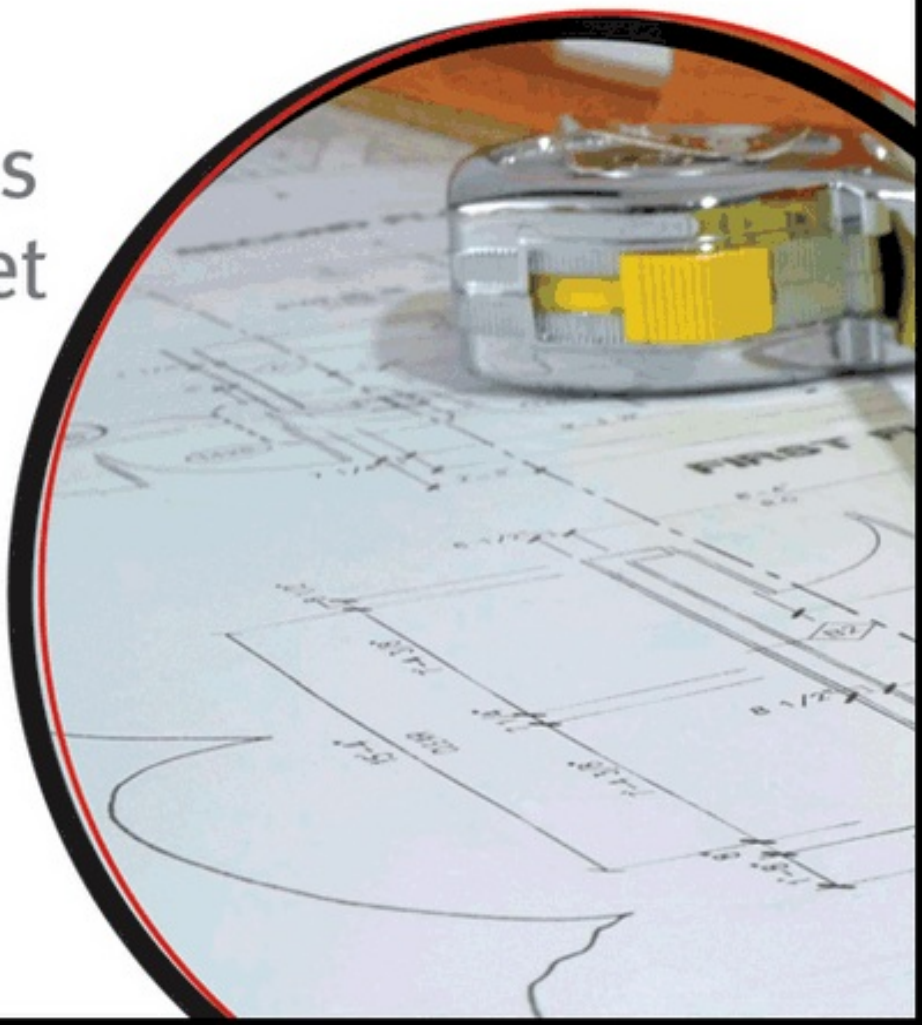
WILEY

KIMBALL GROUP

The Microsoft® Data Warehouse Toolkit

Second Edition

With SQL Server® 2008 R2 and the Microsoft Business Intelligence Toolset



Joy Mundy and Warren Thornthwaite
with Ralph Kimball

Table of Contents

The Microsoft® Data Warehouse Toolkit

With SQL Server 2008 R2 and the Microsoft® Business Intelligence Toolset

Second Edition

Joy Mundy and
Warren Thornthwaite
with Ralph Kimball

WILEY
Wiley Publishing, Inc.

The Microsoft® Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft® Business Intelligence Toolset, Second Edition

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: 978-0-470-64038-8ISBN: 978-1-118-06793-2 (ebk)ISBN: 978-1-118-06795-6 (ebk)ISBN: 978-1-118-06794-9 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number: 2011920894

About the Authors

Joy Mundy has focused on DW/BI systems since 1992 with stints at Stanford, WebTV, 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.

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.

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.

Credits

Executive Editor
Robert Elliott

Project Editors
Sara Shlaer
Ginny Munroe

Technical Editor
Ralph Kimball

Senior Production Editor
Debra Banninger

Copy Editor
Kim Cofer

Editorial Director
Robyn B. Siesky

Editorial Manager
Mary Beth Wakefield

Freelancer Editorial Manager
Rosemarie Graham

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
Katie Crocker

Compositor
Craig Johnson, Happenstance Type-O-Rama

Proofreader
Jen Larsen, Word One

Indexer
Robert Swanson

Cover Image
© Getty Images

Cover Designer
Ryan Sneed

Acknowledgments

First, we want to thank the thousands of you who have read the Kimball Group's *Toolkit* books, attended our courses, and engaged us in consulting projects. We always learn from you, and you've had a profound impact on our thinking and the business intelligence industry.

This book would not have been written without the assistance of many people on the SQL Server product development team. Dave Wickert reviewed the PowerPivot and SharePoint chapters and provided many excellent suggestions for improving the content. Bryan Smith was kind enough to read the chapters on Integration Services and Analysis Services, and those chapters are the better for his assistance. Carolyn Chau reviewed the Reporting Services chapter, Eric Hanson read the relational database chapter, Pej Javaheri commented on the SharePoint chapter, and Raman Iyer read the data mining chapter. Our sincere gratitude to all of them.

Other members of the SQL Server team provided significant assistance reviewing the SQL Server 2005 version of this book, and we were too embarrassed to impose on them a second time. These include Bill Baker, Stuart Ozer, Grant Dickinson, Donald Farmer, Siva Harinath, Jamie MacLennan, John Miller, Ashvini Sharma, Stephen Quinn, and Rob Zare.

Our colleagues at the Kimball Group were invaluable. Their encouragement kept us going while we were writing the book, and their reviews helped us polish and prune material. Ralph Kimball, of course, had a huge impact on the book, not just from his writing and thinking in the business intelligence arena but more directly by helping us improve the book's overall structure and flow.

Sara Shlaer, Ginny Munroe, and Bob Elliott, our editors at Wiley, have been very helpful and encouraging. It’s been a pleasure to work with them.

To our life partners, thanks for being there when we needed you, for giving us the time we needed, and for occasionally reminding us that it was time to take a break. Tony Navarrete and Elizabeth Wright, the book wouldn’t exist without you.

Foreword

In the five years since the first edition was published, Microsoft has made impressive progress in building out its data warehousing and business intelligence tools suite. It is gratifying to those of us who work in this space to see the steady commitment that Microsoft has made to provide usable, professional quality tools. During these five years, Warren and Joy have consulted with dozens of clients, taught scores of classes, answered hundreds if not thousands of questions, had many “schema lunches” where the schema diagrams competed with the food, and have pounded on every module in Microsoft’s DW/BI toolset. This current edition remains a unique reference, combining overall perspectives on what the tools do with accurate assessments of how well they do it. This book teaches judgment, not button clicks!

—Ralph Kimball

Introduction

The goal of this book is to guide the reader down the best path toward designing and building a successful business intelligence system and its underlying data warehouse databases using the Microsoft SQL Server product set.

The Data Warehouse and Business Intelligence System

Data warehousing and business intelligence are techniques to provide business people with the information and tools they need to make both operational and strategic business decisions. We'll break this down a bit so you can really understand the nature and magnitude of what you're about to take on.

First, your customers are the business people in the organization. Not all business people carry the same importance to you — you should be especially concerned with those who make strategic business decisions. One well-made business decision can translate to millions of dollars in many organizations. Your main customers are executives, managers, and analysts throughout the organization. The data warehouse and business intelligence (DW/BI) system is high impact and high profile.

Strategic also means important. These are decisions that can make or break the organization. Therefore, the DW/BI system is a high-risk endeavor. When strategic decisions are made, someone often wins or loses. The DW/BI system is a highly political effort.

Increasingly, the DW/BI system also supports operational decisions, especially where the decision maker needs to see

historical data or integrated data from multiple sources. Many analytic applications have this operational focus. Whether the decision making is strategic or operational, the DW/BI team needs to provide the information necessary to make decisions.

Any given decision will likely require a unique subset of information that generally cannot be predetermined. You'll need to build an information infrastructure that integrates data from across the organization, and potentially from outside the organization, and then cleans, aligns, and restructures the data to make it as flexible and usable as possible. Whereas most transaction system modules work with one type of information, such as billings, orders, or accounts receivable, the DW/BI system must eventually integrate them all. The DW/BI system requires technically sophisticated data gathering and management.

Finally, you need to provide the business decision makers with the tools they need to make use of the data. In this context, "tools" means much more than just software. It means everything the business users need to understand what information is available, find the subsets they need, and structure the data to illuminate the underlying business dynamics. "Tools" includes training, documentation, and support, along with ad hoc query tools, reports, and analytic applications.

Let's review. The DW/BI system:

- Is high profile and high impact
- Is high risk
- Is highly political
- Requires technically sophisticated and complex data gathering and management
- Requires intensive user access, training, and support

Creating and managing the DW/BI system is an extremely challenging task. We want you to take on this task with full knowledge of what you're getting into. In our experience,

it’s easier to deal with all of the challenges if you’re at least somewhat forewarned.

We don’t mean this to discourage you, but rather to warn you before you jump in that the waters are swift and deep. All the reasons that make the data warehouse challenging are also what make it a fun and exciting project.

The Kimball Group

While it’s true that building and managing a successful DW/BI system is a challenge, it’s also true that there are ways to approach it that will increase your likelihood of success. That’s what the Kimball Group is all about. We’ve been working in the DW/BI area for more than 25 years. The authors of this book, who are members of the Kimball Group, have spent their careers working on data warehousing and business intelligence systems as vendors, consultants, implementers, and users. Our motto is “Practical techniques — proven results.” We share a common drive to figure out the best way to build and manage a successful DW/BI system. We are also teachers at heart, with a strong desire to help you succeed and avoid the mistakes we and others have made.

Why We Wrote This Book

Data warehousing and business intelligence have been around in much the same form since at least the 1970s, and continue to enjoy an incredibly long technology lifecycle. In 1995, when the primary authors formed our first consulting organization, one of us voiced the opinion that data warehousing was finished, that the wave had crested and we’d be lucky to get a few more projects before we had to go find real jobs again. Years later, data warehousing and business intelligence are still going strong.

As the DW/BI industry has matured, it's become dominated by single-source providers — a safe choice for risk-averse organizations. The DW/BI technology stack covers everything from esoteric source system knowledge to user interface design and best-practice BI applications. Database vendors are best positioned to provide end-to-end solutions. Since SQL Server 2000 and especially SQL Server 2005, Microsoft has been forcing the concept of a viable, single-source data warehouse system provider into reality, and at an attractive price.

The book you're currently holding is a substantial revision of *The Microsoft Data Warehouse Toolkit with SQL Server 2005*. In addition to updating the content for new features and functionality such as PowerPivot and Master Data Services, the new version updates our previous recommendations with all that we've learned in recent years about building a DW/BI system with the Microsoft tools. The current book is based on the SQL Server 2008 R2 release, but the vast majority of its recommendations are valid for SQL Server 2008 as well. Any technology or recommendation that's new for SQL Server 2008 R2 is clearly identified in the text.

Who Should Read This Book

This book covers the entire DW/BI system lifecycle. As a result, it offers useful guidance to every member of the DW/BI team, from the project manager to the business analyst, data modeler, ETL developer, DBA, BI application developer, and even to the business user. We believe the book will be valuable to anyone working on a Microsoft SQL Server DW/BI program.

The primary audience for this book is the new DW/BI team that's launching a project on the Microsoft SQL Server platform. We don't assume you already have experience in building a DW/BI system. We do assume you have a basic

familiarity with the Microsoft world: operating systems, infrastructure components, and resources. We also assume a basic understanding of relational databases (tables, columns, simple SQL) and some familiarity with the SQL Server relational database, although that's not a requirement. Throughout the book we provide many references to other books and resources.

A second audience is the experienced Kimball Method DW/BI practitioner who's new to the Microsoft SQL Server toolset. We'll point out which sections and chapters will be review for anyone who's read our other *Toolkit* books and practiced our methodology. But we've found that it doesn't hurt to read this material one more time!

Whatever your background, you'll benefit most if you're just starting on a new project. While we do provide suggestions on working with existing data warehouses, in the ideal case you won't have to contend with any existing data warehouse or data marts — at least none that will remain in place after the new system is deployed.

The Kimball Lifecycle

We've all felt the empty pit of panic in our stomach when, deep into a project, we realize the scope and scale of the effort before us will take much more work than we imagined at the outset. Many DW/BI projects begin with the notion that you'll just move some data to a new machine, clean it up a little, and develop some reports. That doesn't sound so bad — six weeks of effort, two months at the most. You charge into the forest and soon realize it's a lot darker and denser than you thought. In fact, you can't even see the road out.

The best way to avoid this sense of panic — and the resulting disaster — is to figure out where you're going before you jump in. It helps to have a roadmap and

directions to lead you safely through unfamiliar territory — one that will tell you the places you have to visit and point out the danger zones on the trip ahead. This book is that roadmap for the Microsoft SQL Server DW/BI system project. This book follows the basic flow of the Kimball Lifecycle described in the book *The Data Warehouse Lifecycle Toolkit, Second Edition* (Wiley, 2008). The steps, tasks, and dependencies of the Lifecycle were crafted based on our collective experience of what works. The Lifecycle is an iterative approach based on four primary principles:

- *Focus on the business:* Concentrate on identifying business requirements and their associated value. Use these efforts to develop solid relationships with the business side and sharpen your business sense and consultative skills.
- *Build an information infrastructure:* Design a single, integrated, easy-to-use, high-performing information foundation that will meet the broad range of business requirements you've identified across the enterprise.
- *Deliver in meaningful increments:* Build the data warehouse in increments that can be delivered in 6 to 12 month timeframes. Use clearly identified business value to determine the implementation order of the increments.
- *Deliver the entire solution:* Provide all the elements necessary to deliver value to the business users. This means a solid, well designed, quality tested, accessible data warehouse database is only the start. You must also deliver ad hoc query tools, reporting applications and advanced analytics, training, support, website, and documentation.

This book helps you follow these four principles by using the Kimball Lifecycle to build your DW/BI system. These four principles are woven into the fabric of the Lifecycle. The secret to understanding the Kimball Lifecycle is that it's

business-based, it takes a dimensional approach to designing data models for end user presentation, and it is a true lifecycle.

Lifecycle Tracks and Task Areas

The DW/BI system is a complex entity, and the methodology to build that system must help simplify that complexity. Figure 1 outlines the Kimball Lifecycle. The 13 boxes show the major task areas involved in building a successful data warehouse and the primary dependencies among those tasks.

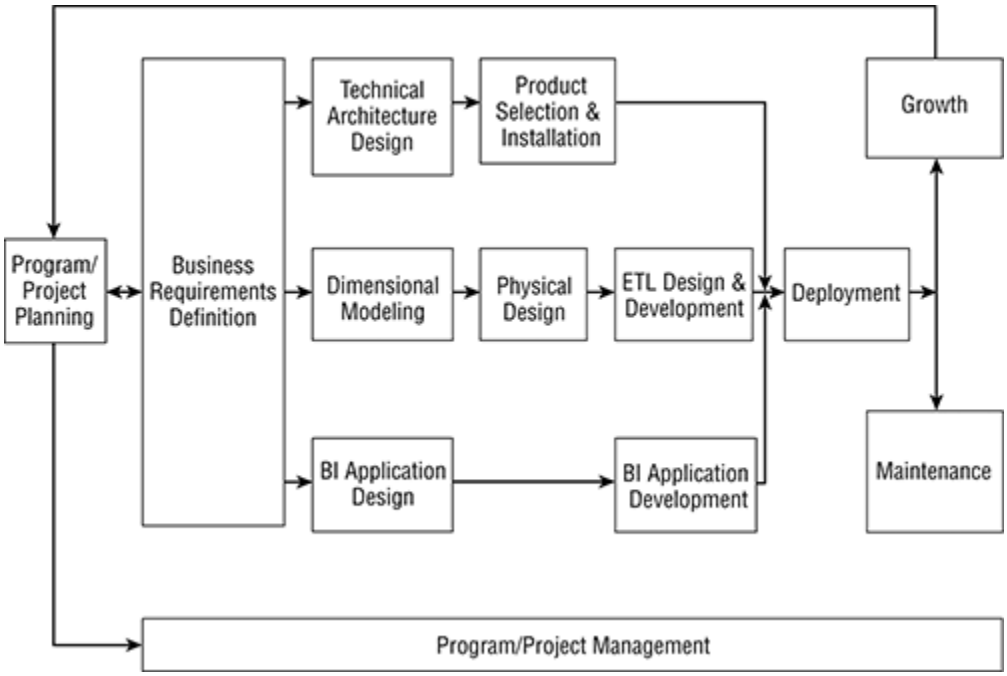


Figure 1: The Business Dimensional Lifecycle

There are several observations to make about the Lifecycle at this level. First, notice the central role of the Business Requirements Definition box. Business requirements provide the foundation for the three tracks that follow. They also influence the project plan, hence the arrow pointing back to the Project Planning box. You usually end up modifying the plan based on a more detailed understanding of the business requirements and priorities.

Second, the three tracks in the middle of the Lifecycle concentrate on three separate areas:

- *The top track is about technology.* These tasks are primarily about determining what functionality you will need, and planning which pieces of Microsoft technology you'll use, and how you'll install and configure them.
- *The middle track is about data.* In the data track you'll design and instantiate the dimensional model, and develop the Extract, Transformation, and Load (ETL) system to populate it. You could think of the data track as "building the data warehouse databases," although your data warehouse will not succeed unless you surround it with the rest of the Lifecycle tasks.
- *The bottom track is about business intelligence applications.* In these tasks you design and develop BI applications for the business users.

The tracks combine when it's time to deploy the system. This is a particularly delicate time because there's only one chance to make a good first impression. Although we've placed maintenance after deployment in the diagram, you need to design your system with the ability and tools for maintaining it. The growth phase of the project links to the arrow heading back to the beginning. This simple arrow has major implications. The Lifecycle's incremental approach is a fundamental element of delivering business value.

Underlying the entire Lifecycle is the Project Management box. The most important thing to remember here is that you need a leader, and that person needs access to senior management. The team leader is ideally one of those difficult-to-find people who can communicate effectively with both technologists and business people, including the most senior executives in the company.

Key Terminology and the Microsoft Toolset

The business intelligence industry is plagued with terminology that's used imprecisely, or in contradictory ways. Some of the most long-standing debates in the industry derive as much from misunderstandings about what others mean by a term, as from true differences in philosophy. Keeping that in mind, we'll try to be clear and consistent even if we don't settle all the historical debates. We highlight some of the key terms here.

As we define each term, we also relate it to the associated Microsoft technologies, most of which are components of SQL Server.

- The *data warehouse* is the "platform for business intelligence." In the Kimball Method, the data warehouse includes everything from the original data extracts to the software and applications that users see. We disagree with other authors who insist that the data warehouse is merely a centralized and highly normalized store of data in the back room, far from the end users. To reduce confusion, in this book we consistently use the phrase "data warehouse/business intelligence system" (DW/BI system) to mean the entire end-to-end system. When we're talking specifically and exclusively about the atomic level user queryable data store, we call it the *data warehouse database*.
- The *business process dimensional model* is a specific discipline for modeling data that is an alternative to normalized modeling. A dimensional model contains the same information as a normalized model but packages the data in a symmetrical format whose design goals are user understandability, business intelligence query performance, and resilience to change. Normalized models, sometimes called *third normal form* models,

were designed to support the high-volume, single-row inserts and updates that define transaction systems, and generally fail at being understandable, fast, and resilient to change. We use the term “business process dimensional model” to refer both to the logical dimensional model that supports a business process and the corresponding physical tables in the database. In other words, dimensional models are both logical and physical.

- The *relational database* is a general purpose technology for storing, managing, and querying data. The SQL Server database engine is Microsoft’s relational database engine. The business process dimensional model can be stored in a relational database. Normalized data models that support transaction processing can also be stored in a relational database.
- The *online analytic processing (OLAP) database* is a technology for storing, managing, and querying data specifically designed to support business intelligence uses. SQL Server Analysis Services is Microsoft’s OLAP database engine. The business process dimensional model can be stored in an OLAP database, but a transactional database cannot, unless it first undergoes transformation to cast it in an explicitly dimensional form.
- An *Extract, Transformation, and Load (ETL)* system is a set of processes that clean, transform, combine, de-duplicate, household, archive, conform, and structure data for use in the data warehouse. These terms are described in this book. Early ETL systems were built using a combination of SQL and other scripts. While this is still true for some smaller ETL systems, larger and more serious systems use a specialized ETL tool. Moving forward, almost every DW/BI system will use an ETL tool such as SQL Server Integration Services because the

benefits are significant and the incremental dollar cost is low or zero.

- *Business intelligence (BI) applications* are predefined applications that query, analyze, and present information to support a business need. There is a spectrum of BI applications, ranging in complexity from a set of predefined static reports, all the way to an analytic application that directly affects transaction systems and the day-to-day operation of the organization. You can use SQL Server Reporting Services to build a reporting application, and a wide range of Microsoft and third-party technologies to build complex analytic applications.
- A *data mining model* is a statistical model, often used to predict future behavior based on data about past behavior or identify closely related subsets of a population called clusters. Data mining is a term for a loose (and ever-changing) collection of statistical techniques or algorithms that serve different purposes. The major categories are clustering, decision trees, neural networks, and prediction. Analysis Services Data Mining is an example of a data mining tool.
- *Ad hoc queries* are formulated by the user on the spur of the moment. The dimensional modeling approach is widely recognized as the best technique to support ad hoc queries because the simple database structure is easy to understand. Microsoft Office, through Excel pivot tables and PowerPivot, is the most popular ad hoc query tool on the market. You can use Reporting Services Report Builder to perform ad hoc querying and report definition. Nonetheless, many systems supplement Excel and Report Builder with a third-party ad hoc query tool for their power users.
- Once again, the *data warehouse/business intelligence (DW/BI) system* is the whole thing: source system

extracts, ETL, dimensional database in both relational and OLAP, BI applications, and an ad hoc query tool. The DW/BI system also includes management tools and practices, user-oriented documentation and training, a security system, and all the other components that we discuss in this book.

Roles and Responsibilities

The DW/BI system requires a number of different roles and skills, from both the business and technical communities, during its lifecycle. In this section, we review the major roles involved in creating a DW/BI system. There is seldom a one-to-one relationship between roles and people. We've worked with teams as small as one person, and as large as forty (and know of much larger teams). The vast majority of DW/BI teams fall between three and ten full-time members, with access to others as required.

It's common for a single DW/BI team to take on both development and operational duties. This is different from most technology project teams, and is related to the highly iterative nature of the DW/BI project development cycle. The following roles are associated with design and development activities:

- The *DW/BI manager* is responsible for overall leadership and direction of the project. The DW/BI manager must be able to communicate effectively with both senior business and IT management. The manager must also be able to work with the team to formulate the overall architecture of the DW/BI system.
- The *project manager* is responsible for day-to-day management of project tasks and activities during system development.
- The *business project lead* is a member of the business community and works closely with the project manager.