Contents

Foreword xix
Preface xxi

Part 1. Foundation 1

Chapter 1. Introduction to Data Warehousing 3

1.1 Why All the Excitement? 4
1.2 The Need for Data Warehousing 5
1.3 Paradigm Shift 6
1.3.1 Computing Paradigm 6
1.3.2 Business Paradigm 7
1.4 Business Problem Definition 9
1.5 Operational and Informational Data Stores* 12
1.6 Data Warehouse Definition and Characteristics 14
1.7 Data Warehouse Architecture 17
1.8 Chapter Summary 21

Chapter 2. Client/Server Computing Model and Data Warehousing 23

2.1 Overview of Client/Server Architecture 23
2.1.1 Host-Based Processing 24
2.1.2 Master-Slave Processing 24
2.1.3 First-Generation Client/Server Processing 25
2.1.4 Second-Generation Client/Server Processing 27
2.2 Server Specialization in Client/Server Computing Environments 29
2.3 Server Functions 30
2.4 Server Hardware Architecture 33
2.5 System Considerations 34
2.6 RISC versus CISC 35
2.7 Multiprocessor Systems 39
2.7.1 SMP Design 41
2.7.2 SMP Features 43
2.7.3 SMP Operating Systems 44
2.8 SMP Implementations 45

Chapter 3. Parallel Processors and Cluster Systems 47

3.1 Distributed-Memory Architecture 47
3.1.1 Shared-Nothing Architectures 48
3.1.2 Shared-Disk Systems 49
3.2 Research Issues 50
3.3 Cluster Systems 52
3.4 Advances in Multiprocessing Architectures 54
3.5 Optimal Hardware Architecture for Query Scalability* 56
3.5.1 Uniformity of Data Access Times 56
3.5.2 System Architecture Taxonomy and Query Execution 56
3.6 Server Operating Systems 59
3.6.1 Operating System Requirements 60
3.6.2 Microkernel Technology 60
3.7 Operating System Implementations 61
3.7.1 UNIX 61
3.7.2 Windows/NT 62
3.7.3 OS/2 63
3.7.4 NetWare 65
3.7.5 OS Summary 65

Chapter 4. Distributed DBMS Implementations 67

4.1 Implementation Trends and Features of Distributed Client/Server DBMS 68
4.1.1 RDBMS Architecture for Scalability 70
4.1.2 RDBMS Performance and Efficiency Features 73
4.1.3 Types of Parallelism 75
4.2 DBMS Connectivity 81
4.3 Advanced RDBMS Features 83
4.4 RDBMS Reliability and Availability 85
4.4.1 Robustness, Transactions Recovery, and Consistency 85
4.4.2 Fault Tolerance 86
4.5 RDBMS Administration 87

Chapter 5. Client/Server RDBMS Solutions 89

5.1 State-of-the-Market Overview 89
5.2 Oracle 90
5.2.1 System Management 92
5.2.2 Oracle Universal Server 92
5.2.3 Oracle ConText Option 93
5.2.4 Oracle Spatial Data Option 95
5.3 Informix 95
5.3.1 Features 97
5.3.2 Informix Universal Server 98
5.4 Sybase 101
5.4.1 SYBASE SQL Server 101
5.4.2 Performance Improvements in SYBASE System 11 103
5.5 IBM 104
5.5.1 Background 104
5.5.2 DB2 Universal Database 104
5.6 Microsoft 108
5.6.1 Background 108
5.6.2 MS SQL Server 108
5.6.3 Data Warehousing and Market Positioning 111

Part 2 Data Warehousing 113

Chapter 6. Data Warehousing Components 115

6.1 Overall Architecture 115
6.2 Data Warehouse Database 116
6.3 Sourcing, Acquisition, Cleanup, and Transformation Tools 117
6.4 Metadata 118
6.5 Access Tools 120
6.5.1 Query and Reporting Tools 121
6.5.2 Applications 122
6.5.3 OLAP 122
6.5.4 Data Mining 122
6.5.5 Data Visualization 124
6.6 Data Marts* 124
6.7 Data Warehouse Administration and Management 126
6.8 Information Delivery System 127

Chapter 7. Building a Data Warehouse 129

7.1 Business Considerations: Return on Investment 130
7.1.1 Approach 130
7.1.2 Organizational Issues 131
7.2 Design Considerations 131
7.2.1 Data Content* 132
7.2.2 Metadata 132
7.2.3 Data Distribution 132
7.2.4 Tools 133
7.2.5 Performance Considerations 133
7.2.6 Nine Decisions in the Design of a Data Warehouse 134
7.3 Technical Considerations 136
7.3.1 Hardware Platforms 137
7.3.2 Data Warehouse and DBMS Specialization 139
7.3.3 Communications Infrastructure 139
7.4 Implementation Considerations 139
7.4.1 Access Tools 140
7.4.2 Data Extraction, Cleanup, Transformation, and Migration 141
7.4.3 Data Placement Strategies 143
7.4.4 Metadata 145
7.4.5 User Sophistication Levels 145
7.5 Integrated Solutions 146
7.6 Benefits of Data Warehousing 148
7.6.1 Tangible Benefits 148
7.6.2 Intangible Benefits 149

Chapter 8. Mapping the Data Warehouse to a Multiprocessor Architecture 151

8.1 Relational Database Technology for Data Warehouse 151
8.1.1 Types of Parallelism 152
8.1.2 Data Partitioning 153
8.2 Database Architectures for Parallel Processing 154
8.2.1 Shared-Memory Architecture 154
8.2.2 Shared-Disk Architecture 155
8.2.3 Shared-Nothing Architecture 156
8.2.4 Combined Architecture 158

8.3 Parallel RDBMS Features 158
8.4 Alternative Technologies 159
8.5 Parallel DBMS Vendors 160
8.5.1 Oracle 161
8.5.2 Informix 161
8.5.3 IBM 162
8.5.4 Sybase 163
8.5.5 Microsoft 164
8.5.6 Other RDBMS Products 166
8.5.7 Specialized Database Products 166
167

Chapter 9. DBMS Schemas for Decision Support 169

9.1 Data Layout for Best Access 170
9.2 Multidimensional Data Model 170
9.3 Star Schema 171
9.3.1 DBA Viewpoint 172
9.3.2 Potential Performance Problems with Star Schemas 172
9.3.3 Solutions to Performance Problems 175
9.4 STARjoin and STARindex 177
9.5 Bitmapped Indexing 179
9.5.1 SYBASE IQ 179
9.5.2 Conclusion 183
9.6 Column Local Storage 184
9.7 Complex Data Types 185

Chapter 10. Data Extraction, Cleanup, and Transformation Tools 187

10.1 Tool Requirements 187
10.2 Vendor Approaches 188
10.3 Access to Legacy Data 190
10.4 Vendor Solutions 192
10.4.1 Prism Solutions 192
10.4.2 SAS Institute 192
10.4.3 Carleton Corporation's Passport and MetaCenter 192
10.4.4 Vality Corporation 196
10.4.5 Evolutionary Technologies 197
10.4.6 Information Builders 201
10.5 Transformation Engines 201
10.5.1 Informatica 201
10.5.2 Constellar 203

Chapter 11. Metadata 205

11.1 Metadata Defined 205
11.2 Metadata Interchange Initiative 206
11.3 Metadata Repository 209
11.4 Metadata Management 211
11.5 Implementation Examples 212
11.5.1 Platinum Repository 212
11.5.2 R&O: The ROCHADE Repository 213
11.5.3 Prism Solutions 213
11.5.4 LogicWorks Universal Directory 217
11.6 Metadata Trends 219

Part 3. Business Analysis 221

Chapter 12. Reporting and Query Tools and Applications 223

12.1 Tool Categories 223
12.1.1 Reporting Tools 223
12.1.2 Managed Query Tools 225
12.1.3 Executive Information System Tools 225
12.1.4 OLAP Tools 225
12.1.5 Data Mining Tools 226
12.2 The Need for Applications 226
12.3 Cognos Impromptu 228
12.4 Applications 233
12.4.1 PowerBuilder 234
12.4.2 Forté 240
12.4.3 Information Builders 243

Chapter 13. On-Line Analytical Processing (OLAP) 247

13.1 Need for OLAP 247
13.2 Multidimensional Data Model 248
13.3 OLAP Guidelines 250
13.4 Multidimensional versus Multirelational OLAP 251
13.5 Categorization of OLAP Tools 251
13.5.1 MOLAP 252
13.5.2 ROLAP 254
13.5.3 Managed Query Environment (MQE) 255
13.6 State of the Market 256
13.6.1 Cognos PowerPlay 256
13.6.2 IBI FOCUS Fusion 259
13.6.3 Pilot Software 261
13.7 OLAP Tools and the Internet 262
13.8 Conclusion 265

Chapter 14. Patterns and Models 267

14.1 Definitions 267
14.1.1 What Is a Pattern? What Is a Model? 267
14.1.2 Visualizing a Pattern 269
14.2 A Note on Terminology 270
14.3 Where Are Models Used? 273
14.3.1 Problem 1: Selection 273
14.3.2 Problem 2: Acquisition 274
14.3.3 Problem 3: Retention 274
14.3.4 Problem 4: Extension 275
14.4 What Is the "Right" Model? 276
14.4.1 The Perfect Model 276
14.4.2 Missing Data 278
14.5 Sampling 278
14.5.1 The Necessity of Sampling 279
14.5.2 Random Sampling 280
14.6 Experimental Design 280
14.6.1 Avoiding Bias 281
14.6.2 More on Sampling 281

14.7 Computer-Intensive Statistics 283
14.7.1 Cross-validation 285
14.7.2 Jackknife and Bootstrap Resampling 286
14.8 Picking the Best Model 288

Chapter 15. Statistics 291

15.1 Data, Counting, and Probability 291
15.1.1 Histograms 292
15.1.2 Types of Categorical Predictors 295
15.1.3 Probability 296
15.1.4 Bayes' Theorem 297
15.1.5 Independence 297
15.1.6 Causality and Collinearity 300
15.1.7 Simplifying the Predictors 301
15.2 Hypothesis Testing 302
15.2.1 Hypothesis Testing on a Real-World Problem 302
15.2.2 Hypothesis Testing, *P* Values, and Alpha 306
15.2.3 Making Mistakes in Rejecting the Null Hypothesis 308
15.2.4 Degrees of Freedom 309
15.3 Contingency Tables, the Chi Square Test, and Noncausal Relationships 309
15.3.1 Contingency Tables 309
15.3.2 The Chi Square Test 310
15.3.3 Sometimes Strong Relationships Are Not Causal 311
15.4 Prediction 312
15.4.1 Linear Regression 312
15.4.2 Other Forms of Regression 313
15.5 Some Current Offerings of Statistics Tools 314
15.5.1 SAS Institute 315
15.5.2 SPSS 315
15.5.3 MathSoft 315

Chapter 16. Artificial Intelligence 317

16.1 Defining Artificial Intelligence 318
16.2 Expert Systems 320
16.3 Fuzzy Logic 323
16.4 The Rise and Fall of AI 327

Part 4. Data Mining

Chapter 17. Introduction to Data Mining 331

17.1 Data Mining Has Come of Age 333
17.2 The Motivation for Data Mining Is Tremendous 333
17.3 Learning from Your Past Mistakes 334
17.4 Data Mining? Don't Need It—I've Got Statistics 335
17.5 Measuring Data Mining Effectiveness: Accuracy, Speed, and Cost 336
17.6 Embedding Data Mining into Your Business Process 337
17.7 The More Things Change, the More They Remain the Same 339
17.8 Discovery versus Prediction 340
17.8.1 Gold in Them Thar Hills 341
341

17.8.2 Discovery—Finding Something You Weren't Looking For 341
17.8.3 Prediction 342
17.9 Overfitting 342
17.10 State of the Industry 343
17.10.1 Targeted Solutions 343
17.10.2 Business Tools 343
17.10.3 Business Analyst Tools 344
17.10.4 Research Analyst Tools 344
17.11 Comparing the Technologies 345
17.11.1 Business Score Card 346
17.11.2 Applications Score Card 347
17.11.3 Algorithmic Score Card 348

Chapter 18. Decision Trees 351

18.1 What Is a Decision Tree? 351
18.2 Business Score Card 352
18.3 Where to Use Decision Trees 353
18.3.1 Exploration 354
18.3.2 Data Preprocessing 354
18.3.3 Prediction 354
18.3.4 Applications Score Card 355
18.4 The General Idea 355
18.4.1 Growing the Tree 355
18.4.2 When Does the Tree Stop Growing? 357
18.4.3 Why Would a Decision Tree Algorithm Prevent the Tree From Growing If There Weren't Enough Data? 357
18.4.4 Decision Trees Aren't Necessarily Finished after They Are Fully Grown 358
18.4.5 Are the Splits at Each Level of the Tree Always Binary Yes/No Splits? 358
18.4.6 Picking the Best Predictors 358
18.4.7 Picking the Right Predictor Value for the Split 359
18.5 How the Decision Tree Works 361
18.5.1 Handling High-Cardinality Predictors in ID3 363
18.5.2 C4.5 Enhances ID3 365
18.5.3 CART Definition 365
18.5.4 Predictors Are Picked as They Decrease the Disorder of the Data 365
18.5.5 CART Splits Unordered Predictors by Imposing Order on Them 367
18.5.6 CART Automatically Validates the Tree 368
18.5.7 CART Surrogates Handle Missing Data 368
18.5.8 CHAID 368
18.6 Case Study: Predicting Wireless Telecommunications Churn with CART 368
18.7 Strengths and Weaknesses 371
18.7.1 Algorithm Score Card 371
18.7.2 State of the Industry 371

Chapter 19. Neural Networks 375

19.1 What Is a Neural Network? 375
19.1.1 Don't Neural Networks Learn to Make Better Predictions? 376
19.1.2 Are Neural Networks Easy to Use? 376
19.1.3 Business Score Card 377
19.2 Where to Use Neural Networks 378
19.2.1 Neural Networks for Clustering 378
19.2.2 Neural Networks for Feature Extraction 379
19.2.3 Applications Score Card 380

19.3 The General Idea
19.3.1 What Does a Neural Network Look Like?
19.3.2 How Does a Neural Network Make a Prediction?
19.3.3 How Is the Neural Network Model Created?
19.3.4 How Complex Can the Neural Network Model Become?
19.3.5 Hidden Nodes Are Like Trusted Advisors to the Output Nodes
19.3.6 Design Decisions in Architecting a Neural Network
19.3.7 Different Types of Neural Networks
19.3.8 Kohonen Feature Maps
19.3.9 How Does the Neural Network Resemble the Human Brain?
19.3.10 A Neural Network Learns to Speak
19.3.11 A Neural Network Learns to Drive
19.3.12 The Human Brain Is Still Much More Powerful
19.4 How the Neural Network Works
19.4.1 How Predictions Are Made
19.4.2 How Backpropagation Learning Works
19.4.3 Data Preparation
19.4.4 Combatting Overfitting
19.4.5 Applying and Training the Neural Network
19.4.6 Explaining the Network
19.5 Case Study: Predicting Currency Exchange Rates
19.5.1 The Problem
19.5.2 Implementation
19.5.3 The results
19.6 Strengths and Weaknessess
19.6.1 Algorithm Score Card
19.6.2 Some Current Market Offerings
19.6.3 Radial-Basis-Function Networks
19.6.4 Genetic Algorithms and Neural Networks
19.6.5 Simulated Annealing and Neural Networks

Chapter 20. Nearest Neighbor and Clustering 407

20.1 Business Score Card
20.2 Where to Use Clustering and Nearest-Neighbor Prediction
20.2.1 Clustering for Clarity
20.2.2 Clustering for Outlier Analysis
20.2.3 Nearest Neighbor for Prediction
20.2.4 Applications Score Card
20.3 The General Idea
20.3.1 There Is No Best Way to Cluster
20.3.2 How Are Tradeoffs Made When Determining Which Records Fall into Which Cluster
20.3.3 Clustering Is the Happy Medium between Homogeneous Clusters and the Lowest Number of Clusters
20.3.4 What Is the Difference between Clustering and Nearest-Neighbor Prediction?
20.3.5 What Is an *n*-Dimensional Space?
20.3.6 How Is the Space for Clustering and Nearest Neighbor Defined?
20.4 How Clustering and Nearest-Neighbor Prediction Work
20.4.1 Looking at an *n*-Dimensional Space
20.4.2 How Is "Nearness" Defined?
20.4.3 Weighting the Dimensions: Distance with a Purpose
20.4.4 Calculating Dimension Weights
20.4.5 Hierarchical and Nonhierarchical Clustering

20.4.6 Nearest-Neighbor Prediction 429
20.4.7 K Nearest Neighbors—Voting Is Better 431
20.4.8 Generalizing the Solution: Prototypes and Sentries 432
20.5 Case Study: Image Recognition for Human Handwriting 433
20.5.1 The Problem 433
20.5.2 Solution Using Nearest-Neighbor Techniques 434
20.6 Strengths and Weaknessess 435
20.6.1 Algorithm Score Card 437
20.6.2 Predicting Future Trends 438

Chapter 21. Genetic Algorithms 441

21.1 What Are Genetic Algorithms? 442
21.1.1 How Do They Relate to Evolution? 443
21.1.2 Genetic Algorithms, Artificial Life, and Simulated Evolution 444
21.1.3 How Can They Be Used in Business? 445
21.1.4 Business Score Card 445
21.2 Where to Use Genetic Algorithms 446
21.2.1 Genetic Algorithms for Optimization 447
21.2.2 Genetic Algorithms for Data Mining 448
21.2.3 Applications Score Card 448
21.3 The General Idea 449
21.3.1 Do Genetic Algorithms Guess the Right Answer? 449
21.3.2 Are Genetic Algorithms Fully Automated? 450
21.3.3 Cost Minimization: Traveling Salesperson 450
21.3.4 Cooperation Strategies: Prisoner's Dilemma 451
21.4 How the Genetic Algorithm Works 452
21.4.1 The Overall Process 452
21.4.2 Survival of the Fittest 453
21.4.3 Mutation 455
21.4.4 Sexual Reproduction and Crossover 456
21.4.5 Exploration versus Exploitation 457
21.4.6 The Schema Theorem 458
21.4.7 Epistasis 459
21.4.8 Classifier Systems 459
21.4.9 Remaining Challenges 461
21.4.10 Sharing: A Solution to Premature Convergence 461
21.4.11 Metalevel Evolution: The Automation of Parameter Choice 462
21.4.12 Parallel Implementation 463
21.5 Case Study: Optimizing Predictive Customer Segments 463
21.6 Strengths and Weaknessess 465
21.6.1 Algorithm Score Card 465
21.6.2 State of the Marketplace 466
21.6.3 Predicting Future Trends 466

Chapter 22. Rule Induction 469

22.1 Business Score Card 470
22.2 Where to Use Rule Induction 471
22.2.1 What Is a Rule? 471
22.2.2 What to Do with a Rule 472
22.2.3 Caveat: Rules Do Not Imply Causality 473
22.2.4 Types of Databases Used for Rule Induction 473

22.2.5 Discovery 474
22.2.6 Prediction 475
22.2.7 Applications Score Card 475
22.3 The General Idea 476
22.3.1 How to Evaluate the Rule 476
22.3.2 Conjunctions and Disjunctions 477
22.3.3 Defining "Interestingness" 479
22.3.4 Other Measures of Usefulness 480
22.3.5 Rules versus Decision Trees 481
22.4 How Rule Induction Works 482
22.4.1 Constructing Rules 484
22.4.2 A Brute-Force Algorithm for Generating Rules 484
22.4.3 Combining Evidence 487
22.5 Case Study: Classifying U.S. Census Returns 488
22.6 Strengths and Weaknesses 488
22.7 Current Offerings and Future Improvements 490
492

Chapter 23. Selecting and Using the Right Technique 493

23.1 Using the Right Technique 493
23.1.1 The Data Mining Process 493
23.1.2 What All the Data Mining Techniques Have in Common 494
23.1.3 Cases in Which Decision Trees Are Like Nearest Neighbors 496
23.1.4 Rule Induction Is Like Decision Trees 496
23.1.5 Could You Do Link Analysis with a Neural Network? 497
23.2 Data Mining in the Business Process 498
23.2.1 Avoiding Some Big Mistakes in Data Mining 499
23.2.2 Understanding the Data 499
23.3 The Case for Embedded Data Mining 503
23.3.1 The Cost of a Distributed Business Process 504
23.3.2 The Best Way to Measure a Data Mining Tool 506
23.3.3 The Case for Embedded Data Mining 508
23.4 How to Measure Accuracy, Explanation, and Integration 510
23.4.1 Measuring Accuracy 510
23.4.2 Measuring Explanation 512
23.4.3 Measuring Integration 512
23.5 What the Future Holds for Embedded Data Mining 513

Part 5. Data Visualization and Overall Perspective 515

Chapter 24. Data Visualization 517

24.1 Data Visualization Principles 518
24.2 Parallel Coordinates 520
24.3 Visualizing Neural Networks 520
24.4 Visualization of Trees 521
24.5 State of the Industry 522
24.5.1 Advanced Visual Systems 522
24.5.2 Alta Analytics 522
24.5.3 Business Objects 524
24.5.4 IBM 525
24.5.5 Pilot Software 526
24.5.6 Silicon Graphics 531

Chapter 25. Putting It All Together 533

25.1 Design for Scalability 534
25.2 Data Quality 536
25.3 Implementation Notes 538
25.3.1 Operational Data Stores 538
25.3.2 Data Marts 539
25.3.3 Star Schema 540
25.4 Making the Most of Your Warehouse 540
25.5 The Data Warehousing Market 541
25.6 Costs and Benefits 544
25.6.1 Big Data—Bigger Returns 546
25.6.2 Law of Diminishing Returns 548
25.7 A Unifying View of Business Information 549
25.8 What's Next 551
25.8.1 Distributed Warehouse Environments 551
25.8.2 Using the Internet or Intranet for Information Delivery 552
25.8.3 Object-Relational Databases 556
25.8.4 Very Large Databases (VLDBs) 558
25.9 Conclusion 559

Appendix A. Glossary 561

Appendix B. Big Data—Better Returns: Leveraging Your Hidden Data Assets to Improve ROI 579

Appendix C. Dr. E. F. Codd's 12 Guidelines for OLAP 597

Appendix D. 10 Mistakes for Data Warehousing Managers to Avoid 599

Bibliography 605
Index 609

Alex Berson
Stephen J. Smith

Data Warehousing, Data Mining, & OLAP

TATA McGRAW-HILL EDITION

Data Warehousing/Data Management

Optimize your organization's data delivery system!

Improving data delivery is a top priority in business computing today. This comprehensive, cutting-edge guide can help—by showing you how to effectively integrate data mining and other powerful data warehousing technologies.

You'll learn how to :

- Use data warehousing to establish a competitive advantage
- Solve business problems faster by exploiting online analytical processing (OLAP)
- Evaluate various data warehousing solutions (including SMP and MPP, parallel database management systems, metadata, OLAP, etc.)
- Leverage your data warehousing utility via the Internet, client/server computing, and various data mining tools

In addition to providing a detailed overview and strategic analysis of the available data warehousing technologies, the book serves as a practical guide to data warehouse database design, star and snowflake schema approaches, multidimensional and multirelational models, advanced indexing techniques, and data mining. You'll also learn how to compare different data mine technologies and products, and understand how they fit into your overall business and data processes. Intended for IS professionals as well as strategic planners, this fascinating book can be well relied upon as the essential reference to the standards, tools, technologies—and possibilities—of data warehousing today.

ABOUT THE AUTHORS

Alex Berson is a senior information technology architect with over 20 years of experience in various areas of information technology including distributed client/server computing, database systems, parallel computing systems, object technology, data communications, and machine learning. He has successfully designed and implemented several large-scale data warehousing projects for major financial services companies. He is the author of several best-selling Mc Graw-Hill books, including *Client/Server Architecture/2E and Sybase and Client/Server Computing.*

Stephen J. Smith is a director and architect responsible for the creation and delivery of two data mining products over the last decade - one for parallel supercomputers and the second for data warehouses with multidimensional databases. He is a well-respected expert in the field of data mining and their integration with the data warehouse.

The McGraw-Hill Companies

Tata McGraw-Hill
Publishing Company Limited
7 West Patel Nagar, New Delhi 110 008

Visit our website at : www.tatamcgrawhill.com

ISBN 0-07-058741-8