

WILEY



KIMBALL  
GROUP

# The Data Warehouse Toolkit

Third Edition

The Definitive Guide  
to Dimensional  
Modeling

Ralph Kimball  
Margy Ross





# **The Data Warehouse Toolkit**



# The Data Warehouse Toolkit

## The Definitive Guide to Dimensional Modeling

Third Edition

Ralph Kimball  
Margy Ross

**WILEY**

**The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition**

Published by

John Wiley & Sons, Inc.

10475 Crosspoint Boulevard

Indianapolis, IN 46256

[www.wiley.com](http://www.wiley.com)

Copyright © 2013 by Ralph Kimball and Margy Ross

Published by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-1-118-53080-1

ISBN: 978-1-118-53077-1 (ebk)

ISBN: 978-1-118-73228-1 (ebk)

ISBN: 978-1-118-73219-9 (ebk)

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 website may provide or recommendations it may make. Further, readers should be aware that Internet websites 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 publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

**Library of Congress Control Number:** 2013936841

**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. John Wiley & Sons, 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 data warehouse and business intelligence industry's thought leader on the dimensional approach. He has educated tens of thousands of IT professionals. The Toolkit books written by Ralph and his colleagues have been the industry's best sellers since 1996. Prior to working at Metaphor and founding Red Brick Systems, Ralph coined the Star workstation, the first commercial product with windows, icons, and a mouse, at Xerox's Palo Alto Research Center (PARC). Ralph has a PhD in electrical engineering from Stanford University.

**Margy Ross** is president of the Kimball Group. She has focused exclusively on data warehousing and business intelligence since 1982 with an emphasis on business requirements and dimensional modeling. Like Ralph, Margy has taught the dimensional best practices to thousands of students; she also coauthored five Toolkit books with Ralph. Margy previously worked at Metaphor and cofounded DecisionWorks Consulting. She graduated with a BS in industrial engineering from Northwestern University.

# Credits

**Executive Editor**

Robert Elliott

**Project Editor**

Maureen Spears

**Senior Production Editor**

Kathleen Wisor

**Copy Editor**

Apostrophe Editing Services

**Editorial Manager**

Mary Beth Wakefield

**Freelancer Editorial Manager**

Rosemarie Graham

**Associate Director of Marketing**

David Mayhew

**Marketing Manager**

Ashley Zurcher

**Business Manager**

Amy Knies

**Production Manager**

Tim Tate

**Vice President and Executive Group****Publisher**

Richard Swadley

**Vice President and Executive Publisher**

Neil Edde

**Associate Publisher**

Jim Minatel

**Project Coordinator, Cover**

Katie Crocker

**Proofreader**

Word One, New York

**Indexer**

Johnna VanHoose Dinse

**Cover Image**

iStockphoto.com / teekid

**Cover Designer**

Ryan Sneed



# Acknowledgments

**F**irst, thanks to the hundreds of thousands 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. Collectively, you have had a profoundly positive impact on the data warehousing and business intelligence industry. Congratulations!

Our Kimball Group colleagues, Bob Becker, Joy Mundy, and Warren Thornthwaite, have worked with us to apply the techniques described in this book literally thousands of times, over nearly 30 years of working together. Every technique in this book has been thoroughly vetted by practice in the real world. We appreciate their input and feedback on this book—and more important, the years we have shared as business partners, along with Julie Kimball.

Bob Elliott, our executive editor at John Wiley & Sons, project editor Maureen Spears, and the rest of the Wiley team have supported this project with skill and enthusiasm. As always, it has been a pleasure to work with them.

To our families, thank you for your unconditional support throughout our careers. Spouses Julie Kimball and Scott Ross and children Sara Hayden Smith, Brian Kimball, and Katie Ross all contributed in countless ways to this book.



# Contents

Introduction .....	xxvii
<b>1 Data Warehousing, Business Intelligence, and Dimensional Modeling Primer .....</b>	<b>1</b>
Different Worlds of Data Capture and Data Analysis .....	2
Goals of Data Warehousing and Business Intelligence .....	3
Publishing Metaphor for DW/BI Managers .....	5
Dimensional Modeling Introduction .....	7
Star Schemas Versus OLAP Cubes .....	8
Fact Tables for Measurements .....	10
Dimension Tables for Descriptive Context .....	13
Facts and Dimensions Joined in a Star Schema .....	16
Kimball's DW/BI Architecture .....	18
Operational Source Systems .....	18
Extract, Transformation, and Load System .....	19
Presentation Area to Support Business Intelligence .....	21
Business Intelligence Applications .....	22
Restaurant Metaphor for the Kimball Architecture .....	23
Alternative DW/BI Architectures .....	26
Independent Data Mart Architecture .....	26
Hub-and-Spoke Corporate Information Factory Inmon Architecture ..	28
Hybrid Hub-and-Spoke and Kimball Architecture .....	29
Dimensional Modeling Myths .....	30
Myth 1: Dimensional Models are Only for Summary Data .....	30
Myth 2: Dimensional Models are Departmental, Not Enterprise .....	31
Myth 3: Dimensional Models are Not Scalable .....	31
Myth 4: Dimensional Models are Only for Predictable Usage .....	31
Myth 5: Dimensional Models Can't Be Integrated .....	32
More Reasons to Think Dimensionally .....	32
Agile Considerations .....	34
Summary .....	35

<b>2</b>	<b>Kimball Dimensional Modeling Techniques Overview. . . . .</b>	<b>37</b>
	Fundamental Concepts. . . . .	37
	Gather Business Requirements and Data Realities. . . . .	37
	Collaborative Dimensional Modeling Workshops. . . . .	38
	Four-Step Dimensional Design Process. . . . .	38
	Business Processes . . . . .	39
	Grain . . . . .	39
	Dimensions for Descriptive Context. . . . .	40
	Facts for Measurements . . . . .	40
	Star Schemas and OLAP Cubes . . . . .	40
	Graceful Extensions to Dimensional Models. . . . .	41
	Basic Fact Table Techniques . . . . .	41
	Fact Table Structure . . . . .	41
	Additive, Semi-Additive, Non-Additive Facts . . . . .	42
	Nulls in Fact Tables . . . . .	42
	Conformed Facts . . . . .	42
	Transaction Fact Tables . . . . .	43
	Periodic Snapshot Fact Tables . . . . .	43
	Accumulating Snapshot Fact Tables . . . . .	44
	Factless Fact Tables . . . . .	44
	Aggregate Fact Tables or OLAP Cubes . . . . .	45
	Consolidated Fact Tables . . . . .	45
	Basic Dimension Table Techniques . . . . .	46
	Dimension Table Structure. . . . .	46
	Dimension Surrogate Keys . . . . .	46
	Natural, Durable, and Supernatural Keys . . . . .	46
	Drilling Down . . . . .	47
	Degenerate Dimensions. . . . .	47
	Denormalized Flattened Dimensions . . . . .	47
	Multiple Hierarchies in Dimensions . . . . .	48
	Flags and Indicators as Textual Attributes. . . . .	48
	Null Attributes in Dimensions . . . . .	48
	Calendar Date Dimensions. . . . .	48
	Role-Playing Dimensions . . . . .	49
	Junk Dimensions . . . . .	49

Snowflaked Dimensions . . . . .	50
Outrigger Dimensions . . . . .	50
Integration via Conformed Dimensions . . . . .	50
Conformed Dimensions . . . . .	51
Shrunken Dimensions . . . . .	51
Drilling Across . . . . .	51
Value Chain . . . . .	52
Enterprise Data Warehouse Bus Architecture . . . . .	52
Enterprise Data Warehouse Bus Matrix. . . . .	52
Detailed Implementation Bus Matrix . . . . .	53
Opportunity/Stakeholder Matrix . . . . .	53
Dealing with Slowly Changing Dimension Attributes . . . . .	53
Type 0: Retain Original . . . . .	54
Type 1: Overwrite . . . . .	54
Type 2: Add New Row . . . . .	54
Type 3: Add New Attribute . . . . .	55
Type 4: Add Mini-Dimension . . . . .	55
Type 5: Add Mini-Dimension and Type 1 Outrigger. . . . .	55
Type 6: Add Type 1 Attributes to Type 2 Dimension. . . . .	56
Type 7: Dual Type 1 and Type 2 Dimensions . . . . .	56
Dealing with Dimension Hierarchies. . . . .	56
Fixed Depth Positional Hierarchies . . . . .	56
Slightly Ragged/Variable Depth Hierarchies . . . . .	57
Ragged/Variable Depth Hierarchies with Hierarchy Bridge Tables . . . . .	57
Ragged/Variable Depth Hierarchies with Pathstring Attributes. . . . .	57
Advanced Fact Table Techniques . . . . .	58
Fact Table Surrogate Keys. . . . .	58
Centipede Fact Tables . . . . .	58
Numeric Values as Attributes or Facts . . . . .	59
Lag/Duration Facts. . . . .	59
Header/Line Fact Tables . . . . .	59
Allocated Facts. . . . .	60
Profit and Loss Fact Tables Using Allocations . . . . .	60
Multiple Currency Facts. . . . .	60
Multiple Units of Measure Facts. . . . .	61

Year-to-Date Facts . . . . .	61
Multipass SQL to Avoid Fact-to-Fact Table Joins . . . . .	61
Timespan Tracking in Fact Tables . . . . .	62
Late Arriving Facts . . . . .	62
Advanced Dimension Techniques . . . . .	62
Dimension-to-Dimension Table Joins . . . . .	62
Multivalued Dimensions and Bridge Tables . . . . .	63
Time Varying Multivalued Bridge Tables . . . . .	63
Behavior Tag Time Series . . . . .	63
Behavior Study Groups . . . . .	64
Aggregated Facts as Dimension Attributes . . . . .	64
Dynamic Value Bands . . . . .	64
Text Comments Dimension . . . . .	65
Multiple Time Zones . . . . .	65
Measure Type Dimensions . . . . .	65
Step Dimensions . . . . .	65
Hot Swappable Dimensions . . . . .	66
Abstract Generic Dimensions . . . . .	66
Audit Dimensions . . . . .	66
Late Arriving Dimensions . . . . .	67
Special Purpose Schemas . . . . .	67
Supertype and Subtype Schemas for Heterogeneous Products . . . . .	67
Real-Time Fact Tables . . . . .	68
Error Event Schemas . . . . .	68
<b>3 Retail Sales . . . . .</b>	<b>69</b>
Four-Step Dimensional Design Process . . . . .	70
Step 1: Select the Business Process . . . . .	70
Step 2: Declare the Grain . . . . .	71
Step 3: Identify the Dimensions . . . . .	72
Step 4: Identify the Facts . . . . .	72
Retail Case Study . . . . .	72
Step 1: Select the Business Process . . . . .	74
Step 2: Declare the Grain . . . . .	74
Step 3: Identify the Dimensions . . . . .	76

Step 4: Identify the Facts . . . . .	76
Dimension Table Details . . . . .	79
Date Dimension . . . . .	79
Product Dimension . . . . .	83
Store Dimension . . . . .	87
Promotion Dimension . . . . .	89
Other Retail Sales Dimensions . . . . .	92
Degenerate Dimensions for Transaction Numbers . . . . .	93
Retail Schema in Action . . . . .	94
Retail Schema Extensibility . . . . .	95
Factless Fact Tables . . . . .	97
Dimension and Fact Table Keys . . . . .	98
Dimension Table Surrogate Keys . . . . .	98
Dimension Natural and Durable Supernatural Keys . . . . .	100
Degenerate Dimension Surrogate Keys . . . . .	101
Date Dimension Smart Keys . . . . .	101
Fact Table Surrogate Keys . . . . .	102
Resisting Normalization Urges . . . . .	104
Snowflake Schemas with Normalized Dimensions . . . . .	104
Outriggers . . . . .	106
Centipede Fact Tables with Too Many Dimensions . . . . .	108
Summary . . . . .	109
<b>4 Inventory . . . . .</b>	<b>111</b>
Value Chain Introduction . . . . .	111
Inventory Models . . . . .	112
Inventory Periodic Snapshot . . . . .	113
Inventory Transactions . . . . .	116
Inventory Accumulating Snapshot . . . . .	118
Fact Table Types . . . . .	119
Transaction Fact Tables . . . . .	120
Periodic Snapshot Fact Tables . . . . .	120
Accumulating Snapshot Fact Tables . . . . .	121
Complementary Fact Table Types . . . . .	122

Value Chain Integration . . . . .	122
Enterprise Data Warehouse Bus Architecture. . . . .	123
Understanding the Bus Architecture . . . . .	124
Enterprise Data Warehouse Bus Matrix. . . . .	125
Conformed Dimensions . . . . .	130
Drilling Across Fact Tables . . . . .	130
Identical Conformed Dimensions. . . . .	131
Shrunk Rollup Conformed Dimension with Attribute Subset . . . . .	132
Shrunk Conformed Dimension with Row Subset . . . . .	132
Shrunk Conformed Dimensions on the Bus Matrix. . . . .	134
Limited Conformity . . . . .	135
Importance of Data Governance and Stewardship . . . . .	135
Conformed Dimensions and the Agile Movement . . . . .	137
Conformed Facts . . . . .	138
Summary . . . . .	139
<b>5 Procurement . . . . .</b>	<b>141</b>
Procurement Case Study . . . . .	141
Procurement Transactions and Bus Matrix . . . . .	142
Single Versus Multiple Transaction Fact Tables . . . . .	143
Complementary Procurement Snapshot. . . . .	147
Slowly Changing Dimension Basics . . . . .	147
Type 0: Retain Original . . . . .	148
Type 1: Overwrite . . . . .	149
Type 2: Add New Row . . . . .	150
Type 3: Add New Attribute . . . . .	154
Type 4: Add Mini-Dimension . . . . .	156
Hybrid Slowly Changing Dimension Techniques. . . . .	159
Type 5: Mini-Dimension and Type 1 Outrigger . . . . .	160
Type 6: Add Type 1 Attributes to Type 2 Dimension. . . . .	160
Type 7: Dual Type 1 and Type 2 Dimensions . . . . .	162
Slowly Changing Dimension Recap . . . . .	164
Summary . . . . .	165



<b>6</b>	<b>Order Management. . . . .</b>	<b>167</b>
	Order Management Bus Matrix . . . . .	168
	Order Transactions . . . . .	168
	Fact Normalization. . . . .	169
	Dimension Role Playing . . . . .	170
	Product Dimension Revisited . . . . .	172
	Customer Dimension . . . . .	174
	Deal Dimension . . . . .	177
	Degenerate Dimension for Order Number . . . . .	178
	Junk Dimensions . . . . .	179
	Header/Line Pattern to Avoid . . . . .	181
	Multiple Currencies . . . . .	182
	Transaction Facts at Different Granularity . . . . .	184
	Another Header/Line Pattern to Avoid . . . . .	186
	Invoice Transactions . . . . .	187
	Service Level Performance as Facts, Dimensions, or Both . . . . .	188
	Profit and Loss Facts. . . . .	189
	Audit Dimension . . . . .	192
	Accumulating Snapshot for Order Fulfillment Pipeline . . . . .	194
	Lag Calculations. . . . .	196
	Multiple Units of Measure . . . . .	197
	Beyond the Rearview Mirror . . . . .	198
	Summary . . . . .	199
<b>7</b>	<b>Accounting . . . . .</b>	<b>201</b>
	Accounting Case Study and Bus Matrix . . . . .	202
	General Ledger Data . . . . .	203
	General Ledger Periodic Snapshot . . . . .	203
	Chart of Accounts . . . . .	203
	Period Close. . . . .	204
	Year-to-Date Facts . . . . .	206
	Multiple Currencies Revisited . . . . .	206
	General Ledger Journal Transactions . . . . .	206

Multiple Fiscal Accounting Calendars . . . . .	208
Drilling Down Through a Multilevel Hierarchy . . . . .	209
Financial Statements . . . . .	209
Budgeting Process . . . . .	210
Dimension Attribute Hierarchies . . . . .	214
Fixed Depth Positional Hierarchies . . . . .	214
Slightly Ragged Variable Depth Hierarchies . . . . .	214
Ragged Variable Depth Hierarchies . . . . .	215
Shared Ownership in a Ragged Hierarchy . . . . .	219
Time Varying Ragged Hierarchies . . . . .	220
Modifying Ragged Hierarchies . . . . .	220
Alternative Ragged Hierarchy Modeling Approaches . . . . .	221
Advantages of the Bridge Table Approach for Ragged Hierarchies . . . . .	223
Consolidated Fact Tables . . . . .	224
Role of OLAP and Packaged Analytic Solutions . . . . .	226
Summary . . . . .	227
<b>8 Customer Relationship Management . . . . .</b>	<b>229</b>
CRM Overview . . . . .	230
Operational and Analytic CRM . . . . .	231
Customer Dimension Attributes . . . . .	233
Name and Address Parsing . . . . .	233
International Name and Address Considerations . . . . .	236
Customer-Centric Dates . . . . .	238
Aggregated Facts as Dimension Attributes . . . . .	239
Segmentation Attributes and Scores . . . . .	240
Counts with Type 2 Dimension Changes . . . . .	243
Outrigger for Low Cardinality Attribute Set . . . . .	243
Customer Hierarchy Considerations . . . . .	244
Bridge Tables for Multivalued Dimensions . . . . .	245
Bridge Table for Sparse Attributes . . . . .	247
Bridge Table for Multiple Customer Contacts . . . . .	248
Complex Customer Behavior . . . . .	249
Behavior Study Groups for Cohorts . . . . .	249

Step Dimension for Sequential Behavior . . . . .	251
Timespan Fact Tables . . . . .	252
Tagging Fact Tables with Satisfaction Indicators . . . . .	254
Tagging Fact Tables with Abnormal Scenario Indicators . . . . .	255
Customer Data Integration Approaches . . . . .	256
Master Data Management Creating a Single Customer Dimension . . . . .	256
Partial Conformity of Multiple Customer Dimensions . . . . .	258
Avoiding Fact-to-Fact Table Joins . . . . .	259
Low Latency Reality Check . . . . .	260
Summary . . . . .	261
<b>9 Human Resources Management . . . . .</b>	<b>263</b>
Employee Profile Tracking . . . . .	263
Precise Effective and Expiration Timespans . . . . .	265
Dimension Change Reason Tracking . . . . .	266
Profile Changes as Type 2 Attributes or Fact Events . . . . .	267
Headcount Periodic Snapshot . . . . .	267
Bus Matrix for HR Processes . . . . .	268
Packaged Analytic Solutions and Data Models . . . . .	270
Recursive Employee Hierarchies . . . . .	271
Change Tracking on Embedded Manager Key . . . . .	272
Drilling Up and Down Management Hierarchies . . . . .	273
Multivalued Skill Keyword Attributes . . . . .	274
Skill Keyword Bridge . . . . .	275
Skill Keyword Text String . . . . .	276
Survey Questionnaire Data . . . . .	277
Text Comments . . . . .	278
Summary . . . . .	279
<b>10 Financial Services . . . . .</b>	<b>281</b>
Banking Case Study and Bus Matrix . . . . .	282
Dimension Triage to Avoid Too Few Dimensions . . . . .	283
Household Dimension . . . . .	286
Multivalued Dimensions and Weighting Factors . . . . .	287

Mini-Dimensions Revisited . . . . .	289
Adding a Mini-Dimension to a Bridge Table. . . . .	290
Dynamic Value Banding of Facts . . . . .	291
Supertype and Subtype Schemas for Heterogeneous Products. . . . .	293
Supertype and Subtype Products with Common Facts . . . . .	295
Hot Swappable Dimensions . . . . .	296
Summary . . . . .	296
<b>11 Telecommunications . . . . .</b>	<b>297</b>
Telecommunications Case Study and Bus Matrix . . . . .	297
General Design Review Considerations. . . . .	299
Balance Business Requirements and Source Realities . . . . .	300
Focus on Business Processes. . . . .	300
Granularity . . . . .	300
Single Granularity for Facts . . . . .	301
Dimension Granularity and Hierarchies . . . . .	301
Date Dimension. . . . .	302
Degenerate Dimensions. . . . .	303
Surrogate Keys. . . . .	303
Dimension Decodes and Descriptions . . . . .	303
Conformity Commitment . . . . .	304
Design Review Guidelines. . . . .	304
Draft Design Exercise Discussion . . . . .	306
Remodeling Existing Data Structures . . . . .	309
Geographic Location Dimension . . . . .	310
Summary . . . . .	310
<b>12 Transportation . . . . .</b>	<b>311</b>
Airline Case Study and Bus Matrix . . . . .	311
Multiple Fact Table Granularities . . . . .	312
Linking Segments into Trips . . . . .	315
Related Fact Tables. . . . .	316
Extensions to Other Industries . . . . .	317
Cargo Shipper . . . . .	317
Travel Services . . . . .	317

Combining Correlated Dimensions . . . . .	318
Class of Service . . . . .	319
Origin and Destination . . . . .	320
More Date and Time Considerations . . . . .	321
Country-Specific Calendars as Outriggers . . . . .	321
Date and Time in Multiple Time Zones . . . . .	323
Localization Recap . . . . .	324
Summary . . . . .	324
<b>13 Education . . . . .</b>	<b>325</b>
University Case Study and Bus Matrix . . . . .	325
Accumulating Snapshot Fact Tables . . . . .	326
Applicant Pipeline . . . . .	326
Research Grant Proposal Pipeline . . . . .	329
Factless Fact Tables . . . . .	329
Admissions Events . . . . .	330
Course Registrations . . . . .	330
Facility Utilization . . . . .	334
Student Attendance . . . . .	335
More Educational Analytic Opportunities . . . . .	336
Summary . . . . .	336
<b>14 Healthcare . . . . .</b>	<b>339</b>
Healthcare Case Study and Bus Matrix . . . . .	339
Claims Billing and Payments . . . . .	342
Date Dimension Role Playing . . . . .	345
Multivalued Diagnoses . . . . .	345
Supertypes and Subtypes for Charges . . . . .	347
Electronic Medical Records . . . . .	348
Measure Type Dimension for Sparse Facts . . . . .	349
Freeform Text Comments . . . . .	350
Images . . . . .	350
Facility/Equipment Inventory Utilization . . . . .	351
Dealing with Retroactive Changes . . . . .	351
Summary . . . . .	352

<b>15</b>	<b>Electronic Commerce. . . . .</b>	<b>353</b>
	Clickstream Source Data. . . . .	353
	Clickstream Data Challenges . . . . .	354
	Clickstream Dimensional Models . . . . .	357
	Page Dimension. . . . .	358
	Event Dimension . . . . .	359
	Session Dimension. . . . .	359
	Referral Dimension. . . . .	360
	Clickstream Session Fact Table . . . . .	361
	Clickstream Page Event Fact Table . . . . .	363
	Step Dimension . . . . .	366
	Aggregate Clickstream Fact Tables. . . . .	366
	Google Analytics . . . . .	367
	Integrating Clickstream into Web Retailer's Bus Matrix . . . . .	368
	Profitability Across Channels Including Web . . . . .	370
	Summary . . . . .	373
<b>16</b>	<b>Insurance. . . . .</b>	<b>375</b>
	Insurance Case Study . . . . .	376
	Insurance Value Chain . . . . .	377
	Draft Bus Matrix . . . . .	378
	Policy Transactions . . . . .	379
	Dimension Role Playing . . . . .	380
	Slowly Changing Dimensions. . . . .	380
	Mini-Dimensions for Large or Rapidly Changing Dimensions. . . . .	381
	Multivalued Dimension Attributes . . . . .	382
	Numeric Attributes as Facts or Dimensions . . . . .	382
	Degenerate Dimension . . . . .	383
	Low Cardinality Dimension Tables . . . . .	383
	Audit Dimension . . . . .	383
	Policy Transaction Fact Table . . . . .	383
	Heterogeneous Supertype and Subtype Products . . . . .	384
	Complementary Policy Accumulating Snapshot . . . . .	384
	Premium Periodic Snapshot . . . . .	385
	Conformed Dimensions. . . . .	386
	Conformed Facts . . . . .	386

Pay-in-Advance Facts . . . . .	386
Heterogeneous Supertypes and Subtypes Revisited . . . . .	387
Multivalued Dimensions Revisited . . . . .	388
More Insurance Case Study Background . . . . .	388
Updated Insurance Bus Matrix . . . . .	389
Detailed Implementation Bus Matrix . . . . .	390
Claim Transactions . . . . .	390
Transaction Versus Profile Junk Dimensions . . . . .	392
Claim Accumulating Snapshot . . . . .	392
Accumulating Snapshot for Complex Workflows . . . . .	393
Timespan Accumulating Snapshot . . . . .	394
Periodic Instead of Accumulating Snapshot . . . . .	395
Policy/Claim Consolidated Periodic Snapshot . . . . .	395
Factless Accident Events . . . . .	396
Common Dimensional Modeling Mistakes to Avoid . . . . .	397
Mistake 10: Place Text Attributes in a Fact Table. . . . .	397
Mistake 9: Limit Verbose Descriptors to Save Space . . . . .	398
Mistake 8: Split Hierarchies into Multiple Dimensions . . . . .	398
Mistake 7: Ignore the Need to Track Dimension Changes . . . . .	398
Mistake 6: Solve All Performance Problems with More Hardware. . . . .	399
Mistake 5: Use Operational Keys to Join Dimensions and Facts. . . . .	399
Mistake 4: Neglect to Declare and Comply with the Fact Grain . . . . .	399
Mistake 3: Use a Report to Design the Dimensional Model . . . . .	400
Mistake 2: Expect Users to Query Normalized Atomic Data . . . . .	400
Mistake 1: Fail to Conform Facts and Dimensions . . . . .	400
Summary . . . . .	401
<b>17 Kimball DW/BI Lifecycle Overview . . . . .</b>	<b>403</b>
Lifecycle Roadmap . . . . .	404
Roadmap Mile Markers . . . . .	405
Lifecycle Launch Activities . . . . .	406
Program/Project Planning and Management . . . . .	406
Business Requirements Definition . . . . .	410
Lifecycle Technology Track . . . . .	416
Technical Architecture Design . . . . .	416
Product Selection and Installation . . . . .	418

Lifecycle Data Track . . . . .	420
Dimensional Modeling. . . . .	420
Physical Design . . . . .	420
ETL Design and Development . . . . .	422
Lifecycle BI Applications Track . . . . .	422
BI Application Specification . . . . .	423
BI Application Development . . . . .	423
Lifecycle Wrap-up Activities . . . . .	424
Deployment. . . . .	424
Maintenance and Growth . . . . .	425
Common Pitfalls to Avoid . . . . .	426
Summary . . . . .	427
<b>18 Dimensional Modeling Process and Tasks. . . . .</b>	<b>429</b>
Modeling Process Overview . . . . .	429
Get Organized . . . . .	431
Identify Participants, Especially Business Representatives . . . . .	431
Review the Business Requirements. . . . .	432
Leverage a Modeling Tool . . . . .	432
Leverage a Data Profiling Tool . . . . .	433
Leverage or Establish Naming Conventions . . . . .	433
Coordinate Calendars and Facilities . . . . .	433
Design the Dimensional Model . . . . .	434
Reach Consensus on High-Level Bubble Chart . . . . .	435
Develop the Detailed Dimensional Model . . . . .	436
Review and Validate the Model . . . . .	439
Finalize the Design Documentation . . . . .	441
Summary . . . . .	441
<b>19 ETL Subsystems and Techniques . . . . .</b>	<b>443</b>
Round Up the Requirements. . . . .	444
Business Needs . . . . .	444
Compliance . . . . .	445
Data Quality . . . . .	445
Security . . . . .	446
Data Integration . . . . .	446



Data Latency . . . . .	447
Archiving and Lineage . . . . .	447
BI Delivery Interfaces . . . . .	448
Available Skills . . . . .	448
Legacy Licenses . . . . .	449
The 34 Subsystems of ETL . . . . .	449
Extracting: Getting Data into the Data Warehouse . . . . .	450
Subsystem 1: Data Profiling . . . . .	450
Subsystem 2: Change Data Capture System . . . . .	451
Subsystem 3: Extract System . . . . .	453
Cleaning and Conforming Data . . . . .	455
Improving Data Quality Culture and Processes . . . . .	455
Subsystem 4: Data Cleansing System . . . . .	456
Subsystem 5: Error Event Schema . . . . .	458
Subsystem 6: Audit Dimension Assembler . . . . .	460
Subsystem 7: Deduplication System . . . . .	460
Subsystem 8: Conforming System . . . . .	461
Delivering: Prepare for Presentation . . . . .	463
Subsystem 9: Slowly Changing Dimension Manager . . . . .	464
Subsystem 10: Surrogate Key Generator . . . . .	469
Subsystem 11: Hierarchy Manager . . . . .	470
Subsystem 12: Special Dimensions Manager . . . . .	470
Subsystem 13: Fact Table Builders . . . . .	473
Subsystem 14: Surrogate Key Pipeline . . . . .	475
Subsystem 15: Multivalued Dimension Bridge Table Builder . . . . .	477
Subsystem 16: Late Arriving Data Handler . . . . .	478
Subsystem 17: Dimension Manager System . . . . .	479
Subsystem 18: Fact Provider System . . . . .	480
Subsystem 19: Aggregate Builder . . . . .	481
Subsystem 20: OLAP Cube Builder . . . . .	481
Subsystem 21: Data Propagation Manager . . . . .	482
Managing the ETL Environment . . . . .	483
Subsystem 22: Job Scheduler . . . . .	483
Subsystem 23: Backup System . . . . .	485
Subsystem 24: Recovery and Restart System . . . . .	486

Subsystem 25: Version Control System . . . . .	488
Subsystem 26: Version Migration System . . . . .	488
Subsystem 27: Workflow Monitor . . . . .	489
Subsystem 28: Sorting System . . . . .	490
Subsystem 29: Lineage and Dependency Analyzer . . . . .	490
Subsystem 30: Problem Escalation System . . . . .	491
Subsystem 31: Parallelizing/Pipelining System . . . . .	492
Subsystem 32: Security System . . . . .	492
Subsystem 33: Compliance Manager . . . . .	493
Subsystem 34: Metadata Repository Manager . . . . .	495
Summary . . . . .	496
<b>20 ETL System Design and Development Process and Tasks . . . . .</b>	<b>497</b>
ETL Process Overview . . . . .	497
Develop the ETL Plan . . . . .	498
Step 1: Draw the High-Level Plan . . . . .	498
Step 2: Choose an ETL Tool . . . . .	499
Step 3: Develop Default Strategies . . . . .	500
Step 4: Drill Down by Target Table . . . . .	500
Develop the ETL Specification Document . . . . .	502
Develop One-Time Historic Load Processing . . . . .	503
Step 5: Populate Dimension Tables with Historic Data . . . . .	503
Step 6: Perform the Fact Table Historic Load . . . . .	508
Develop Incremental ETL Processing . . . . .	512
Step 7: Dimension Table Incremental Processing . . . . .	512
Step 8: Fact Table Incremental Processing . . . . .	515
Step 9: Aggregate Table and OLAP Loads . . . . .	519
Step 10: ETL System Operation and Automation . . . . .	519
Real-Time Implications . . . . .	520
Real-Time Triage . . . . .	521
Real-Time Architecture Trade-Offs . . . . .	522
Real-Time Partitions in the Presentation Server . . . . .	524
Summary . . . . .	526

<b>21</b>	<b>Big Data Analytics . . . . .</b>	<b>527</b>
	Big Data Overview . . . . .	527
	Extended RDBMS Architecture. . . . .	529
	MapReduce/Hadoop Architecture . . . . .	530
	Comparison of Big Data Architectures . . . . .	530
	Recommended Best Practices for Big Data . . . . .	531
	Management Best Practices for Big Data . . . . .	531
	Architecture Best Practices for Big Data . . . . .	533
	Data Modeling Best Practices for Big Data . . . . .	538
	Data Governance Best Practices for Big Data . . . . .	541
	Summary . . . . .	542
	 Index. . . . .	 543



# Introduction

**T**he data warehousing and business intelligence (DW/BI) 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, DW/BI has since been embraced by organizations of all sizes. The industry has built thousands of DW/BI systems. The volume of data continues to grow as warehouses are populated with increasingly atomic data and updated with greater frequency. Over the course of our careers, we have seen databases grow from megabytes to gigabytes to terabytes to petabytes, yet the basic challenge of DW/BI systems has remained remarkably constant. Our job is to marshal an organization's data and bring it to business users for their decision making. Collectively, you've delivered on this objective; business professionals everywhere are making better decisions and generating payback on their DW/BI investments.

Since the first edition of *The Data Warehouse Toolkit* was published, dimensional modeling has been broadly accepted as the dominant technique for DW/BI presentation. Practitioners and pundits alike have recognized that the presentation of data must be grounded in simplicity if it is to stand any chance of success. Simplicity is the fundamental key that allows users to easily understand databases and software to efficiently navigate databases. 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, you establish a coherent design that serves the organization's analytic needs. This dimensionally modeled framework becomes the *platform for BI*. 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 DW/BI initiative.

Dimensional modeling also has emerged as the leading architecture for building integrated DW/BI systems. When you use the conformed dimensions and conformed facts of a set of dimensional models, you have a practical and predictable framework for incrementally building complex DW/BI systems that are inherently distributed.

For all that has changed in our industry, the core dimensional modeling techniques that Ralph Kimball published 17 years ago have withstood the test of time. Concepts such as conformed dimensions, slowly changing dimensions, heterogeneous products, factless fact tables, and the enterprise data warehouse bus matrix

continue to be discussed in design workshops around the globe. The original concepts have been embellished and enhanced by new and complementary techniques. We decided to publish this third edition of Kimball's seminal work because we felt that it would be useful to summarize our collective dimensional modeling experience under a single cover. We have each focused exclusively on decision support, data warehousing, and business intelligence for more than three decades. We want to share the dimensional modeling patterns that have emerged repeatedly during the course of our 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 address the needs of those just starting in dimensional DW/BI and we 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. It's the definitive guide.

## Intended Audience

---

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

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

For those of you who have read the earlier editions 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, including sample enterprise data warehouse bus matrices for nearly every case study. We have developed vignettes for new subject areas, including big data analytics.

The content in this book is somewhat technical. We primarily discuss dimensional modeling in the context of a relational database with nuances for online