Hadoop IN ACTION

Chuck Lam

MANNING

contents

preface *xiii*
acknowledgments *xv*
about this book *xvii*
Author Online *xix*
about the author *xx*
about the cover illustration *xxi*

PART 1 HADOOP–A DISTRIBUTED PROGRAMMING FRAMEWORK.. 1

1 ***Introducing Hadoop 3***

1.1 Why "Hadoop in Action"? 4
1.2 What is Hadoop? 4
1.3 Understanding distributed systems and Hadoop 6
1.4 Comparing SQL databases and Hadoop 7
1.5 Understanding MapReduce 8
Scaling a simple program manually 9 ▪ Scaling the same program in MapReduce 12
1.6 Counting words with Hadoop—running your first program 14
1.7 History of Hadoop 19

1.8 Summary 20
1.9 Resources 20

2 *Starting Hadoop 21*
2.1 The building blocks of Hadoop 21
NameNode 22 ▪ DataNode 22 ▪ Secondary NameNode 23 JobTracker 24 ▪ TaskTracker 24
2.2 Setting up SSH for a Hadoop cluster 25
Define a common account 26 ▪ Verify SSH installation 26 ▪ Generate SSH key pair 26 ▪ Distribute public key and validate logins 27
2.3 Running Hadoop 27
Local (standalone) mode 28 ▪ Pseudo-distributed mode 29 ▪ Fully distributed mode 31
2.4 Web-based cluster UI 34
2.5 Summary 36

3 *Components of Hadoop 37*
3.1 Working with files in HDFS 38
Basic file commands 38 ▪ Reading and writing to HDFS programmatically 42
3.2 Anatomy of a MapReduce program 44
Hadoop data types 46 ▪ Mapper 47 ▪ Reducer 48 ▪ Partitioner—redirecting output from Mapper 49 ▪ Combiner—local reduce 50 ▪ Word counting with predefined mapper and reducer classes 51
3.3 Reading and writing 51
InputFormat 52 ▪ OutputFormat 57
3.4 Summary 58

PART II HADOOP IN ACTION...61

4 *Writing basic MapReduce programs 63*
4.1 Getting the patent data set 64
The patent citation data 64 ▪ The patent description data 65
4.2 Constructing the basic template of a MapReduce program 67
4.3 Counting things 72
4.4 Adapting for Hadoop's API changes 77
4.5 Streaming in Hadoop 80
Streaming with Unix commands 81 ▪ Streaming with scripts 82 Streaming with key/value pairs 86 ▪ Streaming with the Aggregate package 90

4.6 Improving performance with combiners *95*

4.7 Exercising what you've learned *98*

4.8 Summary *100*

4.9 Further resources *101*

5 ***Advanced MapReduce 102***

5.1 Chaining MapReduce jobs *103*
Chaining MapReduce jobs in a sequence 103 ▪ Chaining MapReduce jobs with complex dependency 103 ▪ Chaining preprocessing and postprocessing steps 104

5.2 Joining data from different sources *107*
Reduce-side joining 108 ▪ Replicated joins using DistributedCache 117
Semijoin: reduce-side join with map-side filtering 121

5.3 Creating a Bloom filter *122*
What does a Bloom filter do? 122 ▪ Implementing a Bloom filter 124
Bloom filter in Hadoop version 0.20+ 131

5.4 Exercising what you've learned *131*

5.5 Summary *132*

5.6 Further resources *133*

6 ***Programming Practices 134***

6.1 Developing MapReduce programs *135*
Local mode 135 ▪ Pseudo-distributed mode 140

6.2 Monitoring and debugging on a production cluster *145*
Counters 146 ▪ Skipping bad records 148 ▪ Rerunning failed tasks with IsolationRunner 151

6.3 Tuning for performance *152*
Reducing network traffic with combiner 152 ▪ Reducing the amount of input data 152 ▪ Using compression 153 ▪ Reusing the JVM 155 ▪ Running with speculative execution 156 ▪ Refactoring code and rewriting algorithms 157

6.4 Summary *158*

7 ***Cookbook 160***

7.1 Passing job-specific parameters to your tasks *160*

7.2 Probing for task-specific information *163*

7.3 Partitioning into multiple output files *164*

7.4 Inputting from and outputting to a database *169*

7.5 Keeping all output in sorted order *171*

7.6 Summary *172*

8 *Managing Hadoop 173*

8.1 Setting up parameter values for practical use *174*

8.2 Checking system's health *176*

8.3 Setting permissions *178*

8.4 Managing quotas *179*

8.5 Enabling trash *179*

8.6 Removing DataNodes *180*

8.7 Adding DataNodes *180*

8.8 Managing NameNode and Secondary NameNode *181*

8.9 Recovering from a failed NameNode *183*

8.10 Designing network layout and rack awareness *184*

8.11 Scheduling jobs from multiple users *186*
Multiple JobTrackers 186 ▪ Fair Scheduler 187

8.12 Summary *189*

PART III HADOOP GONE WILD 191

9 *Running Hadoop in the cloud 193*

9.1 Introducing Amazon Web Services 194

9.2 Setting up AWS *194*
Getting your AWS authentication credentials 195 ▪ Getting command line tools 198 ▪ Preparing an SSH key pair 200

9.3 Setting up Hadoop on EC2 201
Configuring security parameters 201 ▪ Configuring cluster type 202

9.4 Running MapReduce programs on EC2 *203*
Moving your code to the Hadoop cluster 204 ▪ Accessing your data from the Hadoop cluster 204

9.5 Cleaning up and shutting down your EC2 instances *209*

9.6 Amazon Elastic MapReduce and other AWS services *209*
Amazon Elastic MapReduce 209 ▪ AWS Import/Export 210

9.7 Summary *211*

10 *Programming with Pig 212*

10.1 Thinking like a Pig *213*
Data flow language 213 ▪ Data types 213 ▪ User-defined functions 214

10.2 Installing Pig *214*

10.3 Running Pig *215*
Managing the Grunt shell 216

10.4 Learning Pig Latin through Grunt *217*

10.5 Speaking Pig Latin *221*
Data types and schemas 222 ▪ Expressions and functions 223
Relational operators 225 ▪ Execution optimization 233

10.6 Working with user-defined functions *233*
Using UDFs 234 ▪ Writing UDFs 234

10.7 Working with scripts *237*
Comments 237 ▪ Parameter substitution 238 ▪ Multiquery execution 239

10.8 Seeing Pig in action—example of computing similar patents *240*

10.9 Summary *245*

11 ***Hive and the Hadoop herd 246***

11.1 Hive *247*
Installing and configuring Hive 248 ▪ Example queries 250 ▪ HiveQL in details 254 ▪ Hive Sum-up 260

11.2 Other Hadoop-related stuff *262*
HBase 262 ▪ ZooKeeper 262 ▪ Cascading 263 ▪ Cloudera 263
Katta 263 ▪ CloudBase 264 ▪ Aster Data and Greenplum 264
Hama and Mahout 264 ▪ search-hadoop.com 264

11.3 Summary *265*

12 ***Case studies 266***

12.1 Converting 11 million image documents from the New York Times archive 267

12.2 Mining data at China Mobile 267

12.3 Recommending the best websites at StumbleUpon 272
Distributed beginnings at StumbleUpon 273 ▪ HBase and StumbleUpon 274 ▪ More Hadoop at StumbleUpon 281

12.4 Building analytics for enterprise search—IBM's Project ES2 282
ES2 architecture 285 ▪ ES2 crawler 287 ▪ ES2 analytics 288
Conclusions 296 ▪ References 297

appendix HDFS file commands 298

index 302