Download from Wow! eBook <www.wowebook.com>

contents

foreword xiii
letter to the HBase community xv
preface xvii
acknowledgments xix
about this book xxi
about the authors xxv
about the cover illustration xxvi

PART 1 HBASE FUNDAMENTALS 1

1 *Introducing HBase* 3

1.1 Data-management systems: a crash course 5

Hello, Big Data 6 ▪ *Data innovation* 7 ▪ *The rise of HBase* 8

1.2 HBase use cases and success stories 8

The canonical web-search problem: the reason for Bigtable's invention 9 ▪ *Capturing incremental data* 10 ▪ *Content serving* 13 ▪ *Information exchange* 14

1.3 Hello HBase 15

Quick install 16 ▪ *Interacting with the HBase shell* 18 *Storing data* 18

1.4 Summary 20

Download from Wow! eBook <www.wowebook.com>

2 *Getting started 21*

2.1 Starting from scratch 22

Create a table 22 ▪ Examine table schema 23 ▪ Establish a connection 24 ▪ Connection management 24

2.2 Data manipulation 25

Storing data 25 ▪ Modifying data 26 ▪ Under the hood: the HBase write path 26 ▪ Reading data 28 ▪ Under the hood: the HBase read path 29 ▪ Deleting data 30 ▪ Compactions: HBase housekeeping 30 ▪ Versioned data 31 ▪ Data model recap 32

2.3 Data coordinates *33*

2.4 Putting it all together 35

2.5 Data models 39

Logical model: sorted map of maps 39 ▪ Physical model: column family oriented 41

2.6 Table scans 42

Designing tables for scans 43 ▪ Executing a scan 45 ▪ Scanner caching 45 ▪ Applying filters 46

2.7 Atomic operations 47

2.8 ACID semantics 48

2.9 Summary 48

3 *Distributed HBase, HDFS, and MapReduce 51*

3.1 A case for MapReduce 52

Latency vs. throughput 52 ▪ Serial execution has limited throughput 53 ▪ Improved throughput with parallel execution 53 ▪ MapReduce: maximum throughput with distributed parallelism 55

3.2 An overview of Hadoop MapReduce 56

MapReduce data flow explained 57 ▪ MapReduce under the hood 61

3.3 HBase in distributed mode 62

Splitting and distributing big tables 62 ▪ How do I find my region? 64 ▪ How do I find the -ROOT- table? 65

3.4 HBase and MapReduce 68

HBase as a source 68 ▪ HBase as a sink 70 ▪ HBase as a shared resource 71

Download from Wow! eBook <www.wowebook.com>

3.5 Putting it all together 75

Writing a MapReduce application 76 ▪ Running a MapReduce application 77

3.6 Availability and reliability at scale 78

HDFS as the underlying storage 79

3.7 Summary 81

PART 2 ADVANCED CONCEPTS.....83

4 *HBase table design 85*

4.1 How to approach schema design 86

Modeling for the questions 86 ▪ Defining requirements: more work up front always pays 89 ▪ Modeling for even distribution of data and load 92 ▪ Targeted data access 98

4.2 De-normalization is the word in HBase land 100

4.3 Heterogeneous data in the same table 102

4.4 Rowkey design strategies 103

4.5 I/O considerations 104

Optimized for writes 104 ▪ Optimized for reads 106 Cardinality and rowkey structure 107

4.6 From relational to non-relational 108

Some basic concepts 109 ▪ Nested entities 110 ▪ Some things don't map 112

4.7 Advanced column family configurations 113

Configurable block size 113 ▪ Block cache 114 ▪ Aggressive caching 114 ▪ Bloom filters 114 ▪ TTL 115 Compression 115 ▪ Cell versioning 116

4.8 Filtering data 117

Implementing a filter 119 ▪ Prebundled filters 121

4.9 Summary 124

5 *Extending HBase with coprocessors 126*

5.1 The two kinds of coprocessors 127

Observer coprocessors 127 ▪ Endpoint Coprocessors 130

5.2 Implementing an observer 131

Modifying the schema 131 ▪ Starting with the Base 132 Installing your observer 135 ▪ Other installation options 137

Download from Wow! eBook <www.wowebook.com>

5.3 Implementing an endpoint 137

Defining an interface for the endpoint 138 ▪ Implementing the endpoint server 138 ▪ Implement the endpoint client 140 Deploying the endpoint server 142 ▪ Try it! 142

5.4 Summary 142

6 *Alternative HBase clients 143*

6.1 Scripting the HBase shell from UNIX 144

Preparing the HBase shell 145 ▪ Script table schema from the UNIX shell 145

6.2 Programming the HBase shell using JRuby 147

Preparing the HBase shell 147 ▪ Interacting with the TwitBase users table 148

6.3 HBase over REST 150

Launching the HBase REST service 151 ▪ Interacting with the TwitBase users table 153

6.4 Using the HBase Thrift gateway from Python 156

Generating the HBase Thrift client library for Python 157 Launching the HBase Thrift service 159 ▪ Scanning the TwitBase users table 159

6.5 Asynchbase: an alternative Java HBase client 162

Creating an asynchbase project 163 ▪ Changing TwitBase passwords 165 ▪ Try it out 176

6.6 Summary 177

PART 3 EXAMPLE APPLICATIONS.....179

7 *HBase by example: OpenTSDB 181*

7.1 An overview of OpenTSDB 182

Challenge: infrastructure monitoring 183 ▪ Data: time series 184 Storage: HBase 185

7.2 Designing an HBase application 186

Schema design 187 ▪ Application architecture 190

7.3 Implementing an HBase application 194

Storing data 194 ▪ Querying data 199

7.4 Summary 202

Download from Wow! eBook <www.wowebook.com>

8 Scaling GIS on HBase 203

8.1 Working with geographic data 203

8.2 Designing a spatial index 206

Starting with a compound rowkey 208 ▪ Introducing the geohash 209 ▪ Understand the geohash 211 ▪ Using the geohash as a spatially aware rowkey 212

8.3 Implementing the nearest-neighbors query 216

8.4 Pushing work server-side 222

Creating a geohash scan from a query polygon 224 ▪ Within query take 1: client side 228 ▪ Within query take 2: WithinFilter 231

8.5 Summary 235

PART 4 OPERATIONALIZING HBASE 237

9 Deploying HBase 239

9.1 Planning your cluster 240

Prototype cluster 241 ▪ Small production cluster (10–20 servers) 242 Medium production cluster (up to ~50 servers) 243 ▪ Large production cluster (>~50 servers) 243 ▪ Hadoop Master nodes 243 ▪ HBase Master 244 ▪ Hadoop DataNodes and HBase RegionServers 245 ZooKeeper(s) 246 ▪ What about the cloud? 246

9.2 Deploying software 248

Whirr: deploying in the cloud 249

9.3 Distributions 250

Using the stock Apache distribution 251 ▪ Using Cloudera's CDH distribution 252

9.4 Configuration 253

HBase configurations 253 ▪ Hadoop configuration parameters relevant to HBase 260 ▪ Operating system configurations 261

9.5 Managing the daemons 261

9.6 Summary 263

10 Operations 264

10.1 Monitoring your cluster 265

How HBase exposes metrics 266 ▪ Collecting and graphing the metrics 266 ▪ The metrics HBase exposes 268 ▪ Application-side monitoring 272

Download from Wow! eBook <www.wowebook.com>