



Community Experience Distilled

Programming MapReduce with Scalding

A practical guide to designing, testing, and implementing complex MapReduce applications in Scala

Antonios Chalkiopoulos

[PACKT] open source*
PUBLISHING community experience distilled

Programming MapReduce with Scalding

A practical guide to designing, testing, and
implementing complex MapReduce applications
in Scala

Antonios Chalkiopoulos

[PACKT] open source 
PUBLISHING community experience distilled

BIRMINGHAM - MUMBAI

Programming MapReduce with Scalding

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: June 2014

Production reference: 1190614

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

ISBN 978-1-78328-701-7

www.packtpub.com

Credits

Author

Antonios Chalkiopoulos

Project Coordinator

Aboli Ambardekar

Reviewers

Ahmad Alkilani

Włodzimierz Bzyl

Tanin Na Nakorn

Sen Xu

Proofreaders

Mario Cecere

Maria Gould

Indexers

Mehreen Deshmukh

Rekha Nair

Tejal Soni

Commissioning Editor

Owen Roberts

Acquisition Editor

Llewellyn Rozario

Graphics

Sheetal Aute

Ronak Dhruv

Valentina Dsilva

Disha Haria

Content Development Editor

Sriram Neelakantan

Technical Editor

Kunal Anil Gaikwad

Production Coordinator

Conidon Miranda

Copy Editors

Sayanee Mukherjee

Alfida Paiva

Cover Work

Conidon Miranda

Cover Image

Sheetal Aute

About the Author

Antonios Chalkiopoulos is a developer living in London and a professional working with Hadoop and Big Data technologies. He completed a number of complex MapReduce applications in Scalding into 40-plus production nodes HDFS Cluster. He is a contributor to Scalding and other open source projects, and he is interested in cloud technologies, NoSQL databases, distributed real-time computation systems, and machine learning.

He was involved in a number of Big Data projects before discovering Scala and Scalding. Most of the content of this book comes from his experience and knowledge accumulated while working with a great team of engineers.

I would like to thank Rajah Chandan for introducing Scalding to the team and being the author of SpyGlass and Stefano Galarraga for co-authoring chapters 5 and 6 and being the author of ScaldingUnit. Both these libraries are presented in this book.

Saad, Gracia, Deepak, and Tamas, I've learned a lot working next to you all, and this book wouldn't be possible without all your discoveries. Finally, I would like to thank Christina for bearing with my writing sessions and supporting all my endeavors.

About the Reviewers

Ahmad Alkilani is a data architect specializing in the implementation of high-performance distributed systems, data warehouses, and BI systems. His career has been split between building enterprise applications and products using a variety of web and database technologies, including .NET, SQL Server, Hadoop, Hive, Scala, and Scalding. His recent interests include building real-time web and predictive analytics and streaming and sketching algorithms.

Currently, Ahmad works at Move.com (<http://www.realtor.com>) and enjoys speaking at various user groups and national conferences, and he is an author on Pluralsight with courses focused on Hadoop and Big Data, SQL Server 2014, and more, targeting the Big Data and streaming spaces.

You can find more information on Ahmad on his LinkedIn profile (<http://www.linkedin.com/in/ahmadalkilani>) or his Pluralsight author page (<http://pluralsight.com/training/Authors/Details/ahmad-alkilani>).

I would like to thank my family, especially my wonderful wife, Farah, and my beautiful son Maher for putting up with my long working hours and always being there for me.

Włodzimierz Bzyl works at the University of Gdańsk. His current interests include web-related technologies and NoSQL databases.

He has a passion for new technologies and introducing his students to them.

He enjoys contributing to open source software and spending time trekking in the Tatra mountains.

Tanin Na Nakorn is a software engineer who is enthusiastic about building consumer products and open source projects that make people's lives easier. He cofounded Thaiware, a software portal in Thailand and GiveAsia, a donation platform in Singapore; he currently builds products at Twitter. You may find him expressing himself on his Twitter handle @tanin and helping on various open source projects at <http://www.github.com/tanin47>.

Sen Xu is a software engineer in Twitter; he was previously a data scientist in Inome Inc.

He worked on designing and building data pipelines on top of traditional RDBMS (MySQL, PostgreSQL, and so on) and key-value store solutions (Hadoop). His interests include Big Data analytics, text mining, record linkage, machine learning, and spatial data handling.

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: Introduction to MapReduce	7
The Hadoop platform	8
MapReduce	8
A MapReduce example	9
MapReduce abstractions	10
Introducing Cascading	11
What happens inside a pipe	13
Pipe assemblies	13
Cascading extensions	14
Summary	15
Chapter 2: Get Ready for Scalding	17
Why Scala?	17
Scala basics	19
Scala build tools	20
Hello World in Scala	21
Development editors	22
Installing Hadoop in five minutes	22
Running our first Scalding job	23
Submitting a Scalding job in Hadoop	24
Summary	27
Chapter 3: Scalding by Example	29
Reading and writing files	29
Best practices to read and write files	31
TextLine parsing	32
Executing in the local and Hadoop modes	32

Understanding the core capabilities of Scalding	33
Map-like operations	33
Join operations	38
Pipe operations	40
Grouping/reducing functions	41
Operations on groups	42
Composite operations	49
A simple example	51
Typed API	51
Summary	52
Chapter 4: Intermediate Examples	53
Logfile analysis	53
Completing the implementation	58
Exploring ad targeting	60
Calculating daily points	62
Calculating historic points	67
Generating targeted ads	67
Summary	69
Chapter 5: Scalding Design Patterns	71
The external operations pattern	71
The dependency injection pattern	75
The late bound dependency pattern	77
Summary	78
Chapter 6: Testing and TDD	79
Introduction to testing	79
MapReduce testing challenges	80
Development lifecycle with testing strategy	81
TDD for Scalding developers	81
Implementing the TDD methodology	82
Decomposing the algorithm	82
Defining acceptance tests	83
Implementing integration tests	83
Implementing unit tests	85
Implementing the MapReduce logic	87
Defining and performing system tests	87
Black box testing	88
Summary	89

Chapter 7: Running Scalding in Production	91
Executing Scalding in a Hadoop cluster	91
Scheduling execution	92
Coordinating job execution	93
Configuring using a property file	94
Configuring using Hadoop parameters	96
Monitoring Scalding jobs	96
Using slim JAR files	98
Scalding execution throttling	100
Summary	101
Chapter 8: Using External Data Stores	103
Interacting with external systems	103
SQL databases	104
NoSQL databases	106
Understanding HBase	107
Reading from HBase	108
Writing in HBase	110
Using advanced HBase features	110
Search platforms	111
Elastic search	111
Summary	113
Chapter 9: Matrix Calculations and Machine Learning	115
Text similarity using TF-IDF	115
Setting a similarity using the Jaccard index	118
K-Means using Mahout	121
Other libraries	125
Summary	125
Index	127

Preface

Scalding is a relatively new Scala DSL that builds on top of the Cascading pipeline framework, offering a powerful and expressive architecture for MapReduce applications. Scalding provides a highly abstracted layer for design and implementation in a componentized fashion, allowing code reuse and development with the Test Driven Methodology.

Similar to other popular MapReduce technologies such as Pig and Hive, Cascading uses a tuple-based data model, and it is a mature and proven framework that many dynamic languages have built technologies upon. Instead of forcing developers to write raw map and reduce functions while mentally keeping track of key-value pairs throughout the data transformation pipeline, Scalding provides a more natural way to express code.

In simpler terms, programming raw MapReduce is like developing in a low-level programming language such as assembly. On the other hand, Scalding provides an easier way to build complex MapReduce applications and integrates with other distributed applications of the Hadoop ecosystem.

This book aims to present MapReduce, Hadoop, and Scalding, it suggests design patterns and idioms, and it provides ample examples of real implementations for common use cases.

What this book covers

Chapter 1, Introduction to MapReduce, serves as an introduction to the Hadoop platform, MapReduce and to the concept of the pipeline abstraction that many Big Data technologies use. The first chapter outlines Cascading, which is a sophisticated framework that empowers developers to write efficient MapReduce applications.

Chapter 2, Get Ready for Scalding, lays the foundation for working with Scala, using build tools and an IDE, and setting up a local-development Hadoop system. It is a hands-on chapter that completes packaging and executing a Scalding application in local mode and submitting it in our Hadoop mini-cluster.

Chapter 3, Scalding by Example, teaches us how to perform map-like operations, joins, grouping, pipe, and composite operations by providing examples of the Scalding API.

Chapter 4, Intermediate Examples, illustrates how to use the Scalding API for building real use cases, one for log analysis and another for ad targeting. The complete process, beginning with data exploration and followed by complete implementations, is expressed in a few lines of code.

Chapter 5, Scalding Design Patterns, presents how to structure code in a reusable, structured, and testable way following basic principles in software engineering.

Chapter 6, Testing and TDD, focuses on a test-driven methodology of structuring projects in a modular way for maximum testability of the components participating in the computation. Following this process, the number of bugs is reduced, maintainability is enhanced, and productivity is increased by testing every layer of the application.

Chapter 7, Running Scalding in Production, discusses how to run our jobs on a production cluster and how to schedule, configure, monitor, and optimize them.

Chapter 8, Using External Data Stores, goes into the details of accessing external NoSQL- or SQL-based data stores as part of a data processing workflow.

Chapter 9, Matrix Calculations and Machine Learning, guides you through the process of applying machine learning algorithms, matrix calculations, and integrating with Mahout algorithms. Concrete examples demonstrate similarity calculations on documents, items, and sets.

What you need for this book

Prior knowledge about Hadoop or Scala is not required to follow the topics and techniques, but it is certainly beneficial. You will need to set up your environment with the JDK, an IDE, and Maven as a build tool. As this is a practical guide you will need to set up a mini Hadoop cluster for development purposes.

Who this book is for

This book is structured in such a way as to introduce Hadoop and MapReduce to a developer who has a basic understanding of these technologies and to leverage existing and well-known tools in order to become highly productive. A more experienced Scala developer will benefit from the Scalding design patterns, and an experienced Hadoop developer will be enlightened by this alternative methodology of developing MapReduce applications with Scalding.

Conventions

In this book, you will find a number of styles of text that distinguish between different kinds of information. Here are some examples of these styles, and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, and user input are shown as follows: "A Map class to map lines into <key, value> pairs; for example, <"INFO", 1>."

A block of code is set as follows:

```
LogLine      = load 'file.logs' as (level, message);
LevelGroup   = group LogLine by level;
Result       = foreach LevelGroup generate group, COUNT(LogLine);
store Result into 'Results.txt';
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
import com.twitter.scalding._

class CalculateDailyAdPoints (args: Args) extends Job(args) {

  val logSchema = List ('datetime, 'user, 'activity, 'data,
    'session, 'location, 'response, 'device, 'error, 'server)

  val logs = Tsv("/log-files/2014/07/01", logSchema )
    .read
    .project('user, 'datetime, 'activity, 'data)
    .groupBy('user) { group => group.sortBy('datetime) }
    .write(Tsv("/analysis/log-files-2014-07-01"))
}
```


Any command-line input or output is written as follows:

```
$ echo "This is a happy day. A day to remember" > input.txt
$ hadoop fs -mkdir -p hdfs:///data/input hdfs:///data/output
$ hadoop fs -put input.txt hdfs:///data/input/
```

New terms and important words are shown in bold.



Warnings or important notes appear in a box like this.



Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or may have disliked. Reader feedback is important for us to develop titles that you really get the most out of.

To send us general feedback, simply send an e-mail to feedback@packtpub.com, and mention the book title via the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide on www.packtpub.com/authors.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for all Packt books you have purchased from your account at <http://www.packtpub.com>. If you purchased this book elsewhere, you can visit <http://www.packtpub.com/support> and register to have the files e-mailed directly to you.

Also you can access the latest code from GitHub at <https://github.com/scalding-io/ProgrammingWithScalding> or <http://scalding.io>.