WILEY

Matt Casters
Roland Bouman
Jos van Dongen

Pentaho® Kettle Solutions

Building Open Source ETL Solutions with Pentaho Data Integration

Pentaho® Kettle Solutions

Pentaho® Kettle Solutions

Building Open Source ETL Solutions with Pentaho Data Integration

Matt Casters
Roland Bouman
Jos van Dongen

Wiley Publishing, Inc.

Pentaho® Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration

Published by
Wiley Publishing, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Published simultaneously in Canada

ISBN: 978-0-470-63517-9
ISBN: 9780470942420 (ebk)
ISBN: 9780470947524 (ebk)
ISBN: 9780470947531 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2010932421

For my wife and kids, Kathleen, Sam and Hannelore.
Your love and joy keeps me sane in crazy times.

—Matt

For my wife, Annemarie, and my children, David, Roos,
Anne and Maarten. Thanks for bearing with me—I love you!

—Roland

For my children Thomas and Lisa, and for Yvonne, to whom
I owe more than words can express.

—Jos

About the Authors

Matt Casters has been an independent business intelligence consultant for many years and has implemented numerous data warehouses and BI solutions for large companies. For the last 8 years, Matt kept himself busy with the development of an ETL tool called Kettle. This tool was open sourced in December 2005 and acquired by Pentaho early in 2006. Since then, Matt took up the position of Chief Data Integration at Pentaho. His responsibility is to continue to be lead developer for Kettle. Matt tries to help the Kettle community in any way possible; he answers questions on the forum and speaks occasionally at conferences all around the world. He has a blog at `http://www.ibridge.be` and you can follow his `@mattcasters` account on Twitter.

Roland Bouman has been working in the IT industry since 1998 and is currently working as a web and business intelligence developer. Over the years he has focused on open source software, in particular database technology, business intelligence, and web development frameworks. He's an active member of the MySQL and Pentaho communities, and a regular speaker at international conferences, such as the MySQL User Conference, OSCON and at Pentaho community events. Roland co-authored the *MySQL 5.1. Cluster Certification Guide* and *Pentaho Solutions*, and was a technical reviewer for a number of MySQL and Pentaho related book titles. He maintains a technical blog at `http://rpbouman.blogspot.com` and tweets as `@rolandbouman` on Twitter.

Jos van Dongen is a seasoned business intelligence professional and well-known author and presenter. He has been involved in software development, business intelligence, and data warehousing since 1991. Before starting his own consulting practice, Tholis Consulting, in 1998, he worked for a top tier systems integrator and a leading management consulting firm. Over the past years, he has successfully implemented BI and data warehouse solutions for a variety of organizations, both commercial and non-profit. Jos covers new BI developments for the Dutch *Database Magazine* and speaks regularly at national and international conferences. He authored one book on open source BI and is co-author of the book *Pentaho Solutions*. You can find more information about Jos on `http://www.tholis.com` or follow `@josvandongen` on Twitter.

Credits

Executive Editor
Robert Elliott

Project Editor
Sara Shlaer

Technical Editors
Jens Bleuel
Sven Boden
Kasper de Graaf
Daniel Einspanjer
Nick Goodman
Mark Hall
Samatar Hassan
Benjamin Kallmann
Bryan Senseman
Johannes van den Bosch

Production Editor
Daniel Scribner

Copy Editor
Nancy Rapoport

Editorial Director
Robyn B. Siesky

Editorial Manager
Mary Beth Wakefield

Marketing Manager
Ashley Zurcher

Production Manager
Tim Tate

Vice President and Executive Group Publisher
Richard Swadley

Vice President and Executive Publisher
Barry Pruett

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Lynsey Stanford

Compositor
Maureen Forys,
Happenstance Type-O-Rama

Proofreader
Nancy Bell

Indexer
Robert Swanson

Cover Designer
Ryan Sneed

Acknowledgments

This book is the result of the efforts of many individuals. By convention, authors receive explicit credit, and get to have their names printed on the book cover. But creating this book would not have been possible without a lot of hard work behind the scenes. We, the authors, would like to express our gratitude to a number of people that provided substantial contributions, and thus help define and shape the final result that is *Pentaho Kettle Solutions*.

First, we'd like to thank those individuals that contributed directly to the material that appears in the book:

- Ingo Klose suggested an elegant solution to generate keys starting from a given offset within a single transformation (this solution is discussed in Chapter 8, "Handling Dimension Tables," subsection "Generating Surrogate Keys Based on a Counter," shown in Figure 8-2).
- Samatar Hassan provided text as well as working example transformations to demonstrate Kettle's RSS capabilities. Samatar's contribution is included almost completely and appears in the RSS section of Chapter 21, "Web Services."
- Thanks to Mike Hillyer and the MySQL documentation team for creating and maintaining the Sakila sample database, which is introduced in Chapter 4 and appears in many examples throughout this book.
- Although only three authors appear on the cover, there was actually a fourth one: We cannot thank Kasper de Graaf of DIKW-Academy enough for writing the Data Vault chapter, which has benefited greatly from his deep expertise on this subject. Special thanks also to Johannes van den Bosch who did a great job reviewing Kasper's work and gave another boost to the overall quality and clarity of the chapter.
- Thanks to Bernd Aschauer and Robert Wintner, both from Aschauer EDV (`http://www.aschauer-edv.at/en`), for providing the examples and screenshots used in the section dedicated to SAP of Chapter 6, "Data Extraction."
- Daniel Einspanjer of the Mozilla Foundation provided sample transformations for Chapter 7, "Cleansing and Conforming."

Thanks for your contributions. This book benefited substantially from your efforts.

Much gratitude goes out to all of our technical reviewers. Providing a good technical review is hard and time-consuming, and we have been very lucky to find a collection of such talented and seasoned Pentaho and Kettle experts willing to find some time in their busy schedules to provide us with the kind of quality review required to write a book of this size and scope.

We'd like to thank the Kettle and Pentaho communities. During and before the writing of this book, individuals from these communities provided valuable suggestions and ideas to all three authors for topics to cover in a book that focuses on ETL, data integration, and Kettle. We hope this book will be useful and practical for everybody who is using or planning to use Kettle. Whether we succeeded is up to the reader, but if we did, we have to thank individuals in the Kettle and Pentaho communities for helping us achieve it.

We owe many thanks to all contributors and developers of the Kettle software project. The authors are all enthusiastic users of Kettle: we love it, because it solves our daily data integration problems in a straightforward and efficient manner without getting in the way. Kettle is a joy to work with, and this is what provided much of the drive to write this book.

Finally, we'd like to thank our publisher, Wiley, for giving us the opportunity to write this book, and for the excellent support and management from their end. In particular, we'd like to thank our Project Editor, Sara Shlaer. Despite the often delayed deliveries from our end, Sara always kept her cool and somehow managed to make deadlines work out. Her advice, patience, encouragement, care, and sense of humor made all the difference and form an important contribution to this book. In addition, we'd like to thank our Executive Editor Robert Elliot. We appreciate the trust he put into our small team of authors to do our job, and his efforts to realize *Pentaho Kettle Solutions*.

—The authors

Writing a technical book like the one you are reading right now is very hard to do all by yourself. Because of the extremely busy agenda caused by the release process of Kettle 4, I probably should never have agreed to co-author. It's only thanks to the dedication and professionalism of Jos and Roland that we managed to write this book at all. I thank both friends very much for their invitation to co-author. Even though writing a book is a hard and painful process, working with Jos and Roland made it all worthwhile.

When Kettle was not yet released as open source code it often received a lukewarm reaction. The reason was that nobody was really waiting for yet another closed source ETL tool. Kettle came from that position to being the most widely deployed open source ETL tool in the world. This happened only thanks to the thousands of volunteers who offered to help out with various tasks. Ever since Kettle was open sourced it became a project with an every growing community. It's impossible to thank this community enough. Without the help of the developers, the translators, the testers, the bug reporters, the folks who participate in the forums, the people with the great ideas, and even the folks who like to complain, Kettle would not be where it is today. I would like to especially thank one important member of our community: Pentaho. Pentaho CEO Richard Daley and his team have done an excellent job in supporting the Kettle project ever

since they got involved with it. Without their support it would not have been possible for Kettle to be on the accelerated growth path that it is on today. It's been a pleasure and a privilege to work with the Pentaho crew.

A few select members of our community also picked up the tough job of reviewing the often technical content of this book. The reviewers of my chapters, Nicholas Goodman, Daniel Einspanjer, Bryan Senseman, Jens Bleuel, Samatar Hassan, and Mark Hall had the added disadvantage that this was the first time that I was going through the process of writing a book. It must not have been pretty at times. All the same they spent a lot of time coming up with insightful additions, spot-on advice, and to the point comments. I do enormously appreciate the vast amount of time and effort that they put into the reviewing. The book wouldn't have been the same without you guys!

—Matt Casters

I'd like to thank both my co-authors, Jos and Matt. It's an honor to be working with such knowledgeable and skilled professionals, and I hope we will collaborate again in the future. I feel our different backgrounds and expertise have truly complemented each other and helped us all to cover the many different subjects covered in this book.

I'd also like to thank the reviewers of my chapters: Benjamin Kallman, Bryan Senseman, Daniel Einspanjer, Sven Boden, and Samatar Hassan. Your comments and suggestions made all the difference and I thank you for your frank and constructive criticism.

Finally, I'd like to thank the readers of my blog at `http://rpbouman.blogspot.com/`. I got a lot of inspiration from the comments posted there, and I got a lot of good feedback in response to the blog posts announcing the writing of *Pentaho Kettle Solutions*.

—Roland Bouman

Back in October 2009, when *Pentaho Solutions* had only been on the shelves for two months and Roland and I agreed never to write another book, Bob Elliot approached us asking us to do just that. Yes, we had been discussing some ideas and already concluded that if there were to be another book, it would have to be about Kettle. And this was exactly what Bob asked us to do: write a book about data integration using Kettle. We quickly found out that Matt Casters was not only interested in reviewing, but in actually becoming a full author as well, an offer we gladly accepted. Looking back, I can hardly believe that we pulled it off, considering everything else that was going on in our lives. So many thanks to Roland and Matt for bearing with me, and thank you Bob and especially Sara for your relentless efforts of keeping us on track.

A special thank you is also warranted for Ralph Kimball, whose ideas you'll find throughout this book. Ralph gave us permission to use the Kimball Group's 34 ETL subsystems as the framework for much of the material presented in his book. Ralph also took the time to review Chapter 5, and thanks to his long list of excellent comments the chapter became a perfect foundation for Parts II, III, and IV of the book.

Finally I'd like to thank Daniel Einspanjer, Bryan Senseman, Jens Bleuel, Sven Boden, Samatar Hassan, and Benjamin Kallmann for being an absolute pain in the neck and thus doing a great job as technical reviewers for my chapters. Your comments, questions and suggestions definitely gave a big boost to the overall quality of this book.

—Jos van Dongen

Contents at a Glance

Contents