# DATA WAREHOUSING FUNDAMENTALS FOR IT PROFESSIONALS

Second Edition

**PAULRAJ PONNIAH**

A JOHN WILEY & SONS, INC., PUBLICATION

# DATA WAREHOUSING FUNDAMENTALS FOR IT PROFESSIONALS

# DATA WAREHOUSING FUNDAMENTALS FOR IT PROFESSIONALS

Second Edition

**PAULRAJ PONNIAH**

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2010 by John Wiley & Sons, Inc. All rights reserved

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

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, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

***Library of Congress Cataloging-in-Publication Data:***

Ponniah, Paulraj.
Data warehousing fundamentals for IT professionals / Paulraj Ponniah.—2nd ed.
p. cm.
Previous ed. published under title: Data warehousing fundamentals.
Includes bibliographical references and index.
ISBN 978-0-470-46207-2 (cloth)
1. Data warehousing. I. Ponniah, Paulraj. Data warehousing fundamentals. II. Title.
QA76.9.D37P66 2010
005.74′5—dc22

2009041789

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To
Vimala, my loving wife
and to
Joseph, David, and Shobi,
my dear children

# CONTENTS

**PREFACE** **xxv**

PART 1 OVERVIEW AND CONCEPTS 1

**1 THE COMPELLING NEED FOR DATA WAREHOUSING** **3**

CHAPTER OBJECTIVES / 3
ESCALATING NEED FOR STRATEGIC INFORMATION / 4
The Information Crisis / 6
Technology Trends / 6
Opportunities and Risks / 8
FAILURES OF PAST DECISION-SUPPORT SYSTEMS / 9
History of Decision-Support Systems / 10
Inability to Provide Information / 10
OPERATIONAL VERSUS DECISION-SUPPORT SYSTEMS / 11
Making the Wheels of Business Turn / 12
Watching the Wheels of Business Turn / 12
Different Scope, Different Purposes / 12
DATA WAREHOUSING—THE ONLY VIABLE SOLUTION / 13
A New Type of System Environment / 13
Processing Requirements in the New Environment / 14
Strategic Information from the Data Warehouse / 14

DATA WAREHOUSE DEFINED / 15
A Simple Concept for Information Delivery / 15
An Environment, Not a Product / 15
A Blend of Many Technologies / 16
THE DATA WAREHOUSING MOVEMENT / 17
Data Warehousing Milestones / 17
Initial Challenges / 18
EVOLUTION OF BUSINESS INTELLIGENCE / 18
BI: Two Environments / 19
BI: Data Warehousing and Analytics / 19
CHAPTER SUMMARY / 20
REVIEW QUESTIONS / 20
EXERCISES / 21

**2 DATA WAREHOUSE: THE BUILDING BLOCKS 23**

CHAPTER OBJECTIVES / 23
DEFINING FEATURES / 24
Subject-Oriented Data / 24
Integrated Data / 25
Time-Variant Data / 26
Nonvolatile Data / 27
Data Granularity / 28
DATA WAREHOUSES AND DATA MARTS / 29
How Are They Different? / 29
Top-Down Versus Bottom-Up Approach / 29
A Practical Approach / 31
ARCHITECTURAL TYPES / 32
Centralized Data Warehouse / 32
Independent Data Marts / 32
Federated / 33
Hub-and-Spoke / 33
Data-Mart Bus / 34
OVERVIEW OF THE COMPONENTS / 34
Source Data Component / 34
Data Staging Component / 37
Data Storage Component / 39
Information Delivery Component / 40
Metadata Component / 41
Management and Control Component / 41

METADATA IN THE DATA WAREHOUSE / 41
Types of Metadata / 42
Special Significance / 42
CHAPTER SUMMARY / 42
REVIEW QUESTIONS / 43
EXERCISES / 43

**3 TRENDS IN DATA WAREHOUSING 45**

CHAPTER OBJECTIVES / 45
CONTINUED GROWTH IN DATA WAREHOUSING / 46
Data Warehousing has Become Mainstream / 46
Data Warehouse Expansion / 47
Vendor Solutions and Products / 48
SIGNIFICANT TRENDS / 50
Real-Time Data Warehousing / 50
Multiple Data Types / 50
Data Visualization / 52
Parallel Processing / 54
Data Warehouse Appliances / 56
Query Tools / 56
Browser Tools / 57
Data Fusion / 57
Data Integration / 58
Analytics / 59
Agent Technology / 59
Syndicated Data / 60
Data Warehousing and ERP / 60
Data Warehousing and KM / 61
Data Warehousing and CRM / 63
Agile Development / 63
Active Data Warehousing / 64
EMERGENCE OF STANDARDS / 64
Metadata / 65
OLAP / 65
WEB-ENABLED DATA WAREHOUSE / 66
The Warehouse to the Web / 67
The Web to the Warehouse / 67
The Web-Enabled Configuration / 69
CHAPTER SUMMARY / 69

REVIEW QUESTIONS / 69
EXERCISES / 70

PART 2 PLANNING AND REQUIREMENTS 71

**4 PLANNING AND PROJECT MANAGEMENT 73**

CHAPTER OBJECTIVES / 73
PLANNING YOUR DATA WAREHOUSE / 74
Key Issues / 74
Business Requirements, Not Technology / 76
Top Management Support / 77
Justifying Your Data Warehouse / 77
The Overall Plan / 78
THE DATA WAREHOUSE PROJECT / 79
How is it Different? / 79
Assessment of Readiness / 81
The Life-Cycle Approach / 81
THE DEVELOPMENT PHASES / 83
Adopting Agile Development / 84
THE PROJECT TEAM / 85
Organizing the Project Team / 85
Roles and Responsibilities / 86
Skills and Experience Levels / 87
User Participation / 88
PROJECT MANAGEMENT CONSIDERATIONS / 90
Guiding Principles / 91
Warning Signs / 92
Success Factors / 92
Anatomy of a Successful Project / 93
Adopt a Practical Approach / 94
CHAPTER SUMMARY / 96
REVIEW QUESTIONS / 96
EXERCISES / 97

**5 DEFINING THE BUSINESS REQUIREMENTS 99**

CHAPTER OBJECTIVES / 99
DIMENSIONAL ANALYSIS / 100
Usage of Information Unpredictable / 100
Dimensional Nature of Business Data / 101
Examples of Business Dimensions / 102

INFORMATION PACKAGES—A USEFUL CONCEPT / 103
Requirements Not Fully Determinate / 104
Business Dimensions / 105
Dimension Hierarchies and Categories / 106
Key Business Metrics or Facts / 107
REQUIREMENTS GATHERING METHODS / 109
Types of Questions / 110
Arrangement of Questions / 111
Interview Techniques / 111
Adapting the JAD Methodology / 113
Using Questionnaires / 115
Review of Existing Documentation / 115
REQUIREMENTS DEFINITION: SCOPE AND CONTENT / 116
Data Sources / 117
Data Transformation / 117
Data Storage / 117
Information Delivery / 118
Information Package Diagrams / 118
Requirements Definition Document Outline / 118
CHAPTER SUMMARY / 119
REVIEW QUESTIONS / 119
EXERCISES / 120

**6 REQUIREMENTS AS THE DRIVING FORCE FOR DATA WAREHOUSING 121**

CHAPTER OBJECTIVES / 121
DATA DESIGN / 122
Structure for Business Dimensions / 123
Structure for Key Measurements / 124
Levels of Detail / 125
THE ARCHITECTURAL PLAN / 125
Composition of the Components / 126
Special Considerations / 127
Tools and Products / 129
DATA STORAGE SPECIFICATIONS / 131
DBMS Selection / 132
Storage Sizing / 132
INFORMATION DELIVERY STRATEGY / 133
Queries and Reports / 134
Types of Analysis / 134
Information Distribution / 135

Real Time Information Delivery / 135
Decision Support Applications / 135
Growth and Expansion / 136
CHAPTER SUMMARY / 136
REVIEW QUESTIONS / 136
EXERCISES / 137

PART 3 ARCHITECTURE AND INFRASTRUCTURE 139

**7 ARCHITECTURAL COMPONENTS 141**

CHAPTER OBJECTIVES / 141
UNDERSTANDING DATA WAREHOUSE ARCHITECTURE / 141
Architecture: Definitions / 142
Architecture in Three Major Areas / 142
DISTINGUISHING CHARACTERISTICS / 143
Different Objectives and Scope / 144
Data Content / 144
Complex Analysis and Quick Response / 145
Flexible and Dynamic / 145
Metadata-Driven / 146
ARCHITECTURAL FRAMEWORK / 146
Architecture Supporting Flow of Data / 146
The Management and Control Module / 147
TECHNICAL ARCHITECTURE / 148
Data Acquisition / 149
Data Storage / 152
Information Delivery / 154
ARCHITECTURAL TYPES / 156
Centralized Corporate Data Warehouse / 156
Independent Data Marts / 156
Federated / 159
Hub-and-Spoke / 159
Data-Mart Bus / 160
CHAPTER SUMMARY / 160
REVIEW QUESTIONS / 160
EXERCISES / 161

**8 INFRASTRUCTURE AS THE FOUNDATION FOR DATA WAREHOUSING 163**

CHAPTER OBJECTIVES / 163

INFRASTRUCTURE SUPPORTING ARCHITECTURE / 164
Operational Infrastructure / 165
Physical Infrastructure / 165
HARDWARE AND OPERATING SYSTEMS / 166
Mainframes / 167
Open System Servers / 168
NT Servers / 168
Platform Options / 168
Server Hardware / 177
DATABASE SOFTWARE / 181
Parallel Processing Options / 182
Selection of the DBMS / 184
COLLECTION OF TOOLS / 184
Architecture First, Then Tools / 186
Data Modeling / 186
Data Extraction / 187
Data Transformation / 187
Data Loading / 187
Data Quality / 187
Queries and Reports / 187
Dashboards / 187
Scorecards / 187
Online Analytical Processing (OLAP) / 188
Alert Systems / 188
Middleware and Connectivity / 188
Data Warehouse Administration / 188
DATA WAREHOUSE APPLIANCES / 188
Evolution of DW Appliances / 189
Benefits of DW Appliances / 190
CHAPTER SUMMARY / 191
REVIEW QUESTIONS / 191
EXERCISES / 192

**9 THE SIGNIFICANT ROLE OF METADATA 193**

CHAPTER OBJECTIVES / 193
WHY METADATA IS IMPORTANT / 193
A Critical Need in the Data Warehouse / 195
Why Metadata Is Vital for End-Users / 198
Why Metadata Is Essential for IT / 199
Automation of Warehousing Tasks / 200
Establishing the Context of Information / 202

METADATA TYPES BY FUNCTIONAL AREAS / 203
Data Acquisition / 204
Data Storage / 205
Information Delivery / 206
BUSINESS METADATA / 207
Content Overview / 207
Examples of Business Metadata / 208
Content Highlights / 209
Who Benefits? / 209
TECHNICAL METADATA / 209
Content Overview / 210
Examples of Technical Metadata / 210
Content Highlights / 211
Who Benefits? / 211
HOW TO PROVIDE METADATA / 212
Metadata Requirements / 212
Sources of Metadata / 214
Challenges for Metadata Management / 215
Metadata Repository / 215
Metadata Integration and Standards / 217
Implementation Options / 218
CHAPTER SUMMARY / 219
REVIEW QUESTIONS / 220
EXERCISES / 220

PART 4 DATA DESIGN AND DATA PREPARATION 223

**10 PRINCIPLES OF DIMENSIONAL MODELING 225**

CHAPTER OBJECTIVES / 225
FROM REQUIREMENTS TO DATA DESIGN / 225
Design Decisions / 226
Dimensional Modeling Basics / 226
E-R Modeling Versus Dimensional Modeling / 230
Use of CASE Tools / 232
THE STAR SCHEMA / 232
Review of a Simple STAR Schema / 232
Inside a Dimension Table / 234
Inside the Fact Table / 236
The Factless Fact Table / 238
Data Granularity / 238

STAR SCHEMA KEYS / 239
Primary Keys / 239
Surrogate Keys / 240
Foreign Keys / 240
ADVANTAGES OF THE STAR SCHEMA / 241
Easy for Users to Understand / 241
Optimizes Navigation / 242
Most Suitable for Query Processing / 243
STARjoin and STARindex / 244
STAR SCHEMA: EXAMPLES / 244
Video Rental / 244
Supermarket / 244
Wireless Phone Service / 244
Auction Company / 244
CHAPTER SUMMARY / 246
REVIEW QUESTIONS / 247
EXERCISES / 247

**11 DIMENSIONAL MODELING: ADVANCED TOPICS 249**

CHAPTER OBJECTIVES / 249
UPDATES TO THE DIMENSION TABLES / 250
Slowly Changing Dimensions / 250
Type 1 Changes: Correction of Errors / 251
Type 2 Changes: Preservation of History / 252
Type 3 Changes: Tentative Soft Revisions / 253
MISCELLANEOUS DIMENSIONS / 255
Large Dimensions / 255
Rapidly Changing Dimensions / 256
Junk Dimensions / 258
THE SNOWFLAKE SCHEMA / 259
Options to Normalize / 259
Advantages and Disadvantages / 260
When to Snowflake / 262
AGGREGATE FACT TABLES / 262
Fact Table Sizes / 264
Need for Aggregates / 266
Aggregating Fact Tables / 266
Aggregation Options / 271
FAMILIES OF STARS / 272
Snapshot and Transaction Tables / 273
Core and Custom Tables / 274

Supporting Enterprise Value Chain or Value Circle / 274
Conforming Dimensions / 275
Standardizing Facts / 276
Summary of Family of STARS / 277
CHAPTER SUMMARY / 277
REVIEW QUESTIONS / 278
EXERCISES / 278

**12 DATA EXTRACTION, TRANSFORMATION, AND LOADING 281**

CHAPTER OBJECTIVES / 281
ETL OVERVIEW / 282
Most Important and Most Challenging / 282
Time Consuming and Arduous / 283
ETL REQUIREMENTS AND STEPS / 284
Key Factors / 285
DATA EXTRACTION / 286
Source Identification / 287
Data Extraction Techniques / 287
Evaluation of the Techniques / 294
DATA TRANSFORMATION / 295
Data Transformation: Basic Tasks / 296
Major Transformation Types / 297
Data Integration and Consolidation / 299
Transformation for Dimension Attributes / 301
How to Implement Transformation / 301
DATA LOADING / 302
Applying Data: Techniques and Processes / 303
Data Refresh Versus Update / 306
Procedure for Dimension Tables / 306
Fact Tables: History and Incremental Loads / 307
ETL SUMMARY / 308
ETL Tool Options / 308
Reemphasizing ETL Metadata / 309
ETL Summary and Approach / 310
OTHER INTEGRATION APPROACHES / 311
Enterprise Information Integration (EII) / 311
Enterprise Application Integration (EAI) / 312
CHAPTER SUMMARY / 313
REVIEW QUESTIONS / 313
EXERCISES / 314

**13 DATA QUALITY: A KEY TO SUCCESS 315**

CHAPTER OBJECTIVES / 315
WHY IS DATA QUALITY CRITICAL? / 316
What Is Data Quality? / 316
Benefits of Improved Data Quality / 319
Types of Data Quality Problems / 320
DATA QUALITY CHALLENGES / 323
Sources of Data Pollution / 323
Validation of Names and Addresses / 325
Costs of Poor Data Quality / 325
DATA QUALITY TOOLS / 326
Categories of Data Cleansing Tools / 327
Error Discovery Features / 327
Data Correction Features / 327
The DBMS for Quality Control / 327
DATA QUALITY INITIATIVE / 328
Data Cleansing Decisions / 329
Who Should Be Responsible? / 330
The Purification Process / 333
Practical Tips on Data Quality / 334
MASTER DATA MANAGEMENT (MDM) / 335
MDM Categories / 335
MDM Benefits / 335
MDM and Data Warehousing / 336
CHAPTER SUMMARY / 336
REVIEW QUESTIONS / 336
EXERCISES / 337

PART 5 INFORMATION ACCESS AND DELIVERY 339

**14 MATCHING INFORMATION TO THE CLASSES OF USERS 341**

CHAPTER OBJECTIVES / 341
INFORMATION FROM THE DATA WAREHOUSE / 342
Data Warehouse Versus Operational Systems / 342
Information Potential / 344
User–Information Interface / 347
Industry Applications / 348
WHO WILL USE THE INFORMATION? / 349
Classes of Users / 349

What They Need / 352
How to Provide Information / 354
INFORMATION DELIVERY / 356
Queries / 357
Reports / 358
Analysis / 359
Applications / 359
INFORMATION DELIVERY TOOLS / 360
The Desktop Environment / 360
Methodology for Tool Selection / 361
Tool Selection Criteria / 364
Information Delivery Framework / 365
INFORMATION DELIVERY: SPECIAL TOPICS / 366
Business Activity Monitoring (BAM) / 366
Dashboards and Scorecards / 367
CHAPTER SUMMARY / 371
REVIEW QUESTIONS / 371
EXERCISES / 372

**15 OLAP IN THE DATA WAREHOUSE 373**

CHAPTER OBJECTIVES / 373
DEMAND FOR ONLINE ANALYTICAL PROCESSING / 374
Need for Multidimensional Analysis / 374
Fast Access and Powerful Calculations / 375
Limitations of Other Analysis Methods / 377
OLAP is the Answer / 379
OLAP Definitions and Rules / 379
OLAP Characteristics / 382
MAJOR FEATURES AND FUNCTIONS / 382
General Features / 383
Dimensional Analysis / 383
What Are Hypercubes? / 386
Drill Down and Roll Up / 390
Slice and Dice or Rotation / 392
Uses and Benefits / 393
OLAP MODELS / 393
Overview of Variations / 394
The MOLAP Model / 394
The ROLAP Model / 395
ROLAP Versus MOLAP / 397

OLAP IMPLEMENTATION CONSIDERATIONS / 398
Data Design and Preparation / 399
Administration and Performance / 401
OLAP Platforms / 402
OLAP Tools and Products / 402
Implementation Steps / 403
Examples of Typical Implementations / 404
CHAPTER SUMMARY / 404
REVIEW QUESTIONS / 405
EXERCISES / 405

**16 DATA WAREHOUSING AND THE WEB 407**

CHAPTER OBJECTIVES / 407
WEB-ENABLED DATA WAREHOUSE / 408
Why the Web? / 408
Convergence of Technologies / 410
Adapting the Data Warehouse for the Web / 411
The Web as a Data Source / 412
Clickstream Analysis / 413
WEB-BASED INFORMATION DELIVERY / 414
Expanded Usage / 414
New Information Strategies / 416
Browser Technology for the Data Warehouse / 418
Security Issues / 419
OLAP AND THE WEB / 420
Enterprise OLAP / 420
Web-OLAP Approaches / 420
OLAP Engine Design / 421
BUILDING A WEB-ENABLED DATA WAREHOUSE / 421
Nature of the Data Webhouse / 422
Implementation Considerations / 423
Putting the Pieces Together / 424
Web Processing Model / 426
CHAPTER SUMMARY / 426
REVIEW QUESTIONS / 426
EXERCISES / 427

**17 DATA MINING BASICS 429**

CHAPTER OBJECTIVES / 429
WHAT IS DATA MINING? / 430

Data Mining Defined / 431
The Knowledge Discovery Process / 432
OLAP Versus Data Mining / 435
Some Aspects of Data Mining / 436
Data Mining and the Data Warehouse / 438
MAJOR DATA MINING TECHNIQUES / 439
Cluster Detection / 440
Decision Trees / 443
Memory-Based Reasoning / 444
Link Analysis / 445
Neural Networks / 447
Genetic Algorithms / 448
Moving into Data Mining / 450
DATA MINING APPLICATIONS / 452
Benefits of Data Mining / 453
Applications in CRM (Customer Relationship Management) / 454
Applications in the Retail Industry / 455
Applications in the Telecommunications Industry / 456
Applications in Biotechnology / 457
Applications in Banking and Finance / 459
CHAPTER SUMMARY / 459
REVIEW QUESTIONS / 459
EXERCISES / 460

PART 6 IMPLEMENTATION AND MAINTENANCE 461

**18 THE PHYSICAL DESIGN PROCESS 463**

CHAPTER OBJECTIVES / 463
PHYSICAL DESIGN STEPS / 464
Develop Standards / 464
Create Aggregates Plan / 465
Determine the Data Partitioning Scheme / 465
Establish Clustering Options / 466
Prepare an Indexing Strategy / 466
Assign Storage Structures / 466
Complete Physical Model / 467
PHYSICAL DESIGN CONSIDERATIONS / 467
Physical Design Objectives / 467
From Logical Model to Physical Model / 469

Physical Model Components / 469
Significance of Standards / 470
PHYSICAL STORAGE / 473
Storage Area Data Structures / 473
Optimizing Storage / 473
Using RAID Technology / 476
Estimating Storage Sizes / 477
INDEXING THE DATA WAREHOUSE / 477
Indexing Overview / 477
B-Tree Index / 479
Bitmapped Index / 481
Clustered Indexes / 482
Indexing the Fact Table / 482
Indexing the Dimension Tables / 483
PERFORMANCE ENHANCEMENT TECHNIQUES / 483
Data Partitioning / 483
Data Clustering / 484
Parallel Processing / 484
Summary Levels / 485
Referential Integrity Checks / 485
Initialization Parameters / 485
Data Arrays / 486
CHAPTER SUMMARY / 486
REVIEW QUESTIONS / 486
EXERCISES / 487

**19 DATA WAREHOUSE DEPLOYMENT 489**

CHAPTER OBJECTIVES / 489
DATA WAREHOUSE TESTING / 490
Front-End / 490
ETL Testing / 490
MAJOR DEPLOYMENT ACTIVITIES / 491
Complete User Acceptance / 491
Perform Initial Loads / 492
Get User Desktops Ready / 493
Complete Initial User Training / 494
Institute Initial User Support / 495
Deploy in Stages / 495
CONSIDERATIONS FOR A PILOT / 497
When is a Pilot Data Mart Useful? / 497

Types of Pilot Projects / 498
Choosing the Pilot / 500
Expanding and Integrating the Pilot / 501
SECURITY / 502
Security Policy / 502
Managing User Privileges / 502
Password Considerations / 503
Security Tools / 504
BACKUP AND RECOVERY / 504
Why Back Up the Data Warehouse? / 505
Backup Strategy / 505
Setting up a Practical Schedule / 506
Recovery / 507
CHAPTER SUMMARY / 508
REVIEW QUESTIONS / 508
EXERCISES / 509

**20 GROWTH AND MAINTENANCE 511**

CHAPTER OBJECTIVES / 511
MONITORING THE DATA WAREHOUSE / 512
Collection of Statistics / 512
Using Statistics for Growth Planning / 514
Using Statistics for Fine-Tuning / 514
Publishing Trends for Users / 515
USER TRAINING AND SUPPORT / 515
User Training Content / 516
Preparing the Training Program / 516
Delivering the Training Program / 518
User Support / 519
MANAGING THE DATA WAREHOUSE / 520
Platform Upgrades / 521
Managing Data Growth / 521
Storage Management / 522
ETL Management / 522
Data Model Revisions / 523
Information Delivery Enhancements / 523
Ongoing Fine-Tuning / 524
CHAPTER SUMMARY / 524
REVIEW QUESTIONS / 525
EXERCISES / 525

**ANSWERS TO SELECTED EXERCISES** 527

**APPENDIX A: PROJECT LIFE CYCLE STEPS AND CHECKLISTS** 531

**APPENDIX B: CRITICAL FACTORS FOR SUCCESS** 535

**APPENDIX C: GUIDELINES FOR EVALUATING VENDOR SOLUTIONS** 537

**APPENDIX D: HIGHLIGHTS OF VENDORS AND PRODUCTS** 539

**APPENDIX E: REAL-WORLD EXAMPLES OF BEST PRACTICES** 549

**REFERENCES** 555

**GLOSSARY** 557

INDEX 565

# PREFACE

## THIS BOOK IS FOR YOU

Are you an information technology professional watching, with great interest, the massive unfolding and spreading of the data warehouse movement during the past decade? Are you contemplating a move into this fast-growing area of opportunity? Are you a systems analyst, programmer, data analyst, database administrator, project leader, or software engineer eager to grasp the fundamentals of data warehousing? Do you wonder how many different books you may have to study to learn the underlying principles and the current practices? Are you lost in the maze of the literature and products on the subject? Do you wish for a single publication on data warehousing, clearly and specifically designed for IT professionals? Do you need a textbook that helps you learn the fundamentals in sufficient depth? If you answered "yes" to any of the above, this book is written specially for you.

This is the *one* definitive book on data warehousing clearly intended for IT professionals. The organization and presentation of the book are specially tuned for IT professionals. This book does not presume to target anyone and everyone remotely interested in the subject for some reason or another, but is written to address the specific needs of IT professionals like you. It does not tend to emphasize certain aspects and neglect other critical ones. The book takes you over the entire spectrum of data warehousing.

As a veteran IT professional with wide and intensive industry experience, as a successful database and data warehousing consultant for many years, and as one who teaches data warehousing fundamentals in the college classroom and at public seminars, I have come to appreciate the precise needs of IT professionals. In every chapter I have incorporated these requirements of the IT community.

## THE SCENARIO

Why have companies rushed into data warehousing? Why is there a tremendous surge in interest? Data warehousing is no longer a purely novel idea just for research and experimentation. It has become a mainstream phenomenon. True, the data warehouse is not in every doctor's office yet, but neither is it confined to only high-end businesses. More than half of all U.S. companies and a large percentage of worldwide businesses have made a commitment to data warehousing.

In every industry across the board, from retail chain stores to financial institutions, from manufacturing enterprises to government departments, and from airline companies to utility businesses, data warehousing has revolutionized the way people perform business analysis and make strategic decisions. Every company that has a data warehouse is realizing the enormous benefits translated into positive results at the bottom line. These companies, now incorporating Web-based technologies, are enhancing the potential for greater and easier delivery of vital information.

Over the past decade, a large number of vendors have flooded the market with numerous data warehousing products. Vendor solutions and products run the gamut of data warehousing and business intelligence—data modeling, data acquisition, data quality, data analysis, metadata, information delivery, and so on. The market is large, mature, and continues to grow.

## CHANGED ROLE OF IT

In this scenario, information technology departments of all progressive companies have perceived a radical change in their roles. IT is no longer required to create every report and present every screen for providing information to the end-users. IT is now charged with the building of information delivery systems and letting the end-users themselves retrieve information in innovative ways for analysis and decision making. Data warehousing and business intelligence environments are proving to be just that type of successful information delivery system.

IT professionals responsible for building data warehouses had to revise their mindsets about building applications. They had to understand that a data warehouse is not a one-size-fits-all proposition. First, they had to get a clear understanding about data extraction from source systems, data transformations, data staging, data warehouse architecture, infrastructure, and the various methods of information delivery. In short, IT professionals, like you, must get a strong grip on the fundamentals of data warehousing.

## WHAT THIS BOOK CAN DO FOR YOU

The book is comprehensive and detailed. You will be able to study every significant topic in planning, requirements, architecture, infrastructure, design, data preparation, information delivery, deployment, and maintenance. The book is specially designed for IT professionals; you will be able to follow the presentation easily because it is built upon the foundation of your background as an IT professional, your knowledge, and the technical terminology familiar to you. It is organized logically, beginning with an overview of concepts, moving on to planning and requirements, then to architecture and infrastructure, on to data design, then to

information delivery, and concluding with deployment and maintenance. This progression is typical of what you are most familiar with in your IT experience and day-to-day work.

The book provides an interactive learning experience. It is not just a one-way lecture. You participate through the review questions and exercises at the end of each chapter. For each chapter, the objectives at the beginning set the theme and the summary at the end highlights the topics covered. You can relate each concept and technique presented in the book to the data warehousing industry and marketplace. You will benefit from the substantial number of industry examples. Although intended as a first course on the fundamentals, this book provides sufficient coverage of each topic so that you can comfortably proceed to the next step of specialization for specific roles in a data warehouse project.

Featuring all the significant topics in appropriate measure, this book is eminently suitable as a textbook for serious self-study, a college course, or a seminar on the essentials. It provides an opportunity for you to become a data warehouse expert.

## ENHANCEMENTS IN THIS SECOND EDITION

This greatly enhanced edition captures the developments and changes in the data warehousing landscape during the past nearly ten years. The underlying purposes and principles of data warehousing have remained the same. However, we notice definitive changes in the details, some finer aspects, and in product innovations. Although this edition succeeds in incorporating all the significant revisions, I have been careful not to disturb the overall logical arrangement and sequencing of the chapters.

The term "business intelligence" has gained a lot more currency. Many practitioners now consider data warehousing to refer to populating the warehouse with data, and business intelligence to refer to using the warehouse data. Data warehousing has made inroads into areas such as Customer Relationship Management, Enterprise Application Integration, Enterprise Information Integration, Business Activity Monitoring, and so on. The size of corporate data warehouses has been rising higher and higher. Some progressive businesses have reaped enormous benefits from data warehouses that are almost in the 500 terabyte range (five times the size of the U.S. Library of Congress archive). The benefits from data warehouses are no longer limited to a selected core of executives, managers, and analysts. Pervasive data warehousing has become the operative principle, providing access and usage to staff at multiple levels. Information delivery through traditional reports and queries is being replaced by interactive dashboards and scorecards.

More specifically, among topics on recent trends and changes, this enhanced edition includes the following:

- Evolution of business intelligence
- Real-time business intelligence
- Data warehouse appliances
- Data warehouse: architectural types
- Data visualization enhancements
- Enterprise application integration (EAI)
- Enterprise information integration (EII)
- Agile data warehouse development

- Data warehousing and KM (knowledge management)
- Data warehousing and ERP (enterprise resource planning)
- Data warehousing and CRM (customer relationship management)
- Improved requirements gathering methods
- Business activity monitoring (BAM)
- Interactive information delivery through dashboards and scorecards
- Additional STAR schema examples
- Master data management
- Examples of typical OLAP (online analytical processing) implementations
- Data mining applications
- Web clickstream analysis
- Highlights of vendors and products
- Real-world examples of best practices

## ACKNOWLEDGMENTS

I wish to acknowledge my indebtedness and to express my gratitude to the authors listed in the reference section at the end of the book. Their insights and observations have helped me cover every topic adequately.

I must also express my appreciation to my students and professional colleagues. My interactions with them have enabled me to shape this textbook according to the needs of IT professionals.

My special thanks are due to the wonderful staff and editors at Wiley, my publishers, who have worked with me and supported me for more than a decade in the publication and promotion of my books.

PAULRAJ PONNIAH, PH.D.

*Milltown, New Jersey*
*October 2009*