

 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

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

Copyright © 2010 by Wiley Publishing, Inc., Indianapolis, Indiana

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

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Web site may provide or recommendations it may make. Further, readers should be aware that Internet Web sites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Control Number: 2010932421

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. Pentaho is a registered trademark of Pentaho, Inc. All other trademarks are the property of their respective owners. Wiley Publishing, Inc. is not associated with any product or vendor mentioned in this book.

*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

Composer

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

Introduction		xxxi
Part I	Getting Started	1
Chapter 1	ETL Primer	3
Chapter 2	Kettle Concepts	23
Chapter 3	Installation and Configuration	53
Chapter 4	An Example ETL Solution—Sakila	73
Part II	ETL	111
Chapter 5	ETL Subsystems	113
Chapter 6	Data Extraction	127
Chapter 7	Cleansing and Conforming	167
Chapter 8	Handling Dimension Tables	207
Chapter 9	Loading Fact Tables	245
Chapter 10	Working with OLAP Data	269
Part III	Management and Deployment	293
Chapter 11	ETL Development Lifecycle	295
Chapter 12	Scheduling and Monitoring	321

Chapter 13	Versioning and Migration	341
Chapter 14	Lineage and Auditing	357
Part IV	Performance and Scalability	375
Chapter 15	Performance Tuning	377
Chapter 16	Parallelization, Clustering, and Partitioning	403
Chapter 17	Dynamic Clustering in the Cloud	433
Chapter 18	Real-Time Data Integration	449
Part V	Advanced Topics	463
Chapter 19	Data Vault Management	465
Chapter 20	Handling Complex Data Formats	497
Chapter 21	Web Services	515
Chapter 22	Kettle Integration	569
Chapter 23	Extending Kettle	593
Appendix A	The Kettle Ecosystem	629
Appendix B	Kettle Enterprise Edition Features	635
Appendix C	Built-in Variables and Properties Reference	637
Index		643



Contents

Introduction	xxxi
Part I Getting Started	1
Chapter 1 ETL Primer	3
OLTP versus Data Warehousing	3
What Is ETL?	5
The Evolution of ETL Solutions	5
ETL Building Blocks	7
ETL, ELT, and EII	8
ELT	9
EII: Virtual Data Integration	10
Data Integration Challenges	11
Methodology: Agile BI	12
ETL Design	14
Data Acquisition	14
Beware of Spreadsheets	15
Design for Failure	15
Change Data Capture	16
Data Quality	16
Data Profiling	16
Data Validation	17
ETL Tool Requirements	17
Connectivity	17
Platform Independence	18
Scalability	18
Design Flexibility	19
Reuse	19
Extensibility	19

	Data Transformations	20
	Testing and Debugging	21
	Lineage and Impact Analysis	21
	Logging and Auditing	22
	Summary	22
Chapter 2	Kettle Concepts	23
	Design Principles	23
	The Building Blocks of Kettle Design	25
	Transformations	25
	Steps	26
	Transformation Hops	26
	Parallelism	27
	Rows of Data	27
	Data Conversion	29
	Jobs	30
	Job Entries	31
	Job Hops	31
	Multiple Paths and Backtracking	32
	Parallel Execution	33
	Job Entry Results	34
	Transformation or Job Metadata	36
	Database Connections	37
	Special Options	38
	The Power of the Relational Database	39
	Connections and Transactions	39
	Database Clustering	40
	Tools and Utilities	41
	Repositories	41
	Virtual File Systems	42
	Parameters and Variables	43
	Defining Variables	43
	Named Parameters	44
	Using Variables	44
	Visual Programming	45
	Getting Started	46
	Creating New Steps	47
	Putting It All Together	49
	Summary	51
Chapter 3	Installation and Configuration	53
	Kettle Software Overview	53
	Integrated Development Environment: Spoon	55
	Command-Line Launchers: Kitchen and Pan	57
	Job Server: Carte	57
	Encr.bat and encr.sh	58
	Installation	58

Java Environment	58
Installing Java Manually	58
Using Your Linux Package Management System	59
Installing Kettle	59
Versions and Releases	59
Archive Names and Formats	60
Downloading and Uncompressing	60
Running Kettle Programs	61
Creating a Shortcut Icon or Launcher for Spoon	62
Configuration	63
Configuration Files and the .kettle Directory	63
The Kettle Shell Scripts	69
General Structure of the Startup Scripts	70
Adding an Entry to the Classpath	70
Changing the Maximum Heap Size	71
Managing JDBC Drivers	72
Summary	72
Chapter 4 An Example ETL Solution—Sakila	73
Sakila	73
The Sakila Sample Database	74
DVD Rental Business Process	74
Sakila Database Schema Diagram	75
Sakila Database Subject Areas	75
General Design Considerations	77
Installing the Sakila Sample Database	77
The Rental Star Schema	78
Rental Star Schema Diagram	78
Rental Fact Table	79
Dimension Tables	79
Keys and Change Data Capture	80
Installing the Rental Star Schema	81
Prerequisites and Some Basic Spoon Skills	81
Setting Up the ETL Solution	82
Creating Database Accounts	82
Working with Spoon	82
Opening Transformation and Job Files	82
Opening the Step's Configuration Dialog	83
Examining Streams	83
Running Jobs and Transformations	83
The Sample ETL Solution	84
Static, Generated Dimensions	84
Loading the dim_date Dimension Table	84
Loading the dim_time Dimension Table	86
Recurring Load	87
The load_rentals Job	88

	The load_dim_staff Transformation	91
	Database Connections	91
	The load_dim_customer Transformation	95
	The load_dim_store Transformation	98
	The fetch_address Subtransformation	99
	The load_dim_actor Transformation	101
	The load_dim_film Transformation	102
	The load_fact_rental Transformation	107
	Summary	109
Part II	ETL	111
Chapter 5	ETL Subsystems	113
	Introduction to the 34 Subsystems	114
	Extraction	114
	Subsystems 1–3: Data Profiling, Change Data Capture, and Extraction	115
	Cleaning and Conforming Data	116
	Subsystem 4: Data Cleaning and Quality Screen Handler System	116
	Subsystem 5: Error Event Handler	117
	Subsystem 6: Audit Dimension Assembler	117
	Subsystem 7: Deduplication System	117
	Subsystem 8: Data Conformer	118
	Data Delivery	118
	Subsystem 9: Slowly Changing Dimension Processor	118
	Subsystem 10: Surrogate Key Creation System	119
	Subsystem 11: Hierarchy Dimension Builder	119
	Subsystem 12: Special Dimension Builder	120
	Subsystem 13: Fact Table Loader	121
	Subsystem 14: Surrogate Key Pipeline	121
	Subsystem 15: Multi-Valued Dimension Bridge Table Builder	121
	Subsystem 16: Late-Arriving Data Handler	122
	Subsystem 17: Dimension Manager System	122
	Subsystem 18: Fact Table Provider System	122
	Subsystem 19: Aggregate Builder	123
	Subsystem 20: Multidimensional (OLAP) Cube Builder	123
	Subsystem 21: Data Integration Manager	123
	Managing the ETL Environment	123
	Summary	126
Chapter 6	Data Extraction	127
	Kettle Data Extraction Overview	128
	File-Based Extraction	128
	Working with Text Files	128
	Working with XML files	133
	Special File Types	134

Database-Based Extraction	134
Web-Based Extraction	137
Text-Based Web Extraction	137
HTTP Client	137
Using SOAP	138
Stream-Based and Real-Time Extraction	138
Working with ERP and CRM Systems	138
ERP Challenges	139
Kettle ERP Plugins	140
Working with SAP Data	140
ERP and CDC Issues	146
Data Profiling	146
Using eobjects.org DataCleaner	147
Adding Profile Tasks	149
Adding Database Connections	149
Doing an Initial Profile	151
Working with Regular Expressions	151
Profiling and Exploring Results	152
Validating and Comparing Data	153
Using a Dictionary for Column Dependency Checks	153
Alternative Solutions	154
Text Profiling with Kettle	154
CDC: Change Data Capture	154
Source Data-Based CDC	155
Trigger-Based CDC	157
Snapshot-Based CDC	158
Log-Based CDC	162
Which CDC Alternative Should You Choose?	163
Delivering Data	164
Summary	164
Chapter 7 Cleansing and Conforming	167
Data Cleansing	168
Data-Cleansing Steps	169
Using Reference Tables	172
Conforming Data Using Lookup Tables	172
Conforming Data Using Reference Tables	175
Data Validation	179
Applying Validation Rules	180
Validating Dependency Constraints	183
Error Handling	183
Handling Process Errors	184
Transformation Errors	186
Handling Data (Validation) Errors	187
Auditing Data and Process Quality	191
Deduplicating Data	192

Handling Exact Duplicates	193
The Problem of Non-Exact Duplicates	194
Building Deduplication Transforms	195
Step 1: Fuzzy Match	197
Step 2: Select Suspects	198
Step 3: Lookup Validation Value	198
Step 4: Filter Duplicates	199
Scripting	200
Formula	201
JavaScript	202
User-Defined Java Expressions	202
Regular Expressions	203
Summary	205
Chapter 8 Handling Dimension Tables	207
Managing Keys	208
Managing Business Keys	209
Keys in the Source System	209
Keys in the Data Warehouse	209
Business Keys	209
Storing Business Keys	210
Looking Up Keys with Kettle	210
Generating Surrogate Keys	210
The “Add sequence” Step	211
Working with auto_increment or IDENTITY Columns	217
Keys for Slowly Changing Dimensions	217
Loading Dimension Tables	218
Snowflaked Dimension Tables	218
Top-Down Level-Wise Loading	219
Sakila Snowflake Example	219
Sample Transformation	221
Database Lookup Configuration	222
Sample Job	225
Star Schema Dimension Tables	226
Denormalization	226
Denormalizing to 1NF with the “Database lookup” Step	226
Change Data Capture	227
Slowly Changing Dimensions	228
Types of Slowly Changing Dimensions	228
Type 1 Slowly Changing Dimensions	229
The Insert / Update Step	229
Type 2 Slowly Changing Dimensions	232
The “Dimension lookup / update” Step	232
Other Types of Slowly Changing Dimensions	237
Type 3 Slowly Changing Dimensions	237
Hybrid Slowly Changing Dimensions	238

More Dimensions	239
Generated Dimensions	239
Date and Time Dimensions	239
Generated Mini-Dimensions	239
Junk Dimensions	241
Recursive Hierarchies	242
Summary	243
Chapter 9 Loading Fact Tables	245
Loading in Bulk	246
STDIN and FIFO	247
Kettle Bulk Loaders	248
MySQL Bulk Loading	249
LucidDB Bulk Loader	249
Oracle Bulk Loader	249
PostgreSQL Bulk Loader	250
Table Output Step	250
General Bulk Load Considerations	250
Dimension Lookups	251
Maintaining Referential Integrity	251
The Surrogate Key Pipeline	252
Using In-Memory Lookups	253
Stream Lookups	253
Late-Arriving Data	255
Late-Arriving Facts	256
Late-Arriving Dimensions	256
Fact Table Handling	260
Periodic and Accumulating Snapshots	260
Introducing State-Oriented Fact Tables	261
Loading Periodic Snapshots	263
Loading Accumulating Snapshots	264
Loading State-Oriented Fact Tables	265
Loading Aggregate Tables	266
Summary	267
Chapter 10 Working with OLAP Data	269
OLAP Benefits and Challenges	270
OLAP Storage Types	272
Positioning OLAP	272
Kettle OLAP Options	273
Working with Mondrian	274
Working with XML/A Servers	277
Working with Palo	282
Setting Up the Palo Connection	283
Palo Architecture	284
Reading Palo Data	285
Writing Palo Data	289
Summary	291

Part III	Management and Deployment	293
Chapter 11	ETL Development Lifecycle	295
	Solution Design	295
	Best and Bad Practices	296
	Data Mapping	297
	Naming and Commentary Conventions	298
	Common Pitfalls	299
	ETL Flow Design	300
	Reusability and Maintainability	300
	Agile Development	301
	Testing and Debugging	306
	Test Activities	307
	ETL Testing	308
	Test Data Requirements	308
	Testing for Completeness	309
	Testing Data Transformations	311
	Test Automation and Continuous Integration	311
	Upgrade Tests	312
	Debugging	312
	Documenting the Solution	315
	Why Isn't There Any Documentation?	316
	Myth 1: My Software Is Self-Explanatory	316
	Myth 2: Documentation Is Always Outdated	316
	Myth 3: Who Reads Documentation Anyway?	317
	Kettle Documentation Features	317
	Generating Documentation	319
	Summary	320
Chapter 12	Scheduling and Monitoring	321
	Scheduling	321
	Operating System–Level Scheduling	322
	Executing Kettle Jobs and Transformations from the Command Line	322
	UNIX-Based Systems: cron	326
	Windows: The at utility and the Task Scheduler	327
	Using Pentaho's Built-in Scheduler	327
	Creating an Action Sequence to Run Kettle Jobs and Transformations	328
	Kettle Transformations in Action Sequences	329
	Creating and Maintaining Schedules with the Administration Console	330
	Attaching an Action Sequence to a Schedule	333
	Monitoring	333
	Logging	333
	Inspecting the Log	333

Logging Levels	335
Writing Custom Messages to the Log	336
E-mail Notifications	336
Configuring the Mail Job Entry	337
Summary	340
Chapter 13 Versioning and Migration	341
Version Control Systems	341
File-Based Version Control Systems	342
Organization	342
Leading File-Based VCSs	343
Content Management Systems	344
Kettle Metadata	344
Kettle XML Metadata	345
Transformation XML	345
Job XML	346
Global Replace	347
Kettle Repository Metadata	348
The Kettle Database Repository Type	348
The Kettle File Repository Type	349
The Kettle Enterprise Repository Type	350
Managing Repositories	350
Exporting and Importing Repositories	350
Upgrading Your Repository	351
Version Migration System	352
Managing XML Files	352
Managing Repositories	352
Parameterizing Your Solution	353
Summary	356
Chapter 14 Lineage and Auditing	357
Batch-Level Lineage Extraction	358
Lineage	359
Lineage Information	359
Impact Analysis Information	361
Logging and Operational Metadata	363
Logging Basics	363
Logging Architecture	364
Setting a Maximum Buffer Size	365
Setting a Maximum Log Line Age	365
Log Channels	366
Log Text Capturing in a Job	366
Logging Tables	367
Transformation Logging Tables	367
Job Logging Tables	373
Summary	374

Part IV	Performance and Scalability	375
Chapter 15	Performance Tuning	377
	Transformation Performance: Finding the Weakest Link	377
	Finding Bottlenecks by Simplifying	379
	Finding Bottlenecks by Measuring	380
	Copying Rows of Data	382
	Improving Transformation Performance	384
	Improving Performance in Reading Text Files	384
	Using Lazy Conversion for Reading Text Files	385
	Single-File Parallel Reading	385
	Multi-File Parallel Reading	386
	Configuring the NIO Block Size	386
	Changing Disks and Reading Text Files	386
	Improving Performance in Writing Text Files	387
	Using Lazy Conversion for Writing Text Files	387
	Parallel Files Writing	387
	Changing Disks and Writing Text Files	387
	Improving Database Performance	388
	Avoiding Dynamic SQL	388
	Handling Roundtrips	388
	Handling Relational Databases	390
	Sorting Data	392
	Sorting on the Database	393
	Sorting in Parallel	393
	Reducing CPU Usage	394
	Optimizing the Use of JavaScript	394
	Launching Multiple Copies of a Step	396
	Selecting and Removing Values	397
	Managing Thread Priorities	397
	Adding Static Data to Rows of Data	397
	Limiting the Number of Step Copies	398
	Avoiding Excessive Logging	398
	Improving Job Performance	399
	Loops in Jobs	399
	Database Connection Pools	400
	Summary	401
Chapter 16	Parallelization, Clustering, and Partitioning	403
	Multi-Threading	403
	Row Distribution	404
	Row Merging	405
	Row Redistribution	406
	Data Pipelining	407
	Consequences of Multi-Threading	408
	Database Connections	408

Order of Execution	409
Parallel Execution in a Job	411
Using Carte as a Slave Server	411
The Configuration File	411
Defining Slave Servers	412
Remote Execution	413
Monitoring Slave Servers	413
Carte Security	414
Services	414
Clustering Transformations	417
Defining a Cluster Schema	417
Designing Clustered Transformations	418
Execution and Monitoring	420
Metadata Transformations	421
Rules	422
Data Pipelining	425
Partitioning	425
Defining a Partitioning Schema	425
Objectives of Partitioning	427
Implementing Partitioning	428
Internal Variables	428
Database Partitions	429
Partitioning in a Clustered Transformation	430
Summary	430
Chapter 17 Dynamic Clustering in the Cloud	433
Dynamic Clustering	433
Setting Up a Dynamic Cluster	434
Using the Dynamic Cluster	436
Cloud Computing	437
EC2	438
Getting Started with EC2	438
Costs	438
Customizing an AMI	439
Packaging a New AMI	442
Terminating an AMI	442
Running a Master	442
Running the Slaves	443
Using the EC2 Cluster	444
Monitoring	445
The Lightweight Principle and Persistence Options	446
Summary	447
Chapter 18 Real-Time Data Integration	449
Introduction to Real-Time ETL	449
Real-Time Challenges	450
Requirements	451

Transformation Streaming	452	
A Practical Example of Transformation Streaming	454	
Debugging	457	
Third-Party Software and Real-Time Integration	458	
Java Message Service	459	
Creating a JMS Connection and Session	459	
Consuming Messages	460	
Producing Messages	460	
Closing Shop	460	
Summary	461	
Part V	Advanced Topics	463
Chapter 19	Data Vault Management	465
Introduction to Data Vault Modeling	466	
Do You Need a Data Vault?	466	
Data Vault Building Blocks	467	
Hubs	467	
Links	468	
Satellites	469	
Data Vault Characteristics	471	
Building a Data Vault	471	
Transforming Sakila to the Data Vault Model	472	
Sakila Hubs	472	
Sakila Links	473	
Sakila Satellites	474	
Loading the Data Vault: A Sample ETL Solution	477	
Installing the Sakila Data Vault	477	
Setting Up the ETL Solution	477	
Creating a Database Account	477	
The Sample ETL Data Vault Solution	478	
Sample Hub: hub_actor	478	
Sample Link: link_customer_store	480	
Sample Satellite: sat_actor	483	
Loading the Data Vault Tables	485	
Updating a Data Mart from a Data Vault	486	
The Sample ETL Solution	486	
The dim_actor Transformation	486	
The dim_customer Transformation	488	
The dim_film Transformation	492	
The dim_film_actor_bridge Transformation	492	
The fact_rental Transformation	493	
Loading the Star Schema Tables	495	
Summary	495	

Chapter 20	Handling Complex Data Formats	497
	Non-Relational and Non-Tabular Data Formats	498
	Non-Relational Tabular Formats	498
	Handling Multi-Valued Attributes	498
	Using the Split Field to Rows Step	499
	Handling Repeating Groups	500
	Using the Row Normaliser Step	500
	Semi- and Unstructured Data	501
	Kettle Regular Expression Example	503
	Configuring the Regex Evaluation Step	504
	Verifying the Match	507
	Key/Value Pairs	508
	Kettle Key/Value Pairs Example	509
	Text File Input	509
	Regex Evaluation	510
	Grouping Lines into Records	511
	Denormaliser: Turning Rows into Columns	512
	Summary	513
Chapter 21	Web Services	515
	Web Pages and Web Services	515
	Kettle Web Features	516
	General HTTP Steps	516
	Simple Object Access Protocol	517
	Really Simple Syndication	517
	Apache Virtual File System Integration	517
	Data Formats	517
	XML	518
	Kettle Steps for Working with XML	518
	Kettle Job Entries for XML	519
	HTML	520
	JavaScript Object Notation	520
	Syntax	521
	JSON, Kettle, and ETL/DI	522
	XML Examples	523
	Example XML Document	523
	XML Document Structure	523
	Mapping to the Sakila Sample Database	524
	Extracting Data from XML	525
	Overall Design: The import_xml_into_db Transformation	526
	Using the XSD Validator Step	528
	Using the “Get Data from XML” Step	530
	Generating XML Documents	537
	Overall Design: The export_xml_from_db Transformation	537
	Generating XML with the Add XML Step	538
	Using the XML Join Step	541

SOAP Examples	544
Using the “Web services lookup” Step	544
Configuring the “Web services lookup” Step	544
Accessing SOAP Services Directly	546
JSON Example	549
The Freebase Project	549
Freebase Versus Wikipedia	549
Freebase Web Services	550
The Freebase Read Service	550
The Metaweb Query Language	551
Extracting Freebase Data with Kettle	553
Generate Rows	554
Issuing a Freebase Read Request	555
Processing the Freebase Result Envelope	556
Filtering Out the Original Row	557
Storing to File	558
RSS	558
RSS Structure	558
Channel	558
Item	559
RSS Support in Kettle	560
RSS Input	561
RSS Output	562
Summary	567
Chapter 22 Kettle Integration	569
The Kettle API	569
The LGPL License	569
The Kettle Java API	570
Source Code	570
Building Kettle	571
Building javadoc	571
Libraries and the Class Path	571
Executing Existing Transformations and Jobs	571
Executing a Transformation	572
Executing a Job	573
Embedding Kettle	574
Pentaho Reporting	574
Putting Data into a Transformation	576
Dynamic Transformations	580
Dynamic Template	583
Dynamic Jobs	584
Executing Dynamic ETL in Kettle	586
Result	587
Replacing Metadata	588
Direct Changes with the API	589
Using a Shared Objects File	589