Professional Hadoop®

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, Kai Sasaki

PROFESSIONAL HADOOP®

PROFESSIONAL

Hadoop®

PROFESSIONAL

Hadoop®

Benoy Antony
Konstantin Boudnik
Cheryl Adams
Branky Shao
Cazen Lee
Kai Sasaki

Professional Hadoop®

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Published simultaneously in Canada

ISBN: 978-1-119-26717-1
ISBN: 978-1-119-26718-8 (ebk)
ISBN: 978-1-119-26720-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2016934264

ABOUT THE AUTHORS

BENOY ANTONY is an Apache Hadoop committer and has contributed features related to security and HDFS. He is the founder of DataApps (http://dataApps.io), a company that specializes in creating applications for big data. He maintains a Hadoop Security wiki at http://HadoopSecurity.org. Benoy is a Hadoop architect at eBay where he focuses on enhancing security and availability on eBay's Hadoop clusters without limiting user productivity. He regularly speaks at conferences like Hadoop Summit.

DR. KONSTANTIN BOUDNIK, co-founder and CEO of Memcore.io, is one of the early developers of Hadoop and a co-author of Apache Bigtop, the open source framework and the community around creation of software stacks for data processing projects. With more than 20 years of experience in software development, big- and fast-data analytic, Git, distributed systems and more, Dr. Boudnik has authored 15 US patents in distributed computing. Dr. Boudnik contributed to a dozen of open source projects in the area of distributed computing and data processing. He has helped and championed a number of successful Apache projects in the area.

CHERYL ADAMS is a senior cloud data and infrastructure architect. Her work includes supporting healthcare data for large government contracts; deploying production-based changes through scripting, monitoring, and troubleshooting; and monitoring environments using the latest tools for databases, web servers, web API, and storage.

BRANKY SHAO is a software engineer at eBay where he is building real time applications with Elasticsearch, Cassandra, Kafka, and Storm. He has been working with the Hadoop ecosystem technologies since 2010. He has extensive experience designing and implementing various software including distributed systems, data integration, framework/APIs, and web applications. He is passionate about open source and is a contributor to the Cascading project.

CAZEN LEE is a software architect at Samsung SDS. He is currently in charge of the Hadoop module for Samsung's big data platform. Prior to joining Samsung, Cazen served as a developer and architect for the integrated data warehouse layer in the financial industry, including work with Samsung Life Insurance and Korea Securities Finance Corp. He is also interested in both machine learning and neural network models.

KAI SASAKI is a Japanese software engineer who is interested in distributed computing and machine learning. Currently he is working at Treasure Data Inc., launched by Japanese entrepreneurs based in Silicon Valley. Although the beginning of his career didn't start with Hadoop or Spark, his interest in middleware and the fundamental technologies that support a lot of these types of big data services and the Internet drove him toward this field. He has been a Spark contributor, developing mainly MLlib and ML libraries. Nowadays, he is trying to research the great potential of combining deep learning and big data. He believes that Spark can play a significant role even in artificial intelligence within the big data era. You can find him on GitHub at https://github.com/Lewuathe.

ABOUT THE TECHNICAL EDITORS

SNEHAL NAGMOTE is a staff software engineer for the search infrastructure team at Walmart Labs. Some of his responsibilities include building data platform applications using the big data stack, and using tools such as Hadoop, Hive, Kafka, Flume, and Spark. Currently, he is focusing on building a near real time indexing data pipeline using Spark Streaming and Kafka.

RENAN PINZON is a software architect at NeoGrid and has been working with Hadoop there for more than three years. He has a lot of experience with mission-critical software and data processing/analysis. He started using Hadoop for real-time processing (HBase + HDFS) and then started to use it in data analysis with RHadoop, Pig, Crunch, and is now moving to Spark. He also has been working with search engines using Apache Solr for real-time indexing and search as well as using Elasticsearch outside of Hadoop. Despite his professional experience being more in software development, he has a strong background in infrastructure, mainly in regard to Hadoop where he has been working tuning applications.

MICHAEL CUTLER has deep experience with the Hadoop ecosystem since building one of the UK's earliest Hadoop Clusters for BSkyB in 2008 after successfully pitching CXO management for innovation funding to explore the tools and techniques, which have now become known as big data. He has real world experience in training predictive models from huge multi-terabyte datasets across diverse business use cases as: automated fraud detection, fault prediction and classification, recommendations, click-stream analysis, large scale business simulations and modeling. Michael was an invited speaker on machine learning at Hadoop World in New York. He is well connected in the open source ecosystem and is a regular speaker at data science and big data events in London.

CREDITS

PROJECT EDITOR
Charlotte Kughen

TECHNICAL EDITORS
Snehal Nagmote
Renan Pinzon
Michael Cutler

PRODUCTION EDITOR
Barath Kumar Rajasekaran

COPY EDITOR
Troy Mott

MANAGER OF CONTENT DEVELOPMENT & ASSEMBLY
Mary Beth Wakefield

PRODUCTION MANAGER
Kathleen Wisor

MARKETING MANAGER
David Mayhew

PROFESSIONAL TECHNOLOGY & STRATEGY DIRECTOR
Barry Pruett

BUSINESS MANAGER
Amy Knies

EXECUTIVE EDITOR
Jim Minatel

PROJECT COORDINATOR, COVER
Brent Savage

PROOFREADER
Nancy Bell

INDEXER
Nancy Guenther

COVER DESIGNER
Wiley

COVER IMAGE
silvrock/Shutterstock

ACKNOWLEDGMENTS

Special thanks to the massive contributions to the Hadoop project by all the volunteers who spent their time to move the Apache Bigtop project forward, helping it to become a true integration hub of the 100% open source Apache data processing stack!

A special thanks also to the volunteers who spent their time to move the Apache Ignite project forward and helping it to become a real core of open source in-memory computing.

And a special thanks goes to Gridgain for their donation of the production grade software to the Apache Software Foundation. It was both a challenge and an honor to transform this project into the Apache TLP.

CONTENTS

INTRODUCTION

Hadoop is an open source project available under the Apache License 2.0. It has the ability to manage and store very large data sets across a distributed cluster of servers. One of the most beneficial features is its fault tolerance, which enables big data applications to continue to operate properly in the event of a failure. Another benefit of using Hadoop is its scalability. This programming logic has the potential to expand from a single server to numerous servers, each with the ability to have local computation and storage options.

WHO IS THIS BOOK FOR?

This book is for anyone using Hadoop to perform a job that is data related, or if you have an interest in redefining how you can obtain meaningful information about any of your data stores. This includes big data solution architects, Linux system and big data engineers, big data platform engineers, Java programmers, and database administrators.

If you have an interest in learning more about Hadoop and how to extract specific elements for further analysis or review, then this book is for you.

WHAT YOU NEED TO USE THIS BOOK

You should have development experience and understand the basics of Hadoop, and should now be interested in employing it in real-world settings.

The source code for the samples is available for download at `www.wrox.com/go/professionalhadoop` or `https://github.com/backstopmedia/hadoopbook`.

HOW THIS BOOK IS STRUCTURED

This book was written in eight chapters as follows:

Chapter 1: Hadoop Introduction

Chapter 2: Storage

Chapter 3: Computation

Chapter 4: User Experience

Chapter 5: Integration with Other Systems

Chapter 6: Hadoop Security

Chapter 7: Ecosystem at Large: Hadoop Stack with Apache Bigtop

Chapter 8: In-Memory Computing in Hadoop Stack

CONVENTIONS

To help you get the most from the text and keep track of what's happening, we've used a number of conventions throughout the book.

As for styles in the text:

- We *highlight* new terms and important words when we introduce them.
- We show code within the text like so: `persistence.properties`.
- We show all code snippets in the book using this style:

```
FileSystem fs = FileSystem.get(URI.create(uri), conf);
InputStream in = null;
try {
```

- We show URLs in text like this:

```
http://<Slave Hostname>:50075
```

SOURCE CODE

As you work through the examples in this book, you may choose either to type in all the code manually, or to use the source code files that accompany the book. All the source code used in this book is available for download at `www.wrox.com`. Specifically for this book, the code download is on the Download Code tab at:

`www.wrox.com/go/professionalhadoop`

You can also search for the book at `www.wrox.com` by ISBN (the ISBN for this book is 9781119267171) to find the code. And a complete list of code downloads for all current Wrox books is available at `www.wrox.com/dynamic/books/download.aspx`.

> **NOTE** *Because many books have similar titles, you may find it easiest to search by ISBN; this book's ISBN is 978-1-119-26717-1.*

Once you download the code, just decompress it with your favorite compression tool. Alternately, you can go to the main Wrox code download page at `www.wrox.com/dynamic/books/download.aspx` to see the code available for this book and all other Wrox books.