

DATASHEET

Pentaho Data Catalog

Automated Discovery, Classification, and Optimization

Know Your Data. Trust Your Data. Build Your Data Products.

Pentaho Data Catalog Helps Business Users Search and Understand Structured and Unstructured Data Everywhere

Without a library catalog, people can't discover the book they need. Without a data catalog, people can't discover the data they need. With Pentaho Data Catalog, you can see all the data you have, whatever form it takes, wherever it sits, check it, classify it and make it available to users.

Get Faster and More Meaningful Data to Users

A modern organization must be data fit. As data grows, so does the need and cost of maintaining the data in business-ready shape. To leverage data for business decisions and enable AI, data must be trusted, high quality and seamlessly available to the data users. Now more than ever, there is a need to discover content across structured and unstructured, on-prem and cloud. Organizations must monitor their data to spot trends and anomalies and maintain data hygiene at the speed of data growth. Policies for governing, life cycle and quality need to be enforced to ensure appropriate high-quality data is available to consumers. Data users and models can easily find and use data via the data catalog, a necessity for modern data-drive organizations.

Powerful Business Glossary Using Machine Learning

Pentaho Data Catalog (PDC) rapidly ingests, profiles and curates structured and unstructured data with both automation and machine learning. Fingerprinting of data and metadata rules are used to contextualize data with the language of the business documented in the business glossary. Its policy manager enables the implementation of governance and security policies.

A powerful rules engine helps determine quality, sensitivity, and usage patterns. Activate your metadata by leveraging monitoring and notification capabilities in the product. Construct a relationship graph across business entities and terms to add semantic understanding to data.

Data fingerprints are analyzed to determine potential duplicates, copies and similarities across data stores to assess data movement, optimization and mastering needs. Data lineage support for Open Lineage provides the ability to track data as it flows through your organization, building trust and enabling a left shift of data quality and remediation activities.

Understand Data

Automatically find, analyze and tag structured and unstructured data across. Contextualize with business glossary and governance policies.

Activate Metadata

Observe data to define measures for data over time. Monitor metadata and act upon changes, trends and anomalies in data. Leverage event-driven architecture to apply remediation before any impact is noticed downstream.

Data-Fit for AI

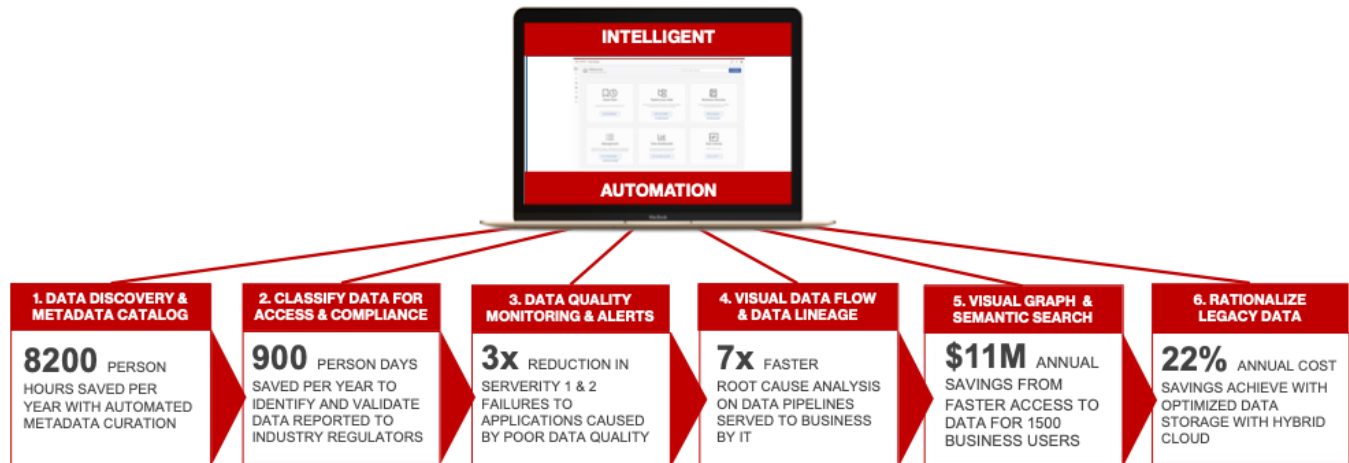
Trusted and high-quality data is made available to decision makers with a shopping experience. Catalog users deliver data to the desired destination with a No Code Data Pipe build experience

Optimization and Compliance

Measure data utilization, value and aging to make optimized storage decisions. Automated classification and characterization enable the application of life cycle, governance and access policies.

“Pentaho Data Catalog allows our staff to find the data they need so they can spend less time on data and more on water. Thank you for making your tool so easy to use!”

Lisa Williams, CDMP Manager and Founder, Office of Data Management at Arizona Department of Water



Source: Pentaho (2022)
 Based on use case across a typical US financial institution fully implemented running on hybrid cloud (AWS, Snowflake, Oracle, Salesforce).

Manage All Your Data

Manage structured and unstructured data connected to multiple and disparate data stores, such as RDBMS Systems, File Systems (NFS, HDFS, SMB) and Object Stores.

Governance for the Enterprise

Bring business and governance vocabulary, policies and standards and their application to data and applications and reports. Determine lineage and usage.

Feature Rich

Bring in reference data, usage characteristics, view semantic relationships, customize properties.

Enterprise Scale

Modern architecture is designed to scale with your data at petabyte scale – without affecting business or systems.

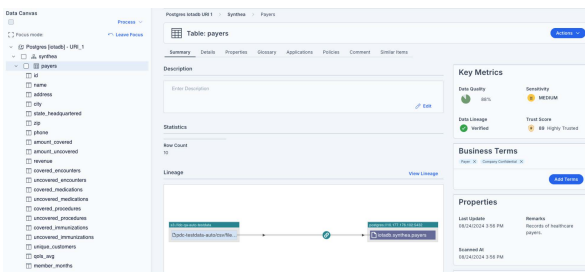


Fig.1 At-a-Glance Data Discovery & Metadata Catalog

Make Better Business Decisions with Better Data

Data with its full context, informed of its characteristics, qualified for accuracy, sensitivity, freshness helps with correctness of business decisions.

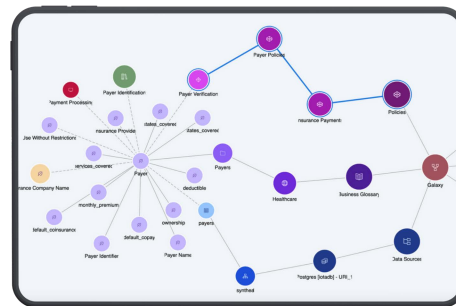


Fig.2 At-a-Glance Visual Galaxy Graph & Semantic Search

Flexibility through Modular Extensibility

Choose the applications you need and build from there – with modules for privacy, security and governance.

Discover Pentaho Data Catalog →

Pentaho Data Catalog Technical Specifications and Compatible Components
[Learn More](#)