PHEMI Central Big Data Warehouse

PHEMI Central™ is a production-ready big data warehouse with built-in privacy, data sharing, and data governance. PHEMI Central delivers the scalability and economics of Hadoop with indexing, cataloging, fine-grained access control, and enterprise-grade data management.

Economically integrate all of your data.

The PHEMI Central Big Data Warehouse, built on Hadoop, unlocks your data silos and makes all of your data available for analytic and operational applications. Big data technology allows you to scale to petabytes of data with cluster economics. PHEMI Central adds simplified deployment and out-of-the-box operations and the ability to integrate immediately with existing data sources and analytics tools. PHEMI's solution simplifies every step, from installation to getting your end users up and running.

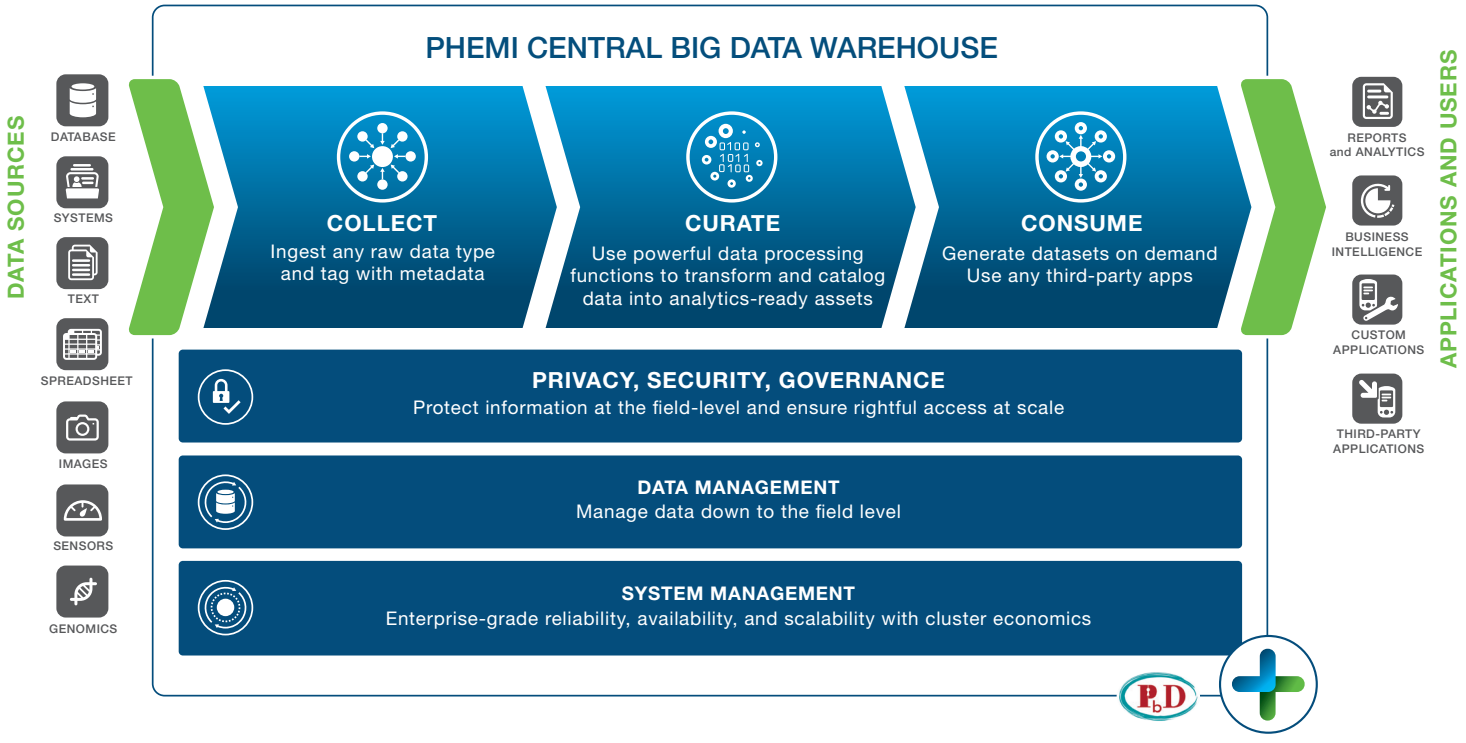
Index, catalog, and enrich your data for findability and performance.

PHEMI Central intelligently curates your data from the moment it arrives. Data is indexed and cataloged on ingest for sub-second lookups. PHEMI Central can cleanse, standardize, encrypt, or otherwise process selected information. Data Processing Functions—programs written in common programming languages that operate right in the datastore—allow you to transform data at any point in time, without having to manage the complexities of distributed computing.

The PHEMI Data Catalog adds the ability to curate virtual datasets for controlled, self-serve, point-and-click access by end users.

Protect and share data with privacy, security, and governance.

PHEMI's privacy, data sharing, and data governance features let you automatically protect private and sensitive data, while being able to share it with collaborators. PHEMI Central's Zero Trust Data implementation controls access to information at the field level, ensuring rightful access at scale. PHEMI Central can anonymize, hide, or redact data for different user authorizations. Access control based on user attributes and metadata enforces rightful access, and advanced data management features like version control, lifecycle management, and data sharing agreement enforcement help you achieve compliance and governance objectives.



PHEMI Big Data. Shared. Simply. Securely.

Drive discovery and fuel innovation with big data economics, while meeting compliance and governance objectives.

COLLECT

Ingest and tag all types and any size of data.

PHEMI Central ingests data from multiple and disparate sources. Data can range from small kilobyte files to large multi-Gigabyte files. Schemaless ingestion is fast.

- Stream data from machine-to-machine data sources using the PHEMI REST API
- Push data directly from data sources or ETL tools using REST
- Upload data manually from a standard web browser window

Data is tagged on ingest with descriptive metadata that directly reflects provenance, privacy policies, and data sharing agreements, and controls the data lifecycle.

CURATE

Extract the greatest possible value from your data with processing, indexing, cataloging, linking, and metadata.

PHEMI Central automatically indexes and catalogs data as it is stored, making it findable and retrievable. Sophisticated metadata tagging is used to describe, manage and govern the data that it stores. After cataloging and indexing, data can be linked based on keywords, graph relationships, and geospatial attributes. Data linking expands the kinds of connections you can make between data items, promotes discovery, and gives you a more complete picture of your data.

CONSUME

Consume data at speed and scale while protecting it.

Describing or tagging information with metadata means that users and applications can query data based on the data's properties, instead of navigating complex directories or schemas. This means data in PHEMI Central is searchable and findable, so you can access your datasets on demand at sub-second speeds, even with petabytes of data.

Multiple users can interact with the system, accessing datasets via SQL, data exports, and applications. Use the analytics tools of your choice, including R, SAP, SAS, SPSS, Stata, Tableau, or SAP Lumira, or write innovative applications that use the REST API. On-demand virtual datasets are lightweight, instantiated only on export, reducing data sprawl.

PHEMI Central leverages three unique innovations

Metadata Framework
Extensible, descriptive, end-to-end metadata enables field-level access control and data management.

DPF Framework
PHEMI Central lets you develop custom programs, called Data Processing Functions (DPFs), that can parse, recognize, extract, cleanse, standardize, encrypt, mask, or otherwise transform selected fields. The complexities of distributed computing are handled automatically by PHEMI's DPF Framework.

Policy Enforcement Engine
Every user request for data is filtered through a rules engine based on metadata and user attributes. PHEMI's Zero Trust Data framework ensures that data remains governance-compliant at all times.

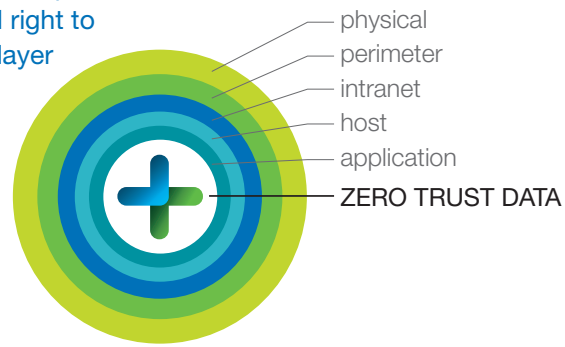
PHEMI

Optimize privacy, security, and governance with Zero Trust Data.

PHEMI Central provides an industry-pioneering set of capabilities to manage the governance of sensitive data, enforced throughout the lifecycle of data. PHEMI Central uses its innovative Zero Trust Data framework to implement the 7 Foundational Principles of Privacy by Design—managing and enforcing data access and privacy policies on a request-by-request basis, across an entire organization or set of organizations.

ZERO TRUST DATA

Defense in depth extended right to the data layer



Zero Trust Data represents a new approach to privacy, data sharing, and governance in a data-driven world. In networking, zero trust means "Never trust; always verify." Access is denied by default. A zero trust approach to data extends this principle to data. Without the correct access credentials, a request for data yields no information.

In a Zero Trust Data strategy, you decouple users from data by using metadata to describe data and attributes to describe users. Metadata and user attributes are combined into a policy-based access control framework, which automatically tests data access requests against policies on a request-by-request basis.

Decoupling data description allows data to remain stable, while accommodating change around users, partnerships, and collaborations. Because access control is localized within the datastore, applications are relieved from the burden of authorization, so they are "thinner," less brittle, and less costly to develop. The overall system is more secure, too, because data can remain secure even if applications are compromised.

Zero Trust Data—an essential for privacy, data security, and governance in a data-driven world.

Privacy by Design

Privacy by Design is a framework for protecting privacy in a big data world. Developed by the former Information and Privacy Commissioner of Ontario, Dr. Ann Cavoukian, it advances the view that privacy must become part of any data system's default mode of operation. The PbD code of conduct for products and organizations using big data consists of seven foundational principles:

1. Privacy should be proactive not reactive, and preventative rather than remedial.
Organizations can't wait to address risks after they materialize. They must prevent them from occurring by tightly defining how data is stored and who is allowed to access it.

2. Privacy must be the default setting.
Any data stored should be automatically protected in any business system or practice. No action should be required to "turn on" privacy.

3. Privacy should be embedded into system design.
No data should be stored—much less accessed—without clear privacy and governance parameters associated with it.

4. Privacy and governance must operate end to end.
It should extend from prior to the first element being collected through the entire lifecycle of the data involved.

5. Data systems should be positive-sum, not zero-sum.
There should be no tradeoffs between privacy and security, or security and usability.

6. Privacy must be transparent.
Users and providers of data, as well as independent parties, should be able to verify that privacy mechanisms are operating as expected.

7. Privacy should be user-centric.
Data systems must keep the interests of the individual foremost through mechanisms like strong privacy defaults, appropriate notice, and usability.

Privacy by Design is recognized as the global privacy standard in a landmark resolution by the International Conference of Data Protection & Privacy Commissioners. Visit privacybydesign.ca.

PHEMI

Ease your entry into big data and grow your solution at your own pace.

Easy deployment

PHEMI Central can be deployed at the customer premise, as a managed service, or as a cloud-based service. Unlike Do-It-Yourself solutions, PHEMI Central is a fully integrated product, enterprise-grade and ready to deploy.

PHEMI Central integrates with your existing applications, ETL tools, analytics tools, and existing databases or Enterprise Data Warehouse. You can preserve your investment in existing infrastructure, easing your way into production and expanding incrementally.

Cluster reliability and economics

The system uses low-cost commodity hardware components and Direct Attached Storage (DAS) disk drives to lower the cost of ownership compared to traditional Enterprise Data Warehouse systems. Storage and compute resources scale linearly from terabytes to petabytes.

All data in the system is replicated three times to ensure availability and resiliency. DAS drives can be hot-swapped without impacting performance or data availability. Larger or faster DAS drives and nodes are absorbed into the system and load-balanced automatically.

Sophisticated system and data management

The system provides clear visibility into system health, diagnostics, troubleshooting, capacity, and digital assets under management. System management capabilities can be integrated with existing tools.

PHEMI Central keeps the system secure with user roles that limit what operations a user can perform. Communication links from data sources or to consuming systems can be encrypted using Secure Sockets Layer (SSL) or Transport Layer Security (TLS). The system maintains a complete, tamperproof audit log of operations and data access.

PHEMI Central also provides system and data management capabilities formerly only available in traditional data warehouses. Data retention policies and data sharing agreements are automatically enforced. Data in the system is immutable: the original data cannot be modified and data is only purged from the system based on the configured retention policy. Robust version control and rollback capabilities mean that data is never lost, corrupted, or overwritten.

Visit www.phemi.com for more information.

SPECIFICATIONS

On-Premise Deployment*

4 Data Nodes. Each:

- 8xCore (2.2GHz)
- 64 GB RAM
- 16 TB Direct Attached Storage

3 Master Nodes. Each:

- 4xCore (2.2 GHz)
- 64 GB RAM
- 300 GB RAID1 Storage
- 1 TB Storage

1 Management Node:

- 4xCore (2.2 GHz)
- 64 GB RAM
- 300 GB RAID1 Storage
- 1 TB Storage

10 Gigabit Ethernet Network

Cloud Deployment*

- Subscribe to PHEMI Central as a managed service running on Amazon Web Services.
- Cloud service grows from 1 TB storage capacity.

Data Ingest Protocols

- SFTP File Transfer
- HTTP/HTTPS Manual Upload
- REST Web Services API
- ODBC/JDBC SQL Interface

Data Export Protocols

- Excel/CSV/TSV Download
- REST Web Services API
- ODBC/JDBC SQL Interface

Analytics Tools

Supports leading analytics tools, including:

- R • SAP • SAS
- SPSS • Stata • Tableau
- Qlikview • Netezza • MySQL

Data Processing Functions

- Excel Reader • CSV Reader
- Variant Call Format (VCF) Reader
- JSON Reader • XML Reader
- Custom DPFs

Hortonworks Certified HDP

Hortonworks Certified HDP YARN READY

* All our deployments align with appropriate privacy and security requirements, including Health Insurance Portability and Accountability Act (HIPAA) and Health Information Technology for Economic and Clinical Health (HITECH) Act as well as Canadian federal and provincial legislation.

PHEMI