Designing, Developing, and Supporting an Enterprise Data Warehouse (EDW) In Healthcare ©

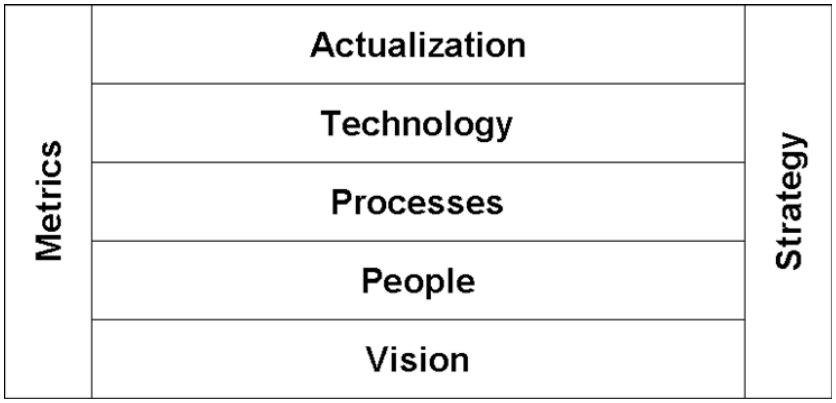
Dale Sanders
Intermountain Health Care

Introduction

The Dutch physicist, Heike Kammerlingh Onnes, discoverer of superconductivity in 1911, posted a sign above the entrance to his laboratory--- “Through measurement, comes knowledge.” In no other field of study, including physics, are measurement and true knowledge more complex, more elusive, or more subjective than that found in healthcare. We are measuring ourselves and in so doing, the observer becomes the observed. The challenge to find the truth is simultaneously fascinating and daunting. The essence of data warehousing is not information technology; information technology is merely the enabler. The essence of data warehousing is measurement, and through this measurement, follows understanding, and through this understanding, follows behavioral change and improvement. At Intermountain Health Care (IHC) in Salt Lake City, UT a team of medical informaticists and information systems professionals recruited from other industries was assembled in 1997 to develop and deploy an enterprise data warehouse (EDW) to measure and better understand IHC’s integrated delivery system. The intent of this chapter is to provide a brief review of transaction-based and analytical-based information systems and the emergence of data warehousing as a sub-specialty in information systems, and discuss the lessons learned in the deployment of IHC’s EDW.

Background

The success of any information system—data warehouse or not—is based on a “Hierarchy of Needs for Information Technology” that is similar conceptually to Maslow’s Hierarchy for human actualization. The success of a data warehouse begins with this sense of IT Actualization, as illustrated below.



Successful IT systems must be founded upon a clear vision of the future for those systems and their role in the enterprise. They must be founded upon an environment that nurtures people that are values based, understand information technology (IT), and fully understand the business and clinical missions that they support. These same people must be allowed to define and operate within a framework of IT processes that facilitates quality, productivity, repeatability, and supportability. Architecting the information technology is the final manifestation of the underlying vision, people, and processes in the journey to IT Actualization and success. All of these steps in the journey must be wrapped in a sense of metrics—measuring the progress towards Actualization---and a systemic strategy that unites each.

Transaction and Analytical Systems: At a high level, there are two basic types of functions supported by information systems—(1) Transaction processing that supports an event-driven clinical or business process, such as patient scheduling, and (2) Analytical processing that supports the longitudinal analysis of information gathered through these same transaction systems. In some cases a transaction system may have little or no need for an analytical capability, though this is very rare. And in some cases, an information system is designed expressly for retrospective data analysis and supports very little in the way of true workflow, e.g., a project time tracking system.

The purest form of an analytical information system is a data warehouse. Data warehouses have existed in various forms and under various names since the early 1980's, though the true origins are difficult to pinpoint. Military command and control and intelligence, manufacturing, banking, finance, and retail markets were among the earliest adopters. Though not yet called "data warehouses", the space and defense intelligence industry created integrated databases as early as the 1960s for the purposes of analysis and decision support, both real-time and off-line. A short and sometimes overlooked period in the history of information systems took place in the early to mid-1990s that also affected the evolution of data warehousing. During this period, there was great emphasis placed on "downsizing" information systems, empowering end users, and distributing processing to the desktop. Client-server computing was competing against entrenched glass house mainframes and was seen as the key to this downsizing and cost reduction. Many companies undertook projects to

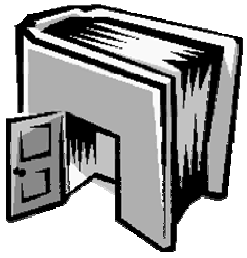
convert mainframe databases and flat files to more modern relational databases, and in so doing, place their data on fewer hardware servers of a common architecture and operating system. History, of course, revealed that client-server computing was actually much more expensive than centralized applications and data, and thin clients. However, despite what some might call the failure of client-server computing, this is the period that created the first data warehouses in private industry.

In reality, a data warehouse is a symptom of two fundamental problems in information systems—(1) The inability to conduct robust analytical processing on information systems designed to support transaction oriented business processes, and (2) Poorly integrated databases that provide a limited and vertical perspective on any particular business process. In a perfect environment, all analytical processing and transaction processing for all workflow processes in an enterprise would be conducted on a single, monolithic information system. Such is the vision of "Enterprise Resource Planning" (ERP) systems, found more and more often in the manufacturing and retail markets. But even in these systems, the vision is elusive, at best, and separate analytical and transaction systems are generally still required to meet the needs of the company. Recognizing that transaction processing and analytical processing require separate IT strategies is an imperative in the architecture of a successful enterprise information system. Unfortunately, in many cases, IT strategies tend to place overwhelming emphasis on the needs of the transaction system and the analytical processing requirements of the enterprise are an afterthought. Yet time and time again, we witness situations in which transaction data is collected quite effectively to support a workflow process, but extracting meaningful reports from this system for analysis is difficult or impossible. Rarely, if ever, is a transaction system deployed that will not require, at some point in its lifetime, the analysis of the data it collects. *Deliberately recognizing this fact in the requirements and design phase of the transaction system will result in a much more elegant solution for the analytical function*. The knowledge gained from the analytical function can be used to improve the front-end data collection process and enhance the design of the transaction system—e.g., improving data validation at the point of collection to improve quality; adding additional data elements for collection deemed important to analysis, etc. In this regard, we can see the constant feedback and interplay between a well-designed information system-- the transaction function supports

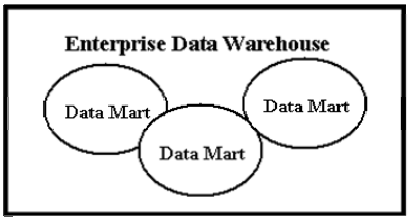
the analytical function which supports the improvement of the transaction system, and so on in a constant cycle of improvement.

As illustrated below, a data warehouse is analogous to a library—a centralized logical and physical collection of data and information that is reused over and over to achieve greater understanding or stimulate new knowledge. A data mart, which is a subset of the data warehouse, is analogous to a section within a library.

- **Data warehouse:** A database repository that integrates data from across the enterprise
  - Analogous to a library
- **Data mart**: A database repository that consolidates or integrates data and supports a single business area or specific reporting requirement
  - Analogous to a section in a library



It is not the Clinical Data Repository

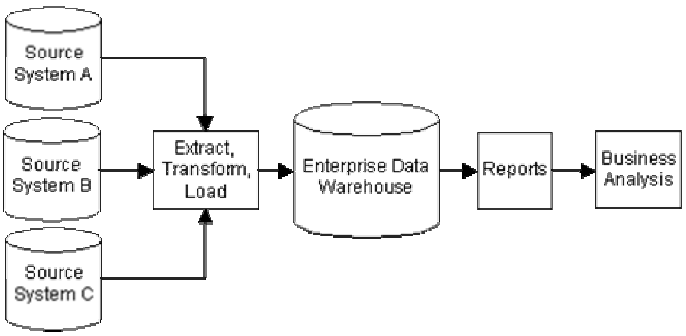


It is difficult to trace the origins of data warehousing because its beginnings evolved slowly and without a formal definition of "What is a data warehouse?" Ralph Kimball is credited with driving the semantics of this specialty in information systems. Prior to his early writings, there was no common language to describe the specialty. [(11)] Consequently, many companies were striving to improve their analysis abilities by integrating data, but doing so through an ad hoc process because no formal language existed to describe anything formal, especially between other companies facing the same challenges. Networking with other professionals about data warehousing did not take off until the mid-1990s, coincidentally with the publication of Kimball's first book on the topic.

In a simplistic style, a data warehouse is merely the integration of data at the technological level—i.e., centralizing the storage of previously disparate data on a single database server under a common relational database management system. In its more complex form, a data warehouse is characterized by the true integration of disparate data *content* under a very formal design and supporting infrastructure with a well-defined purpose for strategic decision support and analytical processing. Either form of a data warehouse has its pros and cons. The technology-driven form is relatively easy and less costly to implement, but very little synergy is derived from the data itself. Today, the term data warehouse is almost exclusively reserved to describe content-driven data integration.

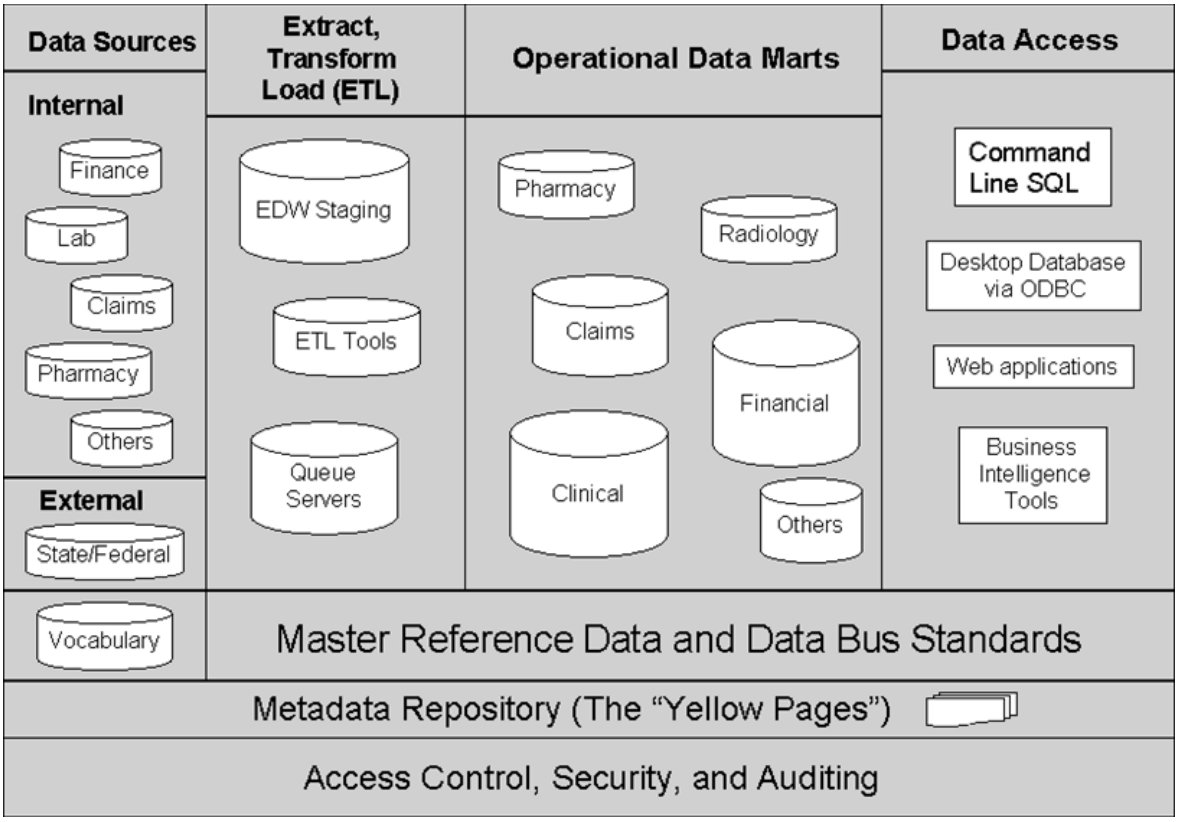
The explosive growth of data warehousing is actually a symptom of a larger problem, i.e., silos of non-integrated, difficult-to-access data, typically stored in legacy information systems. The emergence of data warehouses coincided with improvements in the price/performance ratios of modern database hardware, software, and query tools in the late 1980s, as well as a lingua franca for data warehousing as an information systems specialty. These early attempts at building "data warehouses" were motivated primarily by improving access to data, without regard for improving decision support. However, once data was integrated and easier to access, users discovered that their decision support and data analysis capabilities improved in unexpected ways. *This is a key point: It is not necessary to plan for and predefine all the reports and benefits of those reports expected from a data warehouse. Quite often, the greatest benefits of a data warehouse are not planned for nor predicted a priori*. The unforeseen benefits are realized after the data is integrated and users have the ability to analyze and experiment with the data in ways not previously possible.

The basic data flow diagram for a warehouse is depicted below:



Data is extracted from multiple sources systems, blended together in the extract, transformation, and loading process, and loaded into the EDW in a form that facilitates reporting and analysis.

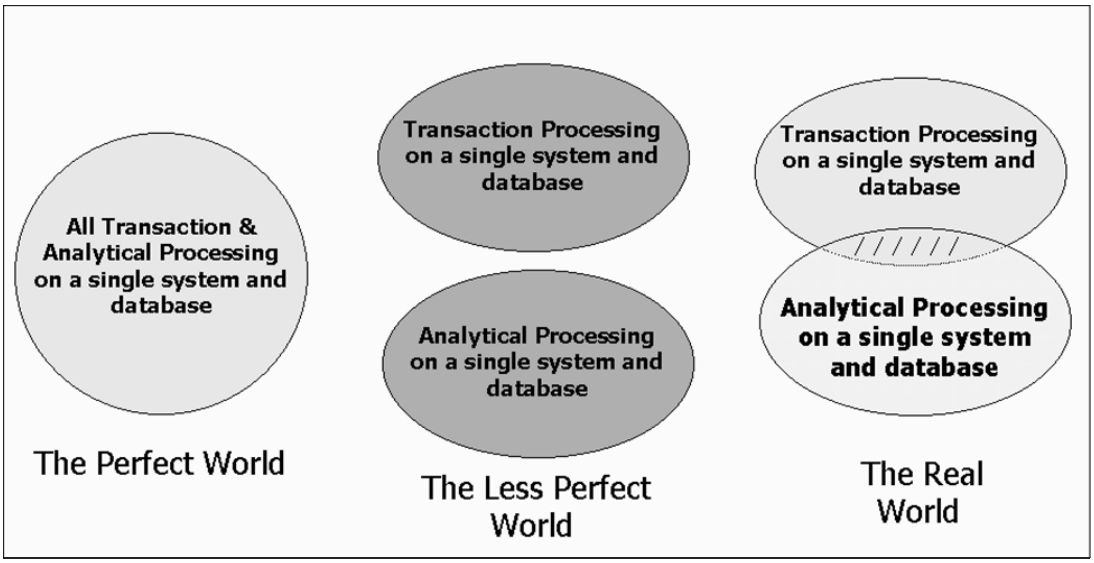
Another, more detailed diagram of a data warehouse architecture is depicted below.



In the above diagram, the flow of data and information is from left to right. Source data can be supplied by information systems that are internal to the company, and by external systems, such as those associated with the state or federal government (e.g., mortality data, cancer registries). A standard vocabulary for consistently mapping similar concepts to the same meaning must be applied to these data sources as they are introduced to the EDW environment. The extract, transformation, and loading (ETL) process pulls data from the source systems, maps the data to the EDW standards for naming and data types, transforms the data into a representation that facilitates the needs of the analysts (pre-calculated aggregates, denormalization, etc.), and loads the data into the operational area of the data warehouse. This process is typically supported by a combination of tools, including ETL tools specifically designed for data

warehousing. A very important type of tool supporting the ETL layer in healthcare are those that apply probabilistic matching between patient demographics and the master patient identifier (MPI), when the MPI is not ubiquitous in the enterprise. Data access is generally achieved through one of four modes: (1) Command line SQL (Structured Query Language), desktop database query tools (e.g., Microsoft Access), (2) Custom web applications that query the EDW, and (4) Business intelligence tools (e.g., Cognos, Crystal Decisions, etc.). Underlying the EDW is master reference data that essentially defines the standards for the "data bus architecture" [7] and allows analysts to query and join data across data marts. The underlying metadata repository should be a web-enabled "Yellow Pages" of the EDW content, documenting information about the data such as the data steward, last load date, update frequency, historical and temporal nature of the data, physical database name of the tables and columns as well as their business definition, the data types, and brief examples of actual data. Access control processes should include the procedures for requesting and approving an EDW account; criteria for determining when access to patient identifiable data will be allowed; and criteria for gaining access to other data in the EDW. Access to patient identifiable data should be closely guarded and, after access has been granted, procedures for auditing that access must be in place.

As discussed earlier, in a theoretical world, all transaction and analytical functions occur on the same information system. In a less perfect world, two distinct information systems are required to support the two functions. In the real world of most companies, there are two distinct information systems to support transaction needs and analytical needs of any given business area, and their analytical capabilities overlap, resulting in redundant reports from the two systems. For obvious reasons, the vision should be to minimize this overlap and redundancy. This concept is depicted below.



As discussed earlier, there are two fundamental motivators when assessing potential data to include in a data warehouse environment: (1) Improving analytical access to data that is "locked" in an information system that is difficult to use; and (2) Linking data from disparate databases, such as that from ambulatory clinics and acute care facilities, to gain a better understanding of the total healthcare environment. These two motivators also play a role in influencing the development strategy for a data warehouse. The best scenario for creating a successful data warehouse is one in which both motivators are important to the project. Typically, if the users of the transaction systems are dissatisfied with their analytical capabilities, they will become strong allies in the development of a data mart that supports their needs. This support can be leveraged to push the project towards successful completion, while the data is also integrated for synergy with other data marts in the warehouse. The enterprise will benefit from the data as well as the vertical business area supported by the data mart—these types of projects are truly win-win and possess a track record of success.

Data warehousing in healthcare evolved across several different environments, as listed below, listed more or less in order of their emergence over time:

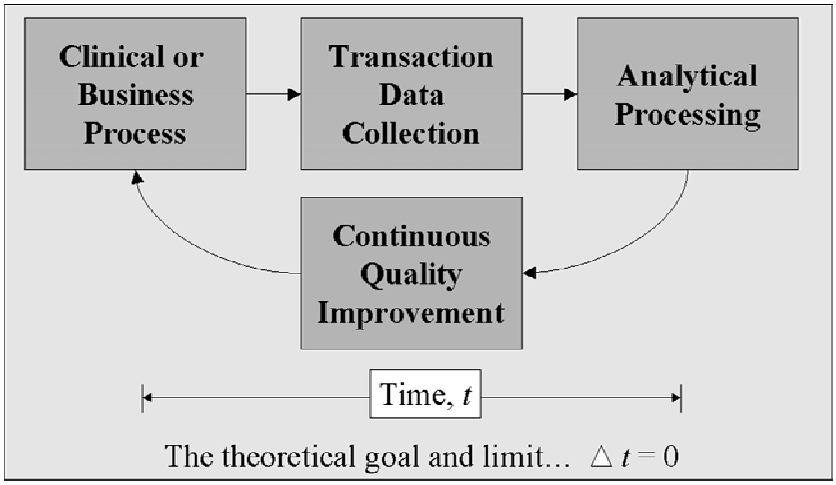
- Research databases, especially those funded by National Institutes of Health and Centers for Disease Control and pharmaceutical companies
- Department of Defense, Veterans Affairs
- Insurance, especially Blue Cross/Blue Shield

- State or federally mandated data integration for registries and outcomes reporting
- Multiple hospital systems
- Integrated delivery systems

It is worthwhile to note that data warehouses are still not prevalent in the settings of small groups of, or individual, hospitals. Several factors contribute to this situation, including the fact that the true power of data warehouses cannot be realized at low volumes of data—enough data must be available to support statistically significant analysis over statistically valid periods of time to identify trends from anomalies. Another, and potentially more serious contributor, is the high cost associated with data warehousing projects. The hardware and software costs have dropped in recent years, especially with the advent of Microsoft-based platforms capable of handling the processing demands of a data warehouse. However, the real costs are associated with IT labor—the design and development labor, especially. And unfortunately, off-the-shelf "turnkey" data warehouses offered by most vendors have not succeeded as hoped; therefore the EDW solutions that truly function as expected are primarily custom built. Off-the-shelf EDW's have not succeeded in health care, or any other major market or industry, because there is very little market overlap between different companies in the profile of the source information systems—different companies use different source systems and different semantics in their data to run their businesses-- creating a "one-size-fits-all" EDW design is essentially impossible.

The fundamental business or clinical purpose of a data warehouse is to enable behavioral change that drives continuous quality improvement, through greater effectiveness, efficiency, or cost reduction. *If a data warehouse is successfully designed, developed, and deployed as an information system, but no accommodations have been made to conduct data analysis, gain knowledge and apply this knowledge to continuous quality improvement, the data warehouse will be a failure.* For this reason, the continuous quality improvement process must be considered an integral part of the data warehousing information technology strategy—neither can succeed without the other. According to the Meta Group, 50% of the business performance metrics delivered via a data warehouse are either directed at individuals not empowered to act on them, or at empowered individuals with no knowledge of how to act on them. The CQI process must be accurately targeted at the right people in the company that can implement

behavioral change. In addition, the continuous quality and process improvement strategy should seek to minimize the time that expires between recognizing that an opportunity has been identified for quality improvement, and the execution of that opportunity. In a theoretical world, that time delay is zero—the process improvement is made at the same time the opportunity is identified. The figure below depicts these relationships.



Risks to Success

In some companies, the rush to deploy data warehouses and data marts has only recreated the problems of the legacy systems, albeit in a more modern form. In the absence of an overall strategy, many of these data warehouses and data marts became silos of inaccessible data in their own right. In general, this modern version of a legacy problem can be attributed to two general causes:

Lack of data standards: An enterprise standard data dictionary for common data formats, coding structures, content, and semantics is critical. The most difficult problem to overcome in any data warehousing effort is the elimination of data homonyms (different attributes with the same name) and data synonyms (same attributes with a different name) between systems. To avoid being crippled by data homonyms and synonyms, it is imperative that these standards be established for core data elements prior to the development of any data marts comprising the data warehouse.

Inadequate metadata: Metadata functions as the EDW's "Yellow Pages" and is analogous to a library's card catalog system The value of metadata increases geometrically as the scope and exposure of the data warehouse expands across business and clinical areas in the company. Metadata is most useful to those analysts who are not intimately familiar with a particular subject area of data, but could benefit significantly in their analysis if they had even a limited understanding of the data content in the unfamiliar subject area. Documentation that accurately describes the contents and structure of the data warehouse to customers and support personnel is critical. Imagine a large library, in which the books and periodicals are not arranged or categorized in any particular order, or a department store that lacks overhead signs or products that are arranged by general category. The manner in which the data warehouse is organized and the communication of this organization to customers is as important as the contents of the warehouse itself.

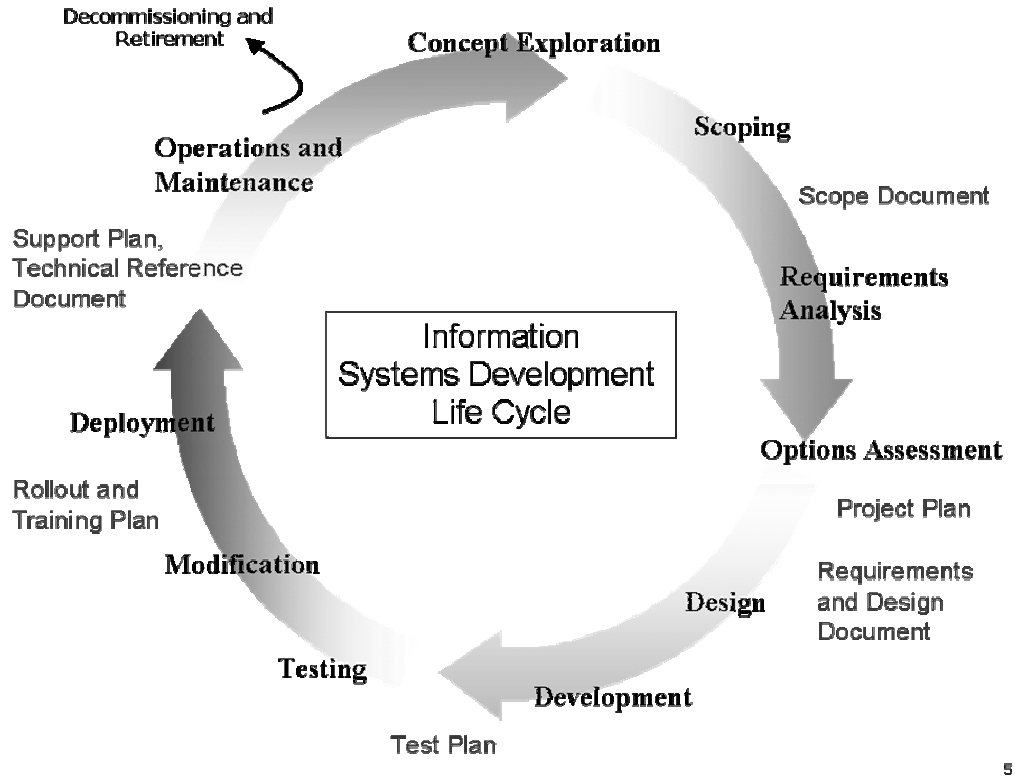
Other risks to success of the EDW are summarized below.

1. Insufficient resources are provided to sustain the operations, maintenance and growth of the data warehouse
2. The warehouse has no support from key business sponsors
3. The organization's information systems infrastructure is not scaleable enough to meet the growing demands for the data warehouse
4. Users are not provided with the tools or training necessary to exploit the data warehouse
5. Individual business areas and data "owners" are not willing to contribute and cooperate within the vision of the EDW
6. Data quality and reliability fail to meet user expectations
7. The EDW implementation team lacks at least one person with experience in all phases of the lifecycle of an EDW
8. The company lacks adequate source information systems. Quite often, companies will engage in a data warehousing project when their transaction source systems are in shambles. These companies would be better served by spending their resources on improving their transaction systems, first.

Our knowledge is bound by the information we have available, or the information for which we are willing to pay. Data warehousing is an interesting investment in knew knowledge—achieving "data synergy." A data warehouse literally enables knowledge and insight that simply did not exist prior to the investment. It is a fascinating thing to witness unfold in the real world, especially healthcare, and participate in the insight and discovery that ensues.

Methodology

The detailed methodology for building a data warehouse is unique from other types of information systems, however, at a high level, a data warehouse lifecycle is the same as any other information system, as depicted in the diagram below. It is important to recognize the different stages and deliverables associated with this lifecycle and manage each differently. A common mistake is the assumption that one person is capable of managing each phase of the lifecycle equally well. The truth is quite different. A data warehouse team must be managed and staffed using a 'division of labor' concept. The team should have at least one person on the staff that has experience through the entire lifecycle of an EDW. The other staff members should have expertise in each of the sub-phases of the lifecycle so that, at the macroscopic level, the skills



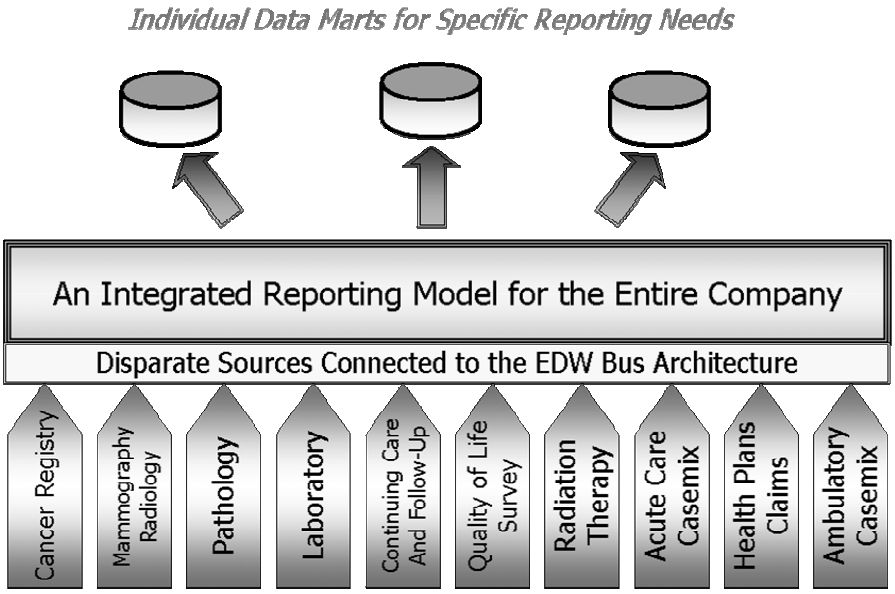
profile of the team fits within the lifecycle like pieces in a puzzle. No part of the lifecycle should be without a competent, experienced member of the team.

In general, three methodologies exist for deploying an enterprise decision support system based on data warehousing and data mart concepts—top down, bottom up, and a combination or hybrid approach.

Top Down Implementation

As the name implies, this approach starts from the enterprise level and works down to the data marts associated with individual business areas. The EDW functions as the source of data for the data marts. Among other tasks, this approach to implementation requires the construction of an enterprise data model and data standards before construction of data marts. Historically, this approach is too slow to respond to the needs of the company and is notorious for a track record of failure.

The diagram below depicts the concept of an EDW in which data marts are populated from a top-down enterprise model.



Bottom Up Implementation

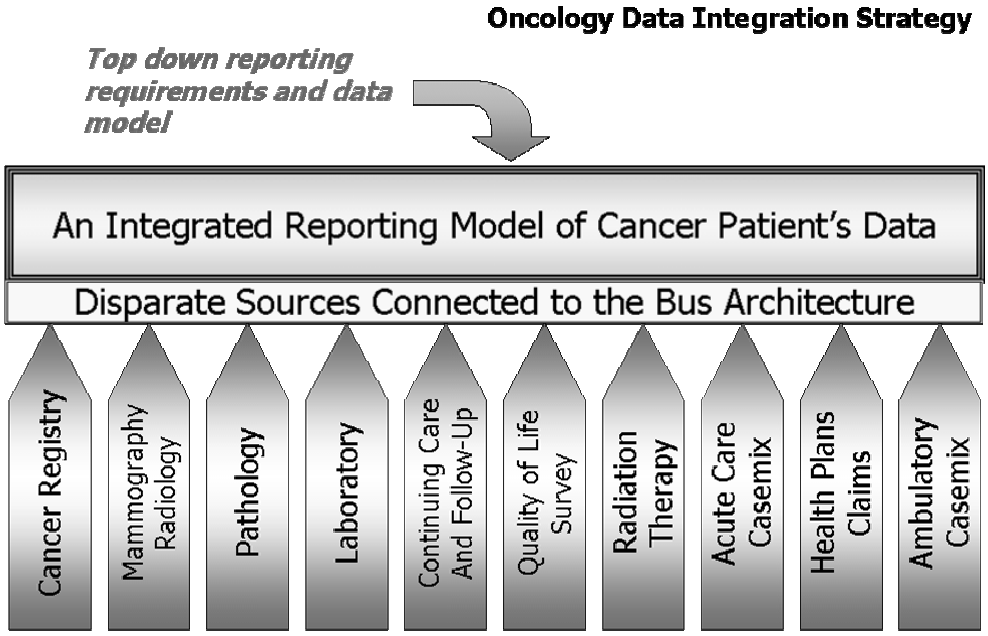
A bottom-up implementation plan focuses on the individual subject areas and constructs individual data marts for each of these areas, with the goal of integrating these individual data marts at some point in the future. This approach generally provides near-term return-on-investment to the individual subject areas, but is also characterized by integration difficulties as the data marts are incorporated into an enterprise data model.

Hybrid Implementation

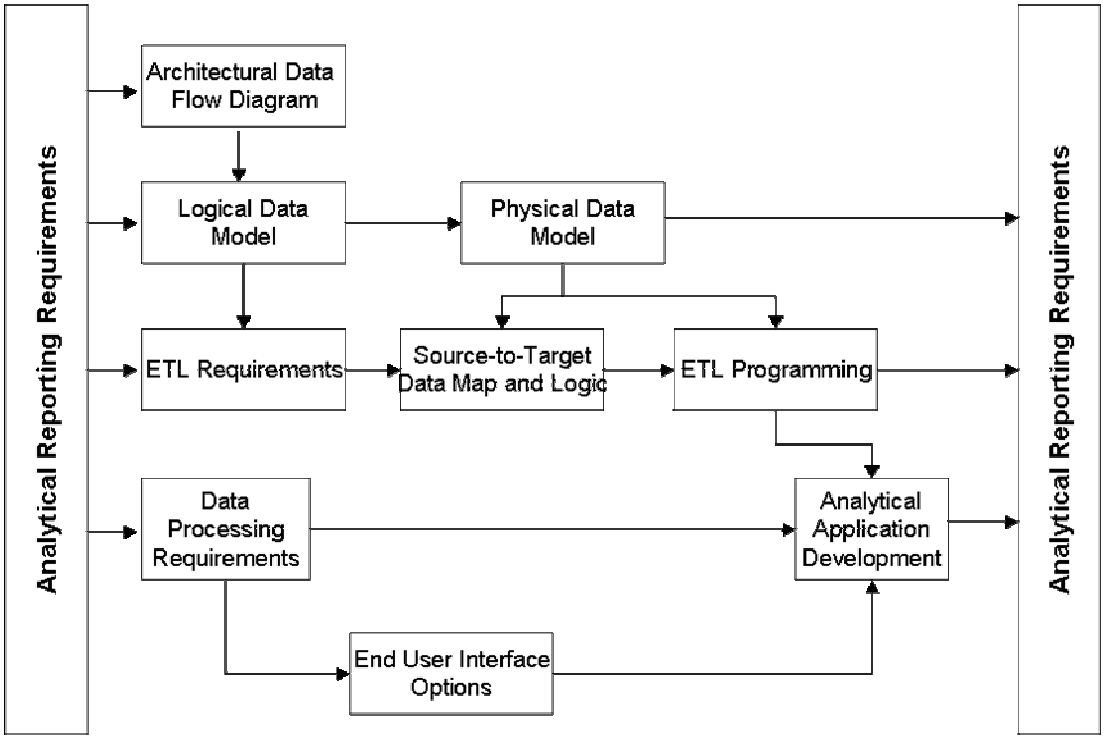
This approach is characterized by a focus on near-term development of data marts, but under a reasonable framework of enterprise standards to facilitate long-term integration and supportability. The greatest area of risk under this option is the deployment of data marts in parallel with the development of enterprise data standards, and the potential for conflict between the two.

Under this strategy, data marts are constructed first, to achieve integration and improve decision support within a specific subject area. In parallel to the construction of these data marts, opportunities are identified for data integration and decision support across the subject area data marts.

This strategy maintains the granularity of the data in the data marts and allow the analysts to decide which version of the “truth” they would prefer to use and when. Under this strategy, there are two types of data marts—(1) Data marts that reflect source systems, and (2) Data marts that are comprised of extracts from the source data marts. In either case, the general definition still applies – a data mart is a subject-oriented subset of the EDW. The diagram below depicts this hybrid methodology, using Oncology as an example subject area.

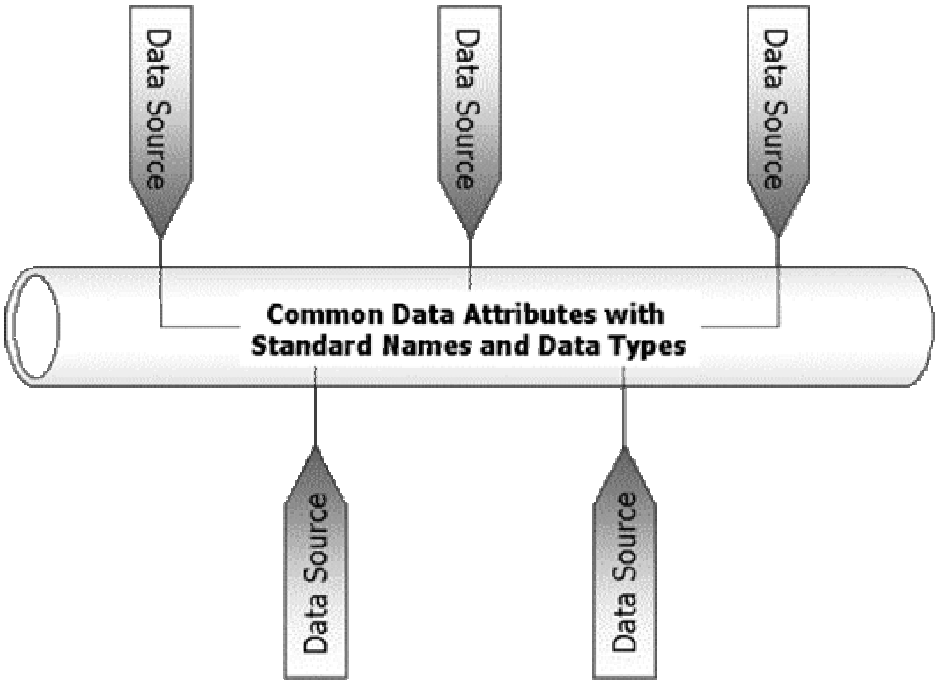


The diagram below depicts the flow of major deliverables and activities associated with the development of a data mart or data warehouse. (7)



Considering the top down aspects of the hybrid methodology, the most important issue is the standardization of data attributes that are common across the enterprise. These common data attributes, also called "core data elements" or "master reference data" by

some organizations, should be defined consistently across the EDW so that each data mart and/or data source can be mapped to this standard as it is loaded into the warehouse. It is this standardization, at the semantic and physical database levels, that will enable the analysts to link their queries across the various data marts in the data warehouse. Kimball et al use the term "data bus architecture" to describe this concept.[7] The diagram below depicts the concept of a data bus architecture—i.e., connecting data marts and other data sources in an EDW to a bus of standard core data elements that enables "communication" via joins across the different data marts.



Examples of common data attributes that comprise the bus architecture in an integrated healthcare delivery system are listed below. All of these common attributes are important, but a master patient identifier and master provider identifier are vitally important to an EDW. If these identifiers are not standardized in the enterprise, the data warehouse will certainly not achieve its full potential--for this reason, it is in the interests of the EDW Team to champion their implementation and standardization.

- Postal Code
- Payer/Carrier Identifier
- Patient Type
- Patient Type
- CPT Code
- Department Identifier
- Gender
- Provider Type
- DRG Code
- Region Identifier
- Age Group
- Race Master

- Patient Identifier
- Medicare Diagnosis Code
- ICD9 Diagnosis Code
- Facility Identifier
- Provider Identifier
- Marital Status
- ICD9 Procedure Code
- Employer Identifier
- Encounter Identifier
- Outcomes Master
- Charge Code
- Employee Identifier

Data Modeling in an EDW

As discussed earlier, in a purely top down implementation strategy, an enterprise data model is constructed first, and then loaded with data from the source systems. Data marts are then extracted from this enterprise. In theory, this is an appealing strategy, but in practice it proves to be too complex and slow to deliver results, for two fundamental reasons—

(1) Creating an enterprise data model is nearly impossible for a large organization, especially in healthcare. The HL7 Reference Information Model is the best example available today of an enterprise data model for health care, but it has its limitations and shortcomings, too. In addition, the HL7 RIM is more reflective of a transaction-based information system, not an analytical system. To function well in the analytical environment of an EDW, the HL7 RIM would, as a minimum, require significant denormalization. Nevertheless, it serves as an excellent reference and theoretical goal and should not be overlooked.

(2) The complexity of loading an enterprise data model with data from the source systems is enormous. Consider that the source systems contain overlapping data concepts—e.g., diagnosis. These overlapping concepts are, many times, completely valid, i.e., the billing department may have a valid reason to code the diagnosis slightly differently than that coded by the provider in the medical record. Loading an enterprise data model would require the data warehouse team to choose which version of the "truth" for diagnosis to load into the enterprise model, or at least provide a way to identify the issues involved in the overlapping concepts.

Star Schemas and Other Data Models

Fundamentally, a third normal form data model best represents a business or clinical environment, but these 3NF data models are not the best models to support analytical processing. In the mid 90s, Ralph Kimball popularized the star schema [(11)], which is now

the de facto standard in data warehouse models. However, the star schema does not reflect the true data environment as well as a traditional data model and, in fact, is more restrictive on analysis than other more traditional data models. In general, a strategy for modeling that frequently succeeds is based on designing a standard 3NF data model that represents the business or clinical area that is the topic of analysis. Then denormalizing this model to facilitate analytic processing, keeping in mind that star schemas are just another method for denormalization. Do not rush to the assumption that star schemas are the best and only solution to your modeling challenges in a data warehouse. They represent only one of several options.

Data Security

As a consequence of the centralized nature of the EDW, the potential for security compromises is enormous, especially if analysts are allowed unrestricted access to the base data in the EDW through command line SQL or desktop query tools that allow data to be downloaded to local desktop computers. In spite of this risk, following a principle of trust is best—trust and empower the analysts and end users with more access to the EDW, rather than less, while holding them accountable for properly handling patient and confidential company data.

During the design and implementation of data marts, any information that can directly identify a patient/member should be physically segregated in the EDW, or logically separated with database views, from confidential clinical information. Access to this identifying information should be strictly controlled with a formal justification process and periodic verification that access to patient identifiable data is still justified. In addition, access to patient identifiable information must be audited to track who accessed the data, the date and time the access took place, and the nature of the query that accessed the data.

In general, security can be implemented at two layers within the EDW architecture. Those layers are:

- Database layer: This layer uses the security features of the database to restrict, grant, and audit access based upon user roles. The data stewards are usually responsible for defining the requirements of the security roles and the database administrators are responsible for implementing and maintaining these roles.

Database attributes may also be used to implement security schemes. For example, if a database table contains a column that identifies the record as belonging to a specific facility, a condition may be added to queries that map the user to a facility. When a query is submitted, a condition is added to the query, which limits returned data to data in the user's facility. *Organize and plan database roles carefully and deliberately. Make certain they are logical, sensible, and manageable and reflect the types of analysts that will be accessing the EDW.* Defining too few roles will not allow for adequate security, yet too many roles will become a confusing and difficult to manage, causing confusion for the EDW Team and analysts.

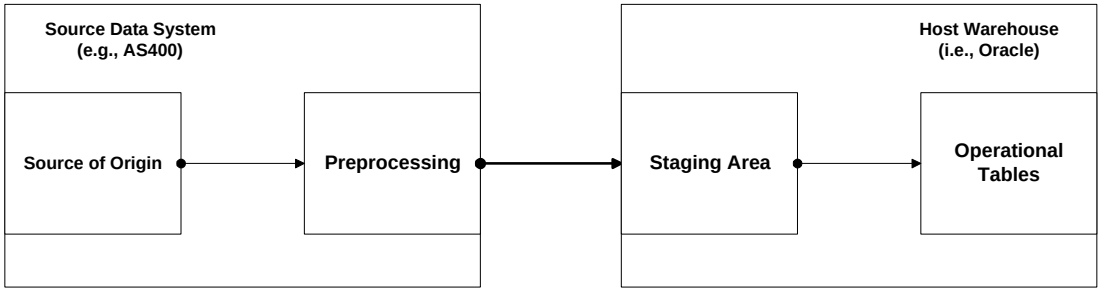
- Application layer: The application layer generally refers to either the desktop query and reporting tool or the web application that is used to access and query the EDW. Business Intelligence (BI) tools possess their own security layers that control access to reports that are published to their directory structures. The strategy for applying these security layers should consider the relationship they have with the roles in the database layer. For example, it would be contrary to allow an analyst or customer access to patient identifiable reports published through a business intelligence tool, while denying similar access rights through the database layer. The directory structures provided by BI tools is also an indirect but important aspect of the EDW's security strategy. The primary purpose of these directory structures is to facilitate the organization and "findability" of reports, but their secondary purpose is certainly security related. Organize the directory structures so that they are also integrated with the strategy of the database roles.

The Lightweight Directory Access Protocol (LDAP) standard is an excellent technology for achieving centralized, role-based security that integrates database and application level security. Business intelligence tools, databases, and web applications in the EDW architecture should take advantage of LDAP's capabilities.

Architectural Issues

The EDW architecture is generally designed to be a read-only data source from the analysts' perspective. Because of the costs of developing and maintaining real-time interfaces, batch interfaces are usually the preferred architecture for populating the EDW, but near-real time updates will probably evolve into a genuine requirement over

the lifecycle of the EDW, so plan accordingly and avoid being surprised—analysts have a growing appetite for the most timely data possible. As a general rule of thumb, the data that populates the EDW should be obtained from a source that is as close as possible to the point of origin for the data—avoid the dependence on intermediate information systems, to supply the EDW when possible. When possible, preprocess your data on the source system because the data is usually most easily manipulated first in its native environment. But, preprocessing on the source system can also have a negative impact on the performance of the source system; if the source system is a production-oriented transaction system, this negative impact can have serious political consequences for the EDW. Preprocessing can also take place within the host data warehouse environment, but preferably in a manner that does not impact the operational response time of the warehouse. A staging area within the warehouse environment should be used for final transformation and quality assurance of the data prior to being loaded into the operational tables. The diagram below depicts this approach.



The EDW is generally designed to function behind the firewall for Intranet and LAN/WAN access only; however, there are emerging requirements in many companies to publish reports from the EDW to an external Internet server. Any processes to transfer data from the EDW to an external Internet server should be accomplished behind the firewall.

Data Quality

Assessing data quality in an objective manner is and will continue to be very complicated; it is inherently subjective. However, a rather elegant algorithm is as follows:

$$\textit{Data Quality} = \textit{Completeness} \times \textit{Validity}$$

Where:

- Completeness is a measure of the robustness and fullness of the data set. It can be objectively measured by counting null values.
- Validity is an inherently subjective measure of the overall accuracy of the data—how well does the content of the data actually reflect the clinical or business process in which it was collected?

The principle of data quality that applies to an EDW is fairly simple: "Use the EDW as a tool for improving data quality *at the source of the data*." The purpose of the EDW is not to improve data quality, per se, though an EDW can facilitate improvement of data quality at the source system. The real purpose of the EDW is to improve access to, and the integration of, data. Contrary to many popular opinions, this principle implies that you should *avoid* extensive "data scrubbing" as part of the EDW operational processes. Data scrubbing at the EDW level tends to treat the symptom, not the underlying cause. The cause of poor data quality usually resides with the source system or the data entry processes that surround it. Also, "data scrubbing" can take on many forms and quickly become a quagmire, both technically and politically.

Another key principle related to data quality and the role of the EDW is, "The EDW shall not lower the quality of the data it stores as a consequence of errors in the EDW extraction, transformation, or loading (ETL) processes." There are many opportunities in the ETL processes of the EDW for inadvertently introducing errors in data—and nothing can be more damaging to the image and reputation of the warehouse than these errors. It is imperative that the EDW Team use extensive peer and design reviews of their ETL processes and code to identify problems before they become problems.

Below are the most common sources of data quality problems in a data warehousing environment.

- Calculation errors (i.e., aggregations, calculations)
- Code translations incorrect (i.e., 1 should be translated to 'M' which equals 'Male', but was translated to 'A')
- Data entry transposition errors (0 vs. O, etc.)
- Data homonyms (same or similar attribute names for different types of data; e.g., Diagnosis code has several different meanings)

- Data mapping errors (i.e., values inserted into the incorrect column)
- Data types mismatched
- Domain constraints violated
- Duplicate records
- Incorrect use of inner and outer join statements during ETL
- Parsing errors
- References to master tables fail
- Referential integrity violations (i.e., a record in a child table which should not exist without an owning record in a corresponding parent table)
- Required columns are not null
- Row counts incorrect
- Data synonyms (different attribute names for the same type of data, e.g., SSN vs. SSNum)
- Truncated fields

An interesting and sometimes unexpected fringe benefit of data warehouse projects is the subsequent, overall improvement of data quality within the company. The publicity and visibility of data errors increases in an integrated EDW environment and the unpleasant consequences of poor data quality also increases. As a result of this phenomenon, the overall motivation to "clean house" and improve data quality in the enterprise increases significantly after the deployment of a successful data warehouse.

Return on Investment

ROI concepts should be applied to the overall business benefits of the EDW, but also to the strategy of development for the EDW; i.e., the data that provides the highest value to the analytical goals of the company should be targeted first for data marts. Determining which data to target as candidates for inclusion in an EDW is typically a challenge for most organizations. The subjective algorithm below provides a framework for approaching this problem.

$$ROI\ of\ EDW\ Data\ Content = \left(\frac{Business\ Value}{Human\ Resource\ Costs + Computing\ Costs} \right) X\ \ Data\ Quality^{*}$$

$$^{*}\ Data\ Quality\ =\ Completeness\ x\ Validity$$

The Business (or Clinical) Value of the data can be assessed by quickly identifying the major sources of transaction data available in the enterprise. In most healthcare organizations, it boils down to systems such as lab, radiology, pharmacy, electronic medical records, finance, materials management, and hospital case mix, et al. These core transaction systems represent the vast majority of the knowledge capital available in the enterprise, from a database perspective, and should be targeted first. The significant deviation in the above algorithm from a standard ROI is the Data Quality variable. Targeting a source system for inclusion in the EDW that possesses low data quality should only be executed if it is a deliberate attempt to improve the quality of data in that system. Clearly, if the data quality for a source system is low, its business value will probably be low.

Measuring *Return* On Investment for an EDW is a difficult endeavor, but that should not deter organizations from deliberately managing and tracking their investment. According to a 1999 Cutter survey [(15)], 17% of companies try to measure ROI for data warehouses; and 48% of these fail completely or give up. This same report found that companies that did conduct an assessment, reported an average ROI for a data warehouse of 180%.

Metadata

Metadata is information about data—Where did it come from? Who generated it? Over what period of time is the data effective? What is the clinical or business definition of a particular database column? The value of metadata to the success of the EDW increases geometrically as the number of data sources and users increases—it could very well be the most strategic, up-front investment to ensure the success of an EDW.

One of the fundamental goals of an EDW is to expose the knowledge of an organization, horizontally, across the organizational chart. Typically, analysts and end users understand the transaction systems that support their vertical domains, very well. In these cases, metadata is not as valuable to an organization because the end users already understand their data. Metadata's true value is realized in horizontal fashion, when analysts in finance use clinical data to better understand the relationships between costs and outcomes, for example. To achieve the vision of an EDW, a metadata repository is absolutely fundamental. No data should be deployed in an EDW without its accompanying metadata. Unfortunately, vendors, especially those associated with ETL tools, have not provided an effective, reasonably priced solution to this problem; therefore, the most effective metadata repositories continue to be "home grown" and will be for the foreseeable future.

Meta reports

Another form of metadata is that associated with the reports generated from the EDW, i.e., metareports. These metareports provide information about the reports themselves and accompany the results of the report, itself, as a cover sheet. The metareport includes information such as:

- Natural language question that the report is answering; e.g., "What is the percentage of patients that received a pre-op biopsy before a definitive surgical procedure?"
- Source(s) of the data: The names of the data marts in the EDW supporting the analysis; e.g., Cancer Registry, Hospital case mix, and Pathology Data Marts. The specific tables and columns in these data marts are also listed, as well as any temporal issues associated with the data.
- Formulas used in statistical calculations and aggregations
- Overall assessment of data quality (Description of completeness and validity)
- Selection criteria used in the query, including temporal criteria
- Names of those involved in the creation and validation of the report
- Date that the report was declared "valid"

Case Study

Intermountain Health Care is an integrated delivery system (acute care, ambulatory clinics, and health plans) headquartered in Salt Lake City, UT. IHC's delivery area is Utah and southern Idaho. In 2000, IHC had 434 thousand patient days in its 22 hospitals, and 5 million outpatient visits, including those at the ambulatory clinics. Total funds available in 2000 were $1.9 billion. IHC employs 22,000 people.

Intermountain Health Care's Enterprise Data Warehouse was deployed as a prototype in 1996, using acute care case mix data. The motivation of the project was two fold: (1) Test the ability to extract data from AS400-based databases and enhance its analytic availability by loading this data into an Oracle database; and (2) Test the ability to develop a web-based interface to this data to support analysis and metadata management. The prototype was developed primarily by a graduate student in medical informatics, with part time assistance from an Oracle database administrator and an AS400 programmer. The prototype was generally considered a success, though it did experience two significant problems that set the project back politically and technically. The ETL programs were very inefficient and error prone, requiring up to 10 days to successfully load the only data mart in the prototype. The ETL processes were also not well validated and introduced significant errors into the data and, as a consequence, end users lost confidence in the quality and reliability of the data. Finally, the EDW server experienced a disk failure that destroyed most of the scripts and database structures and, unfortunately, no backup existed, so the prototype was rebuilt almost from scratch. Despite these hurdles, the prototype EDW received the Smithsonian Award for Innovative Use of Health Care Information in 1997. This award contributed significantly to the internal political support necessary to move forward with a more formal development project.

In a recent study, The Data Warehousing Institute reported that 16% of the 1600 companies surveyed felt that their data warehousing project exceeded their expectations, 42 percent felt that it met their expectations, and 41 percent reported that they were experiencing difficulties. In a recent customer survey at IHC, 89% reported

that the IHC EDW met or exceeded their expectations for supporting their analytic needs. The success of IHC's EDW is largely a reflection of the quality of the transaction systems supplying the EDW. IHC has achieved significant standardization of their core transaction systems, both technologically and semantically, across the enterprise. They also possess a widely implemented master patient identifier and provider identifier. In those cases in which a master patient identifier (MPI) is not available, IHC uses a heuristic matching tool, MPISpy, which matches demographic data to the MPI. Today, the EDW is considered a critical component to achieving IHC's vision of optimum health care quality at the lowest reasonable cost. The IHC EDW contains 1.1 terabytes of storage and 2.1 billion records on a twelve processor IBM Raven server running AIX and Oracle 8i. It supports 50,000 queries and delivers 1.5 billion records per month. Twenty-seven different sources of data supply the EDW. It is supported by 19 FTEs, who are funded by a combination of corporate resources, and individual departments with specific analytic needs. There are 2,250 tables in the Enterprise Data Warehouse. The total investment in information technology and IT staff over the past five years is $11M.

Analytic Examples and Benefits

In a recent attempt to count the number of reports that are regularly generated from IHC's EDW, the inventory stopped at 290, in part because it was difficult to define a "report" and in part because the labor effort required to conduct the inventory was much greater than expected. In less than four years, the EDW evolved from a system that generated a handful of prototype reports to a system that generates literally hundreds of reports supporting critical clinical, business, and regulatory requirements. The high-level types of reports generated from the EDW mirror the structure of IHC-- Health Plans, Health Services, and Internal Operations. Health Services encompasses the operations of acute care hospitals, ambulatory clinics, and homecare.

The Health Services related reports include:

- Quality management—mortality rates, surgical infection rates, prophylactic antibiotics, c-section rates, restraint rates, adverse drug reactions, unplanned readmissions, unplanned return to surgery, etc.
- Joint Commission/Oryx reporting

- Clinical goals (adherence to standard protocols) in focus areas such as cardiovascular, diabetes, asthma, pneumonia, women and newborns, neuromuscular, and oncology,
- Length of stay
- Cost per case

The Internal Operations Reports include:

- Charge correction ratios
- Claims edit ratios
- Co-pay ratios
- Average accounts receivable days
- Lag days by provider
- Invoices without a co-pay
- Data quality reports for a variety of source systems that supply the EDW
- Patient Perception of Quality
- Bad debt
- Materials management—backorders, daily transactions, fill rate, fill/kill, invoices waiting, item usage, invoices in waiting, open purchase orders, price variance, etc.

The Health Plans related reports include:

- Claims analysis by a variety of dimensions—by revenue code, diagnosis code by date range, by DRG, by procedure code, etc.
- Broker sales and management
- Enrollment management
- Health needs appraisals
- Medicaid management
- Cost per member per month
- Underwriting management
- Member management—claims volumes, enrollment, disenrollment, etc.
- Provider operations—Group affiliation, practice type, financials, cost per case, etc.

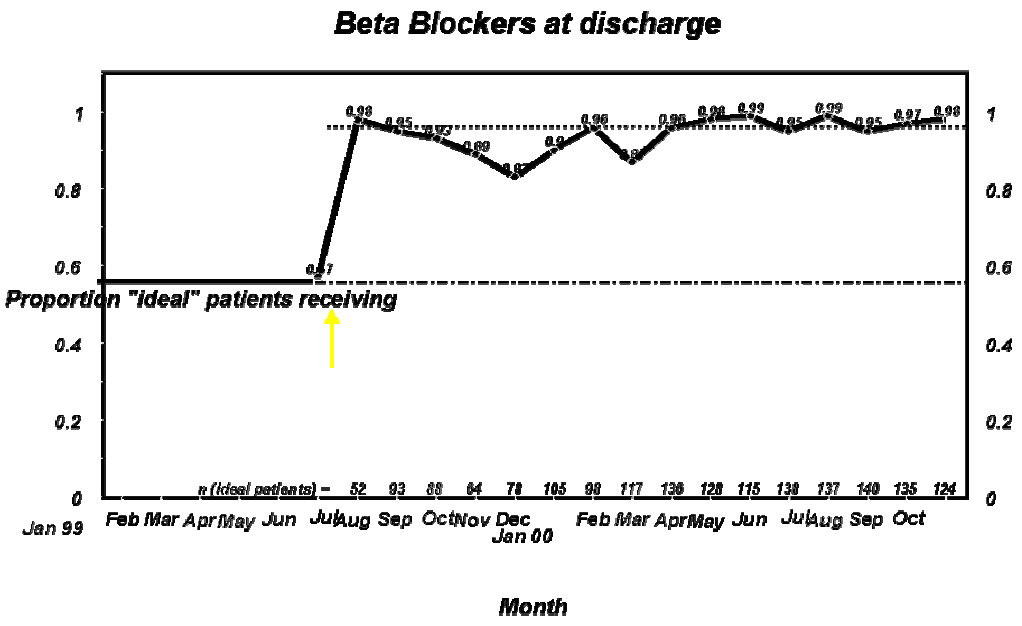
An interesting but not surprising area of ROI for an EDW is that connected to the labor expended by analysts to generate reports. In a 1998 study, IHC (Dr. Diane Tracy, et al;

not published) determined that, prior to the deployment of the EDW, 50% -90% of a Quality Management analyst's time was expended on data collection and integration, leaving only 10% - 50% of their time to facilitate continuous quality improvement. After the deployment of the EDW, the labor split was reversed—the analyst's were able to spend 50-90% of their time on process improvement and behavioral change.

The benefits derived by IHC from these reports are so numerous, it is difficult to choose the "best" examples. Suffice to say that the steady growth in funding committed to the IHC EDW is an indicator that the company is convinced of its clinical and business value. A few examples of the benefits are discussed below.

Ambulatory Clinic Management: Since the deployment of the IDXExtend Data Mart, IHC's ambulatory clinics' "Average Bill Days Outstanding" was reduced from 80 to 43 days. This equates to literally millions of dollars in cost savings per year and contributed significantly to turning the employed Physicians Division into a profitable business area for IHC.

Cardiovascular Clinical Program: After deployment of a dedicated data mart, the Cardiovascular Clinical Program realized significant improvements to standard protocols. For example, Beta Blocker administration upon discharge (CAD w Acute MI) increased from 57% adherence to the protocol to 97% adherence, as indicated in the chart below.



The overall summary of adherence to the CV Discharge Medications protocol is summarized below and illustrates a significant reduction in mortality—almost 1400 more people are alive today as a consequence of the commitment to behavioral change and continuous quality improvement.

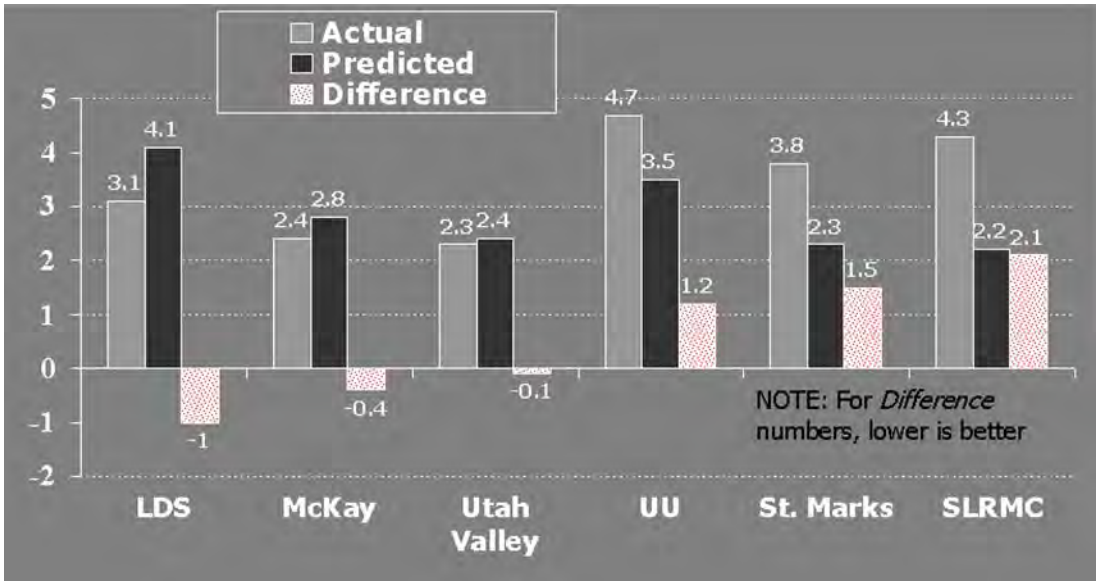
Cardiac discharge meds

	Before	After	National 2000
Beta blockers	57%	91%	41%
ACE / ARB inhibitors	63%	94%	62%
Statins	75%	95%	37%
Antiplatelet	42%	99%	70%
Wafarin (chronic AFib)	10%	90%	<10%

	Mortality at 1 year			Readmissions w/ in 1 year		
	Before	After	Lives	Before	After	Lives
CHF (n = 19,083)	22.7%	17.8%	331	46.5%	38.5%	551
IHD (n = 43,841)	4.5%	3.5%	124	20.4%	17.7%	336
Total			455			887

Based upon data collected by the State of Utah for public dissemination, the quality of the Cardiovascular Clinical Program at IHC's facilities (LDS Hospital, McKay Dee Hospital, and Utah Valley Hospital) is considerably higher, in terms of mortality rate, than those of the non-IHC hospitals in the same market area, as illustrated below.

Death Rate – CV Surgery
Actual vs. Predicted (Year 2000)



Primary Care Diabetes Management: protocol for HPI members: After just 14 months after deployment, the IHC diabetic patients with HbA1c >9.5 decreased from 32.4% to 24.35%, as illustrated below. This data mart, which integrated data from five sources—the ambulatory clinic billing system (IDXExtend), the health plans claims system (internally developed), the clinical data repository (3M Care Innovation Suite), hospital case mix (internally developed), and laboratory results (Sunquest)—was largely responsible for IHC receiving the National Award for Exemplary Health Care Service in Diabetes Patient Management from the National Association of Health Plans.

Performance Measurement	1999	2000	2001
Annual HbA1c*	78.55	83.0%	90.0%
HbA1c > 9.5	34.6%	32.4%	24.35
HbA1c less than 8	59.4%	64.1%	67.3%
Biannual LDL*	65.9%	73.7%	85.2%
LDL less than 130*	39.9%	49.6%	61.3%
Annual Eye Exam*	52.0%	47.9%	56.0%

In addition to the inherent improvement to the quality of life for the patients under the diabetic management program, there were real reductions in cost associated with delivering their healthcare, as summarized below. Approximately 7% of IHC's patient population is diabetic.

Cost impact of diabetes program

Net cost savings per Type II DM patient:

Year 1	*$ 93*	*(loss)*
Year 2	*1*	*(loss)*
Year 3	*764*	
Year 4	*1493*	
Year 5	*2269*	
Year 10	*2688*	
Year 15	*2864*	

Total over 15 years: $ 30,927 per patient

(assumes ~2 point fall in A1c as a result of diabetes management program; impact calculated from medical costs of complications avoided)

Demers *et al.* Computer simulated cost effectiveness of care management strategies on reduction of long-term sequelae in patients with non-insulin dependent diabetes mellitus. *Quality Mgmt in Hlth Care* 1997; 6(1):1-13 (Fall).

Vision, People, Processes, Technology, Strategy, and Metrics

Achieving success with any information system, including an Enterprise Data Warehouse, requires a strategy that encompasses Vision, People, Processes, and Technology, and each of these must be related to a sense of metrics and measurement. Below is a discussion of each, within the context of IHC's EDW.

Vision

The vision of any EDW is data synergy—integrating disparate data so that their combined effect is greater than the sum of their individual effects. The vision of IHC's EDW is to, "Facilitate a standardized analytical understanding of IHC's clinical outcomes, business costs, and insurance services by providing a centralized analytical processing system and the information technology services to support it." Functionally, one aspect of the vision of the EDW is, "The analyst should never have to query the EDW for anything; it should tell them what they want to know, proactively." This implies providing the ability for EDW analysts to define their analytical needs (reports) and thresholds (alerts) within a logic layer that required little or no intervention after set-up. It also implies the ability for data mining tools to run against the EDW in the background with little or no human oversight, looking for new patterns in data, and then alerting analysts to these patterns for further investigation. The vision includes real-time feeds from all source systems supplying the EDW that will enable immediate detection of aggregate trends such as epidemics and bio-terrorism attacks. The vision includes the concept for a single web-enable portal for easily accessing and navigating all the analytical reports generated from the EDW. The fundamental assumption in all these aspects of the vision is, by integrating data in the EDW and thereby creating a more complete view of patients, members, providers, and the processes that surround them, IHC can find the optimum point between highest quality of care and services at lowest reasonable cost to the community and people it serves. This vision is shared by IHC's senior management, which has a long history of embracing information systems and data analysis supporting continuous quality improvement. *The success of the EDW vision is fundamentally based on the foundation of this senior level support.*

People

In IHC and healthcare in general, there are four critical roles and skill areas required to successfully deploy an analytical application; i.e., a valid report that can affect continuous quality improvement:

Analytical Role	Description
Business and/or clinical leader	Understands the process that is targeted for improvement or management, and can influence the process improvement. IHC's success with analytics is directly attributable to the assignment of clinical and business leaders to the analytical environment who are recognized by their peers as influential leaders.
Data manager/steward	Understands the detailed issues surrounding the content and quality of the data that is being used to support the analysis
Statistical data analyst	Expert with "valid" data analysis and presentation techniques
Information technology staff	Data architects that design the data models necessary to support the analysis and programmers that can implement the query and reporting requirements defined by the other three roles.

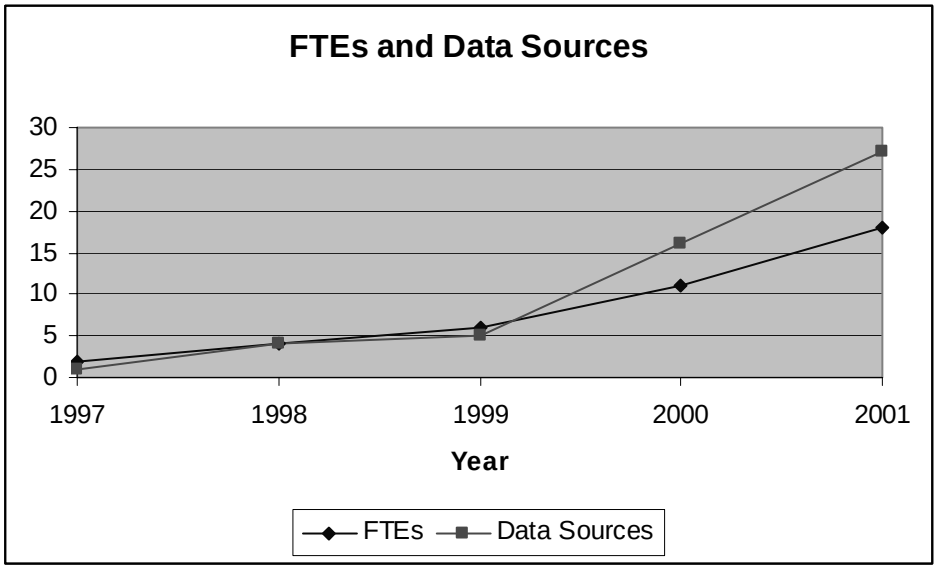
The importance of these four roles increases proportionately with the significance and risk of the decisions being affected by the data analysis. For low-risk decision-making environments, it is common for one person to fill all four roles. For high-risk decision-making environments, IHC typically fills these roles with four individual experts that function collaboratively in check-and-balance fashion. The participation of these four roles is also critical in the political acceptance of the data analysis; i.e., the perception that the analysis is accurate and believable increases significantly if all four of these roles participated in the analysis and endorsed the results.

Deploying a successful information system depends on a combination of information technology expertise, and a fundamental understanding of the domain supported by the information system. IHC's EDW Information Technology team reflects this principle. It is composed of staff members with backgrounds that are heavy in information systems and computer science, along with members who are experienced in the various business and clinical areas supported by the EDW and subsequently cross-trained into information systems or medical informatics. This balance of information technology skills and domain skills creates a complementary environment that results in better EDW-based solutions. In addition to the general balance between information systems skills and domain skills, the EDW IT team is staffed with the following specific skill sets:

- Systems architect: Overall IT architectures associated with data warehouse environments—servers, storage, applications, databases, processes, etc.
- Data architect: Analytical data modeling, physical database design and implementation, SQL programming, etc.
- Programmer/Analyst: Programming web applications, SQL, business intelligence tool, and ETL scripts
- Database administrator: Tuning and operations of the relational databases supporting the EDW
- Systems administrator: Server, storage system, and operating system administration

In the early stages of the EDW lifecycle, the skills emphasis was on analytical data modeling for the EDW, extracting data from the source systems, and loading the EDW. As the EDW matured, emphasis shifted from loading data into the EDW to extracting information from the EDW—i.e., the development of analytical reports for the web and in Crystal Reports. Also, as the EDW grew in volume and complexity, database and system administration skills became more and more critical.

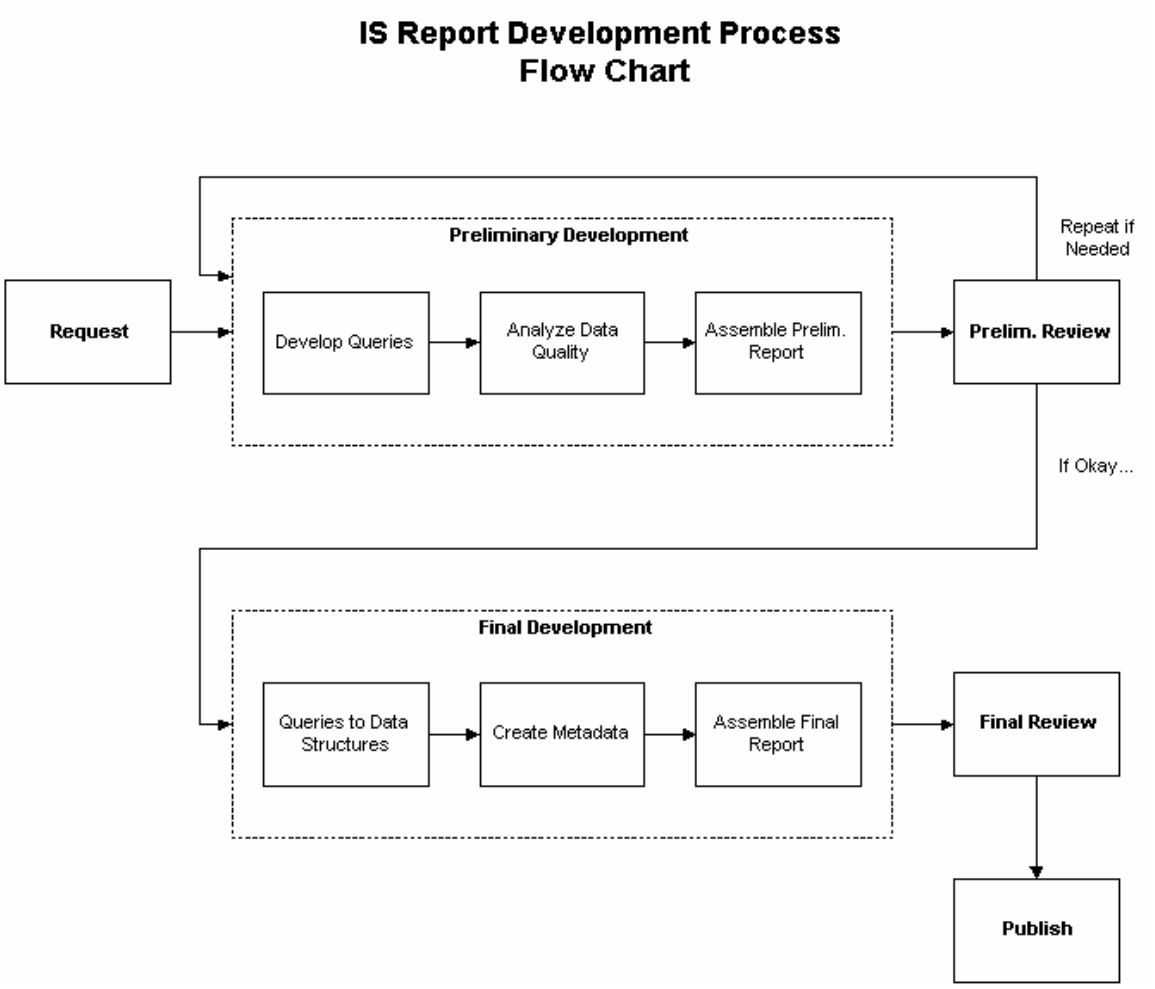
Below is a graph that depicts the growth in FTEs and the number of data sources supplying the EDW, over time.



Processes

The IHC EDW Team follows the overall methods and processes described earlier that are, in general, common to the data warehousing industry. In addition, the EDW Team places significant emphasis on peer reviews, especially in the early stages of the lifecycle of their data mart development efforts.

Another area of specific emphasis on process is that placed upon the development of reports. The diagram below illustrates the overall process.



This process focuses on the iterative development of reports, with a feedback loop that keeps the report in "preliminary" phase until the report is deemed appropriate for final development. Note that one of the final steps in the process is the documentation of the metadata associated with the report—the metareport.

The fundamental principle that drove the process of development was "Data marts first, data warehouse second." Below are the data marts that comprise the IHC EDW and their initial operational dates. Each of these data marts was developed with two strategies in mind: (1) Enhance the analytic requirements of the vertical business or clinical area that depended on the associated transaction system; and (2) Design the data mart in a manner that would facilitate integration with other existing or future data marts, via the standard data bus architecture. The decision regarding the order and priority of development for each data mart was also affected by two strategies: (1) Overall ROI of the data content as described earlier (cost/benefit of development plus consideration of data quality); and (2) The willingness of the business or clinical areas that "owned" the transaction data to participate in the development; some areas were very willing, others perceived the EDW as a threat or saw no strategic value to them by participating.

Data Mart	Analytic Function	Operational Date
Health Needs Appraisal	Support care managers and their high risk and Medicare patients	1Q 1997
Hospital Case Mix	Integrates data from all of IHC's 22 hospitals.	3Q 1997
Patient Perception of Quality	Summary of surveys assessing patient perception of quality associated with hospital inpatient encounters	4Q 1997
Cardiovascular	Supports goals of the CV clinical focus area, Cath Lab research, and reporting for Society of Thoracic Surgeons, and National Registry for Myocardial Infarction	1Q 1998
Labor	Labor and delivery encounters; sourced from case mix, clinical data repository, and labor and delivery monitoring systems	1Q 1998
Newborns	Customized extract from case mix to support Newborn clinical focus area	3Q 1998
HELP Radiology	Radiology management for IHC's urban hospitals	2Q 1998
IDXExtend	Ambulatory clinic billing system for all of IHC's 85 clinics	4Q 1998
Materials Management	Purchase Order management and processing	3Q 1999
Women & Newborns	Data relating to the Women and Newborn Clinical Program. Includes subsets of the Case mix data for Labor and Delivery, and for Newborns; Storkbytes data (Labor and Delivery); NICU EMR; and EVOX (NICU dataset – Extended Vermont Oxford).	3Q 1999
Health Plans	Claims, members, and plan groups	4Q 1999
LAN Desk	Desktop computer inventory management for all of IHC's 18,000 PCs	1Q 2000

Data Mart	Analytic Function	Operational Date
Cancer Registry	Integrates Cancer Registry data from 6 servers across IHC; used to support IHC's oncology clinical focus area	2Q 2000
Primary Care Diabetes	Integrates data from five different sources to support diabetic patient management. Represents the ability to create new knowledge and understanding through data synergy.	2Q 2000
Oryx	JCAHO Reporting	1Q 2001
HELP Pharmacy	In-patient pharmacy orders analysis	2Q 2001
Pharmacy Decision Support	Pharmacy claims management	2Q 2001
Lab Results	Integrates Sunquest data from 5 servers across IHC	3Q 2001
Pathology	Integrates Tamtron data from 7 servers across IHC	3Q 2001
Mammography	Integrates Mammography data with other cancer-related data sources to support the oncology clinical focus area	4Q 2001
Primary Care Asthma	Customized extract from Health Plans Data Mart to support Asthma patient risk management	4Q 2001
Quality Management	Quality Management reports supporting the hospitals	In Development
Radiation Therapy	Integrates Radiation Therapy data with other cancer-related data sources to support the oncology clinical focus area	In Development
Clinical Data Repository	Extracts clinical data from IHC's electronic medical record data repository	In Development

Funding and Prioritization Process

The heart of the funding and prioritization process for the IHC EDW is the Analytical Services Council. The responsibilities of the ASC includes:

- Ensuring analytical consistency across the company
- Encouraging and supporting synergy among analysts and IS resources
- Supporting budget and resource allocation from an enterprise perspective
- Defining "core" standard reports, processes, and stewards
- Resolving conflicting analysis
- Influencing the requirements and design of transaction systems to support analytical requirements
- Defining priorities of analysis efforts

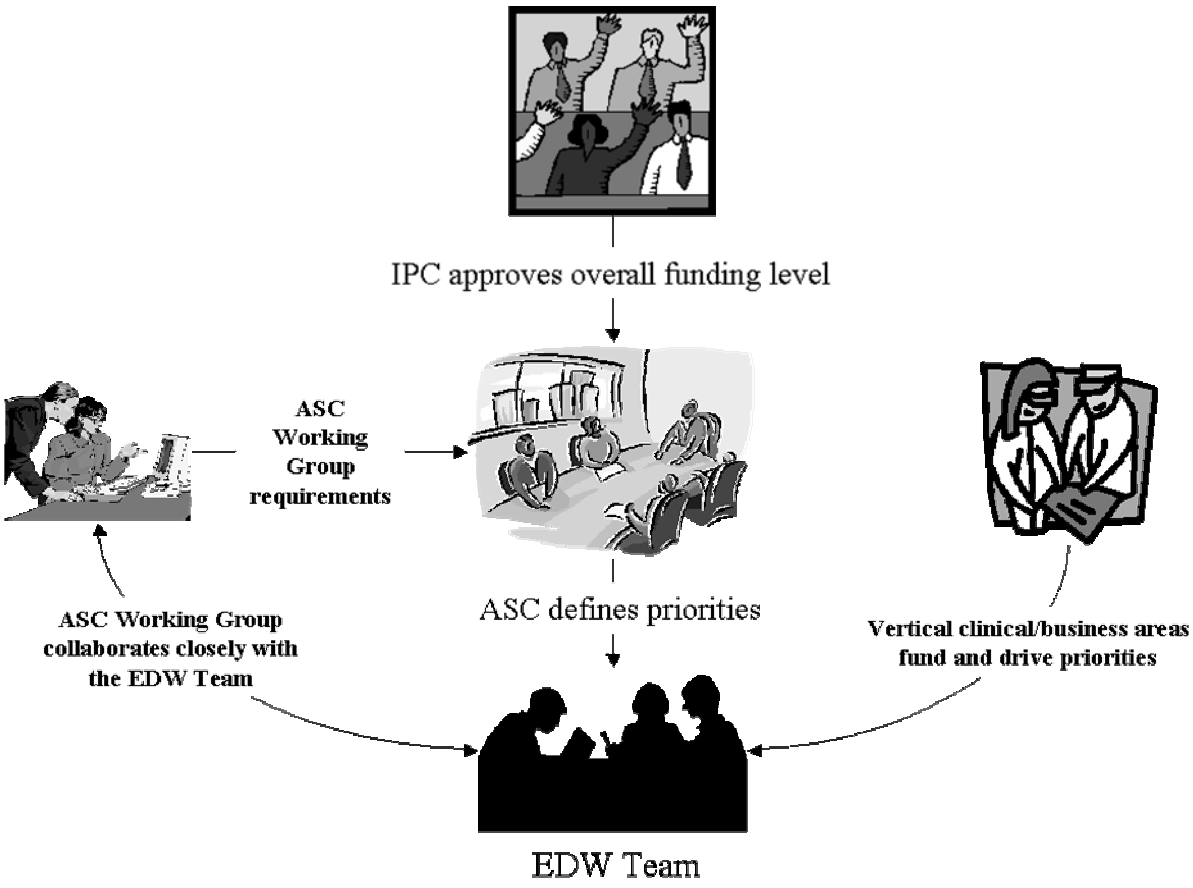
The members of the ASC are senior executives (Assistant Vice Presidents and above) from the following areas of IHC. The CIO is the voting representative from Information Systems.

- Campus-Based Care
- Community-Based Care
- Health Plans
- IHC Operating Regions
- Quality Management
- Strategic Planning (Chair)
- Clinical Programs
- Finance
- Human Resources
- Information Systems
- Shared Services
- The IHC Institute for Health Care Delivery and Research

The ASC meets for two hours each in September, January, March, and June. The general agenda is based on issues related to its core mission: Consistent, repeatable, dependable analytics. The agenda also includes relevant discussions of major project status; spending and value; education of new technology that can benefit the analytic mission of IHC; and potential projects for the next year. In June, requests for new funding are submitted (IHC's fiscal year is based on a calendar year and the budget preparation process begins in earnest in July). If no new funding is required, then justification and the intent for existing resources are reviewed. The ASC is accountable to IHC's Information Planning Council (IPC). The IPC is comprised of many of the same members of the ASC, thus facilitating consistent communications between the two groups. Final approval authority for major funding resides with the IPC. The ASC receives input for new projects and enhancements to the EDW from the ASC Working Group, which is comprised a senior data analysts from the same areas as those represented on the ASC.

Vertical business and clinical areas may also fund EDW-centric projects and applications. In these cases, the EDW Team functions in a pseudo-charge back mode by providing the analytical services required by the individual customer, and these customers define their priorities. This model works well by meeting the bottom-up needs of the vertical customer, yet doing so within the top-down context of the enterprise vision of standards for data integration and analysis.

The relationship between these bodies is illustrated in the diagram, below.



Organizational Alignment

IHC's EDW is aligned under the Senior Vice President for Medical Informatics, who in turn reports to the CIO. The IHC CIO reports to the CFO. The primary business sponsor and executive champion for the EDW is the Senior Vice President for Strategic Planning. This solid-line relationship of the IHC EDW with the executives in Information Systems and Informatics that can influence the transaction-based information systems that supply the EDW was a critical success factor. Without this influence, the data warehouse team lacked the leverage that is frequently necessary to engage the support of the transaction systems that frequently perceived the EDW as a threat. Over time, this perception faded, but in the early stages of the EDW, the perception was strong and serious.

Technology

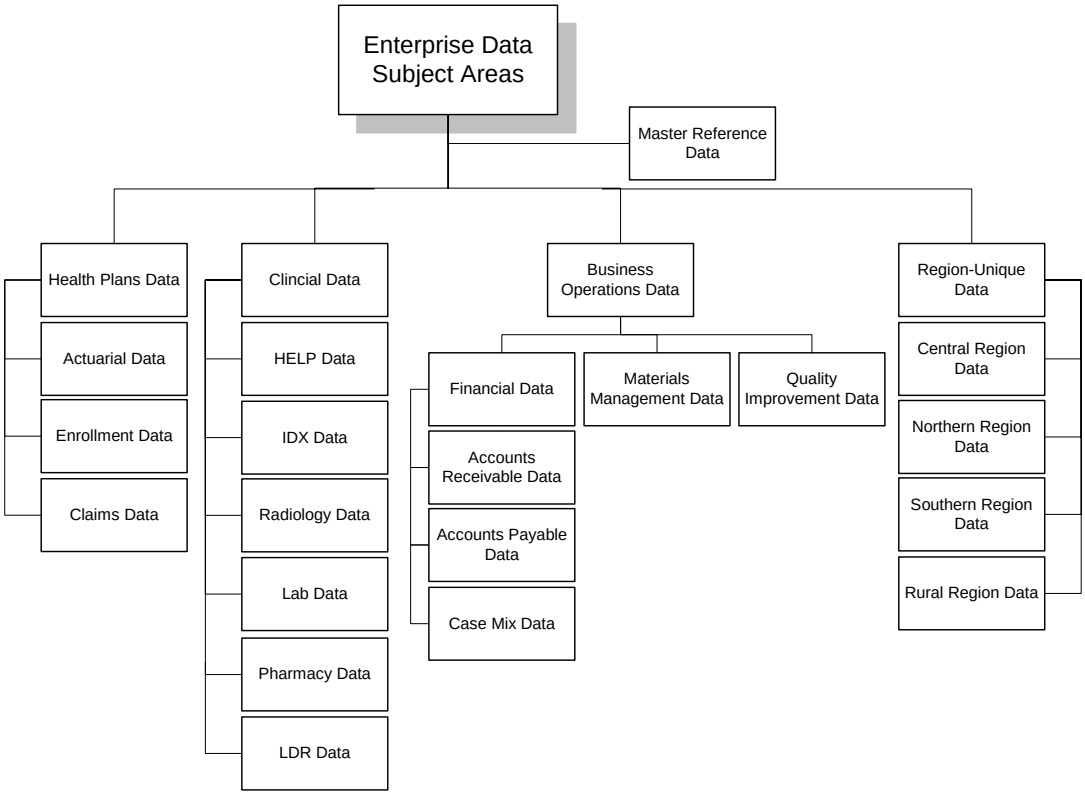
Hardware Architecture: During the prototyping phase, IHC's EDW was deployed on a massively parallel processor (MPP) architecture running Oracle Parallel Server (OPS) and fully mirrored disks. This MPP OPS architecture was retained during the transition to full-scale production. Although very scaleable and flexible, the MPP architecture proved highly complex to operate and tune, and unreliable under increased workloads and data volumes--fairly common characteristics of MPP OPS systems. This architecture was scrapped in favor of a Symmetric Multi-Processor (SMP) architecture with a less expensive RAID-5 storage system. This architecture proved much more reliable, supportable, and cost effective.

The chart below summarizes the current server architecture of IHC's EDW environment.

Server	Function	Architecture
EDW	Production server for the Enterprise Data Warehouse	IBM/AIX, 12 CPU, 300Mhz, 16G RAM, 1.2Tbyte disk SAN
ETL Server	Hosts the extract, transformation, and loading applications	Windows NT, 4 CPU, 550Mhz, 2G RAM, 128G disk
Reporting Servers	Hosts the business intelligence and reporting application	3 each, Windows NT, 4 CPU, 200Mhz, 1G RAM, 8G disk
Queue Server	Used as a "buffer" to cache near-real time feeds from the source systems. These real-time data feeds are stored on the queue server then periodically batch loaded into the EDW.	IBM/AIX, 1 CPU, 332Mhz, 512M RAM, 50G disk
Development Server	Used for development and testing of EDW-based applications. Also used to test database and operating system upgrades.	IBM/AIX, 1 CPU, 332Mhz, 512M RAM, 434G disk

Data Model: IHC's EDW follows the data bus architecture concept that "links" various data marts via semantically and physically standardized data dimensions, such as patient and provider identifier; diagnosis, procedure, etc.. The individual data marts each have their own data model, depending on the nature of the analysis that they support. In some data marts, the data model is very flat—very wide tables with very little normalization that require very few joins in analytical queries. In other data marts, the data model is fairly normalized—second normal to third normal form that can require 6 or more joins in a single query. In these more normalized data marts, indexing and the use of summary and pre-aggregated tables helps alleviate the potential for database performance problems and overly complex queries. Connecting all of these data marts

are the standards of the data bus architecture. At a very high level, the data model for the IHC EDW is depicted conceptually in the diagram, below. There are two key concepts captured subtly in the diagram that should be emphasized: (1) The physical schemas and design of the underlying data structures of the warehouse reflect this high level data model; and (2) The lines and connections between the blocks of the diagram symbolize the standard data types and naming that allows analysts to link data across the different subject areas.



Metrics

Since the essence of a data warehouse is the measurement of the clinical and business processes it supports, it would be ironic if the operations of the warehouse were not also subjected to the scrutiny of measurement. The IHC EDW Team emphasizes the collection of metrics on all things it manages—employees, data, information technology, projects, budgets, and customers. All of these areas of metrics are important, but the two most important metrics to the successful operation of the IHC EDW and its Team

are employee satisfaction and customer satisfaction. These two metrics are each gathered twice per year.

The employee satisfaction surveys appear in two forms—one sponsored by IHC for all employees and the other is unique to the EDW Team and based on a study by the Gallup [17] organization that identified 12 key questions that, in total, provide the best overall measurement of employee fulfillment in a work environment. Those 12 questions are:

1. Do I know what is expected of me at work?
2. Do I have the materials and equipment I need to do my work right?
3. At work, do I have the opportunity to do what I do best every day?
4. In the last seven days, have I received recognition or praise for doing good work?
5. Does my supervisor, or someone at work, seem to care about me as a person?
6. Is there someone at work who encourages my development?
7. At work, do my opinions seem to count?
8. Does the mission/purpose of my company make me feel my job is important?
9. Are my co-workers committed to doing quality work?
10. Do I have a best friend at work?
11. In the last six months, has someone at work talked to me about my progress?
12. This last year, have I had opportunities at work to learn and grow?

EDW Customer satisfaction is assessed by two basic questions, scored on a five-point scale:

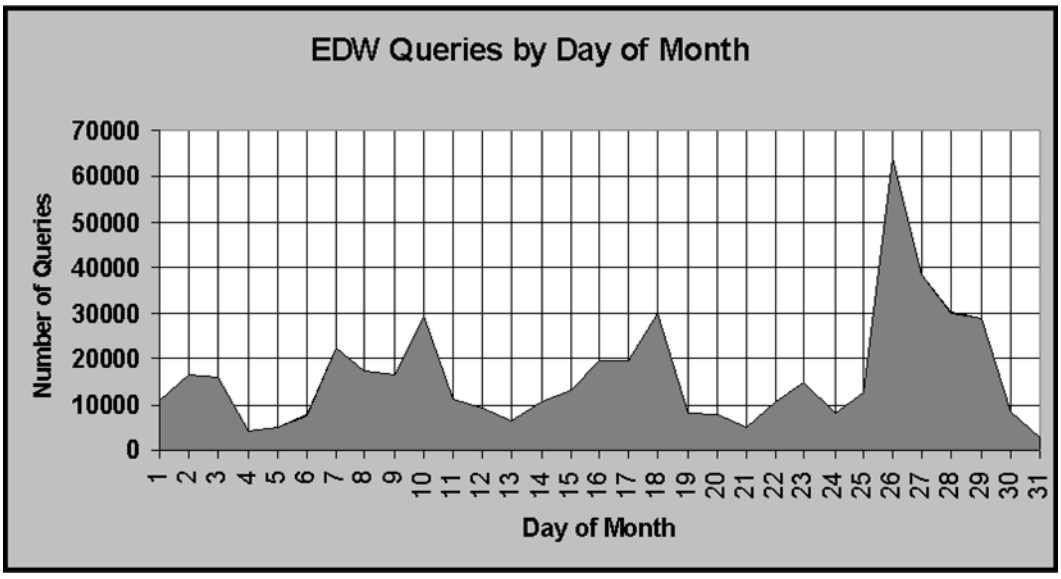
1. Overall, how satisfied are you with the Enterprise Data Warehouse, as an information system (data content, performance, availability, reliability, etc.)?
2. Overall, how satisfied are you with the services provided by the EDW Team (skills, responsiveness, courtesy, etc.)?

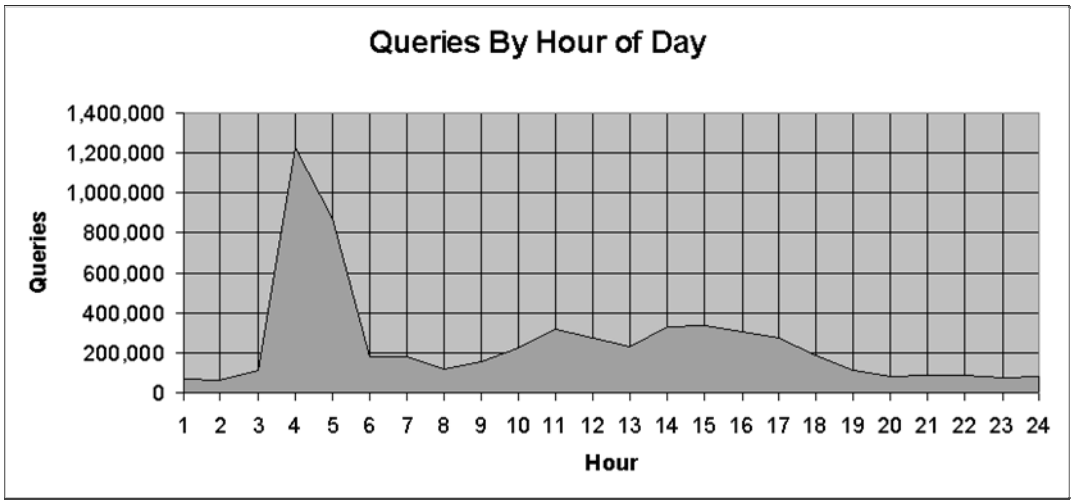
The other metrics gathered by the EDW Team are more typical of information systems, although the metrics required to manage an EDW environment are unique from transaction-based systems. The IHC EDW Team collects metrics on the following, all of

which are trended over time and available via customized web-enabled applications or Crystal Reports.

- Overall EDW Use: Most queried tables, query counts by table, queries user, and queries by application
- EDW backup times
- ETL times for the various data marts
- User Sessions: Minimum, maximum, and average sessions
- Records in the EDW: By table, schema, and in total
- Query response time: Number exceeding 90 minutes and average response time overall
- Total EDW direct log-in accounts
- Tables in the EDW: Number in total and number accessible by end users
- Cells in the EDW: By table and in total (cells = # Rows x # Columns)
- CPU and memory utilization: Peak and average
- Disk utilization: Free and used
- Query metrics: Counts, tables queried, rows returned, average run time; and queries that return no rows (an indicator of content problems; poor understanding of the content; or poor understanding of SQL)
- Database performance metrics: Full table scans; buffer pool management; physical writes and reads; cache hits; pin hits; session waits and timeouts; etc.

Two examples of these EDW operational metrics trended over a one-year time frame are provided below.





Future Plans: IHC's "Strategic To-Do List" for their EDW includes some of the following, more or less in order of priority:

- Integrating aggregate analysis with clinical care process: IHC is experimenting with the impact of trend-based, aggregate data in a clinical workflow setting. Currently, most of the aggregate data analysis produced from the EDW is provided "off-line" to clinicians for retrospective assessment. In the future, this trend-based data, such as that discussed previously under the Diabetes and Cardiovascular Clinical Programs, will be displayed as an integrate part of the clinical medical record.

- Data mining: There are many interpretations and definitions of "data mining", but in IHC's context it is defined as the application of probabilistic pattern recognition algorithms to the background analysis of data. In practice, this means using data mining tools to "crawl" through the EDW, identifying patterns in data that might otherwise escape the detection of human analysts. Data mining has matured in recent years, and among its potential uses is risk profiling for patients that fit a particular pattern of health, e.g., "diabetes", yet are not yet diagnosed or being treated for such. Data mining has been used successfully for a number years by the insurance industry in the detection of fraudulent billing practices.

- Strategic Alerting: This form of alerting is used at the aggregate data level to identify trends as they are developing. Potential applications include the detection of outbreaks from naturally occurring epidemics as well as those perpetrated by bio-terrorism. Strategic alerts can also be used in concert with patient level alerts generated in the electronic medical record. For example, in IHC's HELP system, (hospital-based electronic medical record), the antibiotic assistant alerts doctors to the most effective antibiotic, based upon the patients' clinical profile. By collecting the use of antibiotics in the EDW and assessing their use in aggregate, analysts can determine if the transaction-based alert is truly effective in reducing antibiotic costs or improving clinical outcomes.

- Natural Language Processing (NLP): Some estimates place the amount of text-based data in a healthcare organization as high as 80% of the total data in the enterprise. The ability to process this free-text data and convert it into data that can be examined for trends and common patterns represents the next generation of data analysis in healthcare. In some cases, especially the analysis of pathology reports that are primarily text based, NLP is already having a significant impact on analytics.

- Rules engine: Rules engines in which business logic, or Medical Logic Modules (MLMs), are executed to support transaction-level processes have been common for many years, yet these rules engines have not experienced any significant penetration in the data warehousing architecture of any industry, including healthcare.

- Familial relationships: The Church of Jesus Christ of Latter Day Saints (Mormon Church) maintains the most extensive library of family relationships in the world. This library exists less than two blocks from IHC's corporate headquarters in downtown Salt Lake City. The possibility of combining IHC's extensive clinical records with familial relationships data from the church's archives is intriguing.

- Genetic data: Genome data warehouses exist and clinical data warehouses exist, but to date, very few data warehouses exist that combine the two types of data and attempt to correlate their relationships. Assuming that society can manage this data ethically, healthcare data warehouses will someday contain clinical and genomic data that enables prospective risk and outcomes analysis to levels never before possible.

- Query By Image Content (QBIC): Population-based image analysis is now becoming possible through emerging QBIC technology. Traditionally, data warehouses have not placed great emphasis on capturing image-based data because no capability existed to analyze this data over time, in aggregate. This capability will emerge over the next five years to the point of usefulness in the mission of an EDW.

Summary of Lessons Learned

- Data marts first, data warehouse second: Have a grand vision of the future, and define your enterprise standard data dictionary early, but build the warehouse one step at a time with data marts.

- Maintain the look and feel of the source systems: Following the "Data mart first, warehouse second" philosophy, design data marts so that the data names and relationships resemble the source systems, while still adhering to standards for core data elements. To facilitate an "enterprise" perspective, database views can be created later. The metadata library can also be used as a translator for analysts that are not familiar with this source system perspective.

- Divide business logic from data structures: This is a principle that applies to transaction based systems for years, but is frequently overlooked in data warehousing. Avoid overly complex load processes that attempt to impart significant

business or analytic logic in the data itself. Implement business or clinical logic in one of three ways: (1) Summary tables, (2) A formal rules layer, or (3) Reporting applications. Leave the underlying granular data as "pure" as possible.

- Granularity is good: Grab as much detailed data as possible from the source systems in the ETL process. Inevitably, failing to do so will mean repeated trips back to the source systems, as analysts' desire more and more data.

- The EDW will be blamed for poor data quality in the source systems: This is a natural reaction because data warehouses raise the visibility of poor data quality. Use the EDW as a tool for raising overall data quality, but address data quality at the site of creation in the source systems. In keeping with is principle, avoid complex data scrubbing during the ETL process of the EDW--improve data quality in the source systems first.

- The EDW Team will be called "data thieves" by the source systems: In the early stages of the EDW lifecycle, the stewards of the sources systems will distrust the EDW development Team and their ability to understand and use the data properly. As the EDW matures and the source systems become accustomed to the EDW's existence, this distrust will fade, though never totally disappear. Encourage the stewards of the source systems to take lifecycle ownership of the source system's data, even in the EDW. Invite them into the development process of the data marts and later, the reporting process, as well. Source system stewards understand their data-- acknowledge and embrace this fact and leverage it to benefit the mission of the EDW.

- The EDW will be called a "job robber" by the source systems: The EDW is frequently perceived as a replacement for source systems. The truth is quite the opposite: The EDW depends on transaction systems for its existence. Also, the source systems may perceive as threatening any attempt to migrate analytical reporting from the production system to the EDW. Do not seek to migrate reporting to the EDW simply because it is possible. Migrating reports to the EDW should be motivated by alleviating the processing burden from the source system or to facilitate easier access from the analyst's perspective.

- The EDW will not fit well in the organizational chart: Data warehouses are traditionally difficult to align in the organization because warehouses apply across the enterprise, not to any particular vertical business or clinical area. In any organizational strategy, the EDW should stay aligned in some fashion with the CIO—doing so is critical to the EDW Team's ability to influence the support of the source systems.

- Four roles are required the reporting process: Four roles are necessary to produce a report that is analytically and politically valid—(1) A respected business and/or clinical leader familiar with the process under analysis; (2) A data manager or steward that is familiar with data content and quality issues; (3) A statistician that is familiar with the type of analysis in question and can define valid and invalid interpretations of the results, and can drive the analysis techniques, and (4) Information technology staff from the EDW Team that can support the data modeling and programming needs of the analysis.

- Real data warehousing experience is rare: Hire or contract at least one person for the EDW Team that possesses genuine experience and be wary of anyone that claims a multitude of experience. To fully understand the lifecycle issues of a data warehouse requires at least three years experience with any particular system.

- Data modeling and common sense: Organize and name your database schemas around your business. Schemas should contain similar data, functionally or operationally and reflect this in their names.

- Database tuning basics: Many EDW performance problems can be boiled down to very basic tuning concepts. Publish rules-of-thumb for indexing and partitioning at the on-set, and apply them liberally in every data mart.

- Empower end users: Err on the side of too much access, rather than too little. Assume that analysts are qualified professionals and capable of accessing the base tables in the EDW with free-hand SQL, if they so desire. If this assumption proves incorrect, deal with the problem on an individual, case-by-case basis. If the problem

is more widespread, facilitate training to correct it. Analysts are the customers of the EDW and can be enormously powerful allies and supporters, if they are treated accordingly.

We are just beginning to understand the processes and analytic requirements necessary to implement continuous quality improvement in healthcare—the surface is barely scratched, yet we are witnessing amazing insights already. We can only guess at the potential that lies ahead. It is the most exciting time in the history of our industry—and data warehousing is right at the center of the upcoming revolution in knowledge.

References

1. Lewis, D., “Studies Find Data Warehousing Pays Off”, March 13, 2001, InternetWeek.
2. Siwicki, B. “Managing Data, managing care: How Aetna U.S. Healthcare is using a Massive Data Warehouse to Better Manager Healthcare, Health Data Management May 1999.
3. Shockley, K., “The Long View: Trying to create an integrated customer record? This healthcare data warehouse holds valuable lessons”, Intelligent Enterprise Magazine, May 7, 2001
4. Bourke, M., “Strategy and Architecture of Health Care Information Systems”, New York, NY; Springer Verlag, 1994.
5. Dodge, G., Gorman. T.; “Essential Oracle8i Data Warehousing”, New York, NY, Wiley Computer Publishing, 2000.
6. Adamson, C., Vererable, M., “Data Warehouse Design Solutions”, New York, NY, Wiley Computer Publishing, 1998.
7. Kimball, R., et al, “The Data Warehouse Lifecycle Toolkit”, New York, NY, Wiley Computer Publishing, 1998.
8. Broverman, C. A. “Standards for Clinical Decision Support Systems.” Journal of the Healthcare Information and Management Systems Society, Summer 1999.
9. Ramick, D, “Data Warehousing in Disease Management Programs”, Journal of the Healthcare Information and Management Systems Society, Fall 1999.
10. Verman, R., Harper, J., “Life Cycle of a Data Warehousing Project in Healthcare”, Journal of the Healthcare Information and Management Systems Society, Fall 1999.
11. Kimball, R., “The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses”, Wiley & Sons, New York, NY, 1996.
12. Kelly, S., “Data Warehousing in Action”, Wiley & Sons, New York, NY, 1997.
13. Adelman, S., Moss L., “Data Warehouse Project Management”, Addison Wesley, 2000.
14. Berndt, D. et al, “Healthcare Data Warehousing and Quality Assurance”, IEEE Computer, December 2001.
15. Hall, C., “Data Warehousing for Business Intelligence”, Cutter Consortium, March 1999.

16. Ledbetter, C., “Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse”, Journal of Healthcare Information Management, Fall 2001.
17. Buckingham, M., “First, Break All the Rules: What the World's Greatest Managers Do Differently”, Simon & Schuster, May 1999.
18. Westerman, Paul, “Data Warehousing: Using the Wal-Mart Model”, Morgan Kaufmann, January 2000.

Biography

Dale Sanders is a Senior Medical Informaticist at Intermountain Health Care where he is responsible for the Enterprise Data Warehouse, and supporting Medical Informatics in the IHC Urban Central Region in Salt Lake City, UT. His professional experience in information systems started in 1983, as an officer in the U.S. Air Force. During his tenure in the Air Force, he was involved in a variety of information technology projects focusing on strategic data fusion in complex decision making environments including an assignment on the National Emergency Airborne Command Post, also known as the "Doomsday Plane"; the Looking Glass Airborne Command Post, and support for the U.S.-Soviet Union treaty negotiations in Geneva between President Reagan and Premier Gorbachev. In 1989, he resigned from the Air Force as a captain and joined TRW, Inc. as a systems architect. While at TRW, his assignments included information systems counter-espionage/counter-terrorism for the National Security Agency, nuclear weapons software safety assessment, and large-scale systems design and database integration projects for the U.S. Air Force, and the National Institutes of Health. In 1995, Mr. Sanders formed a small company specializing in systems architecture, database integration, and data warehousing for customers including IBM, Intel, and Motorola. In 1997, he joined Intermountain Health Care. Mr. Sanders was born and raised in Durango, Colorado. He graduated from Ft Lewis College, Colorado in 1983 with degrees in chemistry and biology. In 1984, he graduated from the Air Force's information systems engineering program.

Pearls of Wisdom

- Customer satisfaction is not possible over the long term without employee satisfaction. Employee satisfaction must come first.

- Data warehousing success is all about changing behavior. Many companies spend millions of dollars deploying a data warehouse but fail to realize any real business benefits from the investment because the corporate culture does not have the ability to effect behavioral changes or process improvement. Before investing in a data warehouse, the company should ask itself—"How committed are we to changing our processes and behavior as directed by the knowledge we gain through analytics?"

- The key to success in any business environment is the cross-product of three variables: Quality, Productivity, and Visibility. You must produce a quality product, in volumes high enough to sustain your business, and someone must see the product, value it, and attach your name to it.

- Deploying the technology for an analytical information system is only one-half of the project--don't forget to close the loop of process improvement. The ROI resides in your ability to improve the processes supported by the technology.

- The state of information technology in health care is an amazing mix of the best and worst available. To reach the next level, the state of IT in Health Care must improve by learning from other industries and applying information systems processes and concepts that are common in those industries, especially manufacturing and retail.

- Hire IT staff based on their Values, Technical Skills, and Domain Knowledge, in that order of priority. Technical skills and domain knowledge can be taught; values are more difficult to influence.