Imagine for a moment that you’re an administrator in one of the Schools at Penn, and you need to answer a question: Is there a correlation between student GPA and course load in your School?

Your school has a team of programmers. One of them begins working on the answer to your question, writing routines to pull data from the University’s Student Records System (SRS). But there’s a problem. The official GPA isn’t stored anywhere in SRS; it’s calculated on the fly each time a transcript is printed. The calculation would be hard to reproduce; it’s many pages of complex code. Would an approximation be OK?

You groan. Not really. But then the programmer gives you some interesting news. Penn is building something called a “data warehouse” that may eventually let you—the ultimate user of the data—sit down at a desktop computer and point and click your own way to the answer. And yes, the warehouse will include the official GPA for your students. Eight gigabytes of data have already been loaded, and a pilot project team is hard at work evaluating the results.

The forest and the trees

The problem of getting data to decision-makers isn’t unique to Penn; every large institution has it. Data “warehousing” is an increasingly popular solution to this kind of problem. Most existing computing systems handle specific day-to-day functions, such as posting a student’s grades or printing a paycheck. These “transactional” systems were built to perform functions and store results. In many cases, the ability to retrieve information for analytic purposes was not a design priority. It’s as if you kept all of your bank statements in a grocery bag under the desk. When you need to analyze a trend, the statements are there, but you have to pull them out and find the right ones.

Unlike transactional systems with a single level of detail, data warehouses often have many layers of data, at different levels of summarization. At the lowest level—the “bedrock” layer—there are individual transactions; above that, you might find layers summarizing details for a week, a month or a year. How deep you go into the layers depends on what you need to find out.

Harmonic convergence

Many organizations are implementing data warehouses. At Penn two projects came together to produce the needed momentum.

The first was Project Cornerstone, which outlines Penn’s future information strategy. Cornerstone formulated principles that recognize the need to support administrative processes and provide access to the data people need to do their jobs. Penn would have to move aggressively into the world of relational database technology and client/server computing. The data warehouse was proposed as an integral part of that strategy.

Building on Cornerstone principles, Penn contracted with Oracle Corporation for a multi-platform database environment and a new financial system. The University also purchased an IBM SP2, a parallel-processor version of the popular RS6000 workstation.

Meanwhile, an informal group of information special-
ists around campus—in what has been described as a “skunkworks” project—began meeting to find ways of simplifying access to transactional data. SRS was a prime focus; its data is central to supporting the University’s teaching mission, and its structure, while allowing efficient day-to-day processing, does not lend itself to the task of analysis.

The University’s Data Policy Committee charged these information specialists to recommend solutions to the problem. In February 1994, the group’s “Report on Data Access Strategy” recommended using the warehouse concept.

The Data Policy Committee accepted the recommendation. In a textbook example of cooperation between the Schools and central University administration, the committee created a pilot project team with representatives from Arts and Sciences, Engineering, Wharton, Institutional Research, University Management Information Services (UMIS), and Data Administration. Over the course of the summer and fall, the team worked through the details of the system. There are 2,400 data elements in SRS, and each one needed to be examined individually: How is it used in the real world? Does it belong in the Warehouse? Where? Does it need to be converted or “cleaned up”?

In October, UMIS created a “beta” version of the SRS Warehouse data with 40,000 student records. On December 19, the full version, with data on every student and on every course section in SRS, was made available for the project team’s evaluation. Since that time the data has been updated six nights a week to keep it current.

A work in progress

Having the data available is only part of the task. The other part is providing an easy way to access it. Since the beginning of the project, the Warehouse project team has been evaluating a number of graphical query tools. There are two kinds. At one level there are “row-and-column” tools that let you choose from a list of data elements, indicate your selection criteria, and control how you want the results sorted and totaled. At another level there are tools that provide a multi-dimensional view of the data, simplifying trend analysis.

Choosing a set of query tools has not been easy in Penn’s multi-platform environment. The ideal “suite” of tools would have identical versions for both Windows and Mac, would integrate Penn’s own help text, and would allow users to join a local collection of data to the central collection for a specific query. For the initial pilot, the team selected BusinessObjects as the first-level tool. Site-license negotiations are continuing with several vendors of both report and analysis tools.

Another phase of populating student data in the Warehouse has just been completed. Trend analysis and official statistics need a static, point-in-time snapshot of the data that will produce exactly the same results now or 10 years from now. These snapshots were moved, term by term, into the Warehouse.

In the future, the Warehouse will contain more than student data. Discussions have already begun on adding Sponsored Projects data to the Warehouse. There is no reason to stop there; important information from all transactional systems can be “warehoused.” One advantage will be the ability to convert the “apples and oranges” of diverse systems to a common denominator.

Of course there are risks in putting University data into the Warehouse. Security is important, but it has a technical solution. A greater risk is the possibility that the data may be misinterpreted. Student data is complex. The Warehouse will simplify access to data, but the data must still be approached with knowledge and caution. This is one of the reasons Penn has chosen to “roll out” the

The Data Warehouse saves programming time for Schools and centers and increases the range of possible analyses

Warehouse slowly; the goal is to have a data structure and data documentation that, together, minimize the risk of misinterpretation.

The Data Warehouse will grow by leaps and bounds. What’s there now is only a small piece of a larger puzzle, only the first layer of a single system. Even so, the Warehouse has already begun to produce significant savings in programming time for Schools and centers, and to increase the range of analyses that can be done.

The potential is immense. The foundation of a building has been laid that will—if we’re lucky—never be complete.

TAD DAVIS is a Lead Programmer Analyst for University Management Information Services; DAN SHAPIRO is Director of Institutional Research for the University in the Department of Planning and Institutional Research.