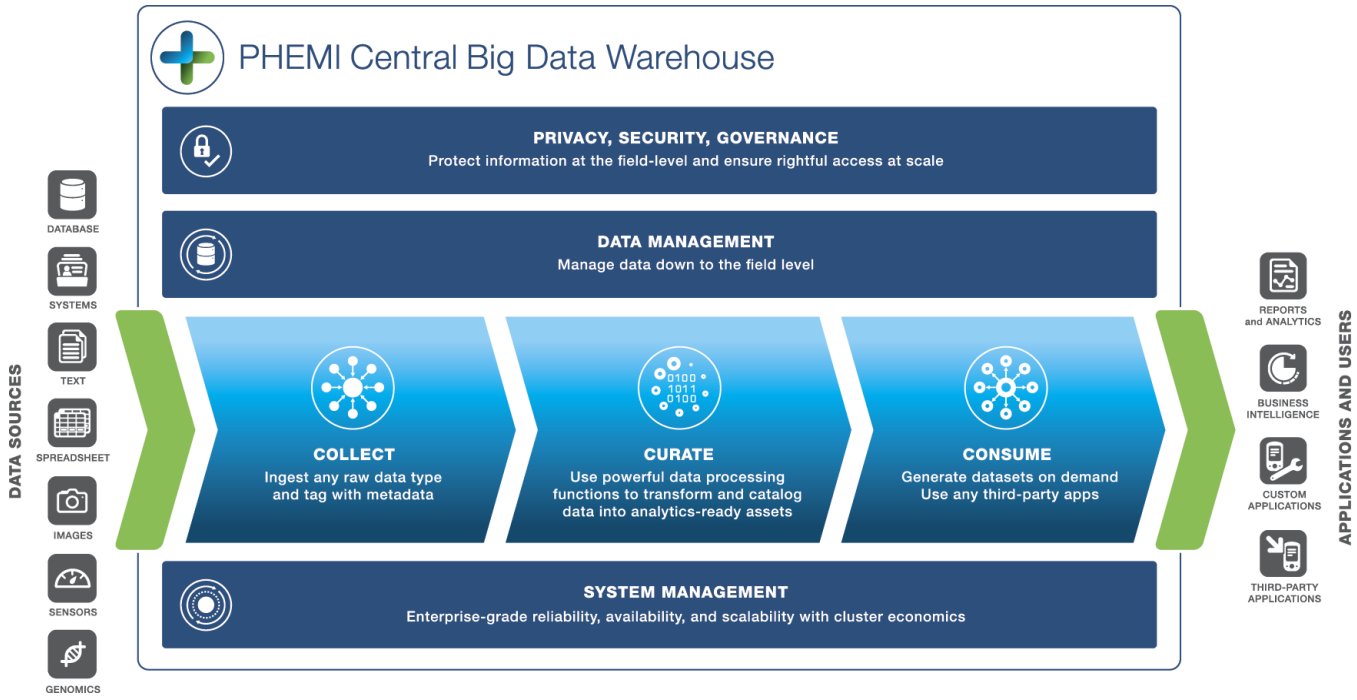
IT Administrator Guide

Contents

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.

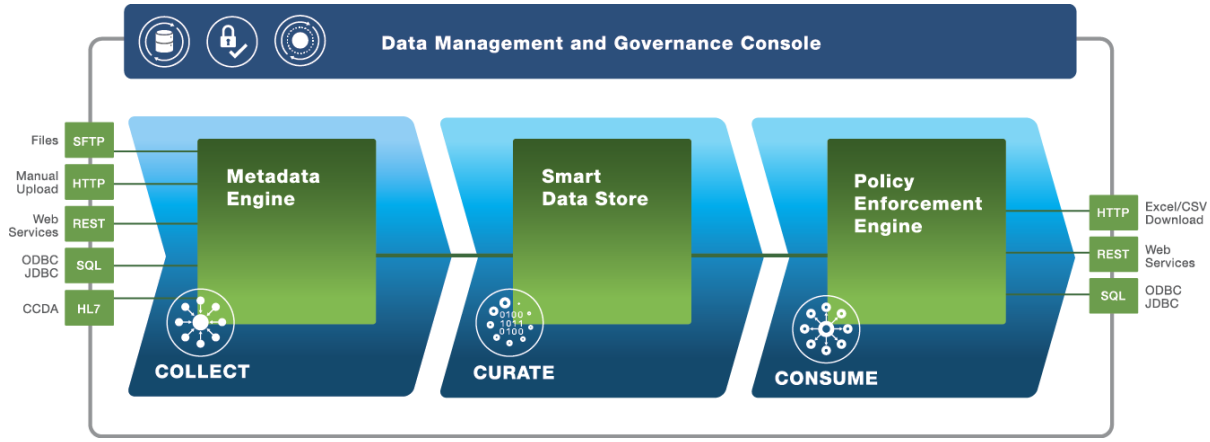
PHEMI Central allows organizations that need to protect and govern the use of their information to take advantage of big data technology to access, catalogue, and analyze their digital assets at speed and scale.



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.

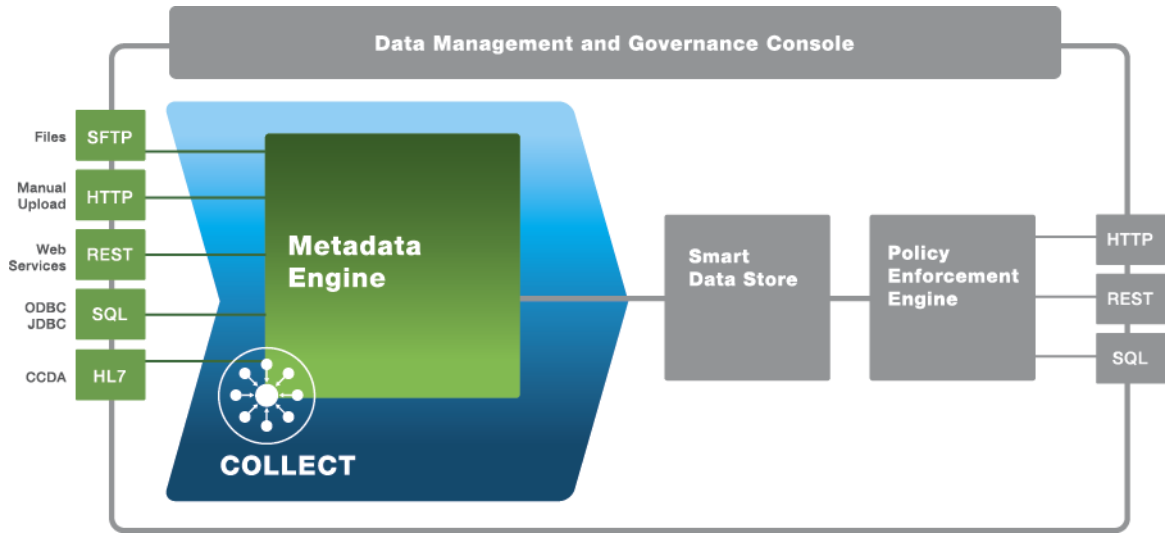


Collection

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:

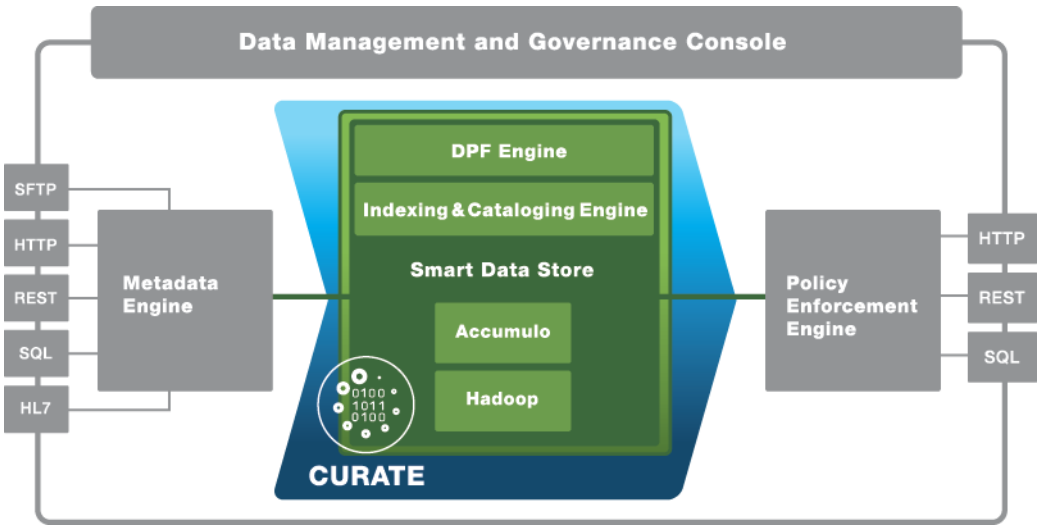
- **Stream**—Machine-to-machine data collections, such as telemetry and hospital bedside monitors, can stream data to PHEMI Central by means of the PHEMI RESTful API.
- **Push**—Data collections and extract, transform, and load (ETL) tools can publish to PHEMI Central using either JDBC or the PHEMI RESTful 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.

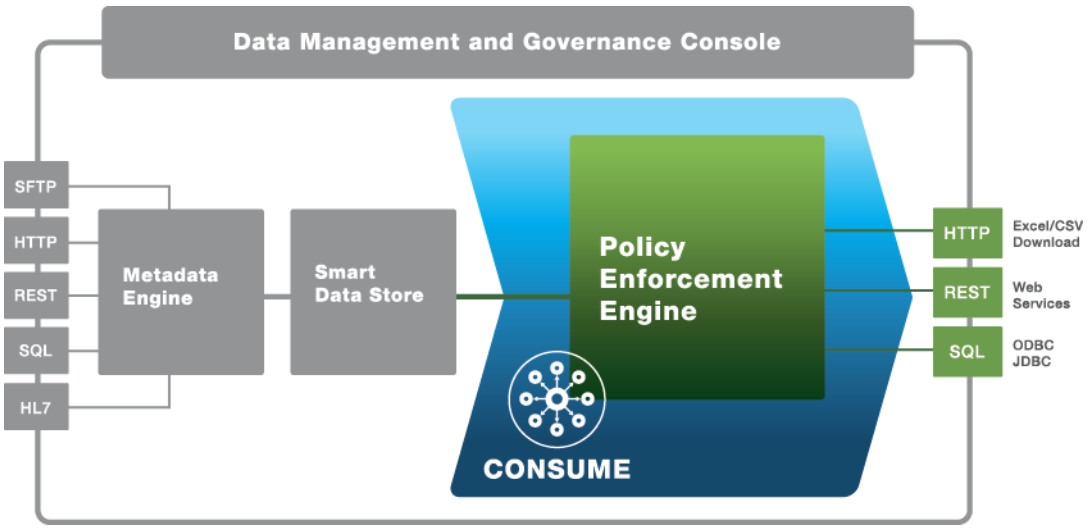
Disparate data collections may have fields that occur in common but are named differently or use different format conventions. For example, one data collection might have a field called "Sex" with values "M" and "F," while another might have a field called "Gender" with values "Male" and "Female." Similarly,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

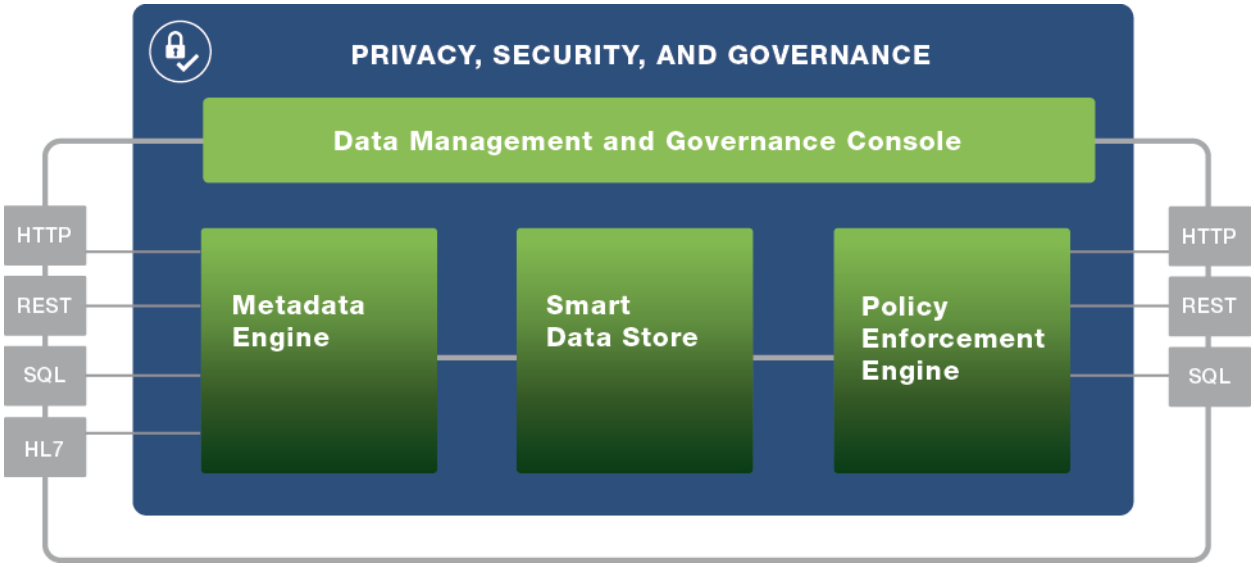


In all cases, PHEMI Central's Policy Enforcement Engine strictly enforces your organization's privacy and security policies to ensure rightful access to data.

Privacy, Security, and Governance

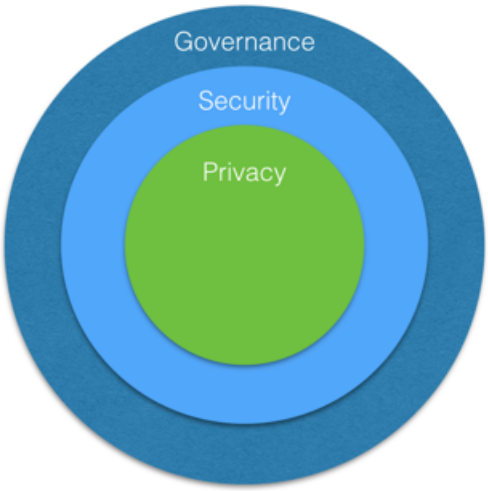
The purpose of privacy, security, and governance is to protect information and ensure rightful access.

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.

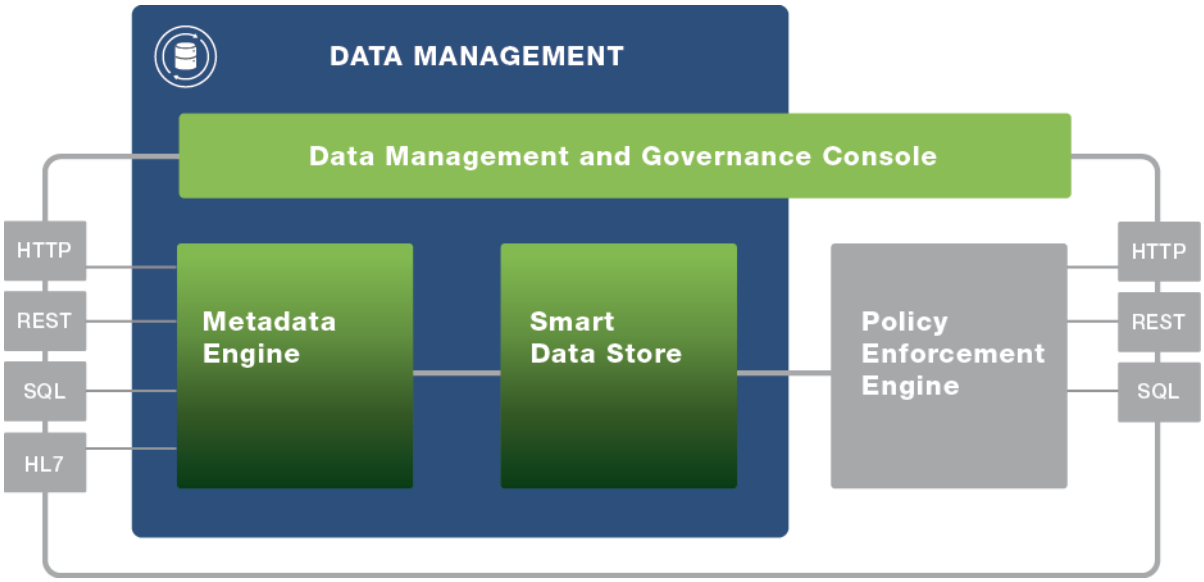
Information governance is about controlling and protecting an organization's data. The data may be sensitive, or perhaps it is important that the data be absolutely accurate, or perhaps the organization must achieve legislative and compliance targets. Data governance includes the process and policies 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.

Retention rules are captured in the Management and Governance Console, and from the rentention rules together with the ingestion timestamp, the system calculates a time to live for every data element. Retention rules also prevent users from deleting data during the configured retention period and helps automatically de-identify, delete, or otherwise process information when the retention period does expire.

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.

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 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 on collection, 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 as 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. The sending system (the data source) submits the data to PHEMI Central, which listens for the data using a web service. Data can also be ingested manually, by using the PHEMI Central Management and Governance Console. The specific characteristics of data ingestion can be specified per data collection as part of the data collection configuration.

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 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, a VCF 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 sensitivity of the data as identified by your organization.