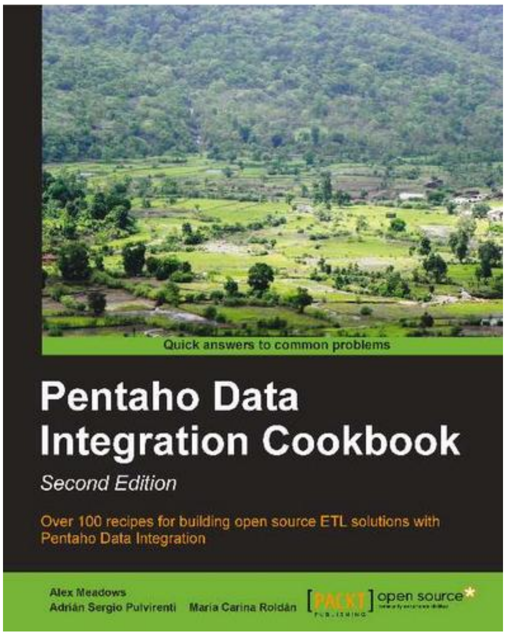
Pentaho Data Integration Cookbook Second Edition

PACKT Publishing

Writers: Alex Meadows, Adrián Sergio Pulvirenti, María Carina Roldán

Paperback: 462

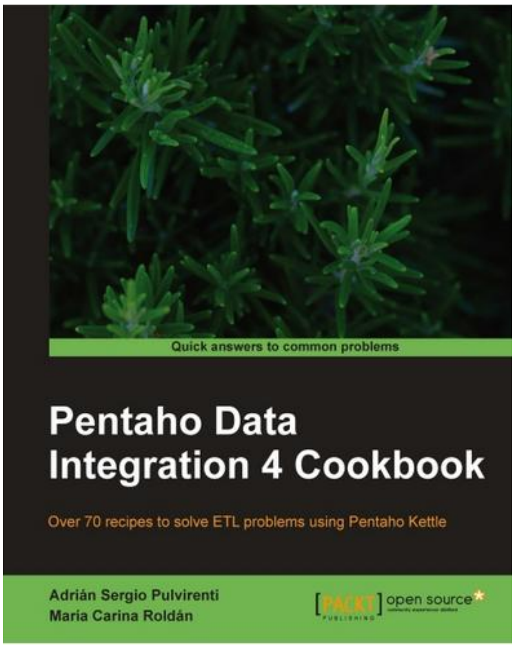
Link to the book page: http://www.packtpub.com/pentaho-data-integration-cookbook-second-edition/book


Quick answers to common problems
Pentaho Data Integration Cookbook
Second Edition
Over 100 recipes for building open source ETL solutions with Pentaho Data Integration
Alex Meadows
Adrián Sergio Pulvirenti
Maria Carina Roldán
[PACKT] PUBLISHING
open source

Pentaho Data Integration or also called Kettle is one of the best open source tool for tasks as extraction, transformation and loading data between different systems. It is integrated within the Pentaho BI suite and covers all necessary to develop and maintain a data warehouse / data mart functionality. Beyond the scope of BI, allows us to deal with and transform data in multiple ways.

This book explains simply and with numerous examples how to get the most out of this tool Pentaho. It is mainly aimed at both developers who have basic knowledge of Kettle, and advanced users who want to know the new possibilities it brings a new version of the tool.

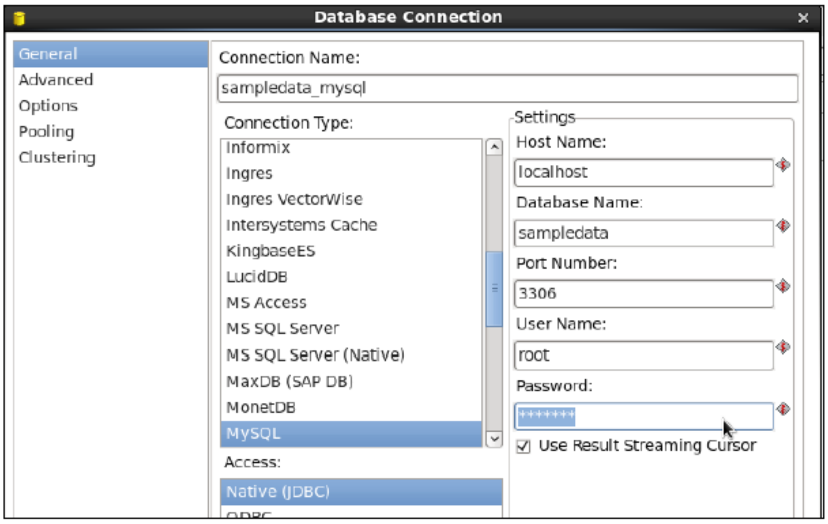
Some years ago, about 2011, one first edition was published with a lot of useful recipes: http://www.packtpub.com/pentaho-data-integration-4-cookbook/book


Quick answers to common problems
Pentaho Data Integration 4 Cookbook
Over 70 recipes to solve ETL problems using Pentaho Kettle
Adrián Sergio Pulvirenti
María Carina Roldán
[PACKT] open source
PUBLISHING

As a general recommendation and for all chapters, I would add a new tip called such "advance trick" where reference is made to some more advanced features. It will always be useful for the user to keep in mind if you handle this case in the future.

* Chapter 1: Working with Databases

This chapter mentions the simplest thing we can do with Kettle: read data from a database. We have 15 recipes, from how to establish a simple connection to database with JDBC or JNDI to more advance recipes like how to build dynamic SQL query options.



Connection Name: sampledata_mysql

Connection Type: MySQL

Access: Native (JDBC)

Host Name: localhost

Database Name: sampledata

Port Number: 3306

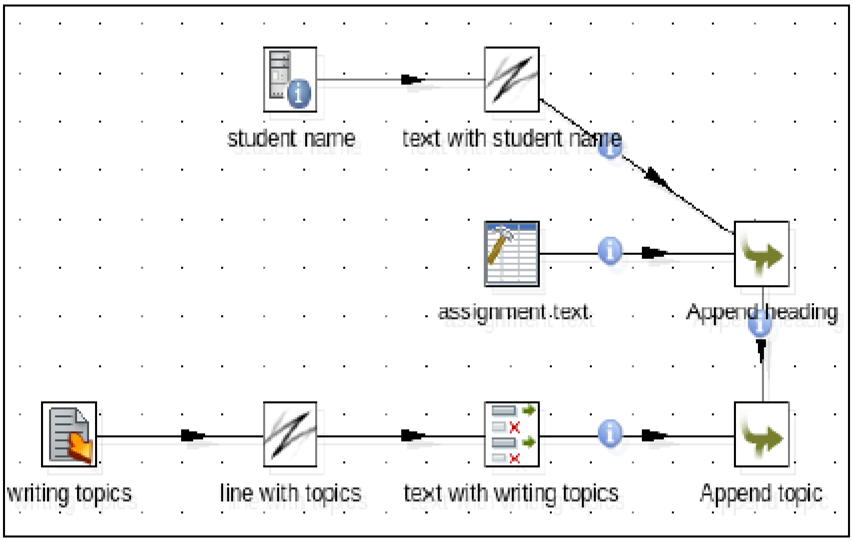
User Name: root

Password: *******

☑ Use Result Streaming Cursor

*Chapter 2: Reading and Writing Files

When we work using PDI, the data source can be very different. In this chapter we have 14 recipes. We will see how to collect and write data in different files .Very interesting recipe that tells how to read data from an instance of Amazon Web Services S3.

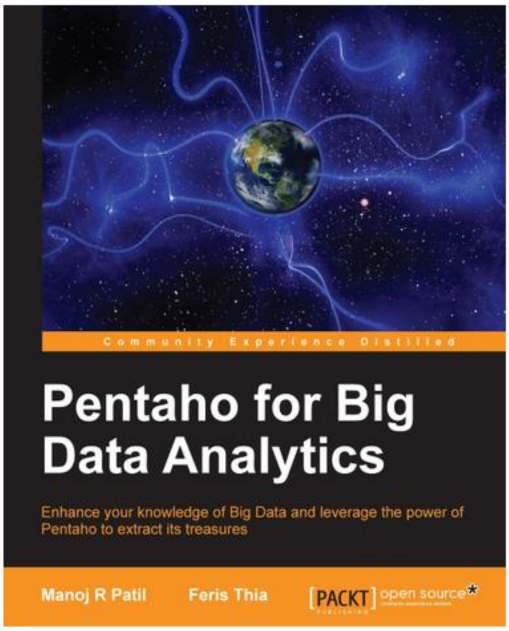


Knowledge of regular expressions is essential to get the most out of this chapter.

*Chapter 3: Working with Big Data and Cloud Sources

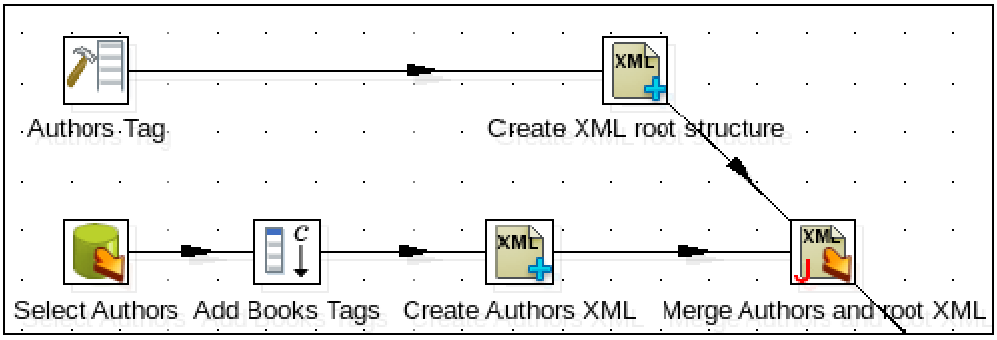
Currently, when we are implementing a data warehouse, we have new possibilities. Depending on the problem to be modeled, we can make use of NOSQL or cloud services like SalesForce.
In this chapter we have 8 recipes on how to get / load data using such technologies.

For those wanting to delve deeper into the world of Big Data and Pentaho, the following book is recommended http://www.packtpub.com/pentaho-for-big-data-analytics/book


Community Experience Distilled
Pentaho for Big Data Analytics
Enhance your knowledge of Big Data and leverage the power of Pentaho to extract its treasures
Manoj R Patil
Feris Thia
[PACKT] open source

*Chapter 4: Manipulating XML Structures

It is very common to find XML files. In this chapter, we have 10 recipes, from reading a single file to validate the contents against a DTD or an XSD schema definition. Generating XML files and reading from an RSS feed generation is also discussed.



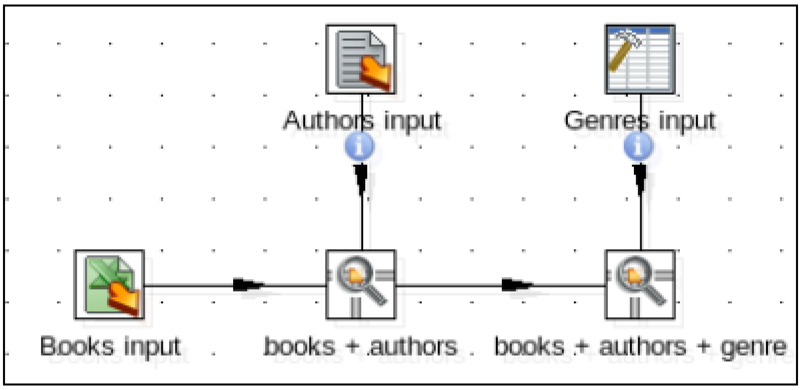
*Chapter 5: File Management

This may be the most interesting chapter to IT staff profiles as systems management. We found 9 recipes that tell how to upload / download files, compare the contents of these, create ZIP files or encrypted files.

Also, as some previous chapter, knowledge of regular expressions will be critical to get the most out of this chapter.

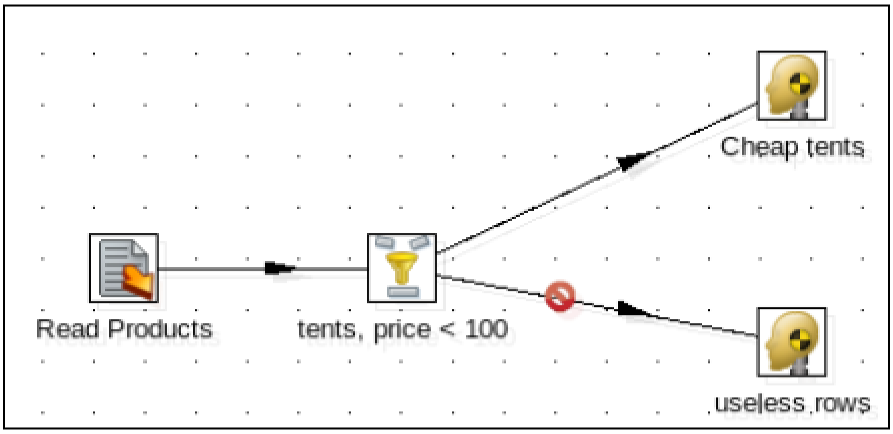
*Chapter 6: Looking for Data

In this chapter we find in 8 recipes and based on various criteria how we can get data from a database, files or web services using PDI.



*Chapter 7: Understanding and Optimizing Data Flows

When we are dealing with data streams, it is common that we find the problem to synchronize that. In this chapter we have 12 recipes to synchronize or redirect our data flows.



*Chapter 8: Executing and Re-using Jobs and Transformations

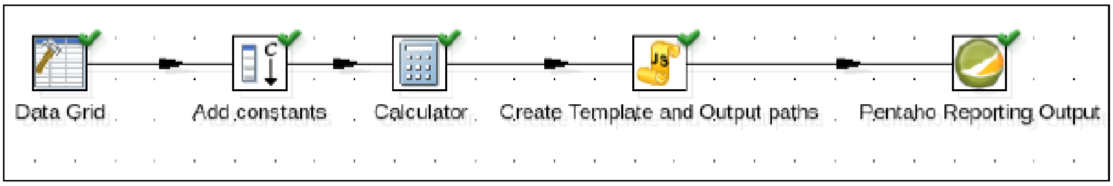
On PDI tool there are 2 ways of structuring our action sequence: jobs and transformations. In this chapter we have 9 recipes to deal with transformations considering parameters and for certain cases.

*Chapter 9: Integrating Kettle and the Pentaho Suite

In my opinion, I find this chapter as the most interesting. It allows you to extract the full potential of Pentaho integrating PDI with the different elements of the suite.

The chapter consists of 6 recipes among which are how to create a report using Pentaho PDI, or how to populate a dashboard created with CDE and using PDI.

In addition to knowledge of PDI, to follow this chapter successfully is required a minimum knowledge of the Pentaho suite, the Design Studio tool and CTools of WebDetails also are recommended http://www.webdetails.pt/ctools.html



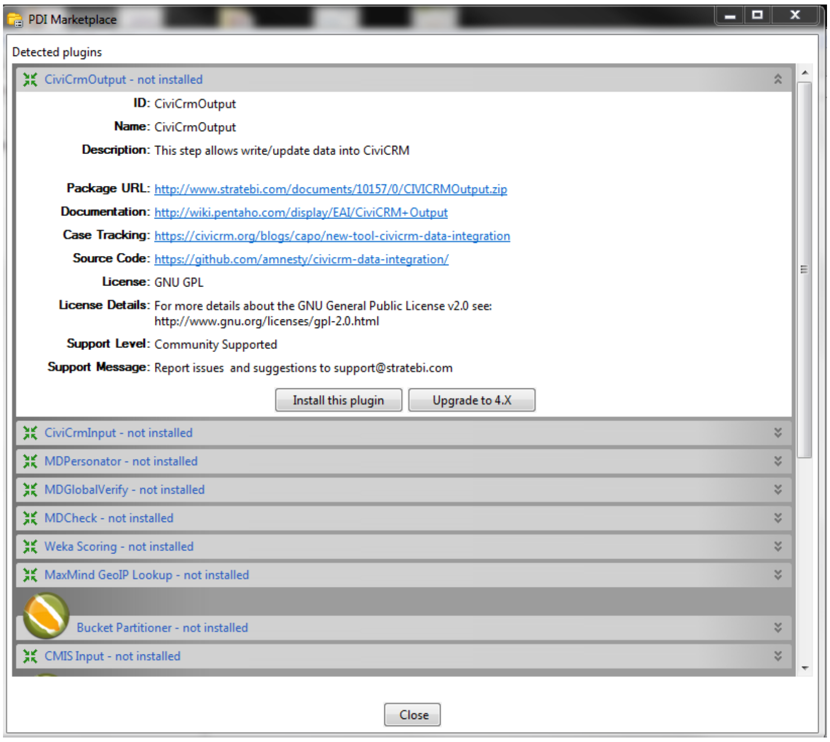
*Chapter 10: Getting the Most Out of Kettle

This chapter contains a variety of recipes that do not fit in any of the other chapters. Specifically, there are 9 recipes, from sending mails with attachments files, processing JSON and one interested recipe for tunning about transformations and jobs.

*Chapter 11: Utilizing Visualization Tools in Kettle

In this chapter there are 4 recipes. One is about adding more functionality to PDI, adding plugins from the MarketPlace . As responsible for maintaining the plugin to extract / load data into CiviCRM using PDI, I can not let pass the opportunity to mention that:

- https://civicrm.org/blogs/capo/new-tool-civicrm-data-integration
- https://github.com/amnesty/civicrm-data-integration



PDI Marketplace

Detected plugins

CiviCrmOutput - not installed

ID: CiviCrmOutput

Name: CiviCrmOutput

Description: This step allows write/update data into CiviCRM

Package URL: http://www.stratebi.com/documents/10157/0/CIVICRMOutput.zip

Documentation: http://wiki.pentaho.com/display/EAI/CiviCRM+Output

Case Tracking: https://civicrm.org/blogs/capo/new-tool-civicrm-data-integration

Source Code: https://github.com/amnesty/civicrm-data-integration/

License: GNU GPL

License Details: For more details about the GNU General Public License v2.0 see:
http://www.gnu.org/licenses/gpl-2.0.html

Support Level: Community Supported

Support Message: Report issues and suggestions to support@stratebi.com

CiviCrmInput - not installed

MDPersonator - not installed

MDGlobalVerify - not installed

MDCheck - not installed

Weka Scoring - not installed

MaxMind GeoIP Lookup - not installed

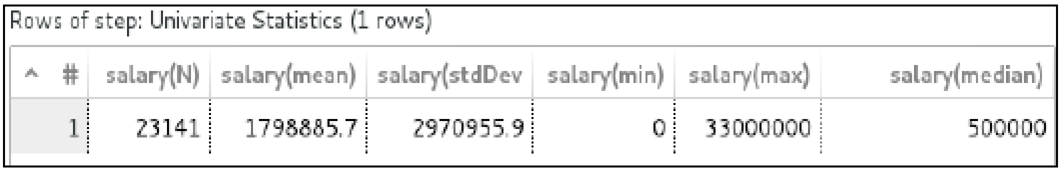
Bucket Partitioner - not installed

CMIS Input - not installed

Also, in this chapter there are other interested recipes as data profiling using PDI and DataCleaner, or display data from our business using AgileBI on a quick way.

*Chapter 12: Data Analytics

The final chapter of the book consists of three interesting recipes on how to obtain information from our data. We will see how to read data from the analytical suite SAS, obtain statistics using PDI steps, and creating a set of random data to the WEKA data mining tool.



Rows of step: Univariate Statistics (1 rows)

#	salary(N)	salary(mean)	salary(stdDev	salary(min)	salary(max)	salary(median)
1	23141	1798885.7	2970955.9	0	33000000	500000