Building the Data Warehouse,
Fourth Edition
Building the Data Warehouse, Fourth Edition

W. H. Inmon
Executive Editor
Robert Elliott

Development Editor
Kevin Shafer

Production Editor
Pamela Hanley

Copy Editor
Kathi Duggan

Editorial Manager
Mary Beth Wakefield

Production Manager
Tim Tate

Vice President & Executive Group Publisher
Richard Swadley

Vice President and Publisher
Joseph B. Wikert

Project Coordinator
Erin Smith

Graphics and Production Specialists
Jonelle Burns
Kelly Emkow
Carrie A. Foster
Joyce Haughey
Jennifer Heleine
Stephanie D. Jumper

Quality Control Technician
Leeann Harney

Proofreading and Indexing
TECHBOOKS Production Services
To Jeanne Friedman and Kevin Gould – friends for all times.
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 xix
Acknowledgments xxvii

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 7
      Lack of Data Credibility 7
      Problems with Productivity 9
      From Data to Information 12
      A Change in Approach 14
      The Architected Environment 16
      Data Integration in the Architected Environment 18
      Who Is the User? 20
   The Development Life Cycle 20
   Patterns of Hardware Utilization 22
   Setting the Stage for Re-engineering 23
   Monitoring the Data Warehouse Environment 25
   Summary 28

Chapter 2 The Data Warehouse Environment 29
   The Structure of the Data Warehouse 33
   Subject Orientation 34
   Day 1 to Day n Phenomenon 39
Star Joins
Supporting the ODS
Requirements and the Zachman Framework
Summary

Chapter 4  **Granularity in the Data Warehouse**

Raw Estimates
Input to the Planning Process
Data in Overflow
Overflow Storage
What the Levels of Granularity Will Be
Some Feedback Loop Techniques
Levels of Granularity — Banking Environment
Feeding the Data Marts
Summary

Chapter 5  **The Data Warehouse and Technology**

Managing Large Amounts of Data
Managing Multiple Media
Indexing and Monitoring Data
Interfaces to Many Technologies
Programmer or Designer Control of Data Placement
Parallel Storage and Management of Data
Metadata Management
Language Interface
Efficient Loading of Data
Efficient Index Utilization
Compaction of Data
Compound Keys
Variable-Length Data
Lock Management
Index-Only Processing
Fast Restore
Other Technological Features
DBMS Types and the Data Warehouse
Changing DBMS Technology
Multidimensional DBMS and the Data Warehouse
Data Warehousing across Multiple Storage Media
The Role of Metadata in the Data Warehouse Environment
Context and Content
  Three Types of Contextual Information
  Capturing and Managing Contextual Information
  Looking at the Past
Refreshing the Data Warehouse
Testing
Summary
<table>
<thead>
<tr>
<th>Chapter 6</th>
<th>The Distributed Data Warehouse</th>
<th>193</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of Distributed Data Warehouses</td>
<td>193</td>
<td></td>
</tr>
<tr>
<td>Local and Global Data Warehouses</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>The Local Data Warehouse</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>The Global Data Warehouse</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>Intersection of Global and Local Data</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>Redundancy</td>
<td>206</td>
<td></td>
</tr>
<tr>
<td>Access of Local and Global Data</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>The Technologically Distributed Data Warehouse</td>
<td>211</td>
<td></td>
</tr>
<tr>
<td>The Independently Evolving Distributed Data Warehouse</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>The Nature of the Development Efforts</td>
<td>213</td>
<td></td>
</tr>
<tr>
<td>Completely Unrelated Warehouses</td>
<td>215</td>
<td></td>
</tr>
<tr>
<td>Distributed Data Warehouse Development</td>
<td>217</td>
<td></td>
</tr>
<tr>
<td>Coordinating Development across Distributed Locations</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>The Corporate Data Model — Distributed</td>
<td>219</td>
<td></td>
</tr>
<tr>
<td>Metadata in the Distributed Warehouse</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Building the Warehouse on Multiple Levels</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Multiple Groups Building the Current Level of Detail</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>Different Requirements at Different Levels</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>Other Types of Detailed Data</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>Metadata</td>
<td>234</td>
<td></td>
</tr>
<tr>
<td>Multiple Platforms for Common Detail Data</td>
<td>235</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>236</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 7</th>
<th>Executive Information Systems and the Data Warehouse</th>
<th>239</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIS — The Promise</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>A Simple Example</td>
<td>240</td>
<td></td>
</tr>
<tr>
<td>Drill-Down Analysis</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>Supporting the Drill-Down Process</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>The Data Warehouse as a Basis for EIS</td>
<td>247</td>
<td></td>
</tr>
<tr>
<td>Where to Turn</td>
<td>248</td>
<td></td>
</tr>
<tr>
<td>Event Mapping</td>
<td>251</td>
<td></td>
</tr>
<tr>
<td>Detailed Data and EIS</td>
<td>253</td>
<td></td>
</tr>
<tr>
<td>Keeping Only Summary Data in the EIS</td>
<td>254</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>255</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 8</th>
<th>External Data and the Data Warehouse</th>
<th>257</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Data in the Data Warehouse</td>
<td>260</td>
<td></td>
</tr>
<tr>
<td>Metadata and External Data</td>
<td>261</td>
<td></td>
</tr>
<tr>
<td>Storing External Data</td>
<td>263</td>
<td></td>
</tr>
<tr>
<td>Different Components of External Data</td>
<td>264</td>
<td></td>
</tr>
<tr>
<td>Modeling and External Data</td>
<td>265</td>
<td></td>
</tr>
<tr>
<td>Secondary Reports</td>
<td>266</td>
<td></td>
</tr>
<tr>
<td>Archiving External Data</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>Comparing Internal Data to External Data</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>268</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 9  Migration to the Architected Environment  269
A Migration Plan  270
The Feedback Loop  278
Strategic Considerations  280
Methodology and Migration  283
A Data-Driven Development Methodology  283
Data-Driven Methodology  286
System Development Life Cycles  286
A Philosophical Observation  286
Summary  287

Chapter 10  The Data Warehouse and the Web  289
Supporting the eBusiness Environment  299
Moving Data from the Web to the Data Warehouse  300
Moving Data from the Data Warehouse to the Web  301
Web Support  302
Summary  302

Chapter 11  Unstructured Data and the Data Warehouse  305
Integrating the Two Worlds  307
Text — The Common Link  308
A Fundamental Mismatch  310
Matching Text across the Environments  310
A Probabilistic Match  311
Matching All the Information  312
A Themed Match  313
Industrially Recognized Themes  313
Naturally Occurring Themes  316
Linkage through Themes and Themed Words  317
Linkage through Abstraction and Metadata  318
A Two-Tiered Data Warehouse  320
Dividing the Unstructured Data Warehouse  321
Documents in the Unstructured Data Warehouse  322
Visualizing Unstructured Data  323
A Self-Organizing Map (SOM)  324
The Unstructured Data Warehouse  325
Volumes of Data and the Unstructured Data Warehouse  326
Fitting the Two Environments Together  327
Summary  330

Chapter 12  The Really Large Data Warehouse  331
Why the Rapid Growth?  332
The Impact of Large Volumes of Data  333
Basic Data-Management Activities  334
The Cost of Storage  335
The Real Costs of Storage  336
The Usage Pattern of Data in the Face of Large Volumes  336
A Simple Calculation 337
Two Classes of Data 338
Implications of Separating Data into Two Classes 339
Disk Storage in the Face of Data Separation 340
Near-Line Storage 341
Access Speed and Disk Storage 342
Archival Storage 343
Implications of Transparency 345
Moving Data from One Environment to Another 346
The CMSM Approach 347
A Data Warehouse Usage Monitor 348
The Extension of the Data Warehouse across Different Storage Media 349
Inverting the Data Warehouse 350
Total Cost 351
Maximum Capacity 352
Summary 354

Chapter 13 The Relational and the Multidimensional Models as a Basis for Database Design 357
The Relational Model 357
The Multidimensional Model 360
Snowflake Structures 361
Differences between the Models 362
The Roots of the Differences 363
Reshaping Relational Data 364
Indirect Access and Direct Access of Data 365
Servicing Future Unknown Needs 366
Servicing the Need to Change Gracefully 367
Independent Data Marts 370
Building Independent Data Marts 371
Summary 375

Chapter 14 Data Warehouse Advanced Topics 377
End-User Requirements and the Data Warehouse 377
The Data Warehouse and the Data Model 378
The Relational Foundation 378
The Data Warehouse and Statistical Processing 379
Resource Contention in the Data Warehouse 380
The Exploration Warehouse 380
The Data Mining Warehouse 382
Freezing the Exploration Warehouse 383
External Data and the Exploration Warehouse 384
Data Marts and Data Warehouses in the Same Processor 384
The Life Cycle of Data 386
Mapping the Life Cycle to the Data Warehouse Environment 387
Testing and the Data Warehouse 388
Chapter 17 Corporate Information Compliance and Data Warehousing 443
Two Basic Activities 445
Financial Compliance 446
  The “What” 447
  The “Why” 449
Auditing Corporate Communications 452
Summary 454

Chapter 18 The End-User Community 457
The Farmer 458
The Explorer 458
The Miner 459
The Tourist 459
The Community 459
Different Types of Data 460
Cost-Justification and ROI Analysis 461
Summary 462

Chapter 19 Data Warehouse Design Review Checklist 463
When to Do a Design Review 464
Who Should Be in the Design Review? 465
What Should the Agenda Be? 465
The Results 465
Administering the Review 466
A Typical Data Warehouse Design Review 466
Summary 488

Glossary 489

References 507
  Articles 507
  Books 510
  White Papers 512

Index 517
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.

Preface for the Third Edition

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 makes 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.

Preface for the Fourth Edition

In the beginning was a theory of database that held that all data should be held in a common source. It was easy to see how this notion came about. Prior to database, there were master files. These master files resided on sequential media and were built for every application that came along. There was absolutely no integration among master files. Thus, the idea of integrating data into a single source — a database — held great appeal.

It was into this mindset that data warehouse was born. Data warehousing was an intellectual threat to those who subscribed to conventional database theory because data warehousing suggested that there ought to be different kinds of databases. And the thought that there should be different kinds of databases was not accepted by the database theoreticians.

Today, data warehousing has achieved the status of conventional wisdom. For a variety of reasons, data warehousing is just what you do. Recently a survey showed that corporate spending on data warehouse and business intelligence surpassed spending on transactional processing and OLTP, something unthinkable a few years back.

The day of data warehouse maturity has arrived.

It is appropriate, then, that the Fourth Edition of the book that began the data warehousing phenomenon has been written.

In addition to the time-honored concepts of data warehousing, this edition contains the data warehouse basics. But it also contains many topics current to today’s information infrastructure.
Following are some of the more important new topics in this edition:

- Compliance (dealing with Sarbanes Oxley, HIPAA, Basel II, and more)
- Near line storage (extending the data warehouse to infinity)
- Multi dimensional database design
- Unstructured data
- End users (who they are and what their needs are)
- ODS and the data warehouse

In addition to having new topics, this edition reflects that larger architecture that surrounds a data warehouse.

Technology has grown up with data warehousing. In the early days of data warehousing, 50 GB to 100 GB of data was considered a large warehouse. Today, some data warehouses are in the petabyte range. Other technology includes advances made in multi-dimensional technology — in data marts and star joins. Yet other technology advances have occurred in the storage of data on storage media other than disk storage.

In short, technology advances have made possible the technological achievements of today. Without modern technology, there would be no data warehouse.

This book is for architects and system designers. The end user may find this book useful as an explanation of what data warehousing is all about. And managers and students will also find this book to be useful.
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.

Guy Hildebrand, a partner like no other
Lynn Inmon, a wife and helpmate like no other
Ryan Sousa, a free thinker for the times
Jim Shank and Nick Johnson, without whom there would be nothing
Ron Powell and Shawn Rogers, friends and inspirations for all times
Joyce Norris Montanari, Intelligent Solutions, an inspiration throughout the ages
John Zachman, Zachman International, a friend and a world class architect
Dan Meers, BillInmon.com, a real visionary and a real friend
Cheryl Estep, independent consultant, who was there at the beginning
Claudia Imhoff, Intelligent Solutions
Jon Geiger, Intelligent Solutions
John Ladley, Meta Group
Bob Terdeman, EMC Corporation
Lowell Fryman, independent consultant
David Fender, SAS Japan
Jim Davis, SAS
Peter Grendel, SAP
Allen Houpt, CA