Community Experience Distilled

Pentaho Analytics for MongoDB

Combine Pentaho Analytics and MongoDB to create powerful analysis and reporting solutions

Bo Borland

[PACKT] PUBLISHING open source* community experience distilled

Pentaho Analytics for MongoDB

Combine Pentaho Analytics and MongoDB to create powerful analysis and reporting solutions

Bo Borland

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

Pentaho Analytics for MongoDB

First published: February 2014

Production Reference: 1180214

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

ISBN 978-1-78216-835-5

`www.packtpub.com`

Cover Image by Andrey Bayda (`http://shutterstock.com`)

Credits

Author

Bo Borland

Reviewers

Rio Bastian

Pooya Esfandiar

Gretchen Moran

Khaled Tannir

Acquisition Editors

James Jones

Meeta Rajani

Content Development Editor

Govindan K

Technical Editors

Dennis John

Gaurav Thingalaya

Project Coordinator

Aboli Ambardekar

Proofreaders

Simran Bhogal

Maria Gould

Ameesha Green

Indexers

Priya Subramani

Hemangini Bari

Monica Ajmera Mehta

Rekha Nair

Graphics

Abhinash Sahu

Yuvraj Mannari

Production Coordinators

Conidon Miranda

Kyle Albuquerque

Cover Work

Kyle Albuquerque

About the Author

Bo Borland is the vice president of field technical sales at Pentaho, a leading Big Data analytics software provider. He has a passion for building teams and helping companies improve performance with data analytics. Prior to joining Pentaho, Bo worked as a management consultant at Deloitte and as a solution architect at Cognos, an IBM company. He founded a successful analytics consulting company, Management Signals, which merged with Pentaho in 2012. His 14 years' experience in professional analytics includes roles in management, sales, consulting, and sales engineering.

Pentaho Corporation is a leading Big Data analytics company headquartered in Orlando, FL, with offices in San Francisco, London, and Portugal. Pentaho tightly couples data integration with full business analytics capabilities into a single, integrated software platform. The company's subscription-based software offers SMEs and global enterprises the ability to reduce the time taken to design, develop, and deploy Big Data analytics solutions.

I wish to thank the wonderful team at Pentaho for their support and encouragement of this project, especially Rebecca Shomair and Monie TenBroeck for proactively reaching out to offer a helping hand, and Gretchen Moran for her valuable assistance as a technical reviewer for this book.

I also want to thank my wife, Alison, and daughters, Lin and Greta, for graciously accepting the fact that many nights and weekends were spent writing this book.

About the Reviewers

Rio Bastian is a happy software developer already working on several IT projects. His interests include business intelligence, data integration, tuning SQL, and tuning Java code. He has also been a Pentaho Business Intelligence trainer for a server company in Indonesia and Malaysia. He is currently focused on the development of an airline customer loyalty program in PT. Aero Systems Indonesia, an IT consultant specializing in airline industries. 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 via `rio.bastian` or via e-mail at `altanovela@gmail.com`.

Pooya Esfandiar is a software engineer and data analyst. He received an MS in Computer Science from the University of British Columbia in 2010 while working at the data management and mining labs. His research interests include simulation, data mining, and machine learning in social networks. He has worked with a few companies and organizations in Iran and Canada to solve business problems in information systems, general software engineering, and data analysis. He once designed an in-house analytics solution using Pentaho and MongoDB and is now working as a software engineer at Amazon.com, Inc.

Gretchen Moran is working as an independent Pentaho consultant on a variety of business analytics and Big Data projects. She has 15 years' experience in the business intelligence industry, developing software and providing services for a number of companies, including Hyperion Solutions and Pentaho Corporation.

Gretchen continues to contribute to Pentaho Corporation's latest software initiatives while managing the daily adventures of her two children, Isabella and Jack, with her husband Doug.

Khaled Tannir has been working with computers since 1980. He began programming with the legendary Sinclair Zx81 and afterwards with all Commodore home computer products (VIC-20, Commodore 64, Commodore 128D, and Amiga 500).

He has a Bachelor's degree in Electronics, a Master's degree in System Information Architectures in which he graduated with a professional thesis, and he completed his education with a Master of Research degree.

He is a Microsoft Certified Solution Developer (MCSD) and has more than 20 years' technical experience leading the development and implementation of software solutions and giving technical presentations. He works as an independent IT consultant and has worked as an infrastructure engineer, senior developer, and enterprise/solution architect for many companies in France and Canada.

With a significant experience in Microsoft .NET/Servers and Oracle Java technologies, he has extensive skills in online/offline application design, system conversions, and multilanguage applications on both the Internet and desktops.

He spends his time researching on new technologies, learning about them, and looking for new adventures in France, North America, and the Middle East. He owns an IT and electronics laboratory with many servers, monitors, open electronics boards (such as Arduino, Netduino, Raspberry Pi, and .NET Gadgeteer), and some smartphone devices based on Windows Phone, Android, and iOS operating systems.

In 2012, he contributed to the EGC 2012 (International Complex Data Mining forum at Bordeaux University, France) and presented his work on *how to optimize data distribution in a cloud computing environment* in a workshop session. This work aims to define an approach to optimize the use of data mining algorithms such as k-means and Apriori in a cloud computing environment.

He is the author of *RavenDB 2.x Beginner's Guide*, *Packt Publishing* and *Optimizing Hadoop MapReduce*, *Packt Publishing*.

He aims to get a PhD degree in Cloud Computing and Big Data and wants to learn more and more about these technologies.

He enjoys taking landscape and night photos, travelling, playing video games, creating funny electronics gadgets with Arduino / .NET Gadgeteer and of course, spending time with his wife and family. You can reach him at `contact@khaledtannir.net`.

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

MongoDB and Pentaho go together like yin and yang. They are emerging as a powerful combination for scalable data storage, processing, and analytics. Leading companies are pairing these complementary technologies together in development labs and production to deliver innovative analytics. These innovations are creating worldwide demand for developers with skills in both Pentaho and MongoDB.

You want to make an impact by creating innovative data storage capabilities or eye-catching data visualizations. Wouldn't it be great if you could quickly ramp up on both technologies to develop a turn-key solution for your organization? However, as with any new and emerging technology combination, the availability of organized knowledge on the combined topic is scarce.

Pentaho Analytics for MongoDB will show you how to develop an analytic solution that you can demonstrate to your colleagues. It is a practical guide to get you started with both Pentaho and MongoDB, beginning with basic MongoDB data modeling and querying and then advancing to data integration, analysis, and reporting with Pentaho. Each chapter guides you through using different components of the Pentaho platform to create analytic models and reports using a sample MongoDB database.

What this book covers

Chapter 1, *Getting Started with Pentaho and MongoDB*, introduces you to the powerful combination of MongoDB and Pentaho and provides step-by-step guidance on how to install and configure both technologies and restore the sample MongoDB data provided with this book.

Chapter 2, *MongoDB Database Fundamentals*, expands on the topic of data modeling and explains MongoDB database concepts essential to querying MongoDB data with Pentaho.

Chapter 3, *Using Pentaho Instaview*, shows you how to visualize data by connecting Pentaho to MongoDB. You use Instaview with the sample MongoDB database to analyze and visualize the website clickstream data.

Chapter 4, *Modifying and Enhancing Instaview Transformations*, introduces Pentaho Data Integration (PDI)—the ETL tool used by Instaview to extract, load, and transform data from various data sources.

Chapter 5, *Modifying and Enhancing Instaview Metadata*, explores metadata by explaining dimensional modeling concepts and how to model metadata to better reflect business requirements.

Chapter 6, *Pentaho Report Designer Fundamentals*, teaches you the basics of Pentaho Report Designer (PRD) to build pixel-perfect reports sourced directly from MongoDB databases.

Chapter 7, *Pentaho Report Designer Prompting and Charting*, expands on the previous chapter by teaching you additional advanced PRD features. You can enhance your report with new queries, charts, and a prompt designed to make the report more interactive.

Chapter 8, *Deploying Pentaho Analytics to the Web*, is all about web-enabling your MongoDB data using Pentaho methods and web interfaces for connecting to, modeling, and analyzing our sample clickstream data in a web browser.

What you need for this book

We need the following software for this book:

- Pentaho Business Analytics v5.0.2 (64-bit for Windows)
- MongoDB v2.2.3 (64-bit for Windows)

This book provides two data sources for use throughout the book, a MongoDB database of sample web clickstream data, and an associated comma-separated (CSV) file containing geographic data. Both files are available as a free download from: `http://www.packtpub.com/support`.

Who this book is for

This book is intended for business analysts, data architects, and developers new to either Pentaho or MongoDB, who want to be able to deliver a complete solution for storing, processing, and visualizing data. It's assumed that you already have experience in defining the data requirements needed to support business processes and exposure to database modeling, SQL query, and reporting techniques.