Community Experience Distilled

Apache Solr for Indexing Data

Enhance your Solr indexing experience with advanced techniques and the built-in functionalities available in Apache Solr

Sachin Handiekar Anshul Johri

[PACKT] PUBLISHING open source* community experience distilled

Apache Solr for Indexing Data

Enhance your Solr indexing experience with advanced techniques and the built-in functionalities available in Apache Solr

Sachin Handiekar

Anshul Johri

[PACKT] PUBLISHING open source community experience distilled

BIRMINGHAM - MUMBAI

Apache Solr for Indexing Data

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: December 2015

Production reference: 1151215

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

ISBN 978-1-78355-323-5

www.packtpub.com

Credits

Authors
Sachin Handiekar
Anshul Johri

Reviewers
Damiano Braga
Florian Hopf

Commissioning Editor
Ashwin Nair

Acquisition Editors
Rebecca Pedley
Reshma Raman

Content Development Editor
Rohit Kumar Singh

Technical Editor
Utkarsha S. Kadam

Copy Editor
Vikrant Phadke

Project Coordinator
Mary Alex

Proofreader
Safis Editing

Indexer
Rekha Nair

Production Coordinator
Manu Joseph

Cover Work
Manu Joseph

About the Authors

Sachin Handiekar is a senior software developer with over 5 years of experience in Java EE development. He graduated in computer science from the University of Greenwich, London, and currently works for a global consulting company, developing enterprise applications using various open source technologies, such as Apache Camel, ServiceMix, ActiveMQ, and ZooKeeper.

He has a lot of interest in open source projects and has contributed code to Apache Camel and developed plugins for the Spring Social, which can be found on GitHub at `https://github.com/sachin-handiekar`.

He also actively writes about enterprise application development on his blog (`http://www.sachinhandiekar.com/`).

Anshul Johri has more than 10 years of technical experience in software engineering. He did his masters in computer science from the computer science department in the University of Pune. Anshul has always been a start-up mindset guy, working on fast-paced development using cutting-edge technologies and doing multiple things at a time. His core strength has always been search technology, whereby Solr plays an important role in his career. Anshul started using Solr around 9 years ago, and since then, he has never looked back. He did better and better with Solr, whether using it or contributing to the open source search community. He has used Solr extensively in all his organizations across various projects.

As mentioned earlier, Anshul has always been a start-up mindset guy. Because of that, he has worked with many start-ups in his career so far, which includes early-age and mid-size start-ups as well. To name a few, they are Ibibo.com, Asklaila.com, Bookadda.com, and so on. His last company was Amazon, where he spent around 2 years building scalable systems for Amazon Prime (a global product). Anshul recently started his own company in India with another friend from Amazon and founded `http://www.rentomo.com/`, a unique concept of a peer-to-peer sharing platform in a trusted community. He heads the technology and other core pillars of his own start-up.

Anshul did the technical review of the book *Indexing with Solr*, published by Packt Publishing.

About the Reviewers

Damiano Braga is the technical search lead at Trulia, where he leads all the backend search and browsing-related projects. He's also an open source contributor and has participated as a speaker at the Lucene Revolution 2014, where he presented Thoth, a real-time Solr monitoring and search analysis engine. He also previously reviewed the book *Apache Solr Search Patterns*, *Packt Publishing*.

Prior to Trulia, Damiano studied and worked for the University of Ferrara (Italy), where he also completed his master's degree in computer science engineering.

Florian Hopf works as a freelance software developer and consultant in Karlsruhe, Germany. He familiarized himself with Lucene-based searching while working with different content management systems on the Java platform. He is responsible for small and large search systems, on both the Internet and Intranet, for web content and application-specific data based on Lucene, Solr, and Elasticsearch. He helps organize the local Java user group as well as the Search Meetup in Karlsruhe. Florian has also written a German book on Elasticsearch. He posts blogs at `http://blog.florian-hopf.de/`.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

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.

PACKTLiB™

https://www2.packtpub.com/books/subscription/packtlib

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read 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 a 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 9 entirely free books. Simply use your login credentials for immediate access.

I would like to dedicate this book to my parents, especially my late mother, Anuradha Johri, who has always been my inspiration and my friend for life. After that I would like to thank my wife, Aparna, who is always there with me in every situation no matter how tough it is; she is someone who always makes me feel complete.

Table of Contents

Preface

Welcome to *Apache Solr for Indexing Data*. Solr is an amazing enterprise tool that gives us a search engine with various possibilities to index data and gives users a better experience. This book will cover the various indexing methods that we can use to improve the indexing process by covering step-by-step examples.

The book is all about indexing in Solr, and we'll cover all the possible topics in Solr that developers can use in their use cases by following simple examples.

What this book covers

Chapter 1, *Getting Started*, covers the basic setup and installation needed to run Solr. It also covers the directory structure and the main configuration files used by Solr.

Chapter 2, *Understanding Analyzers, Tokenizers, and Filters*, shows you the basic building blocks of Solr, such as analyzers, tokenizers, and filters. These help in the indexing of data. This chapter also covers the most commonly used components in detail and how they work together.

Chapter 3, *Indexing Data*, helps you get a better understanding of how indexing works in Solr by building a real-life example that covers various aspects, for example, the copy field, facet, indexing time boosting, and so on.

Chapter 4, *Indexing Data – The Basic Techniques and Using Index Handlers*, covers various techniques by which we can index data in Solr. This chapter explains the various request handlers that are used by Solr to index CSV, JSON, and XML data type documents.

Chapter 5, *Indexing Data Using Structured Datasource Using DIH*, covers how we can use indexed data from a database by using the data import handler available in Solr.