

Contents

Part I Programming Fundamentals of High Performance Distributed Computing

1	Introduction	3
1.1	Distributed Systems	4
1.2	Types of Distributed Systems	8
1.2.1	Distributed Embedded System	8
1.2.2	Distributed Information System	11
1.2.3	Distributed Computing Systems	11
1.3	Distributed Computing Architecture	14
1.4	Distributed File Systems	15
1.4.1	DFS Requirements	16
1.4.2	DFS Architecture	17
1.5	Challenges in Distributed Systems	19
1.6	Trends in Distributed Systems	24
1.7	Examples of HPDC Systems	27
	References	30
2	Getting Started with Hadoop	33
2.1	A Brief History of Hadoop	34
2.2	Hadoop Ecosystem	35
2.3	Hadoop Distributed File System	38
2.3.1	Characteristics of HDFS	39
2.3.2	Namenode and Datanode	41
2.3.3	File System	41
2.3.4	Data Replication	42
2.3.5	Communication	44
2.3.6	Data Organization	45
2.4	MapReduce Preliminaries	46
2.5	Prerequisites for Installation	49
2.6	Single Node Cluster Installation	51

- 2.7 Multi-node Cluster Installation 56
- 2.8 Hadoop Programming 63
- 2.9 Hadoop Streaming 67
- References 71
- 3 Getting Started with Spark 73**
 - 3.1 Overview 73
 - 3.2 Spark Internals 75
 - 3.3 Spark Installation 81
 - 3.3.1 Pre-requisites 81
 - 3.3.2 Getting Started 83
 - 3.3.3 Example: Scala Application 87
 - 3.3.4 Spark with Python 90
 - 3.3.5 Example: Python Application 92
 - 3.4 Deploying Spark 93
 - 3.4.1 Submitting Applications 94
 - 3.4.2 Standalone Mode 95
 - References 99
- 4 Programming Internals of Scalding and Spark 101**
 - 4.1 Scalding 101
 - 4.1.1 Installation 101
 - 4.1.2 Programming Guide 104
 - 4.2 Spark Programming Guide 135
 - References 154
- Part II Case studies using Hadoop, Scalding and Spark**
- 5 Case Study I: Data Clustering using Scalding and Spark 157**
 - 5.1 Introduction 157
 - 5.2 Clustering 158
 - 5.2.1 Clustering Techniques 158
 - 5.2.2 Clustering Process 161
 - 5.2.3 K-Means Algorithm 162
 - 5.2.4 Simple K-Means Example 163
 - 5.3 Implementation 165
 - 5.3.1 Scalding Implementation 167
 - Problems 183
 - References 183
- 6 Case Study II: Data Classification using Scalding and Spark 185**
 - 6.1 Classification 186
 - 6.2 Probability Theory 188
 - 6.2.1 Random Variables 188
 - 6.2.2 Distributions 189

- 6.2.3 Mean and Variance 190
- 6.3 Naive Bayes 191
 - 6.3.1 Probabilty Model 191
 - 6.3.2 Parameter Estimation and Event Models 194
 - 6.3.3 Example 195
- 6.4 Implementation of Naive Bayes Classifier 197
 - 6.4.1 Scalding Implementation 199
 - 6.4.2 Results 214
- Problems 216
- References 216

- 7 Case Study III: Regression Analysis using Scalding and Spark 219**
 - 7.1 Steps in Regression Analysis 220
 - 7.2 Implementation Details 224
 - 7.2.1 Linear Regression: Algebraic Method 226
 - 7.2.2 Scalding Implementation 228
 - 7.2.3 Spark Implementation 234
 - 7.2.4 Linear Regression: Gradient Descent Method 241
 - 7.2.5 Scalding Implementation 244
 - 7.2.6 Spark Implementation 254
 - Problems 258
 - References 259

- 8 Case Study IV: Recommender System using Scalding and Spark 261**
 - 8.1 Recommender Systems 261
 - 8.1.1 Objectives 262
 - 8.1.2 Data Sources for Recommender Systems 263
 - 8.1.3 Techniques used in Recommender Systems 265
 - 8.2 Implementation Details 267
 - 8.2.1 Spark Implementation 269
 - 8.2.2 Scalding Implementation: 289
 - Problems 300
 - References 300

- Index 303**



<http://www.springer.com/978-3-319-13496-3>

Guide to High Performance Distributed Computing
Case Studies with Hadoop, Scalding and Spark

Srinivasa, K.G.; Muppalla, A.K.

2015, XVII, 304 p. 43 illus., Hardcover

ISBN: 978-3-319-13496-3