Managing Large Financial Data with Ease: CRSP, COMPUSTAT, Etc., with SAS®

Paul J. Ratnaraj    Carol M. Katzman
The Wharton School of the University of Pennsylvania, PA

Abstract

During the past decade use of large and complex data sets across academic environments has grown substantially. This usage explosion has compelled schools to experiment with better access and management methodologies for a diverse and very dynamic user environment. In addition to budgetary constraints, there are constant changes in computing technology. This paper describes the changes that the Wharton School is making to better manage its information delivery strategies and objectives. The paper highlights the transition from using traditional languages like Fortran and in-house developed systems to commercial database management products, to SAS and its newly released SAS/ETS® DATASOURCE procedure. The paper also discusses the merits and disadvantages of the various methodologies and the reasons for using a unified data management and analysis tool for information delivery in an academic environment.

Introduction

The Wharton School of the University of Pennsylvania, one of the premier business schools in the nation has over 2,000 undergraduate students, 1,500 graduate students and 200 standing faculty with varied computing needs. Like other academic institutions use of large research data sets has grown substantially in the past decade. In addition, the data sets have grown significantly in size. What was once the domain of finance and accounting has now become the province of management, marketing, and other disciplines as well. Managing these data sets now requires extensive resources in terms of time, personnel and support.

The wide usage has compelled schools across the nation to experiment with better access and management methodologies. Some have opted for database management products, while others have home-grown systems developed to meet their needs. This paper describes the changes the Wharton School is making to better manage its data and access methodologies. The paper highlights the transition from using traditional languages like Fortran and C and in-house developed systems to using SAS to access and analyze data. The paper also discusses the pros and cons of the various methodologies.

History

Data Sets

Financial data sets generally contain time-series data. Some of the files have data beginning in the 1920's. Many of them are very large and require more than a gigabyte of storage. Below is a brief description of the widely used financial data sets at the Wharton School.

The CRSP data files were developed by the Center for Research in Security Prices (CRSP) by the Graduate School of Business at the University of Chicago. The CRSP data files include Stock files (comprehensive information on securities traded on the stock exchanges), Bond files (Government bond information), and Indices files (summary information on individual and combined exchanges).

Standard and Poor's Compustat standardized data files contain financial, statistical, and market information. They provide over 300 annual and 100 quarterly Income Statements, Balance Sheets, Flow of Funds, and supplemental data items on over 7,000 publicly held companies. Like CRSP, there are multiple files including information on principal products manufactured and industry indexes and composites.

Citibase is a macro-economic database maintained by Citicorp Database Services. It contains over 5,000 citations of monthly, quarterly, and annual time-series data going back to 1946. The data is on a national level, taken from over 100 Government and private publications. Financial, Manufacturing, Price indexes, Population, Labor, and Energy are some of the subject areas included in the time-series data file.

Environment

Based on data usage reports and experiences in Wharton's Consulting Office users can be divided into three groups: a) Faculty and Ph.D. candidates, b) MBA candidates, and c) Undergraduates.

Faculty and Ph.D. candidates are primary users of the data. They tend to be moderately computer literate. They are comfortable with using large subsets of the data and often use all the data. They prefer to have all data on disk for quick and easy access. They like to access and analyze related data sets together -- the price information from CRSP together with the
accounting information from Compustat. Their research and usage tend to span over many months and sometimes many years. Thus, an ideal environment for this group is one in which they are shielded from changes to the data and its access methodologies.

MBA candidates and undergraduates require the most support (ease of use, help with programming, documentation, etc.) while working with the data sets. This group has one basic concern: They want to know what is available, where it resides, and how can they get to it immediately with the least effort. Like faculty and Ph.D. candidates, MBA candidates and undergraduates do not wish to be burdened with having to deal with tapes and other storage media. The usage of data is generally limited to class projects or to assignments. The MBAs are more likely to be familiar with spreadsheets and PC statistical software. Therefore data access for them would include interfaces where they can import the data without leaving their spreadsheet or statistical package environment. The MBAs are reluctant to learn any software or programming language that they perceive will have very little value in their future work environment. The undergraduates would prefer accessing and analyzing the data with user friendly menus.

System Overview

Due to storage constraints and the size of data sets (some of the CRSP data sets require close to a gigabyte of storage), the data was generally converted from ASCII (native format) to binary using Fortran. Fortran, the oldest high-level programming language was used mainly by scientists, engineers, mathematicians, and academics. Fortran was designed to meet their needs, and thus was considered a logical solution for analyzing data since many mathematical and statistical sub-routines were also available in Fortran. Furthermore, due to the size of these data sets and the extensive processing power needed to analyze them, mainframes and mini-computers were needed to store them.

Figure 1 illustrates the software necessary to access and analyze the data. Users have to use Fortran to access data. To access Compustat users can also use FAME software. To analyze data, users could choose between Fortran, FAME, or a statistical package. For users familiar with Fortran, this methodology of access and analysis was not too cumbersome. However, for a non-Fortran user, obtaining the data was a difficult and time consuming task.

For example, a user who wants to analyze data using Minitab on a PC would have to do the following: 1.) develop a Fortran program to obtain data based on his or her parameters. 2.) write the data out in ASCII on the mainframe. 3.) download the data to a PC. and, 4.) read and analyze the data in Minitab. Thus, the user would need some knowledge of the data sets (variables and formats), data types (ASCII, binary), mainframe operating system commands to manipulate files, an editor or word processor to develop the program, Fortran commands (including linking, compiling and executing), data transfer protocols (like FTP), PC commands, and Minitab.

Figure 1

Limitations

Most data distributors ship their data with sequential-access Fortran programs. To avoid the constraints of access speed, complex in-house indexing programs are written. Interactive modules and help screens have been developed to alleviate some of the cumbersome tasks necessary for novice users. Naturally, any changes in the data structures and/or computing configurations involve maintenance of existing code and programs to preserve its existing functionality and performance, as well as increasing functionality and performance throughout its life-cycle [Boehm, 1981]. Tight budgets and time constraints do not allow for major re-writes of the programs which results in a system that resembles a patchwork quilt with diminishing returns in performance and functionality.

Because of the varying needs of users, it is extremely difficult to develop a system that meets most of their requirements. This leads to programming beyond prewritten applications. The deficiency is managed by extensive support both in programming and in the explanation of the data and its characteristics. Using work-study and student help has not been a successful option. Students seldom have the skills and time-schedule necessary to efficiently support the programming needs. Thus, skilled computing personnel hired to plan, develop, and implement new systems and strategies are spending their time supporting the immediate user needs.

The computing environment is changing constantly. Keeping up with the changes on a theoretical level itself is a major task. Supporting these changes for a dynamic and totally diverse group is a monumental undertaking. The current trend in computing is downsizing from the mainframe to Unix and to personal computers -- from command line and batch submission.
to point-and-click. Users exposed to Windows on PCs are now requesting graphical user interfaces (GUIs) to access and analyze data. Patching a GUI on top of Fortran just adds to the complexity of maintaining these large data sets and may not be the best solution. Nor does it address the needs of the future -- empowering users to manage their processes and data.

The Current Environment

System Architecture

After numerous meetings, discussions, informal surveys with users and computing colleagues, we decided that a single framework was essential for data manipulation, i.e., data management and analysis. The functionality of the system was required to be consistent on all platforms, i.e., mainframes, workstations and personal computers. For example, users working on PCs should be able to access the data on a Unix workstation without having to learn the operating system commands of Unix. Adequate documentation to support users should be available. The system should be able to employ the point-and-click technologies currently available. The system should empower users by letting them choose their favorite applications. This would give users the authority and enable them to share responsibilities for solving their problems [McClendon, 1993].

Evaluation

Developing an in-house system that met the varied needs of the users would involve a tremendous effort in personnel, time, cost, and subsequent maintenance efforts. In addition by the time the system was developed it would become obsolete in this constantly changing technology environment. Finally, the best designed in-house system would still be restrictive in one fashion or another defeating our stated goal of empowerment.

Commercial database management systems (DBMS) that offer excellent data management capabilities are widely available. User-friendly and powerful access interfaces can easily be developed. It is also easy to develop subsets in a DBMS. However, lack of good analytical capabilities is a primary disadvantage. A DBMS is not suitable for time series data and is considered very slow for large passes of data. It also requires additional database procedures and skills by the users.

Other commercial software like FAME, DART, and Intelligent Query provide reasonable data manipulation abilities. Like DBMS, they lack sophisticated analytical capabilities. We have been using FAME for the past two years and used DART two years prior. Both require programming to convert data to their respective formats. FAME allowed the programs to read Compustat data but required us to program for the remaining data sets. Both FAME and DART provided interactive tools for users; however, faculty and Ph.D. candidates felt that the software was not flexible enough for their needs. Moreover, we understand FAME is moving towards data redistribution, i.e., supplying us with the data in its native format. This would significantly eliminate the flexibility that we need for our diverse user community. Intelligent Query was very slow for obtaining large amounts of data since it had to make a sequential pass to read the data.

Since its inception, SAS has been used in academic environments for its data analysis capabilities. In the past few years, there has been a considerable trend in academic environments towards using SAS for data management and application development. With release 6.09, SAS has made reading external data files much easier. PROC DATASOURCE extracts time series data from many different kinds of data files distributed by various data vendors and stores them in a SAS data set. This procedure generalizes and supersedes the CITIBASE procedure. Naturally, the converted data set can be used by other SAS procedures. Extracting data takes a few lines of code in SAS, whereas in Fortran it takes hundreds of lines. SAS also met our objective of a single unified tool for both data management and data analysis.

<table>
<thead>
<tr>
<th>Evaluation Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>SAS</td>
</tr>
<tr>
<td>Sequential</td>
</tr>
<tr>
<td>Direct-Access</td>
</tr>
<tr>
<td>Access Speed (Seconds)</td>
</tr>
<tr>
<td>Disk Storage (Mbytes)</td>
</tr>
<tr>
<td>System Maintenance</td>
</tr>
<tr>
<td>Support - User</td>
</tr>
<tr>
<td>Support - Vendor</td>
</tr>
<tr>
<td>GUI - based</td>
</tr>
<tr>
<td>Integration - Analytical tools</td>
</tr>
<tr>
<td>Integration - other Datasets</td>
</tr>
<tr>
<td>Integration - other languages</td>
</tr>
<tr>
<td>Ease of use</td>
</tr>
<tr>
<td>Portability</td>
</tr>
<tr>
<td>Menu Driven</td>
</tr>
<tr>
<td>Installation (avg. #hrs per file)</td>
</tr>
</tbody>
</table>

Table 1

We evaluated PROC DATASOURCE against our existing system. We chose the CRSP Monthly Stock File since it is widely used and users find the data structures very challenging. All testing was done on a DECstation 5000/25 workstation as a single user.
6.08 for Unix was used along with ULTRIX 4.2a. The SAS data set was indexed and compressed. The sequential access Fortran code was supplied by CRSP. The direct-access Fortran code was written in-house.

100 Cusips (Company IDs) out of 6713 observations were culled from the CRSP monthly stock file. 10 Cusips from each of the first alphabets of the company's name (A,B,C,D,I,L,N,P,S, & W) were selected. This ensured that data was selected from all sections of the file, (i.e., top, middle, and bottom). One year of data was written out.

In addition to access-speed of the three data formats -- SAS, Fortran sequential access, and Fortran direct access -- other factors like storage, maintenance, and user support were also evaluated (see Table 1).

Maintenance includes the task of updating code, data dictionaries, and documentation if there are changes in the data structures. When code is supplied by the vendor, maintenance is low as code is updated to accommodate any changes made to the data. Similarly, the onus is on SAS to update its data reading procedures if the vendor makes any changes.

User support refers to training and debugging of code. It is labeled low to high depending on whether the user has to use Fortran. To conceptualize a sequential data set is a lot easier than conceptualizing an indexed data set. The user will also need other support in terms of transferring data to other software or platforms. SAS has excellent telephone support, documentation, and help screens to aid the user. The user can also rely on a wide network of SAS users in the academic community. With SAS/ASSIST the user has available the point-and-click technologies.

Integration with analytical tools is the ability to manage data and analyze it within the same program. For non-Fortran users, SAS provides the ability to obtain and analyze the data without using other software. SAS also enables the use of other analytical tools and applications if so desired.

Integration with other data sets is the ability to incorporate other data sets into a user's analysis. For example, a user might need to analyze financial data (Compustat) and equities data (CRSP) simultaneously.

Integration with other languages is the ability to use Fortran and C libraries. Even though SAS offers the capability, it is not as efficient as Fortran and C.

Portability refers to executing the programs across different platforms, i.e., VMS, HP-UX, ULTRIX, DOS, etc. SAS is generally standardized across all platforms. This is not true for Fortran. Other statistical packages like SPSS, BMDP, and SPLUS can easily read SAS data sets.

Storage refers to the CRSP Monthly Stock File. The SAS data set was indexed and compressed. The direct-access storage could have been reduced with better indexing methodologies. We also tested other data sets. On average, converting time series to SAS data sets needed 4 times the storage when compared to data maintained in Fortran binary accessed sequentially.

Installation time refers to time spent on converting ASCII data including updating code (if necessary), building data dictionaries, sub-routine libraries, utilities, and backup of data sets. SAS has numerous procedures to carry out many of the aforementioned tasks.

Implementation

We are implementing the conversion of data into SAS data sets in phases. The first phase included conversion of all data sets except the CRSP Daily Stock File and Nasdaq Stock File to SAS data sets.

We also have the data converted to Fortran sequential access binary for users familiar with Fortran and the old system. Additionally, the Fortran binary copy was maintained as a precaution against any errors that might arise with the SAS data sets, both in terms of integrity as well as its management of memory, CPU, and storage.

We have an X-Windows interface available on all our PCs. We encourage users to use SAS/ASSIST on Unix. The GUI features are very appealing to most beginners.

Simple data extraction procedures have been documented. The users are encouraged to purchase the SAS/ASSIST documentation at the University book store. They also have the option of referring to the documentation in the Consulting office.

We offer training courses in SAS to the Wharton community. Students who attend these courses can use one statistical package both for extracting data as well as analyzing data. They can also use this package for other analyses in their research and coursework.

The second phase will involve converting all data sets to SAS data sets. Some Fortran data sets will be eliminated. We may choose to use SAS/CONNECT on a PC to access the data on Unix.

The Third phase will eliminate all Fortran data sets except for CRSP Daily Stock File and Nasdaq Stock File.

We are continually monitoring the usage of these data sets, and our future implementations are largely dependent upon the findings.
Conclusion

The massive storage required for SAS data sets appears to be the primary disadvantage.

There is substantial overhead in terms of memory and CPU usage. Since we are currently using SAS, we have been able to discount the overhead.

We believe that SAS provides a balanced yet flexible interface for our data management requirements.

SAS is already a strategic tool for analysis at Wharton.

SAS' GUI capabilities, on-line tutorial and help screens reduce user support and training. Users can also obtain additional support from their peers since SAS is widely used.

SAS is used in all kinds of businesses, especially in the financial and accounting worlds. Using SAS enables students to learn and use software relevant to their current studies as well as their future working environment.

Those using SAS to extract data are more likely to continue using SAS as a data management and analytical tool for other needs, which decreases the need of supporting multiple packages.

Faster access to the data is enabled if SAS is the choice for analysis. The SAS system is optimized to deliver data to its own procedures.

The portability of SAS and the ease of transferring data to non-SAS applications allow users to continue their computing tasks in any environment. SAS functionality is consistent across multiple platforms.

Using SAS decreases the dependence on unique code that requires maintenance, upgrades, and skilled personnel.

The pros clearly outweigh the cons.

References


SAS, SAS/ASSIST, SAS/CONNECT, SAS/ETS are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Authors

Paul J. Ratnaraj
Wharton Computing and Information Technology
The Wharton School
3620 Locust Walk
Philadelphia, PA 19104-6301
(215) 898-7602
email: ratnaraj@wharton.upenn.edu

Carol M. Katzman
Wharton Computing and Information Technology
The Wharton School
3620 Locust Walk
Philadelphia, PA 19104-6301
(215) 898-4804
email: katzman@wharton.upenn.edu