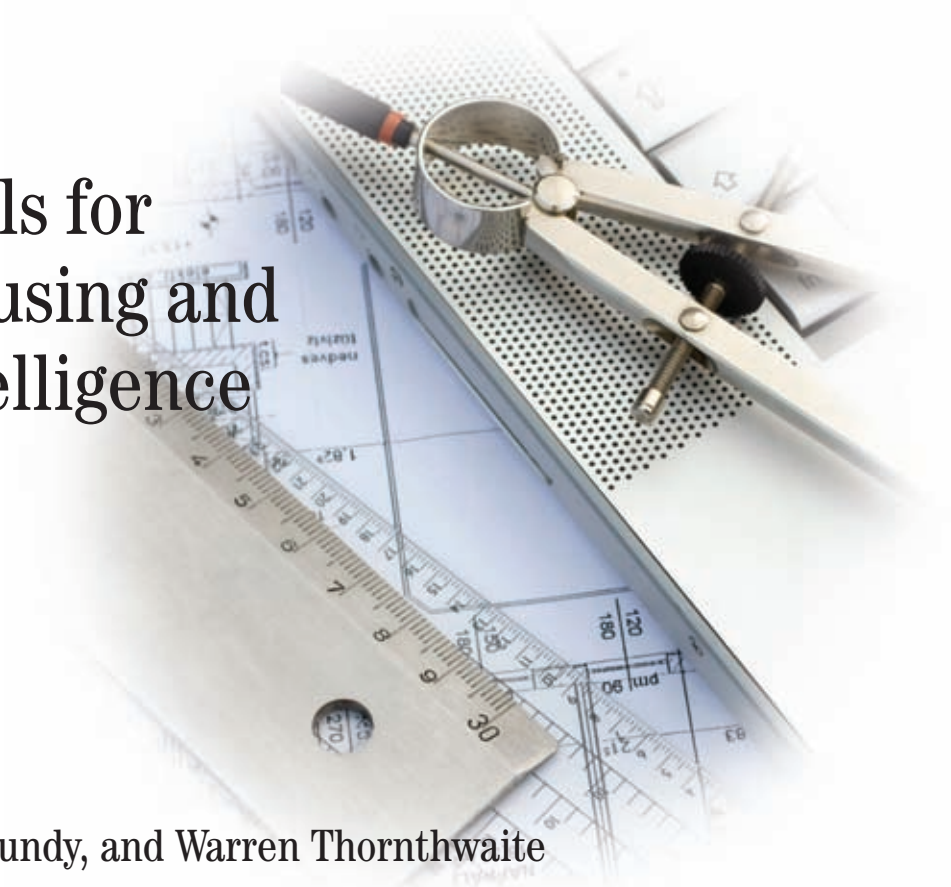
WILEY

KIMBALL GROUP

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

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

Library of Congress Control Number: 2009942822

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 Shlaer

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

First, 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

Contents

Introduction

The Kimball Group's article archive has been the most popular destination on our web site (`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

Beginning 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`).

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.