Corporate Information Factory

Second Edition

W. H. Inmon

Claudia Imhoff

Ryan Sousa

Wiley Computer Publishing

John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

Publisher: Robert Ipsen
Editor: Robert M. Elliott
Assistant Editor: Emilie Herman
Managing Editor: John Atkins
Text Design & Composition: D&G Limited, 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: 0-471-39961-2

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

Some books describe *how* to do things; they're called how-to books or manuals. Other books describe *why* we do things; these include books on philosophy and psychology. Still other books describe *where* things are—for example, atlases. But other books simply describe *what* we should be doing. This is one of those kinds of books.

The usefulness of a *what* book is that it provides direction. There is an old saying among sailors, "When there is no destination, any route will do." This book describes a very substantial destination port for sailors navigating the sea of information. After the corporation understands that there is a port and where it is, it is easy to set the information organization's rudder on the right heading to the appropriate destination, even through stormy seas.

The Origins of Data Warehousing

The world of computers and information technology has grown quickly, sequentially, and in a surprisingly uniform manner. In the era of punch-card and paper-tape systems, we used the computer as a calculating beast of burden, running such systems as accounts payable and accounts receivable.

With the advent of disk storage, cheaper memory, more sophisticated operating systems, and direct end-user interface devices, a whole new style of computing became a reality—online processing. With online processing, the computer changed from a beast of burden to an Arabian stallion. Reservation systems, automated bank tellers, and a host of other new systems became a reality.

Next came the revolution of the end user. Personal computers, spreadsheet applications, and fourth-generation language (4GL) technology opened up computing to an audience that previously had been denied. As the costs of computing plummeted, a Pandora's box was opened; computing now fell outside the classical domain of the information systems organization. Anyone with a budget could begin to take charge of their information destiny.

End users were very happy having such complete control, but this autonomy was a mirage. It soon became apparent that even with unlimited computing power, turning over control to end users created new issues, such as a lack of integration and proper economies of scale. For all of the appeal of autonomous control of processing at the end-user level, the case for centralization was equally valid and appealing.

Simultaneously, people discovered that data derived from processing operational transactions was difficult to access and insufficient for effective decision-making. Historical and integrated data was needed at both a summary and detailed level.

Thus, the data warehouse was born. Shortly thereafter, data volumes and end-user demands and diversity exceeded the pace at which the data warehouse could be tailored and tuned. For all its strength in integrating and managing a common view of corporate data, the data warehouse was not keeping up with the business demands for information. In response, different departments found that a customized subset of the data warehouse—something called a data mart—provided them the needed autonomy to drive the interpretation and use of corporate information.

Companies also discovered that a need for operational data integration existed. The data warehouse—for all it provided—did nothing for the people who needed operational integration. Into the fray came an architectural entity known as the *operational data store*. Finally, with the advent of the Internet and low-cost commodity hardware, the data warehouse, data marts, and operational data stores are scaling to answer increasingly complex questions, using more data, in support of a broader user community. It is no longer uncommon for a user to access hundreds of gigabytes of data, integrated from a dozen sources, just to answer a single question. Additionally, it is becoming increasingly common for the user community to consist of groups outside the company such as business partners or even customers—the ultimate benefactors.

Seeing the Forest for the Trees

Those of us who have been witnesses to some or all of these developments have suffered from two disadvantages in understanding what has transpired. First, we have been too close to the technology and the maturation of the technology to truly grasp its significance. We have marveled at the details without understanding the larger form and function. As such, we, like the six blind men describing the elephant, simply have a limited understanding of that with which we are the most intimate.

The second disadvantage is that of watching this development unfold sequentially day by day. The speed of day-by-day developments has blinded us to the larger picture that is unfolding, and we can only guess what tomorrow will bring.

The objective of this book is to overcome these obstacles by taking a step back and examining the evolution of business information systems across the globe. In the pages that follow, we strive to make sense of this evolution and describe a proven architecture that embraces it. Additionally, we describe the key components of this architecture and how they fit together. This architecture defines the corporate information factory (CIF).

We are particularly interested in what happens when a company attempts to build its systems in a fashion other than what has been suggested by the architecture. The brief history of the corporate information factory has shown that:

- You certainly can build an architecture other than the one described here
- When you do build a variation of the described architecture, there is a price to pay, in terms of:
  - Infrastructure cost
  - Efficient performance
  - Lack of integration
  - Seamlessness of technology
  - End-user satisfaction
  - Responsiveness to change

The corporate information factory is hardly the only way to build systems. But, it is the best way to meet the long-term goals of the information processing company.

In many cases, it will be tempting to violate the architecture; however, systems designers must remember there is a price to pay.

Our purpose in writing this book is to alert readers of a proven way to organize information systems; that when this way is not chosen, they must be willing—and able—to live with the consequences. These consequences can range from the waste of large amounts of development resources to the failure to deliver an effective information resource.

A number of factors led to the evolution of the corporate information factory including:

- Evolving business demands
- Shrinking costs of technology

- Increasing sophistication and breath of the user community
- Growth of hardware, software, and network capabilities

In addition, corporate communities began to move toward very distinct styles of computing. For example, operational, legacy processing set the stage for the data warehouse, which then led to the data mart. Spreadsheets and their many derivatives opened up the desktop to many more analytical capabilities. These various benchmarks led to the evolution of a new mode of corporate computing, which we describe in the chapters to follow.

Why We Wrote the Second Edition

The first edition of this book addressed the larger picture of corporate information systems as they were evolving in the world of information technology. At the time of the first edition, the description found in the book was accurate, to the best of our knowledge. But almost immediately after the book came out, we began to notice that other important components of modern information systems architecture were left out.

The exploration warehouse, alternative storage, and decision-support systems (DSS) applications appeared as major components of information architecture. Many large, vertical multinational corporations began to build multiple data warehouses, and coping with more than one corporate information factory became an issue. In the ever-changing world of technology, the Internet, customer relationship management (CRM), and enterprise resource planning (ERP) applications made an appearance in a big way. Because of all of these changes to the landscape of information systems, it became necessary to publish a second edition of the book, that is at the center of the corporate information factory.

The second edition is predicated on the architecture described in the centerfold of the January 2000, issue of *Data Management Review* magazine and shown in Figure 2.1. That description is the current thinking on what constitutes the corporate information factory. If you want a copy of the centerfold, you can contact *DMR* magazine, or you can download it (for free) from the Web site—www.billinmon.com. Many thanks to Ron Powell and Jean Schauer for their sponsorship in the creation of the centerfold and to IBI Corp. for their financial sponsorship in the creation of the centerfold.

Who Should Read This Book

The corporate information factory can be used in many useful ways by a wide variety of people, such as:

- **The IT manager.** The information technology (IT) manager can use the corporate information factory to predict what the next steps ought to be for systems development and architecture. Instead of spending money unproductively on projects that do not move the organization to the paradigm suggested by the corporate information factory, the manager who understands its implications can use the corporate information factory as a benchmark that tells what the future directions ought to be.
- **The developer.** After a project has begun, the developer can determine whether the project is organized in concert with the corporate information factory. If a design is contrary to the corporate information factory, the designer can make corrections before the design is cast in concrete.
- **The investor.** An easy way to determine how fruitful a technology investment will be is to gauge it against the world described by the corporate information factory. If the architecture of the investment is not aligned with the corporate information factory, then the investor can be alerted to problems with marketplace acceptance.
- **The end user.** At the heart of the corporate information factory is the success of end users who can use it to form their expectations and to assess whether their expectations are out of line. When implemented properly, the corporate information factory makes life very easy and productive for the end user.

How This Book Is Organized

This book is organized to suit the needs of a wide range of readers from novice to experienced in implementing the corporate information factory. If you are new to the corporate information factory, you will want to read this book from beginning to end. Each chapter builds on the previous chapter, providing you a broad understanding of what the corporate information factory is and how to build and manage it. If you are a veteran, you will probably want to read the first two chapters and dive right into whatever chapter suits your needs. In particular, you will probably be interested in Chapter 9, 10, 13, and 17. These chapters introduce some new components and concepts.

The book is divided into four parts. The first part—Chapters 1 and 2—are introductory. They provide you with an overview of the corporate information factory and the drivers in its evolution. The second part—Chapters 3 through 14—reviews the corporate information factory architecture. These chapters review each component of the architecture, how they are combined to deliver decision support capabilities and the implications of varying the architecture. The third part—Chapters 15 through 17— discusses how to build and manage the corporate information factory. The fourth part—Appendix A—provides guidelines for assessing and examining your corporation information factory.

The Evolution of This Book

This book is part of a larger series of books by Bill Inmon. In the first of the books, *Data Architecture: The Information Paradigm*, the notion of a larger architecture was first introduced, and the data warehouse was first mentioned. The next book in the series, *Building the Data Warehouse*, fully explored the data warehouse. The book is now enjoying sales in the second edition. Next came *Using the Data Warehouse*, in which the techniques and considerations of the effective use of the data warehouse were discussed, and the operational data store was introduced. At about the same time, *Building the Operational Data Store* appeared (in the second edition as published in 1999). This book probed the design and technological implications of the ODS. The next book in the series was *Managing the Data Warehouse.* In this book, the assumption is that the data warehouse has already been built and that the issues of cost of data warehousing and complexity of data warehousing are starting to crop up. As people began building, using, and managing their data warehouse environment, they also began asking for more specifics on how to optimize and exploit it. In response, *Data Warehouse Performance* and *Exploration Warehousing* were published.

The second edition of *Corporate Information Factory* in many ways is a capstone book. It brings together the many aspects of the architected information systems environment—the *information ecosystem*—and presents those aspects in an integrated manner.

DEDICATION

It wasn't enough that we asked our families to endure the many days we were away from home working and the long hours we spent at home writing the first edition; we asked if we could do it again. Well, you know what, they said yes. Not only did they say yes, they encouraged our efforts knowing we had much more to say about the corporate information factory. They remain the wind beneath our wings and the joy in our hearts.

ACKNOWLEDGMENTS

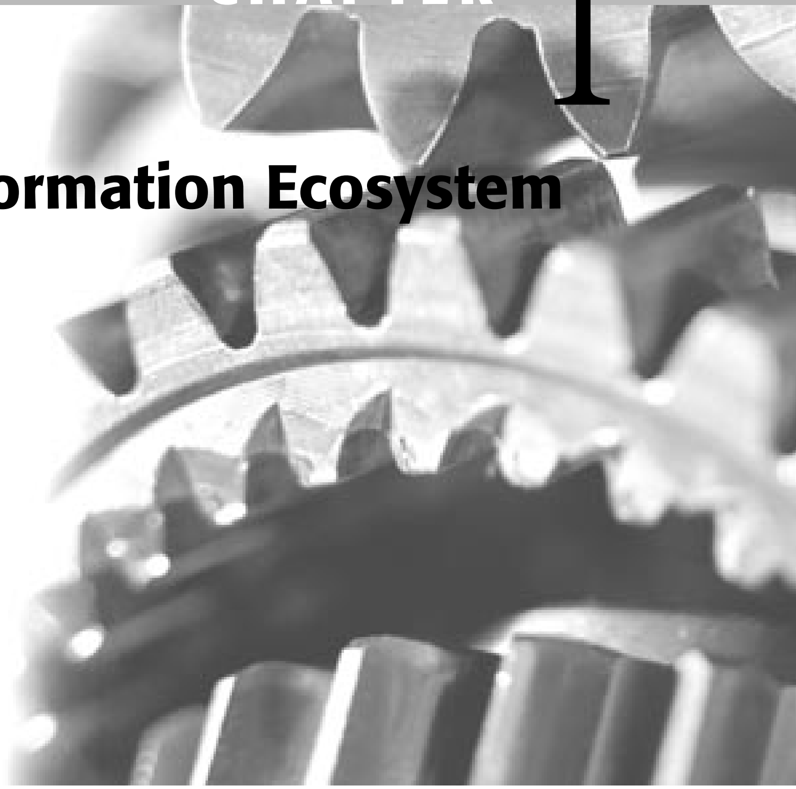
We wish to express thanks to the many colleagues, clients, and friends who have enriched our understanding of the corporate information factory over the years. It is through their collective efforts that ideas become reality, and it is from this reality that we learn and write. We would like to extend a special thanks to:

John Zachman, Zachman International
John Bair, Independent consultant
Lowell Fryman, C/Net
Roger Geiwitz, Independent consultant
Sue Osterfelt, Bank of America
JD Welch, IBM
Dennis McCann, ambeo
Ken Richardson, ambeo
Dale Brocklehurst, ambeo
Joyce Norris-Montanari, Braun Consulting
Jon Geiger, Braun Consulting
Steve Miller, Braun Consulting

Jim Kalustian, Braun Consulting
Mike Evanisko, Braun Consulting
Dave Imhoff, Intelligent Solutions
Rob Geller, Quest
Robert Grim, Independent consultant
Pete Simcox, Genesis
Mark Mays, Arrowhead Consulting
warehouseMCI team
John Ladley, Knowledge Interspace
Doug Laney, Meta Group
Bob Lokken, Knosys
Brian Burnett, AppsCo
Steve Murchie, Microsoft
Bill Baker, Microsoft
Allen Perry, Coglin Mill
Ron Powell, DM Review
Bill Prentice, SAS
Mike Wipperfeld, Informix
Lisa Loftis, Braun Consulting
Steve Hill, Informix
Keven Gould, Sybase
Stephen Gardner, NCR
Ron Swift, NCR
Marc Demarest, Independent consultant
Jeanne Friedman, Independent consultant
Greg Battas, Tandem Computers
Ralph Kimball, Kimball and Associates

CHAPTER 1

Creating an Information Ecosystem



Business is quickly reshaping itself to compete in a global economy governed by the needs of the customer (e.g., individual, business, etc.). The economies gained over the past three decades by automating manual business processes are no longer enough to gain a competitive advantage in today's marketplace. To compete, businesses need to be able to build a new set of capabilities that deliver *best-of-breed* business intelligence and business management solutions that can leverage this legacy environment.

But wait! Perhaps the genesis is already upon us. Your IT department is being bombarded with a growing number of targeted information architectures, technologies, methodologies, terms, and acronyms. Each of these advances promises to deliver competitiveness in one easy step, such as:

- Data warehousing
- Data repository
- Operational data store
- Data marts
- Data mining
- Internet and intranet
- Multidimensional and relational databases

- Exploration processing
- Star schema, snowflake, and relational database design techniques
- High-performance computing (Massively Parallel Processing & Symmetrical Multiprocessing)
- Data acquisition and data delivery
- Online Analytical Processing (OLAP)
- Data warehouse administration
- Metadata management

Each of these advances in modern information technology has promise, but trying to make sense of these point solutions while still getting the job done in a short time frame can be confusing and intimidating. This is largely due to the fact that no model exists that combines these elements of the information primordial pool into a balanced ecosystem that aligns with the evolving needs of the business. An information ecosystem is needed to orchestrate the use of various information technologies and constructs and to foster communication and the cooperative exchange of work, data processing, and knowledge as part of a symbiotic relationship.

Information Ecosystem Briefly Defined

An information ecosystem is a system with different components, each serving a community directly while working in concert with other components to produce a cohesive, balanced information environment. Like nature's ecosystem, an information ecosystem must be adaptable, changing as the inhabitants and participants within its aegis change. Over time, the balance between different components and their relationship to each other changes as well, as the environment changes. Sometimes the effect will appear on seemingly unrelated parts (sometimes disastrously!). Adaptability, change, and balance are the hallmarks of the components of a healthy information ecosystem.

As an example of an information ecosystem, consider a data warehouse working with a data mart to deliver business intelligence capabilities or an operational data store working to deliver business management capabilities. This environment is found in many marketing groups. At first, there is the need for better business intelligence in the form of market segmentation, customer analysis and contact analysis. Then, at some point, marketing wants to take action on the "intelligence" gained. Although the data warehouse and data mart are well suited to support business intelligence, they lack the content and form to drive business-management activities associated with contacting the customer. What is needed is an operational data store to provide near *real-time* access to integrated, current customer information.

As will be discussed in this book, different business needs require that a different set of ecosystem components work in tandem. Ultimately, the information ecosystem will be business-driven, as capabilities delivered (business intelligence and business management) are aligned with the needs of the business (marketing, customer service, product management, etc.). The result is an information environment that allows companies to capitalize on a constantly changing business landscape characterized by customer relationships and customized product delivery.

Shifting Business Landscape

Three fundamental business pressures are fueling the evolution of the information ecosystem: growing consumer demand, increased competition and complexity, and continued demands for improvements in operating efficiencies as seen in Figure 1.1.

Consumer Demand

The first fundamental pressure is growing consumer demand. Consumers expect companies to understand and respect their needs and desires. Because the customer drives the business relationship, business people must hear what the customer has to say and respond by delivering relevant, competitive, and

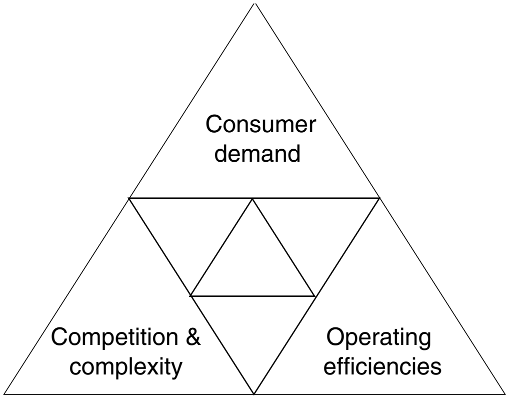


Figure 1.1 The business drivers in today's world.

timely products and services. Companies can no longer expect to sell just a few general products and services to the masses but must tailor many products and services (i.e., mass customize) to the individual consumer. This proposition is called customer relationship management and/or mass customization. The fundamental challenge to many businesses is that their systems, people, and processes are designed around the product. Furthermore, most of these companies have begun to extend these environments with a series of unarchitected point solutions to address their immediate needs for customer management. A healthy information ecosystem will be embodied by an architecture that:

- Leverages this legacy environment
- Delivers new information capabilities that allow companies to thrive in an environment characterized by customer relationships and customized product delivery
- Supports a migration strategy that is evolutionary in nature and delivers incremental value to the business

Competition and Complexity

The second business pressure is that of increased competition and complexity. The ability to refocus and enhance a product mix in response to evolving competition is a critical success factor for any business. The key is to be able to anticipate the needs of the marketplace before your competitors do. Many companies find this difficult or impossible to do, given today's mishmash of technologies, architectures, and systems.

Why is this important? Corporations today are facing more and more deregulation, mergers, and acquisitions, which blur the relationships with their customers. Additionally, globalization of the marketplace and the consumer is opening up businesses to new avenues for expansion and, subsequently, competition. Therefore, it is mandatory for a corporation to quickly restructure itself without losing the ability to compete.

Operating Efficiencies

The third pressure is that of continued improvements in operating efficiencies. The ability to rapidly measure and predict returns on investment is something that corporations find difficult to perform. These measurements indicate the health of the corporation, and the ability to determine them rapidly allows a corporation to change its direction with a minimal loss in time or money. Other examples of improved efficiency include the ability to determine the most efficient channels for contacting customers, to target the best product mix to the best customers, and to identify new product opportunities before the competition does.

Responding to Change

In response to these very real business challenges, companies must be able to support more than just classical business operations (legacy systems that automate manual business processes such as billing, order processing, etc.). Competitive corporations need capabilities to support business intelligence and business management. In this way, they can respond to the dynamics of a quickly changing business landscape, as seen in Figure 1.2.

The information ecosystem provides a context for understanding the needs of your business and taking actions based on those needs while still running the day-to-day business. Additionally, the information ecosystem provides businesses with a comprehensive model for leveraging the growing number of distinctive information constructs and technologies that are required to deliver diverse and pressing business capabilities to support these needs. Figure 1.3 illustrates the central role of the corporate information factory in supporting

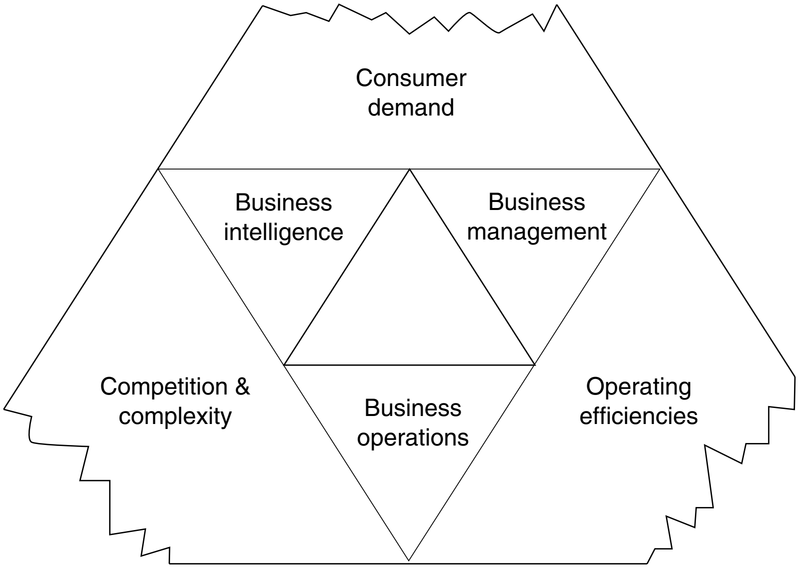


Figure 1.2 The need for business capabilities to compete in a quickly changing business landscape.

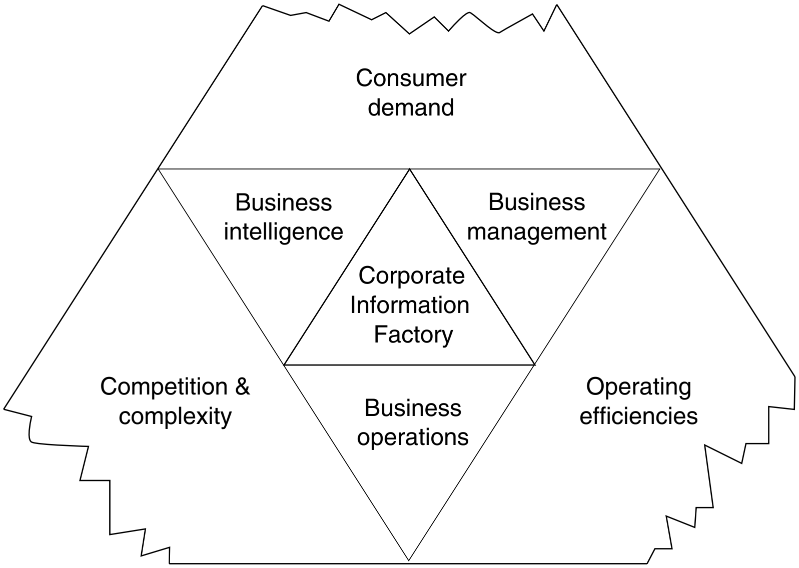


Figure 1.3 The corporate information factory is central to a business and its needed capabilities.

the evolving areas of business proficiency: business operations, business intelligence, and business management.

Business operations are supported by capabilities used to run the day-to-day business. These systems have traditionally made up our legacy environment and have provided a competitive advantage by automating manual business processes to gain economies of scale and speed-to-market. Systems that exemplify business operations include accounts payable, accounts receivable, billing, order processing, compensation, and fulfillment/distribution.

Business intelligence is supported by capabilities that help companies understand what makes the wheels of the corporation turn and help predict the future impact on current decisions. These systems play a key role in the strategic planning process of the corporation. Systems that exemplify business intelligence include medical research, market analysis, customer contact analysis, segmentation, scoring, profitability forecasting, and inventory forecasting.

Business management is supported by capabilities that are needed to effectively manage actions resulting from the business intelligence gained. If business intelligence helps companies understand *what* makes the wheels of the corporation turn, business management helps *direct* the wheels as the business landscape changes. These systems are characterized by robust real-time reporting and tight integration with business intelligence and business operations systems. Systems that exemplify business management include fulfillment management, channel management, inventory management, resource management, and customer information management. These systems generally augment and/or evolve from business operations.

In summary, the information ecosystem provides companies with a complete information solution by complementing traditional business operations with capabilities to deliver business intelligence and business management. In addition, the information ecosystem provides a comprehensive model for making sense and exploiting the growing and diverse information constructs and technologies that are transforming our information paradigm. The physical embodiment of the information ecosystem is the corporate information factory.

Corporate Information Factory

First introduced by W. H. Inmon in the early 1980s, the corporate information factory (CIF) is the physical embodiment of the notion of an information ecosystem. The CIF is at the same time generic in its structure (to the point that it is easily recognizable across different corporations) and is unique to each company as it is shaped by business, culture, politics, economics, and technology. The corporate information factory is made up of the following components:

External world. It is the businesses and people who generate the transactions that fuel the CIF and who produce and benefit from the information produced.

Applications. Applications are the family of systems from which the corporate information factory gathers raw detail data. There are two types of applications: integrated and unintegrated. Integrated applications represent those systems that have been developed according to the guidelines set forth by the corporate information factory. Unintegrated applications are traditionally represented by those core operational systems that have been used to drive day-to-day business activities like order processing, accounts payable, etc. Over time, these unintegrated applications will become integrated as their role transcends beyond traditional business operations to support business management.

Operational data store. It is a subject-oriented, integrated, current-valued, volatile collection of detailed data used to support the up-to-the-second collective tactical decision-making process for the enterprise.

Integration and transformation layer. This is where the data gathered by the applications is refined into a corporate structure.

Data warehouse. It is a subject-oriented, integrated, time-variant (temporal), and nonvolatile collection of summary and detailed data used to support the strategic decision-making process for the enterprise.

Data mart(s). It is a customized subset of data from the data warehouse tailored to support the specified analytical requirements of a given business unit.

Internet/intranet. These are the lines of communication along with data flows and different components that interact with each other.

Metadata. It is the information catalog infrastructure to the CIF. This catalog provides the necessary details to promote data legibility, use, and administration.

Exploration and data mining warehouse. This is where the explorer can go to do analysis and does not have to think about the impact on the resources.

Alternative storage. It is where "overflow" and "bulk" storage can be used, extending the warehouse to infinity. The costs of warehousing are greatly mitigated by moving masses of data to alternative storage.

Decision support systems. These systems are a whole body of applications whose center of existence is the data warehouse. These applications are large and distinctive enough that they form their own component of the corporate information factory.

The different components of the CIF create a foundation for information delivery and decision-making activities that can occur anywhere in the CIF. Many of these activities are in the form of decision-support systems (DSS) that provide the end user with easy-to-use, intuitively simple tools to distill information from data.

People and Processes

The people and processes that work within the structure of the information ecosystem represent the roles, workflow, methods, and tools used in constructing, managing, and using the corporate information factory. Activities that occur here include:

- Customer communications (newsletters, surveys, etc.)
- Request management (logging, prioritizing, and managing)
- Delivery of information (data mart enhancements, corrections)
- Configuration management (versioning of metadata, database design, extraction, programs, transformation programs, etc.)

- Data quality management (performing audits, integrity checks, alerts)
- Systems administration (determining capacity, conducting performance tuning, etc.)

The people and processes of the CIF are perhaps one of the more difficult issues for a corporation because, in planning this function, the corporation must take into consideration its culture, politics, economics, geography, change, and other concerns. For example, companies that have traditionally managed their information systems from a central organization may have challenges supporting data marts that are owned and managed by line-of-business information systems personnel. Alternatively, organizations that have managed information systems at the line-of-business level may have problems giving up the control necessary to form an information systems group to build and manage a corporate data warehouse.

The nature of these variables makes this function a much more customized one for the enterprise and, therefore, harder to implement. There is less uniformity across different corporations in how this aspect of the information ecosystem is implemented than perhaps anywhere else.

Summary

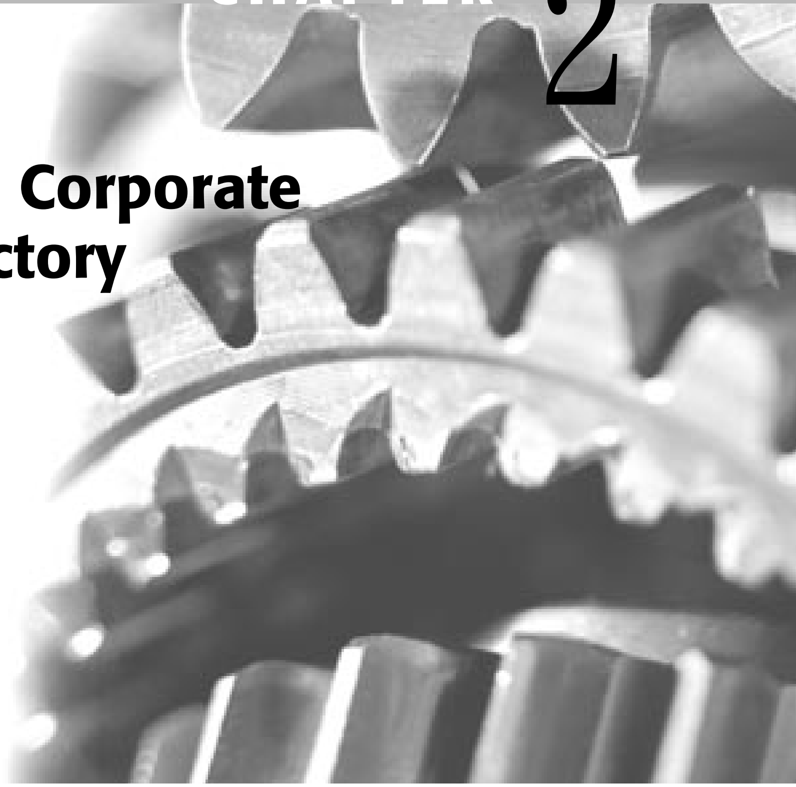
The information ecosystem is a model that supports all of a corporation's information processing. The physical embodiment of the information ecosystem is the corporate information factory. The different components of the corporate information factory have been introduced and defined briefly to give the information systems (IS) architect an idea of how they fit into the overall architecture. Each component must be in balance with the others to avoid a malfunctioning environment, much like nature's ecosystem.

The forces of business coupled with the advances in technology and the symbiotic relationship of technology to the business process cause the world of technology to constantly evolve. In years past when technology was slow and expensive, there was no opportunity for the sophistication that is possible today. But with the decreasing cost of technology, the increasing speed, and new capacities, there are possibilities for the exploitation of technology in the business equation as never before. At the heart of these possibilities is an evolving architecture that has become increasingly apparent, the *corporate information factory*. It has evolved from many systems and technologies now found in the world of corporate information processing.

In the next chapter, we will take a closer look at the corporate information factory, its use, and its evolution.

CHAPTER 2

Introducing the Corporate Information Factory



As discussed in Chapter 1, the corporate information factory (CIF) is an architecture—an infrastructure—for the information ecosystem, consisting of the following components:

- External world
- Applications
- Integration and transformation layer (I & T layer)
- Operational data store (ODS)
- Data warehouse
- Data mart(s)
- Internet and intranet
- Metadata repository
- Exploration and data mining data warehouse
- Alternative storage
- Decision Support Systems (DSS)

The simplest way to understand the CIF is in terms of the data that flows in and the information that flows out of the corporate information factory. Data enters the CIF as detailed, raw data collected by the applications. The raw detailed data is refined by the applications and then passes into a layer of programs that fundamentally integrates and transforms functional data into corporate data. The data passes from the integration and transformation layer into the ODS and the data warehouse. The data warehouse can be fed data from either the ODS or the integration and transformation layer. After the data passes through the data warehouse, data is accessed, analyzed, and transformed into information for various purposes.

The architecture and the flow of data that have been described are very similar to that of an actual factory. Raw and assembly goods enter a factory and are immediately collected by inventory and store management processors. Assembly lines then turn the raw goods into a product. Throughout the manufacturing process, different products are made. Some products are completely finished products; others represent a partial assembly that can be further assembled into many finished products.

Data in the Corporate Information Factory

Key components of the corporate information factory are shown in Figure 2.1. Let's begin with external data. External data enters the corporate information factory from the world outside of the corporation. It is not generated internally, nor is it captured and manipulated at a detailed level internally. Instead, external data represents events and objects outside of the corporation in which the corporation is interested. External data can be used throughout the corporate information factory—at the data mart, data warehouse, ODS, and/or application levels.

Reference data is data that is stored in a shorthand fashion that serves to tie together multiple and diverse users. It is used to speed and standardize processing across many different departments and is typically found at the application level. As reference data passes into the architectural components of the corporate information factory, it takes on a slightly different form, that of historical reference data. The difference between reference data and historical reference data is that reference data represents information that is current and accurate as of the moment of usage. Historical reference data is the historical record of that same reference data, except that it is collected and managed over time. As current reference data changes over time, those changes are collected along with the effective change date in order to create historical reference data. Historical reference data is of great use to the data mart and the data warehouse analyst in that it provides details that help describe data in the data warehouse and data marts.

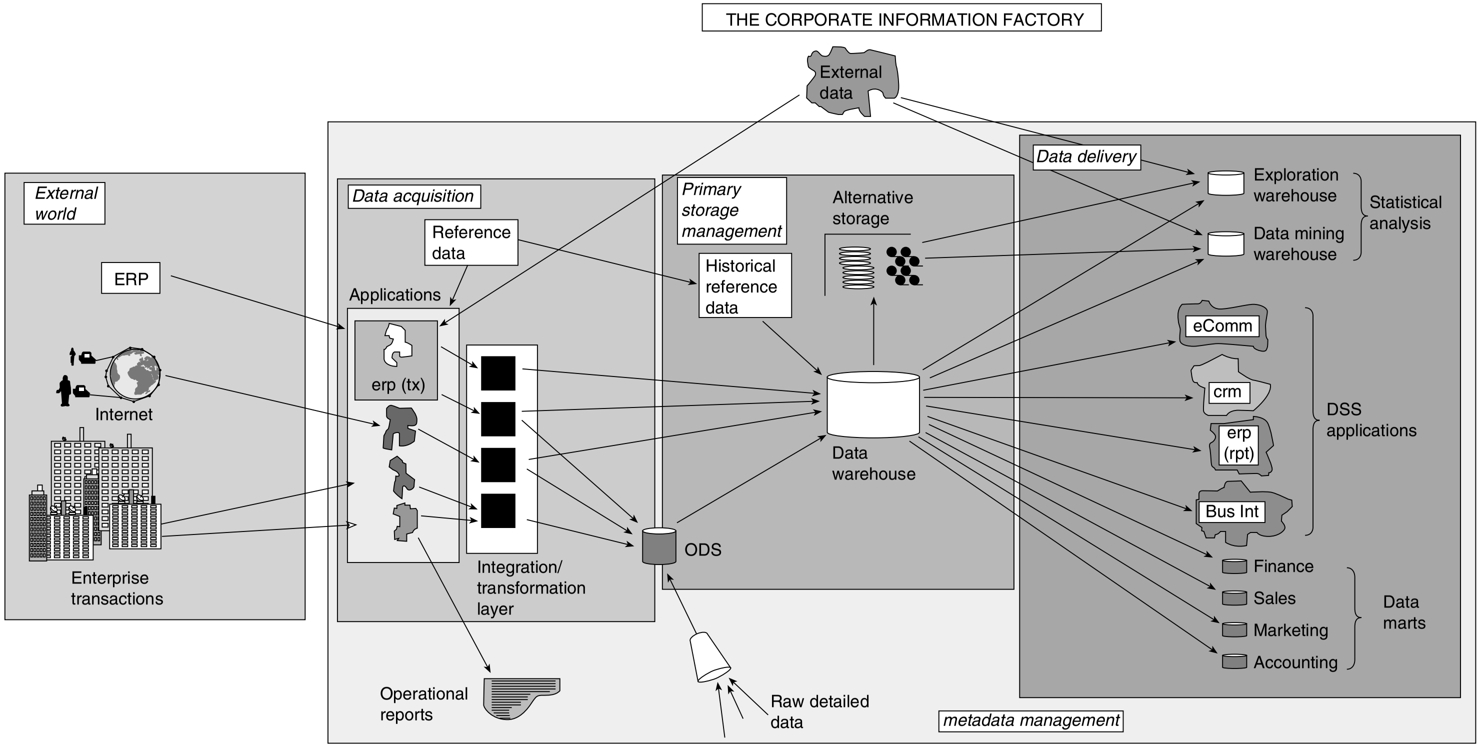


Figure 2.1 The basic structure of the corporate information factory.

A third type of data is raw detailed data. This data is generally captured at the application level and loaded into the data warehouse and ODS via the I & T layer. However, some raw detailed data may be captured and managed directly in the ODS. This happens when the end-user community needs access to data that is not currently being managed by an application. In effect, the ODS becomes the authoritative source of this data and source system to the data warehouse. Some may try to manage this data directly in the data warehouse; however, this is not recommended. This would be like trying to bulldoze a large mound of dirt with a Ferrari. The data warehouse is designed for strategic decision support and lacks the form and function to effectively store and access transaction-level data in real time. Additionally, if the data warehouse became the *authoritative* source of this raw detail data, it is likely that it would quickly become pressured to support operational activities for which it was designed to augment. This is likely to be a terminal condition for the information ecosystem.

Let's take a closer look at external, reference, and historical data.

External Data

A key source of data found in the CIF is that of external data (see Figure 2.2). External data is data originating outside the CIF. Typically, external data is purchased or created by another corporation. It can be of almost any type and volume and can be either structured or unstructured, detailed or summarized. In short, as many types of external data exist as there are internal data.

One fundamental way in which external data differs from internal data is in its capability to be manipulated. When internal data needs to be changed, the programs that capture and shape it can always be altered. In that sense internal data is very malleable.

However, external data is pretty much *what you see is what you get*. Because the sources for the external data lie beyond the CIF, it is beyond the scope of the CIF architect to effect such a change in it. About the only real choice the CIF architect has to make is to either use the external data as is or to reject its use altogether.

The one exception to the alteration of external data is that of modifying a key structure to the external data as it enters the CIF. This happens quite often when trying to match external data to an existing customer. Generally, an attempt is made to match the name and address associated with the external data to a name and address in the customer database. If a match is made, the external key is replaced with the internal customer ID, and the external data is stored.

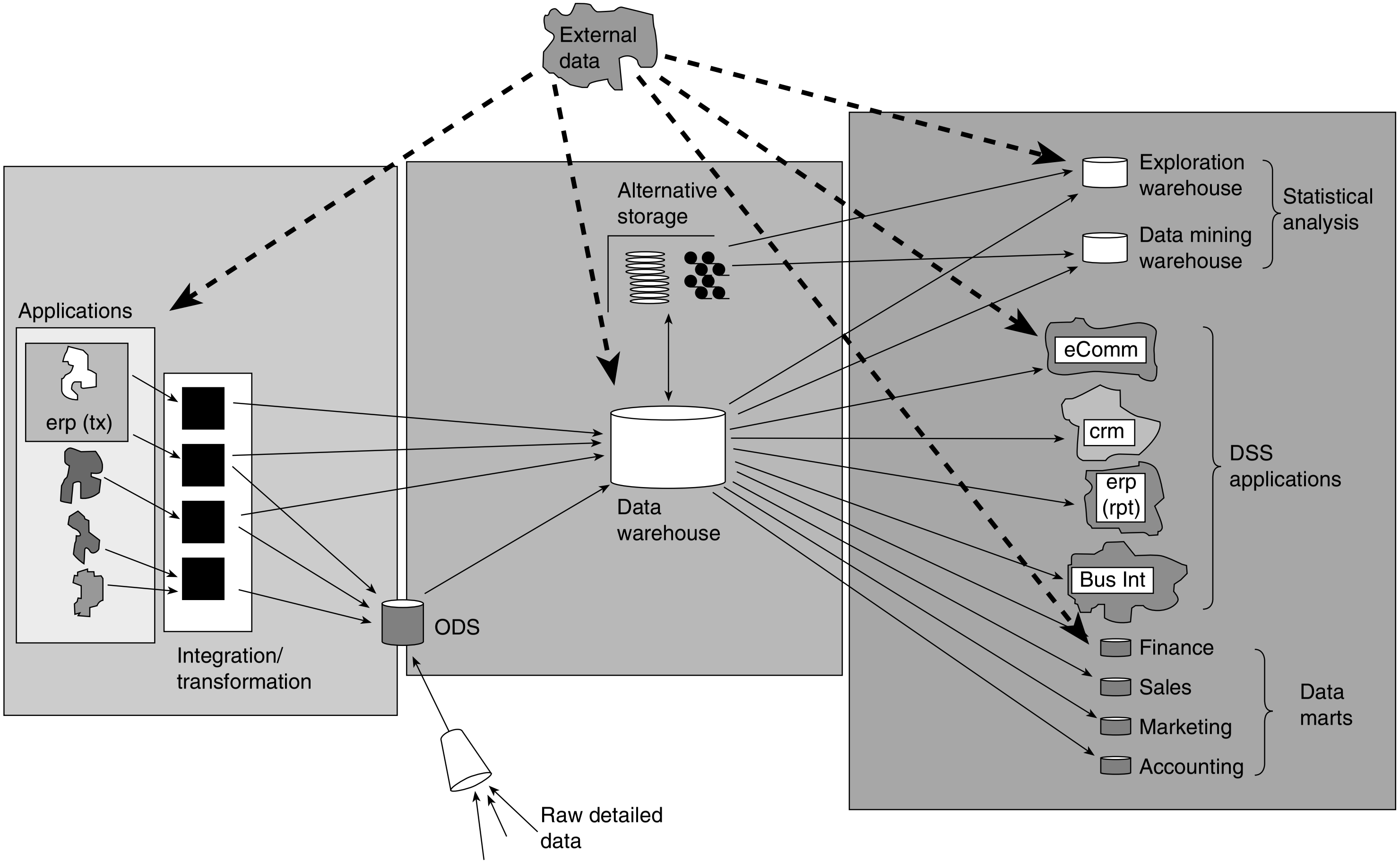


Figure 2.2 External data is an integral part of the CIF environment.

In many cases, the external data will have a key structure that is quite different from the key structure used within the CIF. The external data needs to have its keys modified in order to be used meaningfully within the confines of the CIF.

The modification of the external key can be a simple or a difficult thing to accomplish. In some cases, the external key goes through a simple algorithm to convert it to the CIF key. In other cases, reference tables are used in conjunction with an algorithm. And in the worst case, the conversion is made manually, on a record-by-record basis. The manual approach to key resolution is not viable for massive amounts of data and/or where the manual conversion must be done repeatedly.

External data can be made available to any and all components of the CIF. If the external data is to be used in multiple data marts, it is a good policy to place the external data first in the data warehouse and then transport it individually to the data mart. By placing it first inside the data warehouse, reconcilability of the data is maintained.

The component in which external data is most prominent is the exploration warehouse. In this environment, analysts endeavor to gain new insight about the business that cannot be distilled using internal transactional data. It is not uncommon for these analysts to use the exploration warehouse to identify new market opportunities or to characterize customers so that the business can better respond to their needs.

Reference Data

Some of the most important data any corporation has is reference data. One very popular type of reference data describes valid products and product hierarchies for a company. Reference data fulfills the following roles:

- It allows a corporation to standardize on a commonly used name for important and frequently used information, such as product, state, country, organization, customer segment, and so forth.
- It allows commonly used names to be stored and accessed in a short-hand fashion, which saves disk space.
- It provides the basis for consistent interpretation of corporate data across departments. For example, if reference data existed, we could be reasonably assured that three separate departments analyzing sales volumes for dog food would come up with the same answer. Without this reference data, each department is likely to roll-up products differently, resulting in different sales volumes for dog food.

In short, reference data is one of the most important kinds of data that a corporation has. Figure 2.3 shows the presence of reference data in the CIF.

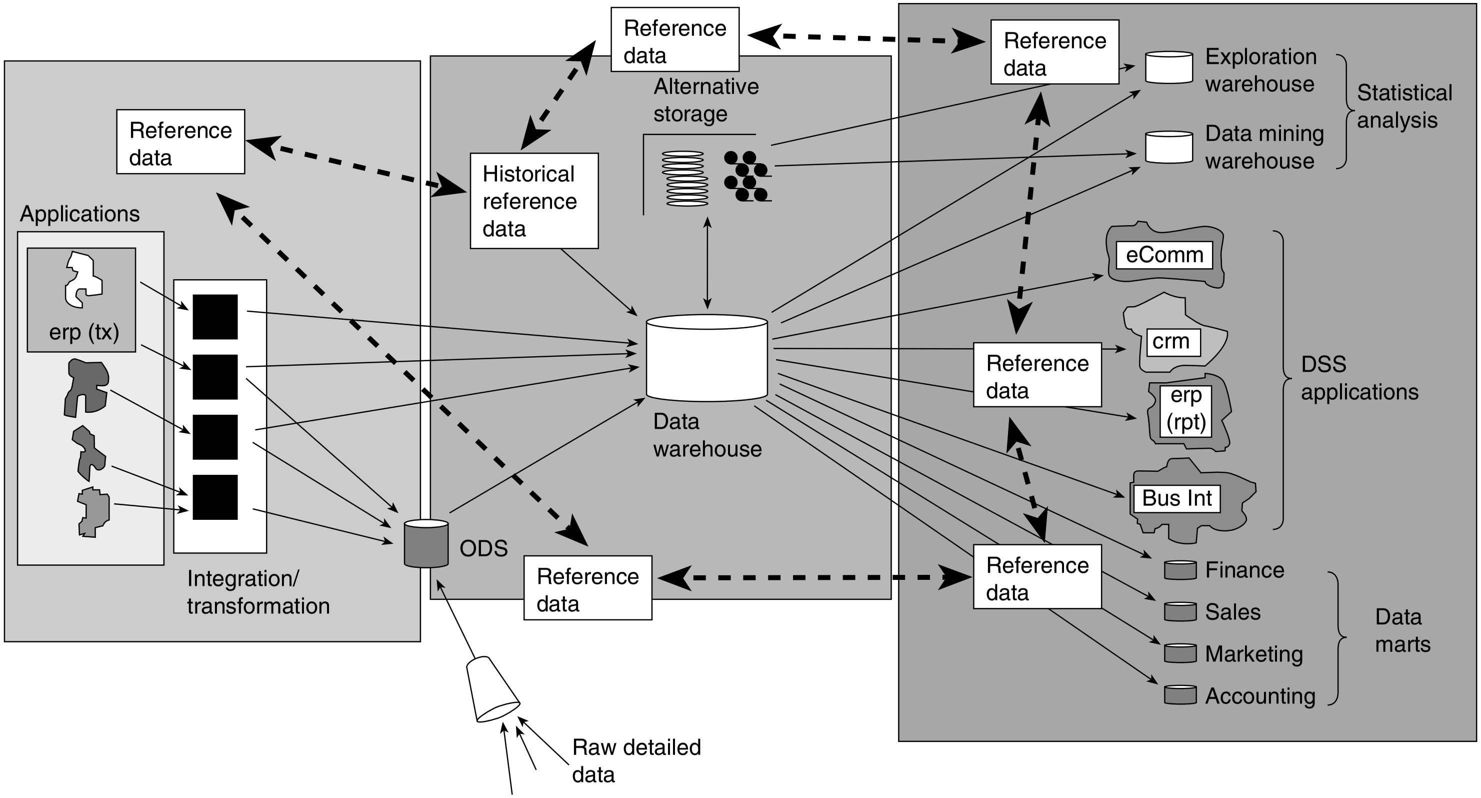


Figure 2.3 Reference data makes up an important part of the CIF. Note that reference data contained in the data warehouse is historical.

Reference data is notoriously unstructured and is, at best, a hit-and-miss proposition. This is in contrast to other forms of data, which are core to running the day-to-day businesses that require and receive great care in systematization. For example, data is needed to invoice a customer correctly. Programs and procedures are written for the update, creation, and deletion of nonreference data. But because reference data is so commonly used, programs and procedures needed for the systematization of reference data are not formalized. Several reasons exist for the lack of formalization:

- The volume of data that constitutes reference data is usually very small compared to other types of data found in the corporation. Reference data consumes only a fraction of a fraction of the space required for regular data. Because of its small size, reference data is often treated as an afterthought.
- Reference data is usually very slow to change. Unlike other types of data, which are constantly being created, deleted, and updated, reference data is very stable. Because of this stability, no one pays attention to the need for systematization of reference data.
- Reference data is often dictated by external sources. There are standard abbreviations for states, countries, and so on. There is no need for systematization of these types of reference data.
- Reference data often belongs to the entire corporation, not just a single department. Because reference data is a common corporate property, no one steps forward to *own* and manage the reference data.

For these reasons and more, reference data is often not managed with the same discipline that other data in the CIF is managed, yet it still requires as careful attention as any other type of data. For at least three reasons, reference data plays a very important role in the world of the CIF:

1. Reference data can simplify I & T layer processing. If reference data in an application is the same as reference data in the data warehouse, then the task of I & T is made much simpler. However, if the I & T layer must completely discard one approach to reference data and create an entirely brand new reference system (which can be done in extreme cases), then the logic of I & T processing becomes very complex and cumbersome.
2. Reference data is one of the primary ways that the different components of the CIF communicate and maintain continuity with each other. Whether you have implemented a data mart, exploration warehouse, ODS, or any other component of the CIF, well-formed and maintained reference data will help to ensure that such measures as revenue by *product group* and households by *customer segment* are consistent across the CIF.

3. Reference data ages over time. In the data warehouse, as reference data ages, a historical record must be kept so that the historical data that resides in the warehouse can have references made to the data that are accurate as of the moment of the creation of the data warehouse record. In other words, because historical data is stored in the data warehouse, an historical reference needs to be kept. If the DSS analyst is going back to 1995 to look at data in the data warehouse, he needs to know what the reference data was for 1995. It will not do to have the DSS analyst looking at 1995 data from the data warehouse where the DSS is trying to use reference tables from 1997. The need for historical referencability is one of the important and peculiar needs of the data warehouse within the context of the CIF.

Historical Data

Even when data has been entered onto a computer system and it is ten seconds old, it is historical in the sense that it represents events now passed. Of course, the event that has passed is much more current than events that may have occurred a week ago or a month ago. Nevertheless, all data entered into a computer system can be thought of as historical data (with the exception of forecast data). The issue is not whether data is historical, but just how historical the data is. The implications of historical data are many, including:

Volume of data. The longer the history is kept, the greater the amount of data.

Business usefulness. The more current a unit of information, the greater the likelihood that it is relevant to current business.

Aggregation of data. The more current the data, the greater the chance that the data will be used at the detailed level. The older the data, the greater the chance that the data will be used at the summary level.

Many other implications of history exist. These are merely the obvious ones. Figure 2.4 shows that the components of the CIF contain different phases of corporate information history.

The applications environment contains very current information, up to 30 days. Of course, the actual time parameters vary across industries and businesses. Some industries may have more than 30 days worth of information; other industries may have less.

The ODS environment has a time period identical to that of the applications. The difference between the ODS and the applications is that the ODS contains integrated corporate data, and the applications do not.

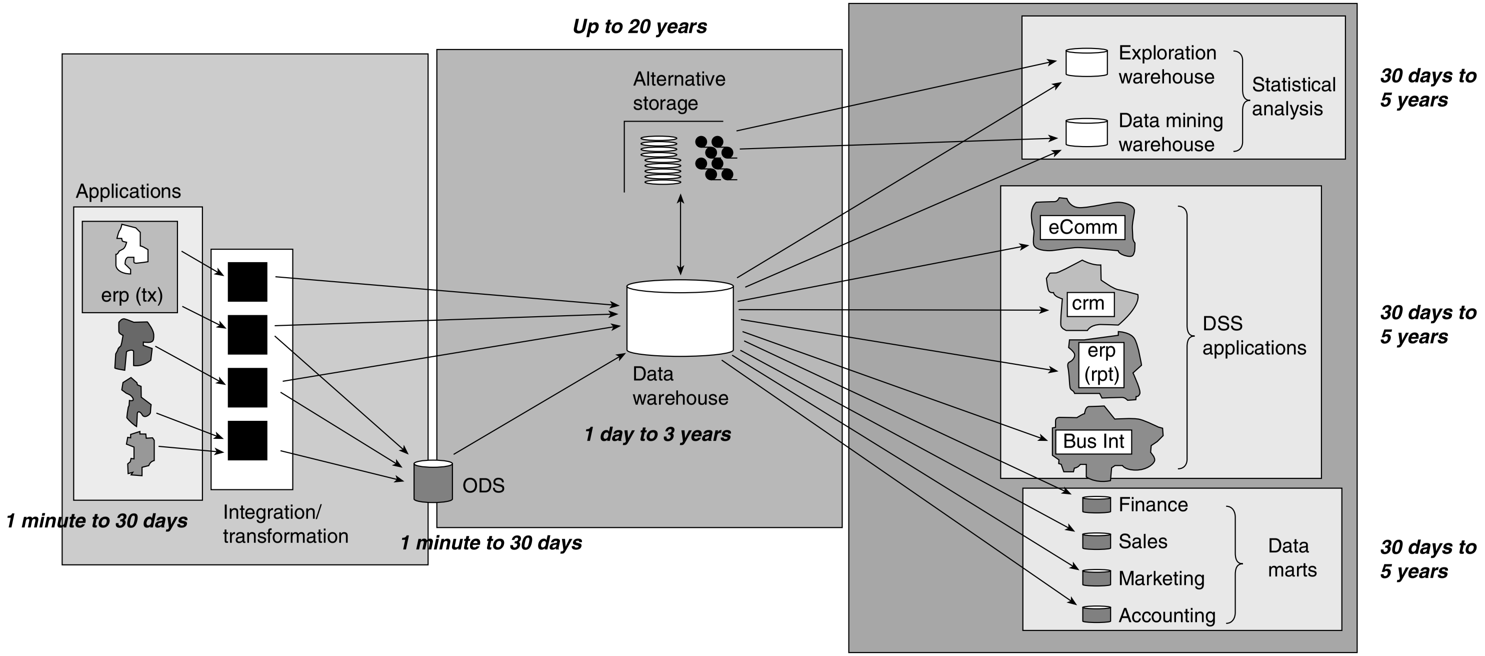


Figure 2.4 The amount of historical data that is found in the CIF differs from component to component.

The data warehouse contains data that is at least 24 hours old, up to 5 to 10 years worth of history. The actual length of time found here is highly dependent on the industry that is being represented by the data warehouse.

The data mart contains the widest variety of data found in the environment. The amount of history contained by a data mart is dependent on:

- The industry the corporation is in
- The company within the industry
- The department within the company

By far the greatest volume of historical data is found in alternative storage. This is where much of the historical transaction data from the data warehouse is archived. Historical data is even found in the exploration and data mining warehouses. Fortunately, use of historical data in these environments is project oriented so history is fairly pruned and temporary. As a result, the exploration and data mining warehouses don't require the large amounts of long-term storage or I & T layer processing to maintain history as do many of the other components of the CIF (data warehouses, alternative storage, data mart, etc.)

Of special interest is where different components overlap. The first overlap is between the applications arena and the ODS. As previously stated, the ODS contains corporate collective data, and the applications contain application-detailed (generally unintegrated or at best functionally integrated) data. There is overlap in the time frame, but no overlap in terms of the integration of the data.

The second overlap is between the data warehouse and the applications. An application may have data stored within it, up to 30 days or so. The data warehouse may have that same data stored. There are a few differences, however. The data warehouse contains data that has been passed through the I & T layer. As such, the data warehouse data may or may not be physically the same as the applications data. The second difference is that the data warehouse historical data is stored along with other historical data of the same ilk. The applications data is stored in an isolated manner.

The Decision-Support System to Operational Feedback Loop

The standard flow of data throughout the CIF is from left to right, that is, from the consumer to the application, from the application to the I & T layer, from the I & T layer to the ODS or the data warehouse, from the ODS to the data warehouse, and from the data warehouse to the data marts. The flow occurs as described in a regular and normal manner. However, another feedback loop is at work, as depicted in Figure 2.5.