

Advanced Information and Knowledge Processing

Advanced Information and Knowledge Processing

Series editors

Xindong Wu

School of Computing and Informatics

University of Louisiana at Lafayette, Lafayette, LA, USA

Lakhmi C. Jain

University of Technology Sydney, Sydney, Australia

SpringerBriefs in Advanced Information and Knowledge Processing presents concise research in this exciting field. Designed to complement Springer's *Advanced Information and Knowledge Processing* series, this Briefs series provides researchers with a forum to publish their cutting-edge research which is not yet mature enough for a book in the *Advanced Information and Knowledge Processing* series, but which has grown beyond the level of a workshop paper or journal article.

Typical topics may include, but are not restricted to:

- Big Data analytics
- Big Knowledge
- Bioinformatics
- Business intelligence
- Computer security
- Data mining and knowledge discovery
- Information quality and privacy
- Internet of things
- Knowledge management
- Knowledge-based software engineering
- Machine intelligence
- Ontology
- Semantic Web
- Smart environments
- Soft computing
- Social networks

SpringerBriefs are published as part of Springer's eBook collection, with millions of users worldwide and are available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines and expedited production schedules to assist researchers in distributing their research fast and efficiently.

More information about this series at <http://www.springer.com/series/16024>

Tomasz Wiktorski

Data-intensive Systems

Principles and Fundamentals
using Hadoop and Spark

 Springer

Tomasz Wiktorski
Department of Electrical Engineering
and Computer Science
Faculty of Science and Technology
University of Stavanger
Stavanger, Norway

ISSN 1610-3947 ISSN 2197-8441 (electronic)
Advanced Information and Knowledge Processing
ISSN 2524-5198 ISSN 2524-5201 (electronic)
SpringerBriefs in Advanced Information and Knowledge Processing
ISBN 978-3-030-04602-6 ISBN 978-3-030-04603-3 (eBook)
<https://doi.org/10.1007/978-3-030-04603-3>

Library of Congress Control Number: 2018962384

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are solely and exclusively licensed 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.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1 Preface	1
1.1 Conventions Used in this Book	2
1.2 Listed Code	2
1.3 Terminology	2
1.4 Examples and Exercises	2
2 Introduction	5
2.1 Growing Datasets	6
2.2 Hardware Trends	7
2.3 The V's of Big Data	8
2.4 NOSQL	9
2.5 Data as the Fourth Paradigm of Science	10
2.6 Example Applications	11
2.6.1 Data Hub	11
2.6.2 Search and Recommendations	12
2.6.3 Retail Optimization	13
2.6.4 Healthcare	13
2.6.5 Internet of Things	14
2.7 Main Tools	15
2.7.1 Hadoop	15
2.7.2 Spark	15
2.8 Exercises	16
References	16
3 Hadoop 101 and Reference Scenario	19
3.1 Reference Scenario	19
3.2 Hadoop Setup	21
3.3 Analyzing Unstructured Data	23
3.4 Analyzing Structured Data	28
3.5 Exercises	30

4	Functional Abstraction	31
4.1	Functional Programming Overview	31
4.2	Functional Abstraction for Data Processing	35
4.3	Functional Abstraction and Parallelism	37
4.4	Lambda Architecture	40
4.5	Exercises	40
	Reference	40
5	Introduction to MapReduce	41
5.1	Reference Code	42
5.2	Map Phase	44
5.3	Combine Phase	45
5.4	Shuffle Phase	46
5.5	Reduce Phase	47
5.6	Embarrassingly Parallel Problems	48
5.7	Running MapReduce Programs	49
5.8	Exercises	50
6	Hadoop Architecture	51
6.1	Architecture Overview	51
6.2	Data Handling	54
6.2.1	HDFS Architecture	54
6.2.2	Read Flow	55
6.2.3	Write Flow	56
6.2.4	HDFS Failovers	57
6.3	Job Handling	57
6.3.1	Job Flow	57
6.3.2	Data Locality	58
6.3.3	Job and Task Failures	60
6.4	Exercises	61
7	MapReduce Algorithms and Patterns	63
7.1	Counting, Summing, and Averaging	64
7.2	Search Assist	67
7.3	Random Sampling	70
7.4	Multiline Input	71
7.5	Inverted Index	74
7.6	Exercises	75
	References	76
8	NOSQL Databases	77
8.1	NOSQL Overview and Examples	77
8.1.1	CAP and PACELC Theorem	78
8.2	HBase Overview	79
8.3	Data Model	80

- 8.4 Architecture 80
 - 8.4.1 Regions 81
 - 8.4.2 HFile, HLog, and Memstore 81
 - 8.4.3 Region Server Failover 81
- 8.5 MapReduce and HBase 82
 - 8.5.1 Loading Data 82
 - 8.5.2 Running Queries 82
- 8.6 Exercises 83
- References 84
- 9 Spark 85**
 - 9.1 Motivation 85
 - 9.2 Data Model 86
 - 9.2.1 Resilient Distributed Datasets and DataFrames 86
 - 9.2.2 Other Data Structures 87
 - 9.3 Programming Model 88
 - 9.3.1 Data Ingestion 88
 - 9.3.2 Basic Actions—Count, Take, and Collect 89
 - 9.3.3 Basic Transformations—Filter, Map,
and reduceByKey 90
 - 9.3.4 Other Operations—flatMap and Reduce 91
 - 9.4 Architecture 93
 - 9.5 SparkSQL 95
 - 9.6 Exercises 96

List of Figures

Fig. 2.1	Examples of big datasets. <i>Source</i> Troester (2012); European Organization for Nuclear Research (2015); The Internet Archive (2015); Amazon Web Services (2015)	6
Fig. 2.2	Historical capacity versus throughput for HDDs. <i>Source</i> Leventhal (2009).	8
Fig. 2.3	Sending data versus computation	9
Fig. 2.4	The timescale of science paradigms	10
Fig. 3.1	Ambari in Hortonworks Sandbox	22
Fig. 3.2	Simple MapReduce workflow overview	25
Fig. 3.3	MapReduce program running	27
Fig. 3.4	Overview of Hive workflow	29
Fig. 3.5	Sample content from hadoop.csv file	30
Fig. 3.6	Hive program running	30
Fig. 4.1	Imperative program structure	32
Fig. 4.2	Functional program structure	33
Fig. 4.3	Parallelism of map function	39
Fig. 4.4	Lack of parallelism of reduce function	39
Fig. 5.1	Example of map phase	44
Fig. 5.2	Example of combine phase	46
Fig. 5.3	Example of shuffle phase	47
Fig. 5.4	Example of reduce phase	48
Fig. 6.1	Hadoop architecture overview.	52
Fig. 6.2	Major components of Hadoop and data-intensive systems	52
Fig. 6.3	HDFS building blocks	54
Fig. 6.4	HDFS read flow	56
Fig. 6.5	HDFS write flow	56
Fig. 6.6	Basic data flow during job execution	59
Fig. 9.1	Distribution of RDDs on a cluster	87
Fig. 9.2	Spark architecture	94

List of Listings

Listing 3.1	Hortonworks Sandbox setup.	22
Listing 3.2	Accessing Hortonworks Sandbox and copying data.	23
Listing 3.3	Sample content from hadoop.txt file.	24
Listing 3.4	Analyzing unstructured data with Hadoop streaming.	26
Listing 3.5	Analyzing unstructured data with MRJob	26
Listing 3.6	Testing code without any Hadoop installation	27
Listing 3.7	Sample content from the results of the MapReduce job.	27
Listing 3.8	Analyzing structured data with Hive	28
Listing 4.1	Execution of an imperative program in Python to extract domain names from email addresses	33
Listing 4.2	Execution of a functional program in Python to extract domain names from email addresses	34
Listing 4.3	Using lambda function in functional programs.	34
Listing 4.4	Using map high-order function	35
Listing 4.5	Using filter high-order function	35
Listing 4.6	Using reduce high-order function.	36
Listing 4.7	Another example of reduce high-order function.	36
Listing 4.8	Two parallel processes using the same variable	37
Listing 4.9	Example 1 of race condition	37
Listing 4.10	Example 2 of race condition	38
Listing 5.1	Running Hadoop job using Hadoop Streaming	42
Listing 5.2	Running Hadoop job using MRJob	42
Listing 5.3	Counting mapper for Hadoop Streaming	42
Listing 5.4	Counting reducer for Hadoop Streaming	42
Listing 5.5	Counting with MRJob	43
Listing 7.1	Counting amount of emails sent from each domain.	64
Listing 7.2	Running counting job in three different modes	65
Listing 7.3	Example output of count job	65
Listing 7.4	Finding average, maximum, and minimum amount of emails per domain	66

Listing 7.5 Running stats job in four different modes and displaying results. 67

Listing 7.6 Example output of stats job 67

Listing 7.7 Search assist, top 3 following words 68

Listing 7.8 Running search assist in four different modes and displaying results. 69

Listing 7.9 Example output of search assist job 69

Listing 7.10 Random sampling email subjects with given probability and seed. 70

Listing 7.11 Running random sampling in four different modes and displaying results. 71

Listing 7.12 Example output of random sampling job 71

Listing 7.13 Processing inputs with multiple lines 72

Listing 7.14 Running multiline input processing 73

Listing 7.15 Example output of multiline input job 73

Listing 7.16 Calculating inverted index 74

Listing 7.17 Running inverted index 75

Listing 7.18 Example output of inverted index 75

Listing 8.1 Create HBase table. 82

Listing 8.2 Import CSV file to HBase 82

Listing 8.3 Point query in HBase. 83

Listing 8.4 Range query in HBase. 83

Listing 9.1 Loading data to RDD from a variable 89

Listing 9.2 Loading data to RDD from local file system 89

Listing 9.3 Loading data to RDD from HDFS 89

Listing 9.4 Basic actions in Spark—count, take, and collect 89

Listing 9.5 Sample result after applying actions on raw data 90

Listing 9.6 Basic transformations in Spark—filter, map, and reduceByKey. 91

Listing 9.7 Sample result after applying transformations 91

Listing 9.8 Difference between map and flatMap in Spark. 92

Listing 9.9 Difference between map and flatMap in Spark—Results 92

Listing 9.10 Reduce in Spark. 93

Listing 9.11 Reduce in Spark—Results 93

Listing 9.12 Register table for in SparkSQL 95

Listing 9.13 Execute query in SparkSQL. 96

Listing 9.14 Execute query in SparkSQL—Results 96

Listing 9.15 Execute map operation on output from SparkSQL. 96

Listing 9.16 Execute map operation on output from SparkSQL—Results 96