Contents

Part I Foundation

1 Introduction 3
1.1 A Broad View of Data Systems 3
1.1.1 Reading Questions 7
1.2 The Sources of Data 7
1.2.1 Reading Questions 9
1.3 The Forms of Data 9
1.3.1 Reading Questions 12
1.4 Book Organization 13
1.4.1 Exercises 14

2 File Systems and File Processing 17
2.1 File Systems 18
2.1.1 Hierarchical Organization 19
2.1.2 Paths 21
2.1.3 Python File System and Path Facilities 23
2.1.4 Reading Questions 25
2.1.5 Exercises 26
2.2 File Level Operations 27
2.2.1 File Open and Close 29
2.2.2 Text File Encoding 35
2.2.3 Reading Questions 37
2.2.4 Exercises 39
2.3 Processing Files for Data 41
2.3.1 Single Data Item per Line 41
2.3.2 Multiple Data Items per Line 45
2.3.3 Reading Questions 48
2.3.4 Exercises 49
2.4 JSON File Processing 51
2.4.1 Writing Data Structures to JSON 52
2.4.2 Reading Data Structures from JSON 53

2.4.3 Reading Questions 54
2.4.4 Exercises 55

3 Python Native Data Structures 59
3.1 List Patterns 60
3.1.1 Accumulation 61
3.1.2 Unary Vector Operations 61
3.1.3 Binary Vector Operations 63
3.1.4 Filter 64
3.1.5 Reduction 65
3.1.6 Reading Questions 65
3.1.7 Exercises 67
3.2 Dictionaries 70
3.2.1 Reading Questions 72
3.2.2 Exercises 73
3.3 Python Features 76
3.3.1 Functions as Objects 76
3.3.2 Lambda Functions 78
3.3.3 List Comprehensions 80
3.3.4 Reading Questions 82
3.3.5 Exercises 84
3.4 Representing General Data Sets 88
3.4.1 Dictionary of Lists 91
3.4.2 List of Lists 92
3.4.3 List of Dictionaries 94
3.4.4 Reading Questions 96
3.4.5 Exercises 98

4 Regular Expressions 103
4.1 Motivation 104
4.1.1 Reading Questions 106
4.2 Terminology 107
4.2.1 Reading Questions 108
4.3 The Regular Expression Language 108
4.3.1 Literal Characters 109
4.3.2 Single Character Wildcard Matching 110
4.3.3 Repetition 112
4.3.4 Disjunction 113
4.3.5 Boundaries/Anchors 113
4.3.6 Grouping 114
4.3.7 Flags 115
4.3.8 Reading Questions 117
4.3.9 Exercises 118
4.4 Python Programming with Regular Expressions 120
4.4.1 Specifying Patterns 120
4.4.2 The `re` Module Interface 120

4.4.3 Reading Questions 124
4.4.4 Exercises 125

Part II Data Systems: The Data Models

5 Data Systems Models 131
5.1 Data Model Framework 131
5.1.1 Structure 132
5.1.2 Operations 133
5.1.3 Constraints 134
5.1.4 Reading Questions 135
5.2 Tabular Model Overview 135
5.2.1 Structure 136
5.2.2 Operations 137
5.2.3 Constraints 137
5.2.4 Reading Questions 138
5.3 Relational Model Overview 138
5.3.1 Structure 138
5.3.2 Operations 140
5.3.3 Constraints 141
5.3.4 Reading Questions 141
5.4 Hierarchical Model Overview 141
5.4.1 Structure 142
5.4.2 Operations 143
5.4.3 Constraints 143
5.4.4 Reading Questions 144

6 Tabular Model: Structure and Formats 145
6.1 Tidy Data 146
6.1.1 Reading Questions 151
6.1.2 Exercises 151
6.2 Tabular Data Format 153
6.2.1 Format Background 153
6.2.2 Format for Tabular Data 153
6.2.3 Tabular Format File Processing 156
6.2.4 Reading Questions 158
6.2.5 Exercises 158
6.3 Tabular Structure as `pandas` `DataFrame` 160
6.3.1 `DataFrame` Creation 161
6.3.2 Operations Involving Whole Data Frames 166
6.3.3 Reading Questions 171
6.3.4 Exercises 172

7 Tabular Model: Access Operations and `pandas` 175
7.1 Tabular Operations Overview 176
7.1.1 Access Operations 176

7.1.2 Computational Operations 177
7.1.3 Mutation Operations 177
7.1.4 Advanced Operations 178
7.1.5 Reading Questions 179
7.2 Preliminaries and Example Data Sets 179
7.2.1 Reading Questions 182
7.3 Access and Computation Operations 182
7.3.1 Single Column Projection and Vector Operations 182
7.3.2 Multicolumn Projection of a `DataFrame` 187
7.3.3 Row Selection by Slice 188
7.3.4 Row Selection by Condition 190
7.3.5 Combinations of Projection and Selection 193
7.3.6 Iteration over Rows and Columns 199
7.3.7 Reading Questions 200
7.3.8 Exercises 202

8 Tabular Model: Advanced Operations and `pandas` 205
8.1 Aggregating and Grouping Data 206
8.1.1 Aggregating Single `Series` 206
8.1.2 Aggregating a Data Frame 208
8.1.3 Aggregating Selected Rows 210
8.1.4 General Partitioning and GroupBy 212
8.1.5 Indicators Grouping Example 215
8.1.6 Reading Questions 217
8.1.7 Exercises 218
8.2 Mutation Operations for a Data Frame 219
8.2.1 Operations to Delete Columns and Rows 220
8.2.2 Operation to Add a Column 223
8.2.3 Updating Columns 225
8.2.4 Reading Questions 227
8.2.5 Exercises 227
8.3 Combining Tables 229
8.3.1 Concatenating Data Frames Along the Row Dimension . 230
8.3.2 Concatenating Data Frames Along the Column Dimension 234
8.3.3 Joining/Merging Data Frames 237
8.3.4 Reading Questions 243
8.3.5 Exercises 244
8.4 Missing Data Handling 245
8.4.1 Reading Questions 248

9 Tabular Model: Transformations and Constraints 249
9.1 Tabular Model Constraints 250
9.1.1 Reading Questions 251
9.1.2 Exercises 251

9.2 Tabular Transformations 252
9.2.1 Transpose 253
9.2.2 Melt 254
9.2.3 Pivot 260
9.2.4 Reading Questions 268
9.2.5 Exercises 270
9.3 Normalization: A Series of Vignettes 271
9.3.1 Column Values as Mashup 272
9.3.2 One Relational Mapping per Row 274
9.3.3 Columns as Values and Mashups 277
9.3.4 Exactly One Table per Logical Mapping 282
9.3.5 Reading Questions 286
9.4 Recognizing Messy Data 287
9.4.1 Focus on Each Column as Exactly One Variable (TidyData1) 287
9.4.2 Focus on Each Row Giving Exactly One Mapping (TidyData2) 288
9.4.3 Focus on Each Table Representing One Data Set (TidyData3) 288
9.4.4 Reading Questions 289
9.4.5 Exercises 290

10 Relational Model: Structure and Architecture 293
10.1 Background 295
10.1.1 Motivation and Requirements 295
10.1.2 The Relational Database Solution 298
10.1.3 Types of Relational Databases 299
10.1.4 Reading Questions 300
10.2 Structure 300
10.2.1 Single Table Characteristics 301
10.2.2 Multiple Table Characteristics 305
10.2.3 Reading Questions 308
10.3 Database Architecture 309
10.3.1 Reading Questions 312

11 Relational Model: Single Table Operations 313
11.1 Example Data Sets 315
11.1.1 Reading Questions 317
11.2 Projecting Column Fields 318
11.2.1 Single Column Field Projection 319
11.2.2 Multiple Column Field Projection 319
11.2.3 Simple Subquery 321
11.2.4 Ordering Results 322
11.2.5 Reading Questions 325
11.2.6 Exercises 325

11.3 Selecting and Filtering Rows 327
11.3.1 Uniqueness Filtering 327
11.3.2 Row Selection by Filtering 329
11.3.3 Missing Values 334
11.3.4 Additional Examples 335
11.3.5 Reading Questions 336
11.3.6 Exercises 337
11.4 Column-Vector Operations 338
11.4.1 Reading Questions 340
11.4.2 Exercises 340
11.5 Aggregation 341
11.5.1 Counting Rows for Fields 341
11.5.2 Reading Questions 343
11.5.3 Exercises 343
11.6 Partitioning and Aggregating 344
11.6.1 Reading Questions 347
11.6.2 Exercises 347

12 Relational Model: Multiple Tables Operations 349
12.1 Preliminaries and Example Data Set 350
12.1.1 Data Set: The `school` Database Schema 350
12.1.2 Table Relationships 350
12.1.3 SQL Execution Plan 353
12.1.4 Reading Questions 354
12.1.5 Exercises 355
12.2 Overview of Join Operations 356
12.3 Inner Joins 358
12.3.1 Two Table SQL Inner Join 358
12.3.2 [Optional] Cartesian Product-Based Inner Join 362
12.3.3 Inner Join to Fill Redundant Fields 363
12.3.4 Three-Table Join 365
12.3.5 Join Table from a Subquery 367
12.3.6 Reading Questions 369
12.3.7 Exercises 370
12.4 Outer Joins 371
12.4.1 Left and Right Joins 372
12.4.2 Full Outer Join 376
12.4.3 Reading Questions 379
12.4.4 Exercises 380
12.5 Partitioning and Grouping Information 380
12.5.1 Reading Questions 382
12.5.2 Exercises 383
12.6 Subqueries 384
12.6.1 Reading Questions 387
12.6.2 Exercises 388

13 Relational Model: Database Programming 391
13.1 Making Connections 393
13.1.1 The Connection String 394
13.1.2 Connecting and Closing 396
13.1.3 Reading Questions 398
13.1.4 Exercises 399
13.2 Executing Queries and Basic Retrieval of Results 400
13.2.1 Basic Query and Fetching Results 400
13.2.2 Reading Questions 404
13.2.3 Exercises 405
13.3 More Advanced Techniques 406
13.3.1 Record at a Time 406
13.3.2 Chunks 408
13.3.3 Working with Multiple Databases 411
13.3.4 Reading Questions 413
13.3.5 Exercises 414
13.4 Incorporating Variables 414
13.4.1 Python String Composition 415
13.4.2 Binding Variables 417
13.4.3 Reading Questions 420
13.4.4 Exercises 421

14 Relational Model: Design, Constraints, and Creation 425
14.1 Motivation and Process 426
14.2 Designing Tables 428
14.2.1 Functional Dependencies 428
14.2.2 Table Design: Advice and Best Practices 429
14.2.3 Table Primary Key 430
14.2.4 Reading Questions 431
14.2.5 Exercises 431
14.3 Table Fields 432
14.3.1 Single Field Issues 432
14.3.2 Field Relationship Issues 433
14.3.3 Field Data Types 435
14.3.4 Field Design: Advice and Best Practices 435
14.3.5 Reading Questions 436
14.3.6 Exercises 437
14.4 Relationships Between Tables 438
14.4.1 Designing for Many-to-One Relationships 438
14.4.2 Designing for Many-to-Many Relationships 440
14.4.3 Reading Questions 441
14.4.4 Exercises 441
14.5 Table and Schema Creation 442
14.5.1 Fields 443
14.5.2 Table Constraints 444

14.5.3 Programming and Development Advice 447
14.5.4 Reading Questions 450
14.5.5 Exercises 450
14.6 Table Population 452
14.6.1 Examples 453
14.6.2 Programming for Table Population 454
14.6.3 Reading Questions 460
14.6.4 Exercises 460

15 Hierarchical Model: Structure and Formats 463
15.1 Motivation 464
15.2 Representation of Trees 465
15.2.1 Terminology 466
15.2.2 Python Native Data Structures and Nesting 467
15.2.3 Traversals and Paths 472
15.2.4 Reading Questions 472
15.3 JSON 473
15.3.1 Reading Questions 475
15.3.2 Exercises 476
15.4 XML 477
15.4.1 XML Structure 477
15.4.2 Extracting Data from an XML File 481
15.4.3 Reading Questions 484
15.4.4 Exercises 484

16 Hierarchical Model: Operations and Programming 487
16.1 Operations Overview 488
16.1.1 Reading Questions 490
16.2 JSON Procedural Programming 490
16.2.1 Access and Traversal Operations Example 491
16.2.2 Node Creation 497
16.2.3 Node Attribute Updates 498
16.2.4 Reading Questions 500
16.2.5 Exercises 501
16.3 XML Procedural Operations 502
16.3.1 Reading and Traversing XML Data 503
16.3.2 Creating XML Data 517
16.3.3 Further Operations 521
16.3.4 Reading Questions 521
16.3.5 Exercises 522
16.4 XPath 524
16.4.1 Paths in XML Documents 525
16.4.2 Paths and Expressions in XPath 526
16.4.3 XPath Syntax 530
16.4.4 XPath Axes 530
16.4.5 XPath Predicates and Built-in Functions 533

16.4.6 Python Programming with XPath 535
16.4.7 Case Study Example 539
16.4.8 Reading Questions 543
16.4.9 Exercises 544

17 Hierarchical Model: Constraints 547
17.1 Motivation 548
17.1.1 Reading Questions 550
17.2 Well-Formed XML 550
17.2.1 Reading Questions 552
17.3 Document Type Definition 552
17.3.1 Declaring Elements 553
17.3.2 Declaring Attributes and Entities 553
17.3.3 Example DTD Declarations 555
17.3.4 DTD Validation of an XML Document 555
17.3.5 Exercises 558
17.4 XML Schema 559
17.4.1 Root of an XML Schema 559
17.4.2 Declaring Elements and Attributes 560
17.4.3 XSD Types 562
17.4.4 XSD Restrictions 564
17.4.5 An XSD Example 566
17.4.6 Validating an XML Document 568
17.4.7 Exercises 570
17.5 JSON Schema 571
17.5.1 Basics of JSON Schema 572
17.5.2 Validating a JSON Document Using a JSON Schema 576
17.5.3 Exercises 577

Part III Data Systems: The Data Sources

18 Overview of Data Systems Sources 583
18.1 Architecture 584
18.2 Data Sources 585
18.2.1 Local Files 586
18.2.2 Database Systems 587
18.2.3 Web Servers 588
18.2.4 API Service 589
18.2.5 Reading Questions 590

19 Networking and Client–Server 591
19.1 The Network Architecture 592
19.1.1 Host Addressing 594
19.1.2 Packet Switching and Routing 594
19.1.3 Summary Characteristics of the Network 595
19.1.4 Reading Questions 596

19.2 The Network Protocol Stack 597
19.2.1 Media Access Protocol Layer 599
19.2.2 Network Protocol Layer 600
19.2.3 Transport Protocol Layer 601
19.2.4 The Socket Interface 602
19.2.5 Application Protocols 603
19.2.6 Reading Questions 603
19.3 Client–Server Model 604
19.3.1 Server Application 605
19.3.2 Client Application 607
19.3.3 Reading Questions 607

20 The HyperText Transfer Protocol 609
20.1 Identifying Resources with URLs and URIs 611
20.1.1 Host Locations 611
20.1.2 Resource Paths 611
20.1.3 URL Syntax 612
20.1.4 Reading Questions 613
20.2 HTTP Definition 614
20.2.1 Message Format 615
20.2.2 Request Messages 615
20.2.3 Connections and Message Exchange 616
20.2.4 Socket Level Programming Examples 617
20.2.5 Request Header Lines 620
20.2.6 Response Messages 621
20.2.7 Redirection 624
20.2.8 Reading Questions 625
20.2.9 Exercises 626
20.3 Programming HTTP Using `requests` 628
20.3.1 GET Requests 630
20.3.2 POST Requests 632
20.3.3 Response Attributes 635
20.3.4 Reading Questions 637
20.3.5 Exercises 638
20.4 Command Line HTTP with `curl` 640
20.4.1 Basics 640
20.4.2 Sending Custom Request Header Lines 644
20.4.3 Query Parameters 645
20.4.4 POST Requests 645
20.4.5 Exploring Further 647
20.4.6 Exercises 648

21 Interlude: Client Data Acquisition 649
21.1 Encoding and Decoding 650
21.1.1 Python Strings and Bytes 652
21.1.2 Prelude to Format Examples 654

21.1.3 Reading Questions 655
21.1.4 Exercises 656
21.2 CSV Data 658
21.2.1 CSV from File Data 658
21.2.2 CSV from Network Data 659
21.2.3 Reading Questions 664
21.2.4 Exercises 664
21.3 JSON Data 666
21.3.1 JSON from File 666
21.3.2 JSON from Network 667
21.3.3 Reading Questions 671
21.3.4 Exercises 671
21.4 XML Data 673
21.4.1 XML from File Data 673
21.4.2 From Network 674
21.4.3 Reading Questions 677
21.4.4 Exercises 678

22 Web Scraping 681
22.1 HTML Structure and Its Representation of Data Sets 682
22.1.1 HTML Tables 685
22.1.2 HTML Lists 687
22.1.3 Reading Questions 689
22.2 Web Scraping Examples 692
22.2.1 Formulating Requests for HTML 693
22.2.2 Simple Table 694
22.2.3 Wikipedia Table 699
22.2.4 POST to Submit a Form 704
22.2.5 Reading Questions 711
22.2.6 Exercises 713

23 RESTful Application Programming Interfaces 715
23.1 Motivation and Background 716
23.1.1 General API Characteristics 718
General API Characteristics 718
23.1.2 Principles of REpresentational State Transfer (REST) 718
Principles of REpresentational State Transfer (REST) 718
23.1.3 Reading Questions 719
23.2 HTTP for REST API Requests 720
23.2.1 Endpoints 722
23.2.2 Path Parameters 725
23.2.3 Query Parameters 727
23.2.4 Header Parameters 731
23.2.5 POST and POST Body 732
23.2.6 Reading Questions 736
23.2.7 Exercises 738

23.3 Case Study 741
23.3.1 Phase 1: Build a Table of Popular Movies 741
23.3.2 Phase 2: Build Table of Top Cast Given Movie IDs 748
23.3.3 Summary Comments 752
23.3.4 Reading Questions 753
23.3.5 Exercises 753

24 Authentication and Authorization 757
24.1 Background 758
24.1.1 Principals 758
24.1.2 Authentication and Authorization Concepts 759
24.1.3 Impersonation 760
24.1.4 Encryption, Keys, and Signatures 761
24.1.5 Reading Questions 763
24.2 Authentication and Privacy 764
24.2.1 HTTPS 764
24.2.2 HTTP Authentication 767
24.2.3 Authentication Considerations 769
24.2.4 Reading Questions 770
24.2.5 Exercises 771
24.3 Authorization 771
24.3.1 `OAuth2` Background 772
24.3.2 Delegated Authority: Authorization Code Grant Flow 773
24.3.3 OAuth Dance Walkthrough 783
24.3.4 Reading Questions 791
24.3.5 Exercises 792

A Custom Software 797
A.1 The `util` Module 798
A.1.1 `buildURL` 798
A.1.2 `random_string` 799
A.1.3 `getLocalXML` 800
A.1.4 `read_creds` 800
A.1.5 `update_creds` 801
A.1.6 `print_text` 802
A.1.7 `print_data` 804
A.1.8 `print_xml` 805
A.1.9 `print_headers` 806
A.2 The `mysocket` Module 807
A.2.1 `makeConnection` 808
A.2.2 `sendString` 808
A.2.3 `receiveTillClose` 809
A.2.4 `sendBytes` 810
A.2.5 `receiveTillSentinel` 811
A.2.6 `receiveBySize` 812

A.2.7 `sendCRLF` 813
A.2.8 `sendCRLFLines` 814

References 815

Index 819