WILEY

KIMBALL GROUP

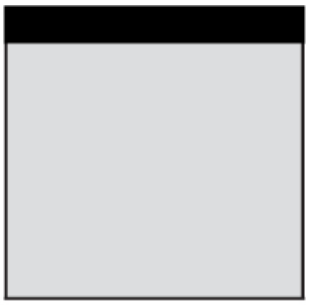
The Data Warehouse Lifecycle Toolkit

Second Edition

Practical Techniques for Building Data Warehouse and Business Intelligence Systems

Ralph Kimball
Margy Ross
Warren Thornthwaite
Joy Mundy
Bob Becker

Table of Contents



The Data Warehouse Lifecycle Toolkit

Second Edition

Ralph Kimball
Margy Ross
Warren Thornthwaite
Joy Mundy
Bob Becker

Wiley Publishing, Inc.

The Data Warehouse Lifecycle Toolkit, Second Edition

Published by

Wiley Publishing, Inc.

10475 Crosspoint Boulevard

Indianapolis, IN 46256

www.wiley.com

Published byWiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-0-470-14977-5

Library of Congress Cataloging-in-Publication Data:

The data warehouse lifecycle toolkit / Ralph Kimball... [et al.]. -- 2nd ed.

p. cm.

Includes index.

ISBN 978-0-470-14977-5 (paper/website)

1. Data warehousing. I. Kimball, Ralph.

QA76.9.D37D38 2007

005.74--dc22

2007040691

About the Authors

The authors' professional careers have followed remarkably similar paths. Each author has focused on data warehousing and business intelligence (DW/BI) consulting and education for more than fifteen years. Most worked together at Metaphor Computer Systems, a pioneering decision support vendor, in the 1980s. All the authors are members of the Kimball Group and teach for Kimball University. They contribute regularly to *Intelligent Enterprise* magazine and other industry publications; most have previously written books in the *Toolkit* series.

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 trained more than 10,000 IT professionals. Ralph has his Ph.D. in Electrical Engineering from Stanford University.

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

Warren Thornthwaite began his DW/BI career in 1980. After managing Metaphor's consulting organization, he worked for Stanford University and WebTV. Warren holds a BA in Communications Studies from the University of Michigan and an MBA from the University of Pennsylvania's Wharton School.

Joy Mundy has focused on DW/BI systems since 1992 with stints at Stanford, Web TV, and Microsoft's SQL Server product development organization. Joy graduated from Tufts University with a BA in Economics, and from Stanford University with an MS in Engineering Economic Systems.

Bob Becker has helped clients across a variety of industries with their DW/BI challenges and solutions since 1989, including extensive work with health care organizations. Bob has a BSB in Marketing from the University of Minnesota's School of Business.

Credits

Executive Editor
Robert Elliott

Development Editor
Sara Shlaer

Production Editor
Debra Banninger

Copy Editor
Kim Cofer

Editorial Manager
Mary Beth Wakefield

Production Manager
Tim Tate

Vice President and Executive Group Publisher
Richard Swadley

Vice President and Executive Publisher
Joseph B. Wikert

Project Coordinator, Cover
Lynsey Osborn

Proofreader
Nancy Carrasco

Indexer
Melanie Belkin

Anniversary Logo Design
Richard Pacifico

Cover Image

Acknowledgments

First, thanks to our students, clients, readers, and former colleagues for supporting, teaching, and influencing us. One of the authors recently received a fortune cookie that read, “You learn most when teaching others.” We couldn't agree more. Our Kimball University students have pushed us to provide precise, specific guidance and kept us on our toes with their questions. Similarly, the challenges faced by our Kimball Group consulting clients have become our challenges, and have kept us grounded in reality. Finally, ex-colleagues have contributed to our thinking about the concepts in this book, including Laura Reeves who participated as a co-author of the first edition of the *Lifecycle Toolkit*. Beginning with our associates from the early days at Metaphor, through Red Brick, Stanford University, DecisionWorks Consulting, InfoDynamics, and Microsoft, we've learned lots from each of you.

Thanks to the Wiley team for making this book a reality. Bob Elliott's subtle, yet persistent prodding got the project off the ground. Sara Shlaer did a wonderful job editing our text with an incredible amount of patience, tenacity, and attention to detail. Deb Banninger and the behind-the-scenes folks worked tirelessly to deliver a quality product. We've enjoyed working with all of you.

Finally, thanks to our spouses, partners, and children for putting up with the demands of our careers, while supporting us unconditionally. You've suffered through late nights and missed vacations alongside us. Thanks to Julie Kimball, Sara Kimball Smith, and Brian Kimball, Scott and Katie Ross, Elizabeth Wright, Tony Navarrete, and Pam, Elisa, and Jenna Becker. We couldn't have done it without you!

Introduction

Remarkable transformations have occurred in the nine years since the first edition of *The Data Warehouse Lifecycle Toolkit* was published. The data warehouse industry has reached full maturity and acceptance across the business world. Hardware and software have made mind boggling advances in these nine years. We have replaced “gigabytes” with “terabytes” in all our conversations. Yet somehow the data warehousing task has remained fundamentally unchanged.

Many of you have thousands of data warehouse users in your organizations. You have welcomed operational decision makers to the ranks of data warehouse users to accompany the original cadres of marketing and finance users. In fact, operational urgencies are the hottest aspects of data warehousing, with everybody insisting that they need the data in “real time.” As our data warehouses have become more important and more visible, we have been hammered by privacy, security, and compliance requirements that are non-negotiable. Business users are waking up to the value of high quality data in much the same way that conventional manufacturing has embraced the management of quality. Finally, and perhaps most important, we have a new name for what we do that reflects our true purpose. It is *business intelligence*. To emphasize that point, in most places in this book we refer to the overall system you are building as the *DW/BI system*.

The shift to business intelligence puts initiative in the hands of business users, not IT. But at the same time this shift puts into perfect focus the mission of the data warehouse: It is the necessary platform for business intelligence. The data warehouse does the hard work of wrangling the data out of the source systems, cleaning it,

and organizing it so that normal business users can understand it. Of course we strive for world class business intelligence, but world class business intelligence is only possible if you have a world class data warehouse. And conversely, a data warehouse without business intelligence will fail spectacularly.

This book is a relentlessly practical field guide for designers, managers, and owners of the DW/BI system. We have tried to distinguish this book from other DW/BI books by making the content very concrete and actionable. It's okay to be dazzled by the landscape but we want you to make it all the way to the finish line. This book describes a coherent framework that goes all the way from the original scoping of an overall enterprise DW/BI system, through the detailed steps of developing and deploying, to the final steps of planning the next phases.

There are tens of thousands of functioning data warehouse installations across the world. Many DW/BI owners have developed a complete lifecycle perspective. Probably the biggest insight that comes from this perspective is that each DW/BI system is continuously evolving and dynamic. It cannot be static. It never stops transforming. New business requirements arise. New managers and executives place unexpected demands on the system. New data sources become available. At the very least, the DW/BI system needs to evolve as fast as the surrounding organization evolves. Stable organizations will place modest demands on the system to evolve. Dynamic, turbulent organizations will make the task more challenging.

Given this churning, evolving nature of the DW/BI system, we need design techniques that are flexible and adaptable. We need to be half DBA and half MBA. We need to opportunistically hook together little pieces from individual business processes into larger pieces, making enterprise data warehouses. And we need our changes to the system

always to be graceful. A graceful change is one that doesn't invalidate previous data or previous applications.

How this Book is Organized

This book has two deep underlying themes. The first is the Kimball Lifecycle approach. You might ask “What makes the Kimball Lifecycle different from any other methodology?” The shortest answer is that we build DW/BI systems by starting with the business users and figuring out what they need to do their jobs. Then, with those results in mind, we systematically work backward through the reports, applications, databases, and software, finally arriving at the most physical layers of the implementation. This contrasts strongly with technology driven approaches, which proceed in the opposite direction. In the early days of the 1990s, some IT shops didn't know what to make of our business and user oriented approach. But as we publish this book in 2008, the very name “business intelligence” says it all. The user and the business drive the data warehouse.

The second theme is the “bus architecture.” We will show you how to build a succession of individual business process iterations that will, in time, create an enterprise DW/BI system. In this book, you will see a heavy reliance on dimensional modeling as a way to present data to business users. We recommend this approach for only one reason: It is demonstrably the best organization of data to meet the business user's desires for simplicity and high query performance. We thank you in advance for following the dimensional approach that is developed in this book. In the end, you are free to present data to users in any way you think appropriate. But we urge you to constantly revisit the fundamental goal of user satisfaction. We have learned to

be humble in the presence of business users. It's not our opinion that matters; it's theirs.

This book captures these perspectives. We will give you actionable skills and actionable tools for getting your job done. Along the way, we hope to give you the perspective and judgment we have accumulated in building DW/BI systems since 1982.

Who Should Read this Book

The primary reader of this book should be a designer or a manager who really needs to get about the business of building and managing a “data warehouse that is a platform for business intelligence applications.” Because that is quite a mouthful, we have consistently referred to this overall system with the name “DW/BI” to drive home the point that you are responsible for getting the data all the way from the original source systems to the business users' screens.

Although the book contains some introductory material, we think the book will be of most use to an IT professional who has already had some exposure to data warehousing. An appropriate next book, which would concentrate more deeply on dimensional modeling, would be *The Data Warehouse Toolkit, Second Edition*, by Ralph Kimball and Margy Ross, published in 2002.

You may have developed your experience and formed your opinions by designing and delivering a real data warehouse. That is the best background of all! There is no substitute for having had the responsibility of delivering an effective DW/BI system. We the authors have all had the humbling experience of presenting our “baby” to a crowd of demanding business users. It is sometimes hard to accept

the reality that most users have real jobs that don't involve technology. They may not even like technology particularly. But business users will use our technology if it is easy to use and provides obvious value.

This book is rather technical. The discussion of design techniques and architectures will undoubtedly introduce terminology that you have not encountered. We have combed this book carefully to make sure that the more technical topics are ones we think you must understand. We have tried not to get bogged down in detail for its own sake. There is a glossary of DW/BI terms at the back of the book that will briefly explain the most insidious terms that we all have to live with.

Although we hope you read this book in its entirety to understand the complete Kimball Lifecycle, we highlight the target audience at the start of each chapter, so you can best judge what to read carefully, and what to skim. Hopefully, your experiences and opinions will give you your own personal framework on which to hang all these ideas. After reading Chapter 1, you will see that there are three parallel threads that must be pursued in building a DW/BI system: the technology, the data, and the business intelligence applications. We even show these three threads in the "You Are Here" diagrams at the beginning of each chapter. Although these threads clearly affect each other, they should be developed in parallel and asynchronously.

However, because a book is necessarily a linear thing, we have had to present the steps in the Kimball Lifecycle as if they occur in just one fixed order. Hopefully as you work through the book, you will visualize the more realistic and complex real world relationships among the various steps. After reading this book, please return eventually to each individual chapter and re-read it very carefully when your project gets to that particular phase. That is why we called it the *Lifecycle Toolkit*.

How this Book Differs from the First Edition

This second edition of the *Lifecycle Toolkit* is significantly updated and reorganized compared to the first edition. The first three chapters set you up for understanding the complete Kimball Lifecycle process and for making sure your effort has satisfied the requirements for moving forward. We then worked very hard to make the complex discussion of architectures more actionable and more obviously tied to the sequence of the Kimball Lifecycle. In Chapter 4 we carefully describe the complete technical architecture of the DW/BI system, from original data extraction to the final painting of results on the business users' screens. In Chapter 5 we show you how to create specific plans for this technical architecture and select products. Then in Chapters 6 through 12 we systematically expand the three main deliverables (database designs, ETL system, and BI applications) by first describing each one conceptually and then physically. Finally, in the last two chapters we show you how to deploy this amazing edifice into real operational environments and how to think about expanding and growing your DW/BI system beyond the first implementation.

We hope our enthusiasm for data warehousing and business intelligence shows through in this book. The DW/BI challenge is a fascinating and worthy one. Undoubtedly, the labels will change over the years as vendors position their products to be new things that will remove all the old objections. But our mission has remained constant: Bring the data and analyses to the business users so they can make better business decisions.

Chapter 1

Introducing the Kimball Lifecycle

Before delving into the specifics of data warehouse/business intelligence (DW/BI) design, development, and deployment, we want to first introduce the Kimball Lifecycle methodology. The Kimball Lifecycle provides the overall framework that ties together the various activities of a DW/BI implementation. The Lifecycle also ties together the content of this book, setting the stage and providing context for the detailed information that unfolds in the subsequent chapters.

This chapter begins with a historical perspective on the origination and evolution of the Kimball Lifecycle. We introduce the Lifecycle roadmap, describing the major tasks and general guidelines for effectively using the Lifecycle throughout your project. Finally, we review the core vocabulary used in the book.

We recommend that all readers take the time to peruse this brief introductory chapter, even if you are involved in only one facet of the DW/BI project. We believe it is beneficial for the entire team to understand and visualize the big picture and overall game plan. This chapter focuses on the forest; each remaining chapter will turn its attention to the individual trees.

Lifecycle History Lesson

The Kimball Lifecycle methodology first took root at Metaphor Computer Systems in the 1980s. Metaphor was a pioneering decision support vendor; its hardware/software product offering was based on LAN technology with a relational database server and graphical user interface client built on a 32-bit operating system. Nearly a quarter century ago, analysts in large corporations were using Metaphor to build queries and download results into spreadsheets and graphs. Sounds familiar, doesn't it?

Most of this book's authors worked together to implement decision support solutions during the early days at Metaphor. At the time, there were no industry best practices or formal methodologies. But the sequential steps of decision support were as obvious then as they are now; our 1984 training manual described them as *extract, query, analysis*, and *presentation*.

The authors and other Metaphor colleagues began honing techniques and approaches to deal with the idiosyncrasies of decision support. We had been groomed in traditional development methodologies, but we modified and enhanced those practices to address the unique challenges of providing data access and analytics to business users, while considering growth and extensibility for the long haul.

Over the years, the authors have been involved with literally hundreds of DW/BI projects in a variety of capacities, including vendor, consultant, IT project team member, and business user. Many of these projects have been wildly successful, some have merely met expectations, and a few have failed in spectacular ways. Each project taught us a lesson. In addition, we have all had the opportunity to learn from many talented individuals and organizations over the years. Our approaches and techniques have been refined over time—and distilled into *The Data Warehouse Lifecycle Toolkit*.

When we first published this book in 1998, we struggled with the appropriate name for our methodology. Someone suggested calling it the Kimball Lifecycle, but Ralph modestly resisted because he felt that many others, in addition to him, contributed to the overall approach.

We eventually determined that the official name would be the Business Dimensional Lifecycle because this moniker reinforced the unique core tenets of our methods. We felt very strongly that successful data warehousing depends on three fundamental concepts:

- Focus on the business.
- Dimensionally structure the data that's delivered to the business via ad hoc queries or reports.
- Iteratively develop the overall data warehouse environment in manageable lifecycle increments rather than attempting a galactic Big Bang.

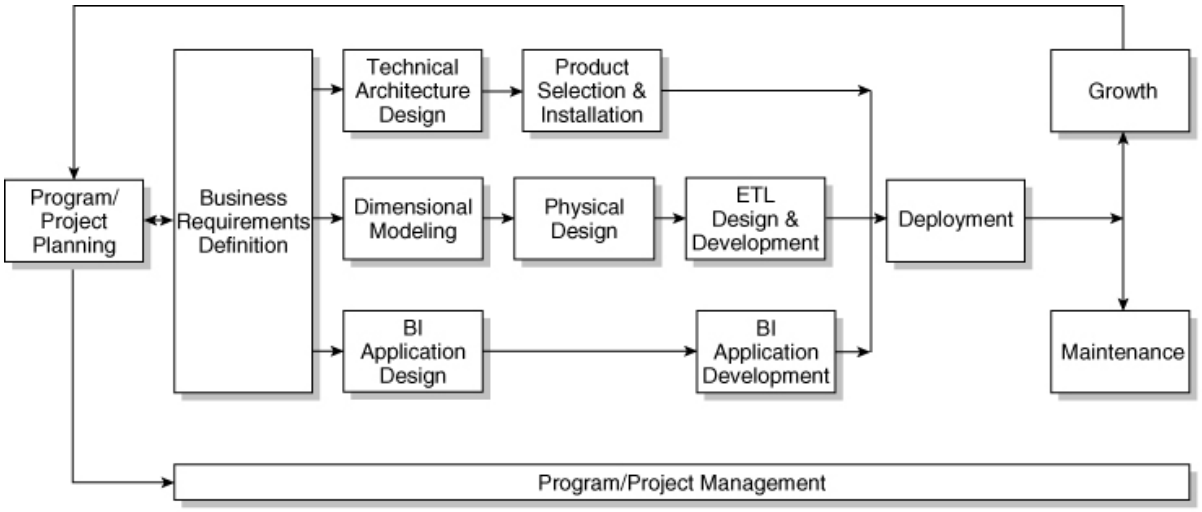
Rewinding back to the 1990s, we were one of the few organizations emphasizing these core principles at the time, so the Business Dimensional Lifecycle name also differentiated our methods from others in the marketplace. Fast forwarding to today, we still firmly believe in these core concepts; however the industry has evolved since the first edition of the *Lifecycle Toolkit* was published. Now nearly everyone else touts these same principles; they've become mainstream best practices. Vocabulary from our approach including dimension tables, fact tables, and slowly changing dimensions have been embedded in the interfaces of many DW/BI tools. While it's both thrilling and affirming that the concepts have been woven into the fiber of our industry, they're no longer differentiators of our approach. Second, despite our thoughtful naming of the Business Dimensional Lifecycle, the result was a mouthful, so most people in the industry simply refer to our methods as the Kimball approach, anyhow. Therefore, we're officially adopting the Kimball Lifecycle nomenclature going forward.

In spite of dramatic advancements in technology and understanding during the last couple of decades, the basic constructs of the Kimball Lifecycle have remained strikingly constant. Our approach to designing, developing, and deploying DW/BI solutions is tried and true. It has been tested with projects across virtually every industry, business function, and platform. The Kimball Lifecycle approach has proven to work again and again. In fact, that's the reasoning behind the Kimball Group's “practical techniques, proven results” motto.

Lifecycle Milestones

The overall Kimball Lifecycle approach to DW/BI initiatives is illustrated in Figure 1.1. Successful implementation of a DW/BI system depends on the appropriate integration of numerous tasks and components. It is not enough to have the perfect data model or best-of-breed technology. You need to coordinate the many facets of a DW/BI project, much like a conductor must unify the many instruments in an orchestra. A soloist cannot carry a full orchestra. Likewise, the DW/BI implementation effort needs to demonstrate strength across all aspects of the project for success. The Kimball Lifecycle is similar to the conductor's score. It ensures that the project pieces are brought together in the right order and at the right time.

Figure 1.1 The Kimball Lifecycle diagram.



The Lifecycle diagram depicts the sequence of high level tasks required for effective DW/BI design, development, and deployment. The diagram shows the overall roadmap, while each box serves as a guidepost or mile/kilometer marker. We'll briefly describe the milestones, as well as provide references to the corresponding chapters in this book for more specific driving instructions.

Program/Project Planning

The Lifecycle begins with program and project planning, as one would expect. Throughout this book, *project* refers to a single iteration of the Kimball Lifecycle from launch through deployment; projects have a finite start and end. On the other hand, *program* refers to the broader, ongoing coordination of resources, infrastructure, timelines, and communication across multiple projects; a program is an overall umbrella encompassing more than one project. It should continuously renew itself and should rarely have an abrupt end.

Which comes first, the program or the project? Much like the classic chicken and egg conundrum, it's not always obvious which comes first. In some organizations, executive

agreement is reached to launch a DW/BI program and then it's a matter of prioritizing to identify the initial project. In other situations, funding is provided for a single project or two, and then the need for program coordination is subsequently realized. There's no single right approach or sequence.

There's much greater consistency around project planning, beginning with the scoping of the DW/BI project. Obviously, you must have a basic understanding of the business's requirements to make appropriate scope decisions; the bi-directional arrow between the project planning and business requirements boxes in Figure 1.1 shows this dependency. Project planning then turns to resource staffing, coupled with project task identification, assignment, duration, and sequencing. The resulting integrated project plan identifies all tasks associated with the Kimball Lifecycle and the responsible parties. It serves as the cornerstone for the ongoing management of your DW/BI project. Chapter 2 details these launch activities, in addition to the ongoing management of the program/project.

Program/Project Management

Program/project management ensures that the Kimball Lifecycle activities remain on track and in sync. Program/project management activities focus on monitoring project status, issue tracking, and change control to preserve scope boundaries. Ongoing management also includes the development of a comprehensive communication plan that addresses both the business and information technology (IT) constituencies. Continuing communication is critical to managing expectations; managing expectations is critical to achieving your DW/BI goals.