rinciples of Database Management

The Practical Guide to Storing, Managing and Analyzing Big and Small Data

Wilfried Lemahieu
KU Leuven, Belgium

Seppe vanden Broucke
KU Leuven, Belgium

Bart Baesens
KU Leuven, Belgium; University of Southampton, United Kingdom

J

CAMBRIDGE
UNIVERSITY PRESS

CONTENTS

About the Authors *page* xvii
Preface xix
Sober: 1000%o Driven by Technology xxiv

Part I Databases and Database Design 1

1 Fundamental Concepts of Database Management 3
1.1 Applications of Database Technology 3
1.2 Key Definitions 4
1.3 File versus Database Approach to Data Management 5
1.3.1 The File-Based Approach 5
1.3.2 The Database Approach 6
1.4 Elements of a Database System 8
1.4.1 Database Model versus Instances 8
1.4.2 Data Model 9
1.4.3 The Three-Layer Architecture 10
1.4.4 Catalog 10
1.4.5 Database Users 11
1.4.6 Database Languages 12
1.5 Advantages of Database Systems and Database Management 12
1.5.1 Data Independence 12
1.5.2 Database Modeling 13
1.5.3 Managing Structured, Semi-Structured, and Unstructured Data 13
1.5.4 Managing Data Redundancy 14
1.5.5 Specifying Integrity Rules 14
1.5.6 Concurrency Control 14
1.5.7 Backup and Recovery Facilities 15
1.5.8 Data Security 15
1.5.9 Performance Utilities 16
Summary 16
Key Terms List 16
Review Questions 17
Problems and Exercises 19

2 Architecture and Categorization of DBMSs 20
2.1 Architecture of a DBMS 20
2.1.1 Connection and Security Manager 21
2.1.2 DDL Compiler 22
2.1.3 Query Processor 22
2.1.3.1 DML Compiler 22
2.1.3.2 Query Parser and Query Rewriter 25
2.1.3.3 Query Optimizer 25
2.1.3.4 Query Executor 25
2.1.4 Storage Manager 25
2.1.4.1 Transaction Manager 25
2.1.4.2 Buffer Manager 26
2.1.4.3 Lock Manager 26
2.1.4.4 Recovery Manager 26
2.1.5 DBMS Utilities 26
2.1.6 DBMS Interfaces 27
2.2 Categorization of DBMSs 27
2.2.1 Categorization Based on Data Model 28
2.2.1.1 Hierarchical DBMSs 28
2.2.1.2 Network DBMSs 28
2.2.1.3 Relational DBMSs 28
2.2.1.4 Object-Oriented DBMSs 28
2.2.1.5 Object-Relational/Extended Relational DBMSs 29
2.2.1.6 XML DBMSs 29
2.2.1.7 NoSQL DBMSs 30
2.2.2 Categorization Based on Degree of Simultaneous Access 30
2.2.3 Categorization Based on Architecture 30
2.2.4 Categorization Based on Usage 31
Summary 33
Key Terms List 33
Review Questions 34
Problems and Exercises 37

3 Conceptual Data Modeling Using the (E)ER Model and UML Class Diagram 38
3.1 Phases of Database Design 38
3.2 The Entity Relationship Model 40
3.2.1 Entity Types 40
3.2.2 Attribute Types 40
3.2.3.1 Domains 41
3.2.3.2 Key Attribute Types 42
3.2.3.3 Simple versus Composite Attribute Types 42
3.2.3.4 Single-Valued versus Multi-Valued Attribute Types 43
3.2.3.5 Derived Attribute Type 43
3.2.4 Relationship Types 43
3.2.4.1 Degree and Roles 44
3.2.4.2 Cardinalities 45
3.2.4.3 Relationship Attribute Types 46
3.2.5 Weak Entity Types 46
3.2.6 Ternary Relationship Types 48
3.2.7 Examples of the ER Model 50
3.2.8 Limitations of the ER Model 51
3.3 The Enhanced Entity Relationship (EER) Model 52
3.3.1 Specialization/Generalization 52
3.3.2 Categorization 54
3.3.3 Aggregation 55
3.3.4 Examples of the EER Model 55
3.3.5 Designing an EER Model 56
3.4 The UML Class Diagram 57
3.4.1 Recap of Object Orientation 57
3.4.2 Classes 58
3.4.3 Variables 58
3.4.4 Access Modifiers 59
3.4.5 Associations 59
3.4.5.1 Association Class 60
3.4.5.2 Unidirectional versus Bidirectional Association 60
3.4.5.3 Qualified Association 61
3.4.6 Specialization/Generalization 62
3.4.7 Aggregation 62
3.4.8 UML Example 63
3.4.9 Advanced UML Modeling Concepts 64
3.4.9.1 Changeability Property 64
3.4.9.2 Object Constraint Language (OCL) 64
3.4.9.3 Dependency Relationship 66
3.4.10 UML versus EER 66
Summary 67
Key Terms List 71
Review Questions 71
Problems and Exercises 75

4 Organizational Aspects of Data Management 79
4.1 Data Management 79
4.1.1 Catalogs and the Role of Metadata 80
4.1.2 Metadata Modeling 80
4.1.3 Data Quality 81
4.1.3.1 Data Quality Dimensions 82
4.1.3.2 Data Quality Problems 84
4.1.4 Data Governance 85
4.2 Roles in Data Management 86
4.2.1 Information Architect 86
4.2.2 Database Designer 87
4.2.3 Data Owner 87
4.2.4 Data Steward 87
4.2.5 Database Administrator 87
4.2.6 Data Scientist 88
Summary 88
Key Terms List 89
Review Questions 89
Problems and Exercises 90

Part II Types of Database Systems 91

5 Legacy Databases 93
5.1 The Hierarchical Model 93
5.2 The CODASYL Model 97
Summary 102
Key Terms List 102
Review Questions 102
Problems and Exercises 103

6 Relational Databases: The Relational Model 104
6.1 The Relational Model 105
6.1.1 Basic Concepts 105
6.1.2 Formal Definitions 106
6.1.3 Types of Keys 108
6.1.3.1 Superkeys and Keys 108
6.1.3.2 Candidate Keys, Primary Keys, and Alternative Keys 108
6.1.3.3 Foreign Keys 109
6.1.4 Relational Constraints 111
6.1.5 Example Relational Data Model 111

6.2 Normalization 111
6.2.1 Insertion, Deletion, and Update Anomalies in an Unnormalized Relational Model 112
6.2.2 Informal Normalization Guidelines 114
6.2.3 Functional Dependencies and Prime Attribute Type 114
6.2.4 Normalization Forms 115
6.2.4.1 First Normal Form (1 NF) 115
6.2.4.2 Second Normal Form (2 NF) 117
6.2.4.3 Third Normal Form (3 NF) 118
6.2.4.4 Boyce-Codd Normal Form (BCNF) 119
6.2.4.5 Fourth Normal Form (4NF) 120
6.3 Mapping a Conceptual ER Model to a Relational Model 121
6.3.1 Mapping Entity Types 121
6.3.2 Mapping Relationship Types 122
6.3.2.1 Mapping a Binary 1:1 Relationship type 122
6.3.2.2 Mapping a Binary 1:N Relationship Type 124
6.3.2.3 Mapping a Binary M:N Relationship Type 126
6.3.2.4 Mapping Unary Relationship Types 127
6.3.2.5 Mapping n-ary Relationship Types 129
6.3.3 Mapping Multi-Valued Attribute Types 130
6.3.4 Mapping Weak Entity Types 131
6.3.5 Putting it All Together 132
6.4 Mapping a Conceptual EER Model to a Relational Model 133
6.4.1 Mapping an EER Specialization 133
6.4.2 Mapping an EER Categorization 136
6.4.3 Mapping an EER Aggregation 137
Summary 138
Key Terms List 139
Review Questions 139
Problems and Exercises 143

7 Relational Databases: Structured Query Language (SQL) 146
7.1 Relational Database Management Systems and SQL 147
7.1.1 Key Characteristics of SQL 147
7.1.2 Three-Layer Database Architecture 149
7.2 SQL Data Definition Language 149
7.2.1 Key DDL Concepts 150
7.2.2 DDL Example 151
7.2.3 Referential Integrity Constraints 154
7.2.4 DROP and ALTER Command 155
7.3 SQL Data Manipulation Language 156
7.3.1 SQL SELECT Statement 156
7.3.1.1 Simple Queries 157
7.3.1.2 Queries with Aggregate Functions 161
7.3.1.3 Queries with GROUP BY/HAVING 163
7.3.1.4 Queries with ORDER BY 165
7.3.1.5 Join Queries 166
7.3.1.6 Nested Queries 172
7.3.1.7 Correlated Queries 175
7.3.1.8 Queries with ALL/ANY 178
7.3.1.9 Queries with EXISTS 181
7.3.1.10 Queries with Subqueries in SELECT/FROM 182
7.3.1.11 Queries with Set Operations 183
7.3.2 SQL INSERT Statement 185
7.3.3 SQL DELETE Statement 185
7.3.4 SQL UPDATE Statement 186
7.4 SQL Views 188
7.5 SQL Indexes 190
7.6 SQL Privileges 191
7.7 SQL for Metadata Management 192
Summary 194
Key Terms List 195
Review Questions 196
Problems and Exercises 205

8 Object-Oriented Databases and Object Persistence 207
8.1 Recap: Basic Concepts of OO 208
8.2 Advanced Concepts of OO 209
8.2.1 Method Overloading 209
8.2.2 Inheritance 210
8.2.3 Method Overriding 211
8.2.4 Polymorphism and Dynamic Binding 212
8.3 Basic Principles of Object Persistence 214
8.3.1 Serialization 214

8.4 OODBMS 216
8.4.1 Object Identifiers 216
8.42 ODMG Standard 217
8.4.3 Object Model 217
8.4.4 Object Definition Language (ODL) 218
8.4.5 Object Query Language (OQL) 221
8.4.5.1 Simple OQL Queries 221
8.4.5.2 SELECT FROM WHERE OQL Queries 221
8.4.5.3 Join OQL Queries 222
8.4.5.4 Other OQL Queries 222
8.4.6 Language Bindings 223
8.5 Evaluating OODBMSs 225
Summary 227
Key Terms List 227
Review Questions 228
Problems and Exercises 229

9 Extended Relational Databases 231
9.1 Limitations of the Relational Model 231
9.2 Active RDBMS Extensions 232
9.2.1 Triggers 233
9.2.2 Stored Procedures 234
9.3 Object-Relational RDBMS Extensions 236
9.3.1 User-Defined Types 236
9.3.1.1 Distinct Data Types 237
9.3.1.2 Opaque Data Types 238
9.3.1.3 Unnamed Row Types 238
9.3.1.4 Named Row Types 239
9.3.1.5 Table Data Types 240
9.3.2 User-Defined Functions 240
9.3.3 Inheritance 242
9.3.3.1 Inheritance at Data Type Level 242
9.3.3.2 Inheritance at Table Type Level 243
9.3.4 Behavior 244
9.3.5 Polymorphism 244
9.3.6 Collection Types 245
9.3.7 Large Objects 247
9.4 Recursive SQL Queries 247
Summary 250
Key Terms List 251
Review Questions 252
Problems and Exercises 253

10 XML Databases 255
10.1 Extensible Markup Language 256
10.1.1 Basic Concepts 256
10.1.2 Document Type Definition and XML Schema Definition 260
10.1.3 Extensible Stylesheet Language 263
10.1.4 Namespaces 266
10.1.5 XPath 267
10.2 Processing XML Documents 267
10.3 Storage of XML Documents 269
10.3.1 The Document-Oriented Approach for Storing XML Documents 270
10.3.2 The Data-Oriented Approach for Storing XML Documents 270
10.3.3 The Combined Approach for Storing XML Documents 270
10.4 Differences Between XML Data and Relational Data 271
10.5 Mappings Between XML Documents and (Object-) Relational Data 272
10.5.1 Table-Based Mapping 272
10.5.2 Schema-Oblivious Mapping 273
10.5.3 Schema-Aware Mapping 275
10.5.4 SQL/XML 276
10.6 Searching XML Data 279
10.6.1 Full-Text Search 280
10.6.2 Keyword-Based Search 280
10.6.3 Structured Search With XQuery 280
10.6.4 Semantic Search With RDF and SPARQL 282
10.7 XML for Information Exchange 284
10.7.1 Message-Oriented Middleware 284
10.7.2 SOAP-Based Web Services 285
10.7.3 REST-Based Web Services 288
10.7.4 Web Services and Databases 289
10.8 Other Data Representation Formats 290
Summary 293
Key Terms List 296
Review Questions 297
Problems and Exercises 298

11 NoSQL Databases 300
11.1 The NoSQL Movement 301
11.1.1 The End of the "One Size Fits All" Era? 301

11.1.2 The Emergence of the NoSQL Movement 302
11.2 Key-Value Stores 304
11.2.1 From Keys to Hashes 304
11.2.2 Horizontal Scaling 305
11.2.3 An Example: Memcached 306
11.2.4 Request Coordination 308
11.2.5 Consistent Hashing 309
11.2.6 Replication and Redundancy 311
11.2.7 Eventual Consistency 312
11.2.8 Stabilization 314
11.2.9 Integrity Constraints and Querying 314
11.3 Tuple and Document Stores 315
11.3.1 Items with Keys 316
11.3.2 Filters and Queries 316
11.3.3 Complex Queries and Aggregation with MapReduce 320
11.3.4 SQL After All... 330
11.4 Column-Oriented Databases 331
11.5 Graph-Based Databases 333
11.5.1 Cypher Overview 335
11.5.2 Exploring a Social Graph 336
11.6 Other NoSQL Categories 341
Summary 342
Key Terms 344
Review Questions 345
Problems and Exercises 347

Part III Physical Data Storage, Transaction Management, and Database Access 349

12 Physical File Organization and Indexing 351
12.1 Storage Hardware and Physical Database Design 351
12.1.1 The Storage Hierarchy 352
12.1.2 Internals of Hard Disk Drives 353
12.1.3 From Logical Concepts to Physical Constructs 356
12.2 Record Organization 359
12.3 File Organization 361
12.3.1 Introductory Concepts: Search Keys, Primary, and Secondary File Organization 362
12.3.2 Heap File Organization 363
12.3.3 Sequential File Organization 363
12.3.4 Random File Organization (Hashing) 365
12.3.4.1 Key-to-Address Transformation 365
12.3.4.2 Factors that Determine the Efficiency of Random File Organization 368
12.3.5 Indexed Sequential File Organization 370
12.3.5.1 Basic Terminology of Indexes 370
12.3.5.2 Primary Indexes 371
12.3.5.3 Clustered Indexes 373
12.3.5.4 Multilevel Indexes 374
12.3.6 List Data Organization (Linear and Nonlinear Lists) 375
12.3.6.1 Linear Lists 375
12.3.6.2 Tree Data Structures 377
12.3.7 Secondary Indexes and Inverted Files 379
12.3.7.1 Characteristics of Secondary Indexes 380
12.3.7.2 Inverted Files 381
12.3.7.3 Multicolumn Indexes 382
12.3.7.4 Other Index Types 383
12.3.8 B-Trees and B+-Trees 384
12.3.8.1 Multilevel Indexes Revisited 384
12.3.8.2 Binary Search Trees 385
12.3.8.3 B-Trees 386
12.3.8.4 B+-Trees 388
Summary 390
Key Terms List 391
Review Questions 392
Problems and Exercises 393

13 Physical Database Organization 395
13.1 Physical Database Organization and Database Access Methods 396
13.1.1 From Database to Tablespace 396
13.1.2 Index Design 398
13.1.3 Database Access Methods 400
13.1.3.1 Functioning of the Query Optimizer 400
13.1.3.2 Index Search (with Atomic Search Key) 402
13.1.3.3 Multiple Index and Multicolumn Index Search 403
13.1.3.4 Index-Only Access 407
13.1.3.5 Full Table Scan 408

13.1.4 Join Implementations 408
13.1.4.1 Nested-Loop Join 409
13.1.4.2 Sort-Merge Join 410
13.1.4.3 Hash Join 410
13.2 Enterprise Storage Subsystems and Business Continuity 411
13.2.1 Disk Arrays and RAID 411
13.2.2 Enterprise Storage Subsystems 413
13.2.2.1 Overview and Classification 414
13.2.2.2 DAS (Directly Attached Storage) 416
13.2.2.3 SAN (Storage Area Network) 416
13.2.2.4 NAS (Network Attached Storage) 417
13.2.2.5 NAS Gateway 418
13.2.2.6 iSCSI/Storage Over IP 419
13.2.3 Business Continuity 421
13.2.3.1 Contingency Planning, Recovery Point, and Recovery Time 421
13.2.3.2 Availability and Accessibility of Storage Devices 422
13.2.3.3 Availability of Database Functionality 422
13.2.3.4 Data Availability 423
Summary 426
Key Terms List 426
Review Questions 427
Problems and Exercises 429

14 Basics of Transaction Management 430
14.1 Transactions, Recovery, and Concurrency Control 431
14.2 Transactions and Transaction Management 432
14.2.1 Delineating Transactions and the Transaction Lifecycle 432
14.2.2 DBMS Components Involved in Transaction Management 433
14.2.3 The Logfile 435
14.3 Recovery 436
14.3.1 Types of Failures 436
14.3.2 System Recovery 436
14.3.3 Media Recovery 438
14.4 Concurrency Control 439
14.4.1 Typical Concurrency Problems 439
14.4.1.1 Lost Update Problem 440
14.4.1.2 Uncommitted Dependency Problem (aka Dirty Read Problem) 440
14.4.1.3 Inconsistent Analysis Problem 441
14.4.1.4 Other Concurrency-Related Problems 442
14.4.2 Schedules and Serial Schedules 442
14.4.3 Serializable Schedules 442
14.4.4 Optimistic and Pessimistic Schedulers 443
14.4.5 Locking and Locking Protocols 444
14.4.5.1 Purposes of Locking 444
14.4.5.2 The Two-Phase Locking Protocol (2PL) 446
14.4.5.3 Cascading Rollbacks 447
14.4.5.4 Dealing with Deadlocks 448
14.4.5.5 Isolation Levels 449
14.4.5.6 Lock Granularity 450
14.5 The ACID Properties of Transactions 452
Summary 453
Key Terms List 453
Review Questions 454
Problems and Exercises 456

15 Accessing Databases and Database APIs 458
15.1 Database System Architectures 459
15.1.1 Centralized System Architectures 459
15.1.2 Tiered System Architectures 460
15.2 Classification of Database APIs 462
15.2.1 Proprietary versus Universal APIs 463
15.2.2 Embedded versus Call-Level APIs 464
15.2.3 Early Binding versus Late Binding 465

15.3 Universal Database APIs 466
15.3.1 ODBC 466
15.3.2 OLE DB and ADO 467
15.3.3 ADO.NET 468
15.3.4 Java DataBase Connectivity (JDBC) 471
15.3.5 Intermezzo: SQL Injection and Access Security 477
15.3.6 SQLJ 479
15.3.7 Intermezzo: Embedded APIs versus Embedded DBMSs 480
15.3.8 Language-Integrated Querying 482
15.4 Object Persistence and Object-Relational Mapping APIs 483
15.4.1 Object Persistence with Enterprise JavaBeans 484
15.4.2 Object Persistence with the Java Persistence API 488
15.4.3 Object Persistence with Java Data Objects 495
15.4.4 Object Persistence in Other Host Languages 498
15.5 Database API Summary 502
15.6 Database Access in the World Wide Web 504
15.6.1 Introduction: the Original Web Server 504
15.6.2 The Common Gateway Interface: Toward Dynamic Web Pages 504
15.6.3 Client-Side Scripting: The Desire for a Richer Web 507
15.6.4 JavaScript as a Platform 508
15.6.5 DBMSs Adapt: REST. Other Web Services, and a Look Ahead 509
Summary 512
Key Terms List 513
Review Questions 513
Problems and Exercises 515

16 Data Distribution and Distributed Transaction Management 516
16.1 Distributed Systems and Distributed Databases 517
16.2 Architectural Implications of Distributed Databases 518
16.3 Fragmentation, Allocation, and Replication 519
16.3.1 Vertical Fragmentation 520
16.3.2 Horizontal Fragmentation (Sharding) 521
16.3.3 Mixed Fragmentation 521
16.3.4 Replication 523
16.3.5 Distribution and Replication of Metadata 524
16.4 Transparency 524
16.5 Distributed Query Processing 525
16.6 Distributed Transaction Management and Concurrency Control 528
16.6.1 Primary Site and Primary Copy 2PL 529
16.6.2 Distributed 2PL 529
16.6.3 The Two-Phase Commit Protocol (2PC) 530
16.6.4 Optimistic Concurrency and Loosely Coupled Systems 532
16.6.5 Compensation-Based Transaction Models 534
16.7 Eventual Consistency and BASE Transactions 538
16.7.1 Horizontal Fragmentation and Consistent Hashing 538
16.7.2 The CAP Theorem 539
16.7.3 BASE Transactions 540
16.7.4 Multi-Version Concurrency Control and Vector Clocks 541
16.7.5 Quorum-Based Consistency 542
Summary 544
Key Terms 545
Review Questions 546
Problems and Exercises 547

Part IV Data Warehousing, Data Governance, and (Big) Data Analytics 549

17 Data Warehousing and Business Intelligence 551
17.1 Operational versus Tactical/ Strategic Decision-Making 552
17.2 Data Warehouse Definition 553
17.3 Data Warehouse Schemas 554
17.3.1 Star Schema 555

17.3.2 Snowflake Schema 556
17.3.3 Fact Constellation 557
17.3.4 Specific Schema Issues 557
17.3.4.1 Surrogate keys 557
17.3.4.2 Granularity of the Fact Table 558
17.3.4.3 Factless Fact Tables 559
17.3.4.4 Optimizing the Dimension Tables 559
17.3.4.5 Defining Junk Dimensions 560
17.3.4.6 Defining Outrigger Tables 561
17.3.4.7 Slowly Changing Dimensions 561
17.3.4.8 Rapidly Changing Dimensions 563
17.4 The Extraction, Transformation, and Loading (ETL) Process 565
17.5 Data Marts 567
17.6 Virtual Data Warehouses and Virtual Data Marts 569
17.7 Operational Data Store 571
17.8 Data Warehouses versus Data Lakes 571
17.9 Business Intelligence 572
17.9.1 Query and Reporting 573
17.9.2 Pivot Tables 573
17.9.3 On-Line Analytical Processing (OLAP) 574
17.9.3.1 MOLAP 574
17.9.3.2 ROLAP 575
17.9.3.3 HOLAP 575
17.9.3.4 OLAP Operators 575
17.9.3.5 OLAP Queries in SQL 577
Summary 583
Key Terms List 584
Review Questions 585
Problems and Exercises 587

18 Data Integration, Data Quality, and Data Governance 590
18.1 Data and Process Integration 591
18.1.1 Convergence of Analytical and Operational Data Needs 591
18.1.2 Data Integration and Data Integration Patterns 593
18.1.2.1 Data Consolidation: Extract, Transform, Load (ETL) 593
18.1.2.2 Data Federation: Enterprise Information Integration (EII) 595
18.1.2.3 Data Propagation: Enterprise Application Integration (EAI) 596
18.1.2.4 Data Propagation: Enterprise Data Replication (EDR) 597
18.1.2.5 Changed Data Capture (CDC), Near-Real-Time ETL, and Event Processing 598
18.1.2.6 Data Virtualization 598
18.1.2.7 Data as a Service and Data in the Cloud 599
18.1.3 Data Services and Data Flows in the Context of Data and Process Integration 601
18.1.3.1 Business Process Integration 602
18.1.3.2 Patterns for Managing Sequence Dependencies and Data Dependencies in Processes 604
18.1.3.3 A Unified View on Data and Process Integration 606
18.2 Searching Unstructured Data and Enterprise Search 610
18.2.1 Principles of Full-Text Search 610
18.2.2 Indexing Full-Text Documents 611
18.2.3 Web Search Engines 613
18.2.4 Enterprise Search 616
18.3 Data Quality and Master Data Management 617
18.4 Data Governance 618
18.4.1 Total Data Quality Management (TDQM) 619
18.4.2 Capability Maturity Model Integration (CMMI) 619
18.4.3 Data Management Body of Knowledge (DMBOK) 620

18.4.4 Control Objectives for Information and Related Technology (COBIT) 620
18.4.5 Information Technology Infrastructure Library 621
18.5 Outlook 621
18.6 Conclusion 622
Key Terms List 622
Review Questions 623
Problems and Exercises 625

19 Big Data 626
19.1 The 5 Vs of Big Data 627
19.2 Hadoop 630
19.2.1 History of Hadoop 630
19.2.2 The Hadoop Stack 631
19.2.2.1 The Hadoop Distributed File System 631
19.2.2.2 MapReduce 635
19.2.2.3 Yet Another Resource Negotiator 641
19.3 SQL on Hadoop 643
19.3.1 HBase: The First Database on Hadoop 644
19.3.2 Pig 648
19.3.3 Hive 649
19.4 Apache Spark 652
19.4.1 Spark Core 653
19.4.2 Spark SQL 654
19.4.3 MLlib, Spark Streaming, and GraphX 656
19.5 Conclusion 659
Key Terms List 660
Review Questions 660
Problems and Exercises 662

20 Analytics 664
20.1 The Analytics Process Model 665
20.2 Example Analytics Applications 667
20.3 Data Scientist Job Profile 668
20.4 Data Pre-Processing 669
20.4.1 Denormalizing Data for Analysis 669
20.4.2 Sampling 670
20.4.3 Exploratory Analysis 671
20.4.4 Missing Values 671
20.4.5 Outlier Detection and Handling 672
20.5 Types of Analytics 673
20.5.1 Predictive Analytics 673
20.5.1.1 Linear Regression 673
20.5.1.2 Logistic Regression 675
20.5.1.3 Decision Trees 677
20.5.1.4 Other Predictive Analytics Techniques 681
20.5.2 Evaluating Predictive Models 682
20.5.2.1 Splitting Up the Dataset 682
20.5.2.2 Performance Measures for Classification Models 684
20.5.2.3 Performance Measures for Regression Models 687
20.5.2.4 Other Performance Measures for Predictive Analytical Models 688
20.5.3 Descriptive Analytics 689
20.5.3.1 Association Rules 689
20.5.3.2 Sequence Rules 691
20.5.3.3 Clustering 692
20.5.4 Social Network Analytics 695
20.5.4.1 Social Network Definitions 696
20.5.4.2 Social Network Metrics 696
20.5.4.3 Social Network Learning 699
20.6 Post-Processing of Analytical Models 700
20.7 Critical Success Factors for Analytical Models 701
20.8 Economic Perspective on Analytics 702
20.8.1 Total Cost of Ownership (TCO) 702
20.8.2 Return on Investment 702
20.8.3 In- versus Outsourcing 704
20.8.4 On-Premises versus Cloud Solutions 705
20.8.5 Open-Source versus Commercial Software 706

20.9 Improving the ROI of Analytics 708
20.9.1 New Sources of Data 708
20.9.2 Data Quality 711
20.9.3 Management Support 712
20.9.4 Organizational Aspects 712
20.9.5 Cross-Fertilization 713
20.10 Privacy and Security 714
20.10.1 Overall Considerations Regarding Privacy and Security 714
20.10.2 The RACI Matrix 715
20.10.3 Accessing Internal Data 716
20.10.3.1 Anonymization 717
20.10.3.2 SQL Views 719
20.10.3.3 Label-Based Access Control 719
20.10.4 Privacy Regulation 721
20.11 Conclusion 723
Key Terms List 724
Review Questions 725
Problems and Exercises 729

Appendix Using the Online Environment 731
Glossary 741
Index 770