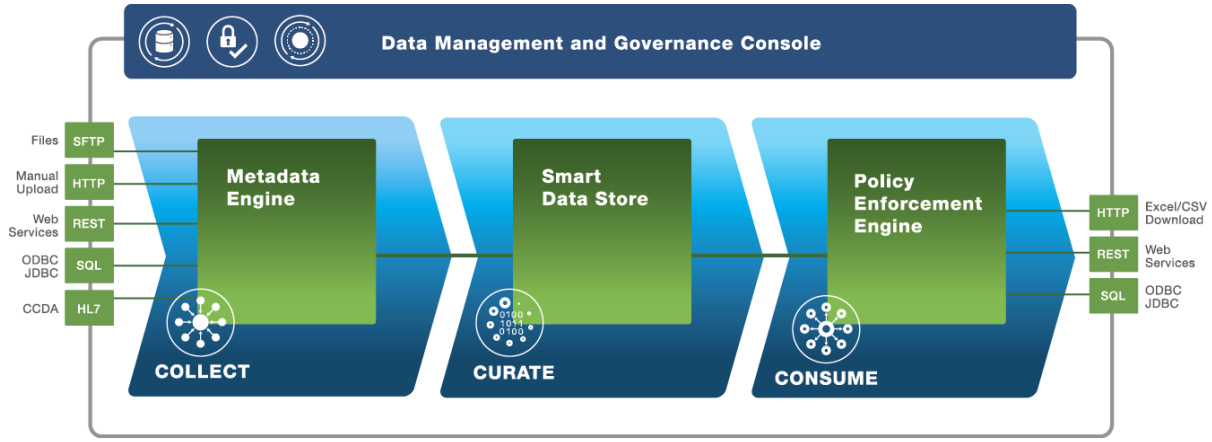
Data in PHEMI Central

Contents

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.

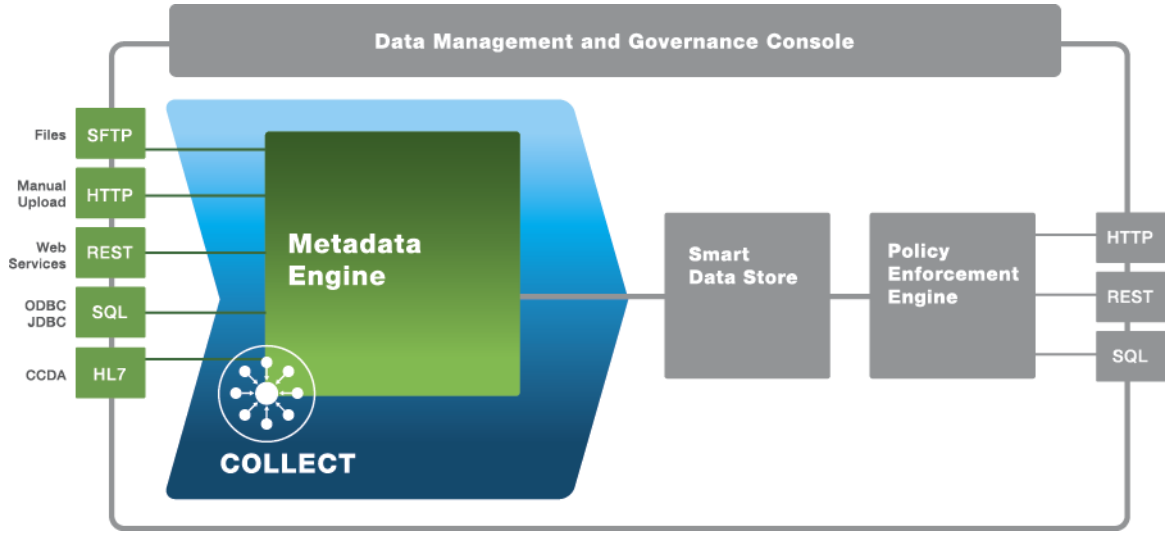


Collection

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, eports, e-mails, customer feedback, social media, and so on.

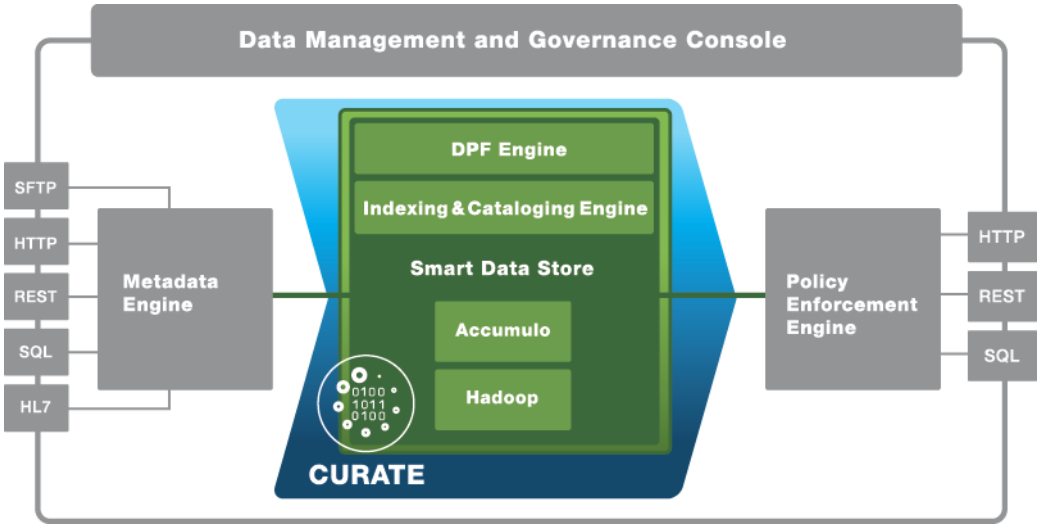


During ingest, PHEMI Central tags each raw data object with metadata that describes the data. Metadata governing digital rights management, retention rules, data sharing agreements, and privacy policies is applied and enforced. Describing information with metadata means that users and applications can query and analyze data based 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.



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 deployed, data must comply with it. If the design does not scale or 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 can offer 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 collection 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 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 log message or medical report) 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 embedding the rules and policies governing the data collection, as well as configured properties such as the data collection ownership, retention policy (time to live), and what visibility the element should have. For example, de-identification, encryption, and masking, along with other privacy restrictions.

Standard PHEMI DPFs libraries are included that index and describe structured data, such as spreadsheet files, database records, or XML/ JSON documents. User-defined DPFs can also 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, calculating aggregates and converting streaming data for trending and predictive analysis. For parsing 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 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 these types or fields.

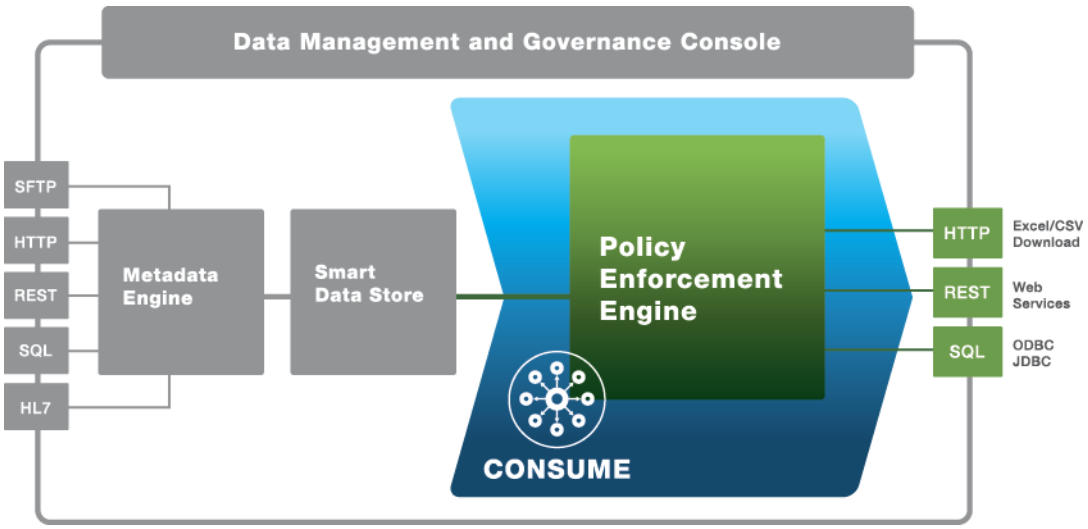
Disparate data collections may have fields that occur in common but are named differently or use different format conventions. For example, one data collection might have a field called "Sex" with values "M" and "F," while another might have a field called "Gender" with values "Male" and "Female." Similarly, different medical imaging can 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. 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. Queries 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 download data from a dataset 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 enforce rightful access to data.