



relational database design and implementation

THIRD
EDITION

JAN L.
HARRINGTON

MK[®]
MORGAN KAUFMANN

TP311.13
H299
E.3

Relational Database Design and Implementation: Clearly Explained

Third Edition

Jan L. Harrington



E2009003615



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Morgan Kaufmann Publishers is an imprint of Elsevier



MORGAN KAUFMANN PUBLISHERS

AN IMPRINT OF ELSEVIER SCIENCE

Morgan Kaufmann Publishers is an imprint of Elsevier.
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

This book is printed on acid-free paper.

Copyright © 2009 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. All trademarks that appear or are otherwise referred to in this work belong to their respective owners. Neither Morgan Kaufmann Publishers nor the authors and other contributors of this work have any relationship or affiliation with such trademark owners nor do such trademark owners confirm, endorse or approve the contents of this work. Readers, however, should contact the appropriate companies for more information regarding trademarks and any related registrations.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, scanning, or otherwise—without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, E-mail: permissions@elsevier.com. You may also complete your request online via the Elsevier homepage (<http://elsevier.com>), by selecting "Support & Contact" then "Copyright and Permission" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data

Harrington, Jan L.

Relational database design and implementation : clearly explained /
Jan L. Harrington.—3rd ed.
p. cm.

Rev. ed of: Relational database design clearly explained, 1998.

Includes bibliographical references and index.

ISBN 978-0-12-374730-3

1. Relational databases. 2. Database design. I. Harrington, Jan L.

Relational database design clearly explained. II. Title.

QA76.9.D26H38 2009

005.75'6—dc22

2009022380

ISBN: 978-0-12-374730-3

For information on all Morgan Kaufmann publications,

visit our Web site at www.mkp.com or
www.elsevierdirect.com

Printed in the United States of America

09 10 11 12 13 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

Relational Database Design and Implementation

Third Edition

Preface to the Third Edition

My favorite opening line for the database courses I teach is “Probably the most misunderstood term in all of business computing is *database*, followed closely by the word *relational*.” At that point, some students look a bit smug because they are absolutely, positively sure that they know what a database is and that they also know what it means for a database to be “relational.” Unfortunately, the popular press, with the help of some PC software developers, long ago distorted the meaning of both terms, which led many businesses to think that designing a database is a task that could be left to any clerical worker who had taken a one-week course on using database software. As you will see throughout this book, however, nothing could be further from the truth.

Note: The media has given us a number of nonsense computer terms such as *telephone modem* (we’re modulating an analog signal, not a telephone), *software program* (the two words mean pretty much the same thing), and *cable modem* and *DSL modem* (they’re not modems; they don’t modulate and demodulate analog signals; they are more properly termed *codecs* that code and decode digital signals). It’s all in an attempt to make computer jargon easier for people to understand, but it has generally had the effect of introducing misunderstandings.

This book is intended for anyone who has been given the responsibility for designing or maintaining a relational database. It will teach you how to look at the environment your database serves and to tailor the design of the database to the environment. It will also teach you how to design the database so it provides accurate and consistent data, avoiding the problems that are common to poorly designed databases. In addition, you will learn about design compromises that you might choose to make in the interest of database application performance and the consequences of making such choices.

If you are a college instructor, you may choose to use this book as a text in an undergraduate database management course. I’ve been doing that for a number of years (along with *SQL Clearly Explained*, this book’s companion volume) and find that students learn from it quite well. They appreciate the straightforward language rather than a text that forces them to struggle with overly academic sentence structures. They also like the many real-world examples that appear throughout the book.

Changes in the Third Edition

The core of this book—Parts II and III, the bulk of the content of the previous editions—remains mostly unchanged from the second edition. Relational database theory has been relatively stable for more than 30 years (with the exception of the addition of sixth normal form) and requires very little updating from one edition to the next, although

it has been seven years since the second edition appeared. The major changes are the discussions of fifth and sixth normal forms. The first two case studies in Part III have been updated; the third case study is new.

The chapter on object-relational databases has been removed from this edition, as well as object-relational examples in the case studies. There are two reasons for this. First, support for objects within a relational environment has largely been provided as a part of the SQL standard rather than as changes to underlying relational database theory. Second, the direction that SQL's object-relational capabilities have taken since the second edition appeared involves a number of features that violate relational design theory, and presenting them in any depth in this book would be more confusing than helpful.

By far the biggest change, however, is the addition of the new Parts I and IV. Part I contains three chapters that provide a context for database design. Database requirements don't magically appear at the point an organization needs a database, although looking at the previous editions of this book, you might think they did. Chapter 1 presents several organizational aspects of database management, including the hardware architectures on which today's databases run, and a look at service-oriented architecture (SOA), an information systems technique in which databases, like other IT functions, become services provided throughout an organization.

Chapter 2 provides an overview of several systems analysis methods to show you how organizations arrive at database requirements. In Chapter 3 you'll discover why we care about good database design. (It really *does* matter!)

Part IV provides an overview of a variety of database implementation issues that you may need to consider as you design a relational database. The topics include concurrency control (keeping the database consistent while multiple users interact with it at the same time), data warehousing (understanding issues that may arise when your operational database data are destined for data mining), data quality (ensuring that data are as accurate and consistent as possible), and XML (understanding how today's databases support XML).

The addition of Parts I and IV also make this book better suited for use as a textbook in a college course. When I used the second edition as a text in my classes, I added supplementary readings to cover that material. It's nice to have it all in once place!

The material about older data models that was presented in Chapter 3 in the second edition has been moved into an appendix. None of the material in the body of the book depends on it any longer. You can read it if you are interested in knowing what preceded the relational data model, but you won't lose anything significant in terms of relational databases if you skip it.

What You Need to Know

When the first edition of this book appeared in 1999, you needed only basic computer literacy to understand just about everything the book discussed. The role of networking in database architectures has grown so much in the past decade that in addition to computer literacy, you now need to understand some basic network hardware and software concepts (e.g., the Internet, interconnection devices such as routers and switches, and servers).

Note: It has always been a challenge to decide whether to teach students about systems analysis and design before or after database management. Now we worry about where a networking course should come in the sequence. It's tough to understand databases without networking, but at the same time, some aspects of networking involve database issues.

Acknowledgments

As always, getting this book onto paper involved an entire cast of characters, all of whom deserve thanks for their efforts. First are the people at Morgan Kaufmann:

- Rick Adams, my editor of many years. (His official title is Senior Acquisitions Editor).
- Heather Scherer, Rick's capable assistant
- Marilyn Rash, the project manager. We've worked together on a number of books over many years and it's always a pleasure.
- Eric DeCicco, the designer of the wonderful cover.
- The folks who clean up after me: Debbie Prato, copyeditor, and Samantha Molineaux, proofreader.
- Ted Laux, the indexer.
- Greg deZam-O'Hare and Sarah Binns who pulled it all together at the end.

A special thanks goes out to my colleague, Dr. Craig Fisher, who is a well-known expert on data quality. He provided me with a wealth of resources on that topic, which he thinks should be a part of everyone's IT education.

JLH



Contents

Preface.....	xv
Acknowledgments.....	xix

PART I INTRODUCTION

CHAPTER 1 The Database Environment.....	3
Defining a Database.....	4
Lists and Files.....	4
Databases.....	5
Data “Ownership”.....	6
Service-Oriented Architecture.....	7
Database Software: DBMSs.....	8
Database Hardware Architecture.....	10
Centralized.....	10
Client/Server.....	13
Distributed.....	14
The Web.....	16
Remote Access.....	17
Other Factors in the Database Environment.....	18
Security.....	18
Government Regulations and Privacy.....	20
Legacy Databases.....	21
For Further Reading.....	23
CHAPTER 2 Systems Analysis and Database Requirements.....	25
Dealing with Resistance to Change.....	26
The Structured Design Life Cycle.....	27
Conducting the Needs Assessment.....	28
Assessing Feasibility.....	32
Generating Alternatives.....	34
Evaluating and Choosing an Alternative.....	35
Creating Design Requirements.....	36
Alternative Analysis Methods.....	36
Prototyping.....	36
Spiral Methodology.....	38
Object-Oriented Analysis.....	38
For Further Reading.....	42

PART II DATABASE DESIGN THEORY

CHAPTER 3	Why Good Design Matters	45
	Effects of Poor Database Design	45
	Unnecessary Duplicated Data and Data Consistency.....	47
	Data Insertion Problems.....	48
	Data Deletion Problems	49
	Meaningful Identifiers.....	50
CHAPTER 4	Entities and Relationships	51
	Entities and Their Attributes.....	51
	Entity Identifiers	53
	Single-Valued versus Multivalued Attributes	54
	Avoiding Collections of Entities.....	56
	Documenting Entities and Their Attributes.....	58
	Entities and Attributes for Antique Opticals	60
	Domains.....	61
	Documenting Domains.....	61
	Practical Domain Choices.....	62
	Basic Data Relationships.....	64
	One-to-One Relationships	64
	One-to-Many Relationships.....	66
	Many-to-Many Relationships.....	67
	Weak Entities and Mandatory Relationships.....	67
	Documenting Relationships	68
	Basic Relationships for Antique Opticals.....	71
	Dealing with Many-to-Many Relationships.....	72
	Composite Entities	73
	Documenting Composite Entities.....	74
	Resolving Antique Opticals' Many-to-Many Relationships	75
	Relationships and Business Rules	77
	Data Modeling versus Data Flow	77
	Schemas.....	80
	For Further Reading.....	83
CHAPTER 5	The Relational Data Model	85
	Understanding Relations.....	86
	Columns and Column Characteristics.....	86
	Rows and Row Characteristics	87
	Types of Tables.....	87
	A Notation for Relations.....	88

Primary Keys	88
Primary Keys to Identify People	89
Avoiding Meaningful Identifiers	90
Concatenated Primary Keys	91
All-Key Relations	92
Representing Data Relationships	93
Referential Integrity	95
Foreign Keys and Primary Keys in the Same Table	95
Views	96
The View Mechanism	96
Why Use Views?	97
The Data Dictionary	97
Sample Data Dictionary Tables	98
A Bit of History	99
For Further Reading	101
CHAPTER 6 Normalization	103
Translating an ER Diagram into Relations	103
Normal Forms	105
First Normal Form	106
Understanding Repeating Groups	106
Handling Repeating Groups	107
Problems with First Normal Form	109
Second Normal Form	111
Understanding Functional Dependencies	111
Using Functional Dependencies to Reach 2NF	112
Problems with 2NF Relations	113
Third Normal Form	114
Transitive Dependencies	114
Boyce-Codd Normal Form	116
Fourth Normal Form	117
Multivalued Dependencies	118
Fifth Normal Form	119
Projections and Joins	120
Understanding 5NF	122
Sixth Normal Form	125
For Further Reading	126
CHAPTER 7 Database Structure and Performance Tuning	127
Joins and Database Performance	128
Indexing	132

Deciding Which Indexes to Create.....	133
Clustering.....	134
Partitioning.....	135
Horizontal Partitioning.....	136
Vertical Partitioning.....	136
For Further Reading.....	137
CHAPTER 8 Codd's Rules for Relational Database Design.....	139
Rule 1: The Information Rule.....	140
Rule 2: The Guaranteed Access Rule.....	141
Rule 3: Systematic Treatment of Null Values.....	142
Rule 4: Dynamic Online Catalog Based on the Relational Model.....	143
Rule 5: The Comprehensive Data Sublanguage Rule.....	144
Rule 6: The View Updating Rule.....	145
Rule 7: High-Level Insert, Update, Delete.....	145
Rule 8: Physical Data Independence.....	146
Rule 9: Logical Data Independence.....	147
Rule 10: Integrity Independence.....	147
Rule 11: Distribution Independence.....	148
Rule 12: Nonsubversion Rule.....	149
CHAPTER 9 Using SQL to Implement a Relational Design.....	151
Database Structure Hierarchy.....	151
Naming and Identifying Structural Elements.....	153
Schemas.....	154
Creating a Schema.....	154
Identifying the Schema You Want to Use.....	155
Domains.....	156
Tables.....	157
Column Data Types.....	158
Default Values.....	160
Not Null Constraints.....	161
Primary Keys.....	161
Foreign Keys.....	161
Additional Column Constraints.....	170
Views.....	170
Deciding Which Views to Create.....	170
View Updatability Issues.....	171
Creating Views.....	172
Temporary Tables.....	173

Creating Temporary Tables	174
Loading Temporary Tables with Data	174
Disposition of Temporary Table Rows	174
Creating Indexes	175
Modifying Database Elements	176
Adding Columns	176
Adding Table Constraints	177
Modifying Columns	177
Deleting Table Elements	179
Renaming Table Elements	179
Deleting Database Elements	179
CHAPTER 10 Using CASE Tools for Database Design	181
CASE Capabilities	182
ER Diagram Reports	183
Data Flow Diagrams	186
The Data Dictionary	188
Code Generation	191
Sample Input and Output Designs	193
The Drawing Environment	195
For Further Reading	196
CHAPTER 11 Database Design Case Study 1: Mighty-Mite Motors	197
Corporate Overview	197
Product Development Division	199
Manufacturing Division	200
Marketing and Sales Division	201
Current Information Systems	202
Reengineering Project	204
New Information Systems Division	204
Basic System Goals	205
Current Business Processes	206
Designing the Database	215
Examining the Data Flows	215
The ER Diagram	218
Creating the Tables	223
Generating the SQL	225
CHAPTER 12 Database Design Case Study 2: East Coast Aquarium	231
Organizational Overview	231
Animal Tracking Needs	233
The Volunteer Organization	236

The Volunteers Database.....	238
Creating the Application Prototype	238
Creating the ER Diagram	248
Designing the Tables	251
Generating the SQL.....	251
The Animal Tracking Database.....	254
Highlights of the Application Prototype	255
Creating the ER Diagram	260
Creating the Tables.....	264
Generating the SQL.....	265

PART III RELATIONAL DESIGN PRACTICE

CHAPTER 13 Database Design Case Study 3: SmartMart	275
The Merchandising Environment.....	275
Product Requirements.....	276
In-Store Sales Requirements	276
Web Sales Requirements	276
Personnel Requirements.....	277
Putting Together an ERD	277
Stores, Products, and Employees.....	277
In-Store Sales.....	281
Web Sales	282
Creating the Tables.....	284
Generating the SQL.....	286

PART IV DATABASE IMPLEMENTATION ISSUES

CHAPTER 14 Concurrency Control.....	299
The Multiuser Environment.....	299
Transactions	300
Logging and Rollback.....	300
Recovery.....	303
Problems with Concurrent Use.....	304
Lost Update #1	304
Lost Update #2	305
Inconsistent Analysis.....	307
Dirty Reads.....	309
Nonrepeatable Read	309
Phantom Read	310
Solution 1: Classic Locking	311

Read or Exclusive Locks	311
Read or Shared Locks	315
Two-Phase Locking	316
Locks and Transaction Length	317
Solution 2: Optimistic Concurrency Control (Optimistic Locking)	318
Solution #3: Multiversion Concurrency Control (Timestamping)	318
Transaction Isolation Levels	319
Web Database Concurrency Control Issues	320
Distributed Database Issues	321
For Further Reading	322

CHAPTER 15 Database Security	323
Sources of External Security Threats	324
Physical Threats	324
Hackers and Crackers	325
Types of Attacks	326
Sources of Internal Threats	327
Employee Threats	327
External Remedies	329
Securing the Perimeter: Firewalls	329
Handling Malware	331
Buffer Overflows	331
Physical Server Security	332
User Authentication	333
VPNs	335
Combating Social Engineering	336
Handling Other Employee Threats	338
Internal Solutions	338
Internal Database User IDs and Passwords	338
Authorization Matrices	339
Granting and Revoking Access Rights	341
Who Has Access to What	343
Backup and Recovery	344
Backup	345
Disaster Recovery	347
The Bottom Line: How Much Security Do You Need?	348
For Further Reading	349

CHAPTER 16 Data Warehousing	351
Scope and Purpose of a Data Warehouse.....	352
Obtaining and Preparing the Data	354
Data Modeling for the Data Warehouse.....	356
Dimensional Modeling Basics	356
Dates and Data	358
Data Warehouse Appliances	358
For Further Reading.....	361
 CHAPTER 17 Data Quality.....	363
Why Data Quality Matters.....	363
Recognizing and Handling Incomplete Data.....	364
Missing Rows.....	365
Missing Column Data	365
Missing Primary Key Data.....	366
Recognizing and Handling Incorrect Data	366
Wrong Codes.....	367
Wrong Calculations	367
Wrong Data Entered into the Database.....	368
Violation of Business Rules	369
Recognizing and Handling Incomprehensible Data.....	369
Multiple Values in a Column	369
Orphaned Foreign Keys.....	370
Recognizing and Handling Inconsistent Data.....	371
Inconsistent Names and Addresses	371
Inconsistent Business Rules	371
Inconsistent Granularity.....	372
Unenforced Referential Integrity	373
Inconsistent Data Formatting	373
Preventing Inconsistent Data on an Organizational Level.....	373
Employees and Data Quality.....	374
For Further Reading.....	375
 CHAPTER 18 XML.....	377
XML Syntax	377
XML Document Correctness.....	380
XML Schemas.....	380
XML Support in Relational DBMSs.....	382

DB2..... 382

Oracle 384

For Further Reading..... 385

APPENDIX HISTORICAL ANTECEDENTS..... 387

GLOSSARY 407

INDEX..... 413