

MARCH 2006



YOUR SAS® TECHNOLOGY REPORT

THE POWER TO KNOW.

Dear Readers,

We strive to make this newsletter as helpful and informative as possible. That's why we recently conducted a random survey of you, our readers. Thank you to those who responded to the survey. Your comments and suggestions are invaluable.

One common thread among all the replies is the desire for more technical tips. So, beginning in April, you'll see an increase in the number of tips in each newsletter. These tips come from SAS technical support and are designed to help you in your everyday work.

And speaking of help, in this issue, you'll find information on data integration and SAS Forecast Studio, plus a Live Web class on business intelligence. And don't forget about the events running in the left column. You're sure to find something useful in the list.

Happy learning!

A handwritten signature in black ink that reads 'Shelley Sessoms'.

Shelley Sessoms

Editor, *Your SAS Technology Report*

FAQ

Q: How Can I Use PROC TRANSPOSE to Get One Record Per BY Group?

A: Because of the structure of a data set, using the BY statement does not always guarantee that PROC TRANSPOSE will create an output data set with one record per BY group. In some cases it is necessary to use a BY variable that is unique for every record (if one does not already exist in the data set, use `_N_`), and then transpose the data set a second time.

```
data test;
  n=_n_;
  input x y z;
cards;
1 2 3
1 2 3
2 3 4
3 4 5
3 4 5
4 4 5
;

/* In order to keep the implied ID variable _NAME_ */
/* from causing errors in the second transpose, */
/* eliminate it using the drop= data set option */
proc transpose out=first(drop=_name_);
  by n x;
  var y z;

proc print; run;

proc transpose data=first out=second prefix=a;
  by x;
  var coll;

proc print; run;
```

Power Talk with Mike Zdeb

Mike Zdeb, assistant professor at the [University at Albany School of Public Health](#), has only taken one programming course in his life. But that hasn't stopped him from being a die-hard SAS user for more than 15 years.

"My first introduction to SAS came at the New York State Health Department," says Zdeb. "We received a SAS license in the late 1980s, and it was our first statistical or data management package."

Zdeb is a completely self-taught SAS user. "I have never taken a SAS course," he admits. "I was just very curious about the package, and I did an awful lot of reading."

"I came at SAS not from the perspective of a programmer, but rather like a researcher or statistician," comments Zdeb. He views his lack of a programming background as a positive thing.

"It turned out to be good for me because most of the people at the health department who used SAS were not programmers. They were more like me: epidemiologists or public health workers who needed some kind of tool to help them organize, manage, present and analyze data," says Zdeb. "Because of my curiosity and seeing how apparent it was that SAS could help both me and my co-workers, I ended up being the SAS educator for the health department."

Zdeb ended up teaching classes and writing short manuals for the health department. "It was easier for me to explain SAS concepts to my co-workers because I was just like them, not like the systems people who were better at explaining how to run SAS on our various platforms rather than at how to use it to accomplish specific tasks."

Practical experience

"Once I started learning SAS, I did everything with it," notes Zdeb. "I gave up using anything else because SAS was so intuitive."

While at the health department, Zdeb also taught part-time at the University at Albany School of Public Health. After he retired from the health department in 2003, he continued to teach as an assistant professor and has now taught SAS for the past 10 years.

He has also continued to assist staff at the health department on SAS-related projects. Zdeb recalls an early project at the school involving SAS.

"Back in the 1970s, I wrote a paper on the probabilities of developing cancer. The paper ended up in the *American Journal of Epidemiology*. When I originally wrote the paper, there was no SAS, and no personal computer," he notes. "I had to do everything in FORTRAN running on a mainframe. One of my master's students was working on a thesis project, so we reran the entire probability aspect of the paper using SAS. The student got a thesis paper out of it, and I was able to present the paper at a Northeast SAS Users Group [NESUG] conference."

Zdeb notes that the original graphics for the paper were all hand drawn by an artist. When the paper was redone with SAS, the images were all done automatically using SAS'GRAPH®.

Teach what you know

At the School of Public Health, Zdeb teaches an introductory three-credit SAS course. He also teaches an advanced one-credit course that takes what was taught in the introductory class and shows students how to apply their knowledge to the various problems they may encounter in the work force. His students learn macros, graphics and SQL in the advanced course. Zdeb would like to include more in the course, but “there’s just not enough time to teach everything that SAS can do.”

Many of Zdeb’s students are physicians who have no experience using analytic software, but they are interested in being able to work with their data on their own. He encourages his students to take the [SAS Certification Exam](#), and several of them have taken and passed the exam. You can read their [stories](#) on the Web. One student came to the school from [Mongolia](#) to get a Master’s degree in public health. He attended Zdeb’s classes and passed the certification exam. “My students have taken the intro course, gotten an internship and used SAS in real life. Then they take the exam and do pretty well,” says Zdeb.

Zdeb also shares his SAS knowledge with people outside of the university. Many people find his [Web site](#) and then e-mail him with SAS questions or requests to use his notes. Former students also call or e-mail him with questions. “I tell my students my courses come with a lifetime warranty, and they really take me up on it.”

Zdeb also receives notes and e-mails full of gratitude. “I just received a Christmas card from an ex-student saying she loves her job and that whenever she has a question, she looks first at my notes and then grabs her book,” he notes.

Conference involvement

Another way Zdeb stays active in SAS is through his participation in regional and international conferences. He attended the first NESUG conference in 1988 and has been a consistent participant ever since. He is actively involved in the group and serves as the CD creator. He began putting the proceedings on CD in 1997 and has continued every year since. “I enjoy the regional conferences because it’s all done by volunteers,” says Zdeb. “I have been attending SUGI more and more lately. In fact, I presented papers at the past two SUGIs and will do so again at [SUGI 31](#) in San Francisco.”

In 1998, Zdeb also wrote a SAS book, *[Maps Made Easy Using SAS](#)*. In it, he includes abundant real-world examples and a tutorial approach to help new users create maps easily and quickly.

As a self-proclaimed SAS proselytizer, Zdeb has used and taught SAS for years...and he knows what he’s talking about. His experience and knowledge make him a valuable resource, both for SAS users and SAS staff, and for that we’re eternally grateful. Thanks for your use of SAS over the years, Mike!

New Introduction to SAS Business Intelligence Live Web class!

Learn how to use the SAS® 9 BI Server throughout your organization. Find out how anyone can turn data into useful information for better business decisions. Using a Web browser and telephone, the Live Web class provides you with an instructor who will guide your learning and answer your questions in real-time – all from the comfort of your desktop. Register today for the upcoming April 12-13 class.

Read more <http://support.sas.com/training/us/crs/lwsbio1.html>

Data Integration

For years, the key to success for any business intelligence solution has been the process known as extract, transform, and load (ETL). Selecting the right tool to meet an organization's need for gathering data from disparate sources, and transforming the data before loading it into a target destination, was the critical factor in building a data warehouse or a data mart in order to support an organization's BI projects. ETL became so important that the process became synonymous with the tool, and the technology became known as ETL technology, which spawned many ETL tools. Numerous organizations have struggled with selecting several tools over a period of time to gain access to new data sources. Some organizations failed to see the benefits of the tools over custom coding, which results in small armies of programmers building and maintaining code. Consequently, organizations ended up with several ETL tools acquired via mergers and acquisitions or by allowing departments to operate unchecked with their tool-of-choice. The problems with using several tools (because of individual tool limitations) or custom code significantly increase the total cost of ownership in terms of maintenance, training, and time lost in regaining familiarity with a rarely used tool. In addition, using several tools can lead to very fragmented metadata, which makes topics like compliance a chore rather than something that is automatically delivered through self-documenting metadata.

In addition to the proliferation of ETL tools, building and maintaining a data warehouse or a data mart is no longer the only activity occurring in organizations when it comes to their data, and BI (while still a powerful driver) no longer stands alone. Organizations are increasingly finding additional non-warehouse projects such as system migration, system consolidation, and system synchronization as a result of mergers, de-mergers, acquisitions, and an overall need to update older systems to their modern equivalents. Some of these, when taken in their near real-time and more batch form, are supported by an ETL process. Many others such as master data management and real-time synchronization/data quality to maintain integrity of operational systems, are fast emerging as a critical theme in most organizations, demand new technologies.

This new expanded scope has led to the emergence of the topic **Data Integration**. Data integration should be a strategic topic in all organizations because it affects everything that an organization does. It's time to move from an ad-hoc approach and look at data integration as something that can contribute significantly to your competitive advantage. It's time to think about standardizing as much of your data integration, including ETL, with one "system-neutral" vendor in order to leverage the synergies across the spectrum of data integration such as shared business rules and metadata. In addition, you will see reduced training costs, reduced maintenance costs, and benefits on the operational and BI activities of having one, consistent, integrated set of technologies. It's time to ensure that the experiences of the ETL era are understood and to establish one way to success.

The Details

Data integration can be considered as the convergence of multiple technologies and the emergence of some new ones. Broadly speaking, data integration brings together technologies that are typically needed for the operational side of the business with technologies that are needed for the BI/Decision Support side of the business. Data integration deals with all types of data that has to be incorporated into a unified whole in

an organization. Data integration cannot be seen as just “a means to an end” because, in many cases, data integration supports operational processes or keeps operational systems in sync, but is not necessarily directly driving things like BI and analytics do. Perhaps, it is this shift in focus that most characterizes data integration, and is the reason why data integration technologies that come from RDBMS vendors are somewhat limited; they are still too focused on the BI world. With a cohesive, data integration strategy, the major focus needs to be on the non-BI aspects that are affecting organizations today, at a time when many organizations have not yet resolved the issue of ETL/data integration for the purpose of supporting data warehouses and data marts.

If we agree that this is an important topic, how do we move forward?

As with all things, a data-integration strategy brings some “buy” vs. “build” choices to organizations. Because of the way that the portfolios of most vendors in the market have evolved (that is, through mergers and acquisitions), there is a third method: buy and integrate the tools even if they are provided by a single vendor. This is the same tool integration that would be required if you take a “piecemeal” approach and buy from several vendors to meet all your needs. Organizations that want to establish a data integration strategy should *learn how all the capabilities were added* to a portfolio (integrated through in-house development or purchased through acquisition), and if things such as metadata, business rules, etc., can be shared, not just if they exist. If they can’t be shared, when and what will the migration steps be? Organizations should be careful not to be “taken in” by descriptions of manual steps in order to get the bigger picture. Manual steps introduce overhead and risk, and many hidden costs and risks can suddenly become apparent.

High-Level Guide to Data Integration

If you are new to data integration, you might be wondering just what you should expect of a data integration solution. To start, let’s establish what data integration is NOT. Data integration is not about Enterprise Application Integration middleware, although it does make use of that in some cases. It’s also not about message queues and application servers; even though these are important parts of the infrastructure that will support certain aspects of data integration. There seems to be an impending desire to force these topics and what is commonly seen as “Middleware” into the broader data integration domain because it suits specific vendors. Do not be confused—tying your infrastructure to your data integration vendor creates a lock-in that might be difficult to ever get out of.

Let’s build the Data Integration Landscape beginning with what most people know about data integration, today, and develop it to include emerging topics and technologies.

Data Connectivity and Metadata: Although not part of the data integration landscape, the topics of data connectivity and metadata are very important in any data integration strategy, because they are pervasive in all the other parts of the landscape as key-enabling technologies and should be given equal consideration. Any data integration solution should provide both native access using standard utilities and open standard access (such as ODBC) to all major structured data sources such as relational databases, flat files, ERP systems, and mark-up languages such as XML for reading and writing. In addition, the connectivity should facilitate the access of information on many different systems such as z/OS, UNIX, and Windows, preferably without having to make

use of intermediate files and extracts. Support for the reading and writing of data from message queues and the ability to receive and send data to/from Web services should also be provided by the solution. In the longer term, the solution needs to continue evolving in order to support unstructured data sources.

Metadata should be pervasive through all types of data integration. Data integration is about relating multiple data sources and bringing them together to make your data more valuable. Metadata provides the definition across data sources that make this possible. In addition, metadata enables you to trace what moved when, how it was changed, what business rules were applied, and what impact those changes might have. These are critical issues facing all organizations. Failure to place enough emphasis on metadata will result in problems later on; often at great cost to an organization.

Data Quality / Real-Time Data Quality Integration: Any data integration solution should include an INTEGRATED data-quality solution to support data-quality processes such as profiling; householding; deduplication; data-quality, business-rule creation; and cleansing of data (where required). These rules should also be callable through custom exits, messages placed in message queues, or Web services to trigger the process and deliver what can be referred to as Real-Time Data Quality Integration. A classic example is the checking of names and addresses at the point of entry into an ERP system, through the use of a custom exit, to build-in data quality from the start.

Data Warehousing/Data Marts (ETL): Any data integration solution needs to provide the capability to both build and maintain data warehouses/data marts via the ETL process. This solution needs to leverage the data connectivity capabilities that were previously mentioned and have fully integrated metadata. Such a solution should also include SUPPORTED. Here, SUPPORT means technical support and help from professional services as a part of the solution, and extensions through custom coding so that organizations have the flexibility to do more than the tool delivers but will not lose the support of the vendor when they use custom code, thus reducing risk. In addition, the solution must allow for the re-use of data-quality business rules that are provided by the data-quality part of the data integration offering. Data quality must take “center stage” in any integration strategy.

Data Migration: Any data integration solution needs to provide the capability to migrate data from multiple existing systems to one or more new or existing systems. You could argue that, in its most primitive form, this is just the application of the ETL process with data quality (Why migrate bad data forward? Why not clean it up and enrich it and deliver business value from what is normally perceived as strictly an IT project?) with the target not being a data warehouse or a data mart. Organizations should be looking to build up data-quality business rules over time (and from data migration project-to-project) that can be applied whenever a migration takes place in order to get re-usable, immediate, and low-cost business benefits. These same rules should be usable when supporting data warehouse and data mart creation/maintenance.

While a one off migration might often take place, it is also likely that, on the operational side where the source system might live, it will be very hard to achieve. This is because organizations often have business applications running from the operational system to be migrated so that movement forward will first involve migrating the data to a new system and verifying its correctness (again, this is where metadata becomes vitally

important), before establishing an ongoing data synchronization process between the old and the new, and placing the business application on top of the new system for acceptance testing. After you are satisfied that the data in the new system is up-to-date and that the business application is operating as expected on the new system, the old system can be turned off and data synchronization ended.

Data Synchronization: Any data integration solution needs to be able to reflect that the changes that are made in one system are also made in other systems in the organization. There are two types of data synchronization. The first type is the movement of “changes” made in one or more systems to other systems in batch/near real-time. The second type is the movement of “changes” made in one or more systems to other systems in real-time. The first type of data synchronization is just another application of the ETL process using change-data capture and a scheduled process to move data around. This process can be scheduled nightly, every 30 minutes, every 5 minutes, or even lower depending on the needs of the organization and the amount of data to be moved. However, it typically involves the movement of multiple “transactions” or “records” at once. The second type of data synchronization involves the movement of “individual transactions” or “records” to synchronize status across multiple systems as the transactions occur and in real-time. Technologies such as message queues and brokers are often used in such circumstances. Here, a real-time server needs to be invoked by using custom exits, messages placed in message queues, change brokers, or Web services to trigger the process.

Again, it is important to note the importance of data quality in data synchronization. Although bad data in one system is not good, the proliferation of bad data through data synchronization can have a devastating effect. Organizations should ensure that any data synchronization also includes the application of data-quality business rules to maintain the quality of data throughout all systems.

Master Data Management (MDM): Any data integration solution needs to provide the capability to handle the new and emerging topic of master data management. *Master Data Management* is the practice of creating a single “perceived” truth through mapping multiple disparate definitions of items such as names of customers and products, which are held in various systems, so that people can ask for “customers” and have all the customers names returned in a common format that uses a standard, company-accepted definition for any application without having to understand the underlying structure in the various silos throughout the organization. Tied closely to MDM are emerging topics such as **Customer Data Integration (CDI)** and **Product Data Integration (PDI)** that build on the basic technology and deliver a number of common mappings and definitions to get organizations up-and-running, quickly. Where MDM is a topic of concern, organizations should look for the development of these more advanced solution areas that incorporate a true metadata management framework in traditional reference data management and speed up the time to deployment, thereby reducing overall costs. Ultimately, CDI and PDI are examples of real implementations that solve specific problems in the broader MDM space. Many organizations will have to solve one of these specific sets of problems first. However, the more forward-thinking organizations will have a broad MDM strategy that leverages many of the same technologies and capabilities to achieve common results within their enterprise. If a vendor says they do MDM but they do not deal with topics like CDI or PDI, then you

might be getting a very limited solution that will require a lot of ongoing, manual, and expensive custom development.

Data Federation / Enterprise Information Integration (EII): Any data integration offering needs to support Data Federation or EII, which is basically a form of data integration that keeps data in place and allows it (the data) to be integrated and surfaced as needed. Due to its dynamic nature, this technique can lend itself to potential problems where there is no need to access large amounts of data or data from many underlying systems. Data Federation and EII, along with Data Synchronization, are often the underlying technologies that are employed with MDM and, also, often used with BI solutions where more operational or real-time views of data are required.

Conclusion

There is no doubt that organizations will set priorities on which data integration tasks take precedence, but how many will avoid the mistakes made with ETL in the past and ensure that a short-term strategy is backed by long-term integrated possibilities? SAS in conjunction with DataFlux, a wholly owned subsidiary of SAS that focuses on the data quality aspects of data integration and real-time data integration, delivers a variety of integrated solutions to meet various needs that can be brought together, incrementally and in a variety of ways, to suit the needs of your organization. You can start out with a solution to address Master Data Management, or you can start with technologies to build data warehouses and data marts or to carry out rudimentary data profiling. The important thing is that, whichever choice you make and whichever direction you subsequently take, SAS and DataFlux can deliver all the technologies that you need to establish a Data Integration strategy while realizing the benefits of shared business rules, shared metadata, and integrated technologies with the associated cost reductions because your employees need less training, and you can re-use business rules. In addition, you'll have less inherent tool and metadata integration and fewer maintenance and management problems, which are alleviated by an integrated comprehensive approach.

If you are not doing so already, today might be a good time to start deciding where the future of your data integration strategy lies. All the topics in the preceding landscape should live and work together to give you maximum benefit and value. Previously established piecemeal standards need to be challenged, and time is not on your side. The most successful organizations will have a clear and precise strategy in place for data integration as a fundamental cornerstone of their competitive differentiators. Those who succeed will be the leaders who can address all their needs by using one integrated offering, thereby having the flexibility to react to new challenges quickly (much re-use if a non-piecemeal comprehensive approach is taken). Those who hesitate will be quickly left behind in a sea of complexity and cost.

Data Integration should be complete, flexible, integrated, and proven. SAS and DataFlux provide all these strengths and are ready to help you address your challenges today.

SAS® Forecast Studio

This white paper provides a detailed overview of SAS Forecast Studio, a key component of SAS Forecast Server. The paper walks you through the process of generating automatic forecasts, viewing results, building models, publishing results, reporting and more. Read this paper to learn how SAS Forecast Server speeds the statistical forecasting process by providing a convenient, user-friendly interface for all the forecasting options available in SAS.

Read more <http://www.sas.com/ctx/whitepapers/whitepapers.jsp?code=333>

SAS Programming in the Pharmaceutical Industry

By: Jack Shostak

List price: 49.95 USD

360 pages

ISBN: 1-59047-793-6

Publisher: SAS Press

Publication Date: August 2005

Description:

At last! A real-world reference guide for clinical trial SAS programming, packed with solutions that programmers can apply to their day-to-day problems. Discover key techniques and tools available within Base SAS (including the macro language and PROC SQL), SAS/GRAPH®, and SAS/STAT® that can be used to resolve many common issues in working with clinical trial data. Organized to reflect the statistical programmer's work flow, this user-friendly text begins with an introduction to the working environment, then presents chapters on importing and massaging data into analysis data sets, producing clinical trial output, and exporting data. Valuable plug-and-play programming examples are provided throughout. Whether you're a novice seeking an introduction to SAS programming for the pharmaceutical industry or a junior-level programmer exploring new approaches to problem solving, you'll find a wealth of practical suggestions to help you sharpen your skills.

SAS Products Addressed: Base SAS, SAS Enterprise Guide, SAS/GRAPH

Releases: 9.1.3, 9.1.2, 9.1, 9.0, 8.2

Operating Systems: 64-bit Enabled AIX, 64-bit Enabled HP-UX, 64-bit Enabled Solaris, ABI+ for Intel Architecture, AIX, HP-UX, HP-UX IPF, IRIX, Linux, Linux on Itanium, Microsoft Windows for IPF, OS/2, OpenVMS Alpha, OpenVMS VAX, Solaris, Tru64 UNIX, VM/CMS, Windows, Windows NT Workstation, z/OS

Read more or order today!

<http://www.sas.com/apps/pubscat/bookdetails.jsp?&pc=59827&promo=EN>

SAS OnlineTutor - Self-paced e-learning for organizations

Deliver **cost-effective** computer-based SAS training to **multiple users** across your organization with SAS OnlineTutor software. SAS OnlineTutor is self-paced training for basic to advanced programmers that is:

- Installed on individual workstations or on a network
- SCORM conformant - for Learning Management Systems
- Licensed on an annual basis
- 508 compliant to meet the requirements of users with disabilities

With 24/7 easy access and a variety of topics, users can focus on learning the topics they need, when they need them. Certificates of completion are available for each lesson.

SAS OnlineTutor licenses are available for:

[Basic and Intermediate SAS programming](#)

[Advanced SAS programming](#)

Try it yourself today!

[Take a tour](#)

More information:

Contact your [local SAS office](#) for more information and pricing. (In the U.S., call 800-333-7660.) Or e-mail us at training@sas.com.

FAQ

Q: Printing text at the end of every page but the last one in PROC REPORT

A: A COMPUTE AFTER block is processed before the COMPUTE AFTER `_PAGE_` on the last page.

```
data new;
  set sashelp.class
      sashelp.class
      sashelp.class
      sashelp.class;
run;
```

```
PROC REPORT nowd data=new;
```

```
  /* Set a hold variable to any value */
  compute after;
    myval='x';
  endcomp;
```

```
  /* Check the hold variable for the value set */
  /* to determine if this is the last page */
  compute after _page_;
    length text $ 20;
    if myval='x' then text=' ';
    else text='continued';
    line text $12.;
  endcomp;
run;
```

Webcasts and Events

Focus on Building the BI Competency Center

March 15

1:00 p.m. ET

This seminar, featuring BI visionary Claudia Imhoff and Aiman Zeid of SAS, will show you how to exploit the benefits of BI Competency Centers.

SUGI 31

March 26-29

San Francisco

Don't miss the chance to network with, and learn from, SAS users and personnel from around the globe.

SAS Users Appreciation Reception

March 27, 6:00-7:30 p.m.

San Francisco

This reception, located in the Demo Area on Level 1, celebrates SAS users and our shared success, so we hope to see you there!

F2006

June 5-6

Cary, NC

Learn the latest forecasting theories, trends and best practices from world-renowned forecasting experts at F2006.

JMP® User Conference

June 20-21

Cary, NC

Attend exciting and insightful sessions, plus roundtable discussions, a Scripting workshop, a Genomics Discovery event and exclusive training courses.