



Beginning Apache Pig

Big Data Processing Made Easy

—
Balaswamy Vaddeman

Apress®

Beginning Apache Pig

Big Data Processing Made Easy



Balaswamy Vaddeman

Apress®

Beginning Apache Pig: Big Data Processing Made Easy

Balaswamy Vaddeman
Hyderabad, Andhra Pradesh, India

ISBN-13 (pbk): 978-1-4842-2336-9

ISBN-13 (electronic): 978-1-4842-2337-6

DOI 10.1007/978-1-4842-2337-6

Library of Congress Control Number: 2016961514

Copyright © 2016 by Balaswamy Vaddeman

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director: Welmoed Spahr

Lead Editor: Celestin Suresh John

Technical Reviewer: Manoj R. Patil

Editorial Board: Steve Anglin, Pramila Balan, Laura Berendson, Aaron Black,

Louise Corrigan, Jonathan Gennick, Robert Hutchinson, Celestin Suresh John,

Nikhil Karkal, James Markham, Susan McDermott, Matthew Moodie, Natalie Pao,

Gwenan Spearing

Coordinating Editor: Prachi Mehta

Copy Editor: Kim Wimpsett

Compositor: SPi Global

Indexer: SPi Global

Artist: SPi Global

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com.

Apress and friends of ED books may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Special Bulk Sales–eBook Licensing web page at www.apress.com/bulk-sales.

Any source code or other supplementary materials referenced by the author in this text are available to readers at www.apress.com. For detailed information about how to locate your book's source code, go to www.apress.com/source-code/. Readers can also access source code at SpringerLink in the Supplementary Material section for each chapter.

Printed on acid-free paper

The six most important people in my life:
The late Kammari Rangaswamy (Teacher)
The late Niranjanamma (Mother)
Devaiah (Father)
Radha (Wife)
Sai Nirupam (Son)
Nitya Maithreyi (Daughter)

Contents at a Glance

About the Author	xix
About the Technical Reviewer	xxi
Acknowledgments	xxiii
■ Chapter 1: MapReduce and Its Abstractions	1
■ Chapter 2: Data Types.....	21
■ Chapter 3: Grunt	33
■ Chapter 4: Pig Latin Fundamentals	41
■ Chapter 5: Joins and Functions.....	69
■ Chapter 6: Creating and Scheduling Workflows Using Apache Oozie.....	89
■ Chapter 7: HCatalog.....	103
■ Chapter 8: Pig Latin in Hue.....	115
■ Chapter 9: Pig Latin Scripts in Apache Falcon	123
■ Chapter 10: Macros	137
■ Chapter 11: User-Defined Functions.....	147
■ Chapter 12: Writing Eval Functions	157
■ Chapter 13: Writing Load and Store Functions.....	171
■ Chapter 14: Troubleshooting	187
■ Chapter 15: Data Formats.....	201

■ CONTENTS AT A GLANCE

■ **Chapter 16: Optimization**..... 209

■ **Chapter 17: Hadoop Ecosystem Tools**..... 225

■ **Appendix A: Built-in Functions**..... 249

■ **Appendix B: Apache Pig in Apache Ambari**..... 257

■ **Appendix C: HBaseStorage and ORCStorage Options** 261

Index..... 265

Contents

About the Author	xix
About the Technical Reviewer	xxi
Acknowledgments	xxiii
■ Chapter 1: MapReduce and Its Abstractions	1
Small Data Processing	1
Relational Database Management Systems	3
Data Warehouse Systems	3
Parallel Computing	4
GFS and MapReduce	4
Apache Hadoop	4
Problems with MapReduce	13
Cascading	13
Apache Hive.....	15
Apache Pig.....	16
Summary	20
■ Chapter 2: Data Types.....	21
Simple Data Types	22
int	22
long.....	22
float	22
double	23
chararray	23

boolean	23
bytearray	23
datetime	23
biginteger	24
bigdecimal	24
Summary of Simple Data Types	24
Complex Data Types	24
map	25
tuple	26
bag	26
Summary of Complex Data Types	27
Schema	28
Casting	28
Casting Error	29
Comparison Operators	29
Identifiers	30
Boolean Operators	31
Summary	31
■ Chapter 3: Grunt	33
Invoking the Grunt Shell	33
Commands	34
The fs Command	34
The sh Command	35
Utility Commands	36
help	36
history	36
quit	36
kill	37

set.....	37
clear.....	38
exec.....	38
run.....	39
Summary of Commands.....	39
Auto-completion.....	40
Summary.....	40
■ Chapter 4: Pig Latin Fundamentals	41
Running Pig Latin Code	41
Grunt Shell.....	41
Pig -e	42
Pig -f	42
Embed Pig Code in a Java Program.....	42
Hue	44
Pig Operators and Commands.....	44
Load.....	45
store	47
dump	48
version.....	48
Foreach Generate	48
filter	50
Limit.....	51
Assert	51
SPLIT.....	52
SAMPLE	53
FLATTEN.....	53
import.....	54
define.....	54
distinct.....	55

■ CONTENTS

RANK.....	55
Union	56
ORDER BY	57
GROUP	59
Stream	61
MAPREDUCE	62
CUBE.....	63
Parameter Substitution	65
-param.....	65
-paramfile.....	66
Summary.....	67
■ Chapter 5: Joins and Functions.....	69
Join Operators.....	70
Equi Joins	70
cogroup	72
CROSS	73
Functions.....	74
String Functions	74
Mathematical Functions	76
Date Functions.....	78
EVAL Functions	80
Complex Data Type Functions.....	81
Load/Store Functions.....	82
Summary.....	87
■ Chapter 6: Creating and Scheduling Workflows Using Apache Oozie.....	89
Types of Oozie Jobs.....	89
Workflow.....	89

Using a Pig Latin Script as Part of a Workflow	91
Writing job.properties	91
workflow.xml	91
Uploading Files to HDFS	93
Submit the Oozie Workflow	93
Scheduling a Pig Script	94
Writing the job.properties File	94
Writing coordinator.xml	94
Upload Files to HDFS	96
Submitting Coordinator.....	96
Bundle	96
oozie pig Command.....	96
Command-Line Interface.....	98
Job Submitting, Running, and Suspending.....	98
Killing Job.....	98
Retrieving Logs	98
Information About a Job	98
Oozie User Interface	99
Developing Oozie Applications Using Hue	100
Summary	100
■ Chapter 7: HCatalog.....	103
Features of HCatalog.....	103
Command-Line Interface.....	104
show Command.....	105
Data Definition Language Commands	105
dfs and set Commands.....	106

- WebHCatalog 107**
 - Executing Pig Latin Code 108
 - Running a Pig Latin Script from a File 108
 - HCatLoader Example 109
 - Writing the Job Status to a Directory 109
- HCatLoader and HCatStorer 110**
 - Reading Data from HCatalog 110
 - Writing Data to HCatalog 110
 - Running Code 111
 - Data Type Mapping 112
- Summary 113**
- Chapter 8: Pig Latin in Hue 115**
 - Pig Module 115**
 - My Scripts..... 116
 - Pig Helper 117
 - Auto-suggestion 117
 - UDF Usage in Script..... 118
 - Query History..... 118
 - File Browser 119**
 - Job Browser 121**
 - Summary 122**
- Chapter 9: Pig Latin Scripts in Apache Falcon 123**
 - cluster 124**
 - Interfaces..... 124
 - Locations 125
 - feed 126**
 - Feed Types..... 126
 - Frequency..... 126

Late Arrival.....	127
Cluster	127
process	128
cluster.....	128
Failures.....	128
feed.....	129
workflow.....	129
CLI	129
entity	129
Web Interface	130
Search	131
Create an Entity	131
Notifications	131
Mirror.....	131
Data Replication Using the Falcon Web UI.....	131
Create Cluster Entities	132
Create Mirror Job	132
Pig Scripts in Apache Falcon.....	134
Oozie Workflow	134
Pig Script	135
Summary.....	136
■ Chapter 10: Macros	137
Structure	137
Macro Use Case	138
Macro Types	138
Internal Macro	139
External Macro	140

- dryrun..... 141
- Macro Chaining 141
- Macro Rules 142
 - Define Before Usage..... 142
 - Valid Macro Chaining..... 143
 - No Macro Within Nested Block 143
 - No Grunt Shell Commands..... 143
 - Invisible Relations..... 143
- Macro Examples..... 144
 - Macro Without Input Parameters Is Possible..... 144
 - Macro Without Returning Anything Is Possible..... 144
- Summary 145
- **Chapter 11: User-Defined Functions 147**
 - User-Defined Functions..... 148
 - Java 148
 - JavaScript..... 150
 - Other Languages 152
 - Other Libraries..... 154
 - PiggyBank..... 154
 - Apache DataFu 155
 - Summary 155
- **Chapter 12: Writing Eval Functions 157**
 - MapReduce and Pig Features 157
 - Accessing the Distributed Cache..... 157
 - Accessing Counters..... 158
 - Reporting Progress..... 159
 - Output Schema and Input Schema in UDF..... 159
 - Examples of Output and Input Schemas..... 161

Other EVAL Functions	162
Algebraic.....	162
Accumulator	168
Filter Functions.....	168
Summary.....	169
■ Chapter 13: Writing Load and Store Functions.....	171
Writing a Load Function	171
Loading Metadata.....	174
Improving Loader Performance	176
Converting from bytearray.....	176
Pushing Down the Predicate	177
Writing a Store Function.....	178
Writing Metadata.....	182
Distributed Cache	183
Handling Bad Records	184
Accessing the Configuration	185
Monitoring the UDF Runtime	185
Summary.....	186
■ Chapter 14: Troubleshooting	187
Illustrate	187
describe.....	188
Dump.....	188
Explain.....	188
Plan Types.....	189
Modes.....	193
Unit Testing.....	195
Error Types	197

Counters	198
Summary	199
Chapter 15: Data Formats.....	201
Compression	201
Sequence File	202
Parquet.....	203
Parquet File Processing Using Apache Pig	204
ORC.....	205
Index	207
ACID	207
Predicate Pushdown.....	207
Data Types	207
Benefits	208
Summary	208
Chapter 16: Optimization.....	209
Advanced Joins	209
Small Files	209
User-Defined Join Using the Distributed Cache.....	210
Big Keys.....	212
Sorted Data.....	212
Best Practices	213
Choose Your Required Fields Early	213
Define the Appropriate Schema.....	213
Filter Data	214
Store Reusable Data	214
Use the Algebraic Interface	214
Use the Accumulator Interface	215
Compress Intermediate Data	215

Combine Small Inputs.....	215
Prefer a Two-Way Join over Multiway Joins.....	216
Better Execution Engine	216
Parallelism.....	216
Job Statistics.....	217
Rules	218
Partition Filter Optimizer.....	218
Merge foreach	218
Constant Calculator	219
Cluster Optimization.....	219
Disk Space	219
Separate Setup for Zookeeper.....	220
Scheduler	220
Name Node Heap Size	220
Other Memory Settings	221
Summary.....	222
■ Chapter 17: Hadoop Ecosystem Tools.....	225
Apache Zookeeper.....	225
Terminology	225
Applications	226
Command-Line Interface	227
Four-Letter Commands.....	229
Measuring Time	230
Cascading.....	230
Defining a Source	230
Defining a Sink	232
Pipes.....	233
Types of Operations	233

- Apache Spark 237
 - Core 238
 - SQL 240
- Apache Tez 245
- Presto 245
 - Architecture 246
 - Connectors 247
 - Pushdown Operations..... 247
- Summary 247
- **Appendix A: Built-in Functions 249**
- **Appendix B: Apache Pig in Apache Ambari 257**
 - Modifying Properties 258
 - Service Check 258
 - Installing Pig..... 259
 - Pig Status 259
 - Check All Available Services..... 259
 - Summary 260
- **Appendix C: HBaseStorage and ORCStorage Options 261**
 - HBaseStorage..... 261
 - Row-Based Conditions 261
 - Timestamp-Based Conditions..... 262
 - Other Conditions 262
 - OrcStorage 263
- Index 265**

About the Author



Balaswamy Vaddeman is a thinker, blogger, and serious and self-motivated big data evangelist with 10 years of experience in IT and 5 years of experience in the big data space. His big data experience covers multiple areas such as analytical applications, product development, consulting, training, book reviews, hackathons, and mentoring. He has proven himself while delivering analytical applications in the retail, banking, and finance domains in three aspects (development, administration, and architecture) of Hadoop-related technologies. At a startup company, he developed a Hadoop-based product that was used for delivering analytical applications without writing code.

In 2013 Balaswamy won the Hadoop Hackathon event for Hyderabad conducted by Cloudwick Technologies. Being the top contributor at [Stackoverflow.com](https://stackoverflow.com), he helped countless people on big data topics at multiple web sites such as [Stackoverflow.com](https://stackoverflow.com) and [Quora.com](https://quora.com). With so much passion on big data, he became an independent trainer and consultant so he could train hundreds of people and set up big data teams in several companies.

About the Technical Reviewer



Manoj R. Patil is a big data architect at TatvaSoft, an IT services and consulting firm. He has a bachelor's of engineering degree from COEP in Pune, India. He is a proven and highly skilled business intelligence professional with 17 years of information technology experience. He is a seasoned BI and big data consultant with exposure to all the leading platforms such as Java EE, .NET, LAMP, and so on. In addition to authoring a book on Pentaho and big data, he believes in knowledge sharing, keeps himself busy in corporate training, and is a passionate teacher. He can be reached at on Twitter @manojrpatil and at <https://in.linkedin.com/in/manojrpatil> on LinkedIn.

Manoj would like to thank his family, especially his two beautiful daughters, Ayushee and Ananyaa, for their patience during the review process.