Getting Started with Pentaho Data Integration and Agile BI

About This Document

If you have questions that are not covered in this guide, or if you find errors in the instructions or language, please contact the Pentaho Technical Publications team at documentation@pentaho.com. The Publications team cannot help you resolve technical issues with products.

Support-related questions should be submitted through the Pentaho Customer Support Portal at http://support.pentaho.com.

For information about how to purchase support or enable an additional named support contact, please contact your sales representative, or send an email to sales@pentaho.com.

For information about instructor-led training on the topics covered in this guide, visit http://www.pentaho.com/training.

Limits of Liability and Disclaimer of Warranty

The author(s) of this document have used their best efforts in preparing the content and the programs contained in it. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, express or implied, with regard to these programs or the documentation contained in this book.

The author(s) and Pentaho shall not be liable in the event of incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of the programs, associated instructions, and/or claims.

Trademarks

Pentaho (TM) and the Pentaho logo are registered trademarks of Pentaho Corporation. All other trademarks are the property of their respective owners. Trademarked names may appear throughout this document. Rather than list the names and entities that own the trademarks or insert a trademark symbol with each mention of the trademarked name, Pentaho states that it is using the names for editorial purposes only and to the benefit of the trademark owner, with no intention of infringing upon that trademark.

Company Information

Pentaho Corporation
Citadel International, Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822
Phone: +1 407 812-OPEN (6736)
Fax: +1 407 517-4575
http://www.pentaho.com

E-mail: communityconnection@pentaho.com

Sales Inquiries: sales@pentaho.com

Documentation Suggestions: documentation@pentaho.com

Sign-up for our newsletter:
http://community.pentaho.com/newsletter/

Contents

Introduction

Thank you for evaluating Pentaho Data Integration, GA Release 3.2.0. Pentaho Data Integration, formerly known as Kettle, provides you with powerful Extraction, Transformation and Loading (ETL) capabilities using an innovative, metadata-driven approach. This means that, unlike some competitor products, Pentaho Data Integration is not a code generator. For example, it does not generate PERL or Java. Instead, Pentaho Data Integration provides execution based on a graphical drag-and-drop design, resulting in less maintenance, easier debugging, deployment, and creation. Pentaho Data Integration is easy to use which means you can get started immediately out-of-the-box.

NEW IMAGE

Choosing Your ETL Solution

One of the most difficult decisions in any data warehousing project is whether to populate your data warehouse, manually, using custom code, or to choose a proprietary ETL tool. Proprietary ETL offerings may get your project off the ground faster and can provide dramatic savings in maintenance costs over time but often carry a six-figure price tag just to get started. A custom solution is attractive because it does not require any up-front cost associated with software licensing and you can build it to your solution to your specific needs; however, the ongoing maintenance costs associated with custom solutions often negate initial savings.

Pentaho Data Integration offers the best of both worlds — no upfront software licensing costs and a significant reduction in total cost of ownership compared to a custom-based solution. An annual subscription that provides professional support, certified builds, and IP indemnification is also available at a fraction of the cost of proprietary offerings.

Pentaho Data Integration features and benefits include:

- A graphical interface that allows you to create transformations and jobs (groups of transformations) easily by assembling them from a set of over 100 out-of-the-box steps, including inputs, transforms, and outputs; this interface also allows you to create database schemas, which are simply tables joined together in some way
- The ability to serve as a data source, allowing you to store data in a repository for collaborative development, or be used as a file-based system, without a data repository
- The ability to copy or distribute data to multiple data points;the ability to link columns in data rows to calculate values
- An extensible architecture that allows you to easily develop your own custom interface plug-ins
- File size limited by your system memory only
- Plug-in architecture allows you to extend functionality
- Totally Java-based with broad cross-platform support for Linux, Macintosh, and Windows
- Repository-based, providing easy structured management of models, connections, logs, allowing multiple contributors to co-develop ETL transformations and jobs
- Agile BI features that leverage the power of Pentaho Data Integration allowing an ETL designer to massage data as needed, and, based on input from a business analyst, to go directly into modeling the data, visualizing the data, and finally to providing the data to users for self-serve reporting and analysis
- Enterprise-class performance and scalability, (see Note below)

Note: PDI's ability to handle partitioning and clustering provides it with almost unlimited scalability. You can set up a cluster of nodes running Carte (Carte is a simple Web server that allows you to execute transformations and jobs remotely). These cluster nodes can partition data (with different partitioning methods for distributing the data), work in parallel on fixed or CSV files, sort data, and much more. You can monitor cluster nodes using the Pentaho Enterprise Console or through a Web browser.

Audience and Assumptions

This guide is written for IT managers, database administrators, who are evaluating Pentaho Data Integration or who just want to know more about the product.

You do not need any *specialized* experience, knowledge, or skills associated with ETL to follow the exercises in this evaluation guide; however, a basic understanding of ETL, SQL, and relational database concepts is helpful.

To complete the exercises, you must have an installation of the latest release of Pentaho Data Integration.

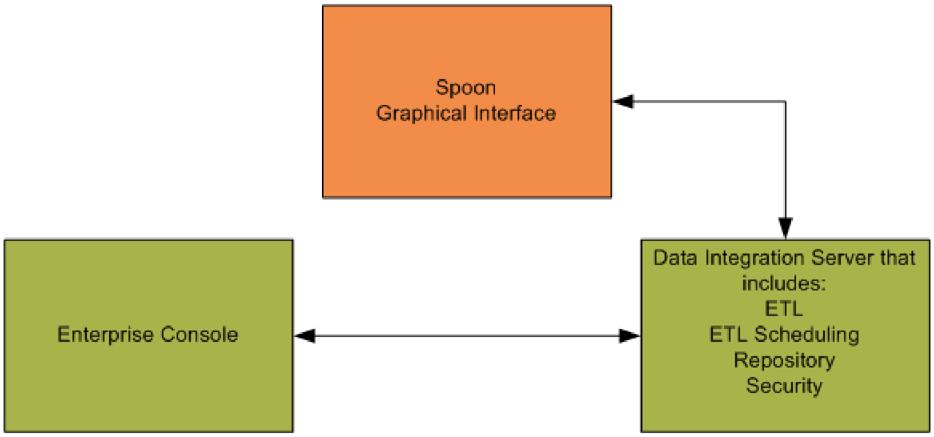
What This Guide Covers

This guide walks you through simple exercises using Pentaho Data Integration. The exercises included begin with setting up simple transformation testing tools and take you through executing a transformation, viewing the output, and troubleshooting. The guide is structured as modules of information so that you can easily skip over any sections that do not interest you.

Module	Description	Time
Pentaho Data Integration Architecture	Architectural overview	3 min.
Downloading and Installing Pentaho Data Integration	Download and installation instructions	5 min.
Navigating the User Interface	Quick introduction to the graphical interface	3 min.
Creating Your First Transformation	Instructions for creating your first transformation	60 min.

Pentaho Data Integration Architecture

The diagram below depicts the Pentaho Data Integration Enterprise Edition current architecture. **Spoon** is the interface that allows you to graphically describe what you want to take place in your transformations without worrying about execution. The **Data Integration Server** houses the ETL engine, facilitates scheduling of jobs and transformations, provides content management to the Enterprise Repository, and provides security integration with LDAP, Active Directory, (or its own version of security). The **Enterprise Console** facilitates licensing for enterprise edition reatures such as remote execution of jobs and transformations.



Downloading Pentaho Data Integration

Context for the current task

Task step.

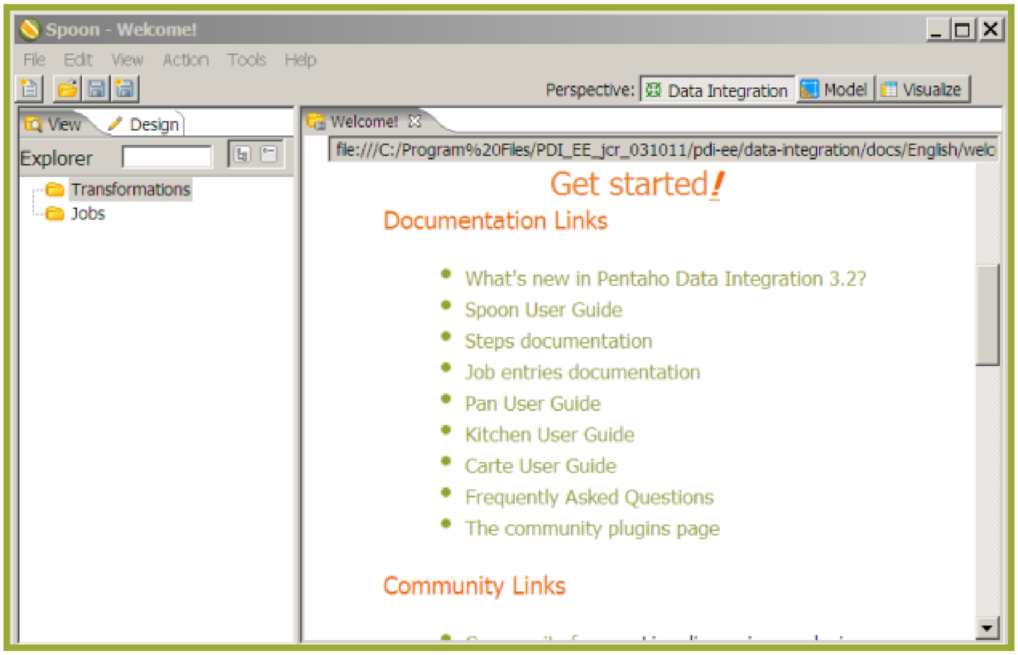
Installing Pentaho Data Integration

Context for the current task

Task step.

Navigating through the Interface

The **Welcome** page contains useful links to documentation, community links for getting involved in the Pentaho Data Integration project, and links to blogs from some of the top contributors to the Pentaho Data Integration project.


Spoon - Welcome!
File Edit View Action Tools Help
Perspective: Data Integration Model Visualize
View Design
Explorer
Transformations
Jobs
Welcome!

file:///C:/Program%20Files/PDI_EE_jcr_031011/pdi-ee/data-integration/docs/English/welc

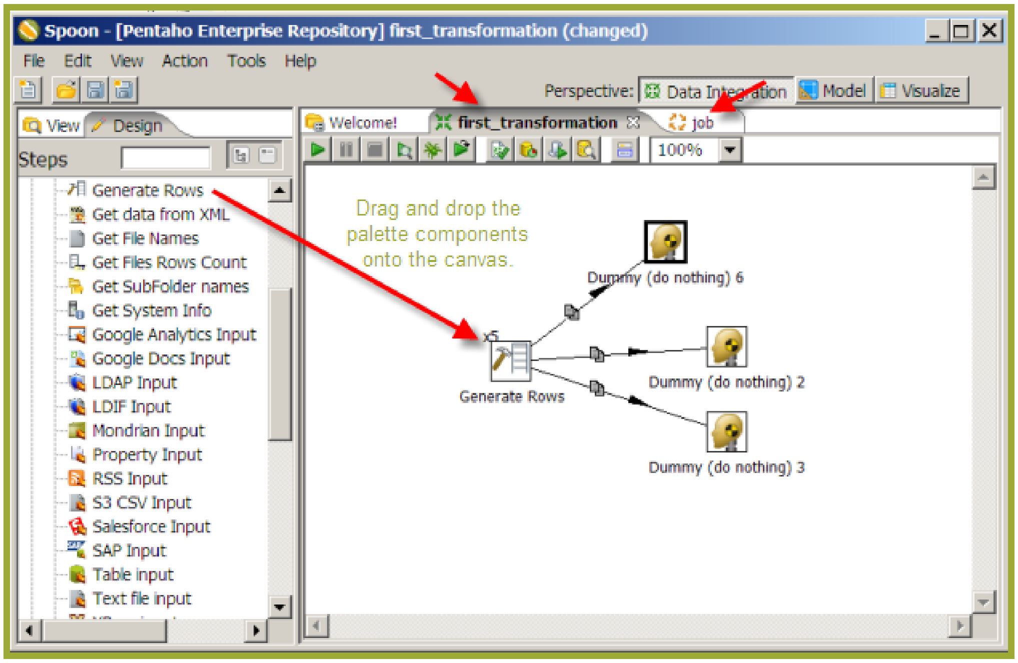
Get started!

Documentation Links

- What's new in Pentaho Data Integration 3.2?
- Spoon User Guide
- Steps documentation
- Job entries documentation
- Pan User Guide
- Kitchen User Guide
- Carte User Guide
- Frequently Asked Questions
- The community plugins page

Community Links

The tabs on the top of the page provides you with a list of all currently open jobs and transformations that you are designing. As you select a particular job or transformation, the left pane provides you with the job entries or steps you can use to design your job or transformation.



Drag and drop the palette components onto the canvas.

Designing a transformation is as simple as dragging the palette component associated with the job entry or step into the canvas on the right.

Creating Your First Transformation

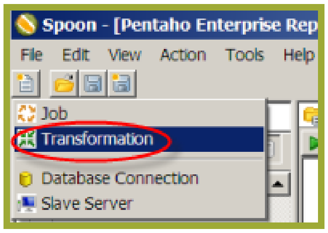
To demonstrate how Pentaho Data Integration works, you will retrieve data from a flat file, resolve missing zip code information using a separate flat file, and load the information into a relational database.

Before you begin, make sure that Pentaho Data Integration is running.

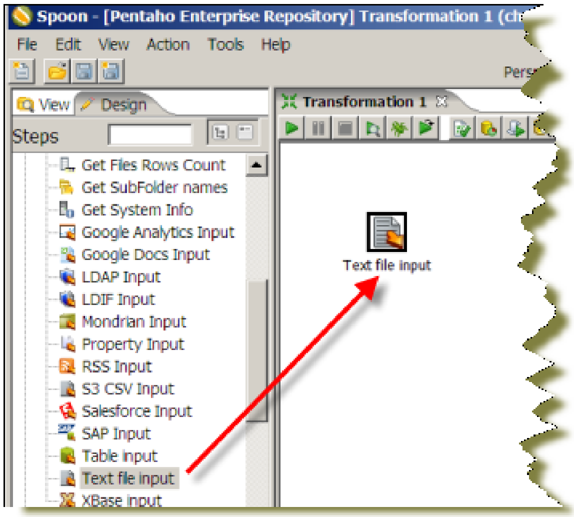
Retrieving Data from a Flat File

intro

1. Click (New) in the upper left corner of the Spoon graphical interface.
2. Select **Transformation** from the list.


Spoon - [Pentaho Enterprise Rep
File Edit View Action Tools Help
Job
Transformation
Database Connection
Slave Server

3. Drag a **Text File** input step onto the canvas on the right.



foo