The Data Warehouse Toolkit:
The Complete Guide to Dimensional Modeling (2nd edition)

by Ralph Kimball, Margy Ross
John Wiley & Sons, Inc.; 2002; 464 pages
ISBN 0-471-20024-7

Review by:
Alexander A.Anisimov, Ural State Technical University
Department of Physics and Technology
anissimov@pochtamt.ru

Overview

Most of the books on data warehousing and business analysis automation are typically concerned with technical issues. Certain books also attempt to describe the overall organization of a data warehousing project – a challenging process that usually requires extraordinary skills and significant efforts.

But there is the vital element of data warehousing methodology that makes DWs valuable for analysts and managers and connects the essence of business-phenomena with corresponding IT-solutions. This backbone of all data warehouses is a multidimensional data model. The problem of dimensional modeling usually remains beyond the scope of references, textbooks and formal methodologies dedicated to data warehousing (probably, due to its complexity). Usually it is assumed that developers and future users should determine dimensional structure during the development process in accordance with business requirements elicited. This assumption is correct, but only up to a point: dimensional structures of different data warehouses in different organizations may be (and should be) different. But at the same time developers should be knowledgeable about certain methodological principles of dimensional modeling that are universal and do not vary from project to project and from company to company. The development of a such methodology and the summarization of experience in this area are challenging tasks, and only few authors have attempted to cover this complicated subject.

"The Complete Guide to Dimensional Modeling" is probably one of the best books, dedicated to this problem, which is written by skilled and clever practitioner.

Content of the Book

The book consists of an introduction and 16 chapters that may be divided into five major blocks. The introductory chapters are dedicated to the basics of dimensional modeling. Chapters 2-8 describe multidimensional data structures, related to certain well-known business-phenomena (sales, inventories, etc.). Chapters 9-15 discuss dimensional modeling from the point of view of certain industries (education, health care, etc.). Chapter 16 (the biggest one in this book – 40 pages) describes the most important issues of data warehousing project management. The book also has a concluding chapter that discusses certain problems the of which the connection to the theory and practice of dimensional modeling is quite indirect.

At first sight one may think that this book describes typical data structures for certain business problems (such as, for example, sales or marketing). But actually each chapter is mostly dedicated to specific methodological issues of dimensional modeling. In other words each chapter should have two titles: one for the business domain described and another for the major methodological issue it is focused on. These methodological issues are probably more important than the business examples, that are actually used as illustrations supporting a better understanding of the theory.

In the second chapter the authors begin by discussing the major steps of dimensional modeling and simplest methodological concepts (in-depth examination of time dimension, generate dimensions, surrogate keys and dimension normalization). They use one of the most understandable business examples – retail sales.

The third chapter describes such methodological concepts as DW bus architecture and matrix, conformed dimensions and facts, semi-additive facts, transaction and accumulating snapshot models. In this chapter inventories management is used as a real-world example.

Chapter 4 (18 pages) is mainly concentrated on handling slowly changing dimensions. Namely it describes three basic approaches to this task ("overwrite the value", "add a dimension row" and "add a dimension column") and two advanced techniques ("predictable changes with multiple version overlays" and "unpredictable changes with single-version overlay"). The authors use procurement for demonstration of these theoretical concepts.

Chapter 5 (37 pages) discusses the introduction of DWs in order management and considers dimensional modeling issues specific to this particular area (like "ship-to/bill-to customer dimension"). Some other typical dimensions and dimensional modeling problems are also discussed (product dimensions, multiple currencies and units of measure, comparison of transaction, periodic snapshot, and accumulating snapshot fact tables).

Chapter 6 covers such issues as name and address parsing, combining of data from multiple sources, modeling of unpredictable hierarchies and large volume dimensions management. Customer Relationship Management is used as a business example in this chapter.

In chapter 7 (14 pages) the authors use general ledger information as an example and discuss modeling of year-to-date facts, multiple fiscal calendars and the notion of consolidated dimensional modeling. Chapter 8 (12 pages) concentrates on human resources dimensional modeling and covers some issues that are specific to this sphere, like handling of survey questionnaire data.

The major techniques discussed in Chapter 9 are arbitrary value banding of facts, point-in-time balances and heterogeneous product schemas. They are demonstrated using financial services as an example. Chapter 10 (12 pages) contains practically no new methodological ideas and describes some organizational issues using telecommunications as a business example. Chapter 11 (14 pages) discusses such concepts as combination of the so called "small dimensions" into a superdimension, synchronization across multiple time zones and country-specific calendars. It describes the use of data warehouses for analysis of transportations. Chapter 12 (12 pages) introduces the concept of factless fact tables and demonstrates it using education as an example.

Chapter 13 (22 pages) describes the use of data warehouses for health care analysis. It describes such issues as multivalued diagnosis dimensions, fact dimensions for sparse facts and bridge table that is used to model multiple diagnoses. Chapter 14 describes the analysis of electronic commerce data (unique characteristics of clickstream data, "clickstream specific" dimensions and data models).

In Chapter 15 the authors bring together the major theoretical concepts introduced in the previous chapters and describe the design of multidimensional data structures for an insurance company.

Concluding Remarks

In my opinion this book can hardly be regarded as a full reference in data warehousing. The reader should initially understand that it is a set of important concept explanations and case studies. So in practical work it will probably be necessary to rethink all these concepts for particular situations.

As for the target audience, this book will be undoubtedly useful for DW project managers and DW data modelers (both as a tutorial and as a reference book). It also may be interesting for field managers and analysts (potential users of data warehouses) who are interested in development of DWs that completely meet their needs: this book will help them to express what is really required for business analysis in nearly technical (and hence quite exact and understandable) terms.