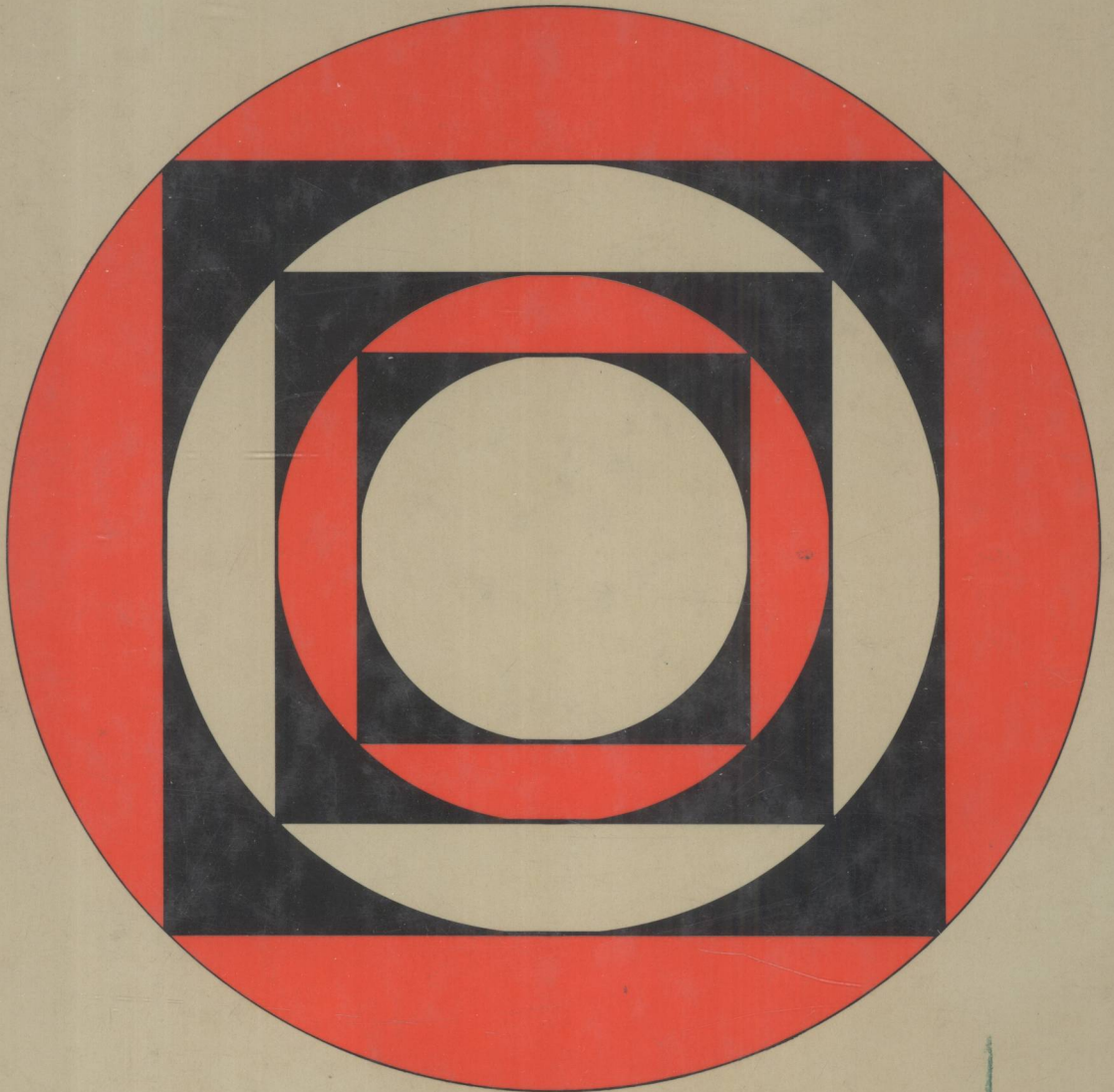


# DATABASE ANALYSIS AND DESIGN

I. T. HAWRYSZKIEWYCZ



---

# **DATABASE ANALYSIS AND DESIGN**

---

---

**I. T. Hawryzkiewicz**

---

Canberra College of Advanced Education, Australia

---

---

**To Daria, Helen, and Adrian**

---

Acquisition Editor  
Project Editor  
Compositor  
Illustrator  
Text Designer  
Cover Design

Alan Lowe  
E. Ann Wood  
Allservice Phototypesetting  
John Foster  
Barbara Ravizza  
Joe DiChiarro

**Library of Congress Cataloging in Publication Data**

Hawryszkiewicz, I. T.

Database analysis and design.

Bibliography: p.

Includes index.

1. Data base management. 2. System design. 3. System analysis. 4. File organization (Computer science)

I. Title.

QA76.9.D3H386 1984 001.64'2 84-1341

ISBN 0-574-21485-2

Copyright © Science Research Associates, Inc. 1984.  
All rights reserved.

Printed in the United States of America.

---

# Preface

---

Database management is now widely practiced and has become an important field of study in computing. The last few years have seen considerable evolution in database practical application and in the theoretic database foundations. This evolution has been accomplished by the development of many areas of study that together fall under the database umbrella. Some of these areas are practical while others are theoretically oriented. Examples are relational theory, data semantics, database management software, or physical file design.

To the new student it is easy to treat these areas of study as unrelated topics of interest. What, after all, is the link between relational theory and physical file design or between data semantics and database management system software. Thus one may become familiar with all these areas of study and yet not be aware of how they combine together.

One goal of this text is to unite areas of database study into a coherent field within the context of database design. One may, of course, ask why use database design for this purpose? To answer we can postulate that database design is to some extent the peak of application of the database field. Ultimately our goal is to design and use databases. If a field of database study is to be ultimately useful then it must somehow fit into the design cycle.

Apart from serving as a framework for unifying fields of database study, database design also provides the criteria to judge good databases. Again the elements of database study must provide designers with the ability to evaluate databases against these criteria. Design criteria met in practice include

- all user relevant relationships can be easily deduced from the chosen database structure
- the database can evolve to meet changes to user requirements without requiring extensive changes to the database structure and programs
- user access requests are met within reasonable performance criteria.

The fields of database study must contribute towards these criteria. It should be noted that database design itself is still not a very structured process. Early designers had few techniques or tools to assist them. In general, early designers used file design techniques for database design; many of these techniques concentrated on physical design criteria and emphasized performance rather than data structure or flexibility.

The last few years, however, have seen the development of many new approaches to database design. The trend has been to methodologies that commence with a formal analysis of the user data structure and then reduce such requirements to a physical design through a sequence of structured steps. Thus we must first study the data that we are to store and then choose its computer representation. Much of this work commenced after the development of relational theory, which provided a formal basis for choosing data structures and can be readily translated into practical terms.

Since then a variety of database analysis and design techniques has developed. Some use the relational model as their basis; others are guided by data abstractions such as entities and relationships, role or generalization, and aggregation to structure databases.

The system design cycle thus provides a basis for integrating the various database fields. Relational theory and semantic models are useful for analyzing and structuring data. The structures developed are then converted to a logical model based on some data model, which in turn is implemented by commercial database software. Hence the relationship between semantic models and commercial software. Finally the implementation may be adjusted, taking into account access requirements and physical structures available to the designer. Hence a relationship evolves from logical structure to physical design. This book follows this design cycle, outlining the database fields of study and the theory behind them as it proceeds. The book itself evolved from the author's experiences during his Doctoral dissertation at the Massachusetts Institute of Technology, where he was first introduced to relational theory by Professor Jack Dennis. Subsequently the author has drawn on his practical experiences in database design and in developing and teaching database courses to college students and in database research, both in Australia and the United States, where he spent time with Professor Edgar Sibley at the University of Maryland. It has also benefited from update courses taught by the author to database practitioners, which over the years has led to many interesting examples with practical orientations. These examples are included throughout the book.

The book has also benefited from the many constructive comments from its reviewers and in this context the author would like to acknowledge the contributions of Marilyn Bohl of IBM, Caroline M. Eastman of Southern Methodist University, and Kenneth M. Hunter of San Francisco State University, and Nancy D. Griffith of Georgia Institute of Technology. Finally, the book would not have been possible without the constant help of SRA staff, in particular the patient guidance of Alan Lowe through its initial stages and the support provided by E. Ann Wood and the SRA editorial staff in the final production stages.

---

# Contents

---

---

## 1 INTRODUCTION 1

---

COMPUTERS AND DATABASES	1
THE EVOLUTION OF DATABASE DESIGN PROCEDURES	2
WHAT ARE THE DESIGN PROCEDURES?	3
WHAT ARE THE DESIGN STAGES?	4
TOOLS FOR ANALYSIS AND DESIGN	6
REQUIREMENTS OF DATABASE SPECIFICATION AND DESIGN	8
TECHNICAL DESIGN OF THE DATABASE	9
TECHNIQUES IN DATABASE DESIGN	10
What Are the Design Techniques?	11
THE BROAD STRUCTURE OF THE TEXT	11
Data Analysis	11
Implementation Models	12
Problems	13

---

## 2 RELATIONAL MODEL 15

---

INTRODUCTION	15
The Relational Model—A Historic Perspective	15
RELATION MODEL—THE BASIC STRUCTURE	16
Terminology	18
Attributes and Domains	18
Properties of Relations	20
Key of Relations	21
Consistency	22
Tuple Operations	22
Anomalies	23
FUNCTIONAL DEPENDENCY	25
Full Functional Dependency	26

NORMAL FORMS	27
Example	29
Boyce–Codd Normal Form	29
FUNCTIONAL DEPENDENCIES AND RELATIONAL DESIGN	32
Properties of Functional Dependencies	32
The Membership Algorithm	34
RELATIONAL LANGUAGES	36
Relational Calculus	36
QUEL	38
SQL	38
Nested Mapping in SQL	40
Expressing Relational Joins	41
Retrievals Using EXISTS	41
Retrievals Using ANY	41
Set Exclusion	42
Compound Conditions	42
Functions	43
Grouping and Partitioning	43
Set Operations	43
Relational Algebra	44
Selection	44
Projection	44
Joining	45
Division	47
SUMMARY	47
PROBLEMS	48

---

**3 RELATIONAL DESIGN** **59**

---

INTRODUCTION	59
THE UNIVERSAL RELATION ASSUMPTION	59
RELATIONAL DESIGN CRITERIA	61
Satisfying Representation Criteria	63
Lossless Decompositions	65
Conditions for Lossless Decompositions	66
Redundancy Criteria	66
RELATIONAL DESIGN PROCEDURES	67
DECOMPOSITION	67
Simple Decomposition Algorithm	67
Limitations of the Single Decomposition Algorithm	68
Multi-valued Dependency	71
Context Dependence of MVDs	74

Fourth Normal Form	76
Decomposition Algorithm 2	76
MVDs and FDs	77
Some Complexities of Decomposition Algorithms	78
SYNTHESIS	80
Semantics of Functional Dependencies	80
Synthesis Algorithms	82
Recent Developments in Relational Theory	86
SUMMARY	88
PROBLEMS	89

---

#### **4 SYNTACTIC AND SEMANTIC DESIGN ISSUES IN DATA ANALYSIS 93**

---

INTRODUCTION	93
THE RELATIONAL MODEL IN SYSTEMS ANALYSIS	94
Semantic Problems in Relational Modeling	95
RECORD-BASED DATA MODELS	97
ENRICHING THE RELATIONAL MODEL	98
SEMANTIC MODELS	98
HOW SEMANTIC MODELS MODEL ENTERPRISES	100
Semantic Abstractions	100
Representing Semantic Models	102
SUMMARY	104
PROBLEMS	105

---

#### **5 SEMANTIC MODELING, I—ENTITIES AND RELATIONSHIPS 107**

---

INTRODUCTION	107
THE ENTITY RELATIONSHIP MODEL	107
Entity and Relationship Sets	109
A Diagrammatic Representation	109
Multivalued Attributes	110
Nonfunctional Dependencies	112
Identifiers	112
Modeling with the E-R Model	115
ENTITIES, RELATIONSHIPS AND RELATIONS	117
CHOICE OF ENTITIES AND RELATIONSHIPS	118
DEPENDENCE BETWEEN ENTITIES	120



Problems with the Use of Composite Keys	121
Improper Use of Composite Keys	123
MULTIPLE RELATIONSHIPS	123
Multiple Relationships and Composite Keys	126
BINARY AND <i>n</i> -ARY RELATIONSHIPS	127
THE ENTITY MODEL	132
UNREPRESENTABLE RELATIONSHIPS	132
MINIMALITY OF RELATIONS	134
SUMMARY	134
PROBLEMS	135

---

**6 SEMANTIC MODELING, II—ROLES AND TYPES** **139**

---

INTRODUCTION	139
RECURSIVE RELATIONSHIPS	141
EXTENSION TO ROLES	143
Modeling Properties of Roles	144
Another Advantage of the Role Concept	148
Nonhomogeneity	148
Nonhomogeneity and Functional Dependency	150
Role-Modeling Structures	150
Role Identifiers and Role Structure	153
ROLES, SOURCES, AND RELATIONS	155
Role Structures and Functional Dependencies	156
Uniform Role Set	156
Nonuniform Role Set	157
Mandatory Roles	157
Optional Roles	157
Using Composite Identifiers for Roles	160
Composite Identifiers and Mandatory Roles	160
Composite Identifiers and Optional Roles	161
Short Cuts	162
TYPES OF ENTITIES	162
Modeling Entity Types	166
Relations and Nonuniform Type Sets	167
Short Cuts	171
AGGREGATE OBJECTS	171
COMBINING ENTITY TYPES AND AGGREGATE OBJECTS	174
SUMMARY	176
PROBLEMS	177

---

**7 FURTHER GENERALIZATIONS AND PROBLEMS  
IN SEMANTIC MODELING**
**187**

INTRODUCTION	187
AGGREGATION AND GENERALIZATION	188
Cluster	191
BINARY MODELING	194
FUNCTIONAL MODEL	197
SUMMARY	199
PROBLEMS	200

---

**8 DATABASE SPECIFICATIONS**
**201**

INTRODUCTION	201
CONVERSION METHODS	203
“NORMALIZING” THE SEMANTIC MODEL	204
FROM “NORMALIZED” SEMANTIC MODELS	
TO RECORD SPECIFICATIONS	207
Logical Structures for Uniform Roles	209
Nonuniform Mandatory Roles	209
Nonuniform Optional Roles	211
Some General Comments on Role and Type Conversion	211
FROM RELATIONAL MODEL TO LOGICAL STRUCTURE	212
MINIMAL LOGICAL RECORD STRUCTURE (MLRS)	216
SPECIFYING ACCESS REQUIREMENTS	216
Access Paths	217
Using Pseudocode	220
QUANTITATIVE DATA SPECIFICATIONS	224
SUMMARY	225
PROBLEMS	226

---

**9 IMPLEMENTATION MODELS,  
I—FILE STRUCTURES**
**227**

INTRODUCTION	227
IMPLEMENTATION MODELS	230
Level Integration	232
Conversion between Levels	232
Generating Mapping Levels	236

## X Database Design

Using the Framework	237
Using the Framework in Design	237
Using the Framework as a Software Philosophy	237
THE PHYSICAL RECORD INTERFACE	237
Access Method Software for Disk Transfer	239
Other Physical Interface Variations	241
Pages	242
Variable Record Size	242
Variable Physical Record Sizes	242
Multiple Buffers	243
The Physical Interface—Implications for Designers	244
LOGICAL RECORD ACCESS	244
Kinds of Logical Processing	245
SEQUENTIAL ACCESS	245
DIRECT ACCESS	246
Hash Access Methods	248
Index Implementations	250
Indexing Techniques	251
<i>Example 1</i>	252
<i>Example 1A</i>	253
Multilevel Indexing	254
Dense and Nondense Indices	255
Maintaining Indices	255
B-Trees	257
Maintaining B-Trees	257
Comparison of Direct Access Methods	260
MULTI-INDEX ACCESS	260
INDEXED SEQUENTIAL ACCESS METHODS	261
LINKING FILES	262
Symbolic Pointers	263
Logical Record Address Pointers	264
AVAILABLE ACCESS METHODS	265
COBOL Input/Output	267
Indexed Sequential File Organization	267
Relative File Organization	270
SUMMARY	271
PROBLEMS	271

---

## 10 IMPLEMENTATION MODELS, II—DATABASE MANAGEMENT SYSTEMS 275

---

INTRODUCTION	275
NATURAL USER INTERFACE	276

Data Models	276
USER VIEWS	278
DATA INDEPENDENCE	279
Database Restructuring	280
INTERFACE SOFTWARE	282
Using the DBMS Software	283
DBMS ARCHITECTURES	285
Three-Level Architectures	285
SCHEMA-SUBSCHEMA Architectures	288
MAINTAINING THE OPERATIONAL ENVIRONMENT	288
Types of Operational Environment	289
Processing Modes	290
On-line Processing	290
<i>Single-Thread On-Line Processing</i>	292
<i>Multithread On-Line Processing</i>	292
<i>Transaction Processor</i>	293
DATABASE FACILITIES TO SUPPORT A MULTI-USER ENVIRONMENT	294
Database Integrity	295
Concurrency Control	296
Recovery	296
Database Privacy	298
Database Distribution	298
Technical Considerations in Distributed Databases	299
Data Dictionaries	300
A Typical Software Structure	301
SUMMARY	302
PROBLEMS	304

---

## 11 RELATIONAL DATABASE MANAGEMENT SYSTEMS 305

---

INTRODUCTION	305
SYSTEM R	307
The SQL Interface	308
Defining the Database	308
Populating and Updating the Database	309
Tuple Insertion	309
Tuple Deletion	309
Tuple Update	309
Amending the Database Definition	310
Removing a Relation	310
Expanding a Relation	310

Defining User Views	310
Embedded SQL	311
Implementation	314
Access Paths	315
Indices	316
Links	316
Some Operational Features	316
QUERY-BY-EXAMPLE	317
Conditional Retrieval	318
Some Options	320
Retrieval from More Than One Relation	321
Negation	322
Functions	323
Set Comparison	323
Insertion, Deletion, and Update	324
Creation and Changes of Database Definition	325
PERSONAL COMPUTER SYSTEMS	326
The User Interface	327
Using the Personal Database	327
Query Processor	328
Procedural Language	329
Report Generator	330
Some Advanced Facilities	330
Creating an Index	330
Joining Two Files	331
SUMMARY	331
PROBLEMS	332

---

**12 NETWORK DATA MODEL 333**

---

INTRODUCTION	333
THE NETWORK MODEL	333
More Elaborate Network Constructs	337
NETWORK MODEL IMPLEMENTATIONS	338
Accessing Records in Network Database	343
THE DBTG IMPLEMENTATION	344
The DBTG Schema	346
AREAS or REALMS	346
RECORD ENTRIES	349
Description of Record Occurrences	349
Areas of Record Placement	349
Control of Record Placement	349
Proposed Development	349

SET ENTRIES	351
Owner and Member Record Types	351
Singular Sets	352
Multimember Set Types	352
Physical Storage Mode	353
Order of Member Records within a Set Occurrence	354
SORTED Orders for Single-Member Set Types	354
SORTED Orders for Multimember Set Types	354
Set Membership Class	355
Structural Constraint	357
Additional Physical Structures	358
ACCESSING A DBTG NETWORK DATABASE	359
Currency	360
The FIND Command	362
The FIND Command and Currency	362
Format 5	362
Finding Members of a Set Occurrence	362
Format 3	363
Format 6 and Format 7	364
<i>Set Occurrence Selection</i>	364
<i>Hierarchical Path Retrieval</i>	365
Format 4—Finding an Owner	367
THE SUBSCHEMA	368
Defining the Subschema	368
Invoking the Subschema	368
SUMMARY	368
PROBLEMS	370

---

**13 HIERARCHICAL DATABASE  
MANAGEMENT SYSTEMS** **379**

---

INTRODUCTION	379
THE HIERARCHICAL MODEL	379
The Hierarchical Data Structure	380
Parents with More than One Child Object Class	381
Accessing the Hierarchical Data Structure	381
<i>Tree Traversal</i>	382
<i>General Selection</i>	383
INFORMATION MANAGEMENT SYSTEM (IMS)	383
The IMS Structure	384
The Physical Database	384
Logical Databases	386

IMS Physical Structures	388
Record Storage	389
Logical Pointers	391
Defining the IMS Database	393
Defining Physical Databases	394
Defining Logical Relationships	395
Defining the Physical Structure	397
Defining Logical Databases	397
Program Communications Blocks	399
Secondary Indexing	400
Accessing an IMS Database	401
Direct Retrieval	402
Sequential Retrieval	402
Updating the Database	403
Some Operational Features	404
GENERALIZED SELECTION	405
SYSTEM 2000	405
The System 2000 Data Structure	405
Populated Database	406
Defining the System 2000 Database	406
Accessing the System 2000 Database	408
Hierarchical Access Commands	409
System 2000 Immediate Access	409
<i>Output without Conditions</i>	410
<i>Output with Conditions</i>	411
Qualified Fields in Descendant Nodes	412
Qualified Fields in Ancestor Nodes	412
Updating the Database	412
Embedded Language Facilities	413
Physical Database Structure	414
Some System 2000 Operational Features	415
SUMMARY	417
PROBLEMS	417

---

**14 THE DESIGN PROCESS****422**

---

INTRODUCTION	422
DESIGN OBJECTIVES	424
DESIGN TECHNIQUES	425
Combining Techniques into a Design Methodology	425
INITIAL DESIGN	426
Choosing the Initial Physical Structure	428
DESIGN ITERATIONS	430

Performance Problems	430
Design Problems	430
Design Tactics	432
LOGICAL DESIGN TACTICS	432
Reducing the Number of Access Steps	432
Derived Relationships	433
Example	433
Duplication of Data Items	434
Combining Files	434
Combining into Nonhomogeneous Files	436
PHYSICAL DESIGN TACTICS	436
Problem PR3—Pointer Manipulations	436
Problem PR4—Excessive Storage Requirements	437
Problem PR5—System Overheads	437
SUMMARY	438

---

<b>15 INITIAL DESIGN</b>	<b>441</b>
--------------------------	------------

---

INTRODUCTION	441
RELATIONAL DATABASE DESIGN	441
CONVERSION TO A NETWORK MODEL	442
Properties of Conversion Rules	443
Conversion of Roles and Types	447
Which Conversion to Use	447
INITIAL NETWORK PHYSICAL DESIGN	450
Example	452
HIERARCHICAL DESIGN	453
Choosing a Hierarchical Design Method	455
Class 1 Design Methods	457
Partitioned Design	459
Partitioning	461
Class 2 Design Methods	463
Class 3 Design Methods	467
Conversion of Roles and Types	469
INITIAL PHYSICAL DESIGN	472
COBOL FILE DESIGN	472
Key Conflicts	474
Conversion of Roles and Types	475
Conversion of Nonuniform Roles and Types	476
Optional Nonuniform Roles and Types	477
Links with Roles	478
Conflicts with Performance Criteria	478
SUMMARY	479
PROBLEMS	480



---

**16 EVALUATING DESIGNS**

---

**481**

INTRODUCTION	481
SELECTING A DBMS	481
MONITORING THE DATABASE	482
PERFORMANCE ESTIMATES	483
Logical and Physical Design Analysis	483
Accuracy of Estimates	484
Generality of Analysis	484
Record Access	484
Analysis Procedures	486
Storage Requirements	488
Logical Analysis	489
Logical Analysis—Access Requirement S	493
Analysis of Batch Processing	493
Logical Analysis—Access Requirement V	494
Physical Analysis	494
Estimating Data Record Transfers	495
Estimating Index Transfers	495
Estimating Total Transfers	496
Physical Analysis—Access Requirement S	496
Physical Analysis—Access Requirement V	496
Variations	497
Estimates for DBMSs	497
ANALYTICAL MODELING METHODS	499
Some Probabilistic Estimates	499
Estimating Response Times	502
Example 1	504
Example 2	506
Example 3	506
Variations in Distribution	507
Extensions to Analytic Modeling	507
SUMMARY	508
PROBLEMS	508

---

**17 CHOOSING DESIGN METHODOLOGIES**

---

**515**

INTRODUCTION	515
CONSTRUCTING A DESIGN METHODOLOGY	515
Selecting Design Techniques	516
Selecting Documentation Methods	516
CHOICE 1: COLLECTING DATA	517