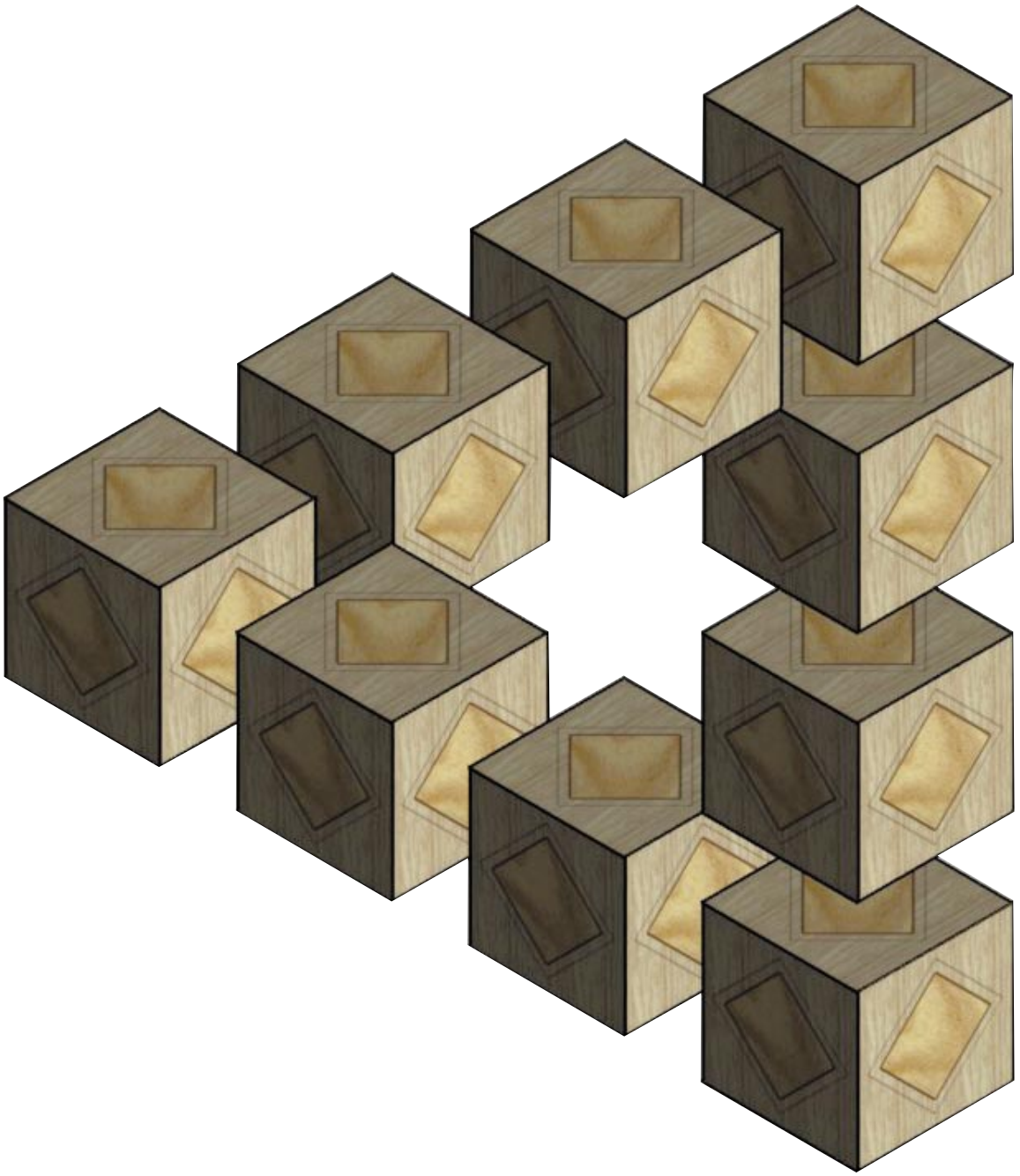
STEPHEN R. GARDNER

BUILDING the Data Warehouse

The tough questions project managers have to ask their companies' executives—and themselves—and the guidelines needed to sort out the answers.

Your company decides to build a data warehouse and you are designated the project manager. What are your first steps? You've read the books, attended the conferences, and perused the trade publications. Now you have to act. There are numerous vendors, all touting the wonders of their products, but you have specific questions that need specific answers, and building a data warehouse is an extremely complex process. Questions you have to weigh fall into the following general categories:

- **Costs.** How much can and should we spend for people, services, hardware, software, tools, and services from partners (vendors)?
- **Time.** How long will it take? How much time do we have?
- **Users.** What do the users—both IT and business—need from the warehouse? Do we have the data we need? Where is that data? Can we get to it? Is it good? Is it consistent across all systems?



A 3D MODEL BY CATHERINE PALMER
(HTTP://WWW.PALMYRA.DEMON.CO.UK)
FROM AN ORIGINAL DRAWING BY OSCAR REUTERSVARD.

- **People.** Who will build and maintain the warehouse? Will it grow so big we need an army of administrators to manage it?
- **Hardware, software, and tools.** What do we use? Where do we get it?
- **Services.** What can we do ourselves? Where do we get help?

One thing you've already read in articles and ads is that there is a lot of confusion in the industry: about what constitutes a data warehouse, how it should be architected, and what types of tools are needed. Depending on the needs of a particular business, a variety of vendors provide appropriate products. The important thing to know is that the questions data warehouse users have today are not necessarily the questions they will have tomorrow. Therefore, a data warehouse solution needs flexibility and scalability to change with the business it is intended to support. Flexibility is generally viewed from the architectural standpoint of a physical model based on third normal form (3NF), or the so-called "whole key," vs. predefined join paths, where every primary key in a dimensional table is a foreign key in the fact table. Due to the constantly changing needs of the warehouse, any architecture not based on the 3NF model can cause the failure of a data warehouse project. Regarding scalability, the main concerns are:

- The amount of data to reside in the warehouse
- The complexity of queries users are using to navigate the warehouse
- The number of users accessing the warehouse concurrently

Before delving deeper into these topics and others, let's first define a data warehouse (see Figure 1). Many definitions are floating around the industry today, but the following is, in my opinion, the most accurate and useful:

"Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business."

Informational vs. Operational Data

One of the first obstacles to overcome when building a data warehouse is the difficulty of understanding the differences between operational and informational data. Operational data is organized around functional organizations within a business. Functionally oriented data is used to satisfy the immediate functional processing requirements of the business user. Such a functional orientation is fine for operational data relevant to that area of the business.

However, the decision support system analyst needs information from across functional departments or business units, that is, data that is subject oriented, with an enterprise view. Subject-oriented, detailed transactional data allows corporate users to drill down into the heart of their business operations, not only to find answers to specific questions but to then show how and why they got each answer. Functionally oriented "stovepipe" systems do not allow this type of analysis.

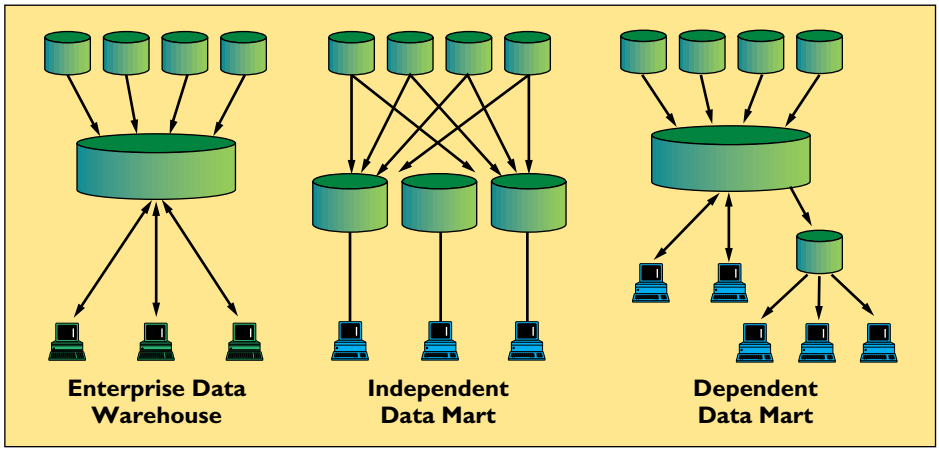


Figure 1. What is a data warehouse?

The first question the project manager has to answer about data warehousing and informational systems is: Who needs what data? Different kinds of users have different kinds of informational needs, and therefore different data and methods of data presentation and access. However, warehouse experience proves that if given the opportunity, users ask for a lot of data, in great detail, from multiple sources. The important thing is to give users the ability to think beyond the perceived IT department limits of technology or existing processes. For example, when asked how they would most like to travel from Los Angeles to New York, most people answer, "By supersonic jet, of course." However, the ideal option is being beamed over or teleported—instantaneous transportation. Most business users think only

within the limitations of the performance offered by their operational systems. Therefore, they simply ask for the same information but faster—not knowing that informational systems can give them access to information they never considered. This informational access means that within today's business environment, people try to answer business questions based on the limitations of existing systems and infrastructures.

Many users ask for data related to their specific functions. Would the same users be interested in cross-functional data if it were available? Answers to simple questions relevant to the business often differ depending on which operational systems' data and data definition were used for the analysis. People in sales look at orders taken, people in production look at units built, and people in finance look at orders billed.

For example, the ordering system may state gross sales, while the billing system may report net sales of discounts. The financial system may deduct allowances for returns or bad debts. This scenario is further complicated for a company with multiple functional systems, say, more than one billing system, that need to be aggregated to arrive at the correct sales totals. Since each functional organization originates, copies, summarizes, and transforms its own data differently, identical business questions can generate different results. Each of these systems keeps different historical data relative to the business. Isn't it valuable to management to know the number of units sold that are actually built and paid for? How can managers effectively run their businesses without this type of information?

Another often-overlooked business requirement is the need to execute cross-functional queries that were not previously possible. Taking data from different aspects of the business and analyzing the trends and correlations can be of infinite value to decision makers. This type of query is not possible in existing operational systems stovepiped along organizational lines. With a data warehouse that provides this cross-functionality, users can answer many crucial business questions once deemed unanswerable.

An enterprise data warehouse allows cross-functional analysis that not only finds out what is happening to the business but allows users to discover why certain things are happening. If certain trends or patterns are looked at within a single department or function of the business, the result and action taken may be completely inappropriate or appropriate only for that limited departmental function. Multiple factors can influence these trends and patterns, and the ability to identify what is occurring and why provides the business a significant competitive advantage.

The project manager should also think about the types of questions he or she would like to ask of the company's data but can't get answered today. The most likely reason a question can't be answered is because the required data sits in disparate systems. A solid informational infrastructure should enable users to get such questions answered.

Reengineering has become the preferred methodology for businesses that need to react to changing environments. However, stovepiped, incompatible operational systems do not offer the flexibility needed to support such reengineering efforts in a timely manner.

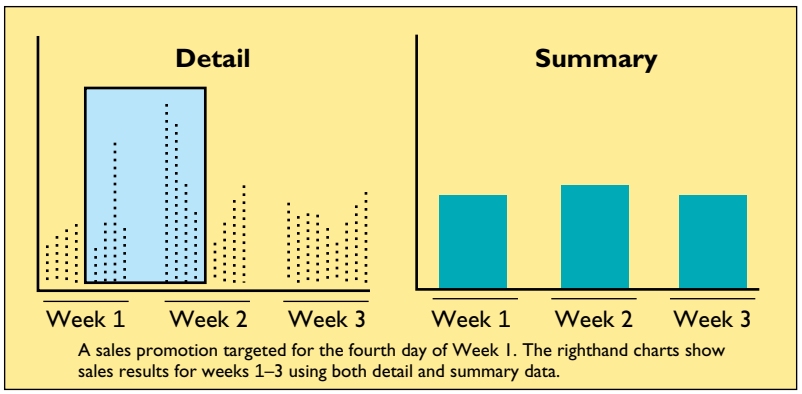


Figure 2. Why do I need detail data?

What Is Data?

Businesses today are inundated with data but have little information. A large number of legacy systems house huge amounts of operational data and even more data archived on nearly inaccessible tapes. This data is not valuable if it can't be readily turned into information, and the best way to get valuable information is to access data at the lowest atomic level of detail available. Historically, few companies have known what to do with detail data or recognized any value in it because of its volume. As a result, it was discarded. A data warehouse allows business users not only to find the answers to their questions but to understand how and why specific answers were received.

The only way to allow detailed, ad hoc analysis of the business is to capture detail and have access to it. The example in Figure 2 shows how summarized data can affect information so much that it becomes

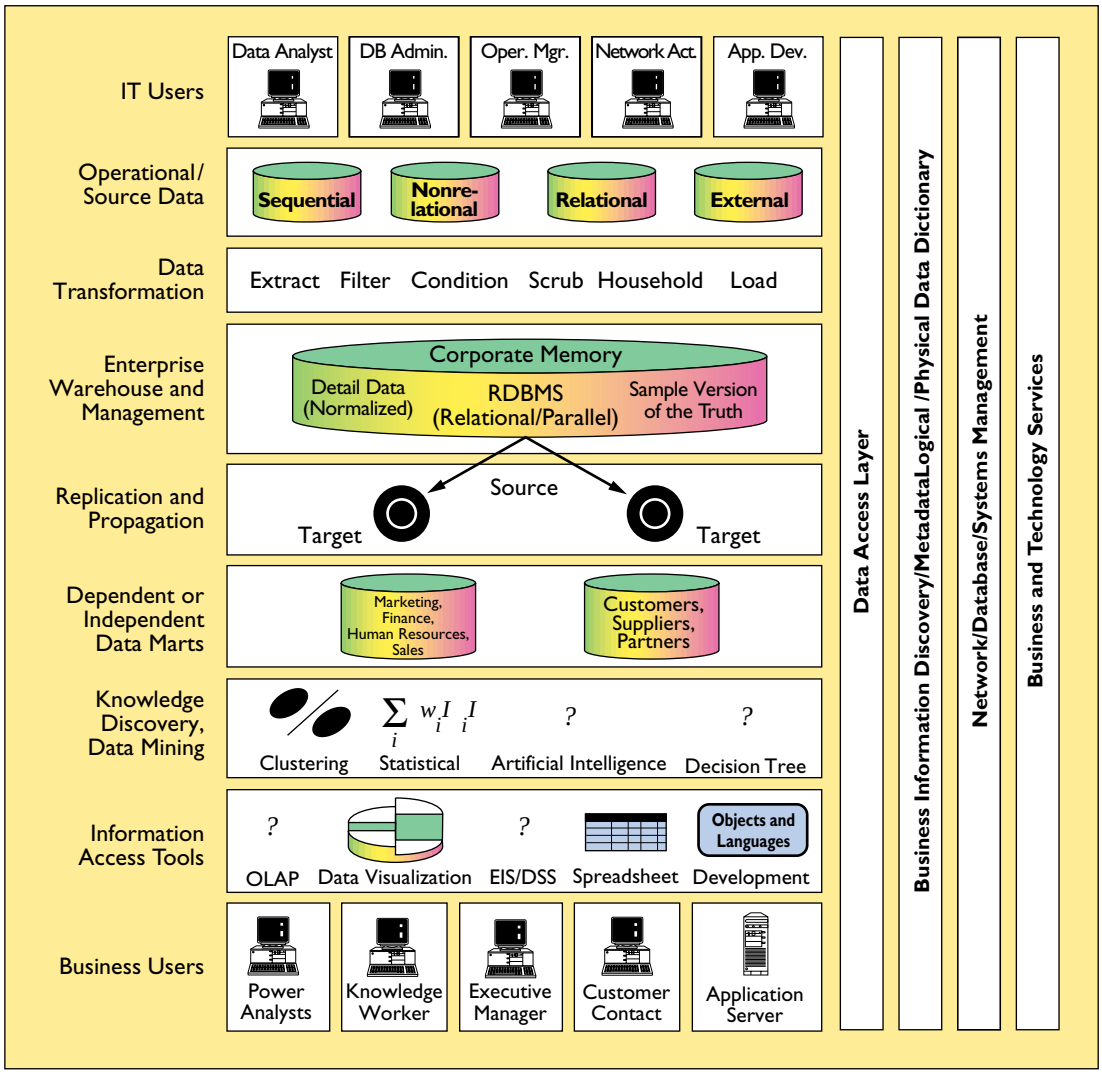


Figure 3. A data warehouse framework

misleading. The charts in Figure 2 represent the result of a promotional program targeted for the end of Week 1. Sales data captured at the daily detail level shows the significant result of a promotional campaign run at the end of Week 1. However, summarizing Week 1 daily sales into a weekly total significantly reduces the promotion's actual result and seems to imply the program was not a success.

Summarizing data is the processing of raw input data or detail data for more compact storage in a form useful for analysis in the particular application recording the data. Summarizing data selects, filters, combines, reorganizes, and manipulates detailed, or atomic, data to produce predetermined and specific categories, totals, and comparisons. It is a fixed model based on a user's unique view of a particular situation at a certain point in time. The hardware, software, and people needed to summarize data may be quite extensive and can quickly use up the support budget.

The response time of standard repetitive queries can be improved significantly if the data for the answer is summarized so answers are stored in a simple table with keyed access. Summarization can also reduce the amount of data stored online. Although this has some advantages, a warehouse project manager should be extremely cautious as to how summary data is applied. While summarizing data highlights some conditions, it hides others.

Summarizing data works only as long as the original business needs for creating the summary remain predictable and constant. When conditions change, summarized data may not meet the new requirements, and if the detail data is not available, the company cannot use its informational assets.

Framework and Methodology

Successfully implementing a data warehouse requires a proven framework, or blueprint. Just as you would not think of building a house without a

blueprint, the data warehouse project manager should carefully consider what framework to use to build a warehouse (see Figure 3) in three basic steps.

Planning. Information discovery services, which identify business problems to be solved, provide a structured process that is the critical first step in building a data warehouse. These services can be independent of each other and can be done in any order or concurrently. Each planning area represents an entry point into the warehousing methodology.

Design and implementation. The data warehousing solution readiness process represents another entry point into the methodology and should take place when warehouse developers are ready to begin their first data warehousing project and each time additional warehouse projects are initiated as the warehouse grows. It provides a comprehensive analysis of a company's current environment.

The solution readiness process validates the effectiveness of the identified solution within the current environment. Solution readiness investigates the elements needed to support the implementation, including data readiness, technology readiness, functional readiness, support readiness, and infrastructure readiness.

This step is intended to protect the business from attempting to implement a solution for which it is not prepared or that might influence other functional areas within the company not included in the planning. Implementation project plans should be adjusted (as needed) based on the results of the assessments.

Support and enhancement. Data warehouse support and enhancement comprises a series of follow-up operational and value processes supporting the operations and maintenance of a data warehouse. These processes serve the following purposes:

- Supporting the day-to-day running of the warehouse solution, ensuring availability and ongoing performance.
- Assisting in expanding the use (and therefore the benefit) of the solution.
- Expanding the system, possibly to include new applications, users, or data or increased use of the solution through the education of end users.
- Helping relaunch the process at the business imperative step, when selling senior management on the project, or if additional needs or applications are discovered for the next project consulting cycle.
- Helping keep the system continually updated and growing, supporting better business decisions in a planned and controlled way to deliver business value.

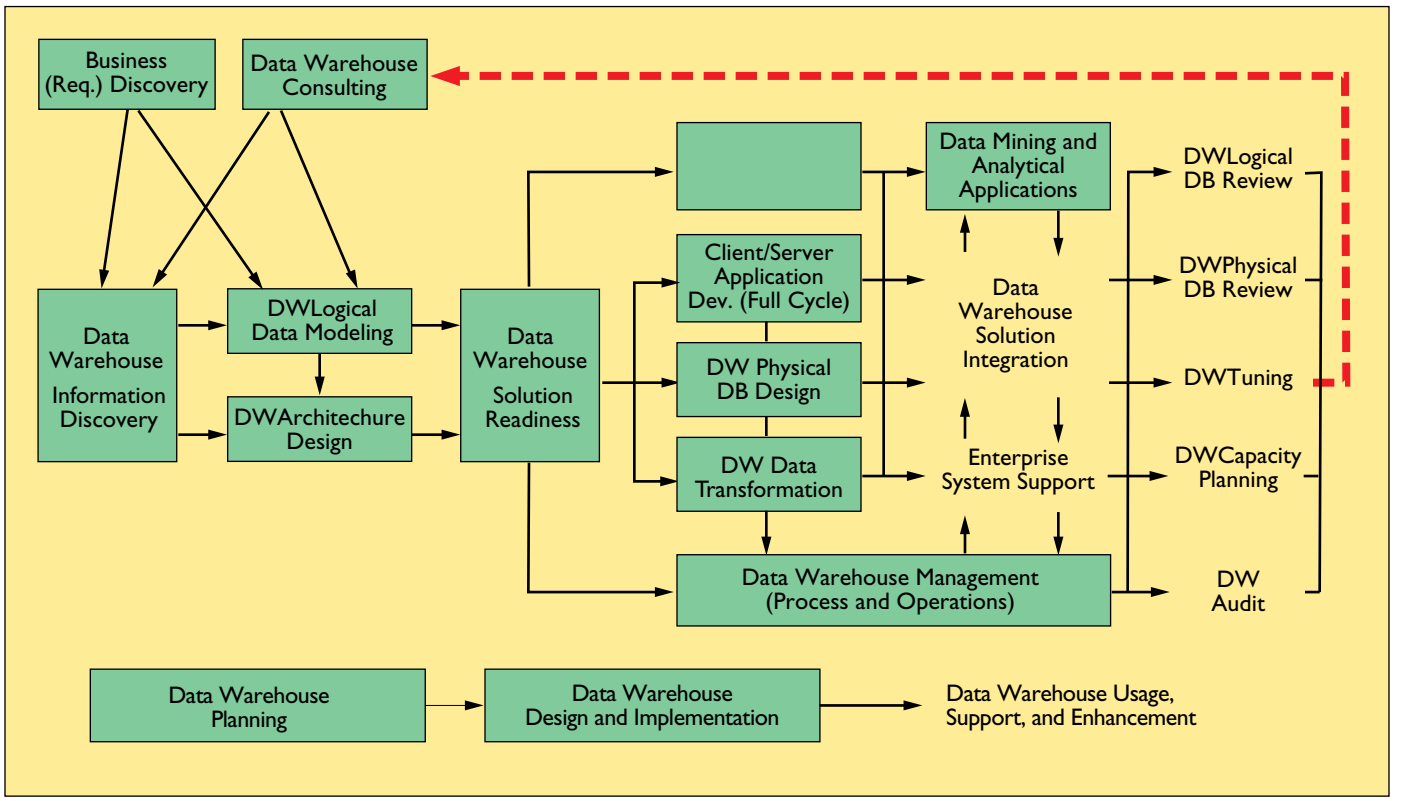


Figure 4. A data warehouse methodology

A warehouse methodology has to address all of these steps. Building a data warehouse is iterative; therefore, it is critical there be multiple entry points into the chosen methodology. Use of a proven methodology, coupled with collaboration between the IT department and business users, will greatly enhance the chances of successfully building the system. Figure 4 shows a proven data warehouse methodology representing an end-to-end solution with multiple entry points. This methodology represents the steps through which service providers and a user company's staff can make decisions regarding the warehouse and then implement and maintain that warehouse.

Health Care Management

Anthem Blue Cross Blue Shield's common repository for claims, revenue, and services provided by hospitals and physicians totals 1.3TB of data for the company's midwestern business operations.

CECILIA CLAUDIO

From a business standpoint, there has never been much dispute about the advantages data warehousing offers a company with large-scale information needs. Less well known is that data warehousing can be a great help to the human condition, especially in terms of health care.

Health plan policyholders are concerned about receiving the best treatment possible for as little cost as possible. With medical costs in the U.S. ever escalating, consumers, physicians, and employers alike often view these goals as mutually exclusive and that it is now nearly impossible to obtain quality, or even adequate, health care for a reasonable price.

Anthem Blue Cross and Blue Shield, one of the largest health care management companies in the U.S., is trying to ensure this perception proves inaccurate. Developing from its beginnings as a one-product—indemnity—one-state—Indiana—health insurer in 1944, Indianapolis-based Anthem has grown into a $6.5-billion (fiscal 1997) integrated health care provider with significant market share in two separate regions—the midwest and the northeast—in the U.S.

This growth has meant a significant IT challenge: how to integrate several disparate data warehouses into a single repository to create a "single version of the truth" for all users. Now, charged with managing Blue Cross Blue Shield plans in Ohio, Kentucky, Indiana, and Connecticut, Anthem faces an insatiable need for information about its more than six million policyholders and 350,000 providers, as well as the care they provide and their costs.

"We're moving from an insurance environment, where customers and health care providers submit claims and we pay them, to managing the health care our customers receive by working closely with the providers," says Anthem senior vice president Bill Milnes. "And we've found that you can't manage health care without managing information."

In 1995, with a $10.6-million contract, Anthem chose NCR Corp.'s Teradata relational database management system (RDBMS) as the platform on which it would build a consolidated database, thus creating a single repository system that would significantly improve companywide access to data. Anthem also chose a 16-node NCR WorldMark 5100M massively parallel processing server to accompany the RDBMS as its common repository for claims, revenue, and services provided by hospitals and physicians, and other vital information for Anthem's midwestern business operations. The consolidated data warehouse contains 1.3TB of data—enough, if put on paper, to fill 27,000 four-drawer filing cabinets.

Better access to information is helping Anthem improve the quality and reduce the cost of care for its policyholders by reducing fraud, negotiating lower rates with providers, accurately managing risks, and saving lives by increasing the knowledge of network physicians. The integrated data warehouse makes it possible for the company to enhance the quality of patient care and deliver more responsive customer service while reducing the costs of health care.

Lives and Money

Anthem has not only accomplished its goal of unifying its disparate corporate parts (it merged with Blue Cross Blue Shield of Connecticut in 1997)

Metadata

Metadata is popularly defined as data about data. In a relational database, metadata is the representation of the objects defined in that database—specifically, the definitions of its tables, columns, databases, views, and any other objects. In data warehousing, "metadata" refers to anything that defines a data warehouse object, such as a table, a column, a query, a report, a business rule, or a transformation algorithm.

Understanding these definitions is critical for all aspects of the data warehouse development process. Metadata management should tightly control everything—from developing programs that extract data from the source operational systems to transforming

but realized financial savings and productivity gains. The time now saved in searching for information in the NCR data warehouse (up and running since mid-1996) also allows users more time to analyze the data they collect. For example, when reviewing data for a particular medical procedure—coronary artery bypass surgery—Anthem found certain providers have superior success rates. Therefore, Anthem now funnels patients to these providers, reducing the procedure's mortality rate from more than 4% to less than 1% for Anthem policyholders while reducing costs to their employers.

The warehousing solution has also helped Anthem's negotiating staff. Armed with detailed data, sorted by region, product, procedure, and price, they can negotiate more favorable contracts with the company's more than 400 provider hospitals.

Anthem's legal department uses the warehouse for fraud detection. For example, an analyst recently uncovered a $37,000 bogus claims check. A particular provider's pattern of payments didn't fit the norm, and the analyst's research revealed that Anthem had actually paid several times for the same services. The provider was then billed for the overpayments.

The data warehouse has also helped Anthem win new business. For example, it used the system to design a custom report for a major prospect showing how costs could be cut in a particular geographic area. The system's ad hoc reporting capability was something the competition could not provide and made the difference in Anthem's winning the account. Anthem now has plans to give external users, such as doctors and hospitals, who need detailed data and direct access to the reporting capabilities of the warehouse.

Reporting Consistency

Before Anthem's mergers and the construction of the enterprise warehouse, the company had plenty of data—and an acute shortage of accurate answers. Users seeking answers to specific queries had to employ data sources from several disparate mainframe environments. This practice proved not only daunting and time-consuming but a breeding ground for inaccuracies and conflicts of data. Invariably, one department's findings would be different from another's. Worst of all, it left little time for data analysis—the lifeblood of the insurance business.

Users performing ad hoc complex queries now find reporting consistency throughout the company on enrollment, utilization, valuation, rate filing, loss ratios, provider profiling, and financial and marketing research. Moreover, with a single version of the truth available for all users at any time, Anthem quickly answers business questions essential for maintaining an edge in the competitive health care management market.

Scalability

Anthem also wanted a system that could grow as the company grew. Teradata was viewed as the most appropriate technology for databases of the size and complexity demanded by the company's midwest operations, with an anticipated growth rate approximately 10 times the current 1.3TB in the next three years. Incorporating common hardware building blocks across high-end symmetric multiprocessing, clustered, and massively parallel processing systems, NCR's scalable WorldMark servers and Teradata RDBMS will allow Anthem to expand its system to accommodate up to thousands of processors and many terabytes of data. C

Cecilia Claudio is a senior vice president and CIO of Farmers Insurance Group in Los Angeles and a former senior vice president and CIO of Anthem Blue Cross and Blue Shield in Indianapolis.

a collection of data into a target data warehouse. The warehouse is viewed as useful only if it provides a competitive advantage, that is, if the data transformed to populate the information store can be used to answer the business questions for which the warehouse was built.

Metadata is the road map or blueprint to that data. Just as a library card catalog points to both the content and the location of books in the library, metadata points to the location and meaning of various informational objects within the warehouse. Similarly, the data warehouse has to maintain a catalog of the items it manages. End users are like library customers, making requests for information based on selections made from a catalog. The process that fulfills their requests must know where the information is located within the data warehouse.

The warehouse must therefore contain a component that fulfills the catalog functions for the information it manages. This catalog is organized to do the following:

- Serve as a map to the locations where information is stored in the warehouse
- Include 2D components for every item, such as the definition needed by the database technology (table name and table owner) and the definition needed by the business user
- Provide a blueprint for the way in which one kind of information is derived from another kind of information
- Provide a blueprint for extractors that take data from operational systems and load it into the data warehouse
- Store the business rules built into the data warehouse
- Store the access control and security rules to support the administration of security
- Make it possible for metadata to track changes over time
- Organize metadata to be versioned to capture its change history
- Store the structure and content of the data warehouse
- Identify clearly and formally the system of record for the data warehouse
- Make available the integration and transformation logic as a regular part of the warehouse's metadata
- Store the history of data refreshment
- Store metrics, so end users can determine whether a request will be large or small before submitting it

Conclusions

Why do some projects fail? There are five primary reasons: lack of partnership between the IT department and business users; incorrect data warehouse architecture; not enough experienced people; improper planning, such as failure to use a proven methodology and a plan to ensure that no details are omitted; and depending on bleeding-edge technology.

On the other hand, experience shows the following precautions encourage successful warehouse implementations: win the highest possible level of executive support; identify a specific business problem to be solved; create a well-defined plan; use proven technology; and employ experienced people. Cooperation and support at a business's executive level is important to success. Executive management can mediate political and funding issues while providing a foundation for collaboration between the IT department and business users.

Choosing a specific business problem to solve and defining requirements and measurements for the solutions help focus the system's direction. Solving this problem by properly implementing a data warehouse ensures success and helps the warehouse grow to help solve even more business problems—resulting in even better business operations.

Be sure a good plan is in place and that project management is top notch. A warehouse project is not a good place for a novice project manager to start. An experienced project manager helps create the plan and then keeps everyone on track.

Employ as many experienced people as possible from both inside and outside the business. They know the pitfalls and should be able to help mentor those who are less experienced.

Building a warehouse is a complex process requiring careful planning and alignment between the IT department and business users. Data warehouses are built to answer specific business problems, not to showcase the wonders of technology. Using the guidelines outlined here can significantly improve your chances of success. C

Stephen R. Gardner (drsrg@bigfoot.com) is director of advanced technology research at NCR Corp. in Seattle.