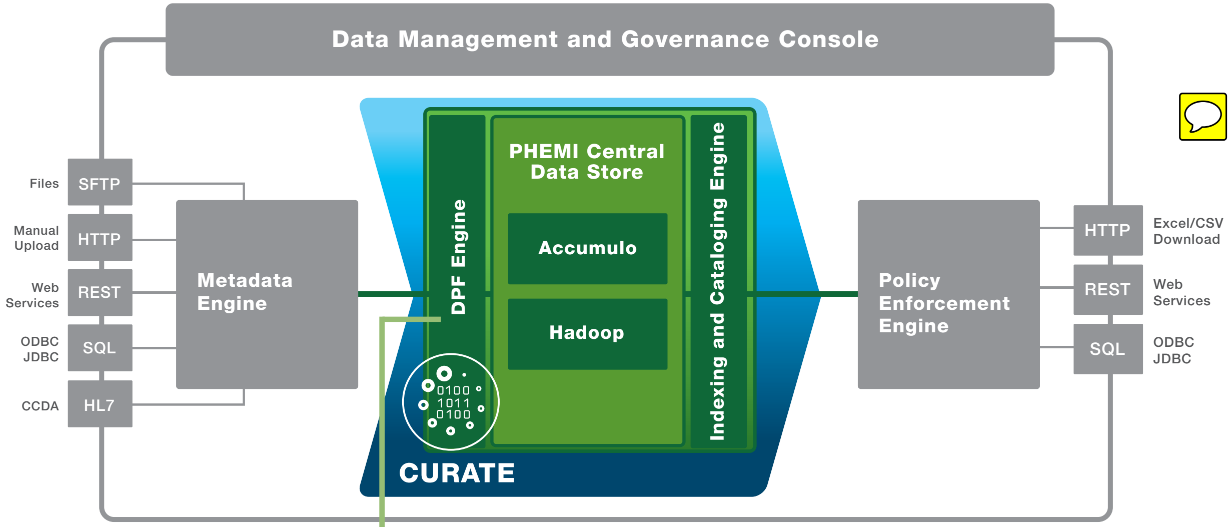
Curating Data

Convert raw data into analytics-ready digital assets.

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, digital assets specifically targeted for 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.



DPFs Provide Power and Flexibility

The DPF supplies the instructions for parsing the raw data, extracting key content and performing data cleansing, 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. 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.

The PHEMI standard DPF library includes:

Excel Reader	VCF Reader	XML Reader	JSON Reader
Ingested Microsoft Excel spreadsheets and comma-separated value (CSV) files are converted into field-level 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.	Icaboreiciis dolore volorer eperiss edictio nsenda incidel esequia sum sum eatqui blanda nobiti. Icaboreiciis dolore volorer eperiss edictio nsenda incidel esequia sum sum eatqui blanda nobiti.	Icaboreiciis dolore volorer eperiss edictio nsenda incidel esequia sum sum eatqui blanda nobiti. Icaboreiciis dolore volorer eperiss edictio nsenda incidel esequia sum sum eatqui blanda nobiti.

DPFs can also analyze streams of machine data to find patterns and exceptions, calculating aggregates and converting streaming data 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.

PHEMI's Unique DPF Concept

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, like Optical Character Recognition (OCR) or image parsing. DPFs can be written by PHEMI, by your organization's in-house programmers, or by third-party developers.

Curating Data continued

With the PHEMI Central Data Store, you can reliably store your digital assets at scale, with powerful features that index, protext, and transform data to be ready for use.

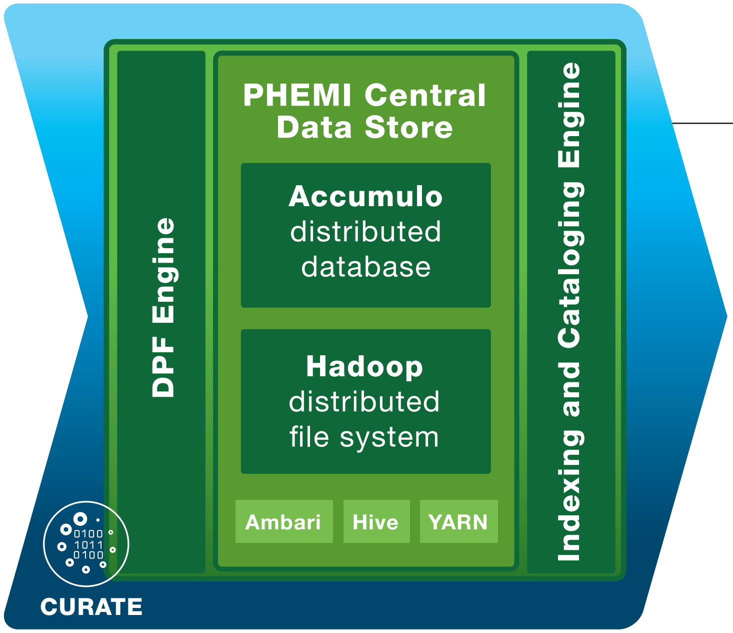
The PHEMI Central Data Store: Beyond Hadoop

PHEMI Central uses well-established and industry-leading big data technologies to reliably store the curated digital assets at scale. PHEMI Central leverages this base "operating system" capability to build powerful features that index, protect, and transform data for business use.

- **Hadoop Distributed File System** (HDFS) provides linear scale and reliable data storage across large cluster of low-cost commodity servers.
- **Accumulo** is a distributed database on top of the HDFS distributed file system. Developed by the NSA, Accumulo provides high-performance storage and retrieval with fine-grained privacy access controls.
- **Ambari** is an open framework to provision, manage and monitor Apache Hadoop clusters.
- **Hive** delivers interactive and batch SQL query capabilities into PHEMI Central in order to interoperate with analytics tools and pre-existing applications.
- **YARN** provides resource management and distributed computing for the PHEMI Central system.

In contrast to conventional Hadoop-based systems, PHEMI Central is a fully integrated enterprise-grade system. Users don't need to worry about digging deep into the world of Hadoop, MapReduce, YARN, Pig, HIVE, Sqoop, HBase, Zookeeper, Accumulo, and so on.

To protect information privacy and security, at no time can users, applications or external systems bypass the PHEMI policy enforcement engine to access data directly.


DPF Engine
PHEMI Central Data Store
Accumulo distributed database
Hadoop distributed file system
Ambari
Hive
YARN
Indexing and Cataloging Engine
0100
1011
0100
CURATE

The PHEMI Central Data Store is central to the process of curating data, providing fully integrated enterprise-grade storage that leverages industry-leading big data technologies to reliably store curated digital assets at scale. Schemaless storage means that organizations have the flexibility to advance quickly and effectively.

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.

> Quickly aggregate new data sources without costly redefinition of old schemas.

Curating Data continued

PHEMI Central comes with a set of powerful built-in features to support data curation. Indexing and cataloging functions are automatic, while user-defined DPFs can extend the built-in capabilities. Data Linking allows a more complete picture of your data. The Data Dictionary allows control of diverse data types, providing consistent interpretation at the point of querying and analysis. Geospatial data capabilities make PHEMI Central a great foundation for geospatial-drive applications and analysis.

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.

Data Linking

PHEMI Central brings together disparate data and data that has been isolated in data silos, so that the organization can extract the greatest value from its information. PHEMI Central links data across different data types and formats with unique identifiers based on powerful graph database capabilities. Data linking allows you to query and analyze a more complete picture of your data, so you can see, at scale and efficiently, relationships between objects in the system.

Data Dictionary

Conventional big data systems store data, but can't catalog or track diverse data types. PHEMI Central provides a powerful data dictionary capability that links data from curation, when data types and fields are identified and tagged, through consumption, when users, tools, and applications query and analyze data. The data dictionary can be used to cleanse data by identifying fields that occur frequently but are named differently, or use different format conventions, across different data sources. For example, different medical imaging systems can use different terminology and conventions for the same concepts and measurements. PHEMI Central allows you to identify and save a common interpretation of these types and fields. Cleansing data with a data dictionary greatly simplifies querying and analysis.

Geospatial Capability

PHEMI Central's efficiently indexes and searches geospatial data, so that organizations can store and analyze data collections rich in geospatial components. PHEMI Central is scalable and fast: a flexible basis on which to build geospatial-driven applications and analysis.

Powerful indexing features allow data to 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.