The Data Warehouse Toolkit

Second Edition

The Complete Guide to Dimensional Modeling

Ralph Kimball
Margy Ross

Wiley Computer Publishing

John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

The Data Warehouse Toolkit
Second Edition

The Data Warehouse Toolkit

Second Edition

The Complete Guide to Dimensional Modeling

Ralph Kimball
Margy Ross

Wiley Computer Publishing

John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

Publisher: Robert Ipsen
Editor: Robert Elliott
Assistant Editor: Emilie Herman
Managing Editor: John Atkins
Associate New Media Editor: Brian Snapp
Text Composition: John Wiley Composition Services

Published by John Wiley and Sons, Inc.

Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Kimball, Ralph.
The data warehouse toolkit : the complete guide to dimensional modeling / Ralph Kimball, Margy Ross. — 2nd ed.
p. cm.
"Wiley Computer Publishing."
Includes index.
ISBN 0-471-20024-7
1. Database design. 2. Data warehousing. I. Ross, Margy, 1959– II. Title.

QA76.9.D26 K575 2002
658.4'038'0285574—dc21 2002002284

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Acknowledgments **xv**
Introduction **xvii**

Chapter 1 Dimensional Modeling Primer **1**
Different Information Worlds 2
Goals of a Data Warehouse 2
The Publishing Metaphor 4
Components of a Data Warehouse 6
Operational Source Systems 7
Data Staging Area 8
Data Presentation 10
Data Access Tools 13
Additional Considerations 14
Dimensional Modeling Vocabulary 16
Fact Table 16
Dimension Tables 19
Bringing Together Facts and Dimensions 21
Dimensional Modeling Myths 24
Common Pitfalls to Avoid 26
Summary 27

Chapter 2 Retail Sales **29**
Four-Step Dimensional Design Process 30
Retail Case Study 32
Step 1. Select the Business Process 33
Step 2. Declare the Grain 34
Step 3. Choose the Dimensions 35
Step 4. Identify the Facts 36

Dimension Table Attributes 38
Date Dimension 38
Product Dimension 42
Store Dimension 45
Promotion Dimension 46
Degenerate Transaction Number Dimension 50
Retail Schema in Action 51
Retail Schema Extensibility 52
Resisting Comfort Zone Urges 54
Dimension Normalization (Snowflaking) 55
Too Many Dimensions 57
Surrogate Keys 58
Market Basket Analysis 62
Summary 65

Chapter 3 Inventory 67
Introduction to the Value Chain 68
Inventory Models 69
Inventory Periodic Snapshot 69
Inventory Transactions 74
Inventory Accumulating Snapshot 75
Value Chain Integration 76
Data Warehouse Bus Architecture 78
Data Warehouse Bus Matrix 79
Conformed Dimensions 82
Conformed Facts 87
Summary 88

Chapter 4 Procurement 89
Procurement Case Study 89
Procurement Transactions 90
Multiple- versus Single-Transaction Fact Tables 91
Complementary Procurement Snapshot 93

Slowly Changing Dimensions 95
Type 1: Overwrite the Value 95
Type 2: Add a Dimension Row 97
Type 3: Add a Dimension Column 100
Hybrid Slowly Changing Dimension Techniques 102
Predictable Changes with Multiple Version Overlays 102
Unpredictable Changes with Single Version Overlay 103
More Rapidly Changing Dimensions 105
Summary 105

Chapter 5 Order Management 107
Introduction to Order Management 108
Order Transactions 109
Fact Normalization 109
Dimension Role-Playing 110
Product Dimension Revisited 111
Customer Ship-To Dimension 113
Deal Dimension 116
Degenerate Dimension for Order Number 117
Junk Dimensions 117
Multiple Currencies 119
Header and Line Item Facts with Different Granularity 121
Invoice Transactions 122
Profit and Loss Facts 124
Profitability—The Most Powerful Data Mart 126
Profitability Words of Warning 127
Customer Satisfaction Facts 127
Accumulating Snapshot for the Order Fulfillment Pipeline 128
Lag Calculations 130
Multiple Units of Measure 130
Beyond the Rear-View Mirror 132
Fact Table Comparison 132
Transaction Fact Tables 133
Periodic Snapshot Fact Tables 134
Accumulating Snapshot Fact Tables 134

Designing Real-Time Partitions 135
Requirements for the Real-Time Partition 136
Transaction Grain Real-Time Partition 136
Periodic Snapshot Real-Time Partition 137
Accumulating Snapshot Real-Time Partition 138
Summary 139

Chapter 6 Customer Relationship Management 141
CRM Overview 142
Operational and Analytical CRM 143
Packaged CRM 145
Customer Dimension 146
Name and Address Parsing 147
Other Common Customer Attributes 150
Dimension Outriggers for a Low-Cardinality Attribute Set 153
Large Changing Customer Dimensions 154
Implications of Type 2 Customer Dimension Changes 159
Customer Behavior Study Groups 160
Commercial Customer Hierarchies 161
Combining Multiple Sources of Customer Data 168
Analyzing Customer Data from Multiple Business Processes 169
Summary 170

Chapter 7 Accounting 173
Accounting Case Study 174
General Ledger Data 175
General Ledger Periodic Snapshot 175
General Ledger Journal Transactions 177
Financial Statements 180
Budgeting Process 180
Consolidated Fact Tables 184
Role of OLAP and Packaged Analytic Solutions 185
Summary 186

Chapter 8 Human Resources Management 187
Time-Stamped Transaction Tracking in a Dimension 188
Time-Stamped Dimension with Periodic Snapshot Facts 191
Audit Dimension 193
Keyword Outrigger Dimension 194
AND/OR Dilemma 195
Searching for Substrings 196
Survey Questionnaire Data 197
Summary 198

Chapter 9 Financial Services 199
Banking Case Study 200
Dimension Triage 200
Household Dimension 204
Multivalued Dimensions 205
Minidimensions Revisited 206
Arbitrary Value Banding of Facts 207
Point-in-Time Balances 208
Heterogeneous Product Schemas 210
Heterogeneous Products with Transaction Facts 215
Summary 215

Chapter 10 Telecommunications and Utilities 217
Telecommunications Case Study 218
General Design Review Considerations 220
Granularity 220
Date Dimension 222
Degenerate Dimensions 222
Dimension Decodes and Descriptions 222
Surrogate Keys 223
Too Many (or Too Few) Dimensions 223
Draft Design Exercise Discussion 223
Geographic Location Dimension 226
Location Outrigger 226
Leveraging Geographic Information Systems 227
Summary 227

Chapter 11 Transportation 229
Airline Frequent Flyer Case Study 230
Multiple Fact Table Granularities 230
Linking Segments into Trips 233
Extensions to Other Industries 234
Cargo Shipper 234
Travel Services 235
Combining Small Dimensions into a Superdimension 236
Class of Service 236
Origin and Destination 237
More Date and Time Considerations 239
Country-Specific Calendars 239
Time of Day as a Dimension or Fact 240
Date and Time in Multiple Time Zones 240
Summary 241

Chapter 12 Education 243
University Case Study 244
Accumulating Snapshot for Admissions Tracking 244
Factless Fact Tables 246
Student Registration Events 247
Facilities Utilization Coverage 249
Student Attendance Events 250
Other Areas of Analytic Interest 253
Summary 254

Chapter 13 Health Care 255
Health Care Value Circle 256
Health Care Bill 258
Roles Played By the Date Dimension 261
Multivalued Diagnosis Dimension 262
Extending a Billing Fact Table to Show Profitability 265
Dimensions for Billed Hospital Stays 266

Complex Health Care Events 267
Medical Records 269
Fact Dimension for Sparse Facts 269
Going Back in Time 271
Late-Arriving Fact Rows 271
Late-Arriving Dimension Rows 273
Summary 274

Chapter 14 Electronic Commerce 277
Web Client-Server Interactions Tutorial 278
Why the Clickstream Is Not Just Another Data Source 281
Challenges of Tracking with Clickstream Data 282
Specific Dimensions for the Clickstream 287
Clickstream Fact Table for Complete Sessions 292
Clickstream Fact Table for Individual Page Events 295
Aggregate Clickstream Fact Tables 298
Integrating the Clickstream Data Mart into the Enterprise Data Warehouse 299
Electronic Commerce Profitability Data Mart 300
Summary 303

Chapter 15 Insurance 305
Insurance Case Study 306
Insurance Value Chain 307
Draft Insurance Bus Matrix 309
Policy Transactions 309
Dimension Details and Techniques 310
Alternative (or Complementary) Policy Accumulating Snapshot 315
Policy Periodic Snapshot 316
Conformed Dimensions 316
Conformed Facts 316
Heterogeneous Products Again 318
Multivalued Dimensions Again 318

More Insurance Case Study Background 319
Updated Insurance Bus Matrix 320
Claims Transactions 322
Claims Accumulating Snapshot 323
Policy/Claims Consolidated Snapshot 324
Factless Accident Events 325
Common Dimensional Modeling Mistakes to Avoid 326
Summary 330

Chapter 16 Building the Data Warehouse 331
Business Dimensional Lifecycle Road Map 332
Road Map Major Points of Interest 333
Project Planning and Management 334
Assessing Readiness 334
Scoping 336
Justification 336
Staffing 337
Developing and Maintaining the Project Plan 339
Business Requirements Definition 340
Requirements Preplanning 341
Collecting the Business Requirements 343
Postcollection Documentation and Follow-up 345
Lifecycle Technology Track 347
Technical Architecture Design 348
Eight-Step Process for Creating the Technical Architecture 348
Product Selection and Installation 351
Lifecycle Data Track 353
Dimensional Modeling 353
Physical Design 355
Aggregation Strategy 356
Initial Indexing Strategy 357
Data Staging Design and Development 358
Dimension Table Staging 358
Fact Table Staging 361

Lifecycle Analytic Applications Track 362
Analytic Application Specification 363
Analytic Application Development 363
Deployment 364
Maintenance and Growth 365
Common Data Warehousing Mistakes to Avoid 366
Summary 369

Chapter 17 Present Imperatives and Future Outlook 371
Ongoing Technology Advances 372
Political Forces Demanding Security and Affecting Privacy 375
Conflict between Beneficial Uses and Insidious Abuses 375
Who Owns Your Personal Data? 376
What Is Likely to Happen? Watching the Watchers . . . 377
How Watching the Watchers Affects Data Warehouse Architecture 378
Designing to Avoid Catastrophic Failure 379
Catastrophic Failures 380
Countering Catastrophic Failures 380
Intellectual Property and Fair Use 383
Cultural Trends in Data Warehousing 383
Managing by the Numbers across the Enterprise 383
Increased Reliance on Sophisticated Key Performance Indicators 384
Behavior Is the New Marquee Application 385
Packaged Applications Have Hit Their High Point 385
Application Integration Has to Be Done by Someone 386
Data Warehouse Outsourcing Needs a Sober Risk Assessment 386
In Closing 387

Glossary 389

Index 419

ACKNOWLEDGMENTS

First of all, we want to thank the thousands of you who have read our *Toolkit* books, attended our courses, and engaged us in consulting projects. We have learned as much from you as we have taught. As a group, you have had a profoundly positive impact on the data warehousing industry. Congratulations!

This book would not have been written without the assistance of our business partners. We want to thank Julie Kimball of Ralph Kimball Associates for her vision and determination in getting the project launched. While Julie was the catalyst who got the ball rolling, Bob Becker of DecisionWorks Consulting helped keep it in motion as he drafted, reviewed, and served as a general sounding board. We are grateful to them both because they helped an enormous amount.

We wrote this book with a little help from our friends, who provided input or feedback on specific chapters. We want to thank Bill Schmarzo of DecisionWorks, Charles Hagensen of Attachmate Corporation, and Warren Thornthwaite of InfoDynamics for their counsel on Chapters 6, 7, and 16, respectively.

Bob Elliott, our editor at John Wiley & Sons, and the entire Wiley team have supported this project with skill, encouragement, and enthusiasm. It has been a pleasure to work with them. We also want to thank Justin Kestelyn, editor-in-chief at *Intelligent Enterprise* for allowing us to adapt materials from several of Ralph's articles for inclusion in this book.

To our families, thanks for being there for us when we needed you and for giving us the time it took. Spouses Julie Kimball and Scott Ross and children Sara Hayden Smith, Brian Kimball, and Katie Ross all contributed a lot to this book, often without realizing it. Thanks for your unconditional support.

INTRODUCTION

The data warehousing industry certainly has matured since Ralph Kimball published the first edition of *The Data Warehouse Toolkit* (Wiley) in 1996. Although large corporate early adopters paved the way, since then, data warehousing has been embraced by organizations of all sizes. The industry has constructed thousands of data warehouses. The volume of data continues to grow as we populate our warehouses with increasingly atomic data and update them with greater frequency. Vendors continue to blanket the market with an ever-expanding set of tools to help us with data warehouse design, development, and usage. Most important, armed with access to our data warehouses, business professionals are making better decisions and generating payback on their data warehouse investments.

Since the first edition of *The Data Warehouse Toolkit* was published, dimensional modeling has been broadly accepted as the dominant technique for data warehouse presentation. Data warehouse practitioners and pundits alike have recognized that the data warehouse presentation must be grounded in simplicity if it stands any chance of success. Simplicity is the fundamental key that allows users to understand databases easily and software to navigate databases efficiently. In many ways, dimensional modeling amounts to holding the fort against assaults on simplicity. By consistently returning to a business-driven perspective and by refusing to compromise on the goals of user understandability and query performance, we establish a coherent design that serves the organization's analytic needs. Based on our experience and the overwhelming feedback from numerous practitioners from companies like your own, we believe that dimensional modeling is absolutely critical to a successful data warehousing initiative.

Dimensional modeling also has emerged as the only coherent architecture for building distributed data warehouse systems. When we use the conformed dimensions and conformed facts of a set of dimensional models, we have a practical and predictable framework for incrementally building complex data warehouse systems that have no center.

For all that has changed in our industry, the core dimensional modeling techniques that Ralph Kimball published six years ago have withstood the test of time. Concepts such as slowly changing dimensions, heterogeneous products,

factless fact tables, and architected data marts continue to be discussed in data warehouse design workshops around the globe. The original concepts have been embellished and enhanced by new and complementary techniques. We decided to publish a second edition of Kimball's seminal work because we felt that it would be useful to pull together our collective thoughts on dimensional modeling under a single cover. We have each focused exclusively on decision support and data warehousing for over two decades. We hope to share the dimensional modeling patterns that have emerged repeatedly during the course of our data warehousing careers. This book is loaded with specific, practical design recommendations based on real-world scenarios.

The goal of this book is to provide a one-stop shop for dimensional modeling techniques. True to its title, it is a toolkit of dimensional design principles and techniques. We will address the needs of those just getting started in dimensional data warehousing, and we will describe advanced concepts for those of you who have been at this a while. We believe that this book stands alone in its depth of coverage on the topic of dimensional modeling.

Intended Audience

This book is intended for data warehouse designers, implementers, and managers. In addition, business analysts who are active participants in a warehouse initiative will find the content useful.

Even if you're not directly responsible for the dimensional model, we believe that it is important for all members of a warehouse project team to be comfortable with dimensional modeling concepts. The dimensional model has an impact on most aspects of a warehouse implementation, beginning with the translation of business requirements, through data staging, and finally, to the unveiling of a data warehouse through analytic applications. Due to the broad implications, you need to be conversant in dimensional modeling regardless whether you are responsible primarily for project management, business analysis, data architecture, database design, data staging, analytic applications, or education and support. We've written this book so that it is accessible to a broad audience.

For those of you who have read the first edition of this book, some of the familiar case studies will reappear in this edition; however, they have been updated significantly and fleshed out with richer content. We have developed vignettes for new industries, including health care, telecommunications, and electronic commerce. In addition, we have introduced more horizontal, cross-industry case studies for business functions such as human resources, accounting, procurement, and customer relationship management.

The content in this book is mildly technical. We discuss dimensional modeling in the context of a relational database primarily. We presume that readers have basic knowledge of relational database concepts such as tables, rows, keys, and joins. Given that we will be discussing dimensional models in a non-denominational manner, we won't dive into specific physical design and tuning guidance for any given database management systems.

Chapter Preview

The book is organized around a series of business vignettes or case studies. We believe that developing the design techniques by example is an extremely effective approach because it allows us to share very tangible guidance. While not intended to be full-scale application or industry solutions, these examples serve as a framework to discuss the patterns that emerge in dimensional modeling. In our experience, it is often easier to grasp the main elements of a design technique by stepping away from the all-too-familiar complexities of one's own applications in order to think about another business. Readers of the first edition have responded very favorably to this approach.

The chapters of this book build on one another. We will start with basic concepts and introduce more advanced content as the book unfolds. The chapters are to be read in order by every reader. For example, Chapter 15 on insurance will be difficult to comprehend unless you have read the preceding chapters on retailing, procurement, order management, and customer relationship management.

Those of you who have read the first edition may be tempted to skip the first few chapters. While some of the early grounding regarding facts and dimensions may be familiar turf, we don't want you to sprint too far ahead. For example, the first case study focuses on the retailing industry, just as it did in the first edition**.** However, in this edition we advocate a new approach, making a strong case for tackling the atomic, bedrock data of your organization. You'll miss out on this rationalization and other updates to fundamental concepts if you skip ahead too quickly.

Navigation Aids

We have laced the book with tips, key concepts, and chapter pointers to make it more usable and easily referenced in the future. In addition, we have provided an extensive glossary of terms.

You can find the tips sprinkled throughout this book by flipping through the chapters and looking for the lightbulb icon.

We begin each chapter with a sidebar of key concepts, denoted by the key icon.

Purpose of Each Chapter

Before we get started, we want to give you a chapter-by-chapter preview of the concepts covered as the book unfolds.

Chapter 1: Dimensional Modeling Primer

The book begins with a primer on dimensional modeling. We explore the components of the overall data warehouse architecture and establish core vocabulary that will be used during the remainder of the book. We dispel some of the myths and misconceptions about dimensional modeling, and we discuss the role of normalized models.

Chapter 2: Retail Sales

Retailing is the classic example used to illustrate dimensional modeling. We start with the classic because it is one that we all understand. Hopefully, you won't need to think very hard about the industry because we want you to focus on core dimensional modeling concepts instead. We begin by discussing the four-step process for designing dimensional models. We explore dimension tables in depth, including the date dimension that will be reused repeatedly throughout the book. We also discuss degenerate dimensions, snowflaking, and surrogate keys. Even if you're not a retailer, this chapter is required reading because it is chock full of fundamentals.

Chapter 3: Inventory

We remain within the retail industry for our second case study but turn our attention to another business process. This case study will provide a very vivid example of the data warehouse bus architecture and the use of conformed dimensions and facts. These concepts are critical to anyone looking to construct a data warehouse architecture that is integrated and extensible.

Chapter 4: Procurement

This chapter reinforces the importance of looking at your organization's value chain as you plot your data warehouse. We also explore a series of basic and advanced techniques for handling slowly changing dimension attributes.

Chapter 5: Order Management

In this case study we take a look at the business processes that are often the first to be implemented in data warehouses as they supply core business performance metrics—what are we selling to which customers at what price? We discuss the situation in which a dimension plays multiple roles within a schema. We also explore some of the common challenges modelers face when dealing with order management information, such as header/line item considerations, multiple currencies or units of measure, and junk dimensions with miscellaneous transaction indicators. We compare the three fundamental types of fact tables: transaction, periodic snapshot, and accumulating snapshot. Finally, we provide recommendations for handling more real-time warehousing requirements.

Chapter 6: Customer Relationship Management

Numerous data warehouses have been built on the premise that we need to better understand and service our customers. This chapter covers key considerations surrounding the customer dimension, including address standardization, managing large volume dimensions, and modeling unpredictable customer hierarchies. It also discusses the consolidation of customer data from multiple sources.

Chapter 7: Accounting

In this totally new chapter we discuss the modeling of general ledger information for the data warehouse. We describe the appropriate handling of year-to-date facts and multiple fiscal calendars, as well as the notion of consolidated dimensional models that combine data from multiple business processes.

Chapter 8: Human Resources Management

This new chapter explores several unique aspects of human resources dimensional models, including the situation in which a dimension table begins to behave like a fact table. We also introduce audit and keyword dimensions, as well as the handling of survey questionnaire data.

Chapter 9: Financial Services

The banking case study explores the concept of heterogeneous products in which each line of business has unique descriptive attributes and performance metrics. Obviously, the need to handle heterogeneous products is not unique to financial services. We also discuss the complicated relationships among accounts, customers, and households.

Chapter 10: Telecommunications and Utilities

This new chapter is structured somewhat differently to highlight considerations when performing a data model design review. In addition, we explore the idiosyncrasies of geographic location dimensions, as well as opportunities for leveraging geographic information systems.

Chapter 11: Transportation

In this case study we take a look at related fact tables at different levels of granularity. We discuss another approach for handling small dimensions, and we take a closer look at date and time dimensions, covering such concepts as country-specific calendars and synchronization across multiple time zones.

Chapter 12: Education

We look at several factless fact tables in this chapter and discuss their importance in analyzing what didn't happen. In addition, we explore the student application pipeline, which is a prime example of an accumulating snapshot fact table.

Chapter 13: Health Care

Some of the most complex models that we have ever worked with are from the health care industry. This new chapter illustrates the handling of such complexities, including the use of a bridge table to model multiple diagnoses and providers associated with a patient treatment.

Chapter 14: Electronic Commerce

This chapter provides an introduction to modeling clickstream data. The concepts are derived from *The Data Webhouse Toolkit* (Wiley 2000), which Ralph Kimball coauthored with Richard Merz.

Chapter 15: Insurance

The final case study serves to illustrate many of the techniques we discussed earlier in the book in a single set of interrelated schemas. It can be viewed as a pulling-it-all-together chapter because the modeling techniques will be layered on top of one another, similar to overlaying overhead projector transparencies.

Chapter 16: Building the Data Warehouse

Now that you are comfortable designing dimensional models, we provide a high-level overview of the activities that are encountered during the lifecycle of a typical data warehouse project iteration. This chapter could be considered a lightning tour of *The Data Warehouse Lifecycle Toolkit* (Wiley 1998) that we coauthored with Laura Reeves and Warren Thornthwaite.

Chapter 17: Present Imperatives and Future Outlook

In this final chapter we peer into our crystal ball to provide a preview of what we anticipate data warehousing will look like in the future.

Glossary

We've supplied a detailed glossary to serve as a reference resource. It will help bridge the gap between your general business understanding and the case studies derived from businesses other than your own.

Companion Web Site

You can access the book's companion Web site at www.kimballuniversity.com. The Web site offers the following resources:

- Register for *Design Tips* to receive ongoing, practical guidance about dimensional modeling and data warehouse design via electronic mail on a periodic basis.
- Link to all Ralph Kimball's articles from *Intelligent Enterprise* and its predecessor, *DBMS Magazine*.
- Learn about Kimball University classes for quality, vendor-independent education consistent with the authors' experiences and writings.

Summary

The goal of this book is to communicate a set of standard techniques for dimensional data warehouse design. Crudely speaking, if you as the reader get nothing else from this book other than the conviction that your data warehouse must be driven from the needs of business users and therefore built and presented from a simple dimensional perspective, then this book will have served its purpose. We are confident that you will be one giant step closer to data warehousing success if you buy into these premises.

Now that you know where we are headed, it is time to dive into the details. We'll begin with a primer on dimensional modeling in Chapter 1 to ensure that everyone is on the same page regarding key terminology and architectural concepts. From there we will begin our discussion of the fundamental techniques of dimensional modeling, starting with the tried-and-true retail industry.

CHAPTER 1

Dimensional Modeling Primer

In this first chapter we lay the groundwork for the case studies that follow. We'll begin by stepping back to consider data warehousing from a macro perspective. Some readers may be disappointed to learn that it is not all about tools and techniques—first and foremost, the data warehouse must consider the needs of the business. We'll drive stakes in the ground regarding the goals of the data warehouse while observing the uncanny similarities between the responsibilities of a data warehouse manager and those of a publisher. With this big-picture perspective, we'll explore the major components of the warehouse environment, including the role of normalized models. Finally, we'll close by establishing fundamental vocabulary for dimensional modeling. By the end of this chapter we hope that you'll have an appreciation for the need to be half DBA (database administrator) and half MBA (business analyst) as you tackle your data warehouse.

Chapter 1 discusses the following concepts:

- **Business-driven goals of a data warehouse**
- **Data warehouse publishing**
- **Major components of the overall data warehouse**
- **Importance of dimensional modeling for the data warehouse presentation area**
- **Fact and dimension table terminology**
- **Myths surrounding dimensional modeling**
- **Common data warehousing pitfalls to avoid**

Different Information Worlds

One of the most important assets of any organization is its information. This asset is almost always kept by an organization in two forms: the operational systems of record and the data warehouse. Crudely speaking, the operational systems are where the data is put in, and the data warehouse is where we get the data out.

The users of an operational system *turn* the wheels of the organization. They take orders, sign up new customers, and log complaints. Users of an operational system almost always deal with one record at a time. They repeatedly perform the same operational tasks over and over.

The users of a data warehouse, on the other hand, *watch* the wheels of the organization turn. They count the new orders and compare them with last week's orders and ask why the new customers signed up and what the customers complained about. Users of a data warehouse almost never deal with one row at a time. Rather, their questions often require that hundreds or thousands of rows be searched and compressed into an answer set. To further complicate matters, users of a data warehouse continuously change the kinds of questions they ask.

In the first edition of *The Data Warehouse Toolkit* (Wiley 1996), Ralph Kimball devoted an entire chapter to describe the dichotomy between the worlds of operational processing and data warehousing. At this time, it is widely recognized that the data warehouse has profoundly different needs, clients, structures, and rhythms than the operational systems of record. Unfortunately, we continue to encounter supposed data warehouses that are mere copies of the operational system of record stored on a separate hardware platform. While this may address the need to isolate the operational and warehouse environments for performance reasons, it does nothing to address the other inherent differences between these two types of systems. Business users are underwhelmed by the usability and performance provided by these pseudo data warehouses. These imposters do a disservice to data warehousing because they don't acknowledge that warehouse users have drastically different needs than operational system users.

Goals of a Data Warehouse

Before we delve into the details of modeling and implementation, it is helpful to focus on the fundamental goals of the data warehouse. The goals can be developed by walking through the halls of any organization and listening to business management. Inevitably, these recurring themes emerge:

- "We have mountains of data in this company, but we can't access it."
- "We need to slice and dice the data every which way."
- "You've got to make it easy for business people to get at the data directly."
- "Just show me what is important."
- "It drives me crazy to have two people present the same business metrics at a meeting, but with different numbers."
- "We want people to use information to support more fact-based decision making."

Based on our experience, these concerns are so universal that they drive the bedrock requirements for the data warehouse. Let's turn these business management quotations into data warehouse requirements.

The data warehouse must make an organization's information easily accessible. The contents of the data warehouse must be understandable. The data must be intuitive and obvious to the business user, not merely the developer. Understandability implies legibility; the contents of the data warehouse need to be labeled meaningfully. Business users want to separate and combine the data in the warehouse in endless combinations, a process commonly referred to as *slicing and dicing*. The tools that access the data warehouse must be simple and easy to use. They also must return query results to the user with minimal wait times.

The data warehouse must present the organization's information consistently. The data in the warehouse must be credible. Data must be carefully assembled from a variety of sources around the organization, cleansed, quality assured, and released only when it is fit for user consumption. Information from one business process should match with information from another. If two performance measures have the same name, then they must mean the same thing. Conversely, if two measures don't mean the same thing, then they should be labeled differently. Consistent information means high-quality information. It means that all the data is accounted for and complete. Consistency also implies that common definitions for the contents of the data warehouse are available for users.

The data warehouse must be adaptive and resilient to change. We simply can't avoid change. User needs, business conditions, data, and technology are all subject to the shifting sands of time. The data warehouse must be designed to handle this inevitable change. Changes to the data warehouse should be graceful, meaning that they don't invalidate existing data or applications. The existing data and applications should not be changed or disrupted when the business community asks new questions or new data is added to the warehouse. If descriptive data in the warehouse is modified, we must account for the changes appropriately.

The data warehouse must be a secure bastion that protects our information assets. An organization's informational crown jewels are stored in the data warehouse. At a minimum, the warehouse likely contains information about what we're selling to whom at what price—potentially harmful details in the hands of the wrong people. The data warehouse must effectively control access to the organization's confidential information.

The data warehouse must serve as the foundation for improved decision making. The data warehouse must have the right data in it to support decision making. There is only one true output from a data warehouse: the decisions that are made after the data warehouse has presented its evidence. These decisions deliver the business impact and value attributable to the warehouse. The original label that predates the data warehouse is still the best description of what we are designing: a decision support system.

The business community must accept the data warehouse if it is to be deemed successful. It doesn't matter that we've built an elegant solution using best-of-breed products and platforms. If the business community has not embraced the data warehouse and continued to use it actively six months after training, then we have failed the acceptance test. Unlike an operational system rewrite, where business users have no choice but to use the new system, data warehouse usage is sometimes optional. Business user acceptance has more to do with simplicity than anything else.

As this list illustrates, successful data warehousing demands much more than being a stellar DBA or technician. With a data warehousing initiative, we have one foot in our information technology (IT) comfort zone, while our other foot is on the unfamiliar turf of business users. We must straddle the two, modifying some of our tried-and-true skills to adapt to the unique demands of data warehousing. Clearly, we need to bring a bevy of skills to the party to behave like we're a hybrid DBA/MBA.

The Publishing Metaphor

With the goals of the data warehouse as a backdrop, let's compare our responsibilities as data warehouse managers with those of a publishing editor-in-chief. As the editor of a high-quality magazine, you would be given broad latitude to manage the magazine's content, style, and delivery. Anyone with this job title likely would tackle the following activities:

- Identify your readers demographically.
- Find out what the readers want in this kind of magazine.
- Identify the "best" readers who will renew their subscriptions and buy products from the magazine's advertisers.