



C o m m u n i t y E x p e r i e n c e D i s t i l l e d

Apache Solr High Performance

Boost the performance of Solr instances and troubleshoot real-time problems

Surendra Mohan

[PACKT] open source*
PUBLISHING community experience distilled

Apache Solr High Performance

Boost the performance of Solr instances and
troubleshoot real-time problems

Surendra Mohan



BIRMINGHAM - MUMBAI

Apache Solr High Performance

Copyright © 2014 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 author 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: March 2014

Production Reference: 1180314

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

ISBN 978-1-78216-482-1

www.packtpub.com

Cover Image by Glain Clarrie (glen.m.carrie@gmail.com)

Credits

Author

Surendra Mohan

Project Coordinator

Puja Shukla

Reviewers

Azaz Desai

Ankit Jain

Mark Kerzner

Ruben Teijeiro

Proofreaders

Simran Bhogal

Ameesha Green

Maria Gould

Acquisition Editor

Neha Nagwekar

Indexers

Monica Ajmera Mehta

Mariammal Chettiyar

Content Development Editor

Poonam Jain

Graphics

Abhinash Sahu

Technical Editor

Krishnaveni Haridas

Production Coordinator

Saiprasad Kadam

Copy Editors

Mradula Hegde

Alfida Paiva

Adithi Shetty

Cover Work

Saiprasad Kadam

About the Author

Surendra Mohan, who has served a few top-notch software organizations in varied roles, is currently a freelance software consultant. He has been working on various cutting-edge technologies such as Drupal and Moodle for more than nine years. He also delivers technical talks at various community events such as Drupal meet-ups and Drupal camps. To know more about him, his write-ups, and technical blogs, and much more, log on to <http://www.surendramohan.info/>.

He has also authored the book *Administrating Solr*, Packt Publishing, and has reviewed other technical books such as *Drupal 7 Multi Sites Configuration* and *Drupal Search Engine Optimization*, Packt Publishing, and titles on Drupal commerce and ElasticSearch, Drupal-related video tutorials, a title on Opsview, and many more.

I would like to thank my family and friends who supported and encouraged me in completing this book on time with good quality.

About the Reviewers

Azaz Desai has more than three years of experience in Mule ESB, jBPM, and Liferay technology. He is responsible for implementing, deploying, integrating, and optimizing services and business processes using ESB and BPM tools. He was a lead writer of *Mule ESB Cookbook*, Packt Publishing, and also played a vital role as a trainer on ESB. He currently provides training on Mule ESB to global clients. He has done various integrations of Mule ESB with Liferay, Alfresco, jBPM, and Drools. He was part of a key project on Mule ESB integration as a messaging system. He has worked on various web services and standards and frameworks such as CXF, AXIS, SOAP, and REST.

Ankit Jain holds a bachelor's degree in Computer Science Engineering from RGPV University, Bhopal, India. He has three years of experience in designing and architecting solutions for the Big Data domain and has been involved with several complex engagements. His technical strengths include Hadoop, Storm, S4, HBase, Hive, Sqoop, Flume, ElasticSearch, Machine Learning, Kafka, Spring, Java, and J2EE.

He also shares his thoughts on his personal blog at <http://ankitasblogger.blogspot.in/>. You can follow him on Twitter at @mynameisanky. He spends most of his time reading books and playing with different technologies. When not at work, Ankit spends time with his family and friends, watching movies, and playing games.

I would like to thank my parents and brother for always being there for me.

Mark Kerzner holds degrees in Law, Maths, and Computer Science. He has been designing software for many years and Hadoop-based systems since 2008. He is the President of SHMsoft, a provider of Hadoop applications for various verticals, and a cofounder of the Hadoop Illuminated training and consulting, as well as the coauthor of the *Hadoop Illuminated* open source book. He has authored and coauthored several books and patents.

I would like to acknowledge the help of my colleagues, in particular Sujee Maniyam, and last but not least, my multitalented family.

Ruben Teijeiro is an experienced frontend and backend web developer who had worked with several PHP frameworks for over a decade. His expertise is focused now on Drupal, with which he had collaborated in the development of several projects for some important organizations such as UNICEF and Telefonica in Spain and Ericsson in Sweden.

As an active member of the Drupal community, you can find him contributing to Drupal core, helping and mentoring other contributors, and speaking at Drupal events around the world. He also loves to share all that he has learned by writing in his blog, <http://drewpull.com>.

I would like to thank my parents for supporting me since I had my first computer when I was eight years old, and letting me dive into the computer world. I would also like to thank my fiancée, Ana, for her patience while I'm geeking around the world.

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: Installing Solr	7
Prerequisites for Solr	7
Installing components	8
Summary	12
Chapter 2: Boost Your Search	13
Scoring	13
Query-time and index-time boosting	15
Index-time boosting	15
Query-time boosting	15
Troubleshoot queries and scores	16
The dismax query parser	18
Lucene DisjunctionMaxQuery	19
Autophrase boosting	20
Configuring autophrase boosting	21
Configuring the phrase slop	21
Boosting a partial phrase	22
Boost queries	22
Boost functions	24
Boost addition and multiplication	24
Function queries	25
Field references	27
Function references	27
Mathematical operations	28
The ord() and rord() functions	29
Other functions	30
Boosting the function query	31
Logarithm	32
Reciprocal	33

Linear	34
Inverse reciprocal	34
Summary	36
Chapter 3: Performance Optimization	37
Solr performance factors	37
Solr caching	38
Document caching	38
Query result caching	39
Filter caching	41
Result pages caching	42
Using SolrCloud	44
Creating a SolrCloud cluster	45
Multiple collections within a cluster	46
Managing a SolrCloud cluster	49
Distributed indexing and searching	51
Stopping automatic document distribution	54
Near real-time search	58
Summary	59
Chapter 4: Additional Performance Optimization Techniques	61
Documents similar to those returned in the search result	62
Sorting results by function values	64
Searching for homophones	67
Ignore the defined words from being searched	69
Summary	72
Chapter 5: Troubleshooting	73
Dealing with the corrupt index	73
Reducing the file count in the index	76
Dealing with the locked index	77
Truncating the index size	77
Dealing with a huge count of open files	79
Dealing with out-of-memory issues	81
Dealing with an infinite loop exception in shards	82
Dealing with expensive garbage collection	83
Bulk updating a single field without full indexation	85
Summary	87
Chapter 6: Performance Optimization with ZooKeeper	89
Getting familiar with ZooKeeper	89
Prerequisites for a distributed server	89
Aid your distributed system using ZooKeeper	91
Setting an ideal node count for ZooKeeper	93

Setting up, configuring, and deploying ZooKeeper	93
Setting up ZooKeeper	94
Configuring ZooKeeper	94
Deploying ZooKeeper	95
Applications of ZooKeeper	99
Summary	100
Appendix: Resources	101
Index	105
