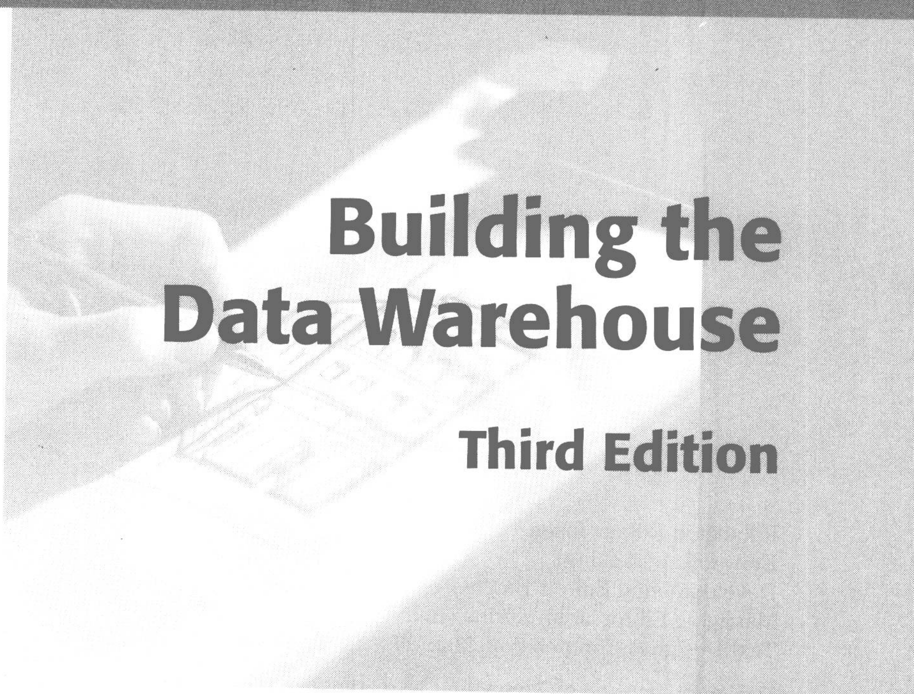
WILEY

TIMELY. PRACTICAL. RELIABLE.

Building the Data Warehouse

Third Edition

W. H. Inmon



Building the Data Warehouse

Third Edition

W. H. Inmon

Wiley Computer Publishing

John Wiley & Sons, Inc.
NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

Publisher: Robert Ipsen
Editor: Robert Elliott
Developmental Editor: Emilie Herman
Managing Editor: John Atkins
Text Design & Composition: MacAllister Publishing Services, LLC

This book is printed on acid-free paper.

Published by John Wiley & Sons, Inc.

Published simultaneously in Canada.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

Library of Congress Cataloging-in-Publication Data:

ISBN: 0-471-08130-2

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Building the Data Warehouse

Third Edition

To Jeanne Friedman—a friend for all times

PREFACE FOR THE SECOND EDITION

Databases and database theory have been around for a long time. Early renditions of databases centered around a single database serving every purpose known to the information processing community—from transaction to batch processing to analytical processing. In most cases, the primary focus of the early database systems was operational—usually transactional—processing. In recent years, a more sophisticated notion of the database has emerged—one that serves operational needs and another that serves informational or analytical needs. To some extent, this more enlightened notion of the database is due to the advent of PCs, 4GL technology, and the empowerment of the end user.

The split of operational and informational databases occurs for many reasons:

- The data serving operational needs is physically different data from that serving informational or analytic needs.
- The supporting technology for operational processing is fundamentally different from the technology used to support informational or analytical needs.
- The user community for operational data is different from the one served by informational or analytical data.
- The processing characteristics for the operational environment and the informational environment are fundamentally different.

Because of these reasons (and many more), the modern way to build systems is to separate the operational from the informational or analytical processing and data.

This book is about the analytical [or the decision support systems (DSS)] environment and the structuring of data in that environment. The focus of the book is on what is termed the "data warehouse" (or "information warehouse"), which is at the heart of informational, DSS processing.

The discussions in this book are geared to the manager and the developer. Where appropriate, some level of discussion will be at the technical level. But, for the most part, the book is about issues and techniques. This book is meant to serve as a guideline for the designer and the developer.

PREFACE FOR THE THIRD EDITION

When the first edition of *Building the Data Warehouse* was printed, the database theorists scoffed at the notion of the data warehouse. One theoretician stated that data warehousing set back the information technology industry 20 years. Another stated that the founder of data warehousing should not be allowed to speak in public. And yet another academic proclaimed that data warehousing was nothing new and that the world of academia had known about data warehousing all along although there were no books, no articles, no classes, no seminars, no conferences, no presentations, no references, no papers, and no use of the terms or concepts in existence in academia at that time.

When the second edition of the book appeared, the world was mad for anything of the Internet. In order to be successful it had to be "e" something—e-business, e-commerce, e-tailing, and so forth. One venture capitalist was known to say, "Why do we need a data warehouse when we have the Internet?"

But data warehousing has surpassed the database theoreticians who wanted to put all data in a single database. Data warehousing survived the dot.com disaster brought on by the short-sighted venture capitalists. In an age when technology in general is spurned by Wall Street and Main Street, data warehousing has never been more alive or stronger. There are conferences, seminars, books, articles, consulting, and the like. But mostly there are companies doing data warehousing, and making the discovery that, unlike the overhyped New Economy, the data warehouse actually delivers, even though Silicon Valley is still in a state of denial.

The third edition of this book heralds a newer and even stronger day for data warehousing. Today data warehousing is not a theory but a fact of life. New technology is right around the corner to support some of the more exotic needs of a data warehouse. Corporations are running major pieces of their business on data warehouses. The cost of information has dropped dramatically because of data warehouses. Managers at long last have a viable solution to the ugliness of the legacy systems environment. For the first time, a corporate "memory" of historical information is available. Integration of data across the corporation is a real possibility, in most cases for the first time. Corporations are learning how to go from data to information to competitive advantage. In short, data warehousing has unlocked a world of possibility.

One confusing aspect of data warehousing is that it is an architecture, not a technology. This frustrates the technician and the venture capitalist alike because these people want to buy something in a nice clean box. But data warehousing simply does not lend itself to being "boxed up." The difference between an architecture and a technology is like the difference between Santa Fe, New Mexico, and adobe bricks. If you drive the streets of Santa Fe you know you are there and nowhere else. Each home, each office building, each restaurant has a distinctive look that says "This is Santa Fe." The look and style that make Santa Fe distinctive are the architecture. Now, that architecture is made up of such things as adobe bricks and exposed beams. There is a whole art to the making of adobe bricks and exposed beams. And it is certainly true that you could not have Santa Fe architecture without having adobe bricks and exposed beams. But adobe bricks and exposed beams by themselves do not make an architecture. They are independent technologies. For example, you have adobe bricks throughout the Southwest and the rest of the world that are not Santa Fe architecture.

Thus it is with architecture and technology, and with data warehousing and databases and other technology. There is the architecture, then there is the underlying technology, and they are two very different things. Unquestionably, there is a relationship between data warehousing and database technology, but they are most certainly not the same. Data warehousing requires the support of many different kinds of technology.

With the third edition of this book, we now know what works and what does not. When the first edition was written, there was some experience with developing and using warehouses, but truthfully, there was not the broad base of experience that exists today. For example, today we know with certainty the following:

- Data warehouses are built under a different development methodology than applications. Not keeping this in mind is a recipe for disaster.
- Data warehouses are fundamentally different from data marts. The two do not mix—they are like oil and water.
- Data warehouses deliver on their promise, unlike many overhyped technologies that simply faded away.
- Data warehouses attract huge amounts of data, to the point that entirely new approaches to the management of large amounts of data are required.

But perhaps the most intriguing thing that has been learned about data warehousing is that data warehouses form a foundation for many other forms of processing. The granular data found in the data warehouse can be reshaped and reused. If there is any immutable and profound truth about data warehouses, it is that data warehouses provide an ideal foundation for many other forms of information processing. There are a whole host of reasons why this foundation is so important:

- There is a single version of the truth.
- Data can be reconciled if necessary.
- Data is immediately available for new, unknown uses.

And, finally, data warehousing has lowered the cost of information in the organization. With data warehousing, data is inexpensive to get to and fast to access.

Databases and database theory have been around for a long time. Early renditions of databases centered around a single database serving every purpose known to the information processing community—from transaction to batch processing to analytical processing. In most cases, the primary focus of the early database systems was operational—usually transactional—processing. In recent years, a more sophisticated notion of the database has emerged—one that serves operational needs and another that serves informational or analytical needs. To some extent, this more enlightened notion of the database is due to the advent of PCs, 4GL technology, and the empowerment of the end user.

The split of operational and informational databases occurs for many reasons:

- The data serving operational needs is physically different data from that serving informational or analytic needs.
- The supporting technology for operational processing is fundamentally different from the technology used to support informational or analytical needs.
- The user community for operational data is different from the one served by informational or analytical data.
- The processing characteristics for the operational environment and the informational environment are fundamentally different.

For these reasons (and many more), the modern way to build systems is to separate the operational from the informational or analytical processing and data.

This book is about the analytical or the DSS environment and the structuring of data in that environment. The focus of the book is on what is termed the data warehouse (or information warehouse), which is at the heart of informational, DSS processing.

What is analytical, informational processing? It is processing that serves the needs of management in the decision-making process. Often known as DSS processing, analytical processing looks across broad vistas of data to detect trends. Instead of looking at one or two records of data (as is the case in operational processing), when the DSS analyst does analytical processing, many records are accessed.

It is rare for the DSS analyst to update data. In operational systems, data is constantly being updated at the individual record level. In analytical processing, records are constantly being accessed, and their contents are gathered for analysis, but little or no alteration of individual records occurs.

In analytical processing, the response time requirements are greatly relaxed compared to those of traditional operational processing. Analytical response time is measured from 30 minutes to 24 hours. Response times measured in this range for operational processing would be an unmitigated disaster.

The network that serves the analytical community is much smaller than the one that serves the operational community. Usually there are far fewer users of the analytical network than of the operational network.

Unlike the technology that serves the analytical environment, operational environment technology must concern itself with data and transaction locking, contention for data, deadlock, and so on.

There are, then, many major differences between the operational environment and the analytical environment. This book is about the analytical, DSS environment and addresses the following issues:

- Granularity of data
- Partitioning of data
- Meta data
- Lack of credibility of data
- Integration of DSS data
- The time basis of DSS data
- Identifying the source of DSS data-the system of record
- Migration and methodology

This book is for developers, managers, designers, data administrators, database administrators, and others who are building systems in a modern data processing environment. In addition, students of information processing will find this book useful. Where appropriate, some discussions will be more technical. But, for the most part, the book is about issues and techniques, and it is meant to serve as a guideline for the designer and the developer.

This book is the first in a series of books relating to data warehouse. The next book in the series is *Using the Data Warehouse* (Wiley, 1994). *Using the Data Warehouse* addresses the issues that arise once you have built the data warehouse. In addition, *Using the Data Warehouse* introduces the concept of a larger architecture and the notion of an operational data store (ODS). An operational data store is a similar architectural construct to the data warehouse, except the ODS applies only to operational systems, not informational systems. The third book in the series is *Building the Operational Data Store* (Wiley, 1999), which addresses the issues of what an ODS is and how an ODS is built.

The next book in the series is *Corporate Information Factory, Third Edition* (Wiley, 2002). This book addresses the larger framework of which the data warehouse is the center. In many regards the CIF book and the DW book are companions. The CIF book provides the larger picture and the DW book provides a more focused discussion. Another related book is *Exploration Warehousing* (Wiley, 2000). This book addresses a specialized kind of processing-pattern analysis using statistical techniques on data found in the data warehouse.

Building the Data Warehouse, however, is the cornerstone of all the related books. The data warehouse forms the foundation of all other forms of DSS processing.

There is perhaps no more eloquent testimony to the advances made by data warehousing and the corporate information factory than the References at the back of this book. When the first edition was published, there were no other books, no white papers, and only a handful of articles that could be referenced. In this third edition, there are many books, articles, and white papers that are mentioned. Indeed the references only start to explore some of the more important works.

ACKNOWLEDGMENTS

The following people have influenced—directly and indirectly—the material found in this book. The author is grateful for the long-term relationships that have been formed and for the experiences that have provided a basis for learning.

Claudia Imhoff, Intelligent Solutions

Jon Geiger, Intelligent Solutions

Joyce Norris Montanari, Intelligent Solutions

John Zachman, Zachman International

John Ladley, Meta Group

Bob Terdeman, EMC Corporation

Dan Meers, BillInmon.com

Cheryl Estep, independent consultant

Lowell Fryman, independent consultant

David Fender, SAS Japan

Jim Davis, SAS

Peter Grendel, SAP

Allen Houpt, CA

ABOUT THE AUTHOR

Bill Inmon, the father of the data warehouse concept, has written 40 books on data management, data warehouse, design review, and management of data processing. Bill has had his books translated into Russian, German, French, Japanese, Portuguese, Chinese, Korean, and Dutch. Bill has published more than 250 articles in many trade journals. Bill founded and took public Prism Solutions. His latest company—Pine Cone Systems—builds software for the management of the data warehouse/data mart environment. Bill holds two software patents. Articles, white papers, presentations, and much more material can be found on his Web site, *www.billinmon.com.*

CONTENTS