PHEMI

# PHEMI Central™ Big Data Warehouse

## Product Description

# PHEMI Central Big Data Warehouse —collect, curate, and consume data with privacy, security, and governance.

For the first time, organizations that need to protect and govern the use of their information can take advantage of big data technology to access, catalog, and analyze their digital assets at speed and scale.

PHEMI

# Table of Contents

# Beyond the “Data Lake”

## Introducing a powerful new way to manage your organization’s data — collect, curate, and consume data at speed and scale.

PHEMI Central is a fully integrated Big Data Warehouse that takes advantage of the power, scalability, and flexibility of Hadoop while providing advanced privacy, security, and governance — all built right in.

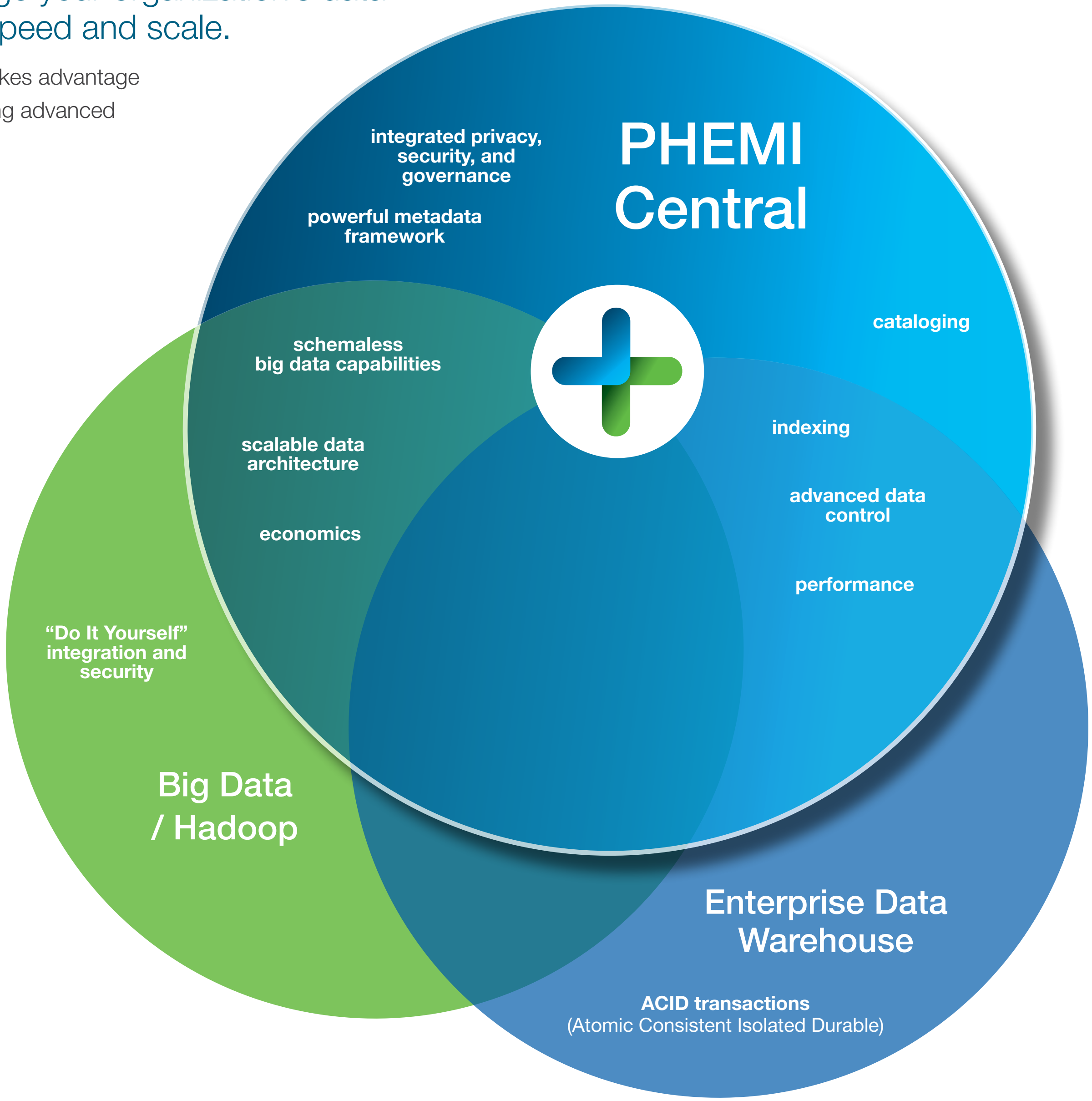
### Beyond the “Data Lake”

Many big data approaches today just pour data in its native format into a data repository, with little oversight or governance. The resulting “data lake” often quickly becomes a “data swamp,” with data that’s almost impossible to protect, control, find, and retrieve.

PHEMI Central’s powerful metadata framework automatically indexes and catalogs all of your digital assets, so that you can find them quickly and easily — making sense of the data lake even as it adds protections to ensure rightful access.

### PHEMI Central — Making Big Data Product-Ready

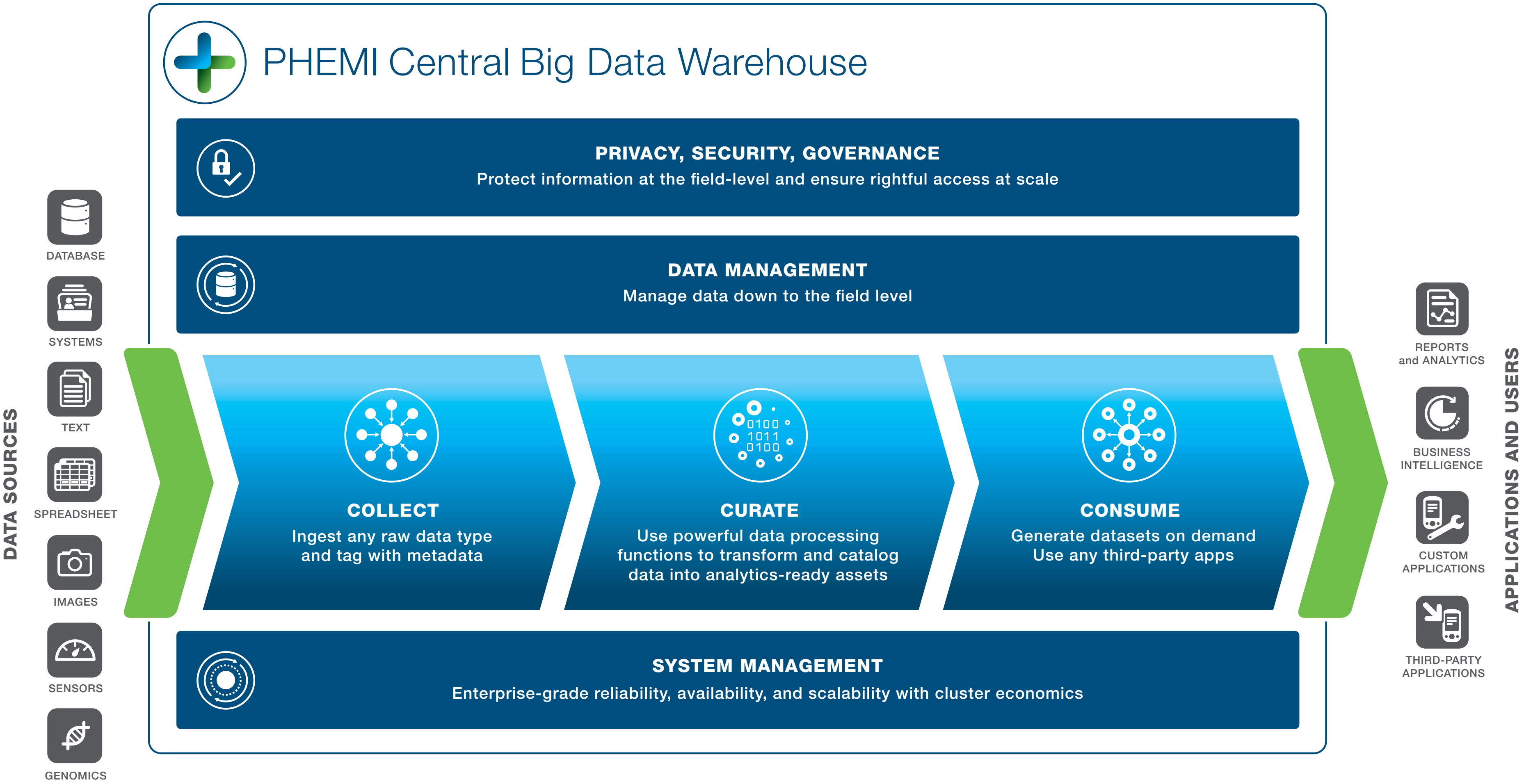
Other big data platforms use Hadoop’s distributed file system capabilities, but PHEMI Central goes beyond “plain old Hadoop,” providing a fully integrated solution that leverages the Hadoop ecosystem but adds enterprise-grade data lifecycle and data management features. PHEMI takes the best of big data technologies and traditional enterprise warehousing systems — then outstrips them both to offer a robust yet flexible solution that can help organizations get control of and get the value from their data.


PHEMI Central
integrated privacy, security, and governance
powerful metadata framework
cataloging
schemaless big data capabilities
scalable data architecture
economics
indexing
advanced data control
performance
“Do It Yourself” integration and security
Big Data / Hadoop
Enterprise Data Warehouse
ACID transactions
(Atomic Consistent Isolated Durable)

# Introducing PHEMI Central Big Data Warehouse

## Collect, curate, and consume your data with privacy, security, and governance.

PHEMI Central is a new class of data warehouse that uses big data technologies to allow your organization to handle any volume and variety of data, while meeting your organization's standards for data management, privacy, and governance.


PHEMI Central Big Data Warehouse
PRIVACY, SECURITY, GOVERNANCE
Protect information at the field-level and ensure rightful access at scale
DATA MANAGEMENT
Manage data down to the field level
COLLECT
Ingest any raw data type and tag with metadata
CURATE
Use powerful data processing functions to transform and catalog data into analytics-ready assets
CONSUME
Generate datasets on demand
Use any third-party apps
SYSTEM MANAGEMENT
Enterprise-grade reliability, availability, and scalability with cluster economics
DATA SOURCES
DATABASE
SYSTEMS
TEXT
SPREADSHEET
IMAGES
SENSORS
GENOMICS
APPLICATIONS AND USERS
REPORTS and ANALYTICS
BUSINESS INTELLIGENCE
CUSTOM APPLICATIONS
THIRD-PARTY APPLICATIONS

# PHEMI Central Functionality

## COLLECT

Ingest all data types at record-setting speeds

PHEMI Central can collect any kind of data—structured (such as database records), semi-structured (such as Microsoft Excel, machine-collected data, or genomic files), or unstructured (such as images or documents). Data can be ingested and aggregated from multiple disparate sources. During collection, PHEMI Central tags each raw data object with metadata, then stores the tagged data in PHEMI's fast, powerful Smart Data Store. The Smart Data Store is key-value–based and schemaless, so data can be ingested at sub-second rates without complex, time-consuming, and brittle schema-mapping exercises.

## CURATE

Find assets at sub-second speeds even with petabytes of data

The PHEMI Central Big Data Warehouse automatically indexes and catalogs all ingested data. Next, user-specifiable Data Processing Functions cleanse, parse, and structure the data—transforming the tagged raw data into analytics-ready digital assets. All data is cataloged and indexed for linking based on key words, graph relationships, and geospatial attributes. Aggregates are computed to accelerate anaytics and application performance. This lets users find specific digital assets at sub-second speed across petabytes of data.

## CONSUME

Access your datasets on demand

With PHEMI Central, you can build datasets on demand without having to manage multiple data marts or complex MapReduce or YARN processes. Whether the dataset is consumed by a user or exported to spreadsheet, application, or analytics tool, PHEMI Central strictly enforces your organization's privacy and security policies to ensure appropriate access to data.

## Privacy, Security, and Governance

Automatically de-identify, encrypt, or mask personal information

PHEMI Central can automatically de-identify, encrypt, or mask personal information and enforce privacy based on sophisticated user access privileges and fine-grained sharing and consent rules. With a Privacy by Design framework at its core, PHEMI Central helps you achieve your organization's governance objectives.

## Data Management

Use a powerful metadata framework to manage digital assets at the field level

PHEMI Central incorporates the advanced data management features of enterprise-grade traditional data warehouses. In addition, PHEMI Central's metadata framework allows organizations to manage data across the entire system at the field level, throughout its lifecycle.

## System Management

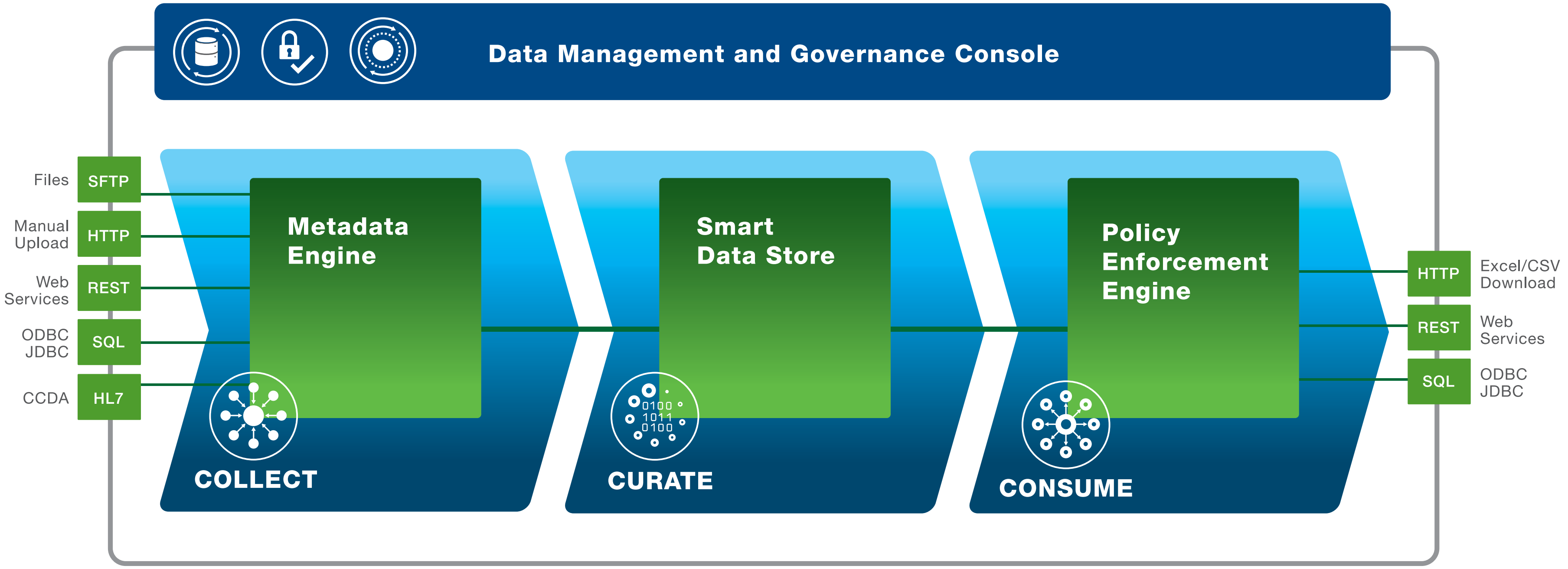
Get cluster reliability and economics at scale

PHEMI Central runs on commercial servers and commodity disk drives, driving down hardware costs and allowing the system to scale from terabytes to petabytes without expensive Storage Area Network (SAN)/Network Attached Storage (NAS) costs or performance bottlenecks. Automatic replication and load balancing means data is always available and performance is optimized across system nodes.

# PHEMI Central Big Data Warehouse is fully integrated and fully adaptable

Designed to adapt to constantly evolving organizational demands, PHEMI Central unlocks data silos and transforms structured and unstructured data into analytics-ready digital assets for users to gain better insights, faster. With increased agility in data collection and increased capability in data inventory and curation, organizations can conceive new applications and rapidly build new solutions to support organizational objectives.

# PHEMI Central System Architecture
## —a fully integrated data management system



- Collect any type of data.
- Curate raw data into analytics-ready digital assets at speed and scale.
- Consume digital assets and build on existing analytics tools and software to create new applications.
- Manage, protect, and govern data easily.
- Get enterprise-grade reliability, availability, and scalability—with cluster economics.

# Collecting Data

## Easily ingest any type of data.

With PHEMI Central, organizations can consolidate their data and eliminate data silos.

### Data Types

Data sources can include any data type from small kilobyte messages to large terabyte files:

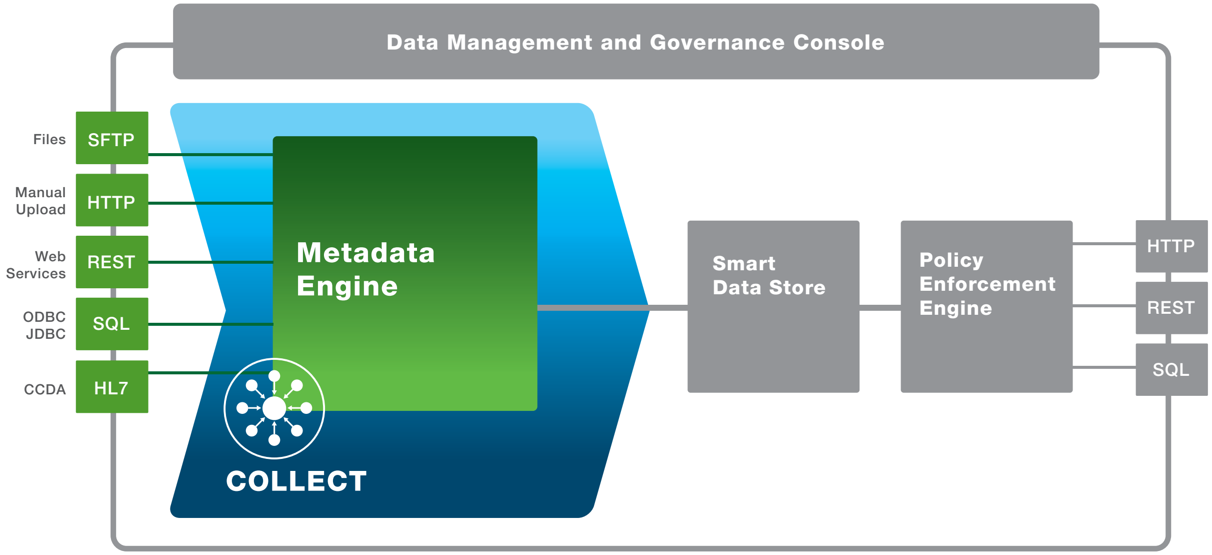
- **Database records** — Data extracted from information systems, databases, etc.
- **Structured non-relational data** — Spreadsheets, GIS datasets, genomics, machine data, XML, JSON, HL7, etc.
- **Semi-structured files** — ECGs, tabular documents, etc.
- **Unstructured files and datasets** — Images,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

Because PHEMI Central does not impose a schema on source systems, the ingest process is faster, less complex, and less brittle.

### Data Import

PHEMI Central can import data in a variety of ways:

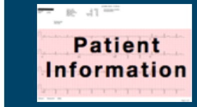
- **Streaming** — Machine-to-machine data sources, such as telemetry and hospital bedside monitors, can stream data to PHEMI Central by means of the PHEMI REST interface.
- **Push** — Data sources and ETL tools can publish to PHEMI Central using either JDBC or the PHEMI REST interface.
- **Pull** — Custom connectors based on PHEMI's REST interface can be deployed to allow PHEMI Central to fetch data from sources.
- **Manual** — 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 REST API. Applications can also be stored and executed, causing external tables or external data to be accessed and pre-processed. Store by Reference and Store by Action operations are ideal for collaboration projects between organizations or when accessing third-party datasets where data sharing agreements restrict data from being replicated.



### Metadata Tagging on Ingest

Immediately on ingest, PHEMI Central catalogs and describes raw data as part of system-wide data management. Metadata governing digital rights management, retention rules, data sharing agreements, and privacy policies are applied and enforced. Metadata fields are fully extensible.

**Digital Asset = Information + Metadata**



+

**Metadata**

Administration
Source, owner…

Structure
Indexes, relationship…

Rights
DRM, access, privacy…

Preservation
Retention policy…

Description
Type, definition…

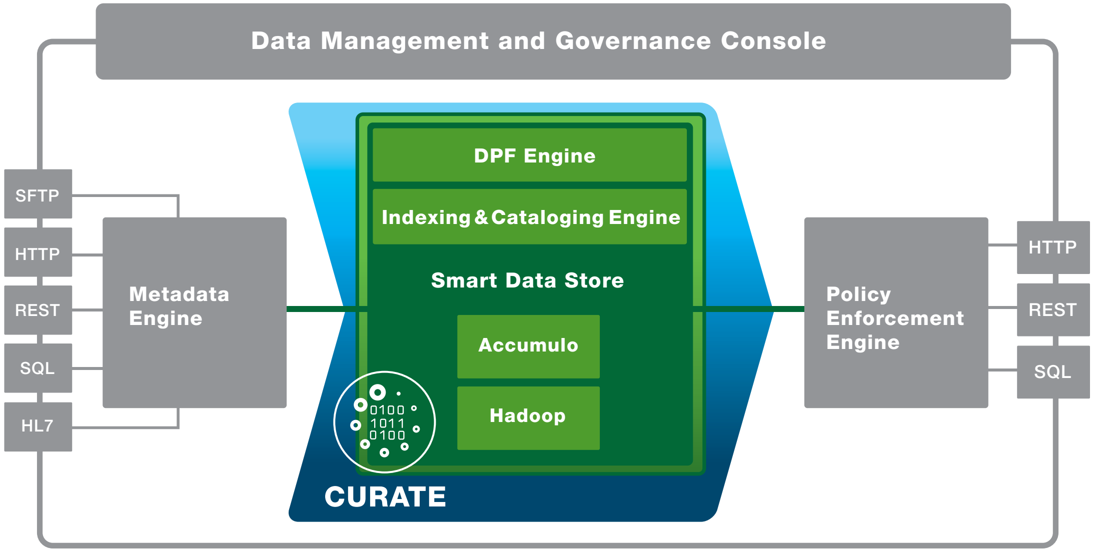
Quality
Confidence…

> Powerful metadata tagging features convert your data into digital assets, making governance and privacy rules easy to apply and enforce across the entire system. Metadata fields are fully extensible.

# Curating Data

## Convert raw data into analytics-ready digital assets.

A Data Processing Function (DPF) is an executable piece of code, written in any modern programming language, 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.



### Data Processing Functions — 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 libraries are included to 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.

**The PHEMI standard DPF library includes:**

| Excel Reader | VCF 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. |

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. 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 and re-executed, to leverage the value of your historical data in new ways.

### PHEMI's Unique DPF Framework

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. DPFs can be written by PHEMI, by your organization's in-house programmers, or by third-party developers.

**PHEMI's innovative DPF framework** enables data scientists and programmers to write rich, customized transform functions in common programming languages using standardized tools. No specialized expertise in MapReduce or YARN is required. Your DPFs can be written by PHEMI, by your organization's in-house programmers, or by third-party developers.

# Curating Data

## Store your digital assets at scale with fully integrated features that index, protect, and transform data to be ready for use.

### PHEMI Central: Built on Hadoop

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

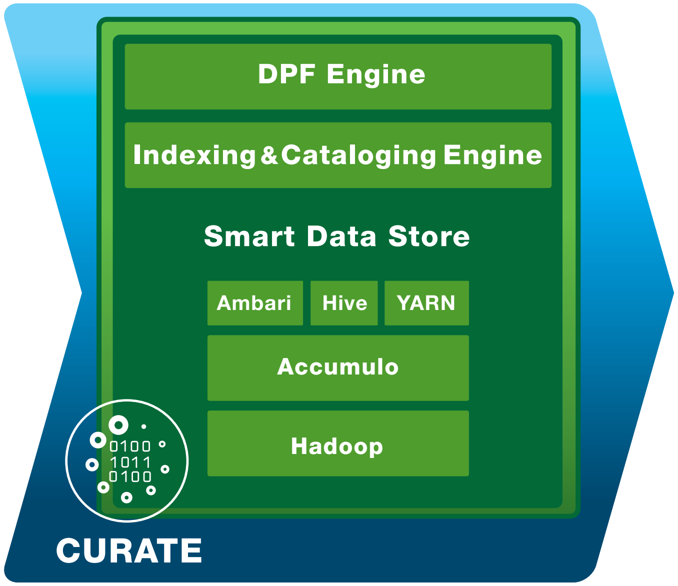
- **Hadoop Distributed File System** (HDFS) provides linear scale and reliable data storage across large cluster of low-cost commodity servers.
- **Accumulo** is a distributed database on top of the HDFS distributed file system. Accumulo provides high-performance storage and retrieval with fine-grained 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 and developers don’t need to worry about digging deep to understand or integrate Hadoop, MapReduce, YARN, Pig, HIVE, Sqoop, HBase, Zookeeper, Accumulo, and so on — PHEMI has already integrated them in PHEMI Central.

### Schemaless Storage

PHEMI Central is schemaless: both raw and curated data items are stored in a binary format that is unaffected by the source or destination schema. This approach means consumers can work without even knowing the source schema, and organizations can quickly aggregate new data sources without costly schema redefinition.

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 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 invisibly manages internal linkages and structures, so users and applications can focus on data use rather than data janitorial work.



PHEMI Central provides enterprise-grade storage of curated digital assets at scale.

With schemaless storage, organizations can extend uses or imagine new uses for data, without concern for migrating rigid or predefined schema.

# Curating Data

## PHEMI Central comes with a set of automatic features to support data curation.

Indexed and cataloged distributed data structures are standard with PHEMI Central, freeing developers from having to create their own indexing and catologing structures. Data linking allows a more complete picture of your data. The data dictionary helps control diverse data types and provide consistent interpretation for queries and analysis. Geospatial data capabilities make PHEMI Central a great foundation for geospatial applications.

### 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 cataloging, while 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 sub-second time, 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 silos, so that the organization can extract the most 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 struggle to 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 across different data sources to cleanse data by identifying fields that occur frequently but are named differently or use different format conventions. 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 query and analysis.

### Geospatial Capability

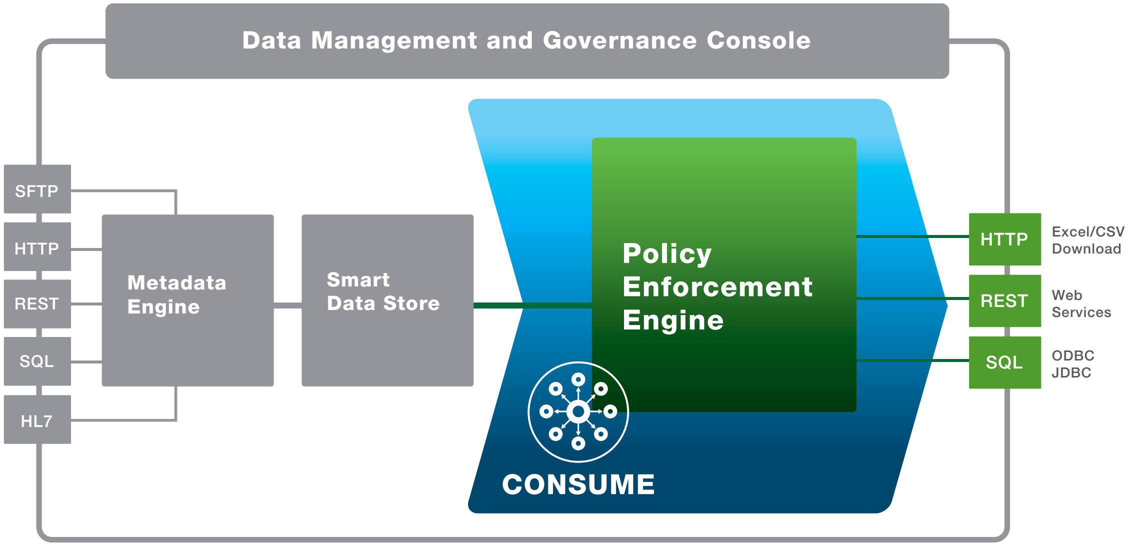
PHEMI Central efficiently indexes and searches geospatial data, so that organizations can store and analyze data collections rich in geospatial components. With its speed and scalability, PHEMI Central is a great foundation for geospatial applications and analysis.

Automatic indexing allows sub-second access to data — no waiting for MapReduce or YARN jobs to complete.

# Consuming Data

## Leverage your existing analytics tools and software to build innovative new applications.

PHEMI Central integrates with your existing infrastructure, applications, and analytics tools to let you immediately take advantage of the big data warehouse. Multiple users can concurrently interact with the system, accessing datasets via SQL, data exports, and custom applications, breaking down the costly data silos spread throughout your organization. Because all digital assets are cataloged and indexed, consuming data is fast. And because the PHEMI Central Policy Enforcement Engine mediates all requests for data, data remains governance-compliant at all times when consumed by or shared with any user, application, or tool.



### Dataset Access

Datasets can be accessed in a variety of ways:

- Export to a Microsoft Excel spreadsheet or to a Comma Separated Value (CSV) file
- Custom application via the PHEMI REST API
- Applications and analytics tools via ODBC/JDBC connectors
- SAP HANA via the SAP Smart Data Access connector

### On-Demand Datasets

Datasets are instantiated when information is needed, without having to predefine or navigate a complex relational database schema. Datasets can bring together any subset of the digital assets in the system, regardless of the data source. Datasets can be created, altered, and discarded and, since they are virtual constructs, they eliminate any need for data marts. Datasets can show, de-identify, or hide data based on attributes of the user and data sharing agreement policies.

### Data Analysis

PHEMI Central supports standard tools, including R, SAP, SAS, SPSS, Stata, Tableau and more, to let organizations leverage their existing software.

### SQL Support

SQL remains the primary method for analyzing and accessing data in many data warehouse architectures. PHEMI Central supports batch and interactive SQL queries.

### Custom Applications

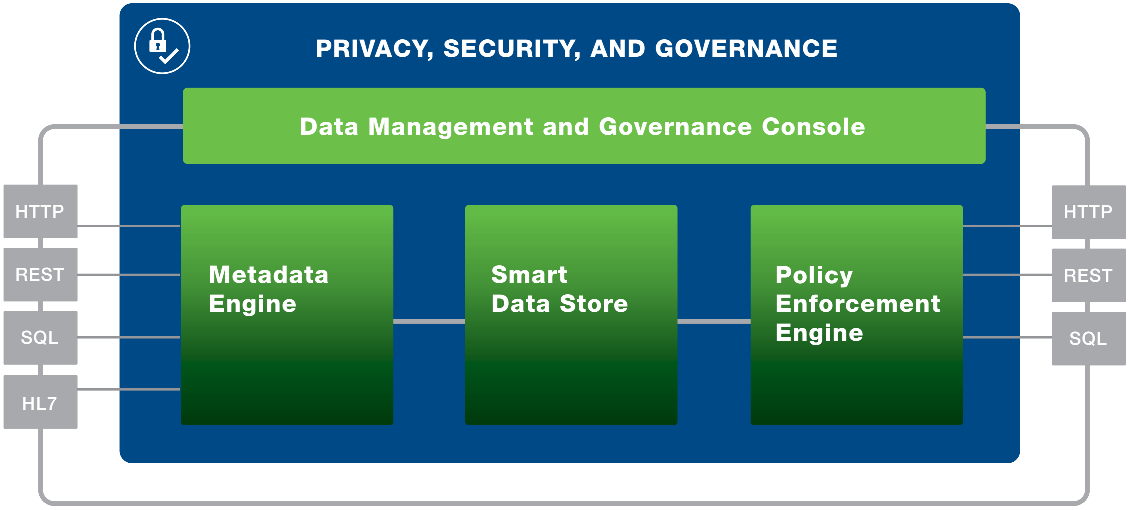
Developers can write custom applications using the PHEMI REST web services interface. The REST interface provides a rich set of statements that can access core system functionality, including flexible queries for data. REST queries, like all queries, are subject to the access controls enforcing information privacy, security, and governance. Organizations can quickly trial innovative new applications, putting information in the hands of their users. This development agility means that applications that gain traction can be hardened and expanded while others are retired.

> PHEMI Central integrates with existing infrastructure, applications, and analytics tools to let you immediately take advantage of the Big Data Warehouse.

# Privacy, Security, and Governance

## Protect sensitive information at scale.

Information governance is the process and policies around the protection, curation, and access to data. The data may be sensitive, or perhaps it is important that the data be absolutely accurate, or perhaps the organization must meet legislative and compliance targets. Data governance encompasses all of privacy protection, data security, and data audit. PHEMI Central helps organizations achieve compliance objectives by providing an industry-pioneering set of capabilities to manage the governance of data. These capabilities are fully configurable and are automatically enforced throughout the data lifecycle.


PRIVACY, SECURITY, AND GOVERNANCE
Data Management and Governance Console
Metadata Engine
Smart Data Store
Policy Enforcement Engine
HTTP
REST
SQL
HL7
HTTP
REST
SQL

## Privacy is Built Right into PHEMI Central

### Privacy by Design Principles

- Proactive, not reactive
- Privacy as default setting
- Privacy embedded in design
- Full functionality — positive sum, not zero sum
- End-to-end security — full life-cycle protection
- Visibility and transparency — keep it open
- Respect for user privacy — keep it user-centric

### PHEMI Central Implementation

- Metadata enables policies to constrain data access
- Assets are immutable
- Metadata and computational access are the core of the system
- Data governance policies enable data use/analysis and do not create restrictions
- Digital assets self-specify how they are managed and handled
- Metadata and auditing provide accountability
- Data steward defines and sets policies on use

### PHEMI Design Innovation

- **By default, all data is inaccessible, and access is only opened by establishing access policies**
- **Data firewalls protect data internally, not just externally**
- **Reliance is on automated operational policies, instead of manual processes**
- **Proper management and control enables positive use of private data**

**Privacy by Design**
Traditional data governance approaches are fragmented, relying on application developers to implement ad hoc security mechanisms. PHEMI Central takes a different approach, designed from the ground up to incorporate an innovative Privacy by Design framework to define, manage, and enforce data sharing agreements and privacy policies across an entire organization.

Because privacy, security, and governance features are one coordinated design across the system, you don't have to rely on a cobbled-together solution of security mechanisms to protect your organization's sensitive data.

**Using *Privacy by Design* to Achieve Big Data Innovation Without Compromising Privacy**

# Privacy, Security, and Governance

## Build your access policies quickly and easily.

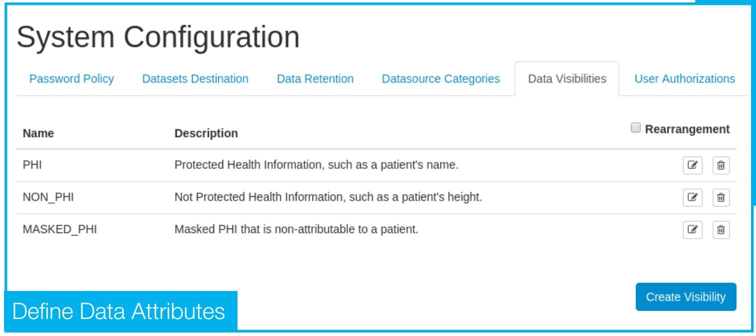
PHEMI Central tags sensitive data to identify its visibility, captures user attributes, and combines them in simple, powerful access rules for attribute based access control. PHEMI Central also manages and enforces data sharing agreements and patient consent.

### Attribute Based Access Control

Users are tagged with attributes that describe their level of authorization. Data is tagged with attributes that describe its level of sensitivity or its requirements for privacy. Together, these two attributes can be combined to allow sophisticated access privileges to data. For example, a data analyst with Level 1 clearance might be able to export fully identified  data, an analyst with Level 2 clearance might only have access to de-identified data, and a Level 3 analyst might have view-only privileges with no export.

### Role Based Access Control

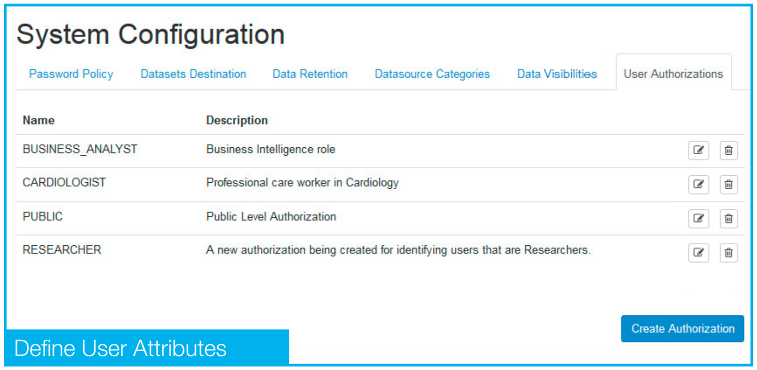
User roles determine what operations a user can perform. For example, only users with a role of administrator can configure the system, while only users with a role of data analyst can query data.



System Configuration

| Name | Description |
|---|---|
| PHI | Protected Health Information, such as a patient's name. |
| NON_PHI | Not Protected Health Information, such as a patient's height. |
| MASKED_PHI | Masked PHI that is non-attributable to a patient. |

Define Data Attributes



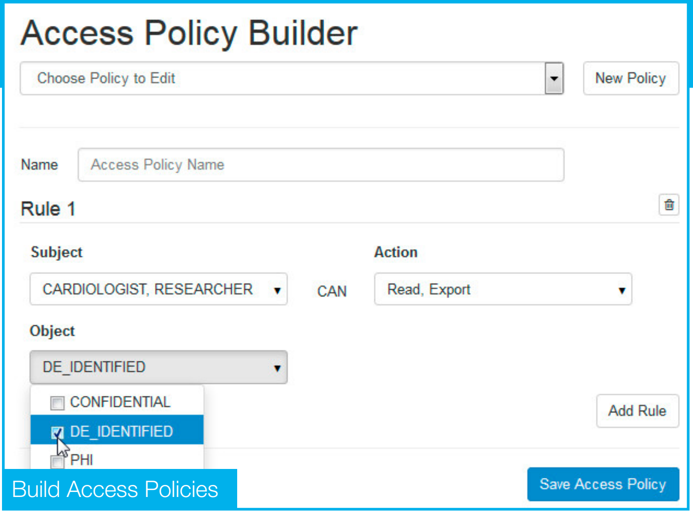
System Configuration

| Name | Description |
|---|---|
| BUSINESS_ANALYST | Business Intelligence role |
| CARDIOLOGIST | Professional care worker in Cardiology |
| PUBLIC | Public Level Authorization |
| RESEARCHER | A new authorization being created for identifying users that are Researchers. |

Define User Attributes

Data attributes and user attributes can be defined and refined through the System Configuration settings. Sets of data attribute definitions can be created along with sets of users with various authorization settings.

Privacy, security, and governance features are one coordinated design across the system — no need for bolted-on, external and fragmented security mechanisms.



Access Policy Builder

Name: Access Policy Name

Rule 1

Subject: CARDIOLOGIST, RESEARCHER CAN Action: Read, Export

Object: DE_IDENTIFIED

☐ CONFIDENTIAL
☑ DE_IDENTIFIED
☐ PHI

Build Access Policies

Data exports can be fully de-identified to reduce the risk of data being compromised. With this capability, a stolen laptop or USB stick with Personal Health Information (a common HITECH violation) will not result in a privacy breach since the exported data has been de-identified.

Attribute based access control reduces complexity and reduces the risk of data breach.

# Privacy, Security, and Governance

Protect both privacy and data consistency with sophisticated automatic data de-identification. Audit operations simply and reliably. Fine-tune data encryption settings.

### Data De-Identification

PHEMI Central can automatically de-identify, encrypt, or mask personal information and enforce privacy based on sophisticated user attributes and fine-grained sharing and consent rules. PHEMI Central stores the fully identified data but strictly controls the rightful use of all digital assets. When the user's attributes and the recorded data sharing agreements dictate, PHEMI Central can invoke a Data Processing Function to de-identify or anonymize any information. Anonymization and de-identification may include disallowing access to personally identifiable information, masking certain information, redacting content, or may involve more sophisticated data dependency algorithms to reduce the risk of re-identification. Centralizing anonymization and de-identification helps reduce data sprawl and reduces the risk of data consistency errors.

### Audit Log

PHEMI Central maintains complete audit logs of system and user operations. They include all create/modify/delete operations, along with a record of all queries made to the system through the REST interface. The audit log files are completely tamperproof for all users. Approved users can filter log files and export the information for downstream analysis and compliance reporting.

### Encryption at Rest

For performance reasons, it is usually unnecessary to encrypt all data. Instead, encryption of only personally identifiable information is advised. PHEMI Central allows you to specify what data must be encrypted when at rest within the system.

### Encryption in Motion

PHEMI Central can encrypt links from data sources and to consuming applications and analytics tools using either Secure Sockets Layer (SSL) or Transport Layer Security (TLS).

### Consent Management

PHEMI Central's unique metadata approach to data management allows consent directives to be explicitly linked to an individual patient's data. Logical control rules can then be applied to enforce the consent directives, such as allowing a patient to direct their data to be used for specific research purposes.

### Data Sharing

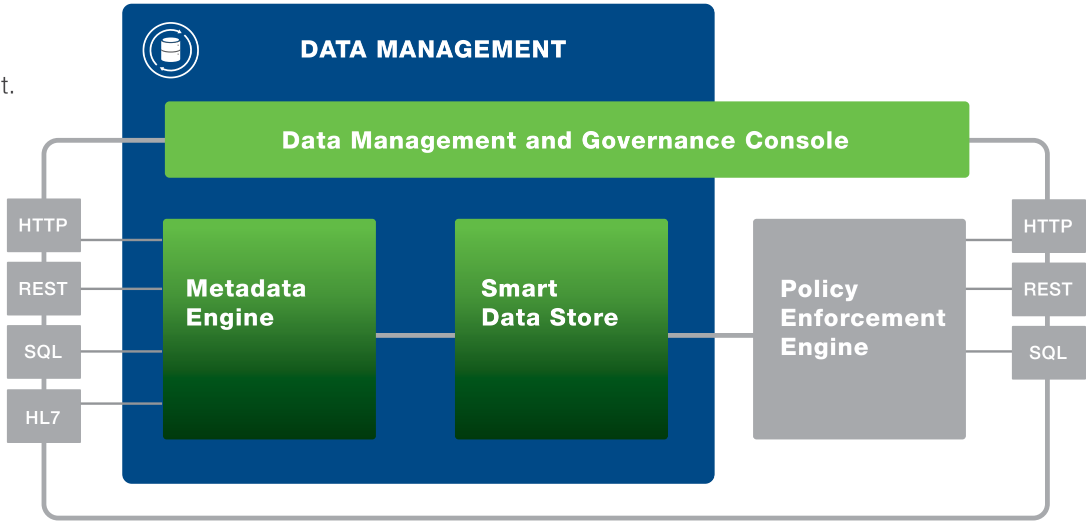
PHEMI Central opens up new opportunities in data sharing, enabling organizations to enforce strict governance requirements while still enabling use of data that is approved and in accordance with established policies. Organizations remain compliant, but governance does not become a roadblock to sharing useful and important data. Combined with de-identification and auditing, organizations gain increased ability to flexibly but safely share data.

PHEMI Central stores fully identified data to make linking across silos easier, but strictly controls the rightful use of all digital assets.

# Data Management

## Manage your data right at the data field level.

Proper management of data through its lifecycle is critical as volumes grow and variety increases. The PHEMI Central Data Management and Governance Console provides the primary interface for data management.



### Powerful Metadata Framework

The power and sophistication of PHEMI Central's data management capability arises from its metadata framework, which extends end to end 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 embedding the rules and policies governing the element, as well as configured properties such as the data source 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 can be enforced per data item, at the individual field level.

In a traditional system, a user has to plan a file system hierarchy or a database schema. Data is forced to comply with this rigid hierarchy, and everyone simply hopes that the design scales and that requirements do not change. When PHEMI Central's metadata framework is deployed with its scalable, distributed key-value store, users no longer need worry about how to structure the system. PHEMI Central structures data automatically, on the fly. Data scales to large volumes at minimal cost while 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 web-service calls based on the properties of the data, rather than having to navigate complex directories or schemas to find the data they seek.

### Governance Rule Enforcement

Most organizations have governance rules and data sharing agreements that stipulate how data may be used and shared. Governance rules are instantiated in PHEMI Central using the PHEMI Central Data Management and Governance Console. PHEMI Central manages and enforces data sharing agreements by flagging the sensitivity of individual digital assets, tracking the retention period, recording rules around version control, and specifying de-identification, encryption, and data access permissions. Automating this function across a variety of data sources and types is critical to managing privacy, security, and governance at scale.

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

# Data Management

### Lifecycle Management

PHEMI Central uses retention rules to manage digital assets throughout their lifecycle, from data creation through curation, usage, and end of life. Retention rules are captured in the Data Management and Governance Console, and the system calculates a time to live for every digital asset based on the retention policy and time of ingestion. The system also prevents users from deleting data during a configured retention period and automatically de-identifies, deletes, or otherwise processes information when the retention period expires.

### Data Immutability

PHEMI Central stores all data in a write-only data system that is never modified. Data is only deleted when its predetermined time to live expires, as specified by 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, and the ability to roll back changes if necessary. This design provides a complete history for audit and compliance requirements.

### DPF Management

Just as the metadata framework manages the tagging of data items throughout the lifetime of data in the system, the PHEMI Central DPF framework manages DPF deployment and execution across the entire system. The DPF framework is very simple and easy for programmers to learn and use: code libraries are uploaded into the system as simple ZIP files and PHEMI Central manages DPF execution across all datasets and all data elements within the system.

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

# System Management

## Get enterprise-grade reliability, availability, and scalability with cluster economics.

### Commodity Hardware

PHEMI Central eliminates the cost and performance bottlenecks associated with expensive Storage Area Network (SAN) or Network Attached Storage (NAS) architectures. The system uses low-cost commodity hardware components and direct-attached disk drives to significantly lower the cost of ownership compared to traditional enterprise data warehouse systems.

### Scalability and Performance

PHEMI Central can easily aggregate structured and unstructured data, scaling storage and compute linearly from terabytes to petabytes with each additional hard drive and node.

### Reliability and Availability

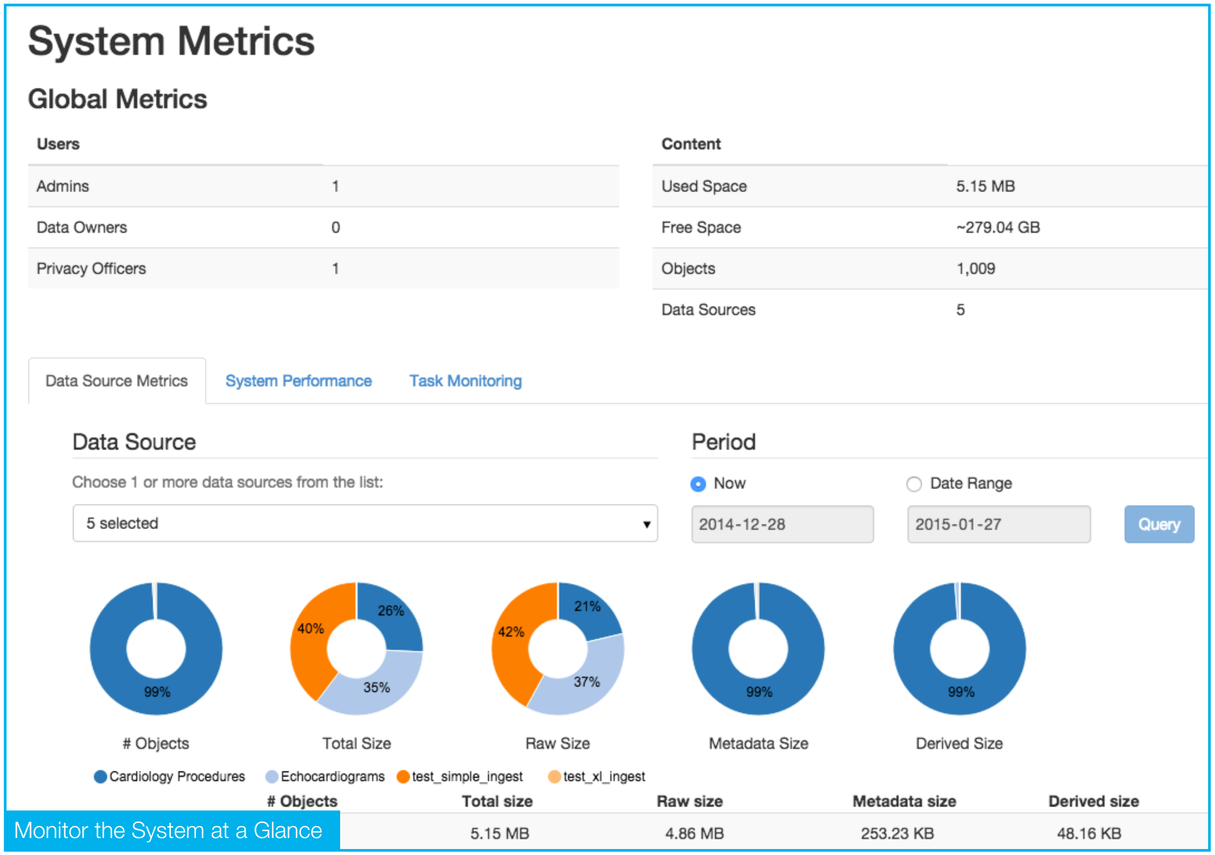
All data stored in PHEMI Central is replicated three times across the system to ensure high availability and resiliency in the event of a hardware failure. Direct-attached hard drives can be hot-swapped without impacting performance or data availability, while larger or faster drives and nodes are absorbed into the system and load balanced automatically.

### LDAP integration

PHEMI Central can interoperate with your existing LDAP or Active Directory identity management systems or use internally-managed PHEMI Central user accounts.

### Maintenance and Support

PHEMI Central provides clear visibility into overall system health, diagnostics, troubleshooting, capacity, and digital assets under management so that IT administrators can quickly configure and test all nodes. Additionally, with Apache Ambari, system management capabilities can be integrated with IT infrastructure monitoring tools such as Microsoft System Center, Teradata Viewpoint, Nagios, and Ganglia.



System Metrics

Global Metrics

| Users | |
|---|---|
| Admins | 1 |
| Data Owners | 0 |
| Privacy Officers | 1 |

| Content | |
|---|---|
| Used Space | 5.15 MB |
| Free Space | ~279.04 GB |
| Objects | 1,009 |
| Data Sources | 5 |

Data Source

Choose 1 or more data sources from the list:

5 selected ▾

Period

● Now ○ Date Range

2014-12-28 2015-01-27

| # Objects | Total size | Raw size | Metadata size | Derived size |
|---|---|---|---|---|
| | 5.15 MB | 4.86 MB | 253.23 KB | 48.16 KB |

Monitor the System at a Glance

See a summary of current system activity and settings in one convenient dashboard view — administrators have clear visibility into overall system health, diagnostics, troubleshooting, capacity, and all digital assets under management. With options to filter by time period, data source, and task, the metrics summary can be fine-tuned to extract relevant information. Potential problems can be spotted early and resolved.

# Deploying PHEMI Central

## PHEMI Central is available in either an on-premise or a cloud solution and can typically be operational within 30 days.

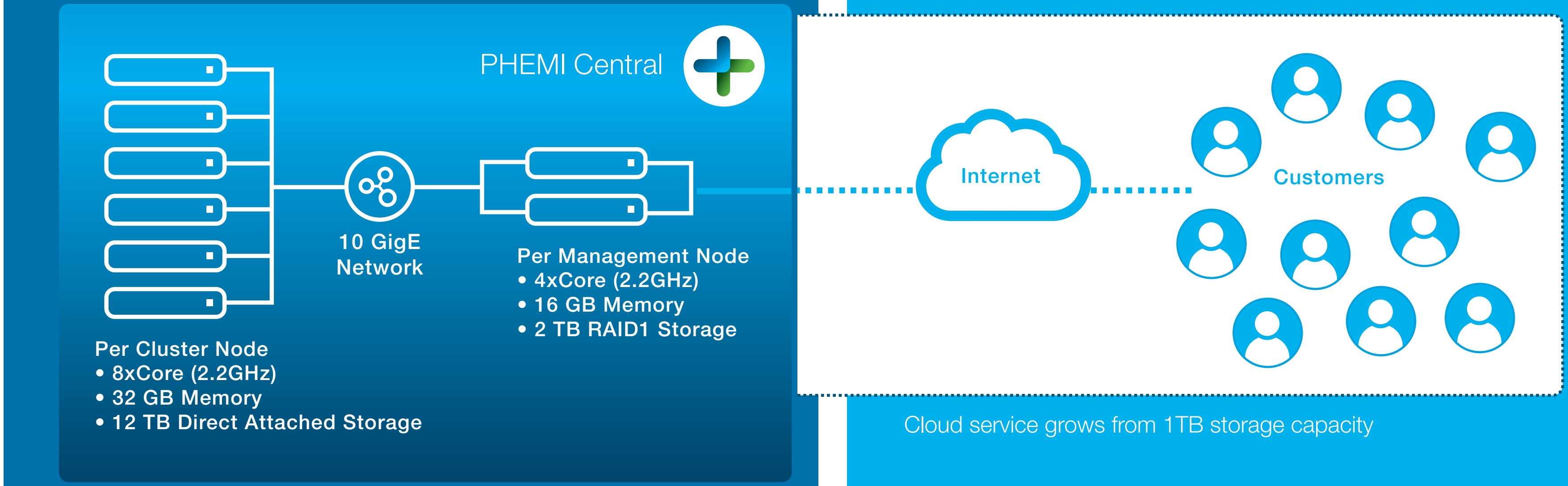
All our deployments align with appropriate privacy and security requirements, including the rules and regulations of Canadian federal and provincial legislation, and the US's Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health (HITECH) Act.

### On-Premise

Run PHEMI Central at your site. Choose between a virtual machine or physical server deployment.

### In the Cloud

Subscribe to PHEMI Central as a managed service in a cloud-based deployment, using industry-leading Amazon Web Services.



PHEMI Central

10 GigE Network

Per Management Node
- 4xCore (2.2GHz)
- 16 GB Memory
- 2 TB RAID1 Storage

Per Cluster Node
- 8xCore (2.2GHz)
- 32 GB Memory
- 12 TB Direct Attached Storage

Sample Configuration

Cloud service grows from 1TB storage capacity

# What's your path to big data?

## With PHEMI Central, deployment of your new data warehouse is painless.

PHEMI Central preserves your investment in your existing solution, while providing a safe, responsible path to the future. Whether you want PHEMI Central to work with your existing data warehouse, or you're looking for the migration path to big data, PHEMI Central is flexible enough to do both.

### Get Started with Big Data

PHEMI Central is a great way to get started with big data. Use PHEMI Central to meet your most pressing needs by starting with a specific application — a disease registry, a hospital readmissions challenge, or a government self-serve portal. Start with a basic deployment, then add applications and data as you need to, increasing storage capacity incrementally.

### Take First Steps Beyond Data Silos

Implement a PHEMI Central deployment to overcome silos of critical information — offer a platform for multiple users and multiple projects. Identify the key data your researchers just can't access today, and make this your first step. Start by giving your researchers a single data store with easy access for a number of applications.

### Extend Your Existing Data Warehouse

We know not all organizations are ready to leave their existing data warehouse behind. Use PHEMI Central to capture new data types and sources. Let your finance department keep using their existing system, but add in PHEMI Central so you can bring in more data — data with formats, types, or volumes your existing system can't accommodate.

Let PHEMI Central feed data into your favorite tools, your current applications, or your existing data warehouse systems.

# Applications

The PHEMI Central big data warehouse allows enterprises to deliver a single application, or deploy a platform for a wide range of big data research and analytics applications. With PHEMI Central, forward-looking organizations can address immediate data capture, storage and sharing needs, and add new data, users, and applications as organization requirements demand.

## Enterprise Information Applications

### Build New Applications

Consolidate data silos and create a warehouse of curated digital assets.

Quickly build innovative new and experimental applications for stakeholders. Those applications that gain traction can be hardened and expanded while others are retired.

Outsource application development without compromising privacy, security, or governance.

### Retire Legacy Systems

Reduce database license costs by retiring and consolidating legacy systems.

Migrate data and point the retired application to the PHEMI Central Big Data Warehouse with commodity hardware economics.

Manage privacy and security of data.

Provide read-only access to data for archive and historic purposes.

### Analyze Machine Data

Collect billions of machine-readable messages from streaming data and *Internet of Things* endpoints.

Convert messages into digital assets, build aggregates, and monitor data for anomalies, trends, and out-of-range readings.

Conduct predictive analysis, trending, and ad hoc analysis via third-party analytics tools.

### Curate Documents

Aggregate Microsoft Word, Excel, PDF, image, and text documents with commodity hardware economics, applying governance rules on all documents to control rightful access.

Perform keyword parsing and Optical Character Recognition where necessary to index text in all files.

Allow users to search for key words, exposing documents based on their access privileges.

### Self Serve Data

Aggregate and curate an inventory of analytics-ready digital assets, applying privacy, security, and governance rules.

Provide a portal for users to authenticate themselves and gain access to appropriate identified and de-identified data based on their access privileges.

Empower developers, data scientists, and data analysts to explore and innovate.

# Applications

## Healthcare Applications

### Personalized Medicine

Aggregate genotype and phenotype data at scale with commodity hardware economics. Integrate mircoarray, Whole Genome Sequences, and microbiome data.

Add clinical data from the hospital EMR and claims data, supplementing it with datasets from patients, researchers, and clinicians.

Apply and enforce data sharing agreements.

Convert all raw data into fine-grained analytics-ready data, across petabytes of de-identified data, and export it for cluster analysis.

Integrate advanced bioinformatics and visualization tools to navigate, annotate, and discover insights.

Search for biomarkers to select the best drugs and treatment for individuals based on their genetic profile.

### Population Health

Convert medical reports into SNOMED, LOINC, RxNorm, and ICD codes.

Link the data and write Data Processing Functions to flag patients for screening.

Integrate data from various hospital information systems.

Protect patient data during population health studies.

### Quality & Outcomes

Identify and close gaps in care. Integrate information from various hospital information systems.

Apply and enforce data sharing agreements.

Convert medical reports into SNOMED, LOINC, RxNorm and ICD codes.

Link the data and analyze treatment plans and outcomes by provider.

Use your existing analytics tools to generate de-identified reports for each provider showing peer benchmarks compared to registry and other in-network providers.

Conduct business analytics on demand to build a clearer view of financial and operational performance and measure the impact of operational changes in real time.

# Applications

## Public Sector Applications

### Open Data

Aggregate data sources.

Apply and enforce data sharing agreements.

Build an inventory of digital assets.

Mark the sensitivity of data such as social insurance or social security numbers (Public, Classified, Secret, Top Secret, etc.).

Link data across various sources.

Allow self-service access to some data and accept dataset applications when users request more sensitive data, de-identifying data when appropriate based on the risk of re-identification.

### Program Review

Collect data from government information systems.

Apply and enforce data sharing agreements.

Write Data Processing Functions to convert Microsoft Excel, database, and unstructured data into structured digital assets.

Use your existing analytics tools to conduct ad hoc analysis, drilling in to program effectiveness and identifying opportunities to improve program efficiency.

### Citizen Services

Build new applications to automate citizen services such as permitting and licensing.

Manage the privacy, security, and governance of applicant data and de-identify information based on the reviewer identity.

Draw on information from various government information systems to supplement the application and apply rules as part of the process workflow to validate or approve the applications.

# Why PHEMI Central?

PHEMI Central goes beyond, offering a perfect big data solution for a variety of organizations.

## Beyond the traditional enterprise data warehouse

- **Aggregate New Data Sources Faster**
  PHEMI Central uses a schemaless architecture to quickly aggregate new data sources, lowering startup cost and complexity.
- **Integrate Any Data Type**
  PHEMI Central easily works with structured, semi-structured, and unstructured data.
- **Lower Cost of Ownership by 60%**
  PHEMI Central uses low-cost commodity hardware, eliminating expensive storage and server hardware.
- **Scale to Petabytes**
  PHEMI Central is built on big data technology — proven to scale to petabytes in production environments worldwide.
- **Enforce Privacy, Security, and Governance across Your Organization**
  Governance rules for privacy and security are enforced within the PHEMI Central Big Data Warehouse, rather than at the application layer, ensuring data custodians — not application developers — retain control over privacy and security.

## Beyond the Hadoop data lake

- **Curate Digital Assets**
  PHEMI Data Processing Functions present a common Python, Java, or C++ programming environment for data scientists, abstracting away the complexities of working with MapReduce and YARN jobs.
- **Catalog All Digital Assets**
  PHEMI Central maintains a catalog and index of all digital assets in the system so that analysts can quickly build datasets across very large and varied collections of data.
- **Enforce Fine-Grained Privacy, Security, and Governance across Your Entire Organization**
  Governance rules for privacy and security are enforced within the PHEMI Central Big Data Warehouse, rather than at the application layer, ensuring data custodians — not application developers — retain control over privacy and security.
- **Focus on Mining Digital Assets**
  PHEMI Central is an enterprise-grade, fully integrated system so you don't have to worry about the confusing world of Apache Hadoop, MapReduce, YARN, Pig, HIVE, Sqoop, HBase or Accumulo.  Instead, you can focus on unlocking your information silos and discovering new insights with your digital assets.

# About PHEMI

PHEMI was founded in 2013 by a team of proven entrepreneurs and industry experts. Headquartered in Vancouver, Canada, the PHEMI team has extensive experience bringing innovative technologies to enterprise-class customers. Industry expertise, including healthcare and security, drives PHEMI Central features, while networking and high performance computing technology expertise drive PHEMI architecture to meet the challenges of big data.

PHEMI Central gives organizations the agility to seamlessly collect data sources, catalog and curate a powerful inventory of secure digital assets, conceive new business applications, and rapidly build new solutions to support strategic objectives.

PHEMI partners with best-in-class technology and service providers to deliver a complete solution to meet any organization’s needs.

Visit www.phemi.com for more information.

www.phemi.com info@phemi.com twitter.com/PHEMIhealth linkedin.com/company/phemi

PHEMI