



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*
PUBLISHING 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



BIRMINGHAM - MUMBAI

Pentaho for Big Data Analytics

Copyright © 2013 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: November 2013

Production Reference: 1181113

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78328-215-9

www.packtpub.com

Cover Image by Jarek Blaminsky (milak6@wp.pl)

Credits

Authors

Manoj R Patil
Feris Thia

Reviewers

Rio Bastian
Paritosh H. Chandorkar
Vikram Takkar

Acquisition Editors

Kartikey Pandey
Rebecca Youe

Commissioning Editor

Mohammed Fahad

Technical Editors

Manan Badani
Pankaj Kadam

Copy Editors

Alisha Aranha
Sarang Chari
Brandt D'Mello
Tanvi Gaitonde
Dipti Kapadia
Laxmi Subramanian

Project Coordinator

Sageer Parkar

Proofreaders

Ameesha Green
Maria Gould

Indexer

Rekha Nair

Graphics

Sheetal Aute
Ronak Dhruv
Disha Haria
Abhinash Sahu

Production Coordinator

Arvindkumar Gupta

Cover Work

Arvindkumar Gupta

About the Authors

Manoj R Patil is the Chief Architect in Big Data at Compassites Software Solutions Pvt. Ltd. where he overlooks the overall platform architecture related to Big Data solutions, and he also has a hands-on contribution to some assignments. He has been working in the IT industry for the last 15 years. He started as a programmer and, on the way, acquired skills in architecting and designing solutions, managing projects keeping each stakeholder's interest in mind, and deploying and maintaining the solution on a cloud infrastructure. He has been working on the Pentaho-related stack for the last 5 years, providing solutions while working with employers and as a freelancer as well.

Manoj has extensive experience in JavaEE, MySQL, various frameworks, and Business Intelligence, and is keen to pursue his interest in predictive analysis.

He was also associated with TalentBeat, Inc. and Persistent Systems, and implemented interesting solutions in logistics, data masking, and data-intensive life sciences.

Thank you Packt Publishing for extending this opportunity and guiding us through this process with your extremely co-operative team! I would also like to thank my beloved parents, lovely wife Manasi, and two smart daughters for their never-ending support, which keeps me going. Special thanks to my friend Manish Patel, my CEO Mahesh Baxi for being inspirational in my taking up this project, my co-author Feris for being committed in spite of his busy schedule, reviewers for reading the book and giving meaningful commentary and reviews, and to all those who directly/indirectly helped me with this book.

Finally I would like to extend an extra special thanks to Mahatria Ra for being an everlasting source of energy.

Feris Thia is a founder of PHI-Integration, a Jakarta-based IT consulting company that focuses on data management, data warehousing and Business Intelligence solutions. As a technical consultant, he has spent the last seven years delivering solutions with Pentaho and the Microsoft Business Intelligence platform across various industries, including retail, trading, finance/banking, and telecommunication.

He is also a member and maintainer of two very active local Indonesian discussion groups related to Pentaho (pentaho-id@googlegroups.com) and Microsoft Excel (the BelajarExcel.info Facebook group).

His current activities include research and building software based on Big Data and the data mining platform, that is, Apache Hadoop, R, and Mahout.

He would like to work on a book with a topic on analyzing customer behavior using the Apache Mahout platform.

I'd like to thank my co-author Manoj R Patil, technical reviewers, and all the folks at Packt Publishing, who have given me the chance to write this book and helped me along the way. I'd also like to thank all the members of the Pentaho Indonesia User Group and Excel Indonesia User Group through the years for being my inspiration for the work I've done.

About the Reviewers

Rio Bastian is a happy software developer already working on several IT projects. He is interested in Data Integration, and tuning SQL and Java code. He has also been a Pentaho Business Intelligence trainer for several companies in Indonesia and Malaysia. Rio is currently working as a software developer in PT. Aero Systems Indonesia, a company that focuses on the development of airline customer loyalty programs. It's an IT consultant company specializing in the airline industry. In his spare time, he tries to share his experience in developing software through his personal blog altanovela.wordpress.com. You can reach him on Skype ([rio.bastian](https://www.skype.com/people/rio.bastian)) or e-mail him at altanovela@gmail.com.

Paritosh H. Chandorkar is a young and dynamic IT professional with more than 11 years of information technology management experience in diverse domains, such as telecom and banking.

He has both strong technical (in Java/JEE) and project management skills. He has expertise in handling large customer engagements. Furthermore, he has expertise in the design and development of very critical projects for clients such as BNP Paribas, Zon TVCabo, and Novell. He is an impressive communicator with strong leadership, coordination, relationship management, analytical, and team management skills. He is comfortable interacting with people across hierarchical levels for ensuring smooth project execution as per client specifications. He is always eager to invest in improving knowledge and skills.

He is currently studying at Manipal University for a full-time M.S. in Software Design and Engineering.

His last designation was Technology Architect at Infosys Ltd.

I would like to thank Manoj R Patil for giving me the opportunity to review this book.

Vikram Takkar is a freelance Business Intelligence and Data Integration professional with nine years of rich, hands-on experience in multiple BI and ETL tools. He has strong expertise in tools such as Talend, Jaspersoft, Pentaho, Big Data-MongoDB, Oracle, and MySQL. He has managed and successfully executed multiple projects in data warehousing and data migration developed for both UNIX and Windows environments.

Apart from this, he is a blogger and publishes articles and videos on open source BI and ETL tools along with supporting technologies. You can visit his blog at www.vikramtakkar.com.

His YouTube channel is www.youtube.com/vtakkar. His Twitter handle is @VikTakkar. You can also follow him on his blog at www.vikramtakkar.com.

www.PacktPub.com

Support files, eBooks, discount offers and more

You might want to visit www.PacktPub.com for support files and downloads related to your book.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<http://PacktLib.PacktPub.com>

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can access, read and search across Packt's entire library of books.

Why Subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print and bookmark content
- On demand and accessible via web browser

Free Access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktLib today and view nine entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface	1
Chapter 1: The Rise of Pentaho Analytics along with Big Data	7
Pentaho BI Suite – components	8
Data	9
Server applications	10
Thin Client Tools	10
Design tools	11
Edge over competitors	12
Summary	12
Chapter 2: Setting Up the Ground	13
Pentaho BI Server and the development platform	13
Prerequisites/system requirements	14
Obtaining Pentaho BI Server (Community Edition)	14
The JAVA_HOME and JRE_HOME environment variables	15
Running Pentaho BI Server	16
Pentaho User Console (PUC)	17
Pentaho Action Sequence and solution	19
The JPivot component example	20
The message template component example	21
The embedded HSQLDB database server	22
Pentaho Marketplace	25
Saiku installation	25
Pentaho Administration Console (PAC)	27
Creating data connections	28
Summary	31

Chapter 3: Churning Big Data with Pentaho	33
An overview of Big Data and Hadoop	33
Big Data	34
Hadoop	35
The Hadoop architecture	35
The Hadoop ecosystem	38
Hortonworks Sandbox	41
Pentaho Data Integration (PDI)	41
The Pentaho Big Data plugin configuration	42
Importing data to Hive	44
Putting a data file into HDFS	50
Loading data from HDFS into Hive (job orchestration)	52
Summary	56
Chapter 4: Pentaho Business Analytics Tools	57
The business analytics life cycle	57
Preparing data	59
Preparing BI Server to work with Hive	59
Executing and monitoring a Hive MapReduce job	60
Pentaho Reporting	61
Data visualization and dashboard building	64
Creating a layout using a predefined template	66
Creating a data source	69
Creating a component	71
Summary	72
Chapter 5: Visualization of Big Data	73
Data visualization	73
Data source preparation	75
Repopulating the nyse_stocks Hive table	75
Pentaho's data source integration	76
Consuming PDI as a CDA data source	77
Visualizing data using CTools	78
Visualizing trends using a line chart	78
Interactivity using a parameter	79
Multiple pie charts	81
Waterfall charts	82
CSS styling	84
Summary	87

Appendix A: Big Data Sets	89
Freebase	89
U.S. airline on-time performance	90
Amazon public data sets	90
Appendix B: Hadoop Setup	91
Hortonworks Sandbox	91
Setting up the Hortonworks Sandbox	91
Hortonworks Sandbox web administration	93
Transferring a file using secure FTP	95
Preparing Hive data	96
The nyse_stocks sample data	98
Index	99
