Pentaho Data Integration

Pentaho Data Integration (PDI) provides the Extract, Transform, and Load (ETL) capabilities that facilitate the process of capturing, cleansing, and storing data using a uniform and consistent format that is accessible and relevant to end users and IoT technologies.

Parent Topic

- Products

Child Topics

- Kettle
- Get Started with the PDI client

  PDI client (also known as Spoon) is a desktop application that enables you to build transformations and schedule and run jobs.
- Use the Data Integration Perspective

  In the Data Integration perspective, workflows are built using steps or entries joined by hops that pass data from one item to the next.
- Use the Schedule Perspective
- Streaming analytics
- Advanced topics
- Troubleshooting

Kettle

If you are new to Pentaho, you may sometimes see or hear Pentaho Data Integration referred to as "Kettle." Pentaho Data Integration began as an open source project called "Kettle." The term, K.E.T.T.L.E, is a recursive term that stands for Kettle Extraction Transformation Transport Load Environment. When Pentaho acquired Kettle, the name was changed to Pentaho Data Integration. Other PDI components such as Spoon, Pan, and Kitchen, have names that were originally meant to support the "culinary" metaphor of ETL offerings.

Parent Topic

- Pentaho Data Integration

Get Started with the PDI client

PDI client (also known as Spoon) is a desktop application that enables you to build transformations and schedule and run jobs.

Common uses of PDI client include:

- Data migration between different databases and applications
- Loading huge data sets into databases taking full advantage of cloud, clustered and massively parallel processing environments
- Data cleansing with steps ranging from very simple to very complex transformations
- Data integration including the ability to leverage real-time ETL as a data source for Pentaho Reporting
- Data warehouse population with built-in support for slowly changing dimensions and surrogate key creation (as described above)

Learn more

Parent Topic

- Pentaho Data Integration

Child Topics

- Use Pentaho Repositories in Pentaho Data Integration

Use Pentaho Repositories in Pentaho Data Integration

The Pentaho Data Integration Client offers several different types of file storage. If your team needs a collaborative ETL (Extract, Transform, and Load) environment, we recommend using a Pentaho Repository. In addition to storing and managing your jobs and transformations, the Pentaho Repository provides full revision history for you to track changes, compare revisions, and revert to previous versions when necessary. These features, along with enterprise security and content locking, make the Pentaho Repository an ideal platform for collaboration.

Learn More

Parent Topic

- Get Started with the PDI client

  PDI client (also known as Spoon) is a desktop application that enables you to build transformations and schedule and run jobs.

Use the Data Integration Perspective

In the Data Integration perspective, workflows are built using steps or entries joined by hops that pass data from one item to the next.

This workflow is built within two basic file types:

- Transformation Steps: to perform ETL tasks.
- Job Entries: to orchestrate ETL activities such as defining the flow, dependencies, and execution preparation.

Learn more

Parent Topic

- Pentaho Data Integration

Use the Schedule Perspective

In the Schedule perspective, you can schedule transformations and jobs to run at specific times.

Learn more

Parent Topic

- Pentaho Data Integration

Streaming analytics

You can retrieve data from a message stream, then ingest it after processing in near real-time.

Learn more

Parent Topic

- Pentaho Data Integration

Advanced topics

The following topics help to extend your knowledge of PDI beyond basic setup and use:

- PDI and Hitachi Content Platform (HCP)

  You can use PDI transformation steps to improve your HCP data quality before storing the data in other formats, such as JSON , XML, or Parquet.
- PDI and Snowflake

  Using PDI job entries for Snowflake, you can load your data into Snowflake and orchestrate warehouse operations.
- Metadata discovery

  You can use to automate the tedious process of manually identifying and determining metadata from Cobol Copybook and databases.
- Use Command Line Tools

  You can use PDI's command line tools to execute PDI content from outside of the PDI client.
- Metadata Injection

  You can insert data from various sources into a transformation at runtime.
- Use Carte Clusters

  You can use Carte to build a simple web server that allows you to run transformations and jobs remotely.
- Embed and Extend PDI

  Develop custom plugins that extend PDI functionality or embed the engine into your own Java applications.
- Partition Data

Split a data set into a number of sub-sets according to a rule that is applied on a row of data.

- Use a Data Service

  Query the output of a step as if the data were stored in a physical table by turning a transformation into a data service.

- Use the Marketplace

  Download, install, and share plugins developed by Pentaho and members of the user community.

- Use Data Lineage

  Track your data from source systems to target applications and take advantage of third-party tools, such as Meta Integration Technology (MITI) and yEd, to track and view specific data.

- Work with Big Data

  Use transformation steps to connect to a variety of Big Data data sources, including Hadoop, NoSQL, and analytical databases such as MongoDB.

- Use Streamlined Data Refinery (SDR)

  You can use SDR to build a simplified and specific ETL refinery composed of a series of PDI jobs that take raw data, augment and blend it through the request form, and then publish it to use in Analyzer.

Parent Topic

- Pentaho Data Integration

Troubleshooting

See our list of common problems and resolutions.

Learn more

Parent Topic

- Pentaho Data Integration