Contents

Foreword xi

Preface xiii

Acknowledgments xix

About the Author xxi

1 **Background and Concepts** 1

Defining Apache Hadoop 1

A Brief History of Apache Hadoop 3

Defining Big Data 4

Hadoop as a Data Lake 5

Using Hadoop: Administrator, User, or Both 6

First There Was MapReduce 7

Apache Hadoop Design Principles 7

Apache Hadoop MapReduce Example 8

MapReduce Advantages 10

Apache Hadoop V1 MapReduce Operation 11

Moving Beyond MapReduce with Hadoop V2 13

Hadoop V2 YARN Operation Design 13

The Apache Hadoop Project Ecosystem 15

Summary and Additional Resources 18

2 **Installation Recipes** 19

Core Hadoop Services 19

Hadoop Configuration Files 20

Planning Your Resources 21

Hardware Choices 21

Software Choices 22

Installing on a Desktop or Laptop 23

Installing Hortonworks HDP 2.2 Sandbox 23

Installing Hadoop from Apache Sources 29

Installing Hadoop with Ambari 40

Performing an Ambari Installation 42

Undoing the Ambari Install 55

Installing Hadoop in the Cloud Using Apache Whirr 56

Step 1: Install Whirr 57

Step 2: Configure Whirr 57

Step 3: Launch the Cluster 59
Step 4: Take Down Your Cluster 61
Summary and Additional Resources 62

3 Hadoop Distributed File System Basics 63
Hadoop Distributed File System Design Features 63
HDFS Components 64
HDFS Block Replication 67
HDFS Safe Mode 68
Rack Awareness 68
NameNode High Availability 69
HDFS Namespace Federation 70
HDFS Checkpoints and Backups 71
HDFS Snapshots 71
HDFS NFS Gateway 72
HDFS User Commands 72
Brief HDFS Command Reference 72
General HDFS Commands 73
List Files in HDFS 75
Make a Directory in HDFS 76
Copy Files to HDFS 76
Copy Files from HDFS 76
Copy Files within HDFS 76
Delete a File within HDFS 76
Delete a Directory in HDFS 77
Get an HDFS Status Report 77
HDFS Web GUI 77
Using HDFS in Programs 77
HDFS Java Application Example 78
HDFS C Application Example 82
Summary and Additional Resources 83

4 Running Example Programs and Benchmarks 85
Running MapReduce Examples 85
Listing Available Examples 86

Running the Pi Example 87
Using the Web GUI to Monitor Examples 89
Running Basic Hadoop Benchmarks 95
Running the Terasort Test 95
Running the TestDFSIO Benchmark 96
Managing Hadoop MapReduce Jobs 97
Summary and Additional Resources 98

5 **Hadoop MapReduce Framework** 101
The MapReduce Model 101
MapReduce Parallel Data Flow 104
Fault Tolerance and Speculative Execution 107
Speculative Execution 108
Hadoop MapReduce Hardware 108
Summary and Additional Resources 109

6 **MapReduce Programming** 111
Compiling and Running the Hadoop WordCount Example 111
Using the Streaming Interface 116
Using the Pipes Interface 119
Compiling and Running the Hadoop Grep Chaining Example 121
Debugging MapReduce 124
Listing, Killing, and Job Status 125
Hadoop Log Management 125
Summary and Additional Resources 128

7 **Essential Hadoop Tools** 131
Using Apache Pig 131
Pig Example Walk-Through 132
Using Apache Hive 134
Hive Example Walk-Through 134
A More Advanced Hive Example 136
Using Apache Sqoop to Acquire Relational Data 139
Apache Sqoop Import and Export Methods 139
Apache Sqoop Version Changes 140
Sqoop Example Walk-Through 142

Using Apache Flume to Acquire Data Streams **148**
Flume Example Walk-Through **151**
Manage Hadoop Workflows with Apache Oozie **154**
Oozie Example Walk-Through **156**
Using Apache HBase **163**
HBase Data Model Overview **164**
HBase Example Walk-Through **164**
Summary and Additional Resources **169**

8 Hadoop YARN Applications 171
YARN Distributed-Shell **171**
Using the YARN Distributed-Shell **172**
A Simple Example **174**
Using More Containers **175**
Distributed-Shell Examples with Shell Arguments **176**
Structure of YARN Applications **178**
YARN Application Frameworks **179**
Distributed-Shell **180**
Hadoop MapReduce **181**
Apache Tez **181**
Apache Giraph **181**
Hoya: HBase on YARN **181**
Dryad on YARN **182**
Apache Spark **182**
Apache Storm **182**
Apache REEF: Retainable Evaluator Execution Framework **182**
Hamster: Hadoop and MPI on the Same Cluster **183**
Apache Flink: Scalable Batch and Stream Data Processing **183**
Apache Slider: Dynamic Application Management **183**
Summary and Additional Resources **184**

9 Managing Hadoop with Apache Ambari 185
Quick Tour of Apache Ambari **186**
Dashboard View **186**

Services View 189
Hosts View 191
Admin View 193
Views View 193
Admin Pull-Down Menu 194
Managing Hadoop Services 194
Changing Hadoop Properties 198
Summary and Additional Resources 204

10 Basic Hadoop Administration Procedures 205
Basic Hadoop YARN Administration 206
Decommissioning YARN Nodes 206
YARN WebProxy 206
Using the JobHistoryServer 207
Managing YARN Jobs 207
Setting Container Memory 207
Setting Container Cores 208
Setting MapReduce Properties 208
Basic HDFS Administration 208
The NameNode User Interface 208
Adding Users to HDFS 211
Perform an FSCK on HDFS 212
Balancing HDFS 213
HDFS Safe Mode 214
Decommissioning HDFS Nodes 214
SecondaryNameNode 214
HDFS Snapshots 215
Configuring an NFSv3 Gateway to HDFS 217
Capacity Scheduler Background 220
Hadoop Version 2 MapReduce Compatibility 222
Enabling ApplicationMaster Restarts 222
Calculating the Capacity of a Node 222
Running Hadoop Version 1 Applications 224
Summary and Additional Resources 225

A Book Webpage and Code Download 227

B Getting Started Flowchart and Troubleshooting Guide 229

Getting Started Flowchart **229**

General Hadoop Troubleshooting Guide **229**

Rule 1: Don't Panic **229**

Rule 2: Install and Use Ambari **234**

Rule 3: Check the Logs **234**

Rule 4: Simplify the Situation **235**

Rule 5: Ask the Internet **235**

Other Helpful Tips **235**

C Summary of Apache Hadoop Resources by Topic 243

General Hadoop Information **243**

Hadoop Installation Recipes **243**

HDFS **244**

Examples **244**

MapReduce **245**

MapReduce Programming **245**

Essential Tools **245**

YARN Application Frameworks **246**

Ambari Administration **246**

Basic Hadoop Administration **247**

D Installing the Hue Hadoop GUI 249

Hue Installation **249**

Steps Performed with Ambari **250**

Install and Configure Hue **252**

Starting Hue **253**

Hue User Interface **253**

E Installing Apache Spark 257

Spark Installation on a Cluster **257**

Starting Spark across the Cluster **258**

Installing and Starting Spark on the Pseudo-distributed Single-Node Installation **260**

Run Spark Examples **260**

Index 261