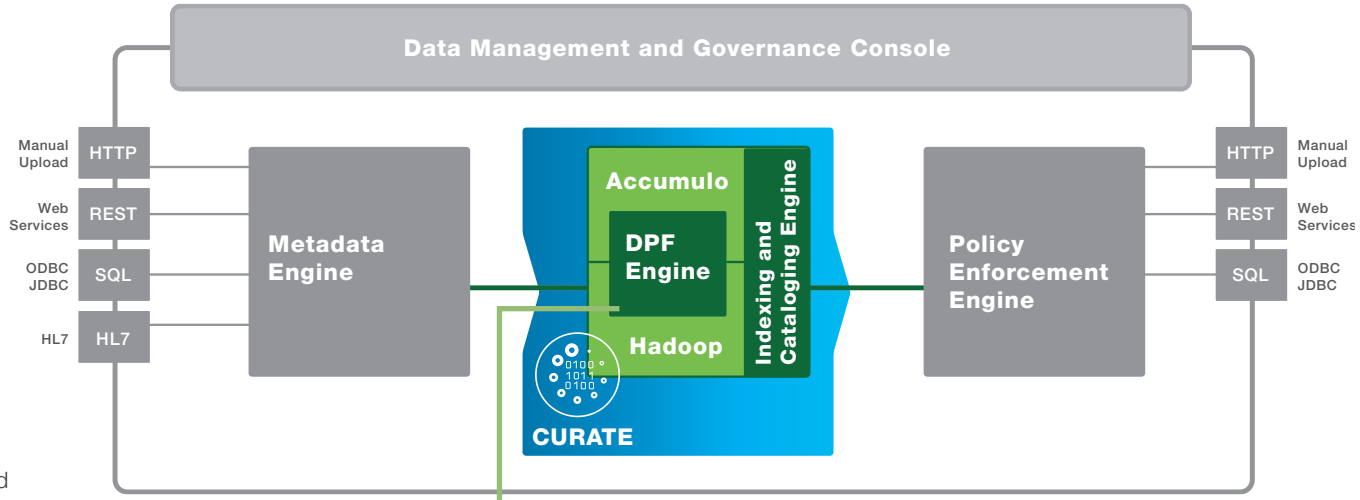
Curating Data

Convert Raw Data into Analytics-Ready Digital Assets

Data Processing Functions

A data processing function (DPF) is an executable piece of code that transforms the original raw data (for example, a log message or medical report) into analytics-ready, governance-compliant digital assets specifically targeted at your organization's needs (such as a temperature reading or blood glucose measurement). The DPF is uploaded as a code archive into PHEMI Central using the Data Management and Governance Console. The code is executed by the PHEMI Central DPF Engine, which uses it to direct curation of the data.



The DPF supplies the instructions for parsing the raw data, extracting key content and performing data cleansing, ontology matching, enhanced indexing and cataloging, and structuring data according to the organization's needs. Standard PHEMI DPFs are included to index and describe structured data, such as database records or strongly typed XML/JSON, for consumption of the data through a REST or SQL interface. 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 data and standardize it 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 the telemetry into an analytics-ready state for trending and predictive analysis. As the organization's needs evolve and as knowledge advances, DPFs can be updated and re-executed, to leverage the value of your historical data in new ways.

The PHEMI standard DPF library includes:

Excel Reader	VCF Reader
Ingested Microsoft Excel spreadsheets and comma-separated value (CSV) files are converted into fine-grained analytics-ready digital assets, with each cell governed by the parent file's data sharing agreement.	Ingested genomic Variant Call Format (VCF) files are converted into a series of analytics-ready variants, with each variant governed by the parent file's data sharing agreement.

The powerful concept of DPFs is unique to PHEMI. DPFs enable data scientists and programmers to write rich, customized transform functions in common programming languages (including Python, Java, and C++) using standard development tools. No specialized expertise in MapReduce or YARN is required. For parsing unstructured documents such as scans or X-rays, the DPF can include specialized parsing functions such as OCR or image parsing. Your DPF can be written by PHEMI, by your organization's in-house programmers, or by third- party developers.

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

Curating Data continued

Indexing and Cataloging

Indexing and cataloging functions occur automatically in PHEMI Central's Indexing and Cataloging Engine, making it easier and faster to find and consume data. User-defined DPFs enable deeper and more sophisticated indexing and cataloguing, and second-order indexes and graph relationships allow data analysts to quickly find and build datasets across petabytes of heterogeneous digital assets. Linking datasets with common keys makes it faster and easier to build meaningful datasets across many sources. These powerful indexing features mean that data can be accessed in milliseconds, without having to wait for MapReduce or YARN jobs to complete.

Schemaless Storage

PHEMI Central's data store is schemaless: both raw and curated data items are stored in a binary format that is unaffected by the source and destination schema. This approach means that organizations can quickly aggregate new data sources without costly redefinition of old schemas. Schemaless storage also permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas. Instead, PHEMI Central uses a flexible, powerful, distributed key-value store and sophisticated metadata tagging to manage, describe, and govern the data it stores. Curated digital assets derived from the raw data are linked to the original raw data, but PHEMI Central's SQL and REST interfaces abstract away from internal linkages and structures, so users and applications can focus on data use rather than data janitorial work.

Powerful indexing features mean that data can be accessed in milliseconds without having to wait for MapReduce or YARN jobs to complete.

Schemaless storage permits the organization to extend uses or imagine new uses for data as knowledge advances and needs evolve, without concern for migrating rigid predefined schemas