

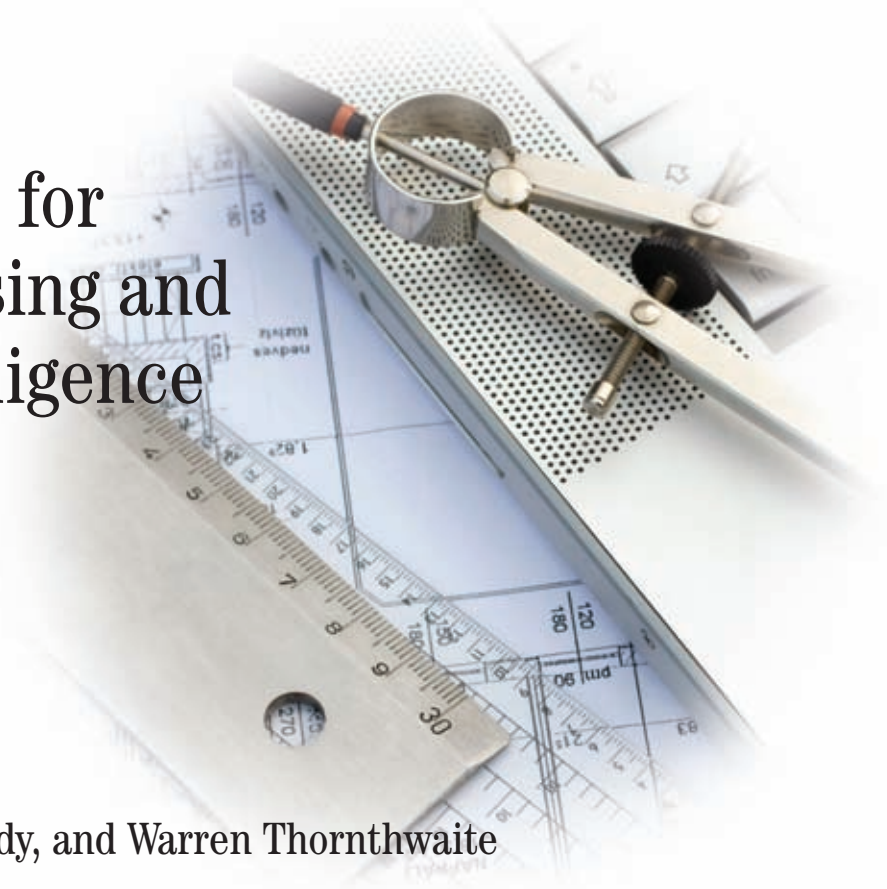


# The Kimball Group Reader

Relentlessly  
Practical Tools for  
Data Warehousing and  
Business Intelligence

**Ralph Kimball**  
**Margy Ross**

with Bob Becker, Joy Mundy, and Warren Thornthwaite





# **The Kimball Group Reader**



# The Kimball Group Reader

Relentlessly Practical Tools  
for Data Warehousing and  
Business Intelligence

Ralph Kimball

Margy Ross

with Bob Becker, Joy Mundy,  
and Warren Thornthwaite



Wiley Publishing, Inc.

**The Kimball Group Reader; Relentlessly Practical Tools for Data Warehousing and Business Intelligence**

Published by  
Wiley Publishing, Inc.  
10475 Crosspoint Boulevard  
Indianapolis, IN 46256  
[www.wiley.com](http://www.wiley.com)

Copyright © 2010 Ralph Kimball and Margy Ross

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-0-470-56310-6

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

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 Sections 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, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. 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/permissions>.

**Limit of Liability/Disclaimer of Warranty:** The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or Web site may provide or recommendations it may make. Further, readers should be aware that Internet Web sites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 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 books.

**Library of Congress Control Number:** 2009942822

**Trademarks:** Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. Wiley Publishing, Inc. is not associated with any product or vendor mentioned in this book.

# About the Authors

**Ralph Kimball** founded the Kimball Group. Since the mid 1980s, he has been the DW/BI industry's thought leader on the dimensional approach and has trained more than 10,000 IT professionals. Prior to working at Metaphor and founding Red Brick Systems, Ralph co-invented the Star workstation at Xerox's Palo Alto Research Center (PARC). Ralph has a Ph.D. in Electrical Engineering from Stanford University.

**Margy Ross** is President of the Kimball Group. She has focused exclusively on DW/BI solutions since 1982 with an emphasis on business requirements analysis and dimensional modeling. Margy graduated with a B.S. in Industrial Engineering from Northwestern University.

# Credits

**Executive Editor**  
Robert Elliott

**Project Editor**  
Sara Schlaer

**Senior Production Editor**  
Debra Banninger

**Copy Editor**  
Kim Cofer

**Editorial Director**  
Robyn B. Siesky

**Editorial Manager**  
Mary Beth Wakefield

**Marketing Manager**  
Ashley Zurcher

**Production Manager**  
Tim Tate

**Vice President and Executive Group Publisher**  
Richard Swadley

**Vice President and Executive Publisher**  
Barry Pruett

**Associate Publisher**  
Jim Minatel

**Project Coordinator, Cover**  
Lynsey Stanford

**Compositor**  
Maureen Forys, Happenstance Type-O-Rama

**Proofreader**  
Nancy Carrasco

**Indexer**  
Johnna VanHoose Dinse

**Cover Image**  
iStockPhoto

**Cover Designer**  
Ryan Sneed



# Acknowledgments

**F**irst, we want to thank the 28,000 current subscribers to the *Kimball Design Tips*, and the uncounted numbers of you who have visited the Kimball Group web site to peruse our magazine article archive. This book brings the design tips and articles together in what we hope is a very usable form.

*The Kimball Group Reader* would not exist without the assistance of our business partners. Kimball Group members Bob Becker, Joy Mundy, and Warren Thornthwaite wrote many of the valuable articles and design tips included in the book. Thanks to Julie Kimball for her insightful comments. Thanks also to former Kimball Group member Bill Schmarzo for his contributions on analytic applications.

Bob Elliott, our executive editor at Wiley Publishing, project editor Sara Shlaer, and the rest of the 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 Doug Henschen, editor of *Intelligent Enterprise*, and Julie Langenkamp, editor of *DM Review*, for allowing us to publish the many articles that appeared in their magazines.

To our families, thank you for your support over the fifteen year span during which we wrote these design tips and articles. Julie Kimball and Scott Ross have been a huge positive force in our professional and personal lives. And, of course, thanks to our children, Sara Kimball Smith, Brian Kimball, and Katie Ross, who have grown into adults over the same time!



# Contents at a Glance

Introduction .....	xxi
<b>1</b> <i>The Reader</i> at a Glance .....	1
<b>2</b> Before You Dive In .....	35
<b>3</b> Project/Program Planning .....	69
<b>4</b> Requirements Definition .....	113
<b>5</b> Data Architecture .....	133
<b>6</b> Dimensional Modeling Fundamentals .....	179
<b>7</b> Dimensional Modeling Tasks and Responsibilities .....	209
<b>8</b> Fact Table Core Concepts .....	233
<b>9</b> Dimension Table Core Concepts .....	285
<b>10</b> More Dimension Patterns and Case Studies .....	333
<b>11</b> Back Room ETL and Data Quality .....	425
<b>12</b> Technical Architecture Considerations .....	513
<b>13</b> Front Room Business Intelligence Applications .....	589
<b>14</b> Maintenance and Growth Considerations .....	651
Index of Articles .....	685
Index .....	693



# Contents

Introduction .....	xxi
<b>1</b> <i>The Reader at a Glance</i> .....	1
Setting Up for Success .....	1
1.1 Resist the Urge to Start Coding .....	1
1.2 Set Your Boundaries .....	4
Tackling DW/BI Design and Development .....	6
1.3 Data Wrangling .....	6
1.4 Myth Busters .....	8
1.5 Dividing the World .....	10
1.6 Essential Steps for the Integrated Enterprise Data Warehouse .....	13
1.7 Drill Down to Ask Why .....	22
1.8 Slowly Changing Dimensions .....	24
1.9 Judge Your BI Tool through Your Dimensions .....	28
1.10 Fact Tables .....	30
1.11 Exploit Your Fact Tables .....	32
<b>2</b> Before You Dive In .....	35
Historical Perspective .....	35
2.1 The Database Market Splits .....	36
2.2 Bringing Up Supermarts .....	38
Dealing with Demanding Realities .....	45
2.3 Brave New Requirements for Data Warehousing .....	46
2.4 Coping with the Brave New Requirements .....	50
2.5 Stirring Things Up .....	55
2.6 Design Constraints and Unavoidable Realities .....	58
2.7 Two Powerful Ideas .....	61
2.8 Data Warehouse Dining Experience .....	65

<b>3</b>	<b>Project/Program Planning</b>	<b>69</b>
	Professional Responsibilities	69
	3.1 Professional Boundaries	69
	3.2 An Engineer's View	72
	3.3 Beware the Objection Removers	76
	3.4 What Does the Central Team Do?	80
	3.5 Avoid DW/BI Isolation	83
	3.6 Implementation Analysis Paralysis	84
	Justification and Sponsorship	85
	3.7 Habits of Effective Sponsors	86
	3.8 TCO Starts with the End User	89
	3.9 Better Business Skills for BI and Data Warehouse Professionals	93
	Kimball Methodology	96
	3.10 Kimball Lifecycle in a Nutshell	96
	3.11 Off the Bench	99
	3.12 The Anti-Architect	100
	3.13 Think Critically When Applying Best Practices	103
	3.14 Eight Guidelines for Low Risk Enterprise Data Warehousing	106
	3.15 Relating to Agile Methodologies	109
	3.16 Is Agile Enterprise Data Warehousing an Oxymoron?	110
<b>4</b>	<b>Requirements Definition</b>	<b>113</b>
	Gathering Requirements	113
	4.1 Alan Alda's Interviewing Tips for Uncovering Business Requirements	113
	4.2 More Business Requirements Gathering Dos and Don'ts	117
	4.3 Overcoming Obstacles When Gathering Business Requirements	119
	4.4 Surprising Value of Data Profiling	121
	Organizing around Business Processes	123
	4.5 Focus on Business Processes, Not Business Departments!	123
	4.6 Identifying Business Processes	124
	4.7 Business Process Decoder Ring	127
	4.8 Relationship between Strategic Business Initiatives and Business Processes	127
	Wrapping Up the Requirements	128
	4.9 The Bottom-Up Misnomer	128

<b>5</b>	<b>Data Architecture</b>	<b>133</b>
	Making the Case for Dimensional Modeling	133
	5.1 Is ER Modeling Hazardous to DSS?	133
	5.2 A Dimensional Modeling Manifesto	137
	5.3 There Are No Guarantees	145
	Enterprise Data Warehouse Bus Architecture	148
	5.4 Divide and Conquer	148
	5.5 The Matrix	151
	5.6 The Matrix: Revisited	155
	5.7 Drill Down into a Detailed Bus Matrix	159
	Integration Instead of Centralization	161
	5.8 Integration for Real People	161
	5.9 Data Stewardship 101: The First Step to Quality and Consistency	165
	5.10 To Be or Not To Be Centralized	168
	Contrast with the Corporate Information Factory	171
	5.11 Differences of Opinion	171
	5.12 Don't Support Business Intelligence with a Normalized EDW	176
<b>6</b>	<b>Dimensional Modeling Fundamentals</b>	<b>179</b>
	Basics of Dimensional Modeling	179
	6.1 Fact Tables and Dimension Tables	179
	6.2 Drilling Down, Up, and Across	183
	6.3 The Soul of the Data Warehouse, Part One: Drilling Down	186
	6.4 The Soul of the Data Warehouse, Part Two: Drilling Across	189
	6.5 The Soul of the Data Warehouse, Part Three: Handling Time	191
	6.6 Graceful Modifications to Existing Fact and Dimension Tables	194
	Dos and Don'ts	196
	6.7 Kimball's Ten Essential Rules of Dimensional Modeling	196
	6.8 What Not to Do	199
	Myths about Dimensional Modeling	201
	6.9 Dangerous Preconceptions	201
	6.10 Fables and Facts	204
<b>7</b>	<b>Dimensional Modeling Tasks and Responsibilities</b>	<b>209</b>
	Design Activities	209
	7.1 Letting the Users Sleep	209

7.2 Staffing the Dimensional Modeling Team . . . . .	216
7.3 Practical Steps for Designing a Dimensional Model . . . . .	217
7.4 The Naming Game . . . . .	220
7.5 When Is the Dimensional Design Done? . . . . .	221
Design Review Activities . . . . .	223
7.6 Fistful of Flaws . . . . .	223
7.7 Rating Your Dimensional Data Warehouse. . . . .	226
<b>8 Fact Table Core Concepts . . . . .</b>	<b>233</b>
Granularity. . . . .	233
8.1 Declaring the Grain. . . . .	233
8.2 Keep to the Grain in Dimensional Modeling . . . . .	236
8.3 Warning: Summary Data May Be Hazardous to Your Health . . . . .	238
8.4 No Detail Too Small . . . . .	239
Types of Fact Tables . . . . .	242
8.5 Fundamental Grains . . . . .	243
8.6 Modeling a Pipeline with an Accumulating Snapshot . . . . .	246
8.7 Combining Periodic and Accumulating Snapshots. . . . .	249
8.8 Modeling Time Spans . . . . .	250
8.9 A Rolling Prediction of the Future, Now and in the Past . . . . .	252
8.10 Factless Fact Tables. . . . .	255
8.11 Factless Fact Tables? Sound Like Jumbo Shrimp? . . . . .	258
8.12 What Didn't Happen . . . . .	259
Parent-Child Fact Tables . . . . .	262
8.13 Managing Your Parents. . . . .	263
8.14 Patterns to Avoid When Modeling Header/Line Item Transactions . . . . .	266
Fact Table Keys and Degenerates. . . . .	268
8.15 Fact Table Surrogate Keys. . . . .	268
8.16 Reader Suggestions on Fact Table Surrogate Keys . . . . .	269
8.17 Another Look at Degenerate Dimensions . . . . .	271
8.18 Creating a Reference Dimension for Infrequently Accessed Degenerates . . . . .	272
Miscellaneous Fact Table Design Patterns. . . . .	273
8.19 Put Your Fact Tables on a Diet . . . . .	273



8.20 Keeping Text Out of the Fact Table . . . . .	275
8.21 Dealing with Nulls in a Dimensional Model. . . . .	276
8.22 Modeling Data as Both a Fact and Dimension Attribute. . . . .	277
8.23 When a Fact Table Can Be Used as a Dimension Table . . . . .	278
8.24 Sparse Facts and Facts with Short Lifetimes. . . . .	280
8.25 Pivoting the Fact Table with a Fact Dimension . . . . .	282
<b>9 Dimension Table Core Concepts . . . . .</b>	<b>285</b>
Dimension Table Keys . . . . .	285
9.1 Surrogate Keys. . . . .	285
9.2 Keep Your Keys Simple. . . . .	289
Date and Time Dimension Considerations . . . . .	290
9.3 It's Time for Time . . . . .	291
9.4 Surrogate Keys for the Time Dimension. . . . .	293
9.5 Latest Thinking on Time Dimension Tables. . . . .	295
9.6 Smart Date Keys to Partition Fact Tables . . . . .	296
9.7 Handling All the Dates . . . . .	298
Miscellaneous Dimension Patterns . . . . .	299
9.8 Data Warehouse Role Models . . . . .	299
9.9 Mystery Dimensions . . . . .	303
9.10 De-Clutter with Junk Dimensions . . . . .	306
9.11 Showing the Correlation Between Dimensions. . . . .	307
9.12 Causal (Not Casual) Dimensions . . . . .	308
9.13 Resist Abstract Generic Dimensions. . . . .	311
9.14 Hot-Swappable Dimensions . . . . .	312
9.15 Accurate Counting with a Dimensional Supplement . . . . .	314
Slowly Changing Dimensions . . . . .	315
9.16 Perfectly Partitioning History with Type 2 SCD. . . . .	316
9.17 Many Alternate Realities . . . . .	316
9.18 Monster Dimensions . . . . .	320
9.19 When a Slowly Changing Dimension Speeds Up . . . . .	322
9.20 When Do Dimensions Become Dangerous? . . . . .	325
9.21 Slowly Changing Dimensions Are Not Always as Easy as 1, 2, and 3. . . . .	326
9.22 Dimension Row Change Reason Attributes . . . . .	330

<b>10</b>	More Dimension Patterns and Case Studies . . . . .	333
	Snowflakes, Outriggers, and Bridges . . . . .	333
	10.1 Snowflakes, Outriggers, and Bridges. . . . .	333
	10.2 A Trio of Interesting Snowflakes. . . . .	336
	10.3 Help for Dimensional Modeling . . . . .	340
	10.4 Managing Bridge Tables . . . . .	343
	10.5 The Keyword Dimension . . . . .	347
	Dealing with Hierarchies . . . . .	351
	10.6 Maintaining Dimension Hierarchies . . . . .	351
	10.7 Help for Hierarchies . . . . .	355
	10.8 Five Alternatives for Better Employee Dimensional Modeling . . . . .	359
	10.9 Alternate Hierarchies . . . . .	365
	Customer Issues. . . . .	366
	10.10 Dimension Embellishments . . . . .	367
	10.11 Wrangling Behavior Tags . . . . .	368
	10.12 Three Ways to Capture Customer Satisfaction. . . . .	371
	Addresses and International Issues . . . . .	374
	10.13 Think Globally, Act Locally . . . . .	374
	10.14 Warehousing without Borders . . . . .	378
	10.15 Spatially Enabling Your Data Warehouse . . . . .	383
	10.16 Multinational Dimensional Data Warehouse Considerations. . . . .	387
	Industry Scenarios and Idiosyncrasies. . . . .	389
	10.17 An Insurance Data Warehouse Case Study. . . . .	389
	10.18 Traveling through Databases. . . . .	393
	10.19 Human Resources Dimensional Models . . . . .	396
	10.20 Not So Fast . . . . .	400
	10.21 The Budgeting Chain . . . . .	403
	10.22 Compliance-Enabled Data Warehouses . . . . .	407
	10.23 Clicking with Your Customer. . . . .	409
	10.24 The Special Dimensions of the Clickstream . . . . .	413
	10.25 Fact Tables for Text Document Searching . . . . .	417
	10.26 Enabling Market Basket Analysis . . . . .	420
<b>11</b>	Back Room ETL and Data Quality. . . . .	425
	Planning the ETL System. . . . .	425
	11.1 Surrounding the ETL Requirements. . . . .	425

11.2 The 34 Subsystems of ETL . . . . .	430
11.3 Doing the Work at Extract Time. . . . .	434
11.4 Is Data Staging Relational? . . . . .	437
11.5 Staging Areas and ETL Tools . . . . .	441
11.6 Should You Use an ETL Tool? . . . . .	442
11.7 Document the ETL System . . . . .	445
11.8 Measure Twice, Cut Once. . . . .	445
11.9 Brace for Incoming . . . . .	449
11.10 Building a Change Data Capture System . . . . .	452
Data Quality Considerations . . . . .	454
11.11 Dealing with Dirty Data . . . . .	454
11.12 An Architecture for Data Quality . . . . .	460
11.13 Indicators of Quality . . . . .	468
11.14 Is Your Data Correct? . . . . .	471
11.15 Eight Recommendations for International Data Quality . . . .	474
11.16 Using Regular Expressions for Data Cleaning . . . . .	477
Populating Fact and Dimension Tables . . . . .	481
11.17 Pipelining Your Surrogates . . . . .	481
11.18 Replicating Dimensions Correctly . . . . .	485
11.19 Identify Dimension Changes Using Cyclic Redundancy Checksums. . . . .	486
11.20 Maintaining Back Pointers to Operational Sources. . . . .	487
11.21 Creating Historical Dimension Rows . . . . .	488
11.22 Backward in Time . . . . .	491
11.23 Early-Arriving Facts . . . . .	494
11.24 Slowly Changing Entities . . . . .	495
11.25 Creating, Using, and Maintaining Junk Dimensions . . . . .	497
11.26 Using the SQL MERGE for Slowly Changing Dimensions . . .	499
11.27 Being Offline as Little as Possible . . . . .	502
Supporting Real Time . . . . .	503
11.28 Working in Web Time . . . . .	503
11.29 Real-Time Partitions . . . . .	507
11.30 The Real-Time Triage . . . . .	510
<b>12 Technical Architecture Considerations . . . . .</b>	<b>513</b>
Overall Technical/System Architecture . . . . .	513
12.1 Can the Data Warehouse Benefit from SOA? . . . . .	513

12.2 Picking the Right Approach to MDM . . . . .	515
12.3 Building Custom Tools for the DW/BI System . . . . .	520
12.4 Welcoming the Packaged App . . . . .	522
12.5 ERP Vendors: Bring Down Those Walls . . . . .	525
12.6 Building a Foundation for Smart Applications . . . . .	528
12.7 RFID Tags and Smart Dust . . . . .	533
Presentation Server Architecture . . . . .	535
12.8 The Aggregate Navigator . . . . .	536
12.9 Aggregate Navigation with (Almost) No Metadata . . . . .	539
12.10 Relating to OLAP . . . . .	546
12.11 Dimensional Relational versus OLAP: The Final Deployment Conundrum . . . . .	549
12.12 Dimensional Modeling for Microsoft Analysis Services . . . . .	553
12.13 Architecting Your Data for Microsoft SQL Server 2005 . . . . .	554
12.14 Microsoft SQL Server Comes of Age for Data Warehousing . . . . .	556
Front Room Architecture . . . . .	560
12.15 The Second Revolution of User Interfaces . . . . .	560
12.16 Designing the User Interface . . . . .	562
Metadata . . . . .	566
12.17 Meta Meta Data Data . . . . .	567
12.18 Creating the Metadata Strategy . . . . .	570
Infrastructure and Security Considerations . . . . .	572
12.19 Watching the Watchers . . . . .	572
12.20 Catastrophic Failure . . . . .	576
12.21 Digital Preservation . . . . .	579
12.22 Creating the Advantages of a 64-Bit Server . . . . .	582
12.23 Server Configuration Considerations . . . . .	583
12.24 Adjust Your Thinking for SANs . . . . .	585
<b>13 Front Room Business Intelligence Applications . . . . .</b>	<b>589</b>
Delivering Value with Business Intelligence . . . . .	589
13.1 The Promise of Decision Support . . . . .	589
13.2 Beyond Paving the Cow Paths . . . . .	593
13.3 Big Shifts Happening in BI . . . . .	596
13.4 Behavior: The Next Marquee Application . . . . .	598

Implementing the Business Intelligence Layer . . . . .	601
13.5 Think Like a Software Development Manager . . . . .	601
13.6 Standard Reports: Basics for Business Users . . . . .	602
13.7 Building and Delivering BI Reports. . . . .	607
13.8 The BI Portal . . . . .	610
13.9 Dashboards Done Right . . . . .	612
13.10 Don't Be Overly Reliant on Your Data Access Tool's Metadata . . . . .	613
Mining Data to Uncover Relationships . . . . .	615
13.11 Digging into Data Mining . . . . .	615
13.12 Preparing for Data Mining . . . . .	617
13.13 The Perfect Handoff . . . . .	621
13.14 Get Started with Data Mining Now. . . . .	625
Dealing with SQL. . . . .	629
13.15 Simple Drill Across in SQL . . . . .	629
13.16 The Problem with Comparisons . . . . .	631
13.17 SQL Roadblocks and Pitfalls . . . . .	635
13.18 Features for Query Tools . . . . .	638
13.19 Turbocharge Your Query Tools . . . . .	640
13.20 Smarter Data Warehouses. . . . .	644
<b>14 Maintenance and Growth Considerations. . . . .</b>	<b>651</b>
Deploying Successfully . . . . .	651
14.1 Don't Forget the Owner's Manual. . . . .	651
14.2 Let's Improve Our Operating Procedures . . . . .	655
14.3 Marketing the DW/BI System . . . . .	656
14.4 Coping with Growing Pains . . . . .	658
Sustaining for Ongoing Impact . . . . .	661
14.5 Data Warehouse Checkups. . . . .	661
14.6 Boosting Business Acceptance. . . . .	667
14.7 Educate Management to Sustain DW/BI Success . . . . .	670
14.8 Getting Your Data Warehouse Back on Track . . . . .	673
14.9 Upgrading Your BI Architecture . . . . .	674
14.10 Four Fixes for Legacy Data Warehouses. . . . .	676
14.11 A Data Warehousing Fitness Program for Lean Times. . . . .	680
Index of Articles . . . . .	685
Index . . . . .	693



# Introduction

**T**he Kimball Group's article archive has been the most popular destination on our web site ([www.kimballgroup.com](http://www.kimballgroup.com)). Stretching back fifteen years to Ralph's original 1995 *DBMS* magazine articles, the archive explores more than 200 topics, sometimes in more depth than provided by our books or courses. In recent years, it has become increasingly difficult to organize this valuable collection because it has grown by accretion; many of the topics were driven by events, opportunities, and advances in the art of data warehousing.

With *The Kimball Group Reader*, we have organized all of the articles in a much more coherent way. But *The Reader* is more than merely a collection of our past magazine articles and design tips verbatim. We have trimmed the redundancy, made sure all the articles are written with the same consistent vocabulary, and updated many of the figures. This is a new and improved remastered compilation of our writings.

## Intended Audience and Goals

The primary reader of this book should be the analyst, designer, modeler, or manager who is delivering a data warehouse in support of business intelligence. The articles in this book trace the entire lifecycle of DW/BI system development, from original business requirements gathering all the way to final deployment. We believe that this collection of articles serves as a superb reference-in-depth for literally hundreds of issues and situations that arise in the development of a DW/BI system.

The articles range from a managerial focus to a highly technical focus, although in all cases, the tone of the articles strives to be educational. These articles have been accessed hundreds of times per day on the Kimball Group web site over a span of 15 years, so we're confident they're useful. This book adds significant value by organizing the archive, and systematically editing the articles to ensure their consistency and relevance.

## Preview of Contents

Following two introductory chapters, the book's organization will look somewhat familiar to readers of *The Data Warehouse Lifecycle Toolkit, Second Edition* (Wiley, 2008) because we've organized the articles topically to correspond with the major milestones of a data warehouse/business intelligence (DW/BI) implementation. Not surprisingly given the word "Kimball" is practically synonymous with dimensional modeling, much of *The Reader* focuses on that topic in particular.

- **Chapter 1: *The Reader* at a Glance.** We begin the book with a series of articles written by Ralph several years ago for *DM Review* magazine. This series succinctly encapsulates

the Kimball approach in a cohesive manner, so it serves as a perfect overview, akin to *CliffsNotes*, for the book.

- **Chapter 2: Before You Dive In.** Long-time readers of Ralph's articles will find that this chapter is a walk down memory lane, as many of the articles are historically significant. Somewhat amazingly, the content is still very relevant even though most of these articles were written in the 1990s.
- **Chapter 3: Project/Program Planning.** With an overview and history lesson under your belt, Chapter 3 moves on to getting the DW/BI program and project launched. We consider both the project team's and sponsoring stakeholders' responsibilities, and then delve into the Kimball Lifecycle approach.
- **Chapter 4: Requirements Definition.** It is difficult to achieve DW/BI success in the absence of business requirements. This chapter delivers specific recommendations for effectively eliciting the business's needs. It stresses the importance of organizing the requirements findings around business processes, and suggests tactics for reaching organizational consensus on appropriate next steps.
- **Chapter 5: Data Architecture.** With a solid understanding of the business requirements, we turn our attention to the data (where we will remain through Chapter 11). This chapter begins with the justification for dimensional modeling. It then describes the Kimball enterprise data warehouse bus architecture, provides rationalization for the requisite integration and stewardship, and then contrasts the Kimball architecture with the Corporate Information Factory's hub-and-spoke.
- **Chapter 6: Dimensional Modeling Fundamentals.** This chapter introduces the basics of dimensional modeling, starting with distinguishing a fact from a dimension, and the core activities of drilling down, drilling across, and handling time in a data warehouse. We also explore familiar fables about dimensional models.
- **Chapter 7: Dimensional Modeling Tasks and Responsibilities.** While Chapter 6 covers the fundamental "what and why" surrounding dimensional modeling, this chapter focuses on the "how, who, and when." Chapter 7 describes the dimensional modeling process and tasks, whether starting with a blank slate or revisiting an existing model.
- **Chapter 8: Fact Table Core Concepts.** The theme for Chapter 8 could be stated as "just the facts, and nothing but the facts." We begin by discussing granularity and the three fundamental types of fact tables, and then turn our attention to fact table keys and degenerate dimensions. The chapter closes with a slew of common fact table patterns, including null, textual, and sparsely populated metrics, as well as facts that closely resemble dimension attributes.
- **Chapter 9: Dimension Table Core Concepts.** We shift our focus to dimension tables in Chapter 9, starting with a discussion of surrogate keys and the ever-present time (or date) dimensions. We then explore role playing, junk, and causal dimension patterns, before launching into a thorough handling of slowly changing dimensions. Hang onto your hats.
- **Chapter 10: More Dimension Patterns and Case Studies.** Chapter 10 complements the previous chapter with more meaty coverage of dimension tables. We describe snowflakes



and outriggers, as well as bridges for handling both multi-valued dimension attributes and ragged variable hierarchies. We discuss nuances often encountered in customer dimensions, along with internationalization issues. The chapter closes with a series of case studies covering insurance, voyage, human resources, finance, electronic commerce, text searching, and retail; we encourage everyone to peruse these vignettes as the patterns and recommendations transcend industry or application boundaries.

- **Chapter 11: Back Room ETL and Data Quality.** We switch gears from designing the target dimensional model to populating it in Chapter 11. Be forewarned: This is a hefty chapter, as you'd expect given the subject matter. We start by describing the 34 subsystems required to extract, transform, and load (ETL) the data, along with the pros and cons of using a commercial ETL tool. From there, we delve into data quality considerations, provide specific guidance for building fact and dimension tables, and discuss the implications of real-time ETL.
- **Chapter 12: Technical Architecture Considerations.** It's taken us until Chapter 12, but we're finally discussing issues surrounding the technical architecture, starting with server oriented architecture (SOA), master data management (MDM), and packaged analytics. Subsequent sections in this chapter focus on the presentation server, including the role of aggregate navigation and online analytical processing (OLAP), user interface design, metadata, infrastructure, and security.
- **Chapter 13: Front Room Business Intelligence Applications.** In Chapter 13, we step into the front room of the DW/BI system where business users are interacting with the data. We describe the lifecycle of a typical business analysis, starting with a review of historical performance but not stopping there. We then turn our attention to standardized BI reports before digging into data mining. The chapter closes by exploring the limitations of SQL for business analysis.
- **Chapter 14: Maintenance and Growth Considerations.** You've made it! Last, but not least, we provide recommendations for successfully deploying the DW/BI system, as well as keeping it healthy for sustained success.

## Navigation Aids

Given the breadth and depth of the articles in *The Kimball Group Reader*, we have very deliberately identified over two dozen articles as “Kimball Classics” because they captured a concept so effectively that we, and many others in the industry, have referred to these articles repeatedly over the past fifteen years. The classic articles are designated with a special icon that looks like this:



We expect most people will read the articles in somewhat random order, rather than digesting the book from front to back. Therefore, we have put special emphasis on *The Reader's* index as we anticipate many of you will delve in by searching the index for a particular technique or modeling situation. Just before sending the book off to the printer, we crawled through the entire *Reader*, asking the question “how would a reader locate this content in an index?” and then constructed index entries accordingly.

## Terminology Notes

We are very proud that the vocabulary established by Ralph has been so durable and broadly adopted. Kimball “marker words” including dimensions, facts, slowly changing dimensions, surrogate keys, fact table grains, factless fact tables, and degenerate dimensions, have been used consistently across the industry for more than a decade. But in spite of our best intentions, a few terms have morphed since their introduction; we have retroactively replaced the old terms with the accepted current ones.

- *Artificial keys* are now called *surrogate keys*.
- *Data mart* has been replaced with *business process dimensional model*, *business process subject area*, or just *subject area*, depending on the context.
- *Data staging* is now known as *extract, transform, and load*.
- *End-user applications* have been replaced by *business intelligence applications*.
- *Helper tables* are now *bridge tables*.

Since most people won't read this book from cover to cover, we need to introduce some common abbreviations up front:

- *DW/BI* is shorthand for the end-to-end *data warehouse/business intelligence* system. This abbreviation is useful for brevity, but it also explicitly links data warehousing and business intelligence as codependent. Finally, it reflects the shift of emphasis from the data warehouse being an end in itself to business intelligence (BI) really driving everything we do. After all, the data warehouse is the platform for all forms of BI.
- Many figures in *The Reader* include the *DD*, *FK*, and *PK* abbreviations, which stand for *degenerate dimension*, *foreign key*, and *primary key* respectively.
- *ETL* means *extract, transform, and load*, the standard paradigm for acquiring data and making it ready for exposure to BI tools.
- *ER* refers to *entity-relationship*. We frequently use *ER* when we discuss third normal form (3NF) or normalized data models, as opposed to dimensional data models.
- *OLAP* stands for *online analytical processing*, typically used to differentiate dimensional models captured in a multidimensional database or *cube* from dimensional models in a relational DBMS called *star schemas*. These relational star schemas are sometimes referred to as *ROLAP*.
- *SCD* is the abbreviation for *slowly changing dimension*, referring to the techniques we've established for handling dimension attribute changes.

# 1

## *The Reader at a Glance*

**B**eginning in late 2007, Ralph wrote a series of articles for *DM Review* magazine (now called *Information Management*). Published over a 16-month time frame, this sequence systematically describes the Kimball approach and classic best practices in a cohesive manner. Rather than scattering these articles topically throughout the book, we opted to present the series nearly in its entirety because it provides an overview of the content that follows in subsequent chapters. You can think of Chapter 1 as *CliffsNotes* for *The Kimball Group Reader*.

The chapter begins with several articles encouraging you to practice restraint and establish appropriate boundaries with other stakeholders when embarking on a data warehouse/business intelligence (DW/BI) project. From there, the series turns its attention to bringing operational data into the data warehouse and then leveraging core dimensional modeling principles to deliver robust analytic capabilities to the business users.

In addition to the articles in this chapter, Ralph also wrote a very detailed article on data quality for *DM Review*. Due to its in-depth coverage, this article is presented in Chapter 11 with other back room extract, transform, and load (ETL) topics.

## **Setting Up for Success**

Before diving into implementing the DW/BI system, make sure you assess the complete set of related requirements, while avoiding the risks of overpromising.

### **1.1 Resist the Urge to Start Coding**

*Ralph Kimball, DM Review, Nov 2007*

The most important first step in designing a DW/BI system, paradoxically, is to stop. Step back for a week, and be absolutely sure you have a sufficiently broad perspective on all the requirements that surround your project. The DW/BI design task is a daunting intellectual challenge, and it is not easy to step far enough back from the problem to protect yourself from embarrassing or career-threatening problems discovered after the project is underway.

Before cutting any code, designing any tables, or making a major hardware or software purchase, take a week to write down thoughtful, high quality answers to the following 10 questions,

each of which is a reality that will come to control your project at some point. These define the classic set of simultaneous constraints faced by every DW/BI effort.

1. *Business requirements.* Are you in touch with the key performance indicators (KPIs) your users actually need to make the decisions currently important to their enterprise? Although all 10 questions are important, understanding the business requirements is the most fundamental and far reaching. If you have a positive answer to this question, you can identify the data assets needed to support decision making, and you will be able to decide which measurement process to tackle first.
2. *Strategic data profiling.* Have you verified that your available data assets are capable of supporting the answers to question number one? The goal of strategic data profiling is to make “go/no go” decisions very early in the DW/BI project as to whether to proceed with a subject area.
3. *Tactical data profiling.* Is there a clear executive mandate to support the necessary business process re-engineering required for an effective data quality culture, perhaps even driving for Six Sigma data quality? The only real way to improve data quality is to go back to the source and figure out why better data isn’t being entered. Data entry clerks are not the cause of poor data quality! Rather, the fixes require an end-to-end awareness of the need for better quality data and a commitment from the highest levels to change how business processes work.
4. *Integration.* Is there a clear executive mandate in your organization to define common descriptors and measures across all your customer-facing processes? All of the organizations within your enterprise that participate in data integration must come to agreement on key descriptors and measures. Have your executives made it clear that this must happen?
5. *Latency.* Do you have a realistic set of requirements from business users for how quickly data must be published by the data warehouse, including as-of-yesterday, many times per day, and truly instantaneous?
6. *Compliance.* Have you received clear guidance from senior management as to which data is compliance-sensitive, and where you must guarantee that you have protected the chain of custody?
7. *Security.* Do you know how you are going to protect confidential as well as proprietary data in the ETL back room, at the users’ desktops, over the web, and on all permanent media?
8. *Archiving.* Do you have a realistic plan for very long term archiving of important data, and do you know what data should be archived?
9. *Supporting business users.* Have you profiled all your user communities to determine their abilities to use spreadsheets, construct database requests in ad hoc query tools, or just view reports on their screens?
10. *IT licenses and skill sets.* Are you prepared to rely on the major technology site licenses your organization has already committed to, and do you have enough staff with advanced skills to exploit the technical choices you make?

Time spent answering these classic DW questions is enormously valuable. Every one of the answers will affect the architecture, choice of approaches, and even the feasibility of your DW/BI project. You dare not start coding before all the team members understand what these answers mean!

The big news is that business users have seized control of the DW. They may not be building the technical infrastructure, but they are quite sure that they own the data warehouse and the BI tools and those tools must meet their needs. This transfer of initiative from IT to the users has been very obvious in the past two or three years. Witness the soul-searching articles and industry speeches exhorting CIOs to show more business leadership and the high CIO turnover as reported in *CIO Magazine* (see the April 1, 2004 issue at [www.cio.com](http://www.cio.com)).

Many of the 10 questions in this article are brought into much clearer focus by increased user ownership of the DW/BI system. Let's focus on the top five new urgent topics, in some cases coalescing our questions:

- *Business requirements.* The DW/BI system needs a permanent "KPI team" continuously in touch with business users' analytic needs and the consequent demand for new data sources to support new KPIs. Also, the system should increasingly support the full gamut of analytic applications, which include not only data delivery, but alerting the users to problems and opportunities, exploring causal factors with additional data sources, testing what-if scenarios to evaluate possible decisions, and tracking the decisions made. The DW/BI system is not just about displaying reports, but rather must be a platform for decision making in the broadest sense. The oldest label for data warehousing, *decision support*, remains surprisingly apt.
- *Strategic data profiling.* The earlier you tell the users bad news about the viability of a proposed data source, the more they will appreciate you. Develop the ability to assess a data source within a day or two. Elevate the data profiling tool to a strategic, must-have status.
- *Tactical data profiling.* The increased awareness of data quality is one of the most remarkable new DW perspectives, certainly driven by business users. But all is for naught if the business is not willing to support a quality culture and the end-to-end business process re-engineering required.
- *Integration and latency.* The user demand for the 360-degree integrated view of the business has been more like an approaching express train than a shock wave. We have been talking about it for almost a decade. But now the demands of integration, coupled with real-time access to information, have combined these two issues into a significant new architectural challenge.
- *Compliance and security.* DW/BI folks in IT often don't have the right instincts for protecting data because the system is supposed to be about exposing data. But this new emphasis on compliance and security can be built systematically into the data flows and the BI tools across the entire DW/BI solution.

The purpose of this first article has been to expose the fundamental design issues every DW/BI design team faces and to bring to the surface the urgent new requirements. In this ongoing series

of articles, I probe each of these areas in some depth, reminding us of the remarkably unchanging aspects of data warehousing, while at the same time trying to catch the winds of change.

## 1.2 Set Your Boundaries

*Ralph Kimball, DM Review, Dec 2007*

In article 1.1, *Resist the Urge to Start Coding*, I encouraged you to pause briefly before charging forward on your ambitious DW/BI project. You were supposed to use this pause to answer a checklist of major environmental questions regarding business requirements, quality data, and whether your organization is ready to attack the hard issues of integration, compliance, and security.

While answering the questions, I hope you talked to all your business user clients and sponsors who may have a stake or a responsibility in the DW/BI system. Before the memory of these conversations fades away, I suggest you make a thorough list of all the promises you made as you were selling the concept of the DW/BI system. It wouldn't surprise me if you said, "Yes, we will..."

- Tie the rolling operational results to the general ledger (GL).
- Implement effective compliance.
- Identify and implement all the key performance indicators (KPIs) needed by marketing, sales, and finance and make them available in the executive dashboard.
- Encourage the business community to add new cost drivers to our system requirements so that they can calculate activity-based costing and accurate profit across the enterprise. And while we are adding these cost drivers, we'll work out all the necessary allocation factors to assign these costs against various categories of revenue.
- Identify and implement all the customer satisfaction indicators needed by marketing.
- Seamlessly integrate all the customer-facing operational processes into a single coherent system.
- Promise to use exclusively the front end, middleware, and back end tools provided by the enterprise resource planning (ERP) vendor whose worldwide license was just signed by our CEO.
- Be the first showcase application for the new service-oriented architecture (SOA) initiative, and we'll implement, manage, and validate the new infrastructure.
- Implement and manage server virtualization for the DW/BI system. And this new system will be "green."
- Implement and manage the storage area network (SAN) for the DW/BI system.
- Implement and manage security and privacy for all data in the DW/BI system, including responsibility for the lightweight directory access protocol (LDAP) directory server and its associated authentication and authorization functions. We'll also make sure that all data accesses by the sales force in the field are secure.
- Define the requirements for long term archiving and recovery of data looking forward 20 years.