# Fundamentals of Data Warehouses

2nd Revised and Extended Edition

*by Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, Panos Vassiliadis*

Springer-Verlag, 2003
214 pages, list price EUR 39.95
ISBN: 3-540-42089-4

Review by:
Vernon Hoffner, Lawrence Technological University
College of Management
Hoffner@ltu.edu

During the last decade the field of data warehousing has grown significantly. Many organizations are either actively looking at this technology or have currently implemented one or more data warehouses or data marts to support corporate decision making. In today's economic environment, the competitive edge frequently comes from the proactive use of the information that companies have been collecting in their operational systems. They are realizing the significant potential this information can have for their organization. The data warehouse provides users with access to these large amounts of integrated, nonvolatile, time variant data that can be used to track business trends, facilitate forecasting and improve strategic decisions.

**Summary of the Book**

The authors state "this book is an introduction and sourcebook for practitioners, graduate students, and researchers interested in the state of the art and the state of the practice in data warehousing." There have been a wide variety of books written for practitioners on the topic of data warehousing in the last few years. However, this is the first book I have seen that focuses on the data warehousing from the point of interest of researchers. Jarke et al. present a good starting point and foundation for someone interested in data warehousing and the related research issues.

Consisting of 178 pages of text, this book is organized into four sections, each consisting of two chapters. The introductory section introduces data warehousing. The first chapter provides a definition and overview of the data warehouse and its components. It also introduces the subject areas that will be covered in more detail in the later chapters. The second chapter presents areas and issues of relevant research. This chapter introduces a focus on modeling and measuring data quality and data warehouse quality, one of the strengths of the text. These topics are covered in more detail in the last section of the book.

The next section looks at the process of obtaining the data and loading it into the data warehouse. In chapter three the authors define source integration and how it can be applied to the process of integrating the schemas of the source data to construct an integrated enterprise schema. We are reminded that this is a continuing process in order to maintain a quality collection of data from the changing source systems. Source integration is also the foundation of the transformation and loading of the data into the data warehouse, the topics of the next chapter. Data warehouse refreshment is the process of integrating, cleansing, and transforming the source data in preparation of physically loading it into the data warehouse. The authors discuss many techniques of data cleaning and provide references to the literature for more details on data cleaning and transformation. The final portion of the chapter covers the process of loading the data into the data warehouse, including the quality factors and design choices necessary to insure a timely and accurate refreshment process.

The third section focuses on the data structures for the data warehouse and the efficient access to that data. Chapter five begins with a short review of the online transaction processing data structure needs to provide a contrast with the data structure needs for online analytical processing. This leads to a description of the multidimensional view of the data needed by the end users and the several ways the multidimensional view can be physically implemented with the database. The authors complete their presentation of the modeling of the data warehouse with a discussion of the role of aggregates in the data warehouse. The role of aggregates is continued in the next chapter, which concentrates on the optimization of query processing. From the perspective of the end users, the response time to their queries is very important, and today everyone is expecting a "quick" reply. The authors describe several methods of improving query performance and reference the underlying research for readers who want or need to pursue additional information in this area.

The last section brings together the topics of the previous chapters and develops an integrated focus on quality in the design and operation of a data warehouse. Chapter seven develops a model for integrating the components of the data warehouse into a unified architecture for an overall perspective on quality of the data warehouse. A major factor in supporting quality is the role played by metadata. The authors then describe how their architecture with supporting metadata can be used to implement a high quality data warehouse. The last chapter describes the application of this quality focus to data warehouse development and operation in the Foundation of Data Warehouse Quality project.

**Target Audience**

The extensive reference to supporting academic and research literature satisfies the authors' goal of making this an excellent sourcebook for graduate students and researchers. Practitioners with good modeling and conceptualization capability will also appreciate this approach. The text covers the essential areas for the development of a quality data warehouse, but it does not provide the cookbook solutions that are provided in the trade press. This style probably will not be greatly appreciated by the majority of practitioners who, in many instances will be looking for the answer to their immediate problem.

**Reviewer's Appreciation**

The writing style and the presentation of the material are a refreshing change for the plethora of popular press books that focus on the experiences of the consultant/writer. The book is easy to read and it provides a good graphical presentation for the conceptual modeling approach to the development of quality data warehouses.

The extensive bibliography is also greatly appreciated. It will provide a good starting point for anyone interested in pursuing a more in-depth study of the supporting literature.

In conclusion, I would recommend this book as one of the required texts for an advanced graduate course in data warehousing.