McGraw-Hill Series on Data Warehousing and Data Management

Alex Berson
Stephen J. Smith

Data Warehousing, Data Mining, & OLAP

数据仓库、数据发掘和联机分析处理

McGraw-Hill Book Co.
世界图书出版公司

世界图书出版公司获得 Addison Wesley Longman、ITP、Springer 和 McGraw-Hill 出版社授权，新近独家重印出版以下计算机科学类图书：

1. Approximation Algorithms for NP-Hard Problems
2. Classical and Object-Oriented Software Engineering 4th ed.
3. Computer Supported Collaborative Writing
4. CSCW and Artificial Intelligence
5. CSCW Requirements and Evaluation
6. Data Stores, Data Warehousing and the Zachman Framework
7. Data Warehousing, Data Mining, & OLAP
8. Design Issues in CSCW
9. Discrete Mathematics for Computer Scientists 2nd ed.
10. IP Multicasting
11. Multimedia Database Systems
12. Operating Systems
13. Software Engineering with C++ and CASE Tools
14. Software Reuse
15. Sybase and Client/Server Computing 2nd ed.
16. TCP/IP and Related Protocols 3rd ed.
17. Working Classes: Data Structures and Algorithms Using C++

ISBN 7-5062-4119-6

9 787506 241199 >

WB4119 定价:85.00 元

Data Warehousing, Data Mining, & OLAP

McGraw-Hill Book

Data Warehousing, Data Mining, and OLAP

Alex Berson
Stephen J. Smith

世界图书出版公司
McGraw-Hill Book Co

Library of Congress Cataloging-in-Publication Data

Berson, Alex.
Data warehousing, data mining, and OLAP / Alex Berson, Stephen J. Smith.
p. cm.
Includes bibliographical references and index.
ISBN 0-07-006272-2
1. Data warehousing. 2. Data mining. 3. Online data processing.
I. Smith, Stephen J. II. Title.
QA76.9.D37B47 1997
005.74—dc21 97-27472
CIP

McGraw-Hill
A Division of The McGraw-Hill Companies

2 3 4 5 6 7 8 9 0 DOC/DOC 9 9 0 2 1 0 9 8

ISBN 0-07-006272-2

The sponsoring editor for this book was John Wyzalek, the editing supervisor was Bernard Onken, and the production supervisor was Pamela A. Pelton. It was set in Century Schoolbook by North Market Street Graphics.

IE ISBN: 0-07-116386-7

Data Warehousing, Data Mining, and OLAP

Other Titles in McGraw-Hill's Data Warehousing and Data Management Series

000748-9	Aiken	*Data Reverse Engineering*
005996-9	Allen, Bambara, and Bambara	*Informix: Client/Server Application Development*
005664-1	Berson	*Client/Server Architecture, 2d ed.*
005203-4	Berson and Anderson	*SYBASE and Client/Server Computing, 2d ed.*
001697-6	Anderson	*Client/Server Database Design with SYBASE*
001737-9	Anderson	*Using DataBlades*
001974-6	Andriole	*Managing Systems Requirements: Methods, Tools, and Cases*
005779-6	Bigus	*Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*
018244-2	Dunham	*Database Performance Tuning Handbook*
021626-6	Fortier	*Database Systems Handbook*
024565-7	Gopaul	*DB2 Common Server Application Development*
912982-X	Jones	*Developing Client/Server Applications with Microsoft Access*
036929-1	Leach	*Software Reuse: Methods, Models, and Costs*
049999-3	Mattison	*Database Management Systems Handbook*
041034-8	Mattison	*Data Warehousing: Strategies, Technologies, and Techniques*
057725-0	Sanders	*The Developer's Handbook to DB2/2 for Common Servers*

To Irina, Vlad, and Michelle
Alex Berson

To Samuel, who needs a chance
Steve Smith

Foreword

Ever since the dawn of business data processing, managers have been seeking ways to increase the utility of their information systems. In the past, much of the emphasis has been on automating the transactions that move an organization through the interlocking cycles of sales, production, and administration. Whether accepting an order, purchasing raw materials, or paying employees, most organizations process an enormous number of transactions and in so doing gather an even larger amount of data about their business.

Despite all the data they have accumulated, what users really want is information. What can they learn from the data about how to satisfy their best customers, how to allocate their resources most efficiently, and how to minimize losses? When there are millions of trees, how can one draw meaningful conclusions about the forest? In conjunction with the increased amount of data, there has been a shift in the primary users of computers, from a limited group of information systems professionals to a much larger group of knowledge workers with expertise in particular business domains, such as finance, marketing, or manufacturing. Data warehousing is a collection of technologies designed to convert heaps of data to usable information. It does this by consolidating data from diverse transactional systems into a coherent collection of consistent, quality-checked databases used only for informational purposes. Not only are data warehouses among the largest databases (frequently more than a terabyte), but they often have large numbers of users with diverse requirements. Consequently, they need carefully thought out architectures that take advantage of the most advanced multitier client/server computing tools.

Data warehouses are used in three primary ways. First, they enhance the traditional information presentation technologies (reports and graphs) by bringing the data necessary for their creation into a single source. This consolidation eliminates one of the biggest sources of error and delay: the fragmentation of data in diverse transaction databases. Second, data warehouses are used to support online analytical processing (OLAP). Whereas traditional query and report tools describe what is in a database, OLAP goes further in helping the user answer why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. For example, an analyst might hypothesize that people with low incomes

and high debt are bad credit risks, and analyze the database with OLAP to verify (or disprove) this assumption.

However, the very size and complexity of data warehouses make it difficult for any user, no matter how knowledgeable in the application of data, to formulate all possible hypotheses that might explain something such as the behavior of a group of customers. How can anyone successfully explore databases containing 100 million rows of data, each with thousands of attributes?

The newest, hottest technology to address these concerns is data mining. Data mining (the third major application of data warehouses) uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases—patterns that ordinary methods might miss.

Data mining is different from OLAP because, rather than verifying a hypothesis, it is used to generate a hypothesis. Say, for example, an analyst wants to identify the risk factors for granting credit. The data mining tool might discover that people with high debt and low incomes are bad credit risks (as before), but it might also discover a pattern that the analyst did not think to try, such as the fact that debt-to-income ratio and age are determinants of risk. Here is where data mining and OLAP complement each other. Before acting on the pattern, the analyst needs to know the financial implication of using the discovered pattern to govern who gets credit. The OLAP tool can allow the analyst to answer those kinds of questions. Together, data warehouses, OLAP, and data mining are transforming the way businesses use data. The resulting insights are producing dramatic returns on investment. A survey that Two Crows Corporation recently conducted provided strong evidence of corporate satisfaction. Of those organizations far enough along to have formed an opinion, all of them plan to continue to expand their present use of data mining.

In this book, Alex Berson and Steve Smith have brought together these different pieces of client/server computing, data warehousing, OLAP, and data mining and have provided an understandable and coherent explanation of how data mining works and how it can be used from the business perspective. I believe that this synergy among data warehouses, OLAP, and data mining will produce a new and significantly improved way of doing business across the enterprise that provides a real competitive advantage to those who make the most effective use of these technologies. This book will be a useful guide.

Herb Edelstein
President, Two Crows Corporation

Preface

The last few years have seen a growing recognition of information as a key business tool. Those who successfully gather, analyze, understand, and act upon information are among the winners in this new information age. Therefore, it is only reasonable to expect the rate of producing and consuming information to grow. We can define *information* as that which resolves uncertainty. We can further say that *decisionmaking* is the progressive resolution of uncertainty and is a key to a purposeful behavior by any mechanism (or organism). In general, the current business market dynamics make it abundantly clear that, for any company, information is the very key to survival.

If we look at the evolution of the information processing technologies, we can see that while the first generation of client/server systems brought data to the desktop, not all of this data was easy to understand, unfortunately, and as such, it was not very useful to end users. As a result, a number of new technologies have emerged that are focused on improving the information content of the data to empower the knowledge workers of today and tomorrow. Among these technologies are data warehousing, metadata repositories, online analytical processing (OLAP), and data mining. In some ways, these technologies are the manifestation of the maturity of the client/server computing model and its applicability to a wide variety of business problems.

Therefore, this book is about the need, the value, and the technological means of acquiring and using information in the information age.

From that perspective, this book is intended to become the handbook and the guide for anybody who's interested in, planning, or working on data warehousing and related issues. This audience is quite large, and includes both technology and business people. Among them are information technology managers, business analysts, marketing managers, product planners, client/server application developers, systems and database administrators, information security officers, data center operations staff, and data networking specialists. Data warehousing and its advantages, features, and usage are discussed against the background of the evolution of the computing models, hardware and software innovations for parallel processing, client/server architecture and implementations, and database management systems. Using these topics as a foundation, the book proceeds to analyze the components of a data warehouse, including

data sourcing and transformation tools, parallel database technology, metadata management, query, reporting, OLAP, data mining tools, and information delivery over the Web. Armed with the knowledge of data warehousing technology, the reader continues into a discussion on the principles of business analysis, models and patterns, and an in-depth analysis of data mining. The book ends with a brief look into the future potential of these technologies.

Why This Book Is Needed

The amount of information related to the subject of data warehousing is tremendous. Moreover, as technologies continue to mature, areas like OLAP, data mining, the World Wide Web, and parallel database technologies continue to attract the attention of developers, strategists, and users alike. At the same time, the amount of information about some of these technologies is still limited, while the hype surrounding other technologies (e.g., data mining) continues to further complicate the choices. To sort through all the available information, to separate hype from reality, and to find a cohesive and complete description of data warehousing and its effect on business is extremely difficult.

The main premise of this book is that, to date, several technologies have matured independently of one another, and no one as yet has thought a great deal about how to put the pieces together. These technologies include the following:

- Data warehouse–enabled relational database systems that are designed to support very large databases (VLDB) at significantly higher levels of performance and manageability.
- Data sourcing and transformation tools that can help acquire, understand, and clean up data stored in legacy and traditional online transaction processing (OLTP) systems before it gets loaded into a data warehouse.
- OLAP that is driven by the business need for an information view that can be rapidly assimilated and manipulated by business users and by the technology need to adapt standards-based solutions (e.g., relational database technology) and to leverage the opportunities available through the Web.
- Data mining that is enabled by new algorithms that provide easy-to-use and understandable techniques and that is driven by the business need to automatically solve well-defined business problems.

Never before has there been an opportunity to combine these technologies into one integrated system. To date, there are a number of very good books on various aspects of data warehousing. Unfortunately, most of the books published to date focus on a specific topic and a specific technology, without recognizing that the technologies have vastly greater value together—each improves the utility of the other like interlocking puzzle pieces. This book reveals how the technologies and architectures work together and what value they provide to the

end user. This book presents the big picture by showing the businessperson how a data warehouse can be made useful to him or her.

Another unique aspect of this book is that it contains a lot of material, some of which can be found in various vendor publications and in specialized research and trade literature. That is especially important because a significant portion of the available information is being changed on a regular basis. Various emerging standards and continuous product updates are examples of the dynamic nature of this material. The technologies and tools described in this book require a detailed knowledge of different hardware and software platforms. Specifically, the hardware platforms described in this book include midrange systems, parallel processors, workstations, and servers. Operating systems include UNIX, Windows/NT, NetWare, and OS/2. Database management systems discussions are focused on key features of SYBASE, ORACLE, INFORMIX, MS SQL Server, DB2, and Red Brick. The book discusses object-relational database technology of universal servers, a star schema design, and the effect of the Web on all components of data warehousing. Readers are also introduced to technologies and products from Arbor Software, Cognos, Constellar, Evolutionary Technologies, Informatica, Information Builders, LogicWorks, MicroStrategies, Prism Solution, and Vality, among many others.

Unfortunately, even if one decides to read all the available literature, it would be very difficult to obtain a clear picture of how all these technologies and products fit together to deliver value to a business enterprise. That is why the authors' personal experiences in developing large-scale data warehousing projects and extended involvement with commercial parallel computing, OLAP, machine learning, artificial intelligence, and the Internet proved to be invaluable in writing this book.

Who This Book Is For

This book has been written as a result of the authors' experiences in participating in several large-scale data warehousing projects and in developing OLAP and data mining solutions for various industry segments.

For the discussion of the architecture, advantages, and benefits of data warehousing, the authors met with many business and IT managers, systems integrators, system administrators, database and data communications specialists, and system programmers, all of whom may be potential readers of this book.

This book can be used as a guide for system integrators, designers of data warehouse and data mart systems, data and database administrators considering the issues of parallel relational database systems, OLAP designers, and those who are planning to implement and support data mining. Webmasters, network specialists, and information security officers will find this book useful for implementing a distributed data warehouse or for deploying Web-enabled analysis tools throughout an enterprise.

Some specific data warehouse components described in the book can help IT managers, system administrators, DBAs, network and communications

specialists, and application developers to make informed decisions when selecting platforms and products to implement a data warehouse or a data mart. The maturity of various OLAP and data mining technologies has enabled the authors to discuss design, implementation, and operational issues at such a level of detail that the book should be an invaluable tool for any professional in solving a whole spectrum of issues and concerns related to data warehousing.

Finally, those readers who are looking into such advanced topics as object-relational database systems, high-performance commercial computing, OLAP, and data mining will find this book extremely useful.

Prerequisite

The authors assume readers have little or no previous knowledge about data warehousing. This book is targeted at two classes of readership: business professionals—including sales and marketing managers, product planners, and financial experts—and technology professionals. Both groups of readers can understand this book—no previous data warehousing experience is necessary. Readers with any degree of knowledge of information technology can benefit from this book. Those who deal with only COBOL batch programs will find this book useful. Those with CICS, SQL, DB2 or any other database expertise, including DBA experience, will benefit. UNIX, Windows and Windows/NT, OS/2, and NetWare application developers, systems and network administrators, and LAN specialists should not have any problems reading this book.

Style Used

The book has been structured as a self-teaching guide. The introduction to data warehousing, its relationship to the client/server architecture, and an overview of data warehousing technology components and their roles is placed in the first part. The rest of the book is dedicated to specific technologies and methodologies designed to implement a data warehouse, with an in-depth discussion of business analysis, OLAP, data mining, and data visualization. The book concludes with a brief look at prevailing trends and directions in the data warehouse market.

The book includes a fair amount of diagrams, figures, examples, and illustrations in an attempt to present a lot of rather complicated material in as simple a form as possible. Data warehousing is a complex, involved, and often-misunderstood subject; so, whenever possible, theoretical issues are explained with practical examples. Therefore, the authors have made a serious effort to explain complex issues of parallel relational database systems, OLAP, and data mining, using both simple examples and theoretical discussions. For those readers interested in theory, the book provides sufficient theoretical overview of star schema design, parallel systems, artificial intelligence, and predictive modeling.

This book is about a very dynamic subject. All material included in the book is current at the time of writing. The authors realize that as data warehousing continues to evolve, and as vendors continue to improve and expand on their product quality and functionality, changes will be necessary. The authors intend to revise the book if a significant development in the data warehousing arena makes it necessary to add, delete, or change parts of the text.

What Is Included

Part I begins with an introduction to the business imperative and the technology roadmap of data warehousing. This part discusses the relationship between a data warehouse and client/server architecture and provides an overview of parallel system architectures and the corresponding developments in the area of database systems.

Part II starts with an in-depth analysis of data warehouse architecture and components, and discusses the design, technical, and implementation considerations of building a data warehouse. This part describes how a relational database technology can be leveraged for the high scalability and very large database support required by a data warehouse. Star schema design, bitmap indexes, and other innovative techniques are also discussed in this part. Finally, this part provides an overview of data extraction, transformation, and cleanup tools, as well as a discussion on the importance of metadata and the issues surrounding its management.

Part III begins to introduce the reader to the technical considerations related to business analysis. Query and reporting tools, OLAP, and the ideas behind models, patterns, statistics, and artificial intelligence are discussed in this part.

Part IV focuses on data mining. Decision trees, neural networks, clustering, nearest neighbor, fuzzy logic, genetic algorithms, and rule induction are among the techniques discussed in this part. In addition, a discussion on how to select the right technique is presented.

Part V concludes with a discussion on data visualization and an in-depth look at the current trends and future directions in the data warehouse arena.

The *appendixes* include an article on the value of data mining, OLAP guidelines, an analysis of common mistakes made when building a data warehouse, and an extensive bibliography.

Acknowledgments

First, I am grateful to Steve Smith for his knowledge, persistence, attention to details, and dedication, without which this book would not have happened. Very special thanks to my many friends and colleagues at Merrill Lynch for providing a creative and challenging atmosphere. Working with people like George Lieberman, Joe Hollander, Scott Ryles, John Ginelli, Tom Musmanno, Steve Wolfe, Guy Pujol, Joe Frediani, and many others gave me an opportunity

to learn and work in a very stimulating and challenging environment on the leading edge of computer technology.

I also have to thank my numerous friends at ADT, Cognizant, Informix, IBM, Pilot Software, and ICS, specifically Peter Meekin, Eric Kim, Larry Johnson, and John Pezzullo.

I am very grateful to Dr. Ramon Barquin for his invaluable help and kindness by allowing me to include his insightful "10 Mistakes . . ." in this book.

I would like to thank all those who have helped me with clarifications, criticism, and valuable information during the writing of this book, including Herb Edelstein, who not only provided many thoughtful insights, but was patient enough to read the entire manuscript and make may useful suggestions. And, of course, this book would never have been finished without the invaluable assistance and thoroughness of McGraw-Hill editors and M.R. Carey of North Market Street Graphics.

Finally, the key reason for this book's existence is my family. My very special thanks to Irina, Vlad, Michelle, and the rest of my family for giving me time to complete the book, for understanding its importance, and for never-ending optimism, support, and love. I am especially grateful to my son Vlad for his help in designing the illustration material (and my personal home page on the Web).

Alex Berson

Since this is my first book I'd like to give credit where credit is due—to my high school teachers, who taught me how to write and started me out in science. I might have figured it out on my own later, but perhaps not. For me, there is no doubt that my ability to write at all rests with the patient encouragement of my teachers Ms. Durish, Mrs. Meys, and Mr. Palmer—and the less than patient but humorous encouragement of Mr. Rullo (though this is probably not exactly what they were expecting when they were making me read Dickens, Sartre, and Hardy).

Alex—thanks for your optimism and good sense to "just keep writing"—this book could, of course, not have happened without you.

Thanks to my parents for making me take piano lessons—which made me want to study instead. Thanks to Noel for giving me the time (sorry I ran a little bit late). Thanks to Debbie, for always checking to see if I was done yet. And special thanks to Samantha, Nathaniel, Emily, Sheri, and Irene.

And, finally, my sincerest appreciation to my colleagues and teachers who have taught me a great deal about what I have written here today: Mario Bourgoin, Joe Yarmus, Kurt Thearling, Emily Stone, Gary Drescher, Brij Masand, Jim Hutchinson, Xiru Zhang, Kris Carlson, Dave Waltz, Danny Hillis, Craig Stanfill. Craig Shaefer, Stewart Wilson, Tommy Poggio, Charles Leiserson, Ron Rivest, Alan Zaslavsky, Jim Clark, Eric Kim, Herb Edelstein and Peter Meekin.

Steve Smith