contents

preface xv
acknowledgments xvii
about this book xviii

PART 1 BACKGROUND AND FUNDAMENTALS 1

1 *Hadoop in a heartbeat 3*

1.1 What is Hadoop? 4
1.2 Running Hadoop 14
1.3 Chapter summary 23

PART 2 DATA LOGISTICS...25

2 *Moving data in and out of Hadoop 27*

2.1 Key elements of ingress and egress 29
2.2 Moving data into Hadoop 30

TECHNIQUE 1 Pushing system log messages into HDFS with Flume 33
TECHNIQUE 2 An automated mechanism to copy files into HDFS 43
TECHNIQUE 3 Scheduling regular ingress activities with Oozie 48
TECHNIQUE 4 Database ingress with MapReduce 53
TECHNIQUE 5 Using Sqoop to import data from MySQL 58

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>

TECHNIQUE 6 HBase ingress into HDFS 68
TECHNIQUE 7 MapReduce with HBase as a data source 70
2.3 Moving data out of Hadoop 73
TECHNIQUE 8 Automated file copying from HDFS 73
TECHNIQUE 9 Using Sqoop to export data to MySQL 75
TECHNIQUE 10 HDFS egress to HBase 78
TECHNIQUE 11 Using HBase as a data sink in MapReduce 79
2.4 Chapter summary 81

3 *Data serialization—working with text and beyond 83*
3.1 Understanding inputs and outputs in MapReduce 84
3.2 Processing common serialization formats 91
TECHNIQUE 12 MapReduce and XML 91
TECHNIQUE 13 MapReduce and JSON 95
3.3 Big data serialization formats 99
TECHNIQUE 14 Working with SequenceFiles 103
TECHNIQUE 15 Integrating Protocol Buffers with MapReduce 110
TECHNIQUE 16 Working with Thrift 117
TECHNIQUE 17 Next-generation data serialization with MapReduce 120
3.4 Custom file formats 127
TECHNIQUE 18 Writing input and output formats for CSV 128
3.5 Chapter summary 136

PART 3 BIG DATA PATTERNS 137

4 *Applying MapReduce patterns to big data 139*
4.1 Joining 140
TECHNIQUE 19 Optimized repartition joins 142
TECHNIQUE 20 Implementing a semi-join 148
4.2 Sorting 155
TECHNIQUE 21 Implementing a secondary sort 157
TECHNIQUE 22 Sorting keys across multiple reducers 162
4.3 Sampling 165
TECHNIQUE 23 Reservoir sampling 165
4.4 Chapter summary 168

5 *Streamlining HDFS for big data 169*
5.1 Working with small files 170
TECHNIQUE 24 Using Avro to store multiple small files 170
5.2 Efficient storage with compression 178

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>

TECHNIQUE 25 Picking the right compression codec for your data 178
TECHNIQUE 26 Compression with HDFS, MapReduce, Pig, and Hive 182
TECHNIQUE 27 Splittable LZOP with MapReduce, Hive, and Pig 187
5.3 Chapter summary 193

6 *Diagnosing and tuning performance problems 194*
6.1 Measuring MapReduce and your environment 195
6.2 Determining the cause of your performance woes 198
TECHNIQUE 28 Investigating spikes in input data 200
TECHNIQUE 29 Identifying map-side data skew problems 201
TECHNIQUE 30 Determining if map tasks have an overall low throughput 203
TECHNIQUE 31 Small files 204
TECHNIQUE 32 Unsplittable files 206
TECHNIQUE 33 Too few or too many reducers 208
TECHNIQUE 34 Identifying reduce-side data skew problems 209
TECHNIQUE 35 Determining if reduce tasks have an overall low throughput 211
TECHNIQUE 36 Slow shuffle and sort 213
TECHNIQUE 37 Competing jobs and scheduler throttling 215
TECHNIQUE 38 Using stack dumps to discover unoptimized user code 216
TECHNIQUE 39 Discovering hardware failures 218
TECHNIQUE 40 CPU contention 219
TECHNIQUE 41 Memory swapping 220
TECHNIQUE 42 Disk health 222
TECHNIQUE 43 Networking 224
6.3 Visualization 226
TECHNIQUE 44 Extracting and visualizing task execution times 227
6.4 Tuning 229
TECHNIQUE 45 Profiling your map and reduce tasks 230
TECHNIQUE 46 Avoid the reducer 234
TECHNIQUE 47 Filter and project 235
TECHNIQUE 48 Using the combiner 236
TECHNIQUE 49 Blazingly fast sorting with comparators 237
TECHNIQUE 50 Collecting skewed data 242
TECHNIQUE 51 Reduce skew mitigation 243
6.5 Chapter summary 249

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>

PART 4 DATA SCIENCE 251

7 *Utilizing data structures and algorithms* 253

7.1 Modeling data and solving problems with graphs 254

TECHNIQUE 52 Find the shortest distance between two users 256

TECHNIQUE 53 Calculating FoFs 263

TECHNIQUE 54 Calculate PageRank over a web graph 269

7.2 Bloom filters 275

TECHNIQUE 55 Parallelized Bloom filter creation in MapReduce 277

TECHNIQUE 56 MapReduce semi-join with Bloom filters 281

7.3 Chapter summary 284

8 *Integrating R and Hadoop for statistics and more* 285

8.1 Comparing R and MapReduce integrations 286

8.2 R fundamentals 288

8.3 R and Streaming 290

TECHNIQUE 57 Calculate the daily mean for stocks 290

TECHNIQUE 58 Calculate the cumulative moving average for stocks 293

8.4 Rhipe—Client-side R and Hadoop working together 297

TECHNIQUE 59 Calculating the CMA using Rhipe 297

8.5 RHadoop—a simpler integration of client-side R and Hadoop 301

TECHNIQUE 60 Calculating CMA with RHadoop 302

8.6 Chapter summary 304

9 *Predictive analytics with Mahout* 305

9.1 Using recommenders to make product suggestions 306

TECHNIQUE 61 Item-based recommenders using movie ratings 311

9.2 Classification 314

TECHNIQUE 62 Using Mahout to train and test a spam classifier 321

9.3 Clustering with K-means 325

TECHNIQUE 63 K-means with a synthetic 2D dataset 327

9.4 Chapter summary 332

PART 5 TAMING THE ELEPHANT 333

10 *Hacking with Hive* 335

10.1 Hive fundamentals 336

10.2 Data analytics with Hive 338

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>

TECHNIQUE 64 Loading log files 338
TECHNIQUE 65 Writing UDFs and compressed partitioned tables 344
TECHNIQUE 66 Tuning Hive joins 350
10.3 Chapter summary 358

11 *Programming pipelines with Pig 359*
11.1 Pig fundamentals 360
11.2 Using Pig to find malicious actors in log data 362
TECHNIQUE 67 Schema-rich Apache log loading 363
TECHNIQUE 68 Reducing your data with filters and projection 368
TECHNIQUE 69 Grouping and counting IP addresses 370
TECHNIQUE 70 IP Geolocation using the distributed cache 375
TECHNIQUE 71 Combining Pig with your scripts 378
TECHNIQUE 72 Combining data in Pig 380
TECHNIQUE 73 Sorting tuples 381
TECHNIQUE 74 Storing data in SequenceFiles 382
11.3 Optimizing user workflows with Pig 385
TECHNIQUE 75 A four-step process to working rapidly with big data 385
11.4 Performance 390
TECHNIQUE 76 Pig optimizations 390
11.5 Chapter summary 393

12 *Crunch and other technologies 394*
12.1 What is Crunch? 395
12.2 Finding the most popular URLs in your logs 401
TECHNIQUE 77 Crunch log parsing and basic analytics 402
12.3 Joins 405
TECHNIQUE 78 Crunch's repartition join 405
12.4 Cascading 407
12.5 Chapter summary 409

13 *Testing and debugging 410*
13.1 Testing 410
TECHNIQUE 79 Unit Testing MapReduce functions, jobs, and pipelines 413
TECHNIQUE 80 Heavyweight job testing with the LocalJobRunner 421
13.2 Debugging user space problems 424
TECHNIQUE 81 Examining task logs 424
TECHNIQUE 82 Pinpointing a problem Input Split 429

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>

TECHNIQUE 83 Figuring out the JVM startup arguments for a task 433

TECHNIQUE 84 Debugging and error handling 433

13.3 MapReduce gotchas 437

TECHNIQUE 85 MapReduce anti-patterns 438

13.4 Chapter summary 441

appendix A *Related technologies 443*

appendix B *Hadoop built-in ingress and egress tools 471*

appendix C *HDFS dissected 486*

appendix D *Optimized MapReduce join frameworks 493*

index 503

Licensed to Arti Pandeya <arti.pandey@dreamtechpress.com>