One of the most compelling visions of modern computing is the promise of easy access to vast data resources. Microsoft’s Bill Gates speaks of “information at your fingertips” and former Apple CEO John Scully envisioned a “knowledge navigator” automatically sifting through vast data repositories.

Such ideas, while deceptively easy to imagine, are often difficult to achieve. At the Wharton School, a program to bring large financial data sets to faculty and students—The Wharton Research Data System—has taken a significant step toward making easy access a reality.

The financial data sets widely used at Wharton include market research data (such as CRSP, Fama, and Market Indices), corporate data (such as Compustat), and banking and insurance data (such as Best and FDIC). Together, the principal data sets used at Wharton occupy over 12 gigabytes of storage space.

Once the exclusive province of Finance and Accounting, these data sets are now used in Management, Marketing, and other Wharton departments. Wharton data sets range from examining world-wide investment patterns to handicapping the box-office success of upcoming motion pictures. Increasingly, faculty at Wharton also use the data sets for instructional coursework assignments.

The way things were

Data sets have been used at Wharton for many years; however, previous methods for delivering data were far from ideal. In the past, data sets were stored on large VAX/VMS systems, and users had to run Fortran programs to analyze or extract data. An increasing number of users, however, preferred working with familiar desktop tools such as Minitab or a spreadsheet program. But working with the data using desktop tools required that the user be familiar with the formats of the data sets, the VMS operating system, Fortran programming, mainframe to PC file transfer techniques, and the data import format of desktop software. As Michael Phelan, Associate Professor of Statistics, points out, these data access techniques were “functional, but not for the timid.”

Not only was the approach cumbersome for faculty and students, it was difficult for Wharton’s computing staff to support. To increase access speed, Wharton wrote in-house indexing programs. To help new users, Wharton provided interactive modules and help screens. Changes in data format required updating everything written in-house and extensive in-house programming support.

For all the effort required, users—accustomed to point-and-click graphical interfaces on personal computers—were increasingly dissatisfied with this arcane, multiple-step procedure.

Alternatives considered

Several alternatives were considered to provide easier access and improved data set management. Developing an in-house system would be costly, time-consuming, and difficult to maintain as the technology changed. Commercial database management systems provide excellent data management capabilities and convenient access tools, but most lack strong analytical tools and are not suitable for time-series data. Commercial data access packages such as Fame, DART, and Intelligent Query offer good data manipulation tools, but also lack sophisticated analytical abilities and require extensive programming to convert the wide selection of data sets used at the Wharton School.

The new architecture

The solution implemented in the Wharton Research Data System involves the following components:

• Using SAS (and SAS/ASSIST) to extract and analyze data
• Managing data sets centrally while providing network access (through NFS mounting) to the complete series of data on UNIX systems throughout Wharton
• Providing X-Window access to UNIX systems from Wharton’s labs and classroom teaching stations

(continued on next page)
SAS for data analysis

SAS best met Wharton’s objectives of offering a single, unified tool for data management and analysis. While SAS has long been popular for data analysis in the academic environment, release 6.09 greatly simplified reading external data. Extracting data requires only a few lines of code in SAS (versus several hundred lines of Fortran). Once extracted, a data set can be used by a wide range of SAS procedures.

SAS/ASSIST offers a point-and-click graphical user interface, an online tutorial, and help screens. These help users and reduce requirements for training and documentation. SAS also provides a menu-driven VT100 interface that allows students to work with data by dialing in from home or connecting across the Internet.

Universal availability

At the time this project was developed, Wharton was completing a migration away from large, centralized VMS systems toward distributed UNIX systems. By the beginning of the Fall 1994 semester, all Wharton faculty research and student communications and instructional applications had moved onto UNIX workstations.

Wharton provides the complete series of data sets on two UNIX systems—a DEC 5000/260 running Ultrix serves data in the “little-endian” format required by Ultrix, VMS, and MS-DOS and an HP 9000/755 running HP-UX serves data in the “big-endian” format used by HP-UX, Solaris, and most other UNIX systems. Identical directory structures are maintained on both systems.

One of the two systems is NFS mounted (so that it appears as a locally accessible device) on UNIX systems throughout Wharton, including shared departmental systems, personal UNIX systems on faculty desktops, and UNIX systems used by Wharton students for e-mail and instructional applications.

Because the data sets are physically stored on only two systems, the data can be centrally maintained while appearing to be available locally on all systems at the School.

Graphical access with X-Windows

Faculty with desktop UNIX systems use SAS/ASSIST in the X-Windows environment to manipulate and analyze data sets as local resources. For students in computing labs, Wharton uses Exceed’s X-Windows server running under Microsoft Windows to provide the same graphical access to data. Networked classroom teaching stations are configured identically to lab stations, allowing faculty to bring the data sets into the classroom (see sidebar).

Advantages

The Wharton Research Data System offers a number of advantages for Wharton faculty and students, and meets the goals of universal availability, ease of use, and reduced maintenance and support.

The entire collection of data sets is now available as a local resource on UNIX systems throughout Wharton—including shared departmental systems and faculty desktop systems. Students can access these data sets with the same graphical environment by using X-Windows in the student labs. Users can now manage and analyze the data using a single tool. Because the same data tool is used for all data sets, users can easily analyze data across different data sets.

Professor Richard J. Herring, who is working on a
large-scale study of the financial services industry sponsored by the Sloan foundation, states that “the beauty of the system is that wherever you go, whatever system you use at the School, the data is accessible and appears in the identical form.” According to Dr. Herring the key benefit this provides is that it “reduces the time researchers spend extracting data and allows them to concentrate on their analysis.”

The SAS access tools are generally standardized across different platforms—VMS, HP-UX, Ultrix, DOS, etc.—giving users the ability to continue their analyses in a number of computing environments. Many other statistical packages such as SPSS, BMDP, and SPlus can easily read SAS data sets.

In the future, Wharton Computing and Information Technology plans to further enhance this data architecture. In addition to adding more data sets, Wharton plans to provide transparent data access to an even wider range of systems at the School. Wharton is currently testing SAS for Windows and PC-based NFS client software with the intention of mounting data sets onto DOS/Windows systems at the School, providing access to vast data resources with the Microsoft Windows “look and feel” familiar to PC users. This will be another step forward in Wharton’s goal of providing information “at the fingertips” of its faculty and students.

**KENDALL WHITEHOUSE** is an Associate Director for Wharton Computing and Information Technology; **PAUL J. RATNARAJ** is an Information Management Specialist (responsible for the Wharton Research Data System) for Wharton Computing and Information Technology.

---

**Teaching with live data**

Wharton’s primary objective in developing the Wharton Research Data System was to assist in faculty research and student instructional exercises. Because Wharton’s networked classroom teaching stations share the same configuration and connectivity as Wharton’s computing lab stations, this project also allows faculty to use data sets “live” in the classroom.

This past year students in Michael Phelan’s Statistics 701 class were taught regression and time-series analysis using SAS/ASSIST to manipulate economic data. “The seamless access to Wharton’s financial data sets has been a key component of my curriculum development,” reports Dr. Phelan. “It has influenced my lecture style, which now combines a formal lecture with a directed recitation.” According to Dr. Phelan this teaching technique “allows the students to see the data unfold in real time.”

Frequently students will present suggestions for new ways to analyze data. Dr. Phelan points out that to use this technique in the classroom “you need the flexibility to travel down these new paths, but have the focus to be sure you cover the essential material.”

The consistent architecture across faculty desktops, student labs, and classroom teaching stations provides further advantages for both faculty and students. “By working with the same software configuration and data sets that students use for their assignments, I am able to discuss statistical concepts while demonstrating specific techniques in class that students can later use in their own work.”

This rich teaching environment hasn’t been without its difficulties. Using SAS to bring large data sets into the classroom requires careful coordination of a number of computing systems at the School. In the classroom, Dr. Phelan launches an X-session to connect to a UNIX workstation. The X-server, Exceed/4, runs under Microsoft Windows on one of the Novell Netware servers that supports Wharton’s DOS/Windows computer labs. The UNIX workstation is an HP 9000/755 that has the Wharton Research Data System NFS mounted from another HP 9000/755. To display the X-session to the class requires a high-resolution projection panel capable of projecting at a resolution of 1024 by 768 pixels. In classrooms without high-resolution RGB three-gun projectors, Wharton uses an In-Focus PowerView 950 high-resolution LCD panel.

One failure anywhere in the system means Dr. Phelan must resort to traditional teaching materials. “I always make certain I can cover the material for that day’s class using back-up materials if necessary. Although the system usually works as expected, we’ve had one or two unanticipated problems that forced me to abandon ship and take to the blackboard.”

Although Dr. Phelan admits that technological glitches can be frustrating, he concludes that “I can’t imagine teaching any other way.” He claims that not using this technology to teach statistics would be like “trying to teach someone to ride a bicycle by simply describing how to do it.”