HITACHI
Inspire the Next

Pentaho and Machine Learning Orchestration

DATASHEET

The Pentaho platform from Hitachi Vantara streamlines your entire machine learning workflow and enables teams of data scientists, engineers and analysts to train, tune, test and deploy predictive models.

Pentaho Data Integration and its analytics capablities end the gridlock associated with machine learning by enabling smooth team collaboration. Pentaho maximizes limited data science resources and puts predictive models to work on big data faster, regardless of use case, industry or language, and whether models were built in R, Python, Scala or Weka (see Figure 1).

Streamline Four Areas of the Machine Learning Workflow

Most enterprises struggle to put models to work because data professionals often operate in silos and create bottlenecks in the data preparation to model update workflow. The Pentaho platform enables collaboration and removes bottlenecks in four key areas:

1 Prepare Data and Engineer New Features

Pentaho makes it easy to prepare and blend traditional sources like enterprise resource planning (ERM) and customer resource management (CRM) with big data sources like sensors and social media. Pentaho also accelerates notoriously difficult and costly tasks of feature engineering, automating data onboarding, data transformation and data validation in an easy-to-use drag-and-drop environment.

2 Train, Tune and Test Models

Data scientists often apply trial and error to strike the right balance of complexity, performance and accuracy in their models. With integrations for languages like R and Python, and for machine learning and deep learning libraries like Spark Llib, Weka, Tensorflow and Keras, Pentaho allows data scientists to seamlessly train, tune, build and test models faster. Additionally, integration with popular integrated development environments (IDEs) such as Jupyter Notebooks makes this process seamless (see Figure 3).

3 Deploy and Operationalize Models

Pentaho allows data professionals to easily embed models developed by a data scientist as execution steps in an operational workflow. They can leverage existing data and feature engineering efforts, significantly reducing time to deployment. With embeddable APIs, organizations can also include the full power of Pentaho within existing applications.

4 Update Models Regularly

Ventana Research finds that less than a third (31%) of organizations use an automated process to update their models. With Pentaho, data engineers and scientists can retrain existing models with new

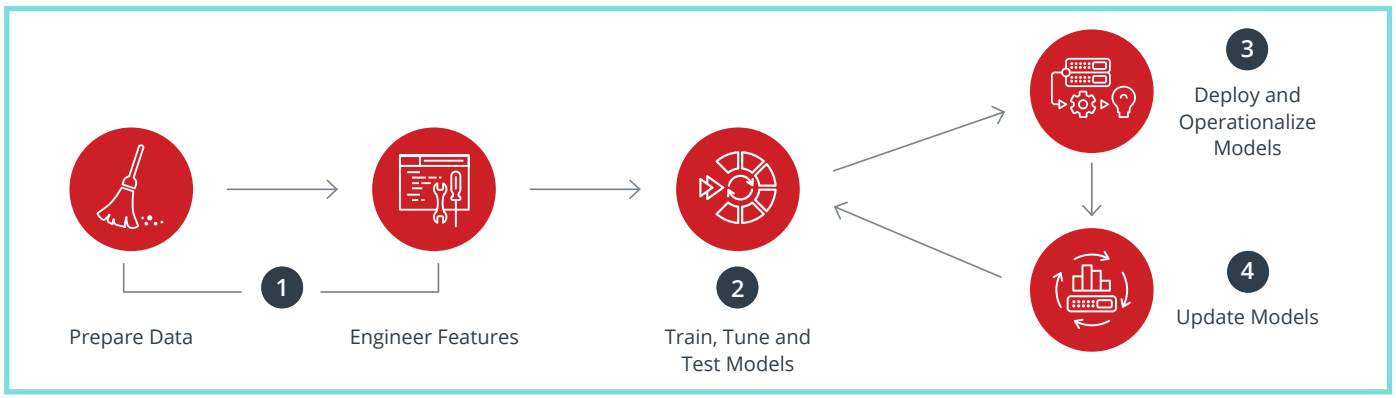

Figure 1. Pentaho addresses the four most important steps in the data science workflow.

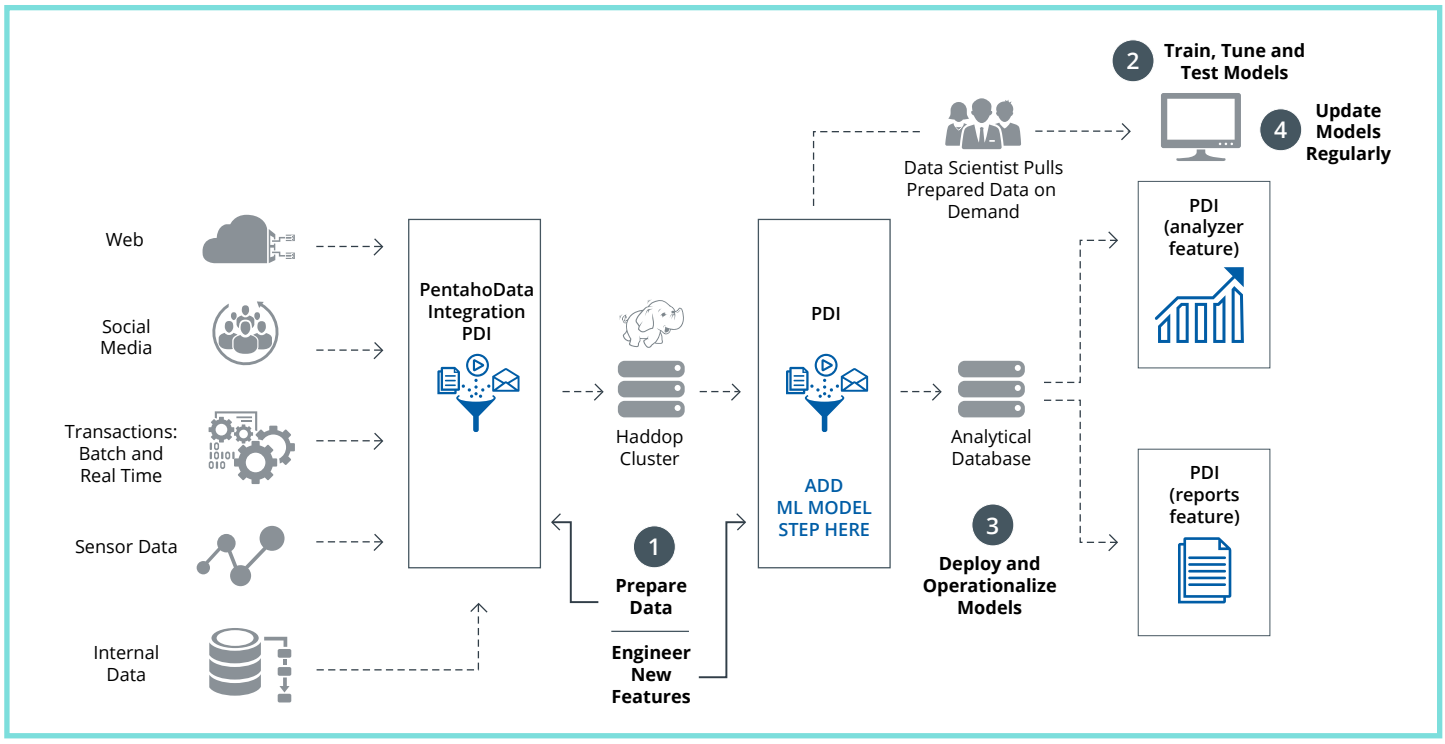

Figure 2. Deploy machine learning models using Pentaho in a complex data environment.

data sets or make feature updates using custom execution steps for R, Python, Spark MLlib and Weka. Prebuilt workflows can automatically update models and archive existing ones.

End-to-End Architecture

Pentaho makes it easy to onboard a wide variety of data sources into your data management environment (see Figure 2). Using our drag-and-drop user interface, you can blend, cleanse and standardize data quickly. Your data scientist can engineer new features and pull this prepared data, on demand, to train, tune and test machine learning models. Your data engineer can then deploy these models into a production environment and transform your business. Finally, to update models, your data scientist can regularly use new training data with the transformations already built in Pentaho.

Figure 3. Integrate various machine learning and deep learning languages and packages.

"Pentaho fills a gap to operationalize the data integration process for advanced and predictive analytics. We have embedded Pentaho for over seven years to provide remote and onboard analytics for maritime fleets and ships and have several years' experience using Pentaho Data Integration. With Weka and R integration, we are now helping clients blend a 360-degree view of all equipment data sources to enable early prediction of potential machinery failure."

– Ken Krooner, President, CAT Marine Asset Intelligence

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
HitachiVantara.com | community.HitachiVantara.com

Contact Information
USA: 1-800-446-0744
GLOBAL: 1-858-547-4526
HitachiVantara.com/contact

P-021-D BTD June 2019