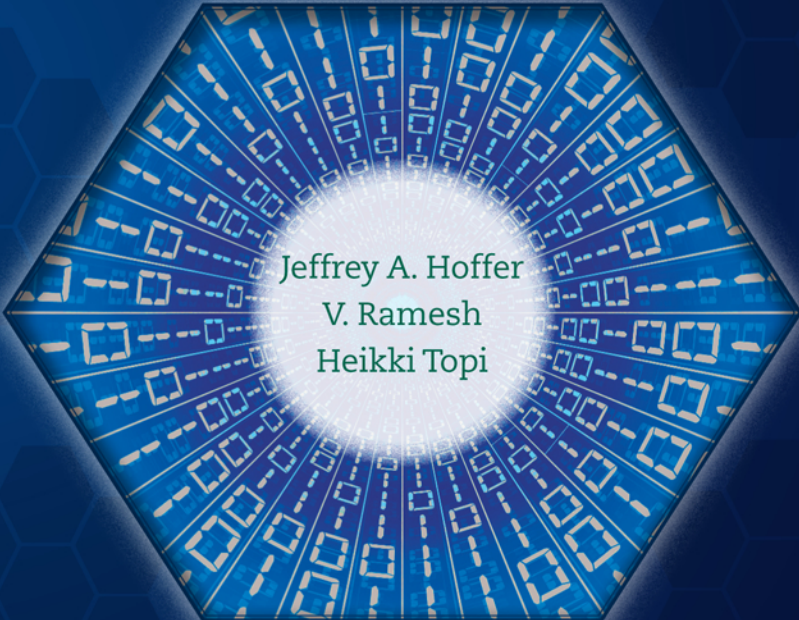


Modern

Database 12e

Management



Jeffrey A. Hoffer

V. Ramesh

Heikki Topi

OTHER MIS TITLES OF INTEREST

Introductory MIS:

Managing Information Technology, 7/e
Brown, DeHayes, Hoffer, Martin & Perkins ©2012

Experiencing MIS, 6/e
Kroenke & Boyle ©2016

Using MIS, 8/e
Kroenke & Boyle ©2016

MIS Essentials, 4/e
Kroenke ©2015

Management Information Systems, 14/e
Laudon & Laudon ©2016

Essentials of Management Information Systems, 11/e
Laudon & Laudon ©2015

IT Strategy, 3/e
McKeen & Smith ©2015

Processes, Systems, and Information: An Introduction to MIS, 2/e
McKinney & Kroenke ©2015

Information Systems Today, 7/e
Valacich & Schneider ©2016

Introduction to Information Systems, 2/e
Wallace ©2015

Database:

Hands-on Database, 2/e
Conger ©2014

Modern Database Management, 12/e
Hoffer, Ramesh & Topi ©2016

Database Systems: Introduction to Databases and Data Warehouses
Jukic, Vrbsky & Nestorov ©2014

Essentials of Database Management
Hoffer, Topi & Ramesh ©2014

Database Concepts, 7/e
Kroenke & Auer ©2015

Database Processing, 14/e
Kroenke & Auer ©2016

Systems Analysis and Design:

Modern Systems Analysis and Design, 7/e
Hoffer, George & Valacich ©2014

Systems Analysis and Design, 9/e
Kendall & Kendall ©2014

Essentials of Systems Analysis and Design, 6/e
Valacich, George & Hoffer ©2015

Decision Support Systems:

Business Intelligence, 3/e
Sharda, Delen & Turban ©2014

Decision Support and Business Intelligence Systems, 10/e
Sharda, Delen & Turban ©2014

Data Communications & Networking:

Applied Networking Labs, 2/e
Boyle ©2014

Digital Business Networks
Dooley ©2014

Business Data Networks and Security, 10/e
Panko & Panko ©2015

Electronic Commerce:

E-Commerce: Business, Technology, Society, 11/e
Laudon & Traver ©2015

Enterprise Resource Planning:

Enterprise Systems for Management, 2/e
Motiwalla & Thompson ©2012

Project Management:

Project Management: Process, Technology and Practice
Vaidyanathan ©2013

This page intentionally left blank

Twelfth Edition

MODERN DATABASE MANAGEMENT

This page intentionally left blank

Twelfth Edition

MODERN DATABASE MANAGEMENT

Jeffrey A. Hoffer

University of Dayton

V. Ramesh

Indiana University

Heikki Topi

Bentley University

PEARSON

Boston Columbus Indianapolis New York San Francisco Hoboken
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Vice President, Business Publishing: Donna Battista
Editor-in-Chief: Stephanie Wall
Acquisitions Editor: Nicole Sam
Program Manager Team Lead: Ashley Santora
Program Manager: Denise Weiss
Editorial Assistant: Olivia Vignone
Vice President, Product Marketing: Maggie Moylan
Director of Marketing, Digital Services and Products:
Jeanette Koskinas
Field Marketing Manager: Lenny Ann Raper
Senior Strategic Marketing Manager: Erin Gardner
Product Marketing Assistant: Jessica Quazza
Project Manager Team Lead: Jeff Holcomb
Project Manager: Ilene Kahn
Operations Specialist: Diane Peirano
Creative Director: Blair Brown
Senior Art Director: Janet Slowik

Interior and Cover Designer: Emily/Integra Software Solutions
Cover Image: Dreaming Andy/Fotolia
Vice President, Director of Digital Strategy & Assessment:
Paul Gentile
Manager of Learning Applications: Paul Deluca
Digital Editor: Brian Surette
Digital Studio Manager: Diane Lombardo
Digital Studio Project Manager: Robin Lazrus
Digital Studio Project Manager: Alana Coles
Digital Studio Project Manager: Monique Lawrence
Digital Studio Project Manager: Regina DaSilva
Full-Service Project Management and Composition:
George Jacob/Integra Software Solutions Pvt., Ltd
Printer/Binder: Edwards Brothers
Cover Printer: Phoenix Color
Text Font: 10/12 PalatinoLTStd Roman

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on the appropriate page within text.

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided “as is” without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Trademarks

Microsoft[®], Windows[®], and Microsoft Office[®] are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

Copyright © 2016, 2013, 2011 by Pearson Education, Inc. All rights reserved. Manufactured in the United States of America. This publication is protected by Copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsoned.com/permissions/.

Acknowledgements of third party content appear on the appropriate page within the text, which constitutes an extension of this copyright page.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

Library of Congress Cataloging-in-Publication Data

Hoffer, Jeffrey A.
Modern database management/Jeffrey A. Hoffer, University of Dayton,
V. Ramesh, Indiana University, Heikki Topi, Bentley University. —Twelfth edition.
pages cm
ISBN 978-0-13-354461-9 — ISBN 0-13-354461-3
1. Database management. I. Ramesh, V. (Venkataraman) II. Topi, Heikki. III. Title.
QA76.9.D3M395 2015
005.74—dc23

2015003732

10 9 8 7 6 5 4 3 2 1

PEARSON

ISBN 10: 0-13-354461-3
ISBN 13: 978-0-13-354461-9

To Patty, for her sacrifices, encouragement, and support for more than 30 years of being a textbook author widow. To my students and colleagues, for being receptive and critical and for challenging me to be a better teacher.

—J.A.H.

To Gayathri, for her sacrifices and patience these past 25 years. To my parents, for letting me make the journey abroad, and to my cat, Raju, who was a part of our family for more than 20 years.

—V.R.

To Anne-Louise, for her loving support, encouragement, and patience. To Leila and Saara, whose laughter and joy of life continue to teach me about what is truly important. To my teachers, colleagues, and students, from whom I continue to learn every day.

—H.T.

This page intentionally left blank

BRIEF CONTENTS

Part I The Context of Database Management 1

Chapter 1 The Database Environment and Development Process 2

Part II Database Analysis 51

Chapter 2 Modeling Data in the Organization 53

Chapter 3 The Enhanced E-R Model 114

Part III Database Design 153

Chapter 4 Logical Database Design and the Relational Model 155

Chapter 5 Physical Database Design and Performance 206

Part IV Implementation 241

Chapter 6 Introduction to SQL 243

Chapter 7 Advanced SQL 289

Chapter 8 Database Application Development 337

Chapter 9 Data Warehousing 374

Part V Advanced Database Topics 417

Chapter 10 Data Quality and Integration 419

Chapter 11 Big Data and Analytics 445

Chapter 12 Data and Database Administration 485

Glossary of Acronyms 534

Glossary of Terms 536

Index 544

Available Online at www.pearsonhighered.com/hoffer

Chapter 13 Distributed Databases 13-1

Chapter 14 Object-Oriented Data Modeling 14-1

Appendices

Appendix A Data Modeling Tools and Notation A-1

Appendix B Advanced Normal Forms B-1

Appendix C Data Structures C-1

This page intentionally left blank

CONTENTS

Preface xxv

Part I The Context of Database Management 1

An Overview of Part One 1

Chapter 1 The Database Environment and Development Process 2

Learning Objectives 2

Data Matter! 2

Introduction 3

Basic Concepts and Definitions 5

Data 5

Data Versus Information 5

Metadata 6

Traditional File Processing Systems 7

File Processing Systems at Pine Valley Furniture Company 8

Disadvantages of File Processing Systems 8

PROGRAM-DATA DEPENDENCE 8

DUPPLICATION OF DATA 9

LIMITED DATA SHARING 9

LENGTHY DEVELOPMENT TIMES 9

EXCESSIVE PROGRAM MAINTENANCE 9

The Database Approach 9

Data Models 9

ENTITIES 10

RELATIONSHIPS 11

Relational Databases 11

Database Management Systems 11

Advantages of the Database Approach 11

PROGRAM-DATA INDEPENDENCE 11

PLANNED DATA REDUNDANCY 12

IMPROVED DATA CONSISTENCY 12

IMPROVED DATA SHARING 12

INCREASED PRODUCTIVITY OF APPLICATION DEVELOPMENT 13

ENFORCEMENT OF STANDARDS 13

IMPROVED DATA QUALITY 13

IMPROVED DATA ACCESSIBILITY AND RESPONSIVENESS 14

REDUCED PROGRAM MAINTENANCE 14

IMPROVED DECISION SUPPORT 14

Cautions About Database Benefits 14

Costs and Risks of the Database Approach 14

NEW, SPECIALIZED PERSONNEL 15

INSTALLATION AND MANAGEMENT COST AND COMPLEXITY 15

CONVERSION COSTS 15

NEED FOR EXPLICIT BACKUP AND RECOVERY 15

ORGANIZATIONAL CONFLICT 15

Components of the Database Environment 15



The Database Development Process 17

- Systems Development Life Cycle 18
 - PLANNING—ENTERPRISE MODELING 18
 - PLANNING—CONCEPTUAL DATA MODELING 18
 - ANALYSIS—CONCEPTUAL DATA MODELING 18
 - DESIGN—LOGICAL DATABASE DESIGN 19
 - DESIGN—PHYSICAL DATABASE DESIGN AND DEFINITION 20
 - IMPLEMENTATION—DATABASE IMPLEMENTATION 20
 - MAINTENANCE—DATABASE MAINTENANCE 20
- Alternative Information Systems (IS) Development Approaches 21
- Three-Schema Architecture for Database Development 22
- Managing the People Involved in Database Development 24

Evolution of Database Systems 24

- 1960s 26
- 1970s 26
- 1980s 26
- 1990s 26
- 2000 and Beyond 27

The Range of Database Applications 27

- Personal Databases 28
- Multitier Client/Server Databases 28
- Enterprise Applications 29

Developing a Database Application for Pine Valley Furniture Company 31

- Database Evolution at Pine Valley Furniture Company 32
- Project Planning 33
- Analyzing Database Requirements 34
- Designing the Database 36
- Using the Database 39
- Administering the Database 40
- Future of Databases at Pine Valley 41
 - Summary 41 • Key Terms 42 • Review Questions 42 • Problems and Exercises 44 • Field Exercises 45 • References 46 • Further Reading 46 • Web Resources 47*
 - ▶ **CASE: Forondo Artist Management Excellence Inc. 48**



Part II Database Analysis 51

An Overview of Part Two 51

Chapter 2 Modeling Data in the Organization 53

Learning Objectives 53

Introduction 53

The E-R Model: An Overview 56

- Sample E-R Diagram 56

- E-R Model Notation 58

Modeling the Rules of the Organization 59

- Overview of Business Rules 60

 - THE BUSINESS RULES PARADIGM 60



Scope of Business Rules	61
GOOD BUSINESS RULES	61
GATHERING BUSINESS RULES	62
Data Names and Definitions	62
DATA NAMES	62
DATA DEFINITIONS	63
GOOD DATA DEFINITIONS	63
Modeling Entities and Attributes	65
Entities	65
ENTITY TYPE VERSUS ENTITY INSTANCE	65
ENTITY TYPE VERSUS SYSTEM INPUT, OUTPUT, OR USER	65
STRONG VERSUS WEAK ENTITY TYPES	66
NAMING AND DEFINING ENTITY TYPES	67
Attributes	69
REQUIRED VERSUS OPTIONAL ATTRIBUTES	69
SIMPLE VERSUS COMPOSITE ATTRIBUTES	70
SINGLE-VALUED VERSUS MULTIVALUED ATTRIBUTES	70
STORED VERSUS DERIVED ATTRIBUTES	71
IDENTIFIER ATTRIBUTE	71
NAMING AND DEFINING ATTRIBUTES	72
Modeling Relationships	74
Basic Concepts and Definitions in Relationships	75
ATTRIBUTES ON RELATIONSHIPS	76
ASSOCIATIVE ENTITIES	76
Degree of a Relationship	78
UNARY RELATIONSHIP	78
BINARY RELATIONSHIP	80
TERNARY RELATIONSHIP	81
Attributes or Entity?	82
Cardinality Constraints	84
MINIMUM CARDINALITY	84
MAXIMUM CARDINALITY	84
Some Examples of Relationships and Their Cardinalities	85
A TERNARY RELATIONSHIP	86
Modeling Time-Dependent Data	86
Modeling Multiple Relationships Between Entity Types	89
Naming and Defining Relationships	90
E-R Modeling Example: Pine Valley Furniture Company	92
Database Processing at Pine Valley Furniture	94
Showing Product Information	95
Showing Product Line Information	95
Showing Customer Order Status	96
Showing Product Sales	97
<i>Summary</i>	98 • <i>Key Terms</i>
<i>Problems and Exercises</i>	99 • <i>Review Questions</i>
<i>Field Exercises</i>	99 • <i>100</i> • <i>Further Reading</i>
<i>References</i>	110 • <i>Web Resources</i>
<i>111</i>	111 • <i>112</i>
▶ CASE: Forondo Artist Management Excellence Inc.	112

Chapter 3 The Enhanced E-R Model 114

Learning Objectives	114
Introduction	114



Representing Supertypes and Subtypes 115

 Basic Concepts and Notation 116

 AN EXAMPLE OF A SUPERTYPE/SUBTYPE RELATIONSHIP 117

 ATTRIBUTE INHERITANCE 118

 WHEN TO USE SUPERTYPE/SUBTYPE RELATIONSHIPS 118

 Representing Specialization and Generalization 119

 GENERALIZATION 119

 SPECIALIZATION 120

 COMBINING SPECIALIZATION AND GENERALIZATION 121

Specifying Constraints in Supertype/Subtype Relationships 122

 Specifying Completeness Constraints 122

 TOTAL SPECIALIZATION RULE 122

 PARTIAL SPECIALIZATION RULE 122

 Specifying Disjointness Constraints 123

 DISJOINT RULE 123

 OVERLAP RULE 123

 Defining Subtype Discriminators 124

 DISJOINT SUBTYPES 124

 OVERLAPPING SUBTYPES 125

 Defining Supertype/Subtype Hierarchies 125

 AN EXAMPLE OF A SUPERTYPE/SUBTYPE HIERARCHY 126

 SUMMARY OF SUPERTYPE/SUBTYPE HIERARCHIES 127

EER Modeling Example: Pine Valley Furniture Company 128

Entity Clustering 131

Packaged Data Models 134

 A Revised Data Modeling Process with Packaged Data Models 136

 Packaged Data Model Examples 138

Summary 143 • Key Terms 144 • Review Questions 144 • Problems and Exercises 145 • Field Exercises 148 • References 148 • Further Reading 148 • Web Resources 149

 ▶ **CASE: Forondo Artist Management Excellence Inc. 150**



Part III Database Design 153

An Overview of Part Three 153

Chapter 4 Logical Database Design and the Relational Model 155



Learning Objectives 155

Introduction 155

The Relational Data Model 156

 Basic Definitions 156

 RELATIONAL DATA STRUCTURE 157

 RELATIONAL KEYS 157

 PROPERTIES OF RELATIONS 158

 REMOVING MULTIVALUED ATTRIBUTES FROM TABLES 158

 Sample Database 158

Integrity Constraints 160

 Domain Constraints 160

 Entity Integrity 160

 Referential Integrity 162

Creating Relational Tables	163
Well-Structured Relations	164
Transforming EER Diagrams into Relations	165
Step 1: Map Regular Entities	166
COMPOSITE ATTRIBUTES	166
MULTIVALUED ATTRIBUTES	167
Step 2: Map Weak Entities	167
WHEN TO CREATE A SURROGATE KEY	169
Step 3: Map Binary Relationships	169
MAP BINARY ONE-TO-MANY RELATIONSHIPS	169
MAP BINARY MANY-TO-MANY RELATIONSHIPS	170
MAP BINARY ONE-TO-ONE RELATIONSHIPS	170
Step 4: Map Associative Entities	171
IDENTIFIER NOT ASSIGNED	172
IDENTIFIER ASSIGNED	172
Step 5: Map Unary Relationships	173
UNARY ONE-TO-MANY RELATIONSHIPS	173
UNARY MANY-TO-MANY RELATIONSHIPS	174
Step 6: Map Ternary (and n -ary) Relationships	175
Step 7: Map Supertype/Subtype Relationships	176
Summary of EER-to-Relational Transformations	178
Introduction to Normalization	178
Steps in Normalization	179
Functional Dependencies and Keys	179
DETERMINANTS	181
CANDIDATE KEYS	181
Normalization Example: Pine Valley Furniture Company	182
Step 0: Represent the View in Tabular Form	182
Step 1: Convert to First Normal Form	183
REMOVE REPEATING GROUPS	183
SELECT THE PRIMARY KEY	183
ANOMALIES IN 1NF	184
Step 2: Convert to Second Normal Form	185
Step 3: Convert to Third Normal Form	186
REMOVING TRANSITIVE DEPENDENCIES	186
Determinants and Normalization	187
Step 4: Further Normalization	187
Merging Relations	188
An Example	188
View Integration Problems	188
SYNONYMS	189
HOMONYMS	189
TRANSITIVE DEPENDENCIES	189
SUPERTYPE/SUBTYPE RELATIONSHIPS	190
A Final Step for Defining Relational Keys	190
<i>Summary</i>	192
<i>Key Terms</i>	194
<i>Review Questions</i>	194
<i>Problems and Exercises</i>	195
<i>Field Exercises</i>	204
<i>References</i>	204
<i>Further Reading</i>	204
<i>Web Resources</i>	204
▶ CASE: Forondo Artist Management Excellence Inc.	205



Chapter 5 Physical Database Design and Performance 206

- Learning Objectives 206
- Introduction 206
- The Physical Database Design Process 207
 - Physical Database Design as a Basis for Regulatory Compliance 208
 - Data Volume and Usage Analysis 209
- Designing Fields 210
 - Choosing Data Types 211
 - CODING TECHNIQUES 212
 - HANDLING MISSING DATA 213
- Denormalizing and Partitioning Data 213
 - Denormalization 213
 - OPPORTUNITIES FOR AND TYPES OF DENORMALIZATION 214
 - DENORMALIZE WITH CAUTION 216
 - Partitioning 217
- Designing Physical Database Files 219
 - File Organizations 221
 - HEAP FILE ORGANIZATION 221
 - SEQUENTIAL FILE ORGANIZATIONS 221
 - INDEXED FILE ORGANIZATIONS 221
 - HASHED FILE ORGANIZATIONS 224
 - Clustering Files 227
 - Designing Controls for Files 227
- Using and Selecting Indexes 228
 - Creating a Unique Key Index 228
 - Creating a Secondary (Nonunique) Key Index 228
 - When to Use Indexes 229
- Designing a Database for Optimal Query Performance 230
 - Parallel Query Processing 230
 - Overriding Automatic Query Optimization 231
 - Summary 232 • Key Terms 233 • Review Questions 233 • Problems and Exercises 234 • Field Exercises 237 • References 237 • Further Reading 237 • Web Resources 238*
- ▶ **CASE: Forondo Artist Management Excellence Inc. 239**



Part IV Implementation 241

An Overview of Part Four 241



Chapter 6 Introduction to SQL 243

- Learning Objectives 243
- Introduction 243
- Origins of the SQL Standard 245
- The SQL Environment 247
- Defining a Database in SQL 251
 - Generating SQL Database Definitions 252
 - Creating Tables 253
 - Creating Data Integrity Controls 255
 - Changing Table Definitions 256
 - Removing Tables 257

- Inserting, Updating, and Deleting Data 257
 - Batch Input 259
 - Deleting Database Contents 259
 - Updating Database Contents 259
- Internal Schema Definition in RDBMSs 260
 - Creating Indexes 260
- Processing Single Tables 261
 - Clauses of the SELECT Statement 262
 - Using Expressions 264
 - Using Functions 265
 - Using Wildcards 267
 - Using Comparison Operators 267
 - Using Null Values 268
 - Using Boolean Operators 268
 - Using Ranges for Qualification 271
 - Using Distinct Values 271
 - Using IN and NOT IN with Lists 273
 - Sorting Results: The ORDER BY Clause 274
 - Categorizing Results: The GROUP BY Clause 275
 - Qualifying Results by Categories: The HAVING Clause 276
 - Using and Defining Views 277
 - MATERIALIZED VIEWS 281
 - [Summary](#) 281 • [Key Terms](#) 282 • [Review Questions](#) 282 • [Problems and Exercises](#) 283 • [Field Exercises](#) 286 • [References](#) 287 • [Further Reading](#) 287 • [Web Resources](#) 287
 - ▶ [CASE: Forondo Artist Management Excellence Inc.](#) 288



Chapter 7 Advanced SQL 289

- Learning Objectives 289
- Introduction 289
- Processing Multiple Tables 290
 - Equi-join 291
 - Natural Join 292
 - Outer Join 293
 - Sample Join Involving Four Tables 295
 - Self-Join 297
 - Subqueries 298
 - Correlated Subqueries 303
 - Using Derived Tables 305
 - Combining Queries 306
 - Conditional Expressions 308
 - More Complicated SQL Queries 308
- Tips for Developing Queries 310
 - Guidelines for Better Query Design 312
- Ensuring Transaction Integrity 314
- Data Dictionary Facilities 315
- Recent Enhancements and Extensions to SQL 317
 - Analytical and OLAP Functions 317
 - New Data Types 319

- New Temporal Features in SQL 319
- Other Enhancements 320
- Triggers and Routines 321
 - Triggers 321
 - Routines and other Programming Extensions 323
 - Example Routine in Oracle's PL/SQL 325
- Embedded SQL and Dynamic SQL 327
 - Summary 329 • Key Terms 330 • Review Questions 330 • Problems and Exercises 331 • Field Exercises 334 • References 334 • Further Reading 334 • Web Resources 335*
 - ▶ **CASE: Forondo Artist Management Excellence Inc. 336**



Chapter 8 Database Application Development 337

- Learning Objectives 337
- Location, Location, Location! 337
- Introduction 338
- Client/Server Architectures 338
- Databases in a Two-Tier Architecture 340
 - A VB.NET Example 342
 - A Java Example 344
- Three-Tier Architectures 345
- Web Application Components 347
- Databases in Three-Tier Applications 349
 - A JSP Web Application 349
 - A PHP Example 353
 - An ASP.NET Example 355
- Key Considerations in Three-Tier Applications 356
 - Stored Procedures 356
 - Transactions 359
 - Database Connections 359
 - Key Benefits of Three-Tier Applications 359
 - Cloud Computing and Three-Tier Applications 360
- Extensible Markup Language (XML) 361
 - Storing XML Documents 363
 - Retrieving XML Documents 363
 - Displaying XML Data 366
 - XML and Web Services 366
 - Summary 369 • Key Terms 370 • Review Questions 370 • Problems and Exercises 371 • Field Exercises 372 • References 372 • Further Reading 372 • Web Resources 372*
 - ▶ **CASE: Forondo Artist Management Excellence Inc. 373**



Chapter 9 Data Warehousing 374

- Learning Objectives 374
- Introduction 374
- Basic Concepts of Data Warehousing 376
 - A Brief History of Data Warehousing 377
 - The Need for Data Warehousing 377
 - NEED FOR A COMPANY-WIDE VIEW 377
 - NEED TO SEPARATE OPERATIONAL AND INFORMATIONAL SYSTEMS 379

Data Warehouse Architectures	380
Independent Data Mart Data Warehousing Environment	380
Dependent Data Mart and Operational Data Store Architecture: A Three-Level Approach	382
Logical Data Mart and Real-Time Data Warehouse Architecture	384
Three-Layer Data Architecture	387
ROLE OF THE ENTERPRISE DATA MODEL	388
ROLE OF METADATA	388
Some Characteristics of Data Warehouse Data	388
Status Versus Event Data	388
Transient Versus Periodic Data	389
An Example of Transient and Periodic Data	389
TRANSIENT DATA	389
PERIODIC DATA	391
OTHER DATA WAREHOUSE CHANGES	391
The Derived Data Layer	392
Characteristics of Derived Data	392
The Star Schema	393
FACT TABLES AND DIMENSION TABLES	393
EXAMPLE STAR SCHEMA	394
SURROGATE KEY	395
GRAIN OF THE FACT TABLE	396
DURATION OF THE DATABASE	397
SIZE OF THE FACT TABLE	397
MODELING DATE AND TIME	398
Variations of the Star Schema	399
MULTIPLE FACT TABLES	399
FACTLESS FACT TABLES	400
Normalizing Dimension Tables	401
MULTIVALUED DIMENSIONS	401
HIERARCHIES	402
Slowly Changing Dimensions	404
Determining Dimensions and Facts	406
The Future of Data Warehousing: Integration with Big Data and Analytics	408
Speed of Processing	409
Cost of Storing Data	409
Dealing with Unstructured Data	409
<i>Summary</i>	410
<i>Key Terms</i>	410
<i>Review Questions</i>	411
<i>Problems and Exercises</i>	411
<i>Field Exercises</i>	415
<i>References</i>	415
<i>Further Reading</i>	416
<i>Web Resources</i>	416

Part V Advanced Database Topics 417

An Overview of Part Five 417

Chapter 10 Data Quality and Integration 419

Learning Objectives 419

Introduction 419

Data Governance 420

Managing Data Quality	421
Characteristics of Quality Data	422
EXTERNAL DATA SOURCES	423
REDUNDANT DATA STORAGE AND INCONSISTENT METADATA	424
DATA ENTRY PROBLEMS	424
LACK OF ORGANIZATIONAL COMMITMENT	424
Data Quality Improvement	424
GET THE BUSINESS BUY-IN	424
CONDUCT A DATA QUALITY AUDIT	425
ESTABLISH A DATA STEWARDSHIP PROGRAM	426
IMPROVE DATA CAPTURE PROCESSES	426
APPLY MODERN DATA MANAGEMENT PRINCIPLES AND TECHNOLOGY	427
APPLY TQM PRINCIPLES AND PRACTICES	427
Summary of Data Quality	427
Master Data Management	428
Data Integration: An Overview	429
General Approaches to Data Integration	429
DATA FEDERATION	430
DATA PROPAGATION	431
Data Integration for Data Warehousing: The Reconciled Data Layer	431
Characteristics of Data After ETL	431
The ETL Process	432
MAPPING AND METADATA MANAGEMENT	432
EXTRACT	433
CLEANSE	434
LOAD AND INDEX	436
Data Transformation	437
Data Transformation Functions	438
RECORD-LEVEL FUNCTIONS	438
FIELD-LEVEL FUNCTIONS	439
Summary	441
Key Terms	441
Review Questions	441
Problems and Exercises	442
Field Exercises	443
References	443
Further Reading	444
Web Resources	444

Chapter 11 Big Data and Analytics 445

Learning Objectives	445
Introduction	445
Big Data	447
NoSQL	449
Classification of NoSQL Database Management Systems	450
KEY-VALUE STORES	450
DOCUMENT STORES	450
WIDE-COLUMN STORES	451
GRAPH-ORIENTED DATABASES	451
NoSQL Examples	452
REDIS	452
MONGODB	452
APACHE CASSANDRA	452
NEO4J	452
Impact of NoSQL on Database Professionals	452

Hadoop	453
Components of Hadoop	454
THE HADOOP DISTRIBUTED FILE SYSTEM (HDFS)	454
MAPREDUCE	455
PIG	456
HIVE	456
HBASE	457
Integrated Analytics and Data Science Platforms	457
HP HAVEN	457
TERADATA ASTER	457
IBM BIG DATA PLATFORM	457
Putting It All Together: Integrated Data Architecture	458
Analytics	460
Types of Analytics	461
Use of Descriptive Analytics	462
SQL OLAP QUERYING	463
ONLINE ANALYTICAL PROCESSING (OLAP) TOOLS	465
DATA VISUALIZATION	467
BUSINESS PERFORMANCE MANAGEMENT AND DASHBOARDS	469
Use of Predictive Analytics	470
DATA MINING TOOLS	470
EXAMPLES OF PREDICTIVE ANALYTICS	472
Use of Prescriptive Analytics	473
Data Management Infrastructure for Analytics	474
Impact of Big Data and Analytics	476
Applications of Big Data and Analytics	476
BUSINESS	477
E-GOVERNMENT AND POLITICS	477
SCIENCE AND TECHNOLOGY	478
SMART HEALTH AND WELL-BEING	478
SECURITY AND PUBLIC SAFETY	478
Implications of Big Data Analytics and Decision Making	478
PERSONAL PRIVACY VS. COLLECTIVE BENEFITS	479
OWNERSHIP AND ACCESS	479
QUALITY AND REUSE OF DATA AND ALGORITHMS	479
TRANSPARENCY AND VALIDATION	480
CHANGING NATURE OF WORK	480
DEMANDS FOR WORKFORCE CAPABILITIES AND EDUCATION	480
<i>Summary</i>	480 •
<i>Key Terms</i>	481 •
<i>Review Questions</i>	481 •
<i>Problems and Exercises</i>	482 •
<i>References</i>	483 •
<i>Further Reading</i>	484 •
<i>Web Resources</i>	484

Chapter 12 Data and Database Administration 485

Learning Objectives	485
Introduction	485
The Roles of Data and Database Administrators	486
Traditional Data Administration	486
Traditional Database Administration	488
Trends in Database Administration	489
Data Warehouse Administration	491
Summary of Evolving Data Administration Roles	492

- The Open Source Movement and Database Management 492
- Managing Data Security 494
 - Threats to Data Security 495
 - Establishing Client/Server Security 496
 - SERVER SECURITY 496
 - NETWORK SECURITY 496
 - Application Security Issues in Three-Tier Client/Server Environments 497
 - DATA PRIVACY 498
- Database Software Data Security Features 499
 - Views 500
 - Integrity Controls 500
 - Authorization Rules 502
 - User-Defined Procedures 503
 - Encryption 503
 - Authentication Schemes 504
 - PASSWORDS 505
 - STRONG AUTHENTICATION 505
- Sarbanes-Oxley (SOX) and Databases 505
 - IT Change Management 506
 - Logical Access to Data 506
 - PERSONNEL CONTROLS 506
 - PHYSICAL ACCESS CONTROLS 507
 - IT Operations 507
- Database Backup and Recovery 507
 - Basic Recovery Facilities 508
 - BACKUP FACILITIES 508
 - JOURNALIZING FACILITIES 508
 - CHECKPOINT FACILITY 509
 - RECOVERY MANAGER 509
 - Recovery and Restart Procedures 510
 - DISK MIRRORING 510
 - RESTORE/RERUN 510
 - MAINTAINING TRANSACTION INTEGRITY 510
 - BACKWARD RECOVERY 512
 - FORWARD RECOVERY 513
 - Types of Database Failure 513
 - ABORTED TRANSACTIONS 513
 - INCORRECT DATA 513
 - SYSTEM FAILURE 514
 - DATABASE DESTRUCTION 514
 - Disaster Recovery 514
- Controlling Concurrent Access 515
 - The Problem of Lost Updates 515
 - Serializability 515
 - Locking Mechanisms 516
 - LOCKING LEVEL 516
 - TYPES OF LOCKS 517
 - DEADLOCK 518
 - MANAGING DEADLOCK 518

Versioning	519
Data Dictionaries and Repositories	521
Data Dictionary	521
Repositories	521
Overview of Tuning the Database for Performance	523
Installation of the DBMS	523
Memory and Storage Space Usage	523
Input/Output (I/O) Contention	524
CPU Usage	524
Application Tuning	525
Data Availability	526
Costs of Downtime	526
Measures to Ensure Availability	526
HARDWARE FAILURES	527
LOSS OR CORRUPTION OF DATA	527
HUMAN ERROR	527
MAINTENANCE DOWNTIME	527
NETWORK-RELATED PROBLEMS	527
<i>Summary</i>	<i>528</i>
<i>Key Terms</i>	<i>528</i>
<i>Review Questions</i>	<i>529</i>
<i>Problems and Exercises</i>	<i>530</i>
<i>Field Exercises</i>	<i>532</i>
<i>References</i>	<i>532</i>
<i>Further Reading</i>	<i>533</i>
<i>Web Resources</i>	<i>533</i>
<i>Glossary of Acronyms</i>	<i>534</i>
<i>Glossary of Terms</i>	<i>536</i>
<i>Index</i>	<i>544</i>

ONLINE CHAPTERS

Chapter 13 Distributed Databases 13-1

Learning Objectives 13-1

Introduction 13-1

Objectives and Trade-offs 13-4

Options for Distributing a Database 13-6

Data Replication 13-6

SNAPSHOT REPLICATION 13-7

NEAR-REAL-TIME REPLICATION 13-8

PULL REPLICATION 13-8

DATABASE INTEGRITY WITH REPLICATION 13-8

WHEN TO USE REPLICATION 13-8

Horizontal Partitioning 13-9

Vertical Partitioning 13-10

Combinations of Operations 13-11

Selecting the Right Data Distribution Strategy 13-11

Distributed DBMS 13-13

Location Transparency 13-15

Replication Transparency 13-16

Failure Transparency 13-17

Commit Protocol 13-17

Concurrency Transparency 13-18

TIME-STAMPING 13-18

Query Optimization 13-19

Evolution of Distributed DBMSs 13-21

REMOTE UNIT OF WORK 13-22

DISTRIBUTED UNIT OF WORK 13-22

DISTRIBUTED REQUEST 13-23

Summary 13-23 • Key Terms 13-24 • Review Questions 13-24 •

Problems and Exercises 13-25 • Field Exercises 13-26 •

References 13-27 • Further Reading 13-27 •

Web Resources 13-27

Chapter 14 Object-Oriented Data Modeling 14-1

Learning Objectives 14-1

Introduction 14-1

Unified Modeling Language 14-3

Object-Oriented Data Modeling 14-4

Representing Objects and Classes 14-4

Types of Operations 14-7

Representing Associations 14-7

Representing Association Classes 14-11

Representing Derived Attributes, Derived Associations,
and Derived Roles 14-12

Representing Generalization 14-13

Interpreting Inheritance and Overriding 14-18

Representing Multiple Inheritance 14-19

Representing Aggregation 14-19

Business Rules 14-22

Object Modeling Example: Pine Valley Furniture Company 14-23

Summary 14-25 • *Key Terms* 14-26 • *Review Questions* 14-26 •
Problems and Exercises 14-30 • *Field Exercises* 14-37 •
References 14-37 • *Further Reading* 14-38 •
Web Resources 14-38



Appendix A Data Modeling Tools and Notation A-1

Comparing E-R Modeling Conventions A-1

Visio Professional 2013 Notation A-1

ENTITIES A-5

RELATIONSHIPS A-5

CA ERwin Data Modeler 9.5 Notation A-5

ENTITIES A-5

RELATIONSHIPS A-5

SAP Sybase PowerDesigner 16.5 Notation A-7

ENTITIES A-8

RELATIONSHIPS A-8

Oracle Designer Notation A-8

ENTITIES A-8

RELATIONSHIPS A-8

Comparison of Tool Interfaces and E-R Diagrams A-8

Appendix B Advanced Normal Forms B-1

Boyce-Codd Normal Form B-1

Anomalies in Student Advisor B-1

Definition of Boyce-Codd Normal Form (BCNF) B-2

Converting a Relation to BCNF B-2

Fourth Normal Form B-3

Multivalued Dependencies B-5

Higher Normal Forms B-5

Key Terms B-6 • *References* B-6 • *Web Resource* B-6

Appendix C Data Structures C-1

Pointers C-1

Data Structure Building Blocks C-2

Linear Data Structures C-4

Stacks C-5

Queues C-5

Sorted Lists C-6

Multilists C-8

Hazards of Chain Structures C-8

Trees C-9

Balanced Trees C-9

Reference C-12

This page intentionally left blank

PREFACE

This text is designed to be used with an introductory course in database management. Such a course is usually required as part of an information systems curriculum in business schools, computer technology programs, and applied computer science departments. The Association for Information Systems (AIS), the Association for Computing Machinery (ACM), and the International Federation of Information Processing Societies (IFIPS) curriculum guidelines (e.g., IS 2010) all outline this type of database management course. Previous editions of this text have been used successfully for more than 33 years at both the undergraduate and graduate levels, as well as in management and professional development programs.

WHAT'S NEW IN THIS EDITION?

This 12th edition of *Modern Database Management* updates and expands materials in areas undergoing rapid change as a result of improved managerial practices, database design tools and methodologies, and database technology. Later, we detail changes to each chapter. The themes of this 12th edition reflect the major trends in the information systems field and the skills required of modern information systems graduates:

- Given the explosion in interest in the topics of big data and analytics, we have added an entire new chapter (Chapter 11) dedicated to this area. The chapter provides in-depth coverage of big data technologies such as NoSQL, Hadoop, MapReduce, Pig, and Hive and provides an introduction to the different types of analytics (descriptive, predictive, and prescriptive) and their use in business.
- We have also introduced this topic in relevant places throughout the textbook, e.g., in the revised introduction section in Chapter 1 as well as in a new section titled “The Future of Data Warehousing: Integration with Big Data and Analytics” in the data warehousing chapter (Chapter 9).
- Topics such as in-memory databases, in-database analytics, data warehousing in the cloud, and massively parallel processing are covered in sections of Chapter 9 and Chapter 11.
- The Mountain View Community Hospital (MVCH) case study (a staple of many past editions) has been replaced with a simpler mini-case titled “Forondo Artist Management Excellence Inc.” (FAME). The case focuses on the development of a system to support the needs of a small artist management company. The case is presented in the form of stakeholder e-mails describing the current challenges faced by the organization as well as the features they would like to see in a new system. Each chapter presents a set of project exercises that serve as guidelines for deliverables for students.
- We have updated the section on routines in Chapter 7 to provide clarity on the nature of routines and how to use them.
- New material added to Chapter 2 on why data modeling is important provides several compelling reasons for why data modeling is still crucial.

In addition to the new topics covered, specific improvements to the textbook have been made in the following areas:

- Every chapter went through significant edits to streamline coverage to ensure relevance with current technologies and eliminate redundancies.
- End-of-chapter material (review questions, problems and exercises, and/or field exercises) in every chapter has been revised with new questions and exercises.
- The figures in several chapters were updated to reflect the changing landscape of technologies that are being used in modern organizations.
- The Web Resources section in each chapter was updated to ensure that the student has information on the latest database trends and expanded background details on important topics covered in the text.



- We have continued to focus on reducing the length of the printed book, an effort that began with the eighth edition. The reduced length is more consistent with what our reviewers say can be covered in a database course today, given the need for depth of coverage in the most important topics. The reduced length should encourage more students to purchase and read the text, without any loss of coverage and learning. The book continues to be available through CourseSmart, an innovative e-book delivery system, and as an electronic book in the Kindle format.

Also, we continue to provide on the student Companion Web site several custom-developed short videos that address key concepts and skills from different sections of the book. These videos, produced by the textbook authors, help students learn difficult material by using both the printed text and a mini lecture or tutorial. Videos have been developed to support Chapters 1 (introduction to database), 2 and 3 (conceptual data modeling), 4 (normalization), and 6 and 7 (SQL). More will be produced with future editions. Look for special icons on the opening page of these chapters to call attention to these videos, and go to www.pearsonhighered.com/hoffer to find these videos.

FOR THOSE NEW TO *MODERN DATABASE MANAGEMENT*

Modern Database Management has been a leading text since its first edition in 1983. In spite of this market leadership position, some instructors have used other good database management texts. Why might you want to switch at this time? There are several good reasons:

- One of our goals, in every edition, has been to lead other books in coverage of the latest principles, concepts, and technologies. See what we have added for the 12th edition in “What’s New in This Edition?” In the past, we have led in coverage of object-oriented data modeling and UML, Internet databases, data warehousing, and the use of CASE tools in support of data modeling. For the 12th edition, we continue this tradition by providing significant coverage on the important topic of big data and analytics, focusing on what every database student needs to understand about these topics.
- While remaining current, this text focuses on what leading practitioners say is most important for database developers. We work with many practitioners, including the professionals of the Data Management Association (DAMA) and The Data Warehousing Institute (TDWI), leading consultants, technology leaders, and authors of articles in the most widely read professional publications. We draw on these experts to ensure that what the book includes is important and covers not only important entry-level knowledge and skills, but also those fundamentals and mind-sets that lead to long-term career success.
- In the 12th edition of this highly successful book, material is presented in a way that has been viewed as very accessible to students. Our methods have been refined through continuous market feedback for more than 30 years, as well as through our own teaching. Overall, the pedagogy of the book is sound. We use many illustrations that help make important concepts and techniques clear. We use the most modern notations. The organization of the book is flexible, so you can use chapters in whatever sequence makes sense for your students. We supplement the book with data sets to facilitate hands-on, practical learning, and with new media resources to make some of the more challenging topics more engaging.
- Our text can accommodate structural flexibility. For example, you may have particular interest in introducing SQL early in your course. Our text makes this possible. First, we cover SQL in depth, devoting two full chapters to this core technology of the database field. Second, we include many SQL examples in early chapters. Third, many instructors have successfully used the two SQL chapters early in their course. Although logically appearing in the life cycle of systems development as Chapters 6 and 7, part of the implementation section of the text, many instructors have used these chapters immediately after Chapter 1 or in parallel with other early chapters. Finally, we use SQL throughout the book, for example, to illustrate Web application connections to relational databases in Chapter 8 and online analytical processing in Chapter 11.

- We have the latest in supplements and Web site support for the text. See the supplement package for details on all the resources available to you and your students.
- This text is written to be part of a modern information systems curriculum with a strong business systems development focus. Topics are included and addressed so as to reinforce principles from other typical courses, such as systems analysis and design, networking, Web site design and development, MIS principles, and computer programming. Emphasis is on the development of the database component of modern information systems and on the management of the data resource. Thus, the text is practical, supports projects and other hands-on class activities, and encourages linking database concepts to concepts being learned throughout the curriculum the student is taking.

SUMMARY OF ENHANCEMENTS TO EACH CHAPTER

The following sections present a chapter-by-chapter description of the major changes in this edition. Each chapter description presents a statement of the purpose of that chapter, followed by a description of the changes and revisions that have been made for the 12th edition. Each paragraph concludes with a description of the strengths that have been retained from prior editions.

PART I: THE CONTEXT OF DATABASE MANAGEMENT

Chapter 1: The Database Environment and Development Process

This chapter discusses the role of databases in organizations and previews the major topics in the remainder of the text. The primary change in this chapter has been in how we use current examples around the explosion in the amount of data being generated and the benefits that can be gained by harnessing the power data (through analytics) to help set the stage for the entire book. A few new exercises have also been added, and the new Forondo Artist Management Excellence (FAME) case is introduced. After presenting a brief introduction to the basic terminology associated with storing and retrieving data, the chapter presents a well-organized comparison of traditional file processing systems and modern database technology. The chapter then introduces the core components of a database environment. It then goes on to explain the process of database development in the context of structured life cycle, prototyping, and agile methodologies. The presentation remains consistent with the companion textbook, *Modern Systems Analysis and Design* by Hoffer, George, and Valacich. The chapter also discusses important issues in database development, including management of the diverse group of people involved in database development and frameworks for understanding database architectures and technologies (e.g., the three-schema architecture). Reviewers frequently note the compatibility of this chapter with what students learn in systems analysis and design classes. A brief history of the evolution of database technology, from pre-database files to modern object-relational technologies, is presented. The chapter also provides an overview of the range of database applications that are currently in use within organizations—personal, two-tier, multitier, and enterprise applications. The explanation of enterprise databases includes databases that are part of enterprise resource planning systems and data warehouses. The chapter concludes with a description of the process of developing a database in a fictitious company, Pine Valley Furniture. This description closely mirrors the steps in database development described earlier in the chapter.

PART II: DATABASE ANALYSIS

Chapter 2: Modeling Data in the Organization

This chapter presents a thorough introduction to conceptual data modeling with the entity-relationship (E-R) model. The chapter title emphasizes the reason for the entity-relationship model: to unambiguously document the rules of the business that influence database design. New material on why data modeling is important helps set the stage for the rest of the discussion that follows. Specific subsections explain in detail how to name and define elements of a data model, which are essential in

developing an unambiguous E-R diagram. The chapter continues to proceed from simple to more complex examples, and it concludes with a comprehensive E-R diagram for the Pine Valley Furniture Company. In the 12th edition, we have provided three new problems and exercises, and the second part of the new FAME case is introduced. Appendix A provides information on different data modeling tools and notations.

Chapter 3: The Enhanced E-R Model

This chapter presents a discussion of several advanced E-R data model constructs, primarily supertype/subtype relationships. As in Chapter 2, problems and exercises have been revised. The third part of the new FAME case is presented in this chapter. The chapter continues to present thorough coverage of supertype/subtype relationships and includes a comprehensive example of an extended E-R data model for the Pine Valley Furniture Company.

PART III: DATABASE DESIGN

Chapter 4: Logical Database Design and the Relational Model

This chapter describes the process of converting a conceptual data model to the relational data model, as well as how to merge new relations into an existing normalized database. It provides a conceptually sound and practically relevant introduction to normalization, emphasizing the importance of the use of functional dependencies and determinants as the basis for normalization. Concepts of normalization and normal forms are extended in Appendix B. The chapter features a discussion of the characteristics of foreign keys and introduces the important concept of a nonintelligent enterprise key. Enterprise keys (also called surrogate keys for data warehouses) are emphasized as some concepts of object orientation have migrated into the relational technology world. Eight new review questions and problems and exercises are included, and the revision has further clarified the coverage of some of the key concepts and the visual quality of the presentation. The chapter continues to emphasize the basic concepts of the relational data model and the role of the database designer in the logical design process. The new FAME case continues in this chapter.

Chapter 5: Physical Database Design and Performance

This chapter describes the steps that are essential in achieving an efficient database design, with a strong focus on those aspects of database design and implementation that are typically within the control of a database professional in a modern database environment. Five new review questions and problems and exercises are included. In addition, the language of the chapter was streamlined to improve readability. References to Oracle (including the visual coverage of database terminology) were updated to cover the latest version (at the time of this writing), 12c. New coverage of heap file organization was added to the chapter. The chapter contains an emphasis on ways to improve database performance, with references to specific techniques available in Oracle and other DBMSs to improve database processing performance. The discussion of indexes includes descriptions of the types of indexes (primary and secondary indexes, join index, hash index table) that are widely available in database technologies as techniques to improve query processing speed. Appendix C provides excellent background on fundamental data structures for programs of study that need coverage of this topic. The chapter continues to emphasize the physical design process and the goals of that process. The new FAME case continues with questions related to the material covered in this chapter.

PART IV: IMPLEMENTATION

Chapter 6: Introduction to SQL

This chapter presents a thorough introduction to the SQL used by most DBMSs (SQL:1999) and introduces the changes that are included in the latest standard (SQL:2011). This edition adds coverage of the new features of SQL:2011. The coverage of SQL is extensive

and divided into this and the next chapter. This chapter includes examples of SQL code, using mostly SQL:1999 and SQL:2011 syntax, as well as some Oracle 12c and Microsoft SQL Server syntax. Some unique features of MySQL are mentioned. Both dynamic and materialized views are also covered. This revision links Chapter 6 explicitly with the material covered in the new Chapter 11 on big data and analytics. Chapter 6 explains the SQL commands needed to create and maintain a database and to program single-table queries. The revised version of the chapter provides the reader with improved guidance regarding alternate sequences for learning the material. Coverage of dual-table, IS NULL/IS NOT NULL, more built-in functions, derived tables, and rules for aggregate functions and the GROUP BY clause is included or improved. Three review questions and eight problems and exercises have been added to the chapter. The chapter continues to use the Pine Valley Furniture Company case to illustrate a wide variety of practical queries and query results. Questions related to the new FAME case also are available in the context of this chapter.

Chapter 7: Advanced SQL

This chapter continues the description of SQL, with a careful explanation of multiple-table queries, transaction integrity, data dictionaries, triggers and stored procedures (the differences between them are now more clearly explained), and embedded SQL in other programming language programs. All forms of the OUTER JOIN command are covered. Standard SQL (with an updated focus on SQL:2011) is also used. The revised version of the chapter includes a new section on the temporal features introduced in SQL:2011. This chapter illustrates how to store the results of a query in a derived table, the CAST command to convert data between different data types, and the CASE command for doing conditional processing in SQL. Emphasis continues on the set-processing style of SQL compared with the record processing of programming languages with which the student may be familiar. The section on routines has been revised to provide clarified, expanded, and more current coverage of this topic. New and updated problems and exercises have been added to the chapter. The chapter continues to contain a clear explanation of subqueries and correlated subqueries, two of the most complex and powerful constructs in SQL. This chapter also includes relevant FAME case questions.

Chapter 8: Database Application Development

This chapter provides a modern discussion of the concepts of client/server architecture and applications, middleware, and database access in contemporary database environments. The section has been revised to ensure that the applicability of the concepts presented in the chapter is clear in the era of modern devices such as smartphones, tablets, etc. Review questions and problems and exercises have been updated. The chapter focuses on technologies that are commonly used to create two- and three-tier applications. Many figures are included to show the options in multitiered networks, including application and database servers, database processing distribution alternatives among network tiers, and browser (thin) clients. The chapter also presents sample application programs that demonstrate how to access databases from popular programming languages such as Java, VB.NET, ASP.NET, JSP, and PHP. This chapter lays the technology groundwork for the Internet topics presented in the remainder of the text and highlights some of the key considerations in creating three-tier Internet-based applications. The chapter also provides coverage of the role of Extensible Markup Language (XML) and related technologies in data storage and retrieval. Topics covered include basics of XML schemas, XQuery, and XSLT. The chapter concludes with an overview of Web services; associated standards and technologies; and their role in seamless, secure movement of data in Web-based applications. A brief introduction to service-oriented architecture (SOA) is also presented. Security topics, including Web security, are covered in Chapter 12. This chapter includes the final questions related to the new FAME case.

Chapter 9: Data Warehousing

This chapter describes the basic concepts of data warehousing, the reasons data warehousing is regarded as critical to competitive advantage in many organizations, and the database design activities and structures unique to data warehousing. A new section on

the future of data warehousing provides a preview of the topics that will be covered in the new chapter (Chapter 11) on big data and analytics and serves as the link between these two chapters. Some of the material that previously belonged to this chapter is now covered in an expanded fashion in Chapter 11. Topics covered in this chapter include alternative data warehouse architectures and the dimensional data model (or star schema) for data warehouses. Coverage of architectures has been streamlined consistent with trends in data warehousing, and a deep explanation of how to handle slowly changing dimensional data is provided. Operational data store and independent, dependent, and logical data marts are defined.

PART V: ADVANCED DATABASE TOPICS

Chapter 10: Data Quality and Integration

In this chapter, the principles of data governance, which are at the core of enterprise data management (EDM) activities, are introduced. This is followed by coverage of data quality. This chapter describes the need for an active program to manage data quality in organizations and outlines the steps that are considered today to be best practices for data quality management. Quality data are defined, and reasons for poor-quality data are identified. Methods for data quality improvement, such as data auditing, improving data capturing (a key part of database design), data stewardship and governance, TQM principles, modern data management technologies, and high-quality data models are all discussed. The topic of master data management, one approach to integrating key business data, is introduced and explained. Different approaches to data integration are overviewed, and the reasons for each are outlined. The extract, transform, load (ETL) process for data warehousing is discussed in detail.

Chapter 11: Big Data and Analytics

Chapter 11 on big data and analytics is new in this edition, and it extends the coverage of the text in three important ways: First, this chapter provides a systematic introduction to the technologies that are currently discussed under the label *big data* and the impact of these technologies on the overall enterprise data management architecture. Specifically, the chapter focuses on the Hadoop infrastructure and four categories of so-called NoSQL (Not only SQL) database management systems. Second, the chapter offers integrated coverage of analytics, including descriptive, predictive, and prescriptive analytics. The discussion on analytics is linked not only to the coverage of big data but also the material on data warehousing in Chapter 9 and the general discussion on data management in Chapter 1. The chapter also briefly covers approaches and technologies used by analytics professionals, such as OLAP, data visualization, business performance management and dashboards, data mining, and text mining. Third, the chapter integrates the coverage of big data and analytics technologies to the individual, organizational, and societal implications of these capabilities.

Chapter 12: Data and Database Administration

This chapter presents a thorough discussion of the importance and roles of data and database administration and describes a number of the key issues that arise when these functions are performed. This chapter emphasizes the changing roles and approaches of data and database administration, with emphasis on data quality and high performance. We also briefly touch upon the impact of cloud computing on the data/database administration. The chapter contains a thorough discussion of database backup procedures, as well as extensively expanded and consolidated coverage of data security threats and responses and data availability. The data security topics include database security policies, procedures, and technologies (including encryption and smart cards). The role of databases in Sarbanes-Oxley compliance is also examined. We also discuss open source DBMS, the benefits and hazards of this technology, and how to choose an open source DBMS. In addition, the topic of heartbeat queries is included in the coverage of database performance improvements. The chapter continues to emphasize the critical importance of data and database management in managing data as a corporate asset.

Chapter 13: Distributed Databases

This chapter reviews the role, technologies, and unique database design opportunities of distributed databases. The objectives and trade-offs for distributed databases, data replication alternatives, factors in selecting a data distribution strategy, and distributed database vendors and products are covered. This chapter provides thorough coverage of database concurrency access controls. The chapter introduces several technical updates that are related to the significant advancements in both data management and networking technologies, which form the context for a distributed database. The full version of this chapter is available on the textbook's Web site. Many reviewers indicated that they are seldom able to cover this chapter in an introductory course, but having the material available is critical for advanced students or special topics.

Chapter 14: Object-Oriented Data Modeling

This chapter presents an introduction to object-oriented modeling using Object Management Group's Unified Modeling Language (UML). This chapter has been carefully reviewed to ensure consistency with the latest UML notation and best industry practices. UML provides an industry-standard notation for representing classes and objects. The chapter continues to emphasize basic object-oriented concepts, such as inheritance, encapsulation, composition, and polymorphism. The revised version of the chapter also includes brand-new review questions and modeling exercises. As with Chapter 13, Chapter 14 is available on the textbook's Web site.

APPENDICES

In the 12th edition three appendices are available on the Web and are intended for those who wish to explore certain topics in greater depth.

Appendix A: Data Modeling Tools and Notation

This appendix addresses a need raised by many readers—how to translate the E-R notation in the text into the form used by the CASE tool or the DBMS used in class. Specifically, this appendix compares the notations of CA ERwin Data Modeler r9.5, Oracle SQL Data Modeler 4.0, SAP Sybase PowerDesigner 16.5, and Microsoft Visio Professional 2013. Tables and illustrations show the notations used for the same constructs in each of these popular software packages.

Appendix B: Advanced Normal Forms

This appendix presents a description (with examples) of Boyce-Codd and fourth normal forms, including an example of BCNF to show how to handle overlapping candidate keys. Other normal forms are briefly introduced. The Web Resources section includes a reference for information on many advanced normal form topics.

Appendix C: Data Structures

This appendix describes several data structures that often underlie database implementations. Topics include the use of pointers, stacks, queues, sorted lists, inverted lists, and trees.

PEDAGOGY

A number of additions and improvements have been made to end-of-chapter materials to provide a wider and richer range of choices for the user. The most important of these improvements are the following:

1. *Review Questions* Questions have been updated to support new and enhanced chapter material.
2. *Problems and Exercises* This section has been reviewed in every chapter, and many chapters contain new problems and exercises to support updated chapter material.

Of special interest are questions in many chapters that give students opportunities to use the data sets provided for the text. Also, Problems and Exercises have been re-sequenced into roughly increasing order of difficulty, which should help instructors and students find exercises appropriate for what they want to accomplish.

3. **Field Exercises** This section provides a set of “hands-on” mini cases that can be assigned to individual students or to small teams of students. Field exercises range from directed field trips to Internet searches and other types of research exercises.
4. **Case** The 12th edition of this book has a brand new mini case: Forondo Artist Management Excellence Inc. (FAME). In the first three chapters, the case begins with a description provided in the “voice” of one or more stakeholders, revealing a new dimension of requirements to the reader. Each chapter has project assignments intended to provide guidance on the types of deliverables instructors could expect from students, some of which tie together issues and activities across chapters. These project assignments can be completed by individual students or by small project teams. This case provides an excellent means for students to gain hands-on experience with the concepts and tools they have studied.
5. **Web Resources** Each chapter contains a list of updated and validated URLs for Web sites that contain information that supplements the chapter. These Web sites cover online publication archives, vendors, electronic publications, industry standards organizations, and many other sources. These sites allow students and instructors to find updated product information, innovations that have appeared since the printing of the book, background information to explore topics in greater depth, and resources for writing research papers.

We continue to provide several pedagogical features that help make the 12th edition widely accessible to instructors and students. These features include the following:

1. **Learning objectives** appear at the beginning of each chapter, as a preview of the major concepts and skills students will learn from that chapter. The learning objectives also provide a great study review aid for students as they prepare for assignments and examinations.
2. **Chapter introductions and summaries** both encapsulate the main concepts of each chapter and link material to related chapters, providing students with a comprehensive conceptual framework for the course.
3. **The chapter review** includes the Review Questions, Problems and Exercises, and Field Exercises discussed earlier and also contains a Key Terms list to test the student’s grasp of important concepts, basic facts, and significant issues.
4. **A running glossary** defines key terms in the page margins as they are discussed in the text. These terms are also defined at the end of the text, in the Glossary of Terms. Also included is the end-of-book Glossary of Acronyms for abbreviations commonly used in database management.

ORGANIZATION

We encourage instructors to customize their use of this book to meet the needs of both their curriculum and student career paths. The modular nature of the text, its broad coverage, extensive illustrations, and its inclusion of advanced topics and emerging issues make customization easy. The many references to current publications and Web sites can help instructors develop supplemental reading lists or expand classroom discussion beyond material presented in the text. The use of appendices for several advanced topics allows instructors to easily include or omit these topics.

The modular nature of the text allows the instructor to omit certain chapters or to cover chapters in a different sequence. For example, an instructor who wishes to emphasize data modeling may cover Chapter 14 (available on the book’s Web site) on object-oriented data modeling along with or instead of Chapters 2 and 3. An instructor who wishes to cover only basic entity-relationship concepts (but not the enhanced E-R model) may skip Chapter 3 or cover it after Chapter 4 on the relational model.

We have contacted many adopters of *Modern Database Management* and asked them to share with us their syllabi. Most adopters cover the chapters in sequence, but several alternative sequences have also been successful. These alternatives include the following:

- Some instructors cover Chapter 12 on data and database administration immediately after Chapter 5 on physical database design and the relational model.
- To cover SQL as early as possible, instructors have effectively covered Chapters 6 and 7 immediately after Chapter 4; some have even covered Chapter 6 immediately after Chapter 1.
- Many instructors have students read appendices along with chapters, such as reading Appendix on data modeling notations with Chapter 2 or Chapter 3 on E-R modeling, Appendix B on advanced normal forms with Chapter 4 on the relational model, and Appendix C on data structures with Chapter 5.

THE SUPPLEMENT PACKAGE: WWW.PEARSONHIGHERED.COM/HOFFER

A comprehensive and flexible technology support package is available to enhance the teaching and learning experience. All instructor and student supplements are available on the text Web site: www.pearsonhighered.com/hoffer.

For Students

The following online resources are available to students:

- *Complete chapters on distributed databases and object-oriented data modeling* as well as appendices focusing on data modeling notations, advanced normal forms, and data structures allow you to learn in depth about topics that are not covered in the textbook.
- *Accompanying databases* are also provided. Two versions of the Pine Valley Furniture Company case have been created and populated for the 12th edition. One version is scoped to match the textbook examples. A second version is fleshed out with more data and tables. This version is not complete, however, so that students can create missing tables and additional forms, reports, and modules. Databases are provided in several formats (ASCII tables, Oracle script, and Microsoft Access), but formats vary for the two versions. Some documentation of the databases is also provided. Both versions of the PVFC database are also provided on Teradata University Network.
- *Several custom-developed short videos that address key concepts and skills from different sections of the book* help students learn material that may be more difficult to understand by using both the printed text and a mini lecture.



For Instructors

The following online resources are available to instructors:

- The *Instructor's Resource Manual* by Heikki Topi, Bentley University, provides chapter-by-chapter instructor objectives, classroom ideas, and answers to Review Questions, Problems and Exercises, Field Exercises, and Project Case Questions. The Instructor's Resource Manual is available for download on the instructor area of the text's Web site.
- The *Test Item File and TestGen*, by Bob Mills, Utah State University, includes a comprehensive set of test questions in multiple-choice, true/false, and short-answer format, ranked according to level of difficulty and referenced with page numbers and topic headings from the text. The Test Item File is available in Microsoft Word and as the computerized TestGen. TestGen is a comprehensive suite of tools for testing and assessment. It allows instructors to easily create and distribute tests for their courses, either by printing and distributing through traditional methods or by online delivery via a local area network (LAN) server. Test Manager features Screen Wizards to assist you as you move through the program, and the software is backed with full technical support.

- *PowerPoint presentation slides*, by Michel Mitri, James Madison University, feature lecture notes that highlight key terms and concepts. Instructors can customize the presentation by adding their own slides or editing existing ones.
- The *Image Library* is a collection of the text art organized by chapter. It includes all figures, tables, and screenshots (as permission allows) and can be used to enhance class lectures and PowerPoint slides.
- *Accompanying databases* are also provided. Two versions of the Pine Valley Furniture Company case have been created and populated for the 12th edition. One version is scoped to match the textbook examples. A second version is fleshed out with more data and tables. This version is not complete, however, so that students can create missing tables and additional forms, reports, and modules. Databases are provided in several formats (ASCII tables, Oracle script, and Microsoft Access), but formats vary for the two versions. Some documentation of the databases is also provided. Both versions of the PVFC database are also available on Teradata University Network.

COURSESMART eTEXTBOOK

CourseSmart eTextbooks were developed for students looking to save on required or recommended textbooks. Students simply select their eText by title or author and purchase immediate access to the content for the duration of the course using any major credit card. With a CourseSmart eText, students can search for specific keywords or page numbers, take notes online, print out reading assignments that incorporate lecture notes, and bookmark important passages for later review. For more information or to purchase a CourseSmart eTextbook, visit www.coursesmart.com

ACKNOWLEDGMENTS

We are grateful to numerous individuals who contributed to the preparation of *Modern Database Management*, 12th edition. First, we wish to thank our reviewers for their detailed suggestions and insights, characteristic of their thoughtful teaching style. As always, analysis of topics and depth of coverage provided by the reviewers were crucial. Our reviewers and others who gave us many useful comments to improve the text include Tamara Babaian, Bentley University; Gary Baram, Temple University; Bijoy Bordoloi, Southern Illinois University, Edwardsville; Timothy Bridges, University of Central Oklahoma; Traci Carte, University of Oklahoma; Wingyan Chung, Santa Clara University; Jagdish Gangolly, State University of New York at Albany; Jon Gant, Syracuse University; Jinzhu Gao, University of the Pacific; Monica Garfield, Bentley University; Rick Gibson, American University; Chengqi Guo, James Madison University; William H. Hochstettler III, Franklin University; Weiling Ke, Clarkson University; Dongwon Lee, Pennsylvania State University; Ingyu Lee, Troy University; Chang-Yang Lin, Eastern Kentucky University; Brian Mennecke, Iowa State University; Dat-Dao Nguyen, California State University, Northridge; Fred Niederman, Saint Louis University; Lara Preiser-Houy, California State Polytechnic University, Pomona; John Russo, Wentworth Institute of Technology; Ioulia Rytikova, George Mason University; Richard Segall, Arkansas State University; Chelley Vician, University of St. Thomas; and Daniel S. Weaver, Messiah College.

We received excellent input from people in industry, including Todd Walter, Carrie Ballinger, Rob Armstrong, and David Schoeff (all of Teradata Corp); Chad Gronbach and Philip DesAutels (Microsoft Corp.); Peter Gauvin (Ball Aerospace); Paul Longhurst (Overstock.com); Derek Strauss (Gavroshe International); Richard Hackathorn (Bolder Technology); and Michael Alexander (Open Access Technology, International).

We have special admiration for and gratitude to Heikki Topi, Bentley University, author of the *Instructor's Resource Manual*. In addition to his duties as author, Heikki took on this additional task and has been extremely careful in preparing the *Instructor's Resource Manual*; in the process he has helped us clarify and fix various parts of the text. We also want to recognize the important role played by Chelley Vician of the University of St. Thomas, the author of several previous editions of the *Instructor's Resource Manual*; her

work added great value to this book. We also thank Sven Aelterman, Troy University, for his many excellent suggestions for improvements and clarifications throughout the text.

We are also grateful to the staff and associates of Pearson for their support and guidance throughout this project. In particular, we wish to thank retired Executive Editor Bob Horan for his support through many editions of this text book, Project Manager Ilene Kahn, who kept us on track and made sure everything was complete; Acquisitions Editor Nicole Sam; Program Manager Denise Weiss; and Editorial Assistant Olivia Vignone. We extend special thanks to George Jacob at Integra, whose supervision of the production process was excellent.

While finalizing this edition of *MDBM*, we learned that one of the co-authors of previous editions, Dr. Mary Prescott, passed away after many years of battling and beating, but ultimately losing one last round of cancer. Mary was a co-author for the 5th through 9th editions, and she was integrally involved as a reviewer and contributor of ideas and teaching methods for several prior editions. Dr. Prescott was a dedicated and inspiring member of the author team and co-author of many of the innovations in *MDBM*, including the first material on data warehousing and significant updates of the coverage of SQL and data administration. Mary was an outstanding educator and academic administrator at the University of South Florida and University of Tampa. She was also involved in leadership development for Florida Polytechnic University (USF Polytechnic). Mary's multiple talents, built on an academic background of BA and MA in Psychology, MBA with a concentration in Accounting, and PhD in MIS, as well as a crisp writing style, contributed greatly to her significant value to this text. Mary's contributions to *MDBM*, both content and spirit, will be directly and indirectly included in this book for many years to come.

Finally, we give immeasurable thanks to our spouses, who endured many evenings and weekends of solitude for the thrill of seeing a book cover hang on a den wall. In particular, we marvel at the commitment of Patty Hoffer, who has lived the lonely life of a textbook author's spouse through 12 editions over more than 30 years of late-night and weekend writing. We also want to sincerely thank Anne-Louise Klaus for being willing to continue her wholehearted support for Heikki's involvement in the project. Although the book project was no longer new for Gayathri Mani, her continued support and understanding are very much appreciated. Much of the value of this text is due to their patience, encouragement, and love, but we alone bear the responsibility for any errors or omissions between the covers.

Jeffrey A. Hoffer

V. Ramesh

Heikki Topi

This page intentionally left blank

PART I

The Context of Database Management

AN OVERVIEW OF PART ONE

In this chapter and opening part of the book, we set the context and provide basic database concepts and definitions used throughout the text. In this part, we portray database management as an exciting, challenging, and growing field that provides numerous career opportunities for information systems students. Databases continue to become a more common part of everyday living and a more central component of business operations. From the database that stores contact information in your smartphone or tablet to the very large databases that support enterprise-wide information systems, databases have become the central points of data storage that were envisioned decades ago. Customer relationship management and Internet shopping are examples of two database-dependent activities that have developed in recent years. The development of data warehouses and “big data” repositories that provide managers the opportunity for deeper and broader historical analysis of data also continues to take on more importance.

We begin by providing basic definitions of *data*, *database*, *metadata*, *database management system*, *data warehouse*, and other terms associated with this environment. We compare databases with the older file management systems they replaced and describe several important advantages that are enabled by the carefully planned use of databases.

The chapter also describes the general steps followed in the analysis, design, implementation, and administration of databases. Further, this chapter also illustrates how the database development process fits into the overall information systems development process. Database development for both structured life cycle and prototyping methodologies is explained. We introduce enterprise data modeling, which sets the range and general contents of organizational databases. This is often the first step in database development. We introduce the concept of schemas and the three-schema architecture, which is the dominant approach in modern database systems. We describe the major components of the database environment and the types of applications, as well as multitier and enterprise databases. Enterprise databases include those that are used to support enterprise resource planning systems and data warehouses. Finally, we describe the roles of the various people who are typically involved in a database development project. The Pine Valley Furniture Company case is introduced and used to illustrate many of the principles and concepts of database management. This case is used throughout the text as a continuing example of the use of database management systems.

Chapter 1

The Database Environment and Development Process

The Database Environment and Development Process



Visit www.pearsonhighered.com/ to view the accompanying video for this chapter.

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Concisely define each of the following key terms: **data, database, database management system, data model, information, metadata, enterprise data model, entity, relational database, enterprise resource planning (ERP) system, database application, data warehouse, data independence, repository, user view, enterprise data modeling, systems development life cycle (SDLC), prototyping, agile software development, data modeling and design tools, conceptual schema, logical schema, and physical schema.**
- Name several limitations of conventional file processing systems.
- Explain at least 10 advantages of the database approach, compared to traditional file processing.
- Identify several costs and risks of the database approach.
- List and briefly describe nine components of a typical database environment.
- Identify four categories of applications that use databases and their key characteristics.
- Describe the life cycle of a systems development project, with an emphasis on the purpose of database analysis, design, and implementation activities.
- Explain the prototyping and agile-development approaches to database and application development.
- Explain the roles of individuals who design, implement, use, and administer databases.
- Explain the differences among external, conceptual, and internal schemas and the reasons for the three-schema architecture for databases.

DATA MATTER!

The amount of data being generated, stored, and processed is growing by leaps and bounds. According to a McKinsey Global Institute Report (Manyika et al., 2011), it is estimated that in 2010 alone global enterprises stored more than 7 exabytes of data (an exabyte is a billion gigabytes) while consumers stored more than 6 exabytes of new data on devices such as PCs, smartphones, tablets, and notebooks. That is a lot of data! And as more and more of the world becomes digital and products we use every day such as watches, refrigerators, and such become smarter, the amount of data that needs to be generated, stored, and processed will only continue to grow.

The availability of all of this data is also opening up unparalleled opportunities for companies to leverage it for various purposes. A recent study by IBM (IBM, 2011) shows that one of the top priorities for CEOs in the coming years is the ability to use insights and intelligence that can be gleaned from data for competitive advantage. The McKinsey Global Institute Report (Manyika et al., 2011) estimates that by appropriately leveraging the data available to them, U.S. retail industry can see up to a 60 percent increase in net margin and manufacturing can realize up to a 50 percent reduction in product development costs.

The availability of large amounts of data is also fueling innovation in companies and allowing them to think differently and creatively about various aspects of their businesses. Below we provide some examples from a variety of domains:

1. The Memorial Sloan-Kettering Cancer center is using IBM Watson (do you remember Watson beating Ken Jennings in Jeopardy?) to help analyze the information from medical literature, research, past case histories, and best practices to help provide oncologists with evidence-based recommendations (http://www-935.ibm.com/services/multimedia/MSK_Case_Study_IMC14794.pdf).
2. Continental Airlines (now United) invested in a real-time business intelligence capability and was able to dramatically improve its customer service and operations. For example, it can now track if a high-value customer is experiencing a delay in a trip, where and when the customer will arrive at the airport, and the gate the customer must go to make the next connection (Anderson-Lehman, et al., 2004).
3. A leading fast food chain uses video information from its fast food lane to determine what food products to display on its (digital) menu board. If the lines are long, the menu displays items that can be served quickly. If the lines are short, the menu displays higher margin but slower to prepare items (Laskowski, 2013).
4. Nagoya Railroad analyzes data about its customers' travel habits along with their shopping and dining habits to better understand its customers. For example, it was able to identify that young women who used a particular train station for their commute also tended to eat at a particular type of restaurant and buy from certain types of stores. This information allows Nagoya Railroad to create a targeted marketing campaign (<http://public.dhe.ibm.com/common/ssi/ecm/en/ytc03707usen/YTC03707USEN.PDF>).

At the heart of all the above examples is the ability to collect, organize, and manage data. This is precisely the focus of this textbook. This understanding will give you the power to support any business strategy and the deep satisfaction that comes from knowing how to organize data so that financial, marketing, or customer service questions can be answered almost as soon as they are asked. Enjoy!

INTRODUCTION

Over the past two decades, data has become a strategic asset for most organizations. Databases are used to store, manipulate, and retrieve data in nearly every type of organization, including business, health care, education, government, and libraries. Database technology is routinely used by individuals on personal computers and by employees using enterprise-wide distributed applications. Databases are also accessed by customers and other remote users through diverse technologies, such as automated teller machines, Web browsers, smartphones, and intelligent living and office environments. Most Web-based applications depend on a database foundation.

Following this period of rapid growth, will the demand for databases and database technology level off? Very likely not! In the highly competitive environment of today, there is every indication that database technology will assume even greater importance. Managers seek to use knowledge derived from databases for competitive advantage. For example, detailed sales databases can be mined to determine customer buying patterns as a basis for advertising and marketing campaigns. Organizations embed procedures called *alerts* in databases

to warn of unusual conditions, such as impending stock shortages or opportunities to sell additional products, and to trigger appropriate actions.

Although the future of databases is assured, much work remains to be done. Many organizations have a proliferation of incompatible databases that were developed to meet immediate needs rather than based on a planned strategy or a well-managed evolution. Enormous amounts of data are trapped in older, “legacy” systems, and the data are often of poor quality. New skills are required to design and manage data warehouses and other repositories of data and to fully leverage all the data that is being captured in the organization. There is a shortage of skills in areas such as database analysis, database design, database application development, and business analytics. We address these and other important issues in this textbook to equip you for the jobs of the future.

A course in database management has emerged as one of the most important courses in the information systems curriculum today. Further, many schools have added an additional elective course in data warehousing and/or business analytics to provide in-depth coverage of these important topics. As information systems professionals, you must be prepared to analyze database requirements and design and implement databases within the context of information systems development. You also must be prepared to consult with end users and show them how they can use databases (or data warehouses) to build decision models and systems for competitive advantage. And, the widespread use of databases attached to Web sites that return dynamic information to users of these sites requires that you understand not only how to link databases to the Web-based applications but also how to secure those databases so that their contents can be viewed but not compromised by outside users.

In this chapter, we introduce the basic concepts of databases and database management systems (DBMSs). We describe traditional file management systems and some of their shortcomings that led to the database approach. Next, we consider the benefits, costs, and risks of using the database approach. We review the range of technologies used to build, use, and manage databases; describe the types of applications that use databases—personal, multitier, and enterprise; and describe how databases have evolved over the past five decades.

Because a database is one part of an information system, this chapter also examines how the database development process fits into the overall information systems development process. The chapter emphasizes the need to coordinate database development with all the other activities in the development of a complete information system. It includes highlights from a hypothetical database development process at Pine Valley Furniture Company. Using this example, the chapter introduces tools for developing databases on personal computers and the process of extracting data from enterprise databases for use in stand-alone applications.

There are several reasons for discussing database development at this point. First, although you may have used the basic capabilities of a database management system, such as Microsoft Access, you may not yet have developed an understanding of how these databases were developed. Using simple examples, this chapter briefly illustrates what you will be able to do after you complete a database course using this text. Thus, this chapter helps you develop a vision and context for each topic developed in detail in subsequent chapters.

Second, many students learn best from a text full of concrete examples. Although all of the chapters in this text contain numerous examples, illustrations, and actual database designs and code, each chapter concentrates on a specific aspect of database management. We have designed this chapter to help you understand, with minimal technical details, how all of these individual aspects of database management are related and how database development tasks and skills relate to what you are learning in other information systems courses.

Finally, many instructors want you to begin the initial steps of a database development group or individual project early in your database course. This chapter gives you an idea of how to structure a database development project sufficient to begin a course exercise. Obviously, because this is only the first chapter, many of

the examples and notations we will use will be much simpler than those required for your project, for other course assignments, or in a real organization.

One note of caution: You will not learn how to design or develop databases just from this chapter. Sorry! We have purposely kept the content of this chapter introductory and simplified. Many of the notations used in this chapter are not exactly like the ones you will learn in subsequent chapters. Our purpose in this chapter is to give you a general understanding of the key steps and types of skills, not to teach you specific techniques. You will, however, learn fundamental concepts and definitions and develop an intuition and motivation for the skills and knowledge presented in later chapters.

BASIC CONCEPTS AND DEFINITIONS

We define a **database** as an organized collection of logically related data. Not many words in the definition, but have you looked at the size of this book? There is a lot to do to fulfill this definition.

A database may be of any size and complexity. For example, a salesperson may maintain a small database of customer contacts—consisting of a few megabytes of data—on her laptop computer. A large corporation may build a large database consisting of several terabytes of data (a *terabyte* is a trillion bytes) on a large mainframe computer that is used for decision support applications (Winter, 1997). Very large data warehouses contain more than a petabyte of data. (A *petabyte* is a quadrillion bytes.) (We assume throughout the text that all databases are computer based.)

Data

Historically, the term *data* referred to facts concerning objects and events that could be recorded and stored on computer media. For example, in a salesperson's database, the data would include facts such as customer name, address, and telephone number. This type of data is called *structured* data. The most important structured data types are numeric, character, and dates. Structured data are stored in tabular form (in tables, relations, arrays, spreadsheets, etc.) and are most commonly found in traditional databases and data warehouses.

The traditional definition of data now needs to be expanded to reflect a new reality: Databases today are used to store objects such as documents, e-mails, tweets, Facebook posts, GPS information, maps, photographic images, sound, and video segments in addition to structured data. For example, the salesperson's database might include a photo image of the customer contact. It might also include a sound recording or video clip about the most recent product. This type of data is referred to as *unstructured* data, or as multimedia data. Today structured and unstructured data are often combined in the same database to create a true multimedia environment. For example, an automobile repair shop can combine structured data (describing customers and automobiles) with multimedia data (photo images of the damaged autos and scanned images of insurance claim forms).

An expanded definition of **data** that includes structured and unstructured types is "a stored representation of objects and events that have meaning and importance in the user's environment."

Data Versus Information

The terms *data* and *information* are closely related and in fact are often used interchangeably. However, it is useful to distinguish between data and information. We define **information** as data that have been processed in such a way that the knowledge of the person who uses the data is increased. For example, consider the following list of facts:

Baker, Kenneth D.	324917628
Doyle, Joan E.	476193248
Finkle, Clive R.	548429344
Lewis, John C.	551742186
McFerran, Debra R.	409723145

Database

An organized collection of logically related data.

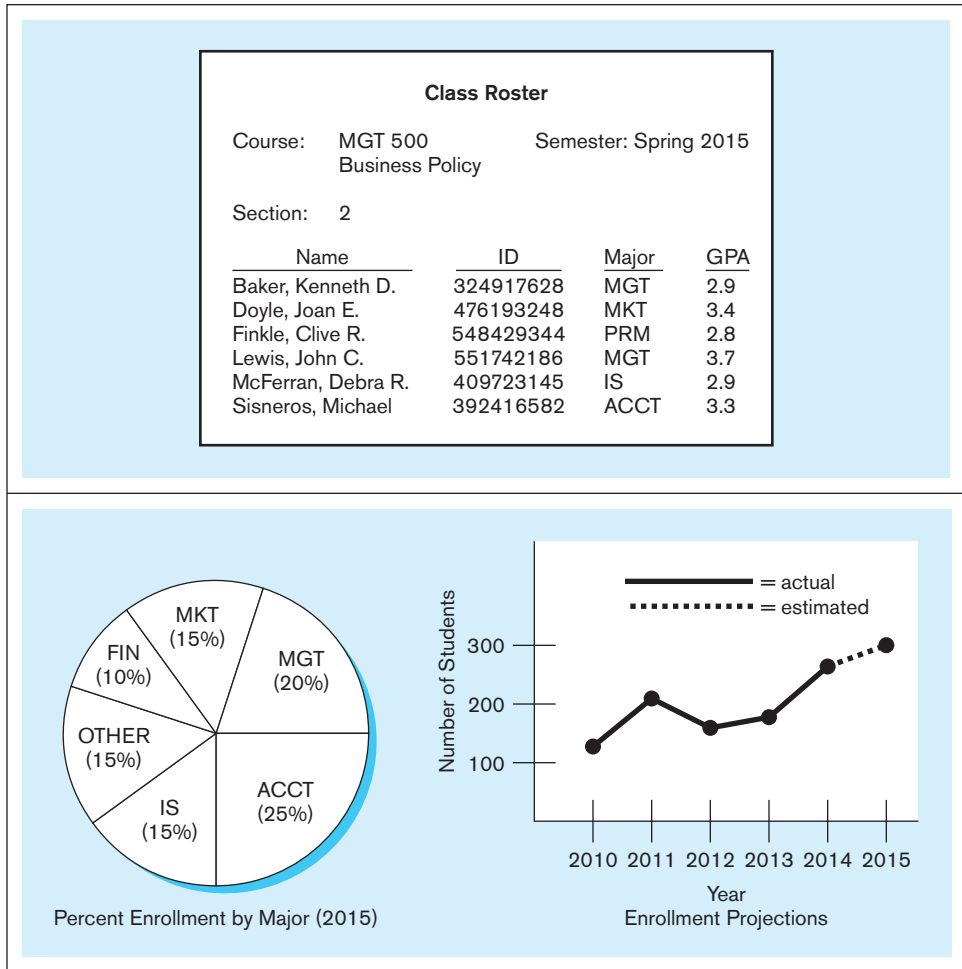
Data

Stored representations of objects and events that have meaning and importance in the user's environment.

Information

Data that have been processed in such a way as to increase the knowledge of the person who uses the data.

FIGURE 1-1 Converting data to information
(a) Data in context



These facts satisfy our definition of data, but most people would agree that the data are useless in their present form. Even if we guess that this is a list of people's names paired with their Social Security numbers, the data remain useless because we have no idea what the entries mean. Notice what happens when we place the same data in a context, as shown in Figure 1-1a.

By adding a few additional data items and providing some structure, we recognize a class roster for a particular course. This is useful information to some users, such as the course instructor and the registrar's office. Of course, as general awareness of the importance of strong data security has increased, few organizations still use Social Security numbers as identifiers. Instead, most organizations use an internally generated number for identification purposes.

Another way to convert data into information is to summarize them or otherwise process and present them for human interpretation. For example, Figure 1-1b shows summarized student enrollment data presented as graphical information. This information could be used as a basis for deciding whether to add new courses or to hire new faculty members.

In practice, according to our definitions, databases today may contain either data or information (or both). For example, a database may contain an image of the class roster document shown in Figure 1-1a. Also, data are often preprocessed and stored in summarized form in databases that are used for decision support. Throughout this text we use the term *database* without distinguishing its contents as data or information.

Metadata

Data that describe the properties or characteristics of end-user data and the context of those data.

Metadata

As we have indicated, data become useful only when placed in some context. The primary mechanism for providing context for data is metadata. **Metadata** are data

TABLE 1-1 Example Metadata for Class Roster

Data Item		Metadata				
Name	Type	Length	Min	Max	Description	Source
Course	Alphanumeric	30			Course ID and name	Academic Unit
Section	Integer	1	1	9	Section number	Registrar
Semester	Alphanumeric	10			Semester and year	Registrar
Name	Alphanumeric	30			Student name	Student IS
ID	Integer	9			Student ID (SSN)	Student IS
Major	Alphanumeric	4			Student major	Student IS
GPA	Decimal	3	0.0	4.0	Student grade point average	Academic Unit

that describe the properties or characteristics of end-user data and the context of that data. Some of the properties that are typically described include data names, definitions, length (or size), and allowable values. Metadata describing data context include the source of the data, where the data are stored, ownership (or stewardship), and usage. Although it may seem circular, many people think of metadata as “data about data.”

Some sample metadata for the Class Roster (Figure 1-1a) are listed in Table 1-1. For each data item that appears in the Class Roster, the metadata show the data item name, the data type, length, minimum and maximum allowable values (where appropriate), a brief description of each data item, and the source of the data (sometimes called the *system of record*). Notice the distinction between data and metadata. Metadata are once removed from data. That is, metadata describe the properties of data but are separate from that data. Thus, the metadata shown in Table 1-1 do not include any sample data from the Class Roster of Figure 1-1a. Metadata enable database designers and users to understand what data exist, what the data mean, and how to distinguish between data items that at first glance look similar. Managing metadata is at least as crucial as managing the associated data because data without clear meaning can be confusing, misinterpreted, or erroneous. Typically, much of the metadata are stored as part of the database and may be retrieved using the same approaches that are used to retrieve data or information.

Data can be stored in files (think Excel sheets) or in databases. In the following sections, we examine the progression from file processing systems to databases and the advantages and disadvantages of each.

TRADITIONAL FILE PROCESSING SYSTEMS

When computer-based data processing was first available, there were no databases. To be useful for business applications, computers had to store, manipulate, and retrieve large files of data. Computer file processing systems were developed for this purpose. Although these systems have evolved over time, their basic structure and purpose have changed little over several decades.

As business applications became more complex, it became evident that traditional file processing systems had a number of shortcomings and limitations (described next). As a result, these systems have been replaced by database processing systems in most business applications today. Nevertheless, you should have at least some familiarity with file processing systems since understanding the problems and limitations inherent in file processing systems can help you avoid these same problems when designing database systems. It should be noted that Excel files, in general, fall into the same category as file systems and suffer from the same drawbacks listed below.



File Processing Systems at Pine Valley Furniture Company

Early computer applications at Pine Valley Furniture used the traditional file processing approach. This approach to information systems design met the data processing needs of individual departments rather than the overall information needs of the organization. The information systems group typically responded to users' requests for new systems by developing (or acquiring) new computer programs for individual applications such as inventory control, accounts receivable, or human resource management. No overall map, plan, or model guided application growth.

Three of the computer applications based on the file processing approach are shown in Figure 1-2. The systems illustrated are Order Filling, Invoicing, and Payroll. The figure also shows the major data files associated with each application. A *file* is a collection of related records. For example, the Order Filling System has three files: Customer Master, Inventory Master, and Back Order. Notice that there is duplication of some of the files used by the three applications, which is typical of file processing systems.

Disadvantages of File Processing Systems

Several disadvantages associated with conventional file processing systems are listed in Table 1-2 and described briefly next. It is important to understand these issues because if we don't follow the database management practices described in this book, some of these disadvantages can also become issues for databases as well.

Database application

An application program (or set of related programs) that is used to perform a series of database activities (create, read, update, and delete) on behalf of database users.

PROGRAM-DATA DEPENDENCE File descriptions are stored within each **database application** program that accesses a given file. For example, in the Invoicing System in Figure 1-2, Program A accesses the Inventory Pricing File and the Customer Master File. Because the program contains a detailed file description for these files, any change to a file structure requires changes to the file descriptions for all programs that access the file.

Notice in Figure 1-2 that the Customer Master File is used in the Order Filling System and the Invoicing System. Suppose it is decided to change the customer address field length in the records in this file from 30 to 40 characters. The file descriptions in each program that is affected (up to five programs) would have to be modified. It is often difficult even to locate all programs affected by such changes. Worse, errors are often introduced when making such changes.

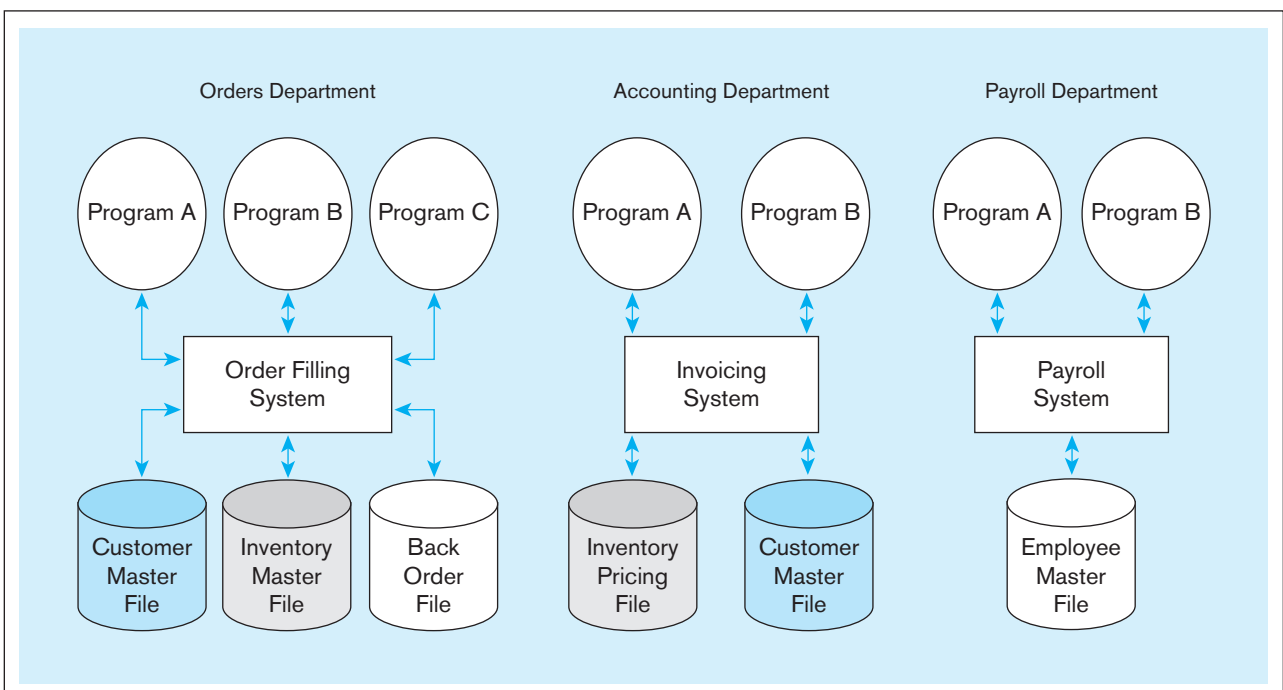


FIGURE 1-2 Old file processing systems at Pine Valley Furniture Company

DUPLICATION OF DATA Because applications are often developed independently in file processing systems, unplanned duplicate data files are the rule rather than the exception. For example, in Figure 1-2, the Order Filling System contains an Inventory Master File, whereas the Invoicing System contains an Inventory Pricing File. These files contain data describing Pine Valley Furniture Company's products, such as product description, unit price, and quantity on hand. This duplication is wasteful because it requires additional storage space and increased effort to keep all files up to date. Data formats may be inconsistent or data values may not agree (or both). Reliable metadata are very difficult to establish in file processing systems. For example, the same data item may have different names in different files or, conversely, the same name may be used for different data items in different files.

LIMITED DATA SHARING With the traditional file processing approach, each application has its own private files, and users have little opportunity to share data outside their own applications. Notice in Figure 1-2, for example, that users in the Accounting Department have access to the Invoicing System and its files, but they probably do not have access to the Order Filling System or to the Payroll System and their files. Managers often find that a requested report requires a major programming effort because data must be drawn from several incompatible files in separate systems. When different organizational units own these different files, additional management barriers must be overcome.

LENGTHY DEVELOPMENT TIMES With traditional file processing systems, each new application requires that the developer essentially start from scratch by designing new file formats and descriptions and then writing the file access logic for each new program. The lengthy development times required are inconsistent with today's fast-paced business environment, in which time to market (or time to production for an information system) is a key business success factor.

EXCESSIVE PROGRAM MAINTENANCE The preceding factors all combined to create a heavy program maintenance load in organizations that relied on traditional file processing systems. In fact, as much as 80 percent of the total information system's development budget might be devoted to program maintenance in such organizations. This in turn means that resources (time, people, and money) are not being spent on developing new applications.

It is important to note that many of the disadvantages of file processing we have mentioned can also be limitations of databases if an organization does not properly apply the database approach. For example, if an organization develops many separately managed databases (say, one for each division or business function) with little or no coordination of the metadata, then uncontrolled data duplication, limited data sharing, lengthy development time, and excessive program maintenance can occur. Thus, the database approach, which is explained in the next section, is as much a way to manage organizational data as it is a set of technologies for defining, creating, maintaining, and using these data.

THE DATABASE APPROACH

So, how do we overcome the flaws of file processing? No, we don't call Ghostbusters, but we do something better: We follow the database approach. We first begin by defining some core concepts that are fundamental in understanding the database approach to managing data. We then describe how the database approach can overcome the limitations of the file processing approach.

Data Models

Designing a database properly is fundamental to establishing a database that meets the needs of the users. **Data models** capture the nature of and relationships among data and are used at different levels of abstraction as a database is conceptualized and designed. The effectiveness and efficiency of a database is directly associated with the structure of the database. Various graphical systems exist that convey this structure and are used to

TABLE 1-2 Disadvantages of File Processing Systems

Program-data dependence
Duplication of data
Limited data sharing
Lengthy development times
Excessive program maintenance

Data model

Graphical systems used to capture the nature and relationships among data.