

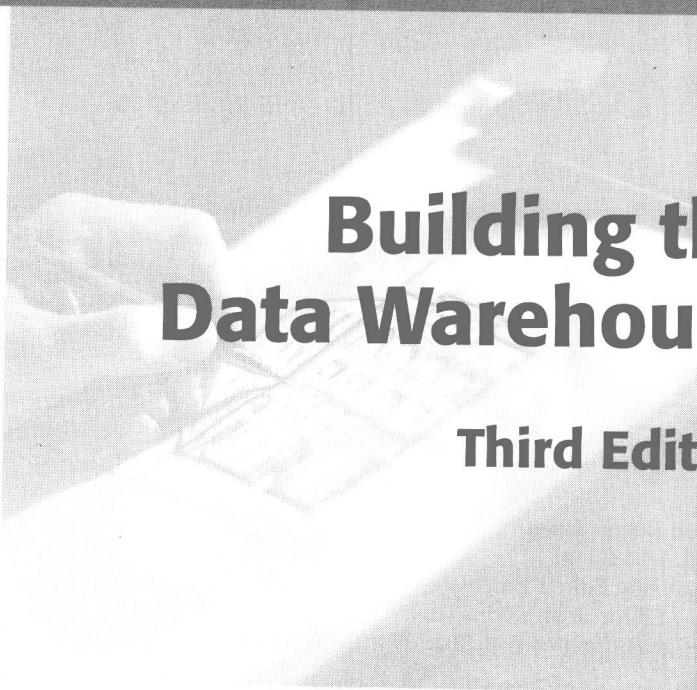
Building the Data Warehouse

Third Edition



J. H. Inmon

311.13
57
.3



Building the Data Warehouse

Third Edition

W. H. Inmon



E200201475

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

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or ALL CAPITAL LETTERS. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

This book is printed on acid-free paper.

Copyright © 2002 by W.H. Inmon. All rights reserved.

Published by John Wiley & Sons, Inc.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ @ WILEY.COM.

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

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.

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 pro-

cessing, 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*.

Preface for the Second Edition	xiii
Preface for the Third Edition	xiv
Acknowledgments	xix
About the Author	xx
 Chapter 1 Evolution of Decision Support Systems	 1
The Evolution	2
The Advent of DASD	4
PC/4GL Technology	4
Enter the Extract Program	5
The Spider Web	6
Problems with the Naturally Evolving Architecture	6
Lack of Data Credibility	6
Problems with Productivity	9
From Data to Information	12
A Change in Approach	15
The Architected Environment	16
Data Integration in the Architected Environment	19
Who Is the User?	19
The Development Life Cycle	21
Patterns of Hardware Utilization	22
Setting the Stage for Reengineering	23
Monitoring the Data Warehouse Environment	25
Summary	28
 Chapter 2 The Data Warehouse Environment	 31
The Structure of the Data Warehouse	35
Subject Orientation	36
Day 1-Day n Phenomenon	41
Granularity	43
The Benefits of Granularity	45
An Example of Granularity	46
Dual Levels of Granularity	49

Exploration and Data Mining	53
Living Sample Database	53
Partitioning as a Design Approach	55
Partitioning of Data	56
Structuring Data in the Data Warehouse	59
Data Warehouse: The Standards Manual	64
Auditing and the Data Warehouse	64
Cost Justification	65
Justifying Your Data Warehouse	66
Data Homogeneity/Heterogeneity	69
Purging Warehouse Data	72
Reporting and the Architected Environment	73
The Operational Window of Opportunity	74
Incorrect Data in the Data Warehouse	76
Summary	77
Chapter 3 The Data Warehouse and Design	81
Beginning with Operational Data	82
Data/Process Models and the Architected Environment	87
The Data Warehouse and Data Models	89
The Data Warehouse Data Model	92
The Midlevel Data Model	94
The Physical Data Model	98
The Data Model and Iterative Development	102
Normalization/Denormalization	102
Snapshots in the Data Warehouse	110
Meta Data	113
Managing Reference Tables in a Data Warehouse	113
Cyclicity of Data-The Wrinkle of Time	115
Complexity of Transformation and Integration	118
Triggering the Data Warehouse Record	122
Events	122
Components of the Snapshot	123
Some Examples	123
Profile Records	124
Managing Volume	126
Creating Multiple Profile Records	127

	Going from the Data Warehouse to the Operational Environment	128
	Direct Access of Data Warehouse Data	129
	Indirect Access of Data Warehouse Data	130
	An Airline Commission Calculation System	130
	A Retail Personalization System	132
	Credit Scoring	133
	Indirect Use of Data Warehouse Data	136
	Star Joins	137
	Supporting the ODS	143
	Summary	145
Chapter 4	Granularity in the Data Warehouse	147
	Raw Estimates	148
	Input to the Planning Process	149
	Data in Overflow?	149
	Overflow Storage	151
	What the Levels of Granularity Will Be	155
	Some Feedback Loop Techniques	156
	Levels of Granularity-Banking Environment	158
	Summary	165
Chapter 5	The Data Warehouse and Technology	167
	Managing Large Amounts of Data	167
	Managing Multiple Media	169
	Index/Monitor Data	169
	Interfaces to Many Technologies	170
	Programmer/Designer Control of Data Placement	171
	Parallel Storage/Management of Data	171
	Meta Data Management	171
	Language Interface	173
	Efficient Loading of Data	173
	Efficient Index Utilization	175
	Compaction of Data	175
	Compound Keys	176
	Variable-Length Data	176
	Lock Management	176

Index-Only Processing	178
Fast Restore	178
Other Technological Features	178
DBMS Types and the Data Warehouse	179
Changing DBMS Technology	181
Multidimensional DBMS and the Data Warehouse	182
Data Warehousing across Multiple Storage Media	188
Meta Data in the Data Warehouse Environment	189
Context and Content	192
Three Types of Contextual Information	193
Capturing and Managing Contextual Information	194
Looking at the Past	195
Refreshing the Data Warehouse	195
Testing	198
Summary	198
Chapter 6 The Distributed Data Warehouse	201
Types of Distributed Data Warehouses	202
Local and Global Data Warehouses	202
The Technologically Distributed Data Warehouse	220
The Independently Evolving Distributed Data Warehouse	221
The Nature of the Development Efforts	222
Completely Unrelated Warehouses	224
Distributed Data Warehouse Development	226
Coordinating Development across Distributed Locations	227
The Corporate Data Model-Distributed	228
Meta Data in the Distributed Warehouse	232
Building the Warehouse on Multiple Levels	232
Multiple Groups Building the Current Level of Detail	235
Different Requirements at Different Levels	238
Other Types of Detailed Data	239
Meta Data	244
Multiple Platforms for Common Detail Data	244
Summary	245
Chapter 7 Executive Information Systems and the Data Warehouse	247
EIS-The Promise	248
A Simple Example	248
Drill-Down Analysis	251