



Choose Your Workflows

Copyright Page

This document supports Pentaho Business Analytics Suite 5.4 GA and Pentaho Data Integration 5.4 GA, documentation revision June 9th, 2015, copyright © 2015 Pentaho Corporation. No part may be reprinted without written permission from Pentaho Corporation. All trademarks are the property of their respective owners.

Help and Support Resources

To view the most up-to-date help content, visit <https://help.pentaho.com>.

If you do not find answers to your questions here, please contact your Pentaho technical support representative.

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, visit <http://www.pentaho.com/training>.

Liability Limits and Warranty Disclaimer

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

The trademarks, logos, and service marks ("Marks") displayed on this website are the property of Pentaho Corporation or third party owners of such Marks. You are not permitted to use, copy, or imitate the Mark, in whole or in part, without the prior written consent of Pentaho Corporation or such third party. Trademarks of Pentaho Corporation include, but are not limited, to "Pentaho", its products, services and the Pentaho logo.

Trademarked names may appear throughout this website. Rather than list the names and entities that own the trademarks or inserting a trademark symbol with each mention of the trademarked name, Pentaho Corporation 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.

Third-Party Open Source Software

For a listing of open source software used by each Pentaho component, navigate to the folder that contains the Pentaho component. Within that folder, locate a folder named licenses. The licenses folder contains HTML files that list the names of open source software, their licenses, and required attributions.

Contact Us

Global Headquarters Pentaho Corporation Citadel International, Suite 460

5950 Hazeltine National Drive Orlando, FL 32822

Phone: +1 407 812-OPEN (6736)

Fax: +1 407 517-4575

<http://www.pentaho.com>

Sales Inquiries: sales@pentaho.com

Introducing Pentaho Workflows and Tracks

We know you are revved-up and eager to get started working with the Pentaho components. Before you do, take a quick peek at the different tracks and workflows that can help you succeed with a well-planned Pentaho implementation.

The route that you take depends on different levels of expertise, what your unique business needs are, the data that you are using, and business interests when it comes to analyzing, reporting, and working with your data. You might find that you travel one track, and then the other, during your evaluation of Pentaho.

All of the products are integrated to work smoothly together, regardless of which track you ultimately choose. We provide specific details within the workflow discussions, however, here are the high-level use cases for each track.

- **Business Analytics (BA) Track** - Great for analysis and reporting. Meant primarily for business users and does not require special skills to successfully use the components involved. This track enables anyone to build Pentaho solutions without using programming or having deep understanding of data structures.
- **Data Integration (DI) Track** - Meant for data design professionals and requires a working knowledge of data structures and modeling, as well as extract, transform, and load (ETL) processes. With this track, you can directly manipulate data from multiple sources, making it scalable and efficient for enterprise-wide analysis and reporting.

Each track has three workflows: one for Evaluation, one for Development, and one for Production.

- **Evaluate and Learn** - If you used the Trial Download on the Pentaho website and want get a hands-on feel for the components that are best for your implementation, follow the Evaluation Workflow.
- **Develop Pentaho Solutions** - After you have figured out which components are best for you and how to use them, the Develop Workflow is the process you use to build, change, and test Pentaho solutions until they meet your production requirements.
- **Go Live for Production** - When your solution is working just right, the Go Live Workflow shows how to move your solution from development to production.

Prepare for the Evaluation

This table guides you through the differences between the Business Analytics and Data Integration tracks, and helps you decide which track to follow for evaluation. You may choose to follow one track, and then the other, while you are exploring the software.

Table 1. Pentaho Track Decision Table

Explore Considerations	Choose Options	
	Business Analytics Evaluation	Data Integration Evaluation
Expertise	<ul style="list-style-type: none">• No special skills required	<ul style="list-style-type: none">• Knowledge of business requirements

Explore Considerations	Choose Options	
	Business Analytics Evaluation	Data Integration Evaluation
	<ul style="list-style-type: none"> • Knowledge of business requirements and what reports and analysis should show 	<ul style="list-style-type: none"> • Understanding of data structures and modeling • Knowledge of extract, transform, and load (ETL) processes
Data Set Description	<ul style="list-style-type: none"> • Single source of data • Data from multiple sources that have been transformed and joined into a single data mart or warehouse • Small data sets 	<ul style="list-style-type: none"> • Multiple sources of data • Data you want to transform and join in one or more data marts or warehouses • Large to enormously vast data sets
Reporting Options	Offers a wide variety of visualization and reporting options.	Offers more limited but focused reporting options that help you visualize and analyze data. BA tools can be used to generate reports based on DI-processed data.
Data Storage Types	<ul style="list-style-type: none"> • Relational databases • CSV data sources • SQL queries 	<ul style="list-style-type: none"> • Relational databases • NoSQL or Hadoop databases • Big Data of any types • Data from a web service
Recommendation	Best used by business analysts, managers, report designers, individual business units within an organization or enterprise	Best used by data scientists, data modelers, data integration and ETL developers, individual business units within an organization or enterprise, and enterprise-wide implementations

Now that you have an idea of which track you want to follow for evaluation, choose an evaluation method. This decision table explains the different options for evaluation so you can pick the option that works best for you.

Table 2. Pentaho Evaluation Decision Table

Explore Considerations	Choose Options		
	Hosted Demo	Custom Prototype	Trial Download
Track	Business Analytics	Business Analytics or Data Integration	Business Analytics or Data Integration
Summary	A cloud-based, hands-on, interactive exploration of Business Analytics reports, analysis, visualizations, and dashboards. Here you can see how easy and fun it can be to use Pentaho.	Work with Pentaho analysts and data integration specialists to plan and build a complimentary custom prototype that illustrates what Pentaho can do with your data. A representative will	Using our trial software, tutorials, and documentation, install and configure your own work environment. Then, build a prototype to get a complete Pentaho experience from installation and administration, through

Explore Considerations	Choose Options		
	Hosted Demo	Custom Prototype	Trial Download
		guide you through the entire process.	creating your first data models and build reports, analysis, dashboards, and data integration ETL transformations.
Data Source	Pentaho sample data in CSV format	Your sample data, including a range of typical data characteristics in CSV format.	Your sample data, including a range of typical data characteristics in the format that you commonly use
Hardware/ Software Requirements	Web Browser	Varies, depending on your requirements.	One computer that meets the server requirements stated here .
Recommendation	<p>Any evaluator who wants an overview of Business Analytics features.</p> <ul style="list-style-type: none"> Recommended for Business Analysts and Report Designers. <p>We recommend that you try out the Custom Prototype or Trial Download after you do the Hosted Demo.</p>	<p>All evaluators, particularly any Big Data or Data Integration evaluators.</p> <ul style="list-style-type: none"> Recommended for evaluators who want to explore Business Analytics and Data Integration features using a subset of their own data. Limited to first-time customers only. 	<p>Any evaluator who wants to independently work with Business Analytics, Data Integration tools, and Big Data.</p> <ul style="list-style-type: none"> Recommended for evaluators who want to explore Business Analytics and Data Integration features using their own data. Technical support is available to help if you have questions.

Pentaho Data Integration

Pentaho Data Integration is a robust extract, transform, and load (ETL) tool that you can use to integrate, manipulate, and visualize your data. You can use PDI to import, transform, and export data from multiple data sources, including flat files, relational databases, Hadoop, NoSQL databases, analytic databases, social media streams, and operational stores. You can also use PDI to clean and enrich the data, move data between databases, and to visualize your data.

- [Evaluate and Learn Pentaho Data Integration](#)
- [Develop Your DI Solution](#)
- [Go Live for Production - DI](#)

Evaluate and Learn Pentaho Data Integration

As you explore Pentaho Data Integration, you will be introduced to the major components, watch videos, work through hands-on examples, and read about the different features. Go at your own pace. Feel free to dig into the documentation or to contact Pentaho [sales support](#) if you have questions.

The PDI evaluation track is divided into several parts.

- [PDI Basics](#)
- [Get Acquainted with Spoon](#)
- [Build Transformations and Jobs](#)
- [Explore Big Data and Streamlined Data Refinery](#)
- [Learn About Kitchen, Pan, and Carte](#)

PDI Basics

This section familiarizes you with PDI and introduces you to basic terminology and concepts. Then, you learn how to start and configure Spoon and take a spin through the interface.

Table 1. PDI Basics Checklist

Task	Do This	Objective
What is PDI and How Does it Fit Into the Pentaho Business Analytics Platform?	<ul style="list-style-type: none">• Get a basic understanding of what PDI does.• View a video that explains how PDI fits into the Business Analytics Platform.	<ul style="list-style-type: none">• Understand PDI's role in the Pentaho Business Analytics platform.
PDI Components	<ul style="list-style-type: none">• Read about Pentaho Data Integration Architecture.• Learn about DI components	<ul style="list-style-type: none">• Understand different components of the Pentaho Integration Architecture.• Learn the primary functions of the Data Integration Server.• Understand the relationship between PDI and Kettle.

Get Acquainted with Spoon

Spoon is the PDI design tool. In this section you will set up Spoon, take a tour of the Spoon interface, and learn about the different Spoon perspectives.

Table 2. Get Acquainted with Spoon Checklist

Task	Do This	Objective
Install and Configure PDI Components	<ul style="list-style-type: none"> • Check out the hardware and software requirements for PDI. • Download Trial version of the Business Analytics Platform. (The platform includes PDI.) • Read the Install All Components documentation to install the software. • Configure the DI Server and Spoon. 	<ul style="list-style-type: none"> • Learn about Pentaho's hardware and software requirements. • Install PDI Using the Wizard • Configure the DI Server and Spoon
Introduction to Spoon	<ul style="list-style-type: none"> • Start the DI Server. • Start Spoon. • Take a tour of the Spoon interface. 	<ul style="list-style-type: none"> • Learn how to start the DI Server and Spoon. • Understand the different parts of the Spoon interface.
PDI Basic Terminology	<ul style="list-style-type: none"> • Read about Terminology and Basic Concepts. 	<ul style="list-style-type: none"> • Become familiar with transformations, steps, hops, and jobs. • Understand the relationship between transformations, steps, and hops. • Know what a job is, and what a job is composed of. • Be able to apply what you have learned about terminology to the Spoon interface.
Get a Different Perspective	<ul style="list-style-type: none"> • Get an overview of Spoon Perspectives. • Review the Model Perspective. • Check out the Visualization Perspective. • View the Schedule Perspective. 	<ul style="list-style-type: none"> • Be able to identify different perspectives. • Learn how to access different perspectives. • Be able to identify the basic types of tasks that you can perform with the different perspectives.

Build Transformations and Jobs

Now that your environment is set up and you are familiar with Spoon, you are ready to build transformations and jobs. Work through the tutorial in this section.

Table 3. Create Your First Checklist

Task	Do This	Objective
Create a Connection to the DI Repository	<ul style="list-style-type: none"> • Create a connection to the DI Repository. 	<ul style="list-style-type: none"> • Learn how to create a connection to the DI Repository.
Create Your First Transformation	<ul style="list-style-type: none"> • Work through the exercise on Creating a Transformation that involves a flat file. Click through the links that are on the bottom of the page to complete the exercise. • Save the exercise. 	<ul style="list-style-type: none"> • Learn how to retrieve data from a flat file using an Input step. • Apply filters and to create a hop. • Load data into a relational database and learn how to test database connections. • Follow an example on how to resolve missing information. • Run the transformation.
Create a Job	<ul style="list-style-type: none"> • Create a job to execute the transformation. 	<ul style="list-style-type: none"> • Be able to articulate why you would create a job. • Create a job for a transformation.
Schedule a Job	<ul style="list-style-type: none"> • Schedule a job to execute the transformation at a later time. 	<ul style="list-style-type: none"> • Schedule a job.
Learn more about commonly-used steps and job entries.	<ul style="list-style-type: none"> • Review commonly-used steps and job entries. 	<ul style="list-style-type: none"> • There are over 330 available job entries and transformation steps. There is a step for virtually anything that you want to do. But, when you are learning about PDI, it is helpful to review the most commonly-used steps and entries.

Explore Big Data and Streamlined Data Refinery

In this section, you will learn how to use transformation steps to connect to a variety of Big Data data sources, including Hadoop, NoSQL, and analytical databases such as MongoDB. You can then try working through the detailed, step-by-step tutorials, and peruse the out-of-the-box steps that Spoon provides. Learn how to work with Streamlined Data Refinery. Read about Instaview. Then, you will have an opportunity to move beyond the basics and learn how to edit transformations and metadata models.

Table 4. Explore Big Data and Streamlined Data Refinery Checklist

Task	Do This	Objective
What is Big Data?	<ul style="list-style-type: none"> • Watch one of our Big Data Videos. • Read about Hadoop Resources. 	<ul style="list-style-type: none"> • Gain an overview of Big Data and PDI. • Check out Cassandra, Splunk, MongoDB, and Hadoop Big Data resources.
Learn about Streamlined Data Refinery	<ul style="list-style-type: none"> • Learn how to work with Streamlined Data Refinery. 	<ul style="list-style-type: none"> • Learn how streamlined data refinery works.
Introduction to DI Big Data Steps and Transformations	<ul style="list-style-type: none"> • Find out what Big Data steps are available, out-of-the-box. 	<ul style="list-style-type: none"> • Review available Big Data transformation steps. • Review Big Data job steps. • Review the YARN steps.
Configure a Hadoop Distribution	<ul style="list-style-type: none"> • Find out which Hadoop Distributions are available and how to configure them. (Note, you should already have a cluster set up to perform this task.) 	<ul style="list-style-type: none"> • Configure a Hadoop Distribution. • Pentaho's Big Data adaptive layer supports over 20 different versions of popular Hadoop distributions such as Apache, Cloudera, Hortonworks, MapR, and Intel.
PDI, Hadoop, Cassandra, MongoDB, and MapR Tutorials	<ul style="list-style-type: none"> • Work through Big Data Tutorials. 	<ul style="list-style-type: none"> • Explore how to load, transform, extract, and report on data in a Hadoop and MapR clusters. • Learn how to write and read data to and from Cassandra, and how to create reports. • Learn how to create MongoDB reports and how to read and write data to and from MongoDB.
Beyond the Basics: Edit Transformations and Metadata Models	<ul style="list-style-type: none"> • Edit Transformations and Metadata Models. 	<ul style="list-style-type: none"> • Edit data transformations • Edit metadata models
Blend Big Data	<ul style="list-style-type: none"> • Watch a video about how to use PDI to blend Big Data. 	<ul style="list-style-type: none"> • Learn about the concept of blending data. • Learn about Pentaho's just-in-time approach Big Data blending.

About Kitchen, Pan, and Carte

These lessons provide an overview of Kitchen and Pan, which are command lines tool for executing jobs and transformations modeled in Spoon. You will also learn about Carte, which is a web server that enables remote execution of jobs and transformations.

Table 5. About Kitchen, Pan, and Carte Checklist

Task	Do This	Objective
What are Kitchen and Pan?	<ul style="list-style-type: none">• Learn about Kitchen and Pan	<ul style="list-style-type: none">• Gain an overview of Kitchen and Pan.
Intro to Kitchen, Pan, and Carte.	<ul style="list-style-type: none">• Learn about Pan's options.• Discover Kitchen options.• Read about Carte clusters.	<ul style="list-style-type: none">• Learn about the capabilities of Kitchen, Pan, and Carte

Learn More

Now that you have completed an initial evaluation of PDI, dig a little deeper. Find out how to:

- Use newer steps and entries, like Spark, [R Script Executor](#), [Pentaho MapReduce](#), and [YARN Kettle Clustering](#).
- [Create other DI Solutions.](#)
- [Administer PDI.](#)
- Integrate with different security protocols, like [Pentaho Security](#), [LDAP](#), [MSAD](#), and [Kerberos](#).
- Check out our [developer's documentation](#).
- Use [Agile BI to create reports](#).
- Learn more about predictive analytics by [watching this video](#).

Next Steps

- Contact Pentaho [sales support](#) to learn more about how Pentaho can be customized to meet your needs. The flexibility of PDI means that you can explore, process, transform, export, and visualize data in a variety of ways.
- Check out the [DI Development](#) workflow.

Develop Your DI Solution

This workflow helps you to set up and configure the DI development and test environments, then build, test, and tune your Pentaho DI Solution prototype. This process is similar to the Trial Download Evaluation experience, except that you will be completely configuring the DI server and working with your own ETL developers. If you need extra help, [Pentaho professional services](#) is available. The end result of this is to learn DI implementation best practices and deploy your DI solution to a production server. Most development and testing for DI occurs in Spoon.

Before you begin developing your DI solution, we recommend that you attend [Pentaho training classes](#) to learn how to install and configure the DI server, as well as how to develop data models.

This section is grouped into five parts that guide you along the happy path for the development of your DI solution.

- [Design Solution](#)
- [Set Up Development and Test Environments](#)
- [Build and Test Solution](#)
- [Tune Solution](#)
- [Test Solution Quality](#)

These parts are iterative and you might bounce between them during development. For example as you tune a job, you might find that although you have built a solution that produces the right results, it takes a long time to run. So, you might need to rebuild and test a transformation to improve efficiency, and then retest it.

Design DI Solution

Design helps you think critically about the problem you want to solve and possible solutions. Consider these questions as you gather your requirements and design the solution.

- **Output:** What does the overall solution look like? What questions are posing and how do you want the answers formatted?
- **Data Sources:** What type(s) of data sources are you querying? Where are they located? How much data do you need to process? Are you using Big Data? Are you using relational or non-relational data sources? Will you have a target data source? If so, where are they located?
- **Content/Processing:** What data quality issues do you have? How is the input data mapped to the output data? Where do you want to process the content, in PDI or in the data source? How hardware will you include in your development environment? Will you need one or more quality assurance test environments or production environments?

Also, consider templates or standards, naming conventions, and other requirements of your end users if you have them. Consider how you will back up your data as well.

Set Up Development Environment

Setting up the environment includes installing and configuring PDI on development computers, configuring clustering if needed, and connecting to data sources. If you have one or more quality assurance environments, you will need to set those up also.

Table 1. PDI Set Up Checklist

Task	Do This	Objective
Verify System Requirements	<ul style="list-style-type: none">• Components Reference• JDBC Drivers Reference	<ul style="list-style-type: none">• Acquire one or more servers that meet the requirements.• Obtain the correct drivers for your system.
Obtain Software and Install PDI	<ul style="list-style-type: none">• Install Pentaho PDI• Start the DI Server• Start Spoon	<ul style="list-style-type: none">• Get the software from your Sales Support representative.• Install the software.• Start the DI Server and Spoon.
Install licenses the DI Server	<ul style="list-style-type: none">• Configure the Licenses	<ul style="list-style-type: none">• Add all relevant Pentaho licenses.
Connect to the DI Repository	<ul style="list-style-type: none">• Connect to the DI Repository	<ul style="list-style-type: none">• Connect to the DI Repository.
Apply Advanced Security (if needed)	<ul style="list-style-type: none">• Review DI Advanced Security documentation	<ul style="list-style-type: none">• Determine whether you need to apply DI Advanced Security.

Build and Test Solution

During this step, you develop transformations, jobs, and models, then test what you have developed. You will tune the transformations, jobs, and models for optimal performance.

Development occurs in the Spoon design tool. Spoon's streamlined design tightly couples the build and test activities so that you can easily perform them iteratively. Spoon has perspectives help you perform ETL and visualize data. Spoon also provides a scheduling perspective that can be used to automate testing. Testing encompasses verifying the quality of transformations and jobs, reviewing visualizations, and debugging issues. One common method of testing is to include steps in a transformation or job that calculate hash totals, checksums, record counts, and so forth to determine whether data is being properly processed. You can also visualize your data in analyzer and report designer and review the results as you develop. This can not only

help you find errors and issues with processing, but can help you get a jump on user acceptance testing if you show these reports to your customers or business analysts to get early feedback.

One basic question, is how to determine the numbers of transformations and jobs needed, as well as the order in which they should be executed. A good rule of thumb is to create one transformation for each combination of source system and target tables. You can often identify combinations in your mapping documents. Once you've identified the number of transformations that you need, you can use the same process to determine that number of jobs that you need. When considering the order of execution for transformations and jobs, consider how referential integrity is enforced. Run target table transformations that have no dependencies first, then run transformations that are depend on those tables next, and so forth.

Table 2. Build and Test Checklist - Spoon

Task	Do This	Objective
Understand the Basics	<ul style="list-style-type: none">• Overview of Process	<ul style="list-style-type: none">• Review information about the process and perspectives.
Review most often used steps and entries	<ul style="list-style-type: none">• Review commonly-used steps and entries.	<ul style="list-style-type: none">• Review available transformations and determine how you can use them for your solution.• Review job step references to identify which steps can be used in your solution.
Create and Run Transformations	<ul style="list-style-type: none">• Creating and running a transformation.	<ul style="list-style-type: none">• Identify the transformations needed for your job and implement them.• Save transformation.• Run transformations locally.
Create and Run a Job	<ul style="list-style-type: none">• Creating and Running a Job	<ul style="list-style-type: none">• Create a job.• Arrange transformations in a job so that they execute logically.• Run a job.
Use Agile BI to test the models and visualize the report	<ul style="list-style-type: none">• Prototype with Data Integration	<ul style="list-style-type: none">• Visualize transformations• Create prototypes• Model schema

You can also choose to primarily use Instaview, which is a perspective in Spoon, to build and test solutions.

Table 3. Build and Test Checklist - Instaview

Task	Do This	Objective
Connect to Data Source	<ul style="list-style-type: none"> • Connect to a Data Source 	<ul style="list-style-type: none"> • Connect to Data Source
Create an Analyzer Report	<ul style="list-style-type: none"> • Create Analyzer Reports 	<ul style="list-style-type: none"> • Create an Analyzer Report.
Apply Filters	<ul style="list-style-type: none"> • Apply Filters 	<ul style="list-style-type: none"> • Filters are applied.
Create Visualizations	<ul style="list-style-type: none"> • Create Visualizations 	<ul style="list-style-type: none"> • Visualizations are created.

Tune Solution

Fine tune transformations and jobs to optimize performance. This involves using various tools such as the DI Operation and Audit Mart to determine where bottlenecks or other performance issues occur, and addressing them.

Table 4. Tune Checklist

Task	Do This	Objective
Review the Performance Tuning Checklist and Make Changes to Transformations and Jobs	<ul style="list-style-type: none"> • Tuning tips 	<ul style="list-style-type: none"> • Get familiar with things that you can do to optimize performance. • Apply tuning tips as needed.
Review audit and other logs	<ul style="list-style-type: none"> • Read about Pentaho Operations Mart 	<ul style="list-style-type: none"> • Learn how to use Pentaho Operations Mart to view audit logs.
Consider other performance tuning options	<ul style="list-style-type: none"> • Transactional Databases • Using Logs 	<ul style="list-style-type: none"> • Learn how to apply transactional databases. • Learn how to use logs to tune transformations and jobs.

Next Steps

- Prepare to [Go Live for Production](#).

Go Live for Production - DI

Go Live is the process by which you migrate a prototype to production. This process is divided into four parts.

- [Set Up Development and Test Environments](#)
- [Deploy Solution](#)
- [Secure and Tune Solution](#)
- [Schedule Runs](#)

Set Up Production Environment

Setting up the environment includes installing the software on production computers, configuring clustering, and connecting to data sources. To set up the environment, install and configure the DI Server, Spoon, and any plugins required. Then set up data sources and clusters.

Table 1. Set up Production Environment Checklist

Task	Do This	Objective
Verify System Requirements	<ul style="list-style-type: none">• Components Reference• JDBC Drivers Reference	<ul style="list-style-type: none">• Acquire one or more servers that meet the requirements.• Obtain the correct drivers for your system.
Obtain Software and Install PDI	<ul style="list-style-type: none">• Install Pentaho PDI• Download and Install the Latest Service Pack• Start the DI Server• Start Spoon• Configure the Licenses (if necessary)	<ul style="list-style-type: none">• Get the software from your Sales Support representative.• Install the software - we recommend using the installation wizard.
Change the DI Server Fully Qualified URL	<ul style="list-style-type: none">• Change the URL	<ul style="list-style-type: none">• Change the server's URL so that you do not have a conflict.
Connect to the DI Repository	<ul style="list-style-type: none">• Connect to the DI Repository	<ul style="list-style-type: none">• Connect to the DI Repository
Set up Clusters	<ul style="list-style-type: none">• Set up Clusters (optional step)	<ul style="list-style-type: none">• Become familiar with clustering.• Set up clusters, if they are needed in your environment.

Task	Do This	Objective
Migrate Configuration Files	Copy shared.xml, repositories.xml, kettle.properties, and jar files from the development environment to the production environment.	<ul style="list-style-type: none"> • System is set up and read for production.

Deploy Solution

Export solutions from the DI Repository that is in the development or test environments, to the DI Repository that is in the production environment.

Table 2. Move Solution Checklist

Task	Do This	Objective
Export and Import DI Repository	<ul style="list-style-type: none"> • Export and Import DI Repository Content 	<ul style="list-style-type: none"> • Export DI Repository content from test environment • Import DI Repository content to production environment

Tune Solution

Fine tune transformations and jobs to optimize performance. This involves using various tools such as the DI Operations and Audit Marts to determine where bottlenecks or other performance issues occur, and attempting to address them.

Table 3. Tune Checklist

Task	Do This	Objective
Review the Performance Tuning Checklist and Make Changes to Transformations and Jobs	<ul style="list-style-type: none"> • Tuning tips 	<ul style="list-style-type: none"> • Get familiar with things that you can do to optimize performance. • Apply tuning tips as needed.
Review audit and other logs	<ul style="list-style-type: none"> • Read about Pentaho Operations Mart 	<ul style="list-style-type: none"> • Learn how to use Pentaho Operations Mart to view audit logs.
Consider other performance tuning options	<ul style="list-style-type: none"> • Transactional Databases • Using Logs 	<ul style="list-style-type: none"> • Learn how to apply transactional databases. • Learn how to use logs to tune transformations and jobs.

Schedule Runs

Use Spoon, Pan, or Kitchen to schedule executions of jobs and transformations.

Table 4. Schedule Runs Checklist

Task	Do This	Objective
Schedule Transformations and Jobs From Spoon	<ul style="list-style-type: none">• Schedule Transformations and Jobs	<ul style="list-style-type: none">• Schedule Transformations and Jobs
Command Line Scripting Through Pan and Kitchen	<ul style="list-style-type: none">• Command Line Scripting	<ul style="list-style-type: none">• Schedule Pan and Kitchen Jobs

Commonly-Used PDI Steps and Entries

Although there are over 330 transformation steps and job entries, some steps and entries are used more often than others. If you are creating a transformation and job, but do not know where to begin, this list might be helpful to you.

Top Ten Transformation Steps

PDI Transformation steps are documented on the [PDI Wiki](#).

- Text File Input
- Table Input
- Excel Input
- Text File Output
- Table Output
- Excel Output
- Select Values
- Filter Rows
- Group By
- Stream Lookup

Other Commonly-Used Transformation Steps

PDI Transformation steps are documented on the [PDI Wiki](#).

- **INPUT:** Generate Rows, Data Grid, Get Data from XML, CSV File Input, Fixed File Input
- **OUTPUT:** XML Output
- **TRANSFORM:** Split Fields, Calculator, Add Constants, Add Sequence, Replacing Strings, Split Fields, Sort Rows, String Operations, Strings Cut
- **SCRIPTING:** User Defined Java Class, Modified Java Script Value, User Defined Java Expression
- **FLOW:** Abort, Append Streams, Block this step until steps finish, Blocking Step, Detect Empty Stream, Dummy, ETL Metadata Injection, Filter Rows, Identify Last Row in a Stream, Java Filter, Job Executor, Prioritize Streams, Single Threader, Switch/Case, Transformation Executor
- **LOOKUP**
- **JOINS:** Join Rows, Merge Join
- **JOB:** Get Variables, Set Variables

Commonly-Used Job Entries

PDI Job entries are documented the [PDI Wiki](#).

- **GENERAL:** Start, Job, Transformation, Success
- **UTILITY:** Abort
- **MAIL:** Mail
- **FILE MANAGEMENT:** Add filenames to result, Compare folders, Convert file between Windows and Unix, Copy Files, Create a folder, Create file, Delete file, Delete filenames from result, Delete files, Delete folders, File Compare, HTTP, Move Files, Process result filenames, Unzip file, Wait for file, Write to file, Zip file
- **UTILITIES:** Write to log

Pentaho Business Analytics

Pentaho Business Analytics is a combined business analytics and data integration platform that allows business users, data scientists, and IT administrators to easily access, explore, and visualize their data. Pentaho empowers business users to make information-driven decisions that positively impact their organization's performance, data scientists to use a full-spectrum of tools to create robust data models, and IT to rapidly deliver a secure, scalable, flexible, and easy to manage business analytics platform for the broadest set of users.

- [Evaluate and Learn Pentaho Business Analytics](#)
- [Develop Your BA Environment](#)
- [Go Live for Production - BA](#)

Evaluate and Learn Pentaho Business Analytics

As you explore Pentaho Business Analytics, you will be introduced to the major components, watch videos, work through hands-on examples, and learn about the different features. Go at your own pace. Feel free to dig into the documentation or to contact Pentaho [sales support](#) if you have questions.

This section is divided into four parts that will help you get familiar with Business Analytics.

Tour the User Console and Create Your First Reports

The **User Console** is a fun, easy to use, web-based design environment where you can analyze data, create interactive reports, dashboard reports, and build integrated dashboards to share business intelligence solutions with others in your organization and on the internet. In addition to its design features, the User Console offers a wide variety of system administration features for configuring the Business Analytics (BA) Server, maintaining Pentaho licenses, setting up security, managing report scheduling, and tailoring system performance to meet your requirements.

If you have installed the trial download on your laptop or desktop machine, you are ready to get started exploring. If you have the software installed on a server, and want to use your machine to point to it, check out the section on [setting up your BA Server](#).

Table 1. Explore and Learn Pentaho Reporting Capabilities

Lesson	Watch, Read, and Do This	Notes
Tour the User Console	<ul style="list-style-type: none">• Pentaho User Console Overview• Tour the User Console	<ul style="list-style-type: none">• Understand the features of the User Console• View the sample reports on the Samples tab of the Getting Started section
Create Your First Reports and Dashboards	<ul style="list-style-type: none">• Creating an Interactive Report• Creating an Analysis Report• Creating a Dashboard• Learn About Analysis, Interactive Reports, and Dashboards	<ul style="list-style-type: none">• Created and saved an Interactive Report• Created and saved an Analysis Report• Created and saved a custom Dashboard

Lesson	Watch, Read, and Do This	Notes
Schedule Your Report	<ul style="list-style-type: none"> • Schedule Reports 	<ul style="list-style-type: none"> • Scheduled a report to run and email automatically. • Received your report through email after the schedule runs.

Explore and Learn Data Source Basics

If you have already worked with the Steel Wheels sample data, and want to learn how to create your own data sources and data models with Pentaho, use the **Data Source Wizard**. The Data Source Wizard helps you define a data source that contains the data you want to use and guides you through the creation of your evaluation data model for use in creating reports.

After you define a data source, you can make it available to other evaluators so they can create reports and analysis by simply picking the data source from the data source list. Any number of reports can be created using a single data source.

Table 2. Explore and Learn Data Sources Basics

Lesson	Watch, Read, and Do This	Notes
Create Your First Data Source	<ul style="list-style-type: none"> • Creating a Data Source • Get Started with Data Models • Tour the Data Source Wizard 	<ul style="list-style-type: none"> • Understand how the BA Server and Data Source Wizard work together to create usable data sources and data models. • Explore the Data Source Wizard interface. • Learn the basics of creating a data source using the Data Source Wizard.
Choose Data Source Types	<ul style="list-style-type: none"> • Choose a Data Source Type 	<ul style="list-style-type: none"> • Learn about the different data source types supported by the Data Source Wizard. <p>We recommend using a CSV data source for evaluation.</p>
Create Your First CSV Data Source	<ul style="list-style-type: none"> • Create CSV Data Source 	<ul style="list-style-type: none"> • Import a CSV data file using the Data Source Wizard. • Create the CSV data source. <p>We recommend creating a report using this new CSV data source, then refining the data model with the Data Source Model Editor as needed.</p>
Refine Your Data Source Model	<ul style="list-style-type: none"> • Edit Data Source Models 	<p>(Optional) Edit your evaluation data source model using the Data Source Model Editor.</p>

Lesson	Watch, Read, and Do This	Notes
Learn about Streamlined Data Refinery	<ul style="list-style-type: none"> Learn how to work with Streamlined Data Refinery. 	<ul style="list-style-type: none"> Learn how streamlined data refinery works.

Learn About Report Designer

Pentaho Report Designer is a report creation tool that you can use by itself, or as part of the BA suite. It allows professionals to create print-quality reports based on data from virtually any type of data source.

These resources will help you get familiar with the Report Designer interface, and guide you through the creation and publishing of a print-quality report.

Table 3. Explore and Learn about Report Designer

Lesson	Watch, Read, or Do This	Notes
Explore the Report Designer Interface	<ul style="list-style-type: none"> Explore Report Designer 	<ul style="list-style-type: none"> Tour the Report Designer interface before you begin building reports.
Report Designer Workflow Overview	<ul style="list-style-type: none"> Report Designer Workflow 	<ul style="list-style-type: none"> Look over the work flow concepts for Report Designer.
Create Your First Report	<ul style="list-style-type: none"> Create Your First Print-Quality Reports 	<ul style="list-style-type: none"> Create a report. Add a chart and parameters to your report. View and then publish your report.
Refine the Look of Your Report	<ul style="list-style-type: none"> Design Print-Quality Reports 	<ul style="list-style-type: none"> Explore more advanced features of Report Designer, beginning with report elements.
Add a PDI Data Source	<ul style="list-style-type: none"> Adding a PDI Data Source 	<ul style="list-style-type: none"> Add a PDI data source and use it to create a report in Report Designer.

Discover More About Pentaho Business Analytics

- [Create Analysis, Interactive Reports, and Dashboards](#) provides in-depth details about creating eye-catching business intelligence deliverables for your user community.
- The [Build Full-Featured Solutions](#) section of our documentation contains a large number of topics on working with the BA Suite. Here you can learn about defining data models, fine-tuning your reports, how to create print-quality reports, or work with Big Data.
- If you are a system administrator, check out the [Configure Pentaho Systems](#) and [Administer BA Server](#) sections. Both have details on configuring and administering your BA Server using the User Console, as well as a section on the variety of things you can do to maintain your server manually.

Next Steps

- Contact [Pentaho](#) to learn more about how Business Analytics can be tailored to meet your business needs.
- Learn how to [Develop Your BA Environment](#).

Develop Your BA Environment

The **Develop Your BA Environment** workflow outlines the processes that you use to set up and configure your BA development server, and to build, change, and test Pentaho solutions until they meet your unique business requirements. This process is similar to the **Trial Download Evaluation** experience, except that you will be completely configuring the BA Server and working with your own report designers and data scientists, as well as [Pentaho professional services](#). The end result of this section is to learn BA implementation best practices and deploy your BA solution to a production server.

Before you begin developing your BA environment, we recommend that you attend [Pentaho training classes](#) to learn how to install and configure the BA server, as well as how to develop data models and applications.

This section is grouped into five parts that guide you along the happy path for the development of your BA environment.

Set Up Your BA Server

This table helps you verify that the hardware and software requirements are met, install the software, and configure the BA Server and BA Design tools.

Table 1. BA Server Set-Up Checklist

Checklist	Steps	Notes
Verify System Requirements	<ul style="list-style-type: none">• Components Reference• JDBC Drivers Reference	In preparation for this task, make sure that you: <ul style="list-style-type: none">• Acquire one or more servers that meet the requirements.• Obtain the correct drivers for your system.
Obtain Software and Install BA	<ul style="list-style-type: none">• Get the software• Install Pentaho BA Suite• Access the User Console	After you have the hardware, you will need to: <ul style="list-style-type: none">• Get the software from your Sales Support representative.• Install the software - we recommend using the installation wizard.• Access the User Console, toured the Administration page, and changed the default administrator password.
Change the BA Server Fully Qualified URL	<ul style="list-style-type: none">• Change the BA Server Fully Qualified URL	If you are going to have many machines pointing to one server, make sure you change the fully qualified URL.

Checklist	Steps	Notes
		<ul style="list-style-type: none"> • Change the URL and verified that you can connect to it.
Configure the BA Server	<ul style="list-style-type: none"> • Manage License Keys • Specify Data Connections for BA Server • Set Up an Email Server • Managing Schedules 	<p>After you have the software installed and log in to the console for the first time, you will need to:</p> <ul style="list-style-type: none"> • Add all relevant Pentaho licenses. • Set up your data connections. • Configure an email server through the Administration page. • Learn how to manage schedules through the Schedules page. These will be set up during the production phase.
Configure BA Design Tools	<ul style="list-style-type: none"> • Configure the BA Design Tools • Specify Data Connections for BA Tools 	<p>If you plan to use BA Design Tools, make sure you do these tasks. Otherwise, skip this step.</p> <ul style="list-style-type: none"> • Configure your BA Design Tools. • Specify the data connections required for each tool. <p>We recommend that you NOT configure these tools on your production server. Rather, configure them on your Dev server, test them, then migrate them over to production.</p>

Import Data Sources and Data Models

Follow the steps in the Data Sources and Data Models Checklist to create data sources and models that are ideal for agile development of your BA solution.

Table 2. Data Sources and Data Models Checklist

Task	Steps	Notes
Choose Data Source Types	<ul style="list-style-type: none"> • Choose a Data Source Type 	<ul style="list-style-type: none"> • Review the information about Data Sources. • Examine the differences between the different data models - relational data models vs. multidimensional data models.
Create Data Sources and Models	<ul style="list-style-type: none"> • Creating a Data Source • Get Started with Data Models • Tour the Data Source Wizard 	<ul style="list-style-type: none"> • Learn how the BA Server and Data Source Wizard work together to create usable data sources and data models.
Create Database Table Data Sources	<ul style="list-style-type: none"> • Create Database Table Data Sources 	<ul style="list-style-type: none"> • Create Data Sources and preliminary data models using the Data Source Wizard.

Create Reports and Further Refine Data Models

We recommend that you work with data scientists and business analysts at this stage of the process to achieve the best results. As you get ready to move to production, use data sources created using either [Schema Workbench](#) or [Metadata Editor](#).

Table 3. Create Reports and Refine Data Models Checklist

Task	Steps	Notes
Create Analysis Reports, Interactive Reports, and Dashboards	<ul style="list-style-type: none">• Learn About Analysis, Interactive Reports, and Dashboards	<ul style="list-style-type: none">• Create Interactive and Analysis reports, as well as a Dashboard.• Verify that the reports show the data that you want.• If they do not, work with data scientists to get the results you need.
Create a Report with Report Designer	<ul style="list-style-type: none">• Create Your First Print-Quality Reports	<ul style="list-style-type: none">• (Optional) Create a report in Report Designer.
Refine Your Data Source Model	<ul style="list-style-type: none">• Edit Data Source Models with the Data Source Model Editor• Use Schema Workbench to Refine Multidimensional Data Models• Use Metadata Editor to Refine Relational Data Models	<ul style="list-style-type: none">• Refine data models to get the data that you want to see in your reports.• Create the reports again to verify the results.• Repeat this process as needed.

Test Environment Quality

If you have decided to perform quality assurance testing, [upload your content](#) to the BA repository, then download it to the QA server. Some organizations have an additional layer of User Acceptance Testing that is performed after QA testing is complete.

Next Steps

- Investigate [Security](#) - decide which security solution works best for your business environment.
- Learn about [Scheduling](#) - decide which schedules to set up on your production server.
- Think about what [content you want to move](#) to your production server.
- Prepare to [Go Live for Production](#).

Go Live for Production - BA

This section explains the process for moving Pentaho content and server settings among multiple BA Server instances. Typically this includes either two or three separate servers with identical configurations: One for BA content development, one for testing and QA if desired, and one for production. We recommend that you work with a member of [Pentaho professional services](#) to help with the preparation work for deployment to production.

Prepare for Going Live

This section is divided into two parts; the first part serves as a checklist for setting up a BA Server, and the second consists of prerequisites that you must perform before going live with your production server. If you already have your BA production server ready to go, skip ahead to the tasks listed in the [Prerequisites for Go Live - Checklist for Production](#) Server table.

The **BA Server Set Up Checklist** helps you verify that the hardware and software requirements are met, get the software installed, and configure the BA Server and BA Design tools.

Table 1. BA Server Set Up Checklist

Task	Do This	Notes
Verify System Requirements	<ul style="list-style-type: none">• Components Reference• JDBC Drivers Reference	<ul style="list-style-type: none">• Acquired one or more servers that meet the requirements.• Obtained the correct drivers for your system.
Obtain Software and Install BA	<ul style="list-style-type: none">• Install Pentaho BA Suite• Download and Install the Latest Service Pack• Access the User Console	<ul style="list-style-type: none">• Installed the software.• Installed the latest service pack.• Accessed the User Console, toured the Administration page, and changed the default administrator password. <p>If you need to, see the next segment to change the fully qualified URL for the BA Server.</p>
Change the BA Server Fully Qualified URL	<ul style="list-style-type: none">• Change the BA Server Fully Qualified URL	<p>If you are going to have many machines pointing to one server, make sure you change the fully qualified URL.</p> <ul style="list-style-type: none">• Changed the URL and verified that you can connect to it.

Task	Do This	Notes
Configure the BA Server	<ul style="list-style-type: none"> • Manage License Keys • Specify Data Connections for BA Server • Set Up an Email Server 	<ul style="list-style-type: none"> • Added all relevant Pentaho licenses. • Set up your data connections. • Configured an email server through the Administration page.

This checklist guides you through the tasks you need to do to ensure a smooth transition from a development server to a live production server.

Table 2. Prerequisites for Go Live - Checklist for Production Server

Task	Do This	Notes
Perform a diff on BA config files	<ul style="list-style-type: none"> • Perform a Diff on BA Configuration Files • Verify and Increase Memory Settings 	<ul style="list-style-type: none"> • Differences between configuration files are identified. • Unified properties file is committed to version control. • Memory settings have been increased as applicable.
Verify Data Sources	<ul style="list-style-type: none"> • Specify Data Connections for BA Server • Define JNDI Connections for the BA Server 	<ul style="list-style-type: none"> • Make sure data sources are able to be promoted. • Established JNDI sources as replacements if needed.
Define Security	Define Security for the BA Server - <ul style="list-style-type: none"> • Use Pentaho Security • Define BA Server Advanced Security 	<ul style="list-style-type: none"> • Security is implemented. • Users, roles, and permissions are defined.
Upload Content	<ul style="list-style-type: none"> • Upload and Download from the BA Repository 	<ul style="list-style-type: none"> • Uploaded files and folders.

Perform a Diff on BA Configuration Files

The most important BA Server configuration settings are stored in the `/server/biserver-ee/pentaho-solutions/system/` directory. There are also a few core settings inside of the Pentaho WAR archive deployed to your application server, though they should not change at all after your initial server setup is complete.

CAUTION:

Do not change the names of any content files, data sources, solution directories, or other file names during the promotion process. Names should be set during solution development and strictly maintained throughout

content promotion. Modifying file names can result in complications with BA content that cannot be immediately detected, which will negatively impact your QA process.

To be absolutely certain that you have all of the BA Server configuration files selected, you should diff the following directories in their entirety:

- /pentaho-solutions/system/
- /WEB-INF/ inside your deployed pentaho.war
- /META-INF/ inside your deployed pentaho.war
- For JBoss deployments, the PentahoHibernate-ds.xml and quartz-ds.xml files in the /server/default/deploy/ directory

NOTE:

There are binaries inside of the plugin directories for Analyzer, Dashboard Designer, Interactive Reporting, and Community Dashboard Framework. While it may be useful to note any binary differences, which would indicate possible version differences among BA Server or plugin deployments, in general you should be most concerned with XML configuration files and properties files. However, if you have done any plugin customization work, you will need to promote those changes as well.

Move Content to Production Server

This checklist is a compilation of best practices for you to use to promote your BA Server settings, data sources, and content.

Before you dive into going from a development server to your live production server, make sure that you have completed all of the tasks in the [Prerequisites](#) checklist.

Task	Do This	Notes
Download Content	<ul style="list-style-type: none">• Upload and Download from the BA Repository	<ul style="list-style-type: none">• All desired content is moved to the production server.
Set Up Schedules and Blockout Times	<ul style="list-style-type: none">• Manage Schedules• Prevent Scheduling by Setting Blockout Times	<ul style="list-style-type: none">• Set up schedules for production server.• Set up blockout times for server maintenance tasks.

Next Steps

These resources will be helpful to you after your Production server is live.

- [Fine-Tune Pentaho Systems](#) - provides guidance on how to maintain and fine-tune your BA Server.
- [Pentaho Training and Education](#)
- [Pentaho Customer Support](#)