CONTENTS

PREFACE xxiv

Part I FOUNDATIONS 1

1 OVERVIEW OF DATABASE SYSTEMS 3
1.1 Managing Data 4
1.2 A Historical Perspective 6
1.3 File Systems versus a DBMS 8
1.4 Advantages of a DBMS 9
1.5 Describing and Storing Data in a DBMS 10
1.5.1 The Relational Model 11
1.5.2 Levels of Abstraction in a DBMS 12
1.5.3 Data Independence 15
1.6 Queries in a DBMS 16
1.7 Transaction Management 17
1.7.1 Concurrent Execution of Transactions 17
1.7.2 Incomplete Transactions and System Crashes 18
1.7.3 Points to Note 19
1.8 Structure of a DBMS 19
1.9 People Who Work with Databases 21
1.10 Review Questions 22

2 INTRODUCTION TO DATABASE DESIGN 25
2.1 Database Design and ER Diagrams 26
2.1.1 Beyond ER Design 27
2.2 Entities, Attributes, and Entity Sets 28
2.3 Relationships and Relationship Sets 29
2.4 Additional Features of the ER Model 32
2.4.1 Key Constraints 32
2.4.2 Participation Constraints 34
2.4.3 Weak Entities 35
2.4.4 Class Hierarchies 37
2.4.5 Aggregation 39

2.5 Conceptual Design With the ER Model 40
2.5.1 Entity versus Attribute 41
2.5.2 Entity versus Relationship 42
2.5.3 Binary versus Ternary Relationships 43
2.5.4 Aggregation versus Ternary Relationships 45
2.6 Conceptual Design for Large Enterprises 46
2.7 The Unified Modeling Language 47
2.8 Case Study: The Internet Shop 49
2.8.1 Requirements Analysis 49
2.8.2 Conceptual Design 50
2.9 Review Questions 51

3 **THE RELATIONAL MODEL** **57**
3.1 Introduction to the Relational Model 59
3.1.1 Creating and Modifying Relations Using SQL 62
3.2 Integrity Constraints over Relations 63
3.2.1 Key Constraints 64
3.2.2 Foreign Key Constraints 66
3.2.3 General Constraints 68
3.3 Enforcing Integrity Constraints 69
3.3.1 Transactions and Constraints 72
3.4 Querying Relational Data 73
3.5 Logical Database Design: ER to Relational 74
3.5.1 Entity Sets to Tables 75
3.5.2 Relationship Sets (without Constraints) to Tables 76
3.5.3 Translating Relationship Sets with Key Constraints 78
3.5.4 Translating Relationship Sets with Participation Constraints 79
3.5.5 Translating Weak Entity Sets 82
3.5.6 Translating Class Hierarchies 83
3.5.7 Translating ER Diagrams with Aggregation 84
3.5.8 ER to Relational: Additional Examples 85
3.6 Introduction to Views 86
3.6.1 Views, Data Independence, Security 87
3.6.2 Updates on Views 88
3.7 Destroying/Altering Tables and Views 91
3.8 Case Study: The Internet Store 92
3.9 Review Questions 94

4 **RELATIONAL ALGEBRA AND CALCULUS** **100**
4.1 Preliminaries 101
4.2 Relational Algebra 102
4.2.1 Selection and Projection 103
4.2.2 Set Operations 104

4.2.3 Renaming 106
4.2.4 Joins 107
4.2.5 Division 109
4.2.6 More Examples of Algebra Queries 110
4.3 Relational Calculus 116
4.3.1 Tuple Relational Calculus 117
4.3.2 Domain Relational Calculus 122
4.4 Expressive Power of Algebra and Calculus 124
4.5 Review Questions 126

5 SQL: QUERIES, CONSTRAINTS, TRIGGERS 130
5.1 Overview 131
5.1.1 Chapter Organization 132
5.2 The Form of a Basic SQL Query 133
5.2.1 Examples of Basic SQL Queries 138
5.2.2 Expressions and Strings in the SELECT Command 139
5.3 UNION, INTERSECT, and EXCEPT 141
5.4 Nested Queries 144
5.4.1 Introduction to Nested Queries 145
5.4.2 Correlated Nested Queries 147
5.4.3 Set-Comparison Operators 148
5.4.4 More Examples of Nested Queries 149
5.5 Aggregate Operators 151
5.5.1 The GROUP BY and HAVING Clauses 154
5.5.2 More Examples of Aggregate Queries 158
5.6 Null Values 162
5.6.1 Comparisons Using Null Values 163
5.6.2 Logical Connectives AND, OR, and NOT 163
5.6.3 Impact on SQL Constructs 163
5.6.4 Outer Joins 164
5.6.5 Disallowing Null Values 165
5.7 Complex Integrity Constraints in SQL 165
5.7.1 Constraints over a Single Table 165
5.7.2 Domain Constraints and Distinct Types 166
5.7.3 Assertions: ICs over Several Tables 167
5.8 Triggers and Active Databases 168
5.8.1 Examples of Triggers in SQL 169
5.9 Designing Active Databases 171
5.9.1 Why Triggers Can Be Hard to Understand 171
5.9.2 Constraints versus Triggers 172
5.9.3 Other Uses of Triggers 172
5.10 Review Questions 173

Part II APPLICATION DEVELOPMENT 183

6 DATABASE APPLICATION DEVELOPMENT 185

6.1 Accessing Databases from Applications 187
6.1.1 Embedded SQL 187
6.1.2 Cursors 189
6.1.3 Dynamic SQL 194
6.2 An Introduction to JDBC 194
6.2.1 Architecture 196
6.3 JDBC Classes and Interfaces 197
6.3.1 JDBC Driver Management 197
6.3.2 Connections 198
6.3.3 Executing SQL Statements 200
6.3.4 ResultSets 201
6.3.5 Exceptions and Warnings 203
6.3.6 Examining Database Metadata 204
6.4 SQLJ 206
6.4.1 Writing SQLJ Code 207
6.5 Stored Procedures 209
6.5.1 Creating a Simple Stored Procedure 209
6.5.2 Calling Stored Procedures 210
6.5.3 SQL/PSM 212
6.6 Case Study: The Internet Book Shop 214
6.7 Review Questions 216

7 INTERNET APPLICATIONS 220

7.1 Introduction 220
7.2 Internet Concepts 221
7.2.1 Uniform Resource Identifiers 221
7.2.2 The Hypertext Transfer Protocol (HTTP) 223
7.3 HTML Documents 226
7.4 XML Documents 227
7.4.1 Introduction to XML 228
7.4.2 XML DTDs 231
7.4.3 Domain-Specific DTDs 234
7.5 The Three-Tier Application Architecture 236
7.5.1 Single-Tier and Client-Server Architectures 236
7.5.2 Three-Tier Architectures 239
7.5.3 Advantages of the Three-Tier Architecture 241
7.6 The Presentation Layer 242
7.6.1 HTML Forms 242
7.6.2 JavaScript 245
7.6.3 Style Sheets 247

7.7 The Middle Tier 251
7.7.1 CGI: The Common Gateway Interface 251
7.7.2 Application Servers 252
7.7.3 Servlets 254
7.7.4 JavaServer Pages 256
7.7.5 Maintaining State 258
7.8 Case Study: The Internet Book Shop 261
7.9 Review Questions 264

Part III STORAGE AND INDEXING 271

8 OVERVIEW OF STORAGE AND INDEXING 273
8.1 Data on External Storage 274
8.2 File Organizations and Indexing 275
8.2.1 Clustered Indexes 277
8.2.2 Primary and Secondary Indexes 277
8.3 Index Data Structures 278
8.3.1 Hash-Based Indexing 279
8.3.2 Tree-Based Indexing 280
8.4 Comparison of File Organizations 282
8.4.1 Cost Model 283
8.4.2 Heap Files 284
8.4.3 Sorted Files 285
8.4.4 Clustered Files 287
8.4.5 Heap File with Unclustered Tree Index 288
8.4.6 Heap File With Unclustered Hash Index 289
8.4.7 Comparison of I/O Costs 290
8.5 Indexes and Performance Tuning 291
8.5.1 Impact of the Workload 292
8.5.2 Clustered Index Organization 292
8.5.3 Composite Search Keys 295
8.5.4 Index Specification in SQL:1999 299
8.6 Review Questions 299

9 STORING DATA: DISKS AND FILES 304
9.1 The Memory Hierarchy 305
9.1.1 Magnetic Disks 306
9.1.2 Performance Implications of Disk Structure 308
9.2 Redundant Arrays of Independent Disks 309
9.2.1 Data Striping 310
9.2.2 Redundancy 311
9.2.3 Levels of Redundancy 312
9.2.4 Choice of RAID Levels 316

9.3 Disk Space Management 316
9.3.1 Keeping Track of Free Blocks 317
9.3.2 Using as File Systems to Manage Disk Space 317
9.4 Buffer Manager 318
9.4.1 Buffer Replacement Policies 320
9.4.2 Buffer Management in DBMS versus OS 322
9.5 Files of Records 324
9.5.1 Implementing Heap Files 324
9.6 Page Formats 326
9.6.1 Fixed-Length Records 327
9.6.2 Variable-Length Records 328
9.7 Record Formats 330
9.7.1 Fixed-Length Records 331
9.7.2 Variable-Length Records 331
9.8 Review Questions 333

10 TREE-STRUCTURED INDEXING 338
10.1 Intuition For Tree Indexes 339
10.2 Indexed Sequential Access Method (ISAM) 341
10.2.1 Overflow Pages, Locking Considerations 344
10.3 B+ Trees: A Dynamic Index Structure 344
10.3.1 Format of a Node 346
10.4 Search 347
10.5 Insert 348
10.6 Delete 352
10.7 Duplicates 356
10.8 B+ Trees in Practice 358
10.8.1 Key Compression 358
10.8.2 Bulk-Loading a B+ Tree 360
10.8.3 The Order Concept 363
10.8.4 The Effect of Inserts and Deletes on Rids 364
10.9 Review Questions 364

11 HASH-BASED INDEXING 370
11.1 Static Hashing 371
11.1.1 Notation and Conventions 373
11.2 Extendible Hashing 373
11.3 Linear Hashing 379
11.4 Extendible vs. Linear Hashing 384
n.5 Review Questions 385

Part IV QUERY EVALUATION 391

12 OVERVIEW OF QUERY EVALUATION 393
12.1 The System Catalog 394
12.1.1 Information in the Catalog 395
12.2 Introduction to Operator Evaluation 397
12.2.1 Three Common Techniques 398
12.2.2 Access Paths 398
12.3 Algorithms for Relational Operations 400
12.3.1 Selection 401
12.3.2 Projection 401
12.3.3 Join 402
12.3.4 Other Operations 404
12.4 Introduction to Query Optimization 404
12.4.1 Query Evaluation Plans 405
12.4.2 Multi-operator Queries: Pipelined Evaluation 407
12.4.3 The Iterator Interface 408
12.5 Alternative Plans: A Motivating Example 409
12.5.1 Pushing Selections 409
12.5.2 Using Indexes 411
12.6 What a Typical Optimizer Does 414
12.6.1 Alternative Plans Considered 414
12.6.2 Estimating the Cost of a Plan 416
12.7 Review Questions 417

13 EXTERNAL SORTING 421
13.1 When Does a DBMS Sort Data? 422
13.2 A Simple Two-Way Merge Sort 423
13.3 External Merge Sort 424
13.3.1 Minimizing the Number of Runs 428
13.4 Minimizing I/O Cost versus Number of I/Os 430
13.4.1 Blocked I/O 430
13.4.2 Double Buffering 432
13.5 Using B+ Trees for Sorting 433
13.5.1 Clustered Index 433
13.5.2 Unclustered Index 434
13.6 Review Questions 436

14 EVALUATING RELATIONAL OPERATORS 439
14.1 The Selection Operation 441
14.1.1 No Index, Unsorted Data 441
14.1.2 No Index, Sorted Data 442
14.1.3 B+ Tree Index 442
14.1.4 Hash Index, Equality Selection 444
14.2 General Selection Conditions 444

14.2.1 CNF and Index Matching 445
14.2.2 Evaluating Selections without Disjunction 445
14.2.3 Selections with Disjunction 446
14.3 The Projection Operation 447
14.3.1 Projection Based on Sorting 448
14.3.2 Projection Based on Hashing 449
14.3.3 Sorting Versus Hashing for Projections 451
14.3.4 Use of Indexes for Projections 452
14.4 The Join Operation 452
14.4.1 Nested Loops Join 454
14.4.2 Sort-Merge Join 458
14.4.3 Hash Join 463
14.4.4 General Join Conditions 467
14.5 The Set Operations 468
14.5.1 Sorting for Union and Difference 469
14.5.2 Hashing for Union and Difference 469
14.6 Aggregate Operations 469
14.6.1 Implementing Aggregation by Using an Index 471
14.7 The Impact of Buffering 471
14.8 Review Questions 472

15 A TYPICAL RELATIONAL QUERY OPTIMIZER 478
15.1 Translating SQL Queries into Algebra 479
15.1.1 Decomposition of a Query into Blocks 479
15.1.2 A Query Block as a Relational Algebra Expression 481
15.2 Estimating the Cost of a Plan 482
15.2.1 Estimating Result Sizes 483
15.3 Relational Algebra Equivalences 488
15.3.1 Selections 488
15.3.2 Projections 488
15.3.3 Cross-Products and Joins 489
15.3.4 Selects, Projects, and Joins 490
15.3.5 Other Equivalences 491
15.4 Enumeration of Alternative Plans 492
15.4.1 Single-Relation Queries 492
15.4.2 Multiple-Relation Queries 496
IS.5 Nested Subqueries 504
15.6 The System R Optimizer 506
15.7 Other Approaches to Query Optimization S07
15.8 Review Questions 507

Part V TRANSACTION MANAGEMENT 517

16 OVERVIEW OF TRANSACTION MANAGEMENT 519
16.1 The ACID Properties 520
16.1.1 Consistency and Isolation 521
16.1.2 Atomicity and Durability 522
16.2 Transactions and Schedules 523
16.3 Concurrent Execution of Transactions 524
16.3.1 rvlotivation for Concurrent Execution 524
16.3.2 Serializability 525
16.3.3 Anomalies Due to Interleaved Execution 526
16.3.4 Schedules Involving Aborted Transactions 529
16.4 Lock-Based Concurrency Control 530
16.4.1 Strict Two-Phase Locking (Strict 2PL) 531
16.4.2 Deadlocks 533
16.5 Performance of Locking 533
16.6 Transaction Support in SQL 535
16.6.1 Creating and Terminating Transactions 535
16.6.2 What Should We Lock? 537
16.6.3 Transaction Characteristics in SQL 538
16.7 Introduction to Crash Recovery 540
16.7.1 Stealing Frames and Forcing Pages 541
16.7.2 Recovery-Related Steps during Normal Execution 542
16.7.3 Overview of ARIES 543
16.7.4 Atomicity: Implementing Rollback 543
16.8 Review Questions 544

17 CONCURRENCY CONTROL 549
17.1 2PL, Serializability, and Recoverability 550
17.1.1 View Serializability 553
17.2 Introduction to Lock Management 553
17.2.1 Implementing Lock and Unlock Requests 554
17.3 Lock Conversions 555
17.4 Dealing With Deadlocks 556
17.4.1 Deadlock Prevention 558
17.5 Specialized Locking Techniques 559
17.5.1 Dynamic Databases and the Phantom Problem 560
17.5.2 Concurrency Control in B+ Trees 561
17.5.3 Multiple-Granularity Locking 564
17.6 ConClurency Control without Locking 566
17.6.1 Optimistic Concurrency Control 566
17.6.2 Timestamp-Based Concurrency Control 569
17.6.3 Multiversion Concurrency Control 572
17.7 Review Questions 573

18 CRASH RECOVERY **579**
18.1 Introduction to ARIES 580
18.2 The Log 582
18.3 Other Recovery-Related Structures 585
18.4 The Write-Ahead Log Protocol 586
18.5 Checkpointing 587
18.6 Recovering from a System Crash 587
18.6.1 Analysis Phase 588
18.6.2 Redo Phase 590
18.6.3 Undo Phase 592
18.7 Media Recovery 595
18.8 Other Approaches and Interaction with Concurrency Control 596
18.9 Review Questions 597

Part VI DATABASE DESIGN AND TUNING **603**

19 SCHEMA REFINEMENT AND NORMAL FORMS **605**
19.1 Introduction to Schema Refinement 606
19.1.1 Problems Caused by Redundancy 606
19.1.2 Decompositions 608
19.1.3 Problems Related to Decomposition 609
19.2 Functional Dependencies 611
19.3 Reasoning about FDs 612
19.3.1 Closure of a Set of FDs 612
19.3.2 Attribute Closure 614
19.4 Normal Forms 615
19.4.1 Boyce-Codd Normal Form 615
19.4.2 Third Normal Form 617
19.5 Properties of Decompositions 619
19.5.1 Lossless-Join Decomposition 619
19.5.2 Dependency-Preserving Decomposition 621
19.6 Normalization 622
19.6.1 Decomposition into BCNF 622
19.6.2 Decomposition into 3NF 625
19.7 Schema Refinement in Database Design 629
19.7.1 Constraints on an Entity Set 630
19.7.2 Constraints on a Relationship Set 630
19.7.3 Identifying Attributes of Entities 631
19.7.4 Identifying Entity Sets 633
19.8 Other Kinds of Dependencies 633
19.8.1 Multivalued Dependencies 634
19.8.2 Fourth Normal Form 636
19.8.3 Join Dependencies 638

19.8.4 Fifth Normal Form 638
19.8.5 Inclusion Dependencies 639
19.9 Case Study: The Internet Shop 640
19.10 Review Questions 642

20 PHYSICAL DATABASE DESIGN AND TUNING 649
20.1 Introduction to Physical Database Design 650
20.1.1 Database Workloads 651
20.1.2 Physical Design and Tuning Decisions 652
20.1.3 Need for Database Tuning 653
20.2 Guidelines for Index Selection 653
20.3 Basic Examples of Index Selection 656
20.4 Clustering and Indexing 658
20.4.1 Co-clustering Two Relations 660
20.5 Indexes that Enable Index-Only Plans 662
20.6 Tools to Assist in Index Selection 663
20.6.1 Automatic Index Selection 663
20.6.2 How Do Index Tuning Wizards Work? 664
20.7 Overview of Database Tuning 667
20.7.1 Tuning Indexes 667
20.7.2 Tuning the Conceptual Schema 669
20.7.3 Tuning Queries and Views 670
20.8 Choices in Tuning the Conceptual Schema 671
20.8.1 Settling for a Weaker Normal Form 671
20.8.2 Denormalization 672
20.8.3 Choice of Decomposition 672
20.8.4 Vertical Partitioning of BCNF Relations 674
20.8.5 Horizontal Decomposition 674
20.9 Choices in Tuning Queries and Views 675
20.10 Impact of Concurrency 678
20.10.1 Reducing Lock Durations 678
20.10.2 Reducing Hot Spots 679
20.11 Case Study: The Internet Shop 680
20.11.11\ming the Database 682
20.12 DBMS Benchmarking 682
20.12.1 Well-Known DBMS Benchmarks 683
20.12.2 Using a Benchmark 684
20.13 Review Questions 685

21 SECURITY AND AUTHORIZATION 692
21.1 Introduction to Database Security 693
21.2 Access Control 694
21.3 Discretionary Access Control 695

21.3.1 Grant and Revoke on Views and Integrity Constraints 704
21.4 Mandatory Access Control 705
21.4.1 Multilevel Relations and Polyinstantiation 707
21.4.2 Covert Channels, DoD Security Levels 708
21.5 Security for Internet Applications 709
21.5.1 Encryption 709
21.5.2 Certifying Servers: The SSL Protocol 712
21.5.3 Digital Signatures 713
21.6 Additional Issues Related to Security 714
21.6.1 Role of the Database Administrator 714
21.6.2 Security in Statistical Databases 715
21.7 Design Case Study: The Internet Store 716
21.8 Review Questions 718

Part VII ADDITIONAL TOPICS 723

22 PARALLEL AND DISTRIBUTED DATABASES 725
22.1 Introduction 726
22.2 Architectures for Parallel Databases 727
22.3 Parallel Query Evaluation 728
22.3.1 Data Partitioning 730
22.3.2 Parallelizing Sequential Operator Evaluation Code 730
22.4 Parallelizing Individual Operations 731
22.4.1 Bulk Loading and Scanning 731
22.4.2 Sorting 732
22.4.3 Joins 732
22.5 Parallel Query Optimization 735
22.6 Introduction to Distributed Databases 736
22.6.1 Types of Distributed Databases 737
22.7 Distributed DBMS Architectures 737
22.7.1 Client-Server Systems 738
22.7.2 Collaborating Server Systems 738
22.7.3 Midclleware Systems 739
22.8 Storing Data in a Distributed DBMS 739
22.8.1 Fragmentation 739
22.8.2 Replication 741
22.9 Distributed Catalog Management 741
22.9.1 Naming Objects 741
22.9.2 Catalog Structure 742
22.9.3 Distributed Data Independence 743
22.10 Distributed Query Processing 743
22.1.0.1 Nonjoin Queries in a Distributed DBMS 744
22.10.2 Joins in a Distributed DBMS 745

22.10.3 Cost-Based Query Optimization 749
22.11 Updating Distributed Data 750
22.11.1 Synchronous Replication 750
22.11.2 Asynchronous Replication 751
22.12 Distributed Transactions 755
22.13 Distributed Concurrency Control 755
22.13.1 Distributed Deadlock 756
22.14 Distributed Recovery 758
22.14.1 Normal Execution and Commit Protocols 758
22.14.2 Restart after a Failure 760
22.14.3 Two-Phase Commit Revisited 761
22.14.4 Three-Phase Commit 762
22.15 Review Questions 763

23 OBJECT-DATABASE SYSTEMS 772
23.1 Motivating Example 774
23.1.1 New Data Types 775
23.1.2 Manipulating the New Data 777
23.2 Structured Data Types 779
23.2.1 Collection Types 780
23.3 Operations on Structured Data 781
23.3.1 Operations on Rows 781
23.3.2 Operations on Arrays 781
23.3.3 Operations on Other Collection Types 782
23.3.4 Queries Over Nested Collections 783
23.4 Encapsulation and ADTs 784
23.4.1 Defining Methods 785
23.5 Inheritance 787
23.5.1 Defining Types with Inheritance 787
23.5.2 Binding Methods 788
23.5.3 Collection Hierarchies 789
23.6 Objects, aIDs, and Reference Types 789
23.6.1 Notions of Equality 790
23.6.2 Dereferencing Reference Types 791
23.6.3 URLs and OIDs in SQL:1999 791
23.7 Database Design for an ORDBMS 792
23.7.1 Collection Types and ADTs 792
23.7.2 Object Identity 795
23.7.3 Extending the ER Model 796
23.7.4 Using Nested Collections 798
23.8 ORDBMS Implementation Challenges 799
23.8.1 Storage and Access Methods 799
23.8.2 Query Processing 801

23.8.3 Query Optimization 80;3
23.9 OODBMS 805
23.9.1 The ODMG Data Model and ODL 805
23.9.2 OQL 807
23.10 Comparing RDBMS, OODBMS, and ORDBMS 809
23.10.1 RDBMS versus ORDBMS 809
23.10.2 OODBMS versus ORDBMS: Similarities 809
23.10.3 OODBMS versus ORDBMS: Differences 810
23.11 Review Questions 811

24 DEDUCTIVE DATABASES 817
24.1 Introduction to Recursive Queries 818
24.1.1 Datalog 819
24.2 Theoretical Foundations 822
24.2.1 Least Model Semantics 823
24.2.2 The Fixpoint Operator 824
24.2.3 Safe Datalog Programs 825
24.2.4 Least Model = Least Fixpoint 826
24.3 Recursive Queries with Negation 827
24.3.1 Stratification 828
24.4 From Datalog to SQL 831
24.5 Evaluating Recursive Queries 834
24.5.1 Fixpoint Evaluation without Repeated Inferences 835
24.5.2 Pushing Selections to Avoid Irrelevant Inferences 837
24.5.3 The Magic Sets Algorithm 838
24.6 Review Questions 841

25 DATA WAREHOUSING AND DECISION SUPPORT 846
25.1 Introduction to Decision Support 848
25.2 OLAP: Multidimensional Data Model 849
25.2.1 Multidimensional Database Design 853
25.3 Multidimensional Aggregation Queries 854
25.3.1 ROLLUP and CUBE in SQL:1999 856
25.4 Window Queries in SQL:1999 859
25.4.1 Framing a Window 861
25.4.2 New Aggregate Functions 862
25.5 Findipg Answers Quickly 862
25.5.1 Top N Queries 863
25.5.2 Online Aggregation 864
25.6 Implementation Techniques for OLAP 865
25.6.1 Bitmap Indexes 866
25.6.2 Join Indexes 868
25.6.3 File Organizations 869

25.7 Data 'Warehousing 870
25.7.1 Creating and Ivlaintaining a Warehouse 871
25.8 Views and Decision Support 872
25.8.1 Views, OLAP, and \Varehousing 872
25.8.2 Queries over Views 873
25.9 View Materialization 873
25.9.1 Issues in View Materialization 874
25.10 Maintaining Materialized Views 876
2.5.10.1 Incremental View Maintenance 876
25.10.2 Maintaining Warehouse Views 879
25.10.3 When Should We Synchronize Views? 881
25.11 Review Questions 882

26 DATA MINING 889
26.1 Introduction to Data Mining 890
26.1.1 The Knowledge Discovery Process 891
26.2 Counting Co-occurrences 892
26.2.1 Frequent Itemsets 892
26.2.2 Iceberg Queries 895
26.3 Mining for Rules 897
26.3.1 Association Rules 897
26.3.2 An Algorithm for Finding Association Rules 898
26.3.3 Association Rules and ISA Hierarchies 899
26.3.4 Generalized Association Rules 900
26.3.5 Sequential Patterns 901
26.3.6 The Use of Association Rules for Prediction 902
26.3.7 Bayesian Networks 903
26.3.8 Classification and Regression Rules 904
26.4 Tree-Structured Rules 906
26.4.1 Decision Trees 907
26.4.2 An Algorithm to Build Decision Trees 908
26.5 Clustering 911
26.5.1 A Clustering Algorithm 912
26.6 Similarity Search over Sequences 913
26.6.1 An Algorithm to Find Similar Sequences 915
26.7 Incremental Mining and Data Streams 916
26.7.1 Incremental Maintenance of Frequent Itemsets 918
26.8 Additional Data Mining Tasks 920
26.9 Review Questions 920

27 INFORMATION RETRIEVAL AND XML DATA 926
27.1 Colliding Worlds: Databases, IR, and XML 927
27.1.1 DBMS versus IR Systems 928

27.2 Introduction to Information Retrieval 929
27.2.1 Vector Space Model 930
27.2.2 TF*jIDF* Weighting of Terms 931
27.2.3 Ranking Document Similarity 932
27.2.4 :Measuring Success: Precision and Recall 934
27.3 Indexing for Text Search 934
27.3.1 Inverted Indexes 935
27.3.2 Signature Files 937
27.4 Web Search Engines 939
27.4.1 Search Engine Architecture 939
27.4.2 Using Link Information 940
27.5 Managing Text in a DBMS 944
27.5.1 Loosely Coupled Inverted Index 945
27.6 A Data Model for XML 945
27.6.1 Motivation for Loose Structure 945
27.6.2 A Graph Model 946
27.7 XQuery: Querying XML Data 948
27.7.1 Path Expressions 948
27.7.2 FLWR Expressions 949
27.7.3 Ordering of Elements 951
27.7.4 Grouping and Generation of Collection Values 951
27.8 Efficient Evaluation of XML Queries 952
27.8.1 Storing XML in RDBMS 952
27.8.2 Indexing XML Repositories 956
27.9 Review Questions 959

28 SPATIAL DATA MANAGEMENT 968
28.1 Types of Spatial Data and Queries 969
28.2 Applications Involving Spatial Data 971
28.3 Introduction to Spatial Indexes 973
28.3.1 Overview of Proposed Index Structures 974
28.4 Indexing Based on Space-Filling Curves 975
28.4.1 Region Quad Trees and Z-Ordering: Region Data 976
28.4.2 Spatial Queries Using Z-Ordering 978
28.5 Grid Files 978
28.5.1 Adapting Grid Files to Handle Regions 981
28.6 R Trees: Point and Region Data 982
28.6.1 Queries 983
28.6.2 Insert and Delete Operations 984
28.6.3 Concurrency Control 986
28.6.4 Generalized Search Trees 987
28.7 Issues in High-Dimensional Indexing 988
28.8 Review Questions 988

29 FURTHER READING **992**
29.1 Advanced Tl"ansaction Processing 993
29.1.1 Transaction Processing Monitors 993
29.1.2 New Transaction Models 994
29.1.3 Real-Time DBMSs 994
29.2 Data Integration 995
29.3 Mobile Databases 995
29.4 Main Memory Databases 996
29.5 Multimedia Databases 997
29.6 Geographic Information Systems 998
29.7 Temporal Databases 999
29.8 Biological Databases 999
29.9 Information Visualization 1000
29.10 Summary 1000

30 THE MINIBASE SOFTWARE **1002**
30.1 What Is Available 1002
30.2 Overview of Minibase Assignments 1003
30.3 Acknowledgments 1004

REFERENCES **1005**

AUTHOR INDEX **1045**

SUBJECT INDEX **1054**