



Contents

1 *Introduction* 1

- 1.1 Categories of Data Processing Software 2
 - Category 1: No Data Independence 2
 - Category 2: Physical Data Independence 3
 - Category 3: Partial Logical Data Independence 4
 - Category 4: Logical and Physical Data Independence 4
 - Category 5: Geographical Data Independence 6
- 1.2 File Overview and Terminology 7
 - File Access Techniques 9
 - File and Program Interdependence 9
 - Data Redundancy 9
- 1.3 Database Overview and Terminology 10
 - Functional Database Description 10
 - Logical Database Description 12
 - Physical Database Description 12
- 1.4 Example Data Processing Application 13
 - SECS Example 13
 - Requirements 14
 - Processes 14
 - Data 14

2 *Secondary Storage Devices and I/O Control* 17

- 2.1 The Storage Hierarchy 18
- 2.2 Tape-Based Devices 20

- Data Tape 20
- Tape Drive 21
- Read/Write Operations 21
- 2.3 Spinning Tracks 26
 - Data Recording 27
 - Read/Write Operations 29
 - Data Placement 31
 - Winchester Technology 32
- 2.4 Esoteric Storage Devices 33
 - Bubble Memories 33
 - Mass Storage Devices 33
 - Content Addressable Filestores 34
 - Optical Discs 34
- 2.5 Operating System Interface 35
 - Low-Level Operating System Functions 36
 - High-Level Operating System Functions 40
- 2.6 Programming Language Interface 42
 - File Opening/Closing 43
 - Data Transfer 43
 - File Manipulation 45
- 2.7 Data Compression 45
 - Elimination of Trailing Blanks 47
 - Run Length Encoding 47
 - Record Differences 48
 - Special Fields 48
 - Variable-Length Codes 49
 - Addressing Problems 50

3 *Serial and Sequential Files 55*

- 3.1 Serial Files—Program Design Methodology 56
 - Notation 56
 - Design Method Overview 58
 - Design Example 59
 - Structure Clashes 64
- 3.2 Master File Update 65
 - Batch Update 67
 - Problem Transactions 70
- 3.3 Sequential File Merging 72
 - Basic Algorithm 73
 - Efficiency Considerations 73

4 *External Sorting 81*

- 4.1 Overview of External Sorting 82
- 4.2 Generating Sorted Partitions 83
 - Internal Sorting 84
 - Replacement Selection 84
 - Natural Selection 85
 - Comparison of Partition-Forming Algorithms 86
- 4.3 Distribution and Merging 87
 - Balanced N-way Merge 89
 - Optimal Merge 92
 - Polyphase Merge 93
 - Comparison of Distribution and Merging Strategies 98

5 *Direct Files 105*

- 5.1 Addressing and Pointers 106
 - Addressing 106
 - Pointers 107
- 5.2 Hashing 109
 - Bucket Size 110
 - Packing Density 111
 - Transformation Function 113
 - Overflow Strategy 115
- 5.3 Table-Assisted Hashing 122
 - Retrievals 123
 - Insertions 123
 - Deletions 124
- 5.4 Extensible Files 124
 - Virtual Hashing 125
 - Dynamic Hashing 127
 - Extendible Hashing 132
 - Deletions 134
 - Order-Preserving Hashing 135
- 5.5 Filestore Reliability through Hashing 136

6 *Indexed Files: Primary Key Indexing of Sequential Files 143*

- 6.1 Static Indexes 147
 - Organization of the Index 147

- Insertions 148
- Physical Organization of Overflow Area 150
- 6.2 Dynamic Indexes 151
 - Binary Trees 152
 - AVL Trees 152
 - Multiway Trees 153
 - B-trees 153
 - B*-trees 170
 - B⁺-trees 175
- 6.3 Comparison of Static and Dynamic Indexes 181

7 *Multikey Processing 185*

- 7.1 Threaded Files 186
- 7.2 Multilists 189
- 7.3 Inverted Files 190
- 7.4 STAIRS: An Application of Inverted Files 193
 - File Structure 195
 - Answering Queries 196
- 7.5 Index Implementation 198
 - Bit Vectors 198
 - Graph Structure 198
 - Comparison of Bit Vectors and Graphs 199
- 7.6 Index Maintenance 201
 - Updating 201
 - Reliability 202
- 7.7 Grid Files 203
 - Design Aims 203
 - Ideal Solution 204
 - Practical Grid File Implementation 204
 - Performance of Grid Files 205

8 *Integrated File Addressing Techniques 211*

- 8.1 Review of Files 211
 - Single-Key File Processing 212
 - Multikey File Processing 212
- 8.2 Choice of File Organization 213
 - Time 214
 - File-Use Ratio 214

- Space 215
- Volatility 215
- 8.3 Integrated Files and Update Anomalies 216
 - Modification Anomalies 216
 - Deletion Anomalies 217
 - Insertion Anomalies 217
- 8.4 Example File Processing Application 218
 - Operations and Data 219
 - File Organization 220
 - Review of Example 224
- 8.5 Advantages and Disadvantages of Files 225
 - Advantages of Using Integrated Files 226
 - Disadvantages of Using Integrated Files 226

9 *Normalization 229*

- 9.1 Associations among Entity Attributes 230
 - Single-Valued Dependencies 231
 - Multivalued Dependencies 236
- 9.2 Normalization: Entity Design through Decomposition 238
 - Domain Key Normal Form 238
 - Unnormal Form 239
 - First Normal Form 240
 - Second Normal Form 240
 - Third Normal Form 241
 - Boyce-Codd Normal Form 244
 - Fourth Normal Form 246
 - Fifth Normal Form 246
 - Normalization and Efficiency 251
- 9.3 SECS Example Database Record Design 253

10 *Database Management Systems 265*

- 10.1 Data Processing with a DBMS 266
 - Advantages of Using a DBMS 266
 - Disadvantages of Using a DBMS 267
- 10.2 Functions of a DBMS 267
 - Maintaining Data Validity 268
 - Efficient and Flexible Data Processing 269
 - Database Administration 270

- 10.3 Architecture of a DBMS 271
 - Components of a DBMS 271
 - Data Independence 272
- 10.4 Database Life Cycle 274
 - Database Modeling 274
 - Database Implementation 275
 - Database Maintenance 276

11 *The Hierarchical and Network Data Models 281*

- 11.1 Properties of the Hierarchical Data Model 282
 - Advantages and Disadvantages of the Hierarchical Model 282
 - SECS Hierarchical Data Definition Trees 283
- 11.2 Hierarchical Data Definition 283
 - A Hierarchical Pseudo DDL 285
 - SECS Hierarchical Data Definition Example 285
- 11.3 Properties of a Hierarchical DML 285
 - Hierarchical Retrieval 288
 - Hierarchical Record Modification 288
 - SECS Hierarchical DML Examples 289
- 11.4 Analysis of the Hierarchical Data Model 289
- 11.5 Properties of the Network Data Model 290
 - M:N Associations in a Network 290
 - SECS Network Data Diagram 291
- 11.6 Network Data Definition 291
 - SECS Network Schema Example 296
 - Network Subschema Definition 296
- 11.7 Properties of Network DML Commands 298
 - Retrieval in a Network 299
 - Record Modification in a Network 300
 - Network Set Maintenance 300
 - SECS Network DML Examples 301
- 11.8 Analysis of the Network Data Model 302

12 *The Relational Data Model 305*

- 12.1 Properties of the Relational Model 305
- 12.2 Relational Data Definition 307

- Relation Specification: Logical Database 307
- Relational Views: Functional Databases 308
- 12.3 Three Levels of Relational DML 309
 - Tuple-by-Tuple DML 309
 - Relational Algebra 309
 - Relational Calculus 317
- 12.4 SECS Relational DML Examples 320
- 12.5 Comparison of Database Data Models 321
 - Comparison of the Three DBMS Data Models 321
 - Multimodel DBMS's 322

13 *Issues in Database Implementation 327*

- 13.1 Integrity of Databases 327
 - Specification of Constraints 328
 - Integrity Enforcement 329
 - Types of Constraints 329
 - Database Recovery 330
- 13.2 Concurrency in DBMS's 331
 - Data Locking 332
 - Process Timestamping 332
- 13.3 DBMS Security 333
 - Security Granularity 334
 - Security Specification 334
 - Security Enforcement 335
 - Security and Statistical Databases 335

14 *Advanced Database Topics 343*

- 14.1 Query Languages and Optimization 343
 - Query Languages: User Interface 343
 - Query Optimization 346
- 14.2 Database Machines 347
 - Associative Memories 348
 - Back-end Database Processing 349
 - Full-Text Information Retrieval 350

14.3 Distributed Database Systems 351
 Rationale for Data Distribution 351
 Design of a Distributed Database 352
 Transaction Processing with Distributed Databases 353

Glossary 359

References 371

Index 381

