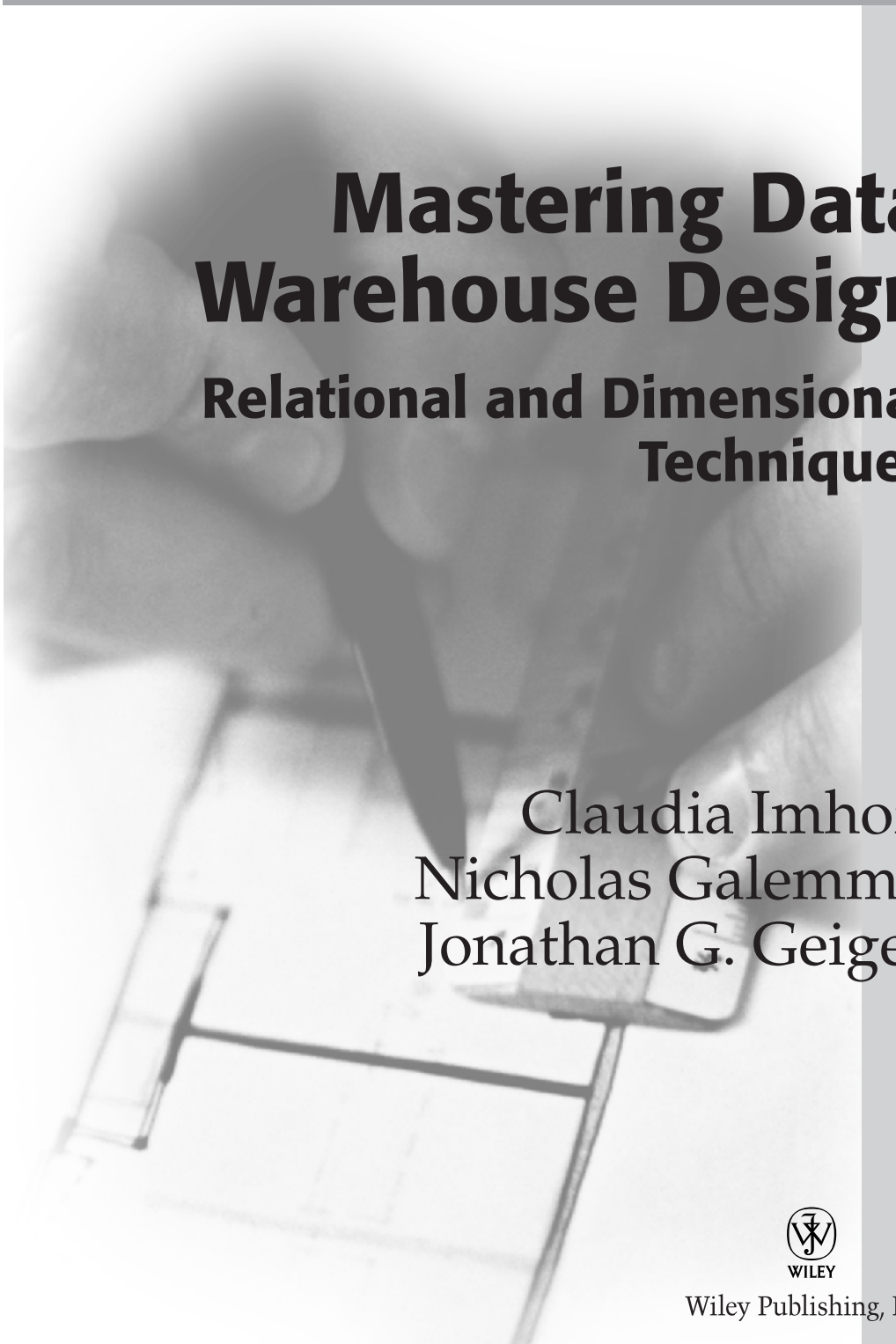
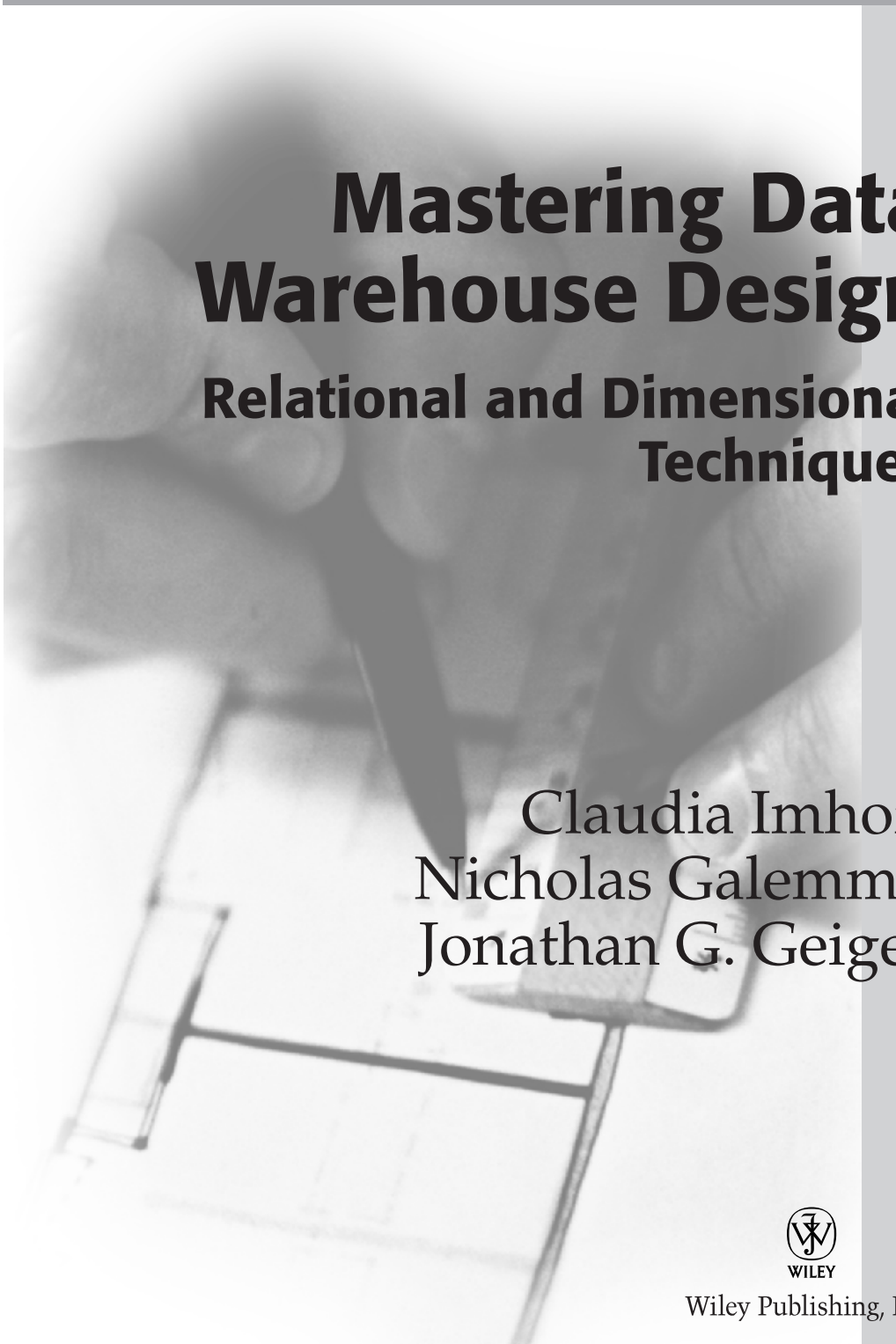


Mastering Data Warehouse Design

Relational and Dimensional Techniques

Claudia Imhoff
Nicholas Galemmo
Jonathan G. Geiger

WILEY
Wiley Publishing, Inc.



Mastering Data Warehouse Design

Relational and Dimensional Techniques

Claudia Imhoff
Nicholas Galemmo
Jonathan G. Geiger

WILEY
Wiley Publishing, Inc.

Vice President and Executive Publisher: Robert Ipsen
Publisher: Joe Wikert
Executive Editor: Robert M. Elliott
Developmental Editor: Emilie Herman
Editorial Manager: Kathryn Malm
Managing Editor: Pamela M. Hanley
Text Design & Composition: Wiley Composition Services

This book is printed on acid-free paper. ♾

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 0-471-32421-3

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

DEDICATION

Claudia: For all their patience and understanding throughout the years, this book is dedicated to David and Jessica Imhoff.

Nick: To my wife Sarah, and children Amanda and Nick Galemmo, for their understanding over the many weekends I spent working on this book. Also to my college professor, Julius Archibald at the State University of New York at Plattsburgh for instilling in me the science and art of computing.

Jonathan: To my wife, Alma Joy, for her patience and understanding of the time spent writing this book, and to my children, Avi and Shana, who are embarking on their respective careers and of whom I am extremely proud.

ACKNOWLEDGMENTS

We gratefully acknowledge the following individuals who directly or indirectly contributed to this book:

Greg Backhus – Helzberg Diamonds

William Baker – Microsoft Corporation

John Crawford – Merrill Lynch

David Gleason – Intelligent Solutions, Inc.

William H. Inmon – Inmon Associates, Inc.

Dr. Ralph S. Kimball- Kimball Associates

Lisa Loftis – Intelligent Solutions, Inc.

Bob Lokken – ProClarity Corporation

Anthony Marino – L'Oreal Corporation

Joyce Norris-Montanari – Intelligent Solutions, Inc.

Laura Reeves – StarSoft, Inc.

Ron Powell – *DM Review* Magazine

Kim Stannick – Teradata Corporation

Barbara von Halle – Knowledge Partners, Inc.

John Zachman – Zachman International, Inc.

We would also like to thank our editors, Bob Elliott, Pamela Hanley, and Emilie Herman, whose tireless prodding and assistance kept us honest and on schedule.

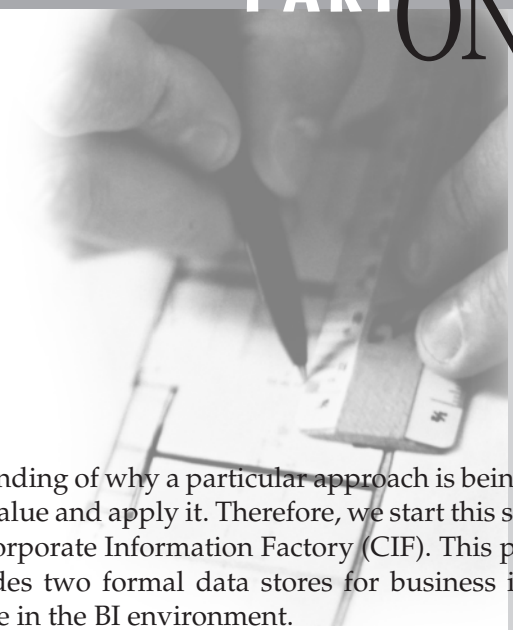
Claudia Imhoff, Ph.D. is the president and founder of Intelligent Solutions (www.IntelSols.com), a leading consultancy on CRM (Customer Relationship Management) and business intelligence technologies and strategies. She is a popular speaker and internationally recognized expert and serves as an advisor to many corporations, universities, and leading technology companies on these topics. She has coauthored five books and over 50 articles on these topics. She can be reached at CImhoff@IntelSols.com.

Nicholas Galemmo was an information architect at Nestlé USA. Nicholas has 27 years' experience as a practitioner and consultant involved in all aspects of application systems design and development within the manufacturing, distribution, education, military, health care, and financial industries. He has been actively involved in large-scale data warehousing and systems integration projects for the past 11 years. He has built numerous data warehouses, using both dimensional and relational architectures. He has published many articles and has presented at national conferences. This is his first book. Mr. Galemmo is now an independent consultant and can be reached at ngalemmo@yahoo.com.

Jonathan G. Geiger is executive vice president at Intelligent Solutions, Inc. Jonathan has been involved in many Corporate Information Factory and customer relationship management projects within the utility, telecommunications, manufacturing, education, chemical, financial, and retail industries. In his 30 years as a practitioner and consultant, Jonathan has managed or performed work in virtually every aspect of information management. He has authored or coauthored over 30 articles and two other books, presents frequently at national and international conferences, and teaches several public seminars. Mr. Geiger can be reached at JGeiger@IntelSols.com.

PART ONE

Concepts



We have found that an understanding of why a particular approach is being promoted helps us recognize its value and apply it. Therefore, we start this section with an introduction to the Corporate Information Factory (CIF). This proven and stable architecture includes two formal data stores for business intelligence, each with a specific role in the BI environment.

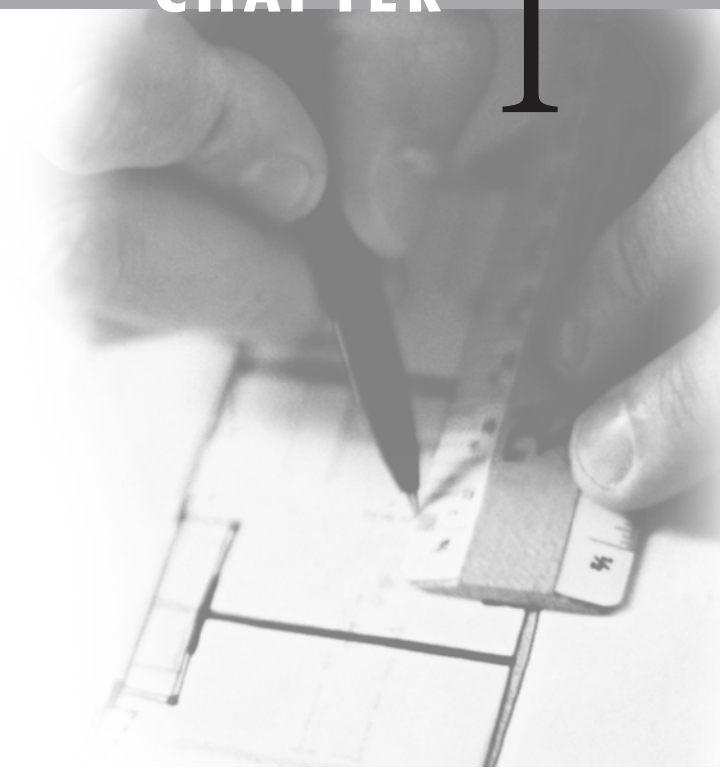
The first data store is the data warehouse. The major role of the data warehouse is to serve as a data repository that stores data from disparate sources, making it accessible to another set of data stores – the data marts. As the collection point, the most effective design approach for the data warehouse is based on an entity-relationship data model and the normalization techniques developed by Codd and Date in their seminal work throughout the 1970's, 80's and 90's for relational databases.

The major role of the data mart is to provide the business users with easy access to quality, integrated information. There are several types of data marts, and these are also described in Chapter 1. The most popular data mart is built to support online analytical processing, and the most effective design approach for it is the dimensional data model.

Continuing with the conceptual theme, we explain the importance of relational modeling techniques, introduce the different types of models that are needed, and provide a process for building a relational data model in Chapter 2. We also explain the relationship between the various data models used in constructing a solid foundation for any enterprise—the business, system, and technology data models—and how they share or inherit characteristics from each other.

CHAPTER 1

Introduction



Welcome to the first book that thoroughly describes the data modeling techniques used in constructing a multipurpose, stable, and sustainable data warehouse used to support business intelligence (BI). This chapter introduces the data warehouse by describing the objectives of BI and the data warehouse and by explaining how these fit into the overall Corporate Information Factory (CIF) architecture. It discusses the iterative nature of the data warehouse construction and demonstrates the importance of the data warehouse data model and the justification for the type of data model format suggested in this book. We discuss why the format of the model should be based on relational design techniques, illustrating the need to maximize nonredundancy, stability, and maintainability. Another section of the chapter outlines the characteristics of a maintainable data warehouse environment. The chapter ends with a discussion of the impact of this modeling approach on the ultimate delivery of the data marts. This chapter sets up the reader to understand the rationale behind the ensuing chapters, which describe in detail how to create the data warehouse data model.

Overview of Business Intelligence

BI, in the context of the data warehouse, is the ability of an enterprise to study past behaviors and actions in order to understand where the organization has

been, determine its current situation, and predict or change what will happen in the future. BI has been maturing for more than 20 years. Let's briefly go over the past decade of this fascinating and innovative history.

You're probably familiar with the technology adoption curve. The first companies to adopt the new technology are called innovators. The next category is known as the early adopters, then there are members of the early majority, members of the late majority, and finally the laggards. The curve is a traditional bell curve, with exponential growth in the beginning and a slowdown in market growth occurring during the late majority period. When new technology is introduced, it is usually hard to get, expensive, and imperfect. Over time, its availability, cost, and features improve to the point where just about anyone can benefit from ownership. Cell phones are a good example of this. Once, only the innovators (doctors and lawyers?) carried them. The phones were big, heavy, and expensive. The service was spotty at best, and you got "dropped" a lot. Now, there are deals where you can obtain a cell phone for about $60, the service providers throw in $25 of airtime, and there are no monthly fees, and service is quite reliable.

Data warehousing is another good example of the adoption curve. In fact, if you haven't started your first data warehouse project, there has never been a better time. Executives today expect, and often get, most of the good, timely information they need to make informed decisions to lead their companies into the next decade. But this wasn't always the case.

Just a decade ago, these same executives sanctioned the development of executive information systems (EIS) to meet their needs. The concept behind EIS initiatives was sound—to provide executives with easily accessible key performance information in a timely manner. However, many of these systems fell short of their objectives, largely because the underlying architecture could not respond fast enough to the enterprise's changing environment. Another significant shortcoming of the early EIS days was the enormous effort required to provide the executives with the data they desired. Data acquisition or the extract, transform, and load (ETL) process is a complex set of activities whose sole purpose is to attain the most accurate and integrated data possible and make it accessible to the enterprise through the data warehouse or operational data store (ODS).

The entire process began as a manually intensive set of activities. Hard-coded "data suckers" were the only means of getting data out of the operational systems for access by business analysts. This is similar to the early days of telephony, when operators on skates had to connect your phone with the one you were calling by racing back and forth and manually plugging in the appropriate cords.

Fortunately, we have come a long way from those days, and the data warehouse industry has developed a plethora of tools and technologies to support the data acquisition process. Now, progress has allowed most of this process to be automated, as it has in today's telephony world. Also, similar to telephony advances, this process remains a difficult, if not temperamental and complicated, one. No two companies will ever have the same data acquisition activities or even the same set of problems. Today, most major corporations with significant data warehousing efforts rely heavily on their ETL tools for design, construction, and maintenance of their BI environments.

Another major change during the last decade is the introduction of tools and modeling techniques that bring the phrase "easy to use" to life. The dimensional modeling concepts developed by Dr. Ralph Kimball and others are largely responsible for the widespread use of multidimensional data marts to support online analytical processing.

In addition to multidimensional analyses, other sophisticated technologies have evolved to support data mining, statistical analysis, and exploration needs. Now mature BI environments require much more than star schemas—flat files, statistical subsets of unbiased data, normalized data structures, in addition to star schemas, are all significant data requirements that must be supported by your data warehouse.

Of course, we shouldn't underestimate the impact of the Internet on data warehousing. The Internet helped remove the mystique of the computer. Executives use the Internet in their daily lives and are no longer wary of touching the keyboard. The end-user tool vendors recognized the impact of the Internet, and most of them seized upon that realization: to design their interface such that it replicated some of the look-and-feel features of the popular Internet browsers and search engines. The sophistication—and simplicity—of these tools has led to a widespread use of BI by business analysts and executives.

Another important event taking place in the last few years is the transformation from technology chasing the business to the business demanding technology. In the early days of BI, the information technology (IT) group recognized its value and tried to sell its merits to the business community. In some unfortunate cases, the IT folks set out to build a data warehouse with the hope that the business community would use it. Today, the value of a sophisticated decision support environment is widely recognized throughout the business. As an example, an effective customer relationship management program could not exist without strategic (data warehouse with associated marts) and a tactical (operational data store and oper mart) decision-making capabilities. (See Figure 1.1)

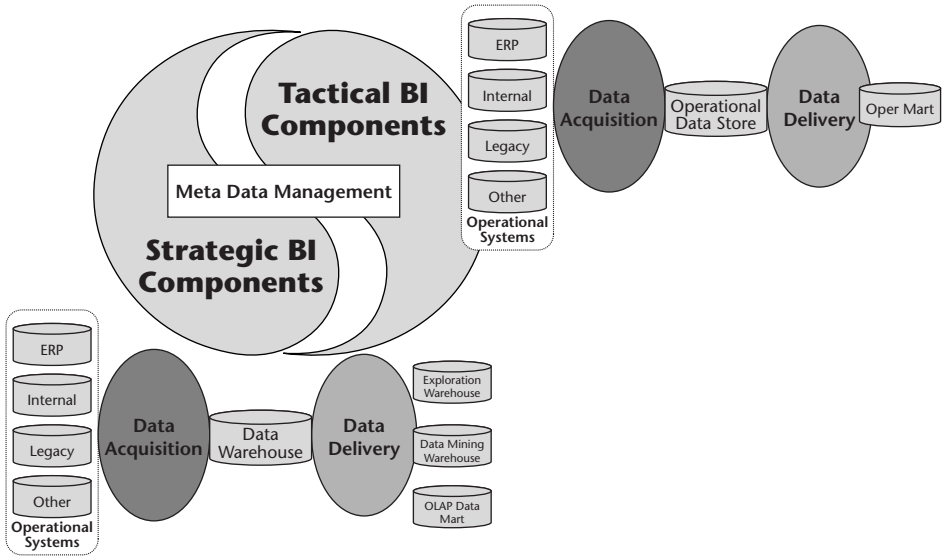


Figure 1.1 Strategic and tactical portions of a BI environment.

BI Architecture

One of the most significant developments during the last 10 years has been the introduction of a widely accepted architecture to support all BI technological demands. This architecture recognized that the EIS approach had several major flaws, the most significant of which was that the EIS data structures were often fed directly from source systems, resulting in a very complex data acquisition environment that required significant human and computer resources to maintain. The Corporate Information Factory (CIF) (see Figure 1.2), the architecture used in most decision support environments today, addressed that deficiency by segregating data into five major databases (operational systems, data warehouse, operational data store, data marts, and oper marts) and incorporating processes to effectively and efficiently move data from the source systems to the business users.

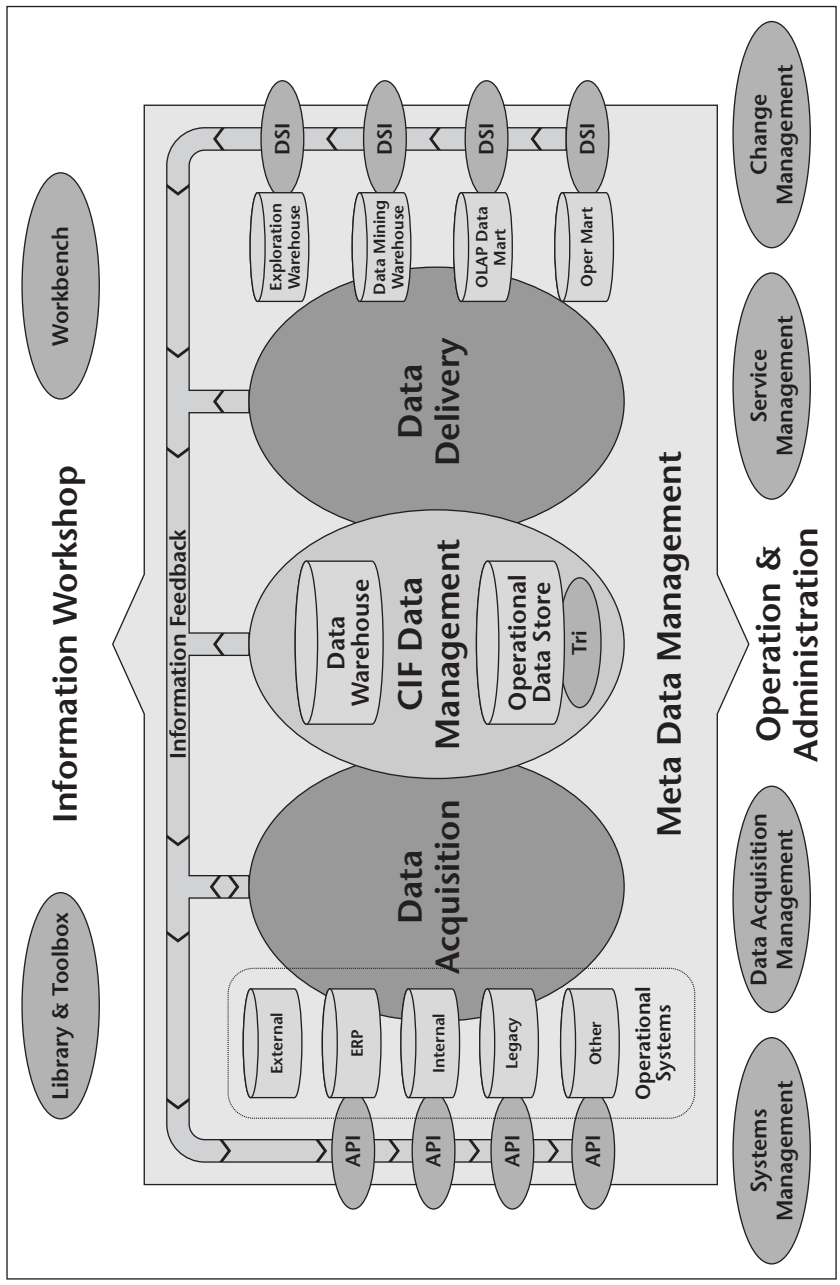


Figure 1.2 The Corporate Information Factory.

These components were further separated into two major groupings of components and processes:

- *Getting data in* consists of the processes and databases involved in acquiring data from the operational systems, integrating it, cleaning it up, and putting it into a database for easy usage. The components of the CIF that are found in this function:
  - The operational system databases (source systems) contain the data used to run the day-to-day business of the company. These are still the major source of data for the decision support environment.
  - The data warehouse is a collection or repository of integrated, detailed, historical data to support strategic decision-making.
  - The operational data store is a collection of integrated, detailed, current data to support tactical decision making.
  - Data acquisition is a set of processes and programs that extracts data for the data warehouse and operational data store from the operational systems. The data acquisition programs perform the cleansing as well as the integration of the data and transformation into an enterprise format. This enterprise format reflects an integrated set of enterprise business rules that usually causes the data acquisition layer to be the most complex component in the CIF. In addition to programs that transform and clean up data, the data acquisition layer also includes audit and control processes and programs to ensure the integrity of the data as it enters the data warehouse or operational data store.
- *Getting information out* consists of the processes and databases involved in delivering BI to the ultimate business consumer or analyst. The components of the CIF that are found in this function:
  - The data marts are derivatives from the data warehouse used to provide the business community with access to various types of strategic analysis.
  - The oper marts are derivatives of the ODS used to provide the business community with dimensional access to current operational data.
  - Data delivery is the process that moves data from the data warehouse into data and oper marts. Like the data acquisition layer, it manipulates the data as it moves it. In the case of data delivery, however, the origin is the data warehouse or ODS, which already contains high-quality, integrated data that conforms to the enterprise business rules.

The CIF didn't just happen. In the beginning, it consisted of the data warehouse and sets of lightly summarized and highly summarized data—initially

a collection of the historical data needed to support strategic decisions. Over time, it spawned the operational data store with a focus on the tactical decision support requirements as well. The lightly and highly summarized sets of data evolved into what we now know are data marts.

Let's look at the CIF in action. Customer Relationship Management (CRM) is a highly popular initiative that needs the components for tactical information (operational systems, operational data store, and oper marts) and for strategic information (data warehouse and various types of data marts). Certainly this technology is necessary for CRM, but CRM requires more than just the technology—it also requires alignment of the business strategy, corporate culture and organization, and customer information in addition to technology to provide long-term value to both the customer and the organization. An architecture such as that provided by the CIF fits very well within the CRM environment, and each component has a specific design and function within this architecture. We describe each component in more detail later in this chapter.

CRM is a popular application of the data warehouse and operational data store but there are many other applications. For example, the enterprise resource planning (ERP) vendors such as SAP, Oracle, and PeopleSoft have embraced data warehousing and augmented their tool suites to provide the needed capabilities. Many software vendors are now offering various plug-ins containing generic analytical applications such as profitability or key performance indicator (KPI) analyses. We will cover the components of the CIF in far greater detail in the following sections of this chapter.

The evolution of data warehousing has been critical in helping companies better serve their customers and improve their profitability. It took a combination of technological changes and a sustainable architecture. The tools for building this environment have certainly come a long way. They are quite sophisticated and offer great benefit in the design, implementation, maintenance, and access to critical corporate data. The CIF architecture capitalizes on these technology and tool innovations. It creates an environment that segregates data into five distinct stores, each of which has a key role in providing the business community with the right information at the right time, in the right place, and in the right form. So, if you're a data warehousing late majority or even a laggard, take heart. It was worth the wait.

What Is a Data Warehouse?

Before we get started with the actual description of the modeling techniques, we need to make sure that all of us are on the same page in terms of what we mean by a data warehouse, its role and purpose in BI, and the architectural components that support its construction and usage.

Role and Purpose of the Data Warehouse

As we see in the first section of this chapter, the overall BI architecture has evolved considerably over the past decade. From simple reporting and EIS systems to multidimensional analyses to statistical and data mining requirements to exploration capabilities, and now the introduction of customizable analytical applications, these technologies are part of a robust and mature BI environment. See Figure 1.3 for the general timeframe for each of these technological advances.

Given these important but significantly different technologies and data format requirements, it should be obvious that a repository of quality, trusted data in a flexible, reusable format must be the starting point to support and maintain any BI environment. The data warehouse has been a part of the BI architecture from the very beginning. Different methodologies and data warehouse gurus have given this component various names such as:

A staging area. A variation on the data warehouse is the "back office" staging area where data from the operational systems is first brought together. It is an informally designed and maintained grouping of data whose only purpose is to feed multidimensional data marts.

The information warehouse. This was an early name for the data warehouse used by IBM and other vendors. It was not as clearly defined as the staging area and, in many cases, encompassed not only the repository of historical data but also the various data marts in its definition.

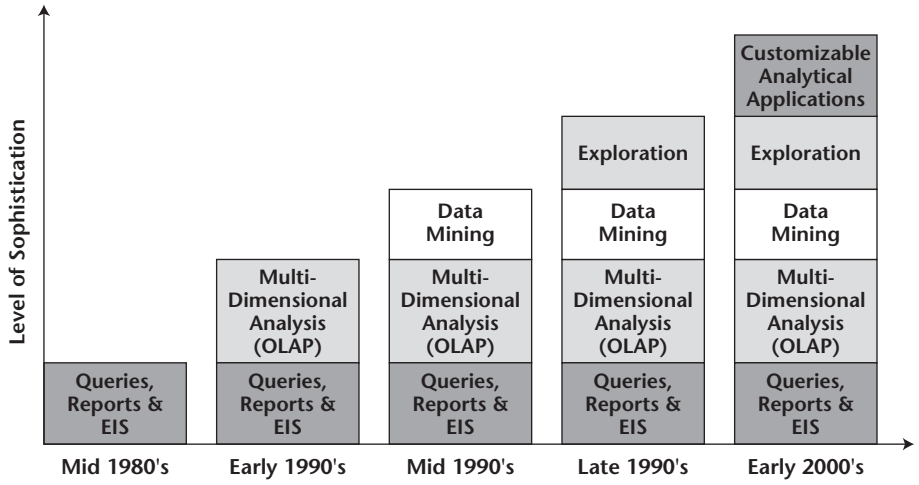


Figure 1.3 Evolving BI technologies.