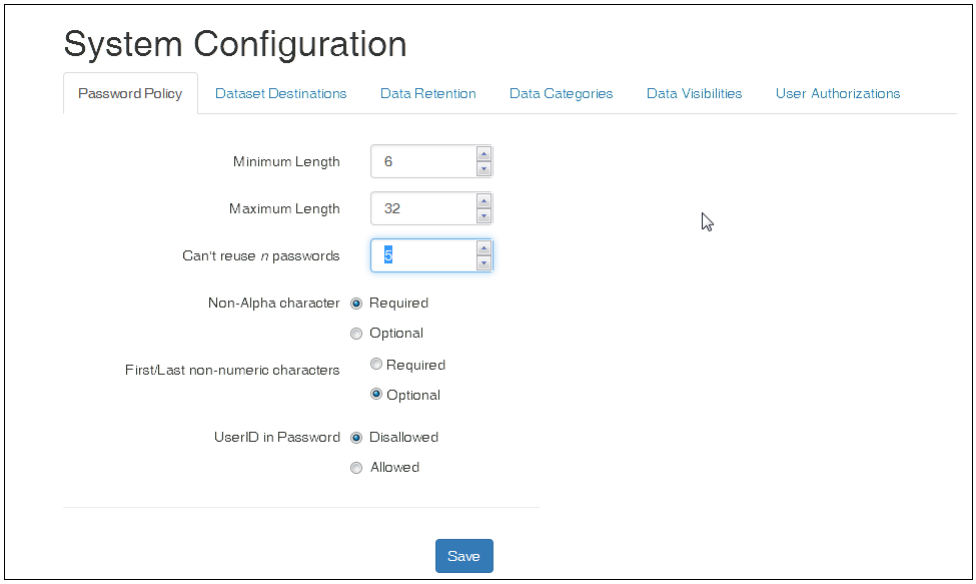

4. Make your edits to the category name.
5. Save the changes by clicking the **Save Category** button.

### Delete a Data Category

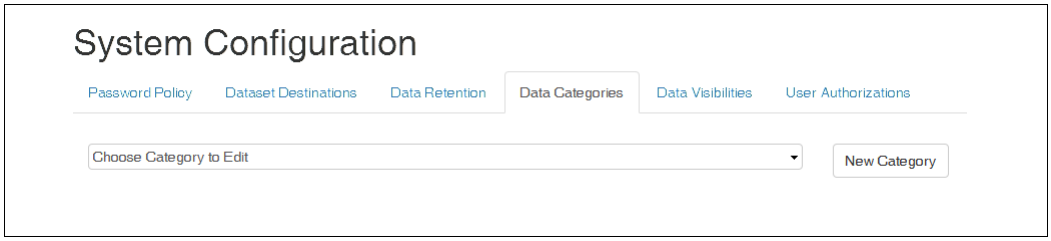
Delete a data category on the **Data Categories** screen of the **System Configuration** page.

To delete a data category:

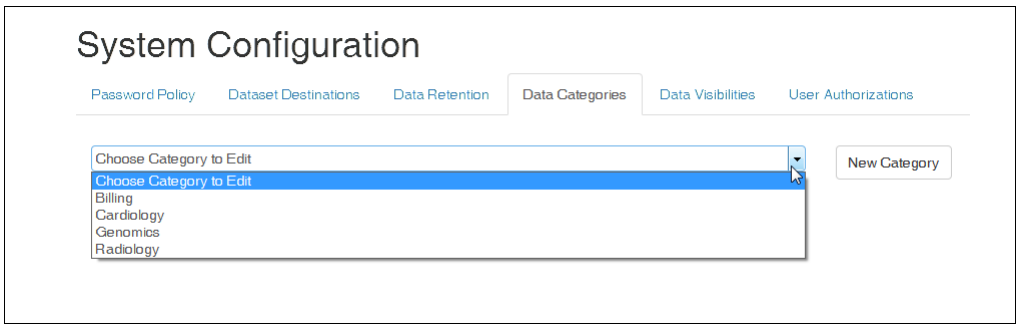
1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.



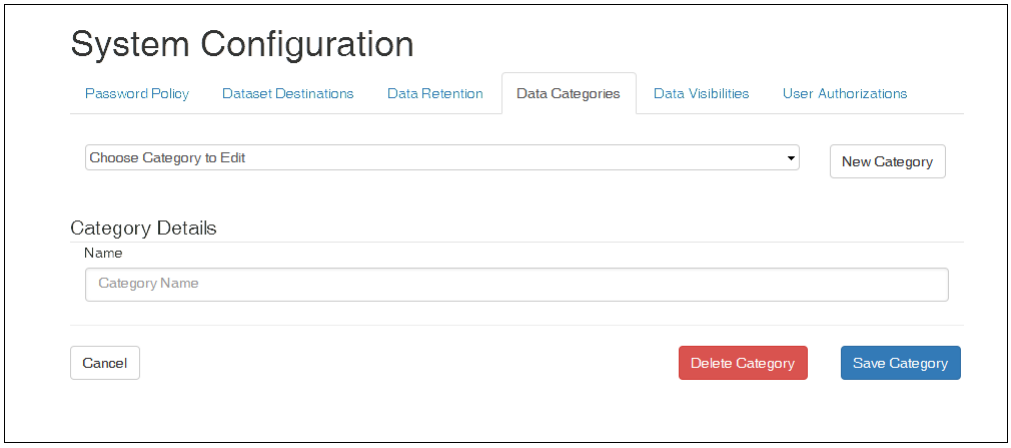

2. Click the **Data Categories** tab.
   The **Data Categories** screen opens.




3. At the right side of the **Choose Category to Edit** field, click the drop-down list and select the category you want to delete.




The **Category Details** screen opens.




4. Click the **Delete Category** button.
5. The system asks you to confirm permanent deletion of the category. Click **Delete**.

## Data Visibilities

Data visibilities identify the privacy requirements for data.

All raw data and derived data stored in PHEMI Central can be tagged with labels that provide information about the data's sensitivity. This sensitivity is described in terms of the visibility the data should have to different system users. The visibility tags you define for your data should reflect the visibility the data is intended to have according to your organization's governance policy.

The visibility of data in your organization should be set out in your organization's governance policy.

*Tell me about governance policies. How do I define a governance policy?*

Working from the governance policy, the PHEMI Administrator configures data visibilities as part of system configuration.

You can define data visibilities to suit your organization's needs. Possible examples are CONFIDENTIAL to identify data that is sensitive and requires a user to a certain level of access or certification to access, PII to identify Personally Identifiable Information, PHI to identify personal health information, or SECRET to identify especially sensitive material.

**Note:** Once defined, the name of a data visibility may not be edited. To change the name of a user visibility, you must delete the visibility and recreate it with the new name.

The data visibilities specified for a data collection are referred to in access policies and matched against user authorizations to determine what actions, if any, a given user is allowed to perform on the data. The access policy can then be applied to a data collection or dataset during system configuration.

*Tell me about user authorizations. Tell me about access policies.*

In addition, any data visibility defined in the system can be used by a Data Processing Function (DPF) to selectively tag any derived data fields. The derived data elements can be assigned different privacy levels given that the

visibilities have been configured in the system. For example, the Social Security Number extracted from a health record web form can be tagged CONFIDENTIAL if that data visibility has been defined in the system. This mechanism provides field-level privacy protection.

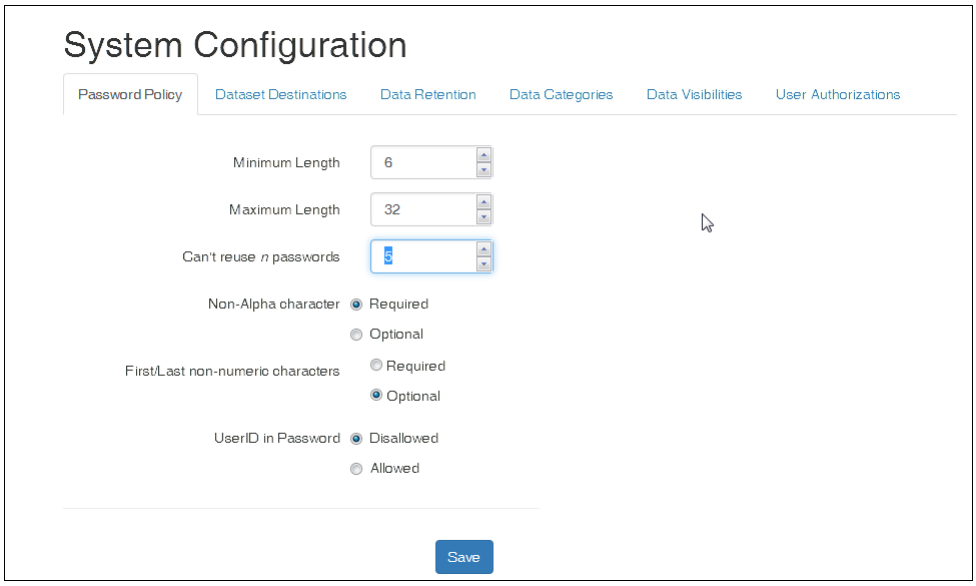
Similarly, the DPF can extract several versions of a field; for example, one version containing the original personally identifying information and another version that is de-identified, perhaps by encrypting it or masking it, and then the DPF apply different data visibilities to each version of the field.

### View Configured Data Visibilities

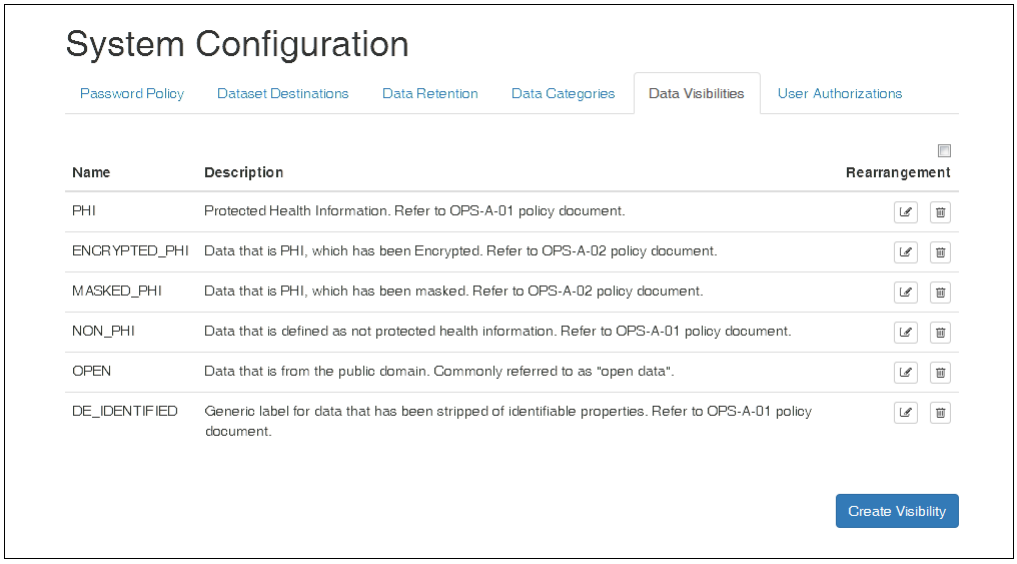
View configured data visibilities on the **Data Visibilities** screen of the **System Configuration** page.

To view the data visibilities that have been configured in the system:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Visibilities** tab.
   The **Data Visibilities** screen opens, listing all the data visibilities configured for the system.



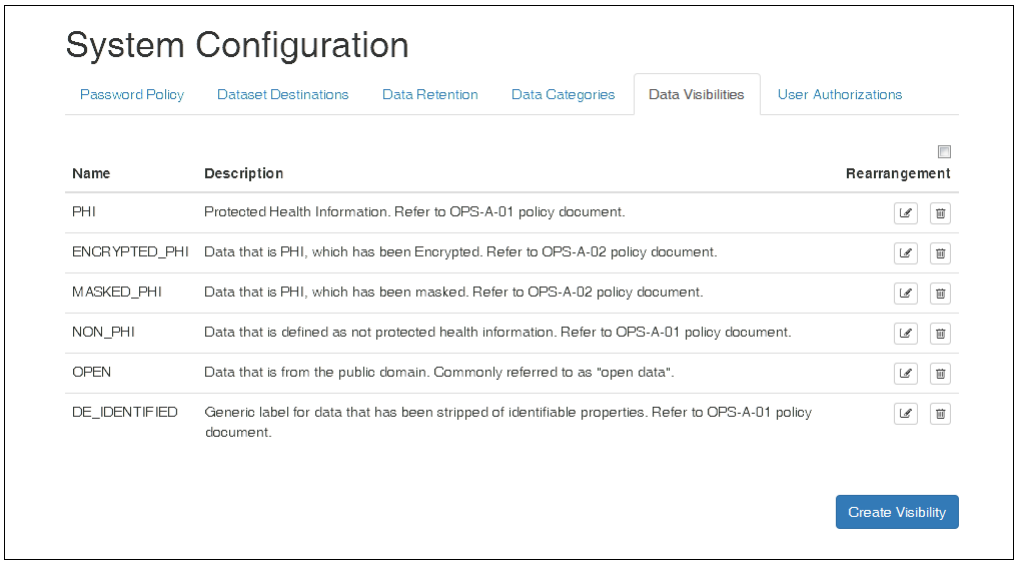

### Define Data Visibilities

Define a data visibility on the **Data Visibilities** screen of the **System Configuration** page.

**Note:** Once defined, the name of a data visibility may not be changed. To change the name, you must delete the visibility and recreate it with the new name.

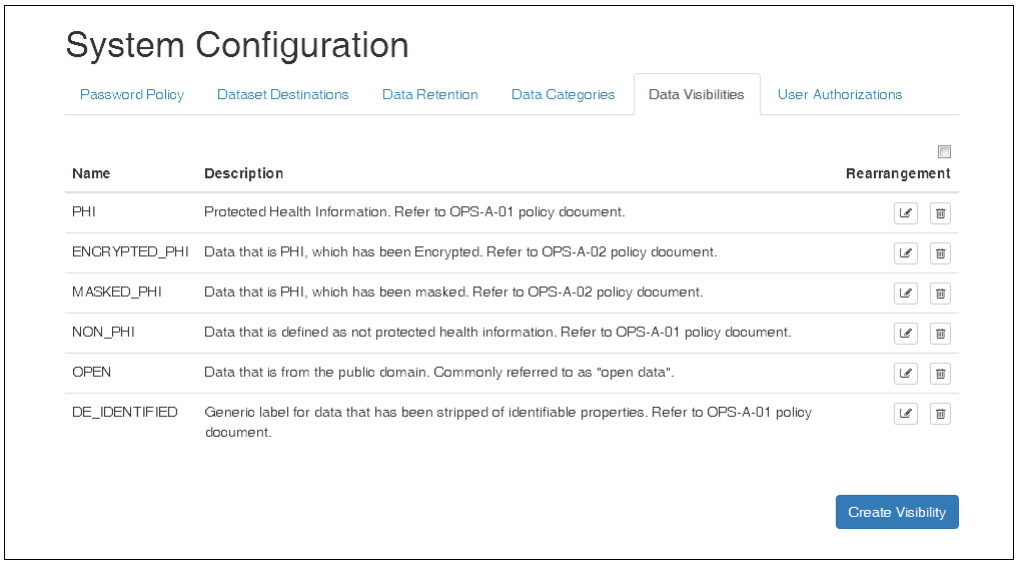
To define data visibilities:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar. The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Visibilities** tab.

   The **Data Visibilities** screen opens.




3. Click the **Create Visibility** button.

   The **Create Visibility** dialog opens.

4. Enter the visibility information.

| Option | Description |
|---|---|
| **Name** | Enter a name for the data visibility; for example, "CONFIDENTIAL," "SECRET" or "PII".<br>Note: The visibility name cannot be subsequently edited. To change the name, you must delete the visibility and recreate it. |
| **Description** | Enter a brief description to describe the intention of the visibility. |

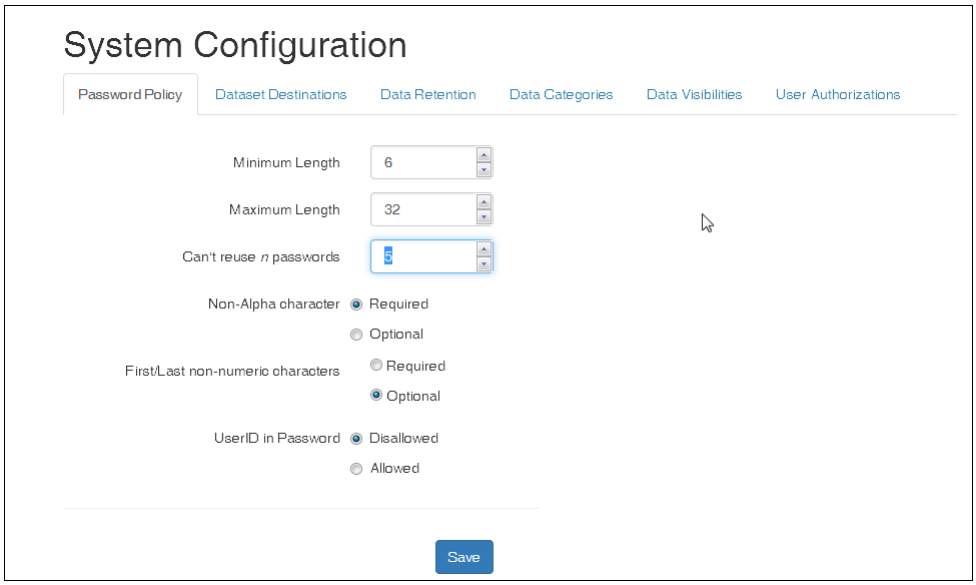
5. Save the visibility information by clicking the **Create** button.

### Change Visibility Description

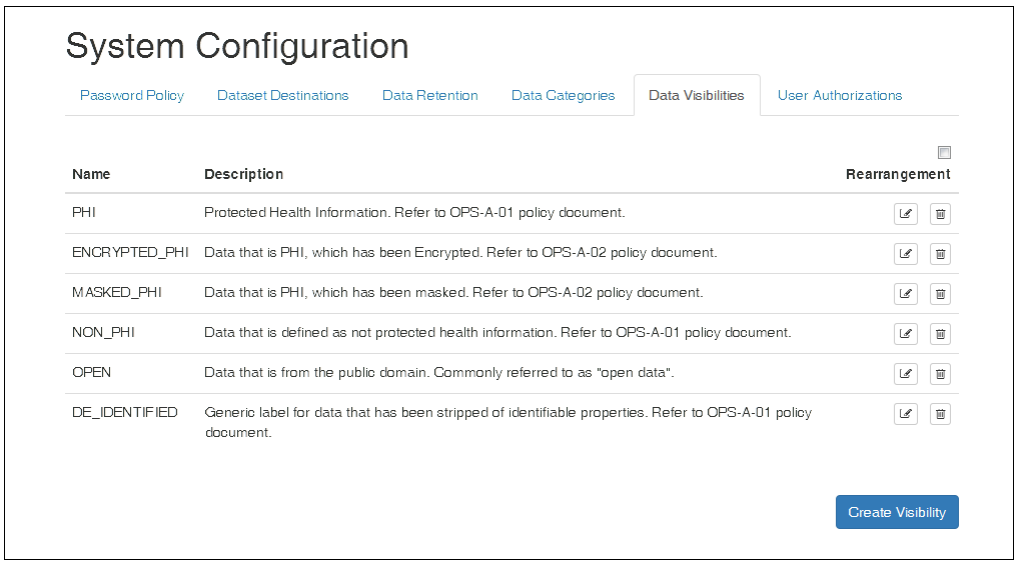
Change a data visibility description on the **Data Visibilities** screen of the **System Configuration** page.

To change data visibility description:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
The **System Configuration** page opens on the **Password Policy** screen.

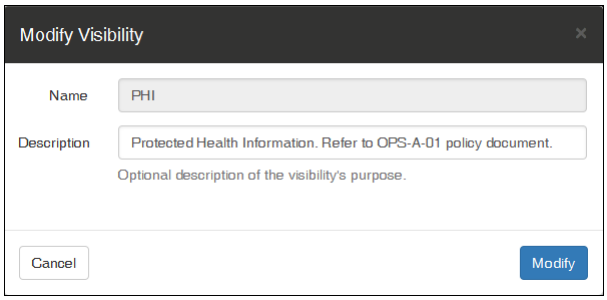



2. Click the **Data Visibilities** tab.
The **Data Visibilities** screen opens.




3. Locate the data visibility you want to change and click the edit icon in the **Rearrangement** column.

   The **Modify Visibility** dialog opens.




4. Modify the description. Click **Modify** to save your changes.

### Rearrange Data Visibilities

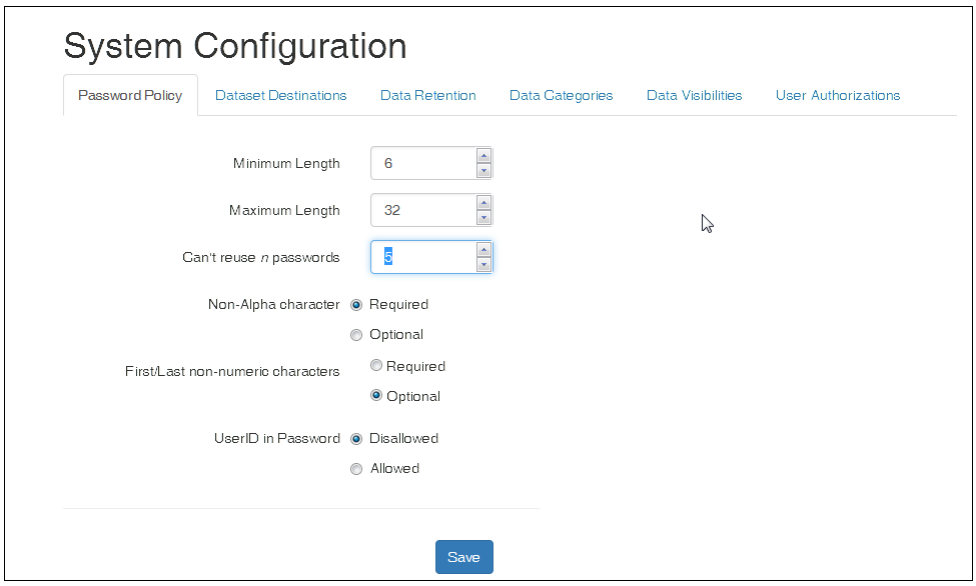
View configured data visibilities on the **Data Visibilities** screen of the **System Configuration** page.

Given that several versions of a field may be produced, your user authorities may entitle you to see more than one version of a field. You can stipulate which data visibilities should be preferred by arranging data visibilities in order of preference. If your user authority and the data's access policies entitle you to see more than one version of a field, the data with first-occurring visibility is given preference.

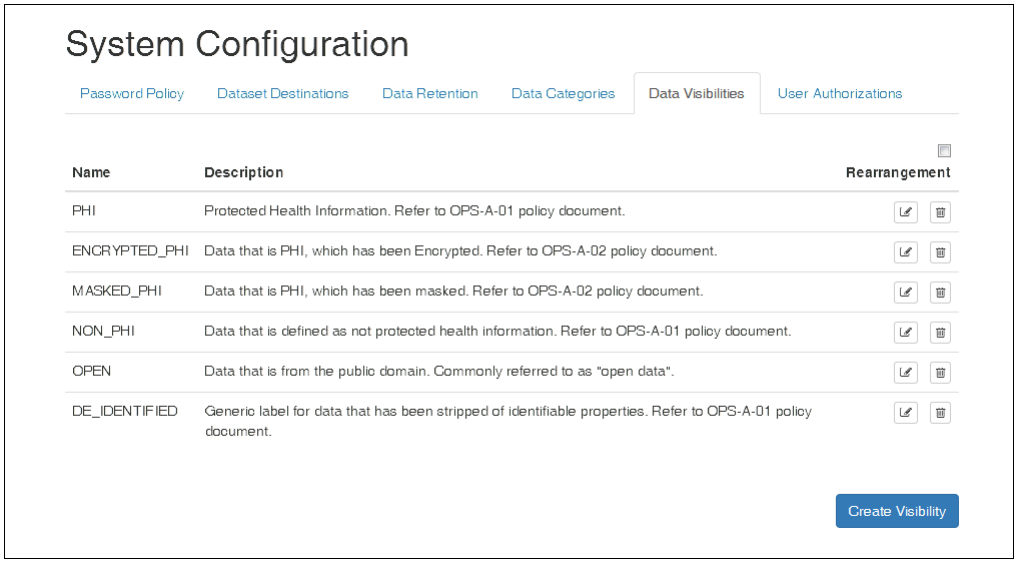
By default, visibilities are ordered by order of configuration. To set the order data visibilities:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Visibilities** tab.
   The **Data Visibilities** screen opens, listing all the data visibilities configured for the system.



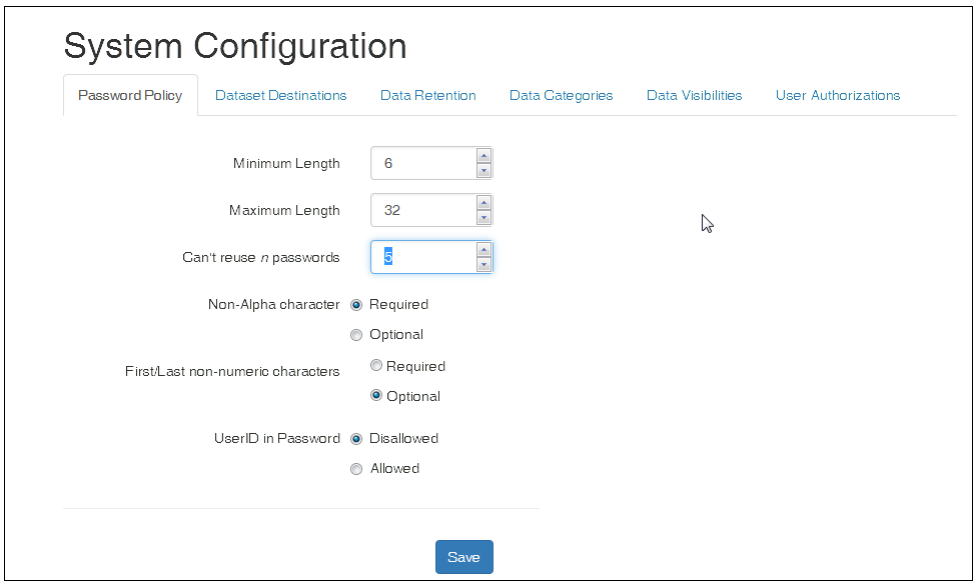

3. Click the **Rearrangement** checkbox.
4. Click in the row you want to rearrange and drag it to the new location.
5. When you have finished rearranging visibilities, click the **Save the New Arrangement** button.

### Delete Visibility

Delete a data visibility on the **Data Visibilities** screen of the **System Configuration** page.

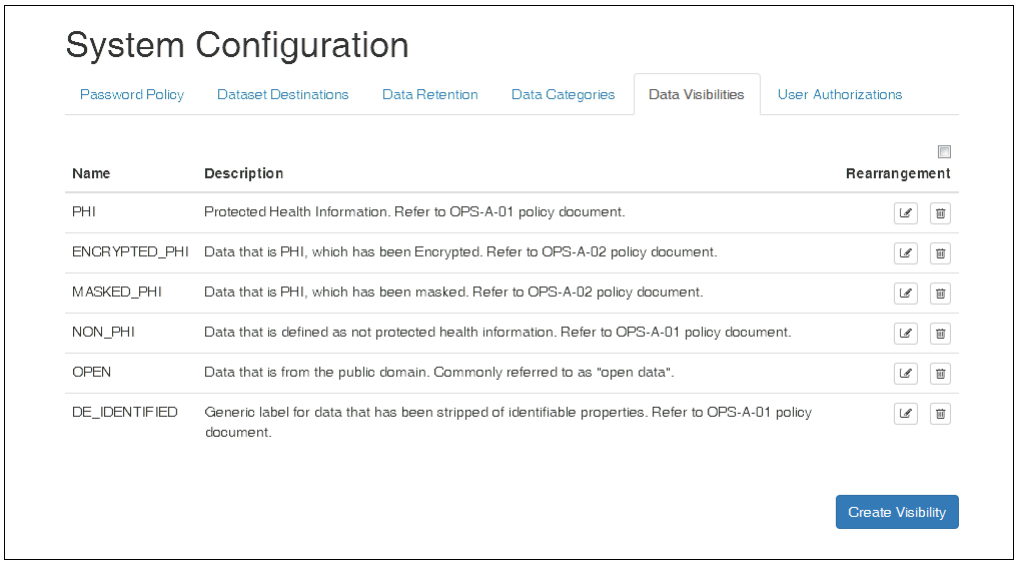
To delete data visibility:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **Data Visibilities** tab.

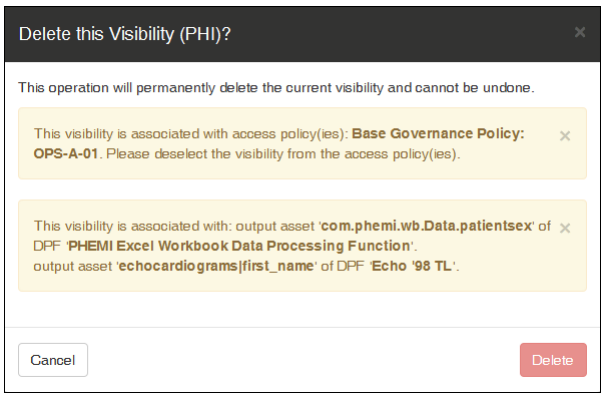
   The **Data Visibilities** screen opens.




3. Locate the data visibility you want to delete and click the trash can icon in the **Rearrangement** column.

   The **Delete this Visibility?** dialog opens.

   Note that data collection configurations and access policies both depend on data visibilities. The system checks for dependencies before it deletes a visibility, and if one of your data collections or one of your access policies refers to the data visibility you want to delete, the system will warn you, listing the data collection(s) and access policy or policies that you need to modify. In this case, the **Delete** button will be unavailable (greyed out). You must eliminate all dependencies before the system will permit you to delete the visibility. When all dependencies have been eliminated, the **Delete** button will become available.




4. When all dependencies have been eliminated, the **Delete** button will become available. Click **Delete** to delete the visibility.

## User Authorizations

User authorizations are configurable attributes you can assign to PHEMI Central users. Authorizations are defined in PHEMI Central by the PHEMI Administrator, who sets them in accordance with the organization's governance policies.

User authorizations are used together with data visibilities to create access policies. The access policy matches the authorization against the data visibility to determine what action, if any, a user may take with respect to accessing the data.

*Tell me about data visibilities. Tell me about access policies.*

Some examples of possible user authorizations are as follows:

- C_LEVEL: The user is a C-Level individual (for example, CEO, COO, CIO, or CTO) with a privileged level of access. Individuals with C_LEVEL authorization, for example, might be permitted to read data with CONFIDENTIAL visibility.
- DOCTOR: A user with DOCTOR authorization might, for example, be permitted to read any information, including personally identifiable information or personal health information.
- ANALYST: A user with ANALYST authorization might be restricted to accessing only the de-identified or nonidentified data.

A user can be assigned multiple authorizations. User authorizations are set by the PHEMI Administrator during system configuration.

All users are assigned the predefined PUBLIC authorization by default. The PUBLIC authorization can subsequently be removed by the PHEMI Administrator.

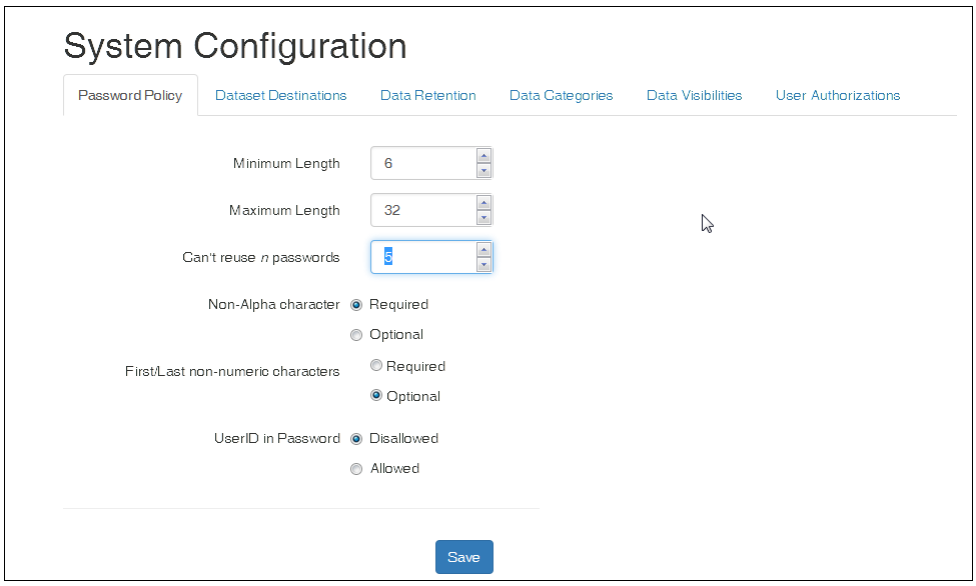
**Note:** Once defined, a user authorization name cannot be changed. To change the authorization name, you must delete the authorization and recreate it with the new name.

### View Defined User Authorizations

View defined user authorizations on the **User Authorizations** screen of the **System Configuration** page.

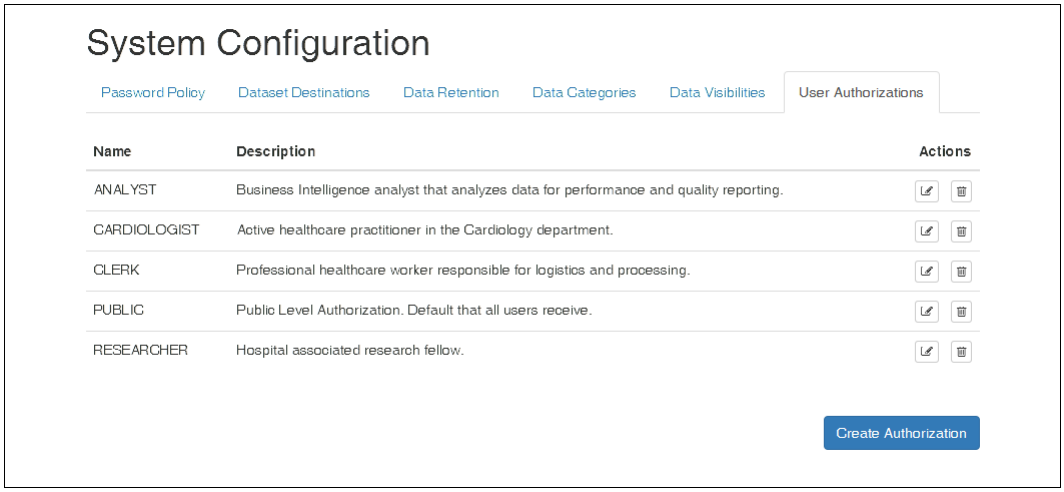
To view defined user authorizations:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar. The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **User Authorizations** tab.

   The **User Authorizations** screen opens, showing you each authorization that has been defined, along with a brief description.




## Define User Authorizations

Define user authorizations on the **User Authorizations** screen of the **System Configuration** page.

The user authorizations you define should reflect your organization's governance policies. Consult your privacy officer, or a person in a similar role, to understand what user authorizations your organization recognizes.

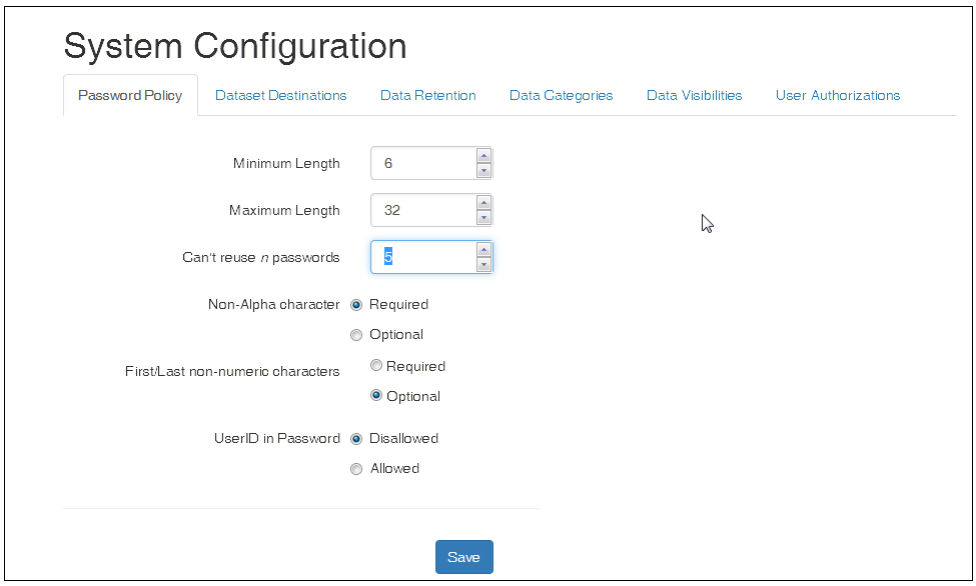
*Tell me about governance policies. How do I define a governance policy?*

**Note:** Once defined, a user authorization name may not be changed. To change the authorization name, you must delete the authorization and recreate it with the new name.

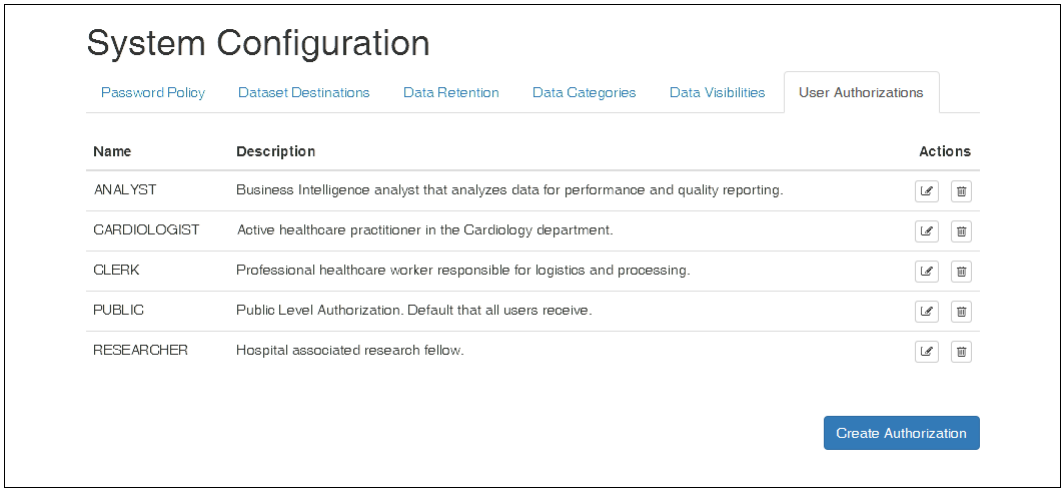
To define a user authorization:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.

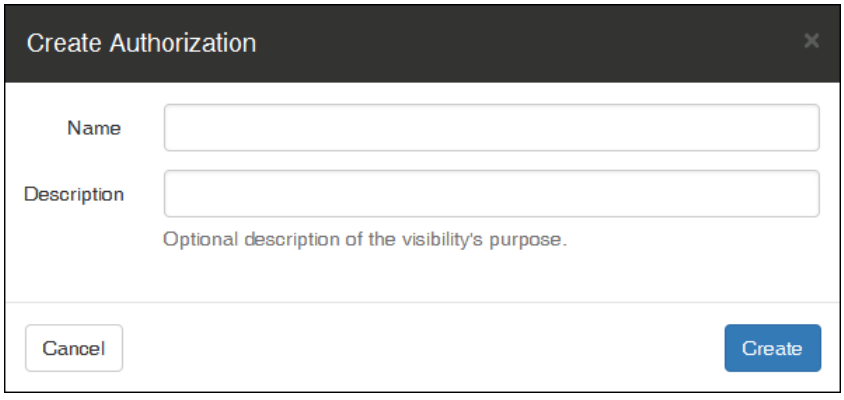
   The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **User Authorizations** tab.
   The **User Authorizations** screen opens.




3. Cick the **Create Authorization** button.
   The **Create Authorization** dialog opens.




4. Specify the authorization information.

| Option | Description |
|---|---|
| **Name** | The authorization label to be applied to users. |

| Option | Description |
|---|---|
| | Note that once defined, the authorization name cannot be changed. To change the name, you must delete the authorization and recreate it. |
| **Description** | A brief description for the authorization. |

5. Click the **Create** button to save the information.

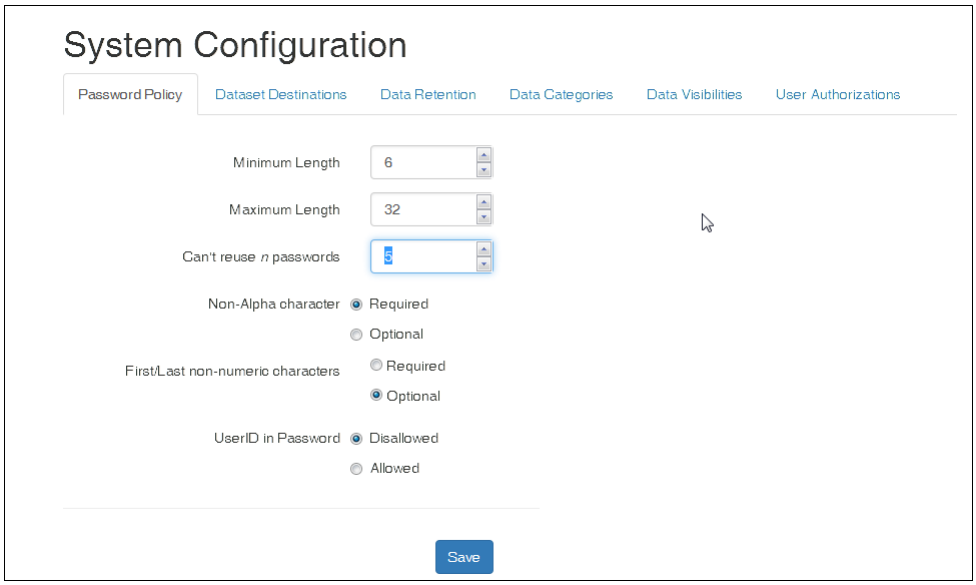
## Change Authorization Description

Change user authorization descriptions on the **User Authorizations** screen of the **System Configuration** page.

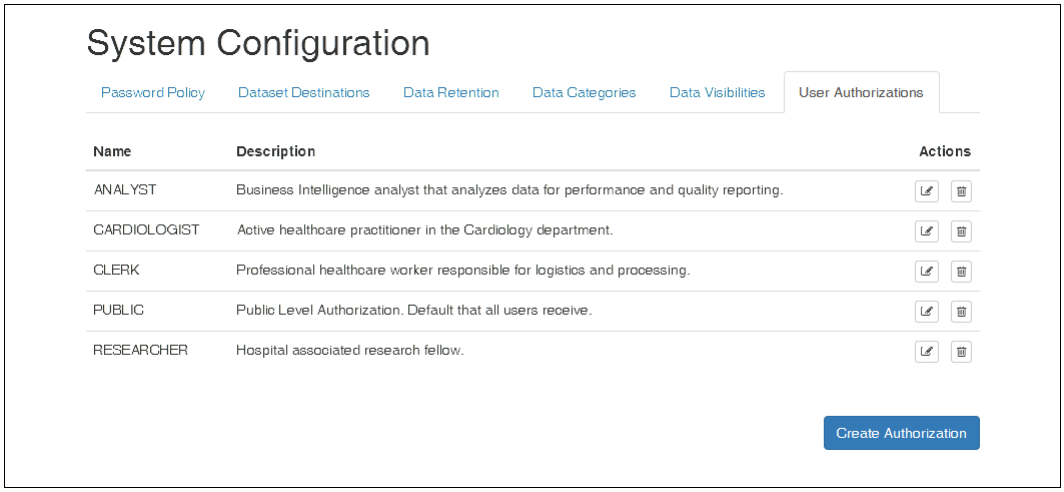
Although you can't edit an authorization name, you can edit the description after the authorization has been created.

To edit an authorization description:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar.
   The **System Configuration** page opens on the **Password Policy** screen.

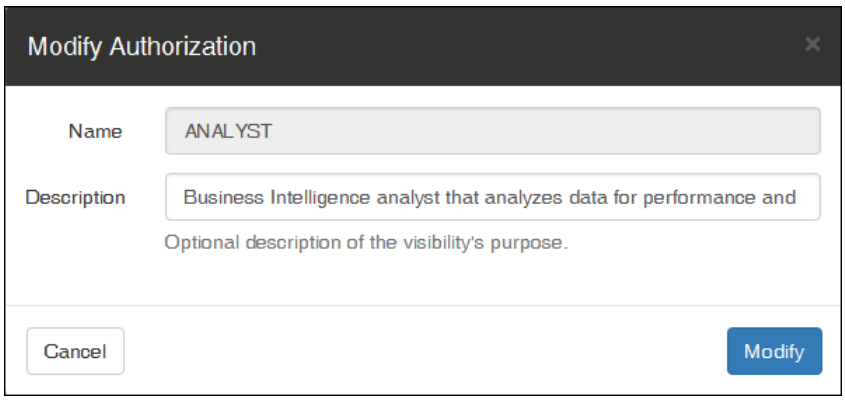



2. Click the **User Authorizations** tab.
   The **User Authorizations** screen opens.




3. Location the authorization you want to edit. Click the edit icon in the **Actions** column.

   The **Modify Authorization** dialog opens.



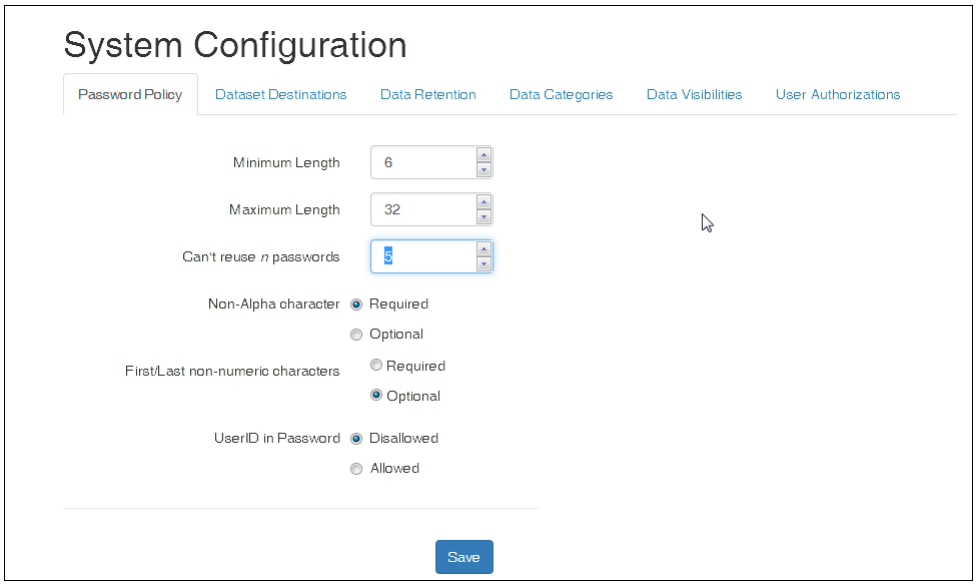

4. Make your change. Click the **Modify** button to save your changes.

### Delete Authorization

Delete a user authorization on the **Data Visibilities** screen of the **System Configuration** page.

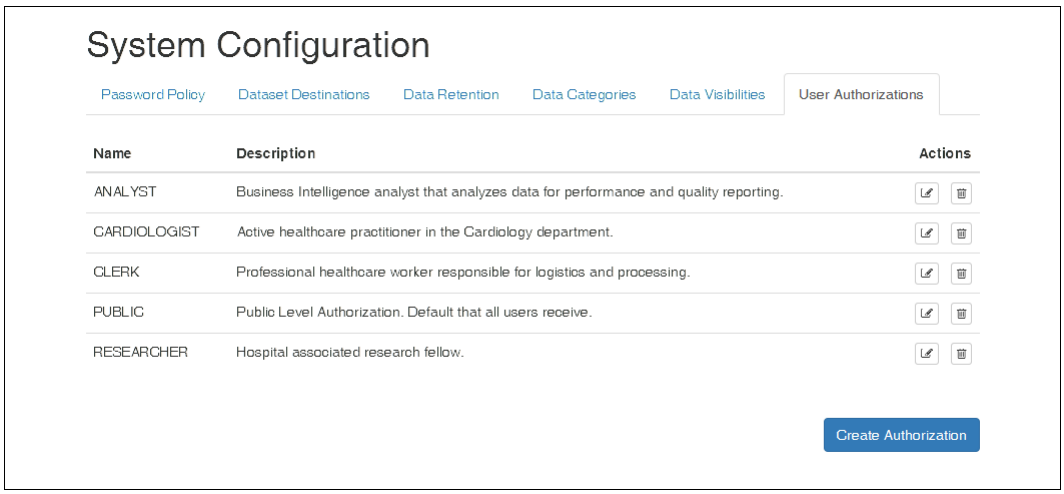
To delete a user authorization:

1. Open the **System Configuration** page, by clicking the **System Configuration** icon in the left navigation bar. The **System Configuration** page opens on the **Password Policy** screen.




2. Click the **User Authorizations** tab.

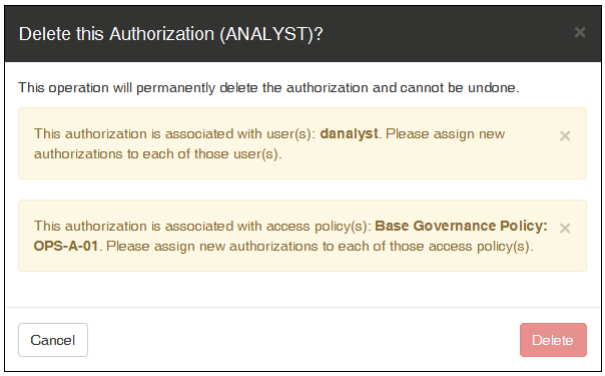
   The **User Authorizations** screen opens.




3. Locate the authorization you want to delete and click the trash can icon in the **Actions** column.

   The **Delete this Authorization?** dialog opens.

   Note that user configurations and access policies both depend on user authorizations. The system checks for dependencies before it deletes an authorization, and if one of your users or one of your access policies refers to the authorization you want to delete, the system will warn you, listing the user(s) and access policy or policies that you need to modify. In this case, the **Delete** button will be unavailable (greyed out). You must eliminate all dependencies before the system will permit you to delete the authorization. When all dependencies have been eliminated, the **Delete** button will become available.




4. When all dependencies have been eliminated, the **Delete** button will become available. Click **Delete** to delete the authorization.

# Glossary of Terms and Concepts

**access policy**

An access policy is a set of rules that specifies how users can consume data stored in PHEMI Central. The access policy lists what user authorizations are required to interact with data tagged with specified visibility. Access policies can be applied to data collections and datasets.

**authorizations**

User authorizations are configurable attributes you can assign to PHEMI Central users. Authorizations are defined in PHEMI Central by the PHEMI Administrator, who sets them in accordance with the organization's governance policies.

**field**

A field is the smallest unit of data storage in PHEMI Central. A field is a single data item, which can range from a single byte up to gigabytes, plus the metadata associated with the data item. Any piece of raw data, regardless of size, is stored in a single field. Elements of derived data (transformed from the raw data) are also each stored individually in fields. Any field can be protected by applying data visibilities. For derived data, each derived item can be individually assigned a visibility (which may be different than that configured for the data collection) by the DPF performing the processing.

**code library**

A code library is a package of executable code that is included in a DPF archive. Whether the code is source or compiled depends on the coding language. Code libraries must be portable and self-contained; that is, all dependencies required for the DPF to function must be bundled inside the library, in the appropriate way, for whatever language is being used.

**data category**

Data categories are a way to classify data into broader groupings. Examples of data categories are "Research Reports," "X-Rays," and "Prescriptions."

**data collection**

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

**Data Processing Function, DPF**

A Data Processing Function, or DPF, is an executable piece of code that supplies the instructions for processing raw data to extract 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 selectively 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.

**data visibilities**

See visibilities.

**dataset**

A dataset is a view, or map, of an underlying set of data. Data items in a dataset can be selected from across multiple data collections. The dataset is a view, or map, to the underlying data. The actual content of the dataset (that is, the dataset's data) is generated when the dataset is executed or when it is queried against.

**derived data**

Derived data is data that has been parsed, extracted, or otherwise enriched or processed by running a DPF on stored raw data. The set of derived data items can be searched, queried, further processed, or exported from the system.

**digital asset**

A digital asset is any piece of data stored with metadata in the system. This may be raw data that has had metadata applied during ingestion, or it may be derived data that has been parsed, indexed, catalogued, and/or enriched with additional metadata.

**DPF archive**

The set of code that makes up a DPF is called a DPF archive. 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 collection, the DPF archive is ``registered`` with the data collection by uploading the DPF archive. Registering the DPF is part of data collection configuration.

**ETL**

Extract, transform, and load. In databases, a set of tools or processes that extracts data from sources, transforms the format or structure for storage, query, and analysis, and loads it into the receiving or consuming system.

**ingestion**

Ingestion is the process by which data is brought into in PHEMI Central. Often, the sending system submits the data to PHEMI Central, which listens for the data, often using a web service. Data can also be ingested manually, by using the PHEMI Central Management and Governance Console.

**JSON**

JSON stands for JavaScript Object Notation. JSON is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate. JSON is used in the body of several REST requests in the PHEMI RESTful API. PHEMI Central also includes a system DPF that can create derived data from JSON objects, providing the objects conform to PHEMI's JSON specification.

**key-value pairs**

A key-value pair is a set of two linked data items: a key which uniquely identifies some item of data, and the value, which is the data itself. PHEMI Central uses key-value store to efficiently store, process, and retrieve data.

**M2M**

M2M is a way of referring to machine-to-machine interfaces, used in machine-to-machine communication.

**manifest file**

A manifest file is a JSON file that specifies the output of a DPF. With the code library, the manifest file makes up the DPF archive that is uploaded to register the DPF with a data collection. The manifest file should include the properties of the DPF along with the details of each derived data item to be generated.

**metadata**

Metadata is information about a piece of data. In PHEMI Central, metadata is information about how a given piece of data is to be managed. When a piece of raw data is ingested into PHEMI Central, information from the connection (for example, the timestamp) together with policy information configured for the data collection (for example, the data visibility) and some derived information (for example, a "time to live," as derived from the timestamp and the data retention policy) is used to create metadata properties that are stored with the data. Further, PHEMI Central also automatically indexes and catalogues all stored data, whether raw or derived; the indexes and catalogues can also be considered a kind of metadata.

**PII**

Personally Identifiable Information, or PII, is a legal concept used in US privacy law and information security to mean information that can be used on its own or with other information to identify, contact, or locate a single person or to identify an individual in context. When thinking about PII, it is important to distinguish legal

requirements to remove attributes uniquely identify an individual from a general technical ability to identify individuals. Because of the versatility and power of modern re-identification algorithms, together with the amount of information freely available from all sources, the absence of PII data does not guarantee that de-identified data cannot be used, perhaps in combination with other data, to identify individuals.

**raw data**

In PHEMI Central, raw data items are files, objects, records, images, and so on that are submitted for ingestion into the system. Raw data is stored exactly as received, along with the metadata generated for it on ingestion.

**REST, RESTful API**

Representational Statement Transfer (REST) is an architectural style that uses HTTP requests and associated methods (POST, PUT, GET, and DELETE) to create, update, read, and delete data. A RESTful API is an application programming interface (API) based on REST.

**VCF**

A Variant Call Format (VCF) file is a text file containing tab-separated marker and genotype data. VCF data is used in bioinformatics to store gene sequence variations. As such, VCF files are very large, and a a VCF file can document hundreds, thousands, and even millions of gene sites in a single file.

**visibilities**

All raw data and derived data stored in PHEMI Central can be tagged with labels that provide information about the data's sensitivity. This sensitivity is described in terms of the visibility the data should have to different system users. The visibility tags you define for your data should reflect the visibility the data is intended to have according to your organization's governance policy.