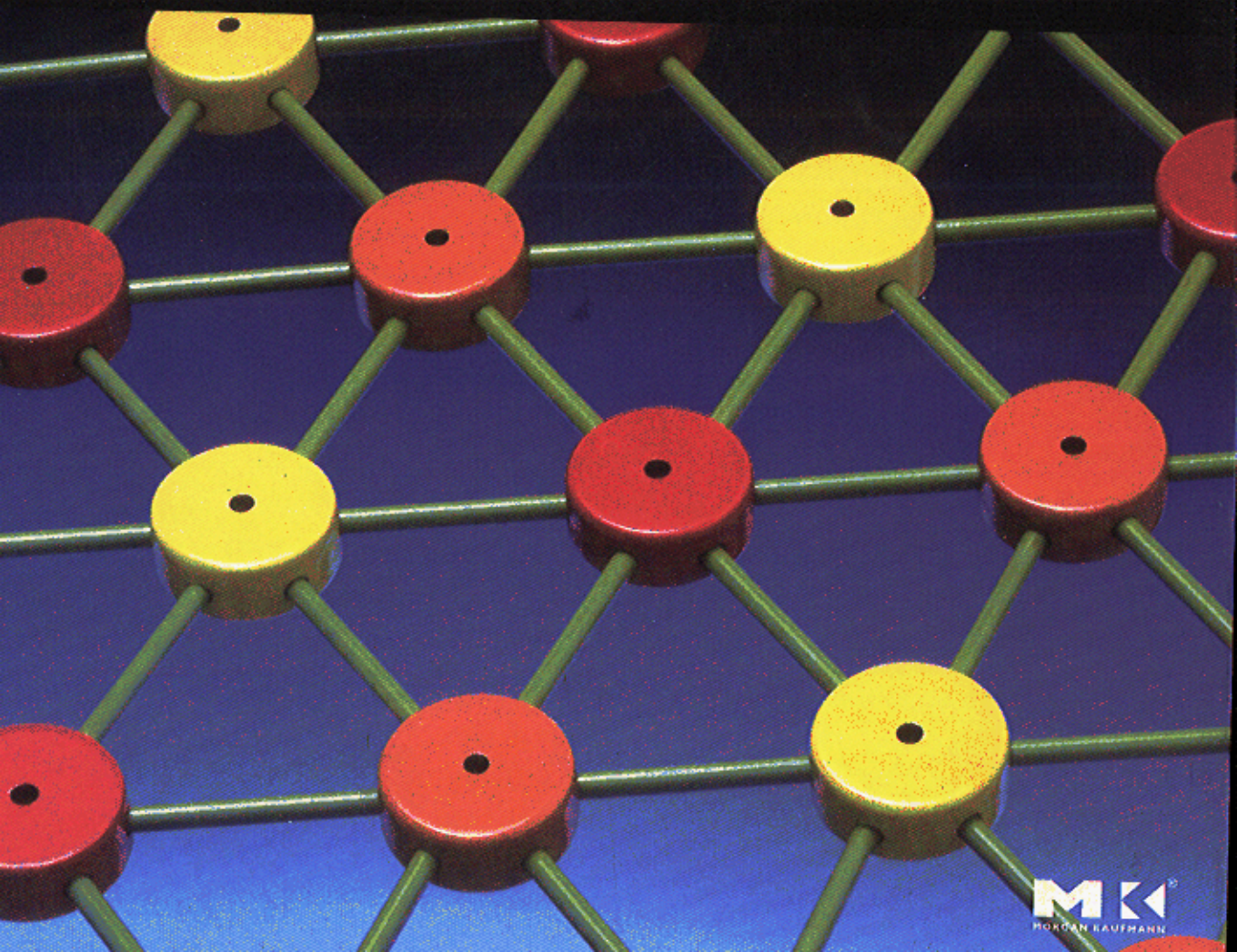




DATA MODELING ESSENTIALS

Third
Edition

Graeme C. **SIMSION** • Graham C. **WITT**



Contents

Preface xxiii

■ ■ ■ Part I

The Basics 1

■ ■ ■ Chapter 1

What Is Data Modeling? 3

1.1 Introduction 3

1.2 A Data-Centered Perspective 3

1.3 A Simple Example 4

1.4 Design, Choice, and Creativity 6

1.5 Why Is the Data Model Important? 8

1.5.1 Leverage 8

1.5.2 Conciseness 9

1.5.3 Data Quality 10

1.5.4 Summary 10

1.6 What Makes a Good Data Model? 10

1.6.1 Completeness 10

1.6.2 NonRedundancy 11

1.6.3 Enforcement of Business Rules 11

1.6.4 Data Reusability 11

1.6.5 Stability and Flexibility 12

1.6.6 Elegance 13

1.6.7 Communication 14

1.6.8 Integration 14

1.6.9 Conflicting Objectives 15

1.7 Performance 15

1.8 Database Design Stages and Deliverables 16

1.8.1 Conceptual, Logical, and Physical Data Models 16

1.8.2 The Three-Schema Architecture and Terminology 17

- 1.9 **Where Do Data Models Fit In? 20**
 - 1.9.1 Process-Driven Approaches 20
 - 1.9.2 Data-Driven Approaches 20
 - 1.9.3 Parallel (Blended) Approaches 22
 - 1.9.4 Object-Oriented Approaches 22
 - 1.9.5 Prototyping Approaches 23
 - 1.9.6 Agile Methods 23
- 1.10 **Who Should Be Involved in Data Modeling? 23**
- 1.11 **Is Data Modeling Still Relevant? 24**
 - 1.11.1 Costs and Benefits of Data Modeling 25
 - 1.11.2 Data Modeling and Packaged Software 26
 - 1.11.3 Data Integration 27
 - 1.11.4 Data Warehouses 27
 - 1.11.5 Personal Computing and User-Developed Systems 28
 - 1.11.6 Data Modeling and XML 28
 - 1.11.7 Summary 28
- 1.12 **Alternative Approaches to Data Modeling 29**
- 1.13 **Terminology 30**
- 1.14 **Where to from Here?—An Overview of Part I 31**
- 1.15 **Summary 32**
- ■ ■ **Chapter 2**
 - Basics of Sound Structure 33**
 - 2.1 **Introduction 33**
 - 2.2 **An Informal Example of Normalization 34**
 - 2.3 **Relational Notation 36**
 - 2.4 **A More Complex Example 37**
 - 2.5 **Determining Columns 40**
 - 2.5.1 One Fact per Column 40
 - 2.5.2 Hidden Data 41
 - 2.5.3 Derivable Data 41
 - 2.5.4 Determining the Primary Key 41
 - 2.6 **Repeating Groups and First Normal Form 43**
 - 2.6.1 Limit on Maximum Number of Occurrences 43
 - 2.6.2 Data Reusability and Program Complexity 43
 - 2.6.3 Recognizing Repeating Groups 44
 - 2.6.4 Removing Repeating Groups 45

- 2.6.5 Determining the Primary Key of the New Table 46
- 2.6.6 First Normal Form 47

2.7 Second and Third Normal Forms 47

- 2.7.1 Problems with Tables in First Normal Form 47
- 2.7.2 Eliminating Redundancy 48
- 2.7.3 Determinants 48
- 2.7.4 Third Normal Form 51

2.8 Definitions and a Few Refinements 53

- 2.8.1 Determinants and Functional Dependency 53
- 2.8.2 Primary Keys 54
- 2.8.3 Candidate Keys 54
- 2.8.4 A More Formal Definition of Third Normal Form 55
- 2.8.5 Foreign Keys 55
- 2.8.6 Referential Integrity 56
- 2.8.7 Update Anomalies 57
- 2.8.8 Denormalization and Unnormalization 58
- 2.8.9 Column and Table Names 59

2.9 Choice, Creativity, and Normalization 60

2.10 Terminology 62

2.11 Summary 63

■ ■ ■ Chapter 3

The Entity-Relationship Approach 65

3.1 Introduction 65

3.2 A Diagrammatic Representation 65

- 3.2.1 The Basic Symbols: Boxes and Arrows 66
- 3.2.2 Diagrammatic Representation of Foreign Keys 67
- 3.2.3 Interpreting the Diagram 68
- 3.2.4 Optionality 69
- 3.2.5 Verifying the Model 70
- 3.2.6 Redundant Arrows 71

3.3 The Top-Down Approach: Entity-Relationship Modeling 72

- 3.3.1 Developing the Diagram Top Down 74
- 3.3.2 Terminology 75

3.4 Entity Classes 76

- 3.4.1 Entity Diagramming Convention 77
- 3.4.2 Entity Class Naming 78
- 3.4.3 Entity Class Definitions 80

3.5 Relationships 82

- 3.5.1 Relationship Diagramming Conventions 82
- 3.5.2 Many-to-Many Relationships 87
- 3.5.3 One-to-One Relationships 92
- 3.5.4 Self-Referencing Relationships 93
- 3.5.5 Relationships Involving Three or More Entity Classes 96
- 3.5.6 Transferability 98
- 3.5.7 Dependent and Independent Entity Classes 102
- 3.5.8 Relationship Names 103

3.6 Attributes 104

- 3.6.1 Attribute Identification and Definition 104
- 3.6.2 Primary Keys and the Conceptual Model 105

3.7 Myths and Folklore 105

- 3.7.1 Entity Classes without Relationships 106
- 3.7.2 Allowed Combinations of Cardinality and Optionality 106

3.8 Creativity and E-R Modeling 106

3.9 Summary 109

■ ■ ■ **Chapter 4**

Subtypes and Supertypes 111

4.1 Introduction 111

4.2 Different Levels of Generalization 111

4.3 Rules versus Stability 113

4.4 Using Subtypes and Supertypes 115

4.5 Subtypes and Supertypes as Entity Classes 116

- 4.5.1 Naming Subtypes 117

4.6 Diagramming Conventions 117

- 4.6.1 Boxes in Boxes 117
- 4.6.2 UML Conventions 118
- 4.6.3 Using Tools That Do Not Support Subtyping 119

4.7 Definitions 119

4.8 Attributes of Supertypes and Subtypes 119

4.9 Nonoverlapping and Exhaustive 120

4.10 Overlapping Subtypes and Roles 123

- 4.10.1 Ignoring Real-World Overlaps 123
- 4.10.2 Modeling Only the Supertype 124
- 4.10.3 Modeling the Roles as Participation in Relationships 124
- 4.10.4 Using Role Entity Classes and One-to-One Relationships 125
- 4.10.5 Multiple Partitions 126

4.11 Hierarchy of Subtypes 127**4.12 Benefits of Using Subtypes and Supertypes 128**

- 4.12.1 Creativity 129
- 4.12.2 Presentation: Level of Detail 129
- 4.12.3 Communication 130
- 4.12.4 Input to the Design of Views 132
- 4.12.5 Classifying Common Patterns 132
- 4.12.6 Divide and Conquer 133

4.13 When Do We Stop Supertyping and Subtyping? 134

- 4.13.1 Differences in Identifiers 134
- 4.13.2 Different Attribute Groups 135
- 4.13.3 Different Relationships 135
- 4.13.4 Different Processes 136
- 4.13.5 Migration from One Subtype to Another 136
- 4.13.6 Communication 136
- 4.13.7 Capturing Meaning and Rules 137
- 4.13.8 Summary 137

4.14 Generalization of Relationships 138

- 4.14.1 Generalizing Several One-to-Many Relationships to a Single Many-to-Many Relationship 138
- 4.14.2 Generalizing Several One-to-Many Relationships to a Single One-to-Many Relationship 139
- 4.14.3 Generalizing One-to-Many and Many-to-Many Relationships 141

4.15 Theoretical Background 142**4.16 Summary 143****Chapter 5****Attributes and Columns 145****5.1 Introduction 145****5.2 Attribute Definition 146**

5.3	Attribute Disaggregation: One Fact per Attribute	147
5.3.1	Simple Aggregation	148
5.3.2	Conflated Codes	150
5.3.3	Meaningful Ranges	151
5.3.4	Inappropriate Generalization	151
5.4	Types of Attributes	152
5.4.1	DBMS Datatypes	152
5.4.2	The Attribute Taxonomy in Detail	154
5.4.3	Attribute Domains	158
5.4.4	Column Datatype and Length Requirements	162
5.4.5	Conversion Between External and Internal Representations	166
5.5	Attribute Names	166
5.5.1	Objectives of Standardizing Attribute Names	166
5.5.2	Some Guidelines for Attribute Naming	168
5.6	Attribute Generalization	171
5.6.1	Options and Trade-Offs	171
5.6.2	Attribute Generalization Resulting from Entity Generalization	172
5.6.3	Attribute Generalization within Entity Classes	173
5.6.4	“First Among Equals”	177
5.6.5	Limits to Attribute Generalization	178
5.7	Summary	180
■ ■ ■	Chapter 6	
	Primary Keys and Identity	183
6.1	Basic Requirements and Trade-Offs	183
6.2	Basic Technical Criteria	185
6.2.1	Applicability	185
6.2.2	Uniqueness	186
6.2.3	Minimality	188
6.2.4	Stability	189
6.3	Surrogate Keys	191
6.3.1	Performance and Programming Issues	191
6.3.2	Matching Real-World Identifiers	191
6.3.3	Should Surrogate Keys Be Visible?	192
6.3.4	Subtypes and Surrogate Keys	193
6.4	Structured Keys	194
6.4.1	When to Use Structured Keys	196
6.4.2	Programming and Structured Keys	197
6.4.3	Performance Issues with Structured Keys	198
6.4.4	Running Out of Numbers	199

6.5 Multiple Candidate Keys 201

6.5.1 Choosing a Primary Key 201

6.5.2 Normalization Issues 201

6.6 Guidelines for Choosing Keys 202

6.6.1 Tables Implementing Independent Entity Classes 202

6.6.2 Tables Implementing Dependent Entity Classes and Many-to-Many Relationships 203

6.7 Partially-Null Keys 204**6.8 Summary 206**■ ■ ■ **Chapter 7****Extensions and Alternatives 207****7.1 Introduction 207****7.2 Extensions to the Basic E-R Approach 209**

7.2.1 Introduction 209

7.2.2 Advanced Attribute Concepts 210

7.3 The Chen E-R Approach 216

7.3.1 The Basic Conventions 216

7.3.2 Relationships with Attributes 217

7.3.3 Relationships Involving Three or More Entity Classes 217

7.3.4 Roles 218

7.3.5 The Weak Entity Concept 219

7.3.6 Chen Conventions in Practice 220

7.4 Using UML Object Class Diagrams 220

7.4.1 A Conceptual Data Model in UML 221

7.4.2 Advantages of UML 222

7.5 Object Role Modeling 227**7.6 Summary 228**■ ■ ■ **Part II****Putting It Together 229**■ ■ ■ **Chapter 8****Organizing the Data Modeling Task 231****8.1 Data Modeling in the Real World 231****8.2 Key Issues in Project Organization 233**

8.2.1 Recognition of Data Modeling 233

8.2.2 Clear Use of the Data Model 234

8.2.3	Access to Users and Other Business Stakeholders	234
8.2.4	Conceptual, Logical, and Physical Models	235
8.2.5	Cross-Checking with the Process Model	236
8.2.6	Appropriate Tools	237
8.3	Roles and Responsibilities	238
8.4	Partitioning Large Projects	240
8.5	Maintaining the Model	242
8.5.1	Examples of Complex Changes	242
8.5.2	Managing Change in the Modeling Process	247
8.6	Packaging It Up	248
8.7	Summary	249
■ ■ ■ Chapter 9		
	The Business Requirements	251
9.1	Purpose of the Requirements Phase	251
9.2	The Business Case	253
9.3	Interviews and Workshops	254
9.3.1	Should You Model in Interviews and Workshops?	255
9.3.2	Interviews with Senior Managers	256
9.3.3	Interviews with Subject Matter Experts	257
9.3.4	Facilitated Workshops	257
9.4	Riding the Trucks	258
9.5	Existing Systems and Reverse Engineering	259
9.6	Process Models	261
9.7	Object Class Hierarchies	261
9.7.1	Classifying Object Classes	263
9.7.2	A Typical Set of Top-Level Object Classes	265
9.7.3	Developing an Object Class Hierarchy	267
9.7.4	Potential Issues	270
9.7.5	Advantages of the Object Class Hierarchy Technique	270
9.8	Summary	270

■■■ Chapter 10.

Conceptual Data Modeling 273

- 10.1 **Designing Real Models 273**
- 10.2 **Learning from Designers in Other Disciplines 275**
- 10.3 **Starting the Modeling 276**
- 10.4 **Patterns and Generic Models 277**
 - 10.4.1 Using Patterns 277
 - 10.4.2 Using a Generic Model 278
 - 10.4.3 Adapting Generic Models from Other Applications 279
 - 10.4.4 Developing a Generic Model 282
 - 10.4.5 When There Is Not a Generic Model 284
- 10.5 **Bottom-Up Modeling 285**
- 10.6 **Top-Down Modeling 288**
- 10.7 **When the Problem Is Too Complex 288**
- 10.8 **Hierarchies, Networks, and Chains 290**
 - 10.8.1 Hierarchies 291
 - 10.8.2 Networks (Many-to-Many Relationships) 293
 - 10.8.3 Chains (One-to-One Relationships) 295
- 10.9 **One-to-One Relationships 295**
 - 10.9.1 Distinct Real-World Concepts 296
 - 10.9.2 Separating Attribute Groups 297
 - 10.9.3 Transferable One-to-One Relationships 298
 - 10.9.4 Self-Referencing One-to-One Relationships 299
 - 10.9.5 Support for Creativity 299
- 10.10 **Developing Entity Class Definitions 300**
- 10.11 **Handling Exceptions 301**
- 10.12 **The Right Attitude 302**
 - 10.12.1 Being Aware 303
 - 10.12.2 Being Creative 303
 - 10.12.3 Analyzing or Designing 303
 - 10.12.4 Being Brave 304
 - 10.12.5 Being Understanding and Understood 304
- 10.13 **Evaluating the Model 305**
- 10.14 **Direct Review of Data Model Diagrams 306**

- 10.15 **Comparison with the Process Model** 308
- 10.16 **Testing the Model with Sample Data** 308
- 10.17 **Prototypes** 309
- 10.18 **The Assertions Approach** 309
 - 10.18.1 Naming Conventions 310
 - 10.18.2 Rules for Generating Assertions 311
- 10.19 **Summary** 319
- **Chapter 11**
- Logical Database Design** 321
- 11.1 **Introduction** 321
- 11.2 **Overview of the Transformations Required** 322
- 11.3 **Table Specification** 325
 - 11.3.1 The Standard Transformation 325
 - 11.3.2 Exclusion of Entity Classes from the Database 325
 - 11.3.3 Classification Entity Classes 325
 - 11.3.4 Many-to-Many Relationship Implementation 326
 - 11.3.5 Relationships Involving More Than Two Entity Classes 328
 - 11.3.6 Supertype/Subtype Implementation 328
- 11.4 **Basic Column Definition** 334
 - 11.4.1 Attribute Implementation: The Standard Transformation 334
 - 11.4.2 Category Attribute Implementation 335
 - 11.4.3 Derivable Attributes 336
 - 11.4.4 Attributes of Relationships 336
 - 11.4.5 Complex Attributes 337
 - 11.4.6 Multivalued Attribute Implementation 337
 - 11.4.7 Additional Columns 339
 - 11.4.8 Column Datatypes 340
 - 11.4.9 Column Nullability 340
- 11.5 **Primary Key Specification** 341
- 11.6 **Foreign Key Specification** 342
 - 11.6.1 One-to-Many Relationship Implementation 343
 - 11.6.2 One-to-One Relationship Implementation 346
 - 11.6.3 Derivable Relationships 347
 - 11.6.4 Optional Relationships 348

- 11.6.5 Overlapping Foreign Keys 350
- 11.6.6 Split Foreign Keys 352
- 11.7 Table and Column Names 354**
- 11.8 Logical Data Model Notations 355**
- 11.9 Summary 357**
- ■ ■ **Chapter 12**
- Physical Database Design 359**
- 12.1 Introduction 359**
- 12.2 Inputs to Database Design 361**
- 12.3 Options Available to the Database Designer 362**
- 12.4 Design Decisions Which Do Not Affect Program Logic 363**
 - 12.4.1 Indexes 363
 - 12.4.2 Data Storage 370
 - 12.4.3 Memory Usage 372
- 12.5 Crafting Queries to Run Faster 372**
 - 12.5.1 Locking 373
- 12.6 Logical Schema Decisions 374**
 - 12.6.1 Alternative Implementation of Relationships 374
 - 12.6.2 Table Splitting 374
 - 12.6.3 Table Merging 376
 - 12.6.4 Duplication 377
 - 12.6.5 Denormalization 378
 - 12.6.6 Ranges 379
 - 12.6.7 Hierarchies 380
 - 12.6.8 Integer Storage of Dates and Times 382
 - 12.6.9 Additional Tables 383
- 12.7 Views 384**
 - 12.7.1 Views of Supertypes and Subtypes 385
 - 12.7.2 Inclusion of Derived Attributes in Views 385
 - 12.7.3 Denormalization and Views 385
 - 12.7.4 Views of Split and Merged Tables 386
- 12.8 Summary 386**

■ ■ ■ Part III

Advanced Topics 389

■ ■ ■ Chapter 13

Advanced Normalization 391

13.1 **Introduction 391**

13.2 **Introduction to the Higher Normal Forms 392**

13.2.1 Common Misconceptions 392

13.3 **Boyce-Codd Normal Form 394**

13.3.1 Example of Structure in 3NF but not in BCNF 394

13.3.2 Definition of BCNF 396

13.3.3 Enforcement of Rules versus BCNF 397

13.3.4 A Note on Domain Key Normal Form 398

13.4 **Fourth Normal Form (4NF) and Fifth Normal Form (5NF) 398**

13.4.1 Data in BCNF but not in 4NF 399

13.4.2 Fifth Normal Form (5NF) 401

13.4.3 Recognizing 4NF and 5NF Situations 404

13.4.4 Checking for 4NF and 5NF with the Business Specialist 405

13.5 **Beyond 5NF: Splitting Tables Based on Candidate Keys 407**

13.6 **Other Normalization Issues 408**

13.6.1 Normalization and Redundancy 408

13.6.2 Reference Tables Produced by Normalization 410

13.6.3 Selecting the Primary Key after Removing Repeating Groups 411

13.6.4 Sequence of Normalization and Cross-Table Anomalies 414

13.7 **Advanced Normalization in Perspective 415**

13.8 **Summary 416**

■ ■ ■ Chapter 14

Modeling Business Rules 417

14.1 **Introduction 417**

14.2 **Types of Business Rules 418**

14.2.1 Data Rules 418

14.2.2 Process Rules 420

14.2.3 What Rules are Relevant to the Data Modeler? 420

14.3 Discovery and Verification of Business Rules 420

14.3.1 Cardinality Rules 420

14.3.2 Other Data Validation Rules 421

14.3.3 Data Derivation Rules 421

14.4 Documentation of Business Rules 422

14.4.1 Documentation in an E-R Diagram 422

14.4.2 Documenting Other Rules 422

14.4.3 Use of Subtypes to Document Rules 424

14.5 Implementing Business Rules 427

14.5.1 Where to Implement Particular Rules 428

14.5.2 Implementation Options: A Detailed Example 433

14.5.3 Implementing Mandatory Relationships 436

14.5.4 Referential Integrity 438

14.5.5 Restricting an Attribute to a Discrete Set of Values 439

14.5.6 Rules Involving Multiple Attributes 442

14.5.7 Recording Data That Supports Rules 442

14.5.8 Rules That May Be Broken 443

14.5.9 Enforcement of Rules Through Primary Key Selection 445

14.6 Rules on Recursive Relationships 446

14.6.1 Types of Rules on Recursive Relationships 447

14.6.2 Documenting Rules on Recursive Relationships 449

14.6.3 Implementing Constraints on Recursive Relationships 449

14.6.4 Analogous Rules in Many-to-Many Relationships 450

14.7 Summary 450

■ ■ ■ Chapter 15

Time-Dependent Data 451

15.1 The Problem 451

15.2 When Do We Add the Time Dimension? 452

15.3 Audit Trails and Snapshots 452

15.3.1 The Basic Audit Trail Approach 453

15.3.2 Handling Nonnumeric Data 458

15.3.3 The Basic Snapshot Approach 458

15.4 Sequences and Versions 462

15.5 Handling Deletions 463

15.6 Archiving 463

- 15.7 **Modeling Time-Dependent Relationships 464**
 - 15.7.1 One-to-Many Relationships 464
 - 15.7.2 Many-to-Many Relationships 466
 - 15.7.3 Self-Referencing Relationships 468
- 15.8 **Date Tables 469**
- 15.9 **Temporal Business Rules 469**
- 15.10 **Changes to the Data Structure 473**
- 15.11 **Putting It into Practice 473**
- 15.12 **Summary 474**
- ■ ■ **Chapter 16**
- Modeling for Data Warehouses and Data Marts 475**
- 16.1 **Introduction 475**
- 16.2 **Characteristics of Data Warehouses and Data Marts 478**
 - 16.2.1 Data Integration: Working with Existing Databases 478
 - 16.2.2 Loads Rather Than Updates 478
 - 16.2.3 Less Predictable Database “Hits” 479
 - 16.2.4 Complex Queries—Simple Interface 479
 - 16.2.5 History 480
 - 16.2.6 Summarization 480
- 16.3 **Quality Criteria for Warehouse and Mart Models 480**
 - 16.3.1 Completeness 480
 - 16.3.2 Nonredundancy 481
 - 16.3.3 Enforcement of Business Rules 482
 - 16.3.4 Data Reusability 482
 - 16.3.5 Stability and Flexibility 482
 - 16.3.6 Simplicity and Elegance 483
 - 16.3.7 Communication Effectiveness 483
 - 16.3.8 Performance 483
- 16.4 **The Basic Design Principle 483**
- 16.5 **Modeling for the Data Warehouse 484**
 - 16.5.1 An Initial Model 484
 - 16.5.2 Understanding Existing Data 485
 - 16.5.3 Determining Requirements 485
 - 16.5.4 Determining Sources and Dealing with Differences 485
 - 16.5.5 Shaping Data for Data Marts 487

- 16.6 **Modeling for the Data Mart 488**
 - 16.6.1 The Basic Challenge 488
 - 16.6.2 Multidimensional Databases, Stars and Snowflakes 488
 - 16.6.3 Modeling Time-Dependent Data 494
- 16.7 **Summary 496**
- ■ ■ **Chapter 17**
- Enterprise Data Models and Data Management 499**
- 17.1 **Introduction 499**
- 17.2 **Data Management 500**
 - 17.2.1 Problems of Data Mismanagement 500
 - 17.2.2 Managing Data as a Shared Resource 501
 - 17.2.3 The Evolution of Data Management 501
- 17.3 **Classification of Existing Data 503**
- 17.4 **A Target for Planning 504**
- 17.5 **A Context for Specifying New Databases 506**
 - 17.5.1 Determining Scope and Interfaces 506
 - 17.5.2 Incorporating the Enterprise Data Model in the Development Life Cycle 506
- 17.6 **Guidance for Database Design 508**
- 17.7 **Input to Business Planning 508**
- 17.8 **Specification of an Enterprise Database 509**
- 17.9 **Characteristics of Enterprise Data Models 511**
- 17.10 **Developing an Enterprise Data Model 512**
 - 17.10.1 The Development Cycle 512
 - 17.10.2 Partitioning the Task 513
 - 17.10.3 Inputs to the Task 514
 - 17.10.4 Expertise Requirements 515
 - 17.10.5 External Standards 515
- 17.11 **Choice, Creativity, and Enterprise Data Models 516**
- 17.12 **Summary 517**
- Further Reading 519**