Lessons from a Successful Data Warehouse Implementation

by John D. Porter and John J. Rome

A data warehouse is often the first client/server application an organization attempts. This was the case at Arizona State University (ASU), where such a project brought together student, financial, and human resources data in an integrated data warehouse. This article discusses the project's history and architecture, issues faced, and lessons learned after three years of work.

To remain competitive in today's business climate, an organization needs a foundation of quality data. Organizations of higher education need this capability as much as Fortune 500 companies. To ensure quality data for tactical and strategic decision-making, many colleges and universities are creating a "data warehouse."[1] A data warehouse is a separate store of data extracted from one or more production databases to produce an authoritative source for decision support. Some critics believe data warehousing contributes to an organization's information problem by adding yet another source of data. However, the success that organizations are experiencing with data warehousing proves it is a solid business strategy for the 1990s.

Building a data warehouse is extremely complex and takes commitment from both the information technology department and the business analysts of the organization. It takes planning, hard work, dedication, and time to create a relational database that delivers the right data to the right user. Arizona State University's data warehouse is not a panacea for every data problem, but it is a very good start toward a permanent solution.

"Building a data warehouse is extremely complex and takes commitment from both the information technology department and the business analysts of the organization."

Data Warehousing—Popular, But Not New

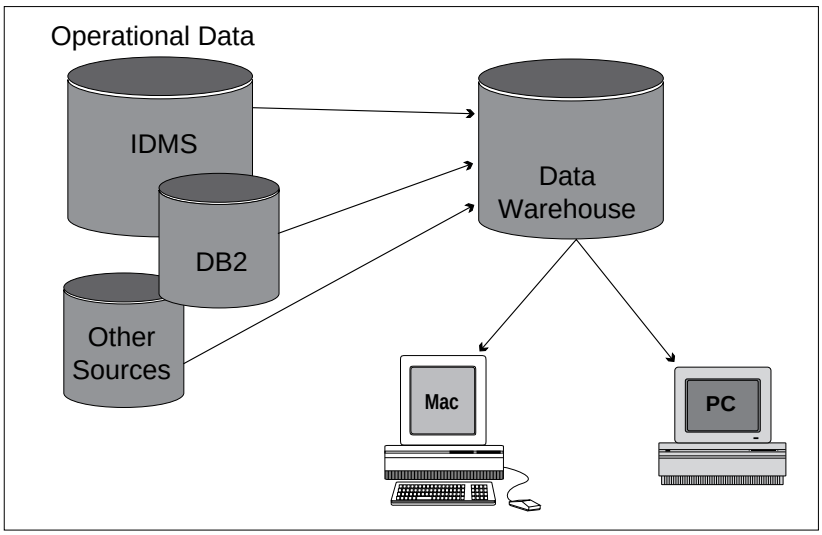
Data warehousing is not new. In fact, it is reminiscent of an old mainframe concept from the mid-1970s: take data out of production databases, clean them up a bit, and load them into an end-user database. International Business Machines Corporation (IBM) first coined the phrase "information warehouse" in 1991. IBM's original concept was met with skepticism because accessing non-relational data stores (such as IDMS®, IMS® or VSAM®) was too complex and degraded operational system performance. Based on these

John Porter *(john.porter@asu.edu) is Director of Institutional Analysis and Data Administration at Arizona State University. He has seventeen years experience in information management and is a leader in the field of institutional research. Two years ago, Porter became ASU's data administrator and assumed leadership for developing the data warehouse. Currently, he is prototyping ways to build inexpensive online analytical processing (OLAP) systems for ASU's executives.*

John Rome *(john.rome@asu.edu) is Assistant Data Administrator at Arizona State University. Involved in data warehousing since its inception, he is the project leader and chief architect of ASU's data warehouse. An expert in database design, Rome spent ten years of a "previous life" as a database administrator and programmer/analyst. He is also exploring inexpensive online analytical processing (OLAP) solutions for executive information systems and accessing distributed data stores through the World Wide Web.*

[1] William H. Inmon, *What is a Data Warehouse?* Tech Topic 1, No. 1 (Sunnyvale, Calif.: PRISM Solutions, Inc., 1992), p. 1.

Figure 1: Diagram of ASU's data warehouse



experiences, experts now agree that a warehouse needs to be a separate data store built with a relational database management system (RDBMS). Names such as "information factory" or "data refinery" surfaced initially, but "data warehouse" is now the generally accepted terminology.

Warehouse definition

The most widely recognized definition of a data warehouse is, "a subject-oriented, integrated, time variant, non-volatile collection of data in support of management's decision-making process."[2] Subject-oriented means the data warehouse focuses on the high-level entities of the business, such as students, courses, accounts, and employees. This is in contrast to operational systems, which deal with processes such as student registration or payment of an invoice. Integrated means the data are stored in consistent formats (e.g., consistent naming conventions, domain constraints, physical attributes, and measurements). ASU's operational systems have four unique coding schemes for ethnicity. In the data warehouse, there is only one coding scheme. Time variant means the data are associated with a point in time (e.g., semester, fiscal year, and pay period). Finally, non-volatile means the data do not change once they are entered into the warehouse.

Use increasing

In higher education, glimpses of data warehousing exist in the file extracts that institutional research departments receive or the user reporting databases that the information technology department creates. Consequently, data warehousing is really an old concept with a new name and better technology. The data warehouse is likely to become the cornerstone of client/server activity in the immediate future.[3] So popular is the notion that a recent META Group report indicates 90 percent of their clients are undertaking warehouse initiatives, up from less than 10 percent just a year ago.[4] Judging from the number of inquiries about ASU's data warehouse, similar trends are occurring in many higher education organizations. In the business market, analysts estimate the industry will grow to $2.1 billion by 1998.[5] Some of the major players vying for this money include IBM, Hewlett-Packard Company, Oracle Corporation, AT&T GIS, Sybase, Inc., and SAS Institute, Inc., as well as vendors already established in this market such as Prism Solutions, Inc. and Red Brick Systems.

ASU's Warehouse Development

Development of ASU's data warehouse started in the summer of 1992 as a client/server "proof of concept" project. Negotiations with RDBMS and UNIX workstation vendors resulted in a one-year lease of a server for the cost of the annual maintenance contract. While getting the warehouse server in place, over twenty companies agreed to provide complimentary copies of their data access software for evaluation. Although many of the access tools were in their adolescence at the time, accessing data was much easier with these graphical user interface (GUI) tools than with the fourth-generation tools then in use.

ASU formed a development team of twelve individuals from the data administration and information technology departments to build the data warehouse. The team selected a representative group of business analysts to serve as pilot users to test the warehouse and access software. During the next few months, the team built a "student" warehouse model based on over 200 questions, which the pilot users considered difficult or critical to answer using current information resources.

During 1993, many of the original data warehouse team members shifted back to their regular duties, leaving a core of about five full-time equivalent employees working on the project. That core has remained intact, receiving additional help from ASU's institutional research office and many of the business analysts who are regular users of the warehouse. Also, the data administration department initiated a formal program to train users on the warehouse. To date, there are over 400 trained warehouse users, with two to three classes being taught each month.

[2] Ibid.

[3] Colin White, "Client/Server Obsession," *Database Programming and Design,* Special Supplement, December 1994, p. 7.

[4] Katherine Hammer, "Will the Data Warehouse Be Warehoused?" *Relational Database Journal,* September-October 1994, p. 32.

[5] Rosemary Cafasso, "Praxis Forges Data Warehouse Plan," *Computerworld,* 10 October 1994, p. 32.

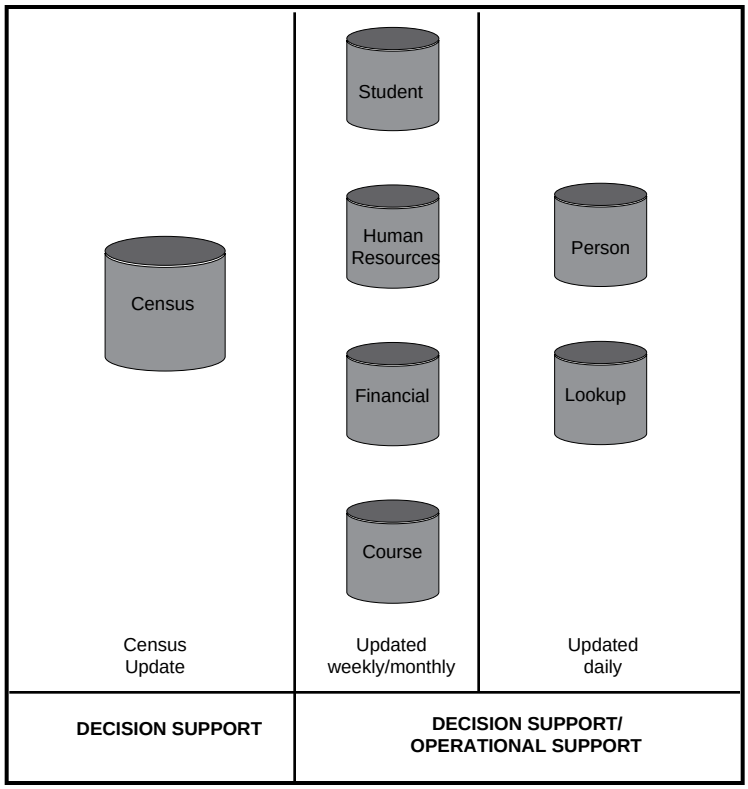
The goal is to train 1,000 employees, approximately 20 percent of ASU's full-time work force.

Warehouse architecture

ASU's data warehouse resides in a client/server environment. As seen in Figure 1, ASU extracts data from the mainframe and loads it into a UNIX server running an RDBMS. ASU's warehouse server is a Sun® Sparc 630™ with 512 megabytes of memory and two processors, running the Sun Solaris 2.3™ operating system. The RDBMS is Sybase SQLServer release 10.x. Users connect through Ethernet to the warehouse over ASU's network backbone via Transmission Control Protocol/Internet Protocol (TCP/IP). The suggested GUI data access tool is BrioQuery™ from Brio Technology, Inc., which runs identically in both the Macintosh and Windows environments. Microsoft Access® is another tool used.

The process of using GUI tools to build structured query language (SQL) requests and bring the results back to a client machine is much different from the 3270 protocols in a mainframe environment. With client/server architecture, once the data are in the workstation, users "own" the data, cutting and pasting at will into their favorite software (e.g., spreadsheet, word processor, graphic tools).

Figure 2: Warehouse design



Modeling and design

As business requirements and database technologies become more sophisticated, the need for data modeling and design increases. ASU uses an "upper" computer-aided software engineering (CASE) tool to design the warehouse. However, the entity/relationship (E/R) diagramming function and the object repository are the only features of the CASE tool used. The E/R diagramming tool creates a graphical representation of the data in the data warehouse and automates the creation of data definition language (DDL), the technical language used to create the warehouse's tables, views, and indexes. The object repository ensures consistent definitions and characteristics of fields in the data warehouse. While an upper CASE tool is not imperative in building a data warehouse, it does help automate the development process and the E/R diagrams produce "road maps" to the data.

Designing a data warehouse is different from designing an operational system (see warehouse design, Figure 2). The warehouse wants data with a high value for decision-making, whereas the data content of an operational system is more requirements driven. ASU's data warehouse contains four primary databases: student, finance, human resource, and course. These databases are updated weekly, biweekly, monthly, and yearly, depending on the data. They also contain "official" data, which are captured on the census date and never changed. Over time, this results in an official and end-of-period look to the historical data in the warehouse. For example, by the end of the semester, the student database provides users official and end-of-semester enrollment data. ASU found that some data in the warehouse need more frequent updating. These data are the valid occurrences of data contained in the code tables and the names and addresses of ASU's customers. These records are updated daily.

There are four basic types of tables in ASU's data warehouse: data tables, lookup tables, virtual tables, and summarized tables. Data tables contain raw data, extracted at the unit record level from the operational system. Lookup tables are code tables, defining the cryptic coding schemes that exist in the operational data. Lookup tables save space, improve flexibility, and allow the description of a code value to change while retaining its meaning. Virtual tables are views into the warehouse data. Views simplify the user's perception of a data warehouse, presenting data in a different way or restricting access to certain data (e.g., class roster appears as a single table, but the data reside physically on multiple tables). Last, summarized

tables contain summarized data. These tables improve response time to frequently queried data and will become the foundation for ASU's online analytical processing systems (OLAP) and executive information systems (EIS).

Database design is a creative process. In fact, given the same set of requirements, two designers usually produce different but acceptable solutions. Often, in database design, it is easier to just do it than to explain exactly what you did. ASU's warehouse team follows the design guidelines listed in the sidebar below.

"A data warehouse must deliver the right data to the right people. However, the data warehouse cannot deliver all the data people want."

ASU's Ongoing Warehouse Data Issues

ASU addressed a number of data issues in the process of building its data warehouse over the past three years, including determining what data to collect, how often to update the data, how to achieve "officialness," and how to resolve data security and privacy issues. ASU's experience is that with a data warehouse, these issues are never completely resolved.

What data to collect

A data warehouse must deliver the right data to the right people. However, the data warehouse cannot deliver all the data people want. People are always asking new questions, so predicting what data they need is difficult. ASU started by asking users what data they wanted or what reports they used. Another good starting point is to look at the data going to the institutional research department. ASU's experience is that once the data warehouse is implemented, users quickly let the development team know what data they want.

Update frequency

A data warehouse must deliver the right data at the right time, but what *is* the right time? The answer is, "it depends." In ASU's data warehouse, data are entered yearly, by census date, monthly, bi-weekly, weekly, and daily. By rule, the more often a table is updated, the more operational the nature of the data. For example, ASU's data warehouse updates daily the addresses of students. Many warehouse users create labels for student mailings and need current address information. Updates to code tables occur daily, too. However, ASU tries to limit the number of data elements updated daily, since there is a cost associated with loading the warehouse. In the future, daily updates to ASU's data warehouse will "replicate" data in operational systems. Replication is an economical industry solution of copying data and making them available to users on a server.

Official vs. current

Making official numbers available in a data warehouse brings consistency and credibility. ASU adds official "numbers" to the warehouse to limit how much users need to understand the impact of timing on the data. To achieve "officialness," an organization selects census or "cut-off" dates for measuring data. With census dates, there is a distinct period of measurement, making historical trends much easier and allowing integration across systems. For some requests, official numbers are better to use (e.g., historical trends), while at other times (e.g., for financial decisions) the most current data are best. At ASU, both numbers are available from the data warehouse. However, to simplify user queries, official and current values appear in separate databases.

Delivering "officialness" in the data warehouse is not as easy as it sounds. The programs that extract and transform the data from the legacy databases must produce numbers that

ASU's Warehouse Design Guidelines

- Identify major subjects or topics as tables in the warehouse
- Add an element of time to the tables—semester, fiscal year, etc.
- Appropriately name fields in the tables or views
- Add derived fields when necessary—calculated age, GPA, etc.
- Duplicate data to decrease the number of tables that must be joined
- Exclude extraneous fields found in operational files—"flags," etc.
- Create logical tables or views for ease of use—class roster, etc.
- Consider security and privacy during design
- Make sure the data model answers the critical business questions

balance with the official numbers released by ASU. Since different algorithms and extract programs exist, there are often differences between the warehouse and official University reports. The problem multiplies because of ten years of data in the warehouse. Creating and validating ten years of official data was difficult. ASU found that going forward in time when building the data warehouse was easier than reconstructing and validating history.

Security and privacy

Security and safeguarding privacy are major concerns when building a data warehouse. Security in a database means protecting data against unauthorized disclosure, alteration, or destruction. Granting *select* (authorization to read only) access to tables or views achieves a certain level of security in a warehouse. At ASU, read-only access to the data warehouse is at the database level. This procedure follows an open access policy for employees approved by ASU's administration in 1993. This policy is based on the notion that giving employees access to data and holding them accountable is better for the organization than withholding the data.

Although many RDBMSs support column-level security, ASU has not implemented this feature, primarily due to the high cost of administering user access. In traditional operating systems, tasks or screens control access, meaning users only have access to a single record or instance of data (e.g., verifying admission status of a student). In a data warehouse, users have access to a table or set of tables in a subject area, which means access goes beyond retrieving single records to retrieving groups of records.

At ASU, the registrar's office is the trustee of the student database, the human resources director is trustee of the human resources database, and so forth. In these databases, read-only access excludes access to name and address. To obtain name and address information, the data trustee grants access to the person database. The user's business need determines whether access is granted. Given the large number of records in ASU's data warehouse, placing name and address in a separate database achieves a certain level of privacy. Additionally, training classes emphasize the Family Educational Rights and Privacy Act (FERPA), and users must sign a document stating they will follow these policies. In addition, users receive training on the appropriate and fair use of information.

ASU's Lessons Learned

ASU's data warehouse development improved access to and the integrity of administrative data, increased organizational awareness of data administration, and improved the quality of decision-making. However, "like a good marriage, a data warehouse is not created instantly; it develops only over time, with thoughtful goal setting to build a strong foundation. And like a nasty divorce, a bad warehouse can extract a lot of money, time, and energy from the parties involved."[6] ASU has high hopes for its marriage. But as with any computing project, ASU learned many lessons along the development path and hopes other organizations will learn from its experience.

ASU's Lessons Learned

- ➤ Develop an enterprise strategy.
- ➤ Identify a project champion.
- ➤ Avoid cost justification.
- ➤ Be ready for technology shortfalls.
- ➤ Make users aware of costs up front.
- ➤ Find ways to capture metadata.
- ➤ Build integrity and integration.
- ➤ Let the warehouse fill operational gaps.
- ➤ Invest in training.
- ➤ Make sure a support structure is in place.

"Security and safeguarding privacy are major concerns when building a data warehouse."

Develop an enterprise strategy

A successful data warehouse requires an enterprise strategy; otherwise the data warehouse may fail. The first step in establishing this strategy is adopting policies promoting sound data management. A data warehouse is easier to build and more useful to the organization when strong data management practices are in place. Before opening the data warehouse, ASU adopted an enterprise data access policy, determining who received access and what type of access a user received.

Policies making the data administration department responsible for data integrity and integration of the data warehouse and enterprise operational systems also were implemented. ASU found these policies essential to navigate the warehouse project around everyone's data "turf." Also, good enterprise thinking must penetrate development culture. Users, business analysts, and the information technology, institutional research, and data administration departments must work together. At ASU, strong collaborative working relationships exist among all these areas. This culture contributed significantly to the success of ASU's warehouse implementation.

[6] Kim Nash, "Data Warehouses Mature with Help," *Computerworld*, 15 May 1995, p. 69.

"Users do not believe how bad their data are until they see them."

Identify a project champion

All computing projects need a champion. ASU's data warehouse champion is the data administration department. Data administration follows the evolution of the data warehouse according to Bill Inmon. Inmon says the data administrator's role has changed dramatically from managing the data dictionary to designing and constructing the data warehouse.[7] ASU's data warehouse put the office of data administration on the map and brings a new awareness of enterprise data to the organization. Users do not believe how bad their data are until they see them. For example, one college uses the data warehouse to verify professional program information and correct mistakes on ASU's operational systems. However, a data warehouse is a double-edged sword for the data administration department. Once users access the warehouse, a "never-ending" list of enhancements quickly appears. At this point, organizations will need to commit additional resources for warehouse development and support, or other data administration functions suffer.

Avoid cost justification

If possible, avoid the traditional cost/benefit analysis in justifying a data warehouse project. Since a data warehouse benefits the entire organization, ascertaining the full benefits is difficult, if not impossible. Fortunately, at ASU, a limited demonstration of the warehouse concept was enough to sell the project. If a more complete cost/benefit analysis were required, the project might never have started. In other words, don't spend too much time justifying the benefits of a warehouse; just start building one!

A data warehouse may be inevitable for most organizations, since there is little chance that a technical breakthrough will make access to legacy data easier or cheaper. The Gartner Group says, "Organizations employing a data warehouse architecture will reduce user-driven access to operational data stores by 75 percent, enhance overall data availability, increase effectiveness and timeliness of business decisions, and decrease resources required by IS to build and maintain reports."[8] But how can all these benefits be quantified?

Be ready for technology shortfalls

Client/server technology is less reliable, secure, and timely than its mainframe predecessor. Data access tools are just beginning to mature. Networks add new layers of complexity, and monitoring performance and tuning of servers is imperfect. The results are gaps in available technology and software, leaving users frustrated and their needs unmet. One such example is matching a cohort on a desktop machine with the data warehouse. Most query and retrieval tools do not support this type of function (joining a local table with server table). If the tool allows this function, joining data is slow, making the match process prohibitive for large databases. Allowing users to create tables containing the IDs of records being tracked on the server solves this problem. However, this solution defeats the benefits of client/server technology, moving emphasis back to the host machine. ASU's experience is that when problems occur with client/server technology, no one, including the vendor, knows how to solve the problem in a timely manner.

Make users aware of costs up front

The information technology department and technology infusion funding traditionally absorbed much of the cost of new technology at ASU. With the data warehouse and client/server computing, the cost of upgrading hardware and buying software for enterprise systems has shifted to the individual or department. Employees seeking access to the warehouse need to know the cost of connecting. At ASU, a "connection checklist" is available, detailing all the steps necessary for access. The checklist includes information on these items: how to obtain a data warehouse account and receive access, what PC or Mac and printer to buy, how to connect to the network, what software to buy, and how to register for training. ASU finds this checklist very helpful.

Find ways to capture metadata

One of the more difficult tasks is providing users and application developers a good data dictionary and source for metadata. Metadata are

ASU Data Warehouse Connection Checklist

- ✓ Data Access Approval
- ✓ PC or Macintosh
- ✓ Printer
- ✓ Ethernet Connection
- ✓ Communications Software
- ✓ Data Access Software
- ✓ Software Installation
- ✓ Training

[7] William H. Inmon, "Winds of Change: A Brief History of Data Administration's Amazing Growth and Development," *Database Programming and Design*, January 1992, pp. 68-69.

[8] Gartner Group, "Data Warehousing," a conference presentation on data warehouse, 1994.

DATA WAREHOUSE	vs.	ADMINISTRATIVE SYSTEMS
• data are read-only		• data are updated
• serves management		• serves operational users
• "time-fixed" data		• "current value" data
• "what if" processing		• processing is repetitive
• data driven		• requirements driven
• response…minutes		• response…seconds

data about the data, including format, encoding/decoding algorithms, domain constraints, and definitions of the data. There are thousands of data elements, and capturing metadata is an endless task. Although this process is time consuming, the dividends are significant. ASU learned developers are more concerned about metadata, while users want data definitions. The problem is exacerbated by the fact that good metadata and data dictionary tools do not exist in client/server technology. ASU found no good tools to help solve the metadata problem. This is a problem that needs to be solved by the client/server industry.

Build integrity and integration capabilities

Integrity and integration are important characteristics of a data warehouse, and the characteristic lacking in most operational systems. These features give the data warehouse credibility, consistency, and real power. When designing these capabilities into ASU's data warehouse, the development team recognized that the integrity of the data varied. In some cases, the development team "scrubbed" the data; in other cases, it was simply too difficult. ASU's experience is that making the data available through the warehouse improves data accuracy in all systems. Knowing the data are observable in the warehouse is an incentive for those inputting data into the operational systems to be accurate.

The data warehouse also requires data that integrate. These are the data that span the high-level subject areas of the warehouse. At ASU, these high-level subject areas are students, financial information, human resources, and courses. Examples of data that integrate or crosswalk the high-level subjects are fiscal year, semester or term, department, course, a person's unique ID, and account number. Data elements that integrate are the very fabric of an operational system. If these elements differ in format, domain, or values between systems, integrating the data in the warehouse is difficult or impossible. When data successfully span high-level subject areas, building a data warehouse is easier and less expensive. ASU's experience is that most integration problems need to be solved in the organization's operating systems before attempting to integrate data on the warehouse.

"ASU's experience is that most integration problems need to be solved in the organization's operating systems before attempting to integrate data on the warehouse."

Let the data warehouse fill operational gaps

Many users tend to look at the data warehouse as another administrative system. This phenomenon happens since the data warehouse is in relational format. While the warehouse can address some of the data shortfalls that operational users experience ("data gaps"), this is not the warehouse's primary role. To help our users understand the difference between the data warehouse and their administrative system, we developed a slide that compares a data warehouse to an administrative operational system on a variety of dimensions (see sidebar above). Every talk or presentation on the data warehouse includes this slide underscoring the differences between the two. ASU reiterates these differences frequently to discourage users from making unreasonable requests of the warehouse. However, the truth is that ASU's data warehouse plays a powerful role in bringing inexpensive, temporary solutions to some operational computing shortfalls.

Invest in training

Training data warehouse users is critical and pays good dividends. In most computing projects, management recognizes the need for training, but does not always fund training. This is true of ASU's data warehouse. With every new database there is a need for another training course, complete with reference materials. Every enhancement or change to the warehouse must be documented and communicated to warehouse users. At ASU, the data administration department assumed responsibility for training

and documentation of the data warehouse. While training users is essential, it distracts from future warehouse development unless new resources are allocated.

Initial training at ASU focuses on the tool, the logic, and the data. While a data warehouse supports many different access tools, training with one tool reduces a trainee's learning curve. After an extensive review of data access tools, ASU chose a tool that works in both the Macintosh and Windows environment. Logic training is important also (e.g., SQL operators, Cartesian join, etc.). While this functionality is inherent in most access tools, training on query logic avoids many questions down the road. Finally, ASU's training concentrates on the data, which is usually what users understand the least. ASU spends up to 60 percent of class time training on data, and hopes to increase this percentage as users become more familiar with access tools and query logic.

"Eventually, the warehouse will serve as a telescope into ASU's distributed data stores."

Make sure a support structure is in place

While training reduces the number of data warehouse questions, a support infrastructure is key to handling other support needs.

At ASU, there is an e-mail address where users can send their questions or problems. Experts on warehouse data, networking, and data access tools receive these messages and respond within 24 hours. ASU logs responses in a searchable database for users to reference in the future. Also, users can telephone a central help line that will send an e-mail message for them.

Second, there is a file transfer protocol (FTP) site available for warehouse users. This site stores PostScript copies of all documents associated with the data warehouse and copies of the data models. This is also a site for sharing common queries built by users or the warehouse team.

Last, there is a Warehouse Users Group. The WUG meets monthly to share findings, educate members about the data warehouse, and provide feedback to the warehouse team. Currently, over seventy-five people attend the monthly meeting. The WUG also gives warehouse users an opportunity to find a "warehouse buddy," so they don't feel alone in ASU's world of data.

Conclusion

After three years of experience, the future of ASU's data warehouse is becoming more clear. Initially, the warehouse served as a resource for accessing information from legacy systems. Eventually, the warehouse will serve as a telescope into ASU's distributed data stores. Some of these data will reside in the data warehouse, while other elements will be "viewed" from the RDBMSs where the data reside. ASU foresees a time when the telescope extends beyond ASU to other organizations with common goals, such as the neighboring Maricopa County Community College District. The real power of the warehouse will be actualized in years to come.

The data warehouse fills an important data administration role in a client/server environment. As distributed application developers move further away from the central computing core, the data elements in the warehouse ensure the integrity of the organization's enterprise data. The definitions and coding standards in the warehouse are what distributed developers follow. The warehouse is the "glue" holding enterprise data stores together until a mature repository comes along.

The most important contribution of ASU's data warehouse is the new focus on data integration. While attempting to achieve integration in the warehouse, ASU conceived a new data model which integrates not only the warehouse, but also the administrative systems. By integrating the warehouse, ASU obtains more powerful data. By integrating the operational systems, ASU gains strategic new levels of customer service.

The bottom line is that data warehousing is here to stay. Warehousing gives organizations the opportunity to "get their feet wet" in client/server technology, distributed solutions, and RDBMS. This is essential for any future mission-critical application, making the data warehouse a low-risk, high-return investment. The question is not simply whether to build a warehouse, but when.

C/E