Paper 132-2011

SAS® Global Users Group Data Warehouse: A Look Behind the Scenes

Gregory S. Nelson
Chair: Data Model Working Group

President and CEO
ThotWave Technologies, Cary, North Carolina

Abstract

Wouldn't it be nice to see how someone built an entire end-to-end system (from data acquisition to data cleansing to a fully integrated data warehouse) using SAS? This presentation shares some of the experience of designing and implementing a data warehouse. Topics include planning, architecture, data model, technology choices, and methods used in the creation of the platform. Special attention is paid to the challenge of data cleansing and enrichment and how we were able to engage the SAS community in helping us analyze this data.

Background

For over three decade, the SAS Global Forum (and its predecessor SAS Users Group International – or SUGI) has been supporting the SAS eco-system of users through an annual conference. The conference has grown from just a few attendees to over 3,000 annually.

In recent years, it has become evident that the volunteer organization that runs the conference (SAS Global Users Group) requires fact-based information so as to make informed decisions about the future direction of the conference and how to best support all of its stakeholders.

Heretofore, the SAS Global User Group Executive Board (SGUGEB) has made assumptions on attendee's patterns of behavior, opinions and even demographics based on anecdotes and gut feel along with limited data made available through registration and post-conference surveys. A small group of volunteers from SGUGEB was formed during the Summer 2005 to focus on what data would be useful to help us drive strategy and the questions we believe could be answered by this platform.

A strategy document was developed to outline these high level findings and a conceptual data model that showed the key subject areas of interest was developed. It focused on the capabilities needed by the stakeholders and the target users, and why these needs exist. Use cases were then outlined to detail how the needs of the stakeholders could be met from the data warehouse.

This paper outlines these efforts and we are sharing the strategy and tactical plans as a way to engage the community and to be transparent about these efforts with SAS Users around the globe in hopes of continually improving the way in which we connect and learn.

Problem Statement

The goal in a simple statement is:

> *to create a data warehouse to support strategic questions involving services delivered to a diverse population of consumers and to measure the effectiveness of our efforts by looking at patterns of attendance, sponsorship trends, conference and speaker effectiveness, and conference outcomes (attitudes and beliefs)*

The data warehouse was to become a planning tool for the SAS Global Forum Executive Board, SAS and potentially the regional user groups for future improvements of their conferences.

The fundamental challenge that we hoped to solve through this effort and the resultant system included:

- To build a fact-based infrastructure that helps us make informed decisions about the future of the organization
- Collect relevant information about the attendees (and prospective attendees) in order to make the conference better, attract new and returning attendees and increase/ maintain sponsorships involvement
- To help grow the volunteer base and future generations of leadership

This project hopes to address the following:

The problem of	*Describing our attendees, sponsors and their organizations in hopes of understanding their attitudes and behaviors about the conference*
affects	*The current attendees, prospective attendees, sponsors, SAS and the SAS Global User Group Executive Board*
the impact of which is	*To drive increased attendance and retention as well as increased sponsorship involvement*
a successful solution would be	*An straight forward to use, secure environment where questions about SAS Global Forum could readily be answered*

We hope that by developing a data warehouse with tools appropriate for the analysis, querying and reporting, we would be able to make fact-based decisions around the future of SAS Global Forum on behalf of our stakeholders. Currently, the data made available to us through registrations and post-conference surveys falls short in addressing the strategic nature of our questions.

Stakeholder and User Descriptions

To effectively provide products and services that meet our stakeholders' and our internal users' real need it is necessary to identify and involve all of the stakeholders as part of the Requirements Modeling process. SAS Global Forum is a conference designed to meet the needs of an entire eco-system of SAS programmers, statisticians, analysts and the people who support them. We want to ensure that the data warehouse captures the true nature of this stakeholder community and that we adequately represent them.

This section provides a profile of the stakeholders and users involved in the project, and the key problems that they perceive to be addressed by the proposed solution.

Stakeholder Summary

Description	Responsibilities
SAS Global User Group ~~Forum~~ Executive Board	Monitors the project's progress Approves funding Confirms business value for this project
SAS Global Forum Executive Board Data Model Working Group	Ensures the data warehouse meets stated and implied requirements Ensures that the system will be maintainable and manageable by the allocated resources Manages all code related to data integration and business intelligence Oversees the provisioning of access to ensure that data security and privacy policies are maintained
SAS Global Forum Current and Prospective Attendees	Decisions made using the data housed in this system will continue to reflect what people really want The conference will provide a revnue and format (and potentially other services) that people choose to attend and return to Companies that send people to our conference feel good about the value and continue to fund their attendance
Sponsors (and vendors and Alliance members)	The future state of the conference is something that sponsors are willing (and excited) to spend their resources on Business value for the sponsors is high
SAS	SAS continues to invest in SAS Global Forum as the premier SAS event Internal groups at SAS recommend SAS Global Forum to their customers, spend energy making it better through content and representation

USER SUMMARY

Description	Responsibilities
SAS Global Forum Executive Board Marketing and Advertising Committee	Primary user of the system Representatives from the SAS Global User Group and SAS Develop models for retention marketing Manage sponsor recruitment Generates standard and ad-hoc reports

Privacy and Confidentiality

One of the critical business requirements that we had early on was that we needed to make sure that we protected the data and all of the information that we might potentially house in this data warehouse. It wasn't just a matter of technology but also policy and governance. As a volunteer organization, we knew that this was even more important given the fact that we all did not work for the same organization and our employment governed by a confidentiality agreement and/or a non-disclosure agreement that restrict sharing of private information.

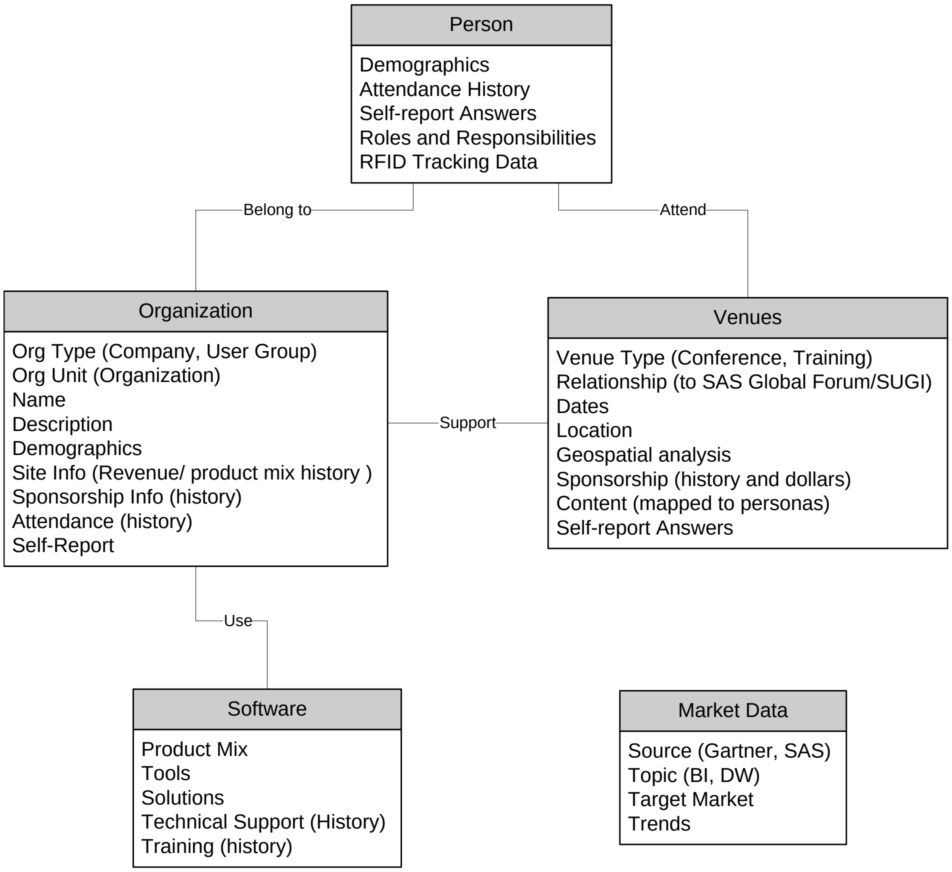
To that end, we had to develop governance processes that ensured that only those that really required de-identified information should have that level of access, but that we also wanted to find a balance. We have not yet found the penultimate solution but suffice it to say, we are proceeding in an extremely cautious manner to avoid having to undo anything later. Only a few people have access to detailed data and when we do distribute data (for example, for a crowd sourcing project described later), it is de-identified so that only a surrogate key is used instead of name, company, address, etc.

Conceptual Data Model

At the outset of any data warehouse project, it is important to define, at a conceptual level, the types of data elements that seem to be useful. This section provides a high level view of the data that we hope to capture in this data warehouse. The conceptual data model is the first step in defining the structure of how data will be housed in the system. It describes the data in terms of "subjects" or topics that we think are important. Later in the process, specifications for how the data will be stored will be developed and documented in the logical and physical data models.

Subject Areas Perspective

The following diagram outlines the high level subjects that we believe would be critical to a robust data model.



These high-level subject areas are described in more detail here including the value of having this data.

Subject	Description
Person	Data from registrations at SAS Global Forum and potentially the regional user groups (such as NESUG, SESUG and WUSS) as well as SAS events and conferences like Data Mining/ Analytics, executive events, SAS days, as well as training classes and books purchased would be captured. The primary unit of analysis would be the person with tables linking past attendance and employment history (companies they listed). Participation at various events could be captured in surveys, RFID tracking information and involvement as a speaker or volunteer. Knowing how attendees participated in a conference helps us model their potential future attendance and motivations.

Organization	Organizations would contain information on sponsor companies, attendee employers and conference organizers (e.g., regional conferences) Organizational level analysis could detect patterns of attendance by various departments and trends in increasing/ decreasing use of SAS. Those companies that also sponsored an event could be tracked here as well. Surveys could be directed at management to understand what motivates companies and their management teams to send and/or sponsor events.
Venue	Primarily held as reference data for comparison of attendance figures in various locations, the venues would provide critical data for geospatial analysis and patterns of attendance between conferences. Additional effort could be spent on conducting text analysis on conference proceedings that would characterize the content domain of the conference.
Software	By looking at what software various companies use (and tech support issues they have had), we could hone-in on the topics people might be interested in and predict future areas of interest.
Market Data	External data that provides third party validation of software usage data along with patterns of general conference attendance. This data would give us baseline information for trends.

Stakeholder Questions

Below we have outlined the questions that could potentially help us drive SAS Global Forum strategy. We did not outline how the data would be captured (survey versus purchased data versus an existing data source) nor did we rank the importance of each question. These are listed here primarily to define the scope of what we are interested in.

Stakeholder	Questions
All stakeholders	Does location matter? How does location affect you? Does time of year affect the decision to attend? How does time of year affect attendance? Does the price of the conference (and sponsorship cost) affect you? What content would keep you coming back (or to attend in the first place)?

	How important is networking (and the ability to connect with employers and agencies)? What is the impact of content (and content changes) to attendance and sponsorships? How do you feel about a frequent attendee id (potentially tied to SAS Certifications) as single source of involvement? How do you feel about RFID for tracking movement at a conference (that is, which sessions did people attend)? Giving the evolving role of SAS in organizations and whom the software impacts, how should SAS Global Forum evolve? How much overlap is there in the various SAS conferences? In other words, How many conferences do individuals attend? Other than the conference, what would be of value that we (SAS Global Forum?) could provide as a service to the SAS community? (e.g., web site, newsletter, etc.)
Attendees	How many SAS Global Forums/ SUGIs have you attended? Which conferences? Why did you come (and not come) in certain years? Have you ever volunteered? How often (which SAS Global Forums)? What role? How did you get involved? Why? What other regional SAS user groups have you attended/ and what roles did you play? Are there competing forces for your not attending SAS Global Forum (regional user groups, training, budget, economic cycles)? What training and/or conferences have you attended (and plan to attend)? What giveaways, perks, things would be of value to you (bags, t-shirts, connect with SAS employees)? What resources would you need to help justify your attendance? What tools could you use? How has the demographic of the attendee changes over the years? What impact does that have on how the conference and other services should evolve?
Regional SAS User Groups, Special Interest Groups, and	What conferences do people attend (both leadership and perception of their attendees)?

SAS-sponsored Venues	Are there opportunities to collaborate around how we engage sponsors? What history do you have on your attendees (and would you share if there was a reciprocal sharing program for de-identified data)? What marketing programs would benefit your organization? What would you like to see SAS Global Forum do differently?
Prospective attendees (SAS customers & sites)	Do you know about SAS Global Forum? Do you or have you a training program for SAS users? Should free registrations be included with licenses? Why don't you currently attend (don't know about it)? Are there competing forces for not attending SAS Global Forum (regional user groups, training, budget, economic cycles)? What would be of value for you to send people to SAS Global Forum?
Prospective attendees (non-returning)	Why did you not come back (i.e., job change, company support, specific complaint)? What could we do to encourage you to come back? Are there competing forces for not attending SAS Global Forum (regional user groups, training)? Would SAS conference package deals allow you to attend?
Prospective attendees (Future SAS Sites)	How SAS is sold (solutions) will affect what they perceive their need to be in training/ conferences. What are they going to want? How do we make it more attractive for solutions owners/ non-traditional attendees to frequent SAS Global Forum?
Sponsors	What motivates sponsors? What would be of value (that is, what would motivate them to continue their support)? What competes with SAS Global Forum for marketing dollars? Would giving sponsors' package deals (combing different types of exposure) for multiple SAS conferences increase sponsorships? What types of linkages would be an incentive (should there be cross promotion)? What are the limitations of what SAS Global Forum currently does that needs to be removed to make SAS Global Forum a more attractive venue? (e.g., getting names/ email; competing entities)

	What would the sponsor like to see SAS Global Forum do (differently?) or additionally? What could SAS Global Forum do to attract non-traditional sponsors? What would SAS like to see SAS Global Forum do? (Marketing, Sales, Alliances, Pubs, etc.) Should recruiting be allowed? (SAS and other stakeholders) Do other conferences allow recruiting (and how?)
SAS Employees	What would SAS like to see SAS Global Forum do? (Marketing, Sales, Alliances, Publications, etc.) What would have to happen to SAS Global Forum to get you to recommend SAS Global Forum to attend / sponsor the conference?

Given the types of questions that we have and breadth of depth of information that would be useful, it was also important to realize that this project would have to evolve over time so that lessons learned early on in the life cycle of the project could be applied to enhance the capabilities. Furthermore, we understand that this system must operate in an environment that interacts with both structured and unstructured data. That is, we intend to obtain data from operational systems such as the SAS Global Forum registration system as well as informal source systems like spreadsheets and surveys, focus groups and external data.

Architecture

Up to this point, we've spent a lot of time thinking about the strategy – or what we hope to get out of this data. Let's now turn our attention to ―how do we do it - that is, the methods and technical bits used to support these tasks.

As we have outlined elsewhere (Nelson, 1999), there are a variety of tasks that are required in order to build a successful warehouse. Now we're ready to start building our physical data model. The tasks now at hand are:

- Extraction
- Data Validation
- Scrubbing (cleansing)
- Integration
- Structuring
- Denormalizing

- Summarizing
- Create fact and dimension tables
- Optimize our indexes and queries
- Create views
- and finally, develop an exploitation methodology that takes advantage of the technology of multidimensional databases, dashboards, reports and other business intelligence capabilities.

In this paper, we wanted to provide a sampling of some of the challenges that we faced and how these were overcome. As with most projects, it has little to do with the technology, but rather how we can capture the right data in a way that is easy to analyze. As this project is entirely staffed through volunteers, we don't have a warehouse full of staff (no pun intended!) We'll speak more on the implementation strategy later, but suffice it to say, our priorities are to establish a baseline data model to answer 80% of the questions and then evolve that over time to fill in the blanks.

Technology choices

Since we are all volunteers from the SAS community, it was an obvious choice for us to utilize tools and technologies available through SAS. However, we did not have to sacrifice in capabilities – because SAS has true end-to-end solutions for data quality, data integration, storage, analytics, reporting and querying as well as information delivery.

In the paragraphs below, we highlight just a few of the choices that we made in selecting the technology options for our warehouse and some of the more interesting challenges.

DATA CLEANSING

Like most projects we started with what we had. In terms of data assets, we had data from those who registered for the conference since 1989 along with some information on speakers, papers, sponsors and post-conference surveys. The focus for this first phase was to get the fundamental unit of analysis right. That is, we wanted the person – name, company, roles – absolutely right before building on to the data warehouse.

Also, it was important to note that prior to 1989, the data was not easily accessible and the value for understanding current trends was clearly diminished. So while we started with attendance data, our plans do include the ability to augment this data with conference survey results, organizational overlay data and sponsor history. Our priority was getting the basic data around "who" attended the conference and enriching that as much as possible to get an accurate picture of what happened at each conference.

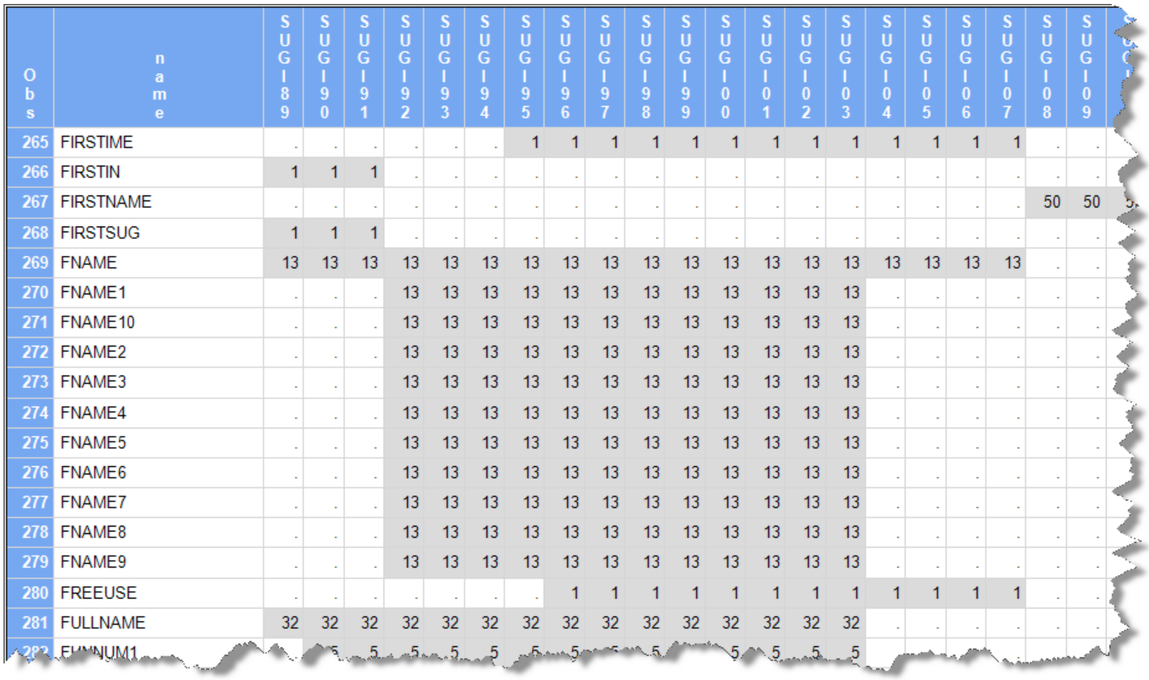
Data Consistency

Initially, we had to evaluate just how good the data was. We knew that for each year, we had names, addresses, and company information. But we also had some demographic information such as title and whether they attended any pre/post conference workshops. Within each conference the data was fairly clean. While we had lots of missing data for attributes like title, we had a fairly consistent picture. However, between conferences, we had little consistency. The registration form changed almost annually making it difficult for us to map titles and companies from year to year. In fact, the same can be said of a person's name. We were at the mercy of the attendee to enter their name, organization, and contact information consistently.

As you might guess, we had a lot of work to do. So our first pass, was to identify which columns were consistently used from year to year. To that end, we developed a routine, which looked at the variables in common across over two decades of registrations. Most of this required manual updates to variable names to achieve consistency. From there, we could see which variables were available in which years. Interestingly, we didn't start collecting email address until 1994. In some cases, we collected data that was no longer needed and there were points in history when a variable was collected only for one or two years.

Additionally, we mapped which variable lengths changed over time so that when we did start to combine data from multiple years, we could do so with confidence.

In the table below, we demonstrate the variable mapping with the variables going down the side, the conference year across the top and the variable length at the intersection.



Obs	name	SUGI89	SUGI90	SUGI91	SUGI92	SUGI93	SUGI94	SUGI95	SUGI96	SUGI97	SUGI98	SUGI99	SUGI00	SUGI01	SUGI02	SUGI03	SUGI04	SUGI05	SUGI06	SUGI07	SUGI08	SUGI09
265	FIRSTIME	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.
266	FIRSTIN	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
267	FIRSTNAME	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	50	50
268	FIRSTSUG	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
269	FNAME	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	13	.	.
270	FNAME1	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
271	FNAME10	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
272	FNAME2	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
273	FNAME3	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
274	FNAME4	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
275	FNAME5	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
276	FNAME6	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
277	FNAME7	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
278	FNAME8	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
279	FNAME9	.	.	.	13	13	13	13	13	13	13	13	13	13	13	13	.	.	.	.	.	.
280	FREEUSE	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	.	.
281	FULLNAME	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	.	.	.	.	.	.
282	FUNNUM1		5	5	5	5	5	5	5	5	5	5	5	5	5	5						

Table 1. Variable Mapping by Year

Ultimately, we decided on a short list of variables to use to merge the data sets together forming a baseline of 21 years of attendance information. In future iterations we will look to prioritize the incremental value of adding fields from some of the past conferences to get at the attendees role, attendance at workshops and purchase history.

Data Profiling

Once we had a consistent set of variables across time, our next task was to evaluate what each of the fields really held and how much cleansing needed to be done. Data profiling involves using analytical techniques to discover the true content, structure and quality of data. Any data management initiative begins with profiling, the current state of the data is analyzed – and a plan to improve the information is begun. Data profiling is a fundamental, yet often overlooked, step that should begin every data-driven initiative.

Each of the columns we pulled in needed some form of cleaning. The interesting exception was how to standardize without losing critical information. As we could assess whether "Census Bureau", "US Census Bureau", "US Department of Census" and "US Census" needed to be rationalized, it was not in our best interest to change someone's name or company!

By using the SAS product: DataFlux, we were able to quickly profile each field and assess how much work we had in front of us.

In the example below, we highlight one of the outputs of the data-profiling step.

Field Name	Company
Ordinal Position	3
Count	59641
Null Count	1429
Blank Count	0
Minimum Value	Maritz Fire
Maximum Value	
Mode	US Bureau of the Census
Pattern Count	7854
Unique Count	14261
Uniqueness	24.5
Primary Key Candidate	no
Data Type	CHARACTER
Data Length	200 chars
Actual Type	string
Minimum Length	1
Maximum Length	66
Non-null Count	58212
Nullable	YES
Decimal Places	0
Percent Null	2.4

Table 2. Data Profile Report from DataFlux

Data Correction and Standardization

Data correction is the logical next step after the data profiling stage. This generally consists of routines and procedures that apply to specific data hygiene issues on the data. Typically, the operations that fall into this category physically change or transform the existing data and write it back to the original field or to a new field in a data set. Data correction typically includes processes such as parsing, standardization, verification, and general data sanitization procedures. Often these procedures are the precursors to more extensive data manipulations.

Furthermore, data elements are often ambiguously represented in a data set. We have data that was collected uniquely for 20+ years – and the variables collected change (for example both the variable name and content? Both are issues, the latter compounds the issue). There are many ways to address this situation and they can all be generally called standardization routines. Standardization could mean physically standardizing the data within the data set, it could mean creating synonym tables or filters, and it could mean correcting undesired permutations before they enter the dataset in the first place. Critical to data standardization is that rules for these permutation transformations are maintained external to an application or data set and can be applied by any of the SAS components that might be deployed across the enterprise.

Above, we show the output from the field: Company. As you can see, there over 58,000 companies listed. However, we know that the actual number of unique companies is hidden in the data. So this is where standardization comes into play.

Standardization is where we say that “Census Bureau” and “US Dept of Census” are the same. In BASE SAS, we would code this like:

```
If company_name = “Census Burea” then standard_company_name =
“United States Department of Census”

else if...
```

However, with DataFlux, we don’t have to code those details, instead we just tell the software to make its best guess as to the mappings, and then we point and click to modify or create new mappings. We can then use this “standardization” in an ETL job or create a one-time standardized output.

So we turned our team loose on creating a standard for Company, Name (even though I said we were not changing names, we create a person ID that linked people – regardless of how they spelled their name), address, country, etc.

Standardization was a critically important component of this process because we needed to be able to uniquely identify people and organizations (see conceptual data model above).

Data Enrichment

Frequently, records in a data table (or tables) are incomplete. This missing data may prevent you from adequately recognizing the true value of the data, or it may be difficult to tie these types of records to other information that may already exist in a system. There are really two aspects to these problems.

The first issue is *data completeness*. A typical example of this is records with a missing ZIP code. If we cannot derive the zip code from the other fields, this might preclude any mailing effort or ZIP code-based geographic mapping.

Second is the issue of *data keying*. While we may have a complete view of the data for certain needs, but there may also be a need for geographical data (i.e. longitude and latitude) for additional business objectives.

- Data Completion - Data completion often takes the form of data consolidation. There may be bits and pieces of data about customers in many different data sets, or you may find that you are missing one important piece of data, (i.e. phone number), for many of your records. For certain pieces of missing data, you may decide to purchase 3rd party data. This may require the generation of match codes on names and addresses between the two data sets to adequately match up the data. The end result is a more complete picture of your customer.

- Data Keying - There may be instances where you need to augment existing data in a record to be able to access additional (more extensive) data. Often the first step is only taken to link the two data sources, and the data from step one may not provide any further benefit. This step is called data keying.

In our case, we wanted to employ both approaches to filling in the blanks. While in this first phase of the data warehouse development, we did not choose to employ techniques to enrich our data (walk before you run!) We do have a number of ideas for how the data could be augmented to include industry codes, gender for attendees, company profile (number of employees, industry / SIC codes), etc.

One thing that we did do, however, was to geocode the primary address of the attendee and/or company. This was important because we wanted to be able to answer questions such as “does conference location matter?” and “how far do attendees tend to travel?, etc.

Geocoding in SAS 9.2 was easy as it is built into the software. We simply provided PROC GEOCODDE the contact fields (address, city, state, zip, country, etc.) and the procedure returns the geocoded values (latitude, longitude, etc.).

```
proc geocode ZIP                   /* Geocoding method        */
     data=WORK.events              /* Address data            */
     out=geocoded_events           /* Output data set         */
         addresszipvar=zip
         addresscityvar=eventcity
         addressstatevar=eventstate
         attributevar=(city, state, STATECODE, STATENAME, COUNTY,
COUNTYNM, MSA, ALIAS_CITY, TIMEZONE); /* Include these variables */
run;
```

DataFlux also has the capability if more complex coding is required. For our case, we found the PROC GEOCODE approach worked just fine.

ETL (Extract, Transform and Load)

As you can see from above, our phase one data structure is fairly simplistic. We have people, companies and attributes of each for those attending the annual conference since 1989. Our first step in the overall program flow was to merge these datasets and do all of the renaming that had to be done as well as rationalizing the variable attributes (such as length, formats, labels, etc.) Next we applied the standardization and enrichment from above and then loaded into a de-normalized structure (one record per person per conference year).

As the data warehouse matures, we'll de-normalize this structure into a star schema for exploitation. Before we get there, there were some additional components that needed to be considered before loading into a final data structure: matching and duplicate records.

- Matching and Duplicate Identification - Data matching is at the core of all the functionality that SAS employs to solve data quality and data integration issues. Typically, organizations will use SAS matching technology to find potentially duplicate records. This, however, proves to be no easy task. Not only can there be several permutations of a given data element, but certain data elements are often missing or are in a different order from record to record. Sometimes data values that should appear as matches are not even remotely similar without a business rule that links them together. A great example of this is nicknames for individuals. If you attempt to match Richard Smith and R. S. David Smithtogether, no string similarity routine will catch the potential match unless you can apply data specific rules to the matching process. In this case, SAS employs a nickname knowledge base that can be used as a look-up table to find non-obvious matches.

- Duplicate Elimination and Consolidation - Matching records or data in this fashion is great for reporting or auditing but the same process can be used to do rather powerful duplicate elimination and consolidation as well. The Match Clusters that are created by DataFlux can be used to eliminate potential duplicate records. This can be performed manually or automatically by applying rules to physically delete all duplicate records (perhaps: saving only a single unique occurrence) or logically delete duplicate records (by appending a delete flag to duplicate records referring back to the surviving record).

- Linking - Match Codes are very powerful tools that can perform a number of manipulations to the data. Two extensions of Match Code technology are the ability to virtually link data across disparate data sets and the ability to household records which means grouping individuals into clusters based on some common relationship—usually this refers to a individuals with a family relationship sharing the same address, but it can be applied to business as well.

Once the matching, de-duplication and linking has been done, we then take our matched records and load our analysis tables.

DATA EXPLOITATION

As it stands today, the data is primarily used to create analysis ready datasets used by our community of volunteers to find interesting patterns. In fact, as we will learn in the next section, we tasked the community to see what interesting patterns of attendance they could find.

Community Collaboration: Crowd-sourced Analytics

As is true of most volunteer organizations, we have a long to-do list and precious few resources. In 2010, we developed a model to truly engage the user community and to utilize the collective knowledge, skills and abilities of entire eco-system to help analyze this data – in support of making the conference better for everyone.

But like any project worth doing, we faced our challenges:

- It was going to take more resources than we had
- Implementation would be slow as we are reliant on a volunteer community to implement
- Quality control / validation NOT maximized as we had constraints around how we implemented code sharing, reuse and version control
- Ideation and execution limited by resource availability (small team)
- Limited access to SAS server (and complex governance process)
- Ensuring that access to de-identified data was limited to only those that were actively contributing to the body of knowledge and contributing to the core code-base in SAS, DataFlux or JMP.

Furthermore, we did not have the budget to go out and hire someone, so we decided to bribe resources through fame and glory – we'd “crowd source” it!

Crowd sourcing is a simple concept - we can find solutions to some of our most difficult questions by using the collective power of a community. Think of this as a grid-enabled, cloud based analytic engine. The engine for innovation is the community!

We can set goals, brainstorm ideas, develop theories and test them using tools that each of us are uniquely qualified to operate. Some people are data junkies - beating data into submission until it squeals the answer. Others are masters of visualization - creating maps, charts and info-graphics that tell a story. There are those that are good at reminding us that we have a problem to solve and focus on our attention on the why of data analysis.

Through the concept of crowd-sourcing and a vibrant community of interested and willing participants, we can make sense out of the world around us.

Our first experiment is about telling the story of the semi-colon people - SAS users. SAS users have helped organize and run an annual conference for over 35 years. During this time we've tried a number of things that has helped to form one of the largest and most successful user run conferences in the world.

As a community of over 30,000 companies world wide, we wanted to put the collective hearts and minds of those users together to help morph this conference into the modern equivalent of SUGI 2.0

This is where individuals or small teams follow a prescribed path to find something useful in the data that we have collected about our users and our conference. Here, we prioritize on the most pressing issues that face our community, conduct root cause analysis, collect data (if needed), perform analysis, share our findings and recommend change.

Benefits

The benefits of this approach were compelling enough to give this a try. It gave us to the opportunity to engage the community we serve, it would help to increase awareness of the "user" aspect of SAS Global Users Group, we could certainly accomplish more than we could alone and it demonstrate the vast capabilities of SAS in the real world (for and with our users).

How it Works

In 6-8 week mini-project formats, the intention is to recruit volunteers who have an interest in solving a specific set of business problems as posed by the Data Model Working Group. From there, the team goes through a fairly standard set of steps aimed at answering the business questions posed:

1. Confirm understanding of business problem
2. Define data requirements
3. Perform impact and feasibility analysis
4. Document schedule and dependencies
5. Design, build and test solution
6. Present findings

At the end of each iteration, teams have an opportunity to showcase their findings and report back to the SAS Global Users Group Marketing & Advertising committee (and other interested parties within the Executive Board).

Annually the most significant and interesting findings will be shared at the annual conference (2012+ - conference section or interactive poster) as well as being recognized through formal citations of work products.

Summary

This paper describes the process of building a data warehouse capable of answering questions about SAS Global Forum and its attendees. Specifically, our questions focused on

- Patterns of attendance
- First timer trends
- Geospatial mapping
- Job roles
- Conference content
- Section attendance
- Improved marketing
- Rationale for attendance
- Costs/ motivators
- Non-attendance
- Sponsorships
- Brand

Our challenge was that we had a very small, volunteer based workforce to get us there. We had a number of constraints that included:

- Resource availability
- Right skill at the right time
- Access to server resources
- Inventiveness of a small community
- Protect an invaluable asset - our data

Through the inventiveness of the Data Model Working Group, SAS Global Users Group Executive Board and a vibrant community of interested and active volunteers, we will continue to evolve this resource in hopes of making SAS Global Users Group and the entire user group ecosystem better for the efforts of those that made this possible.

References and Author Contact Information

References and Recommended Reading

Kurt Bittner and Ian Spence (2002). Use Case Modeling. Addison Wesley Professional, 2-3.

Grasse, D. and Nelson, G. *"Base SAS vs. Data Integration Studio: Understanding ETL and the SAS tools used to support it".* Invited Paper presented at the SAS Users Group International Conference. San Francisco, CA. March, 2006.

Kimball, Ralph. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Wiley & Sons, 1996

Kimball, Ralph. *"The 38 Subsystems of ETL: To create a successful data warehouse, rely on best practices, not intuition."* Intelligent Enterprise." December 4, 2004.

Kimball, Ralph and Conserta, Joe. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004

Kimball, Ralph, Laura Reeves, Margy Ross, and Warren Thornthwaite. The Data Warehouse Lifecycle Toolkit: Tools and Techniques for Designing, Developing, and Deploying Data Warehouses John Wiley & Sons, 1998

Kimball, Ralph and Ross, Margy. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). John Wiley & Sons, 2002

Nelson, G. *"A Pragmatic Programmers Introduction to Data Integration Studio: Hands on Workshop".* Hands on Workshop presented at the SAS Users Group International Conference. San Francisco, CA. March, 2006.

Nelson, Gregory S. "*Implementing a Dimensional Data Warehouse with the SAS System*." Invited Paper presented at the SAS Users Group International Conference. San Diego, CA. March, 1999.

Acknowledgements

I would like to thank Kim Lebouton, Art Carpenter, Duke Owen, Neil Howard, Lauren Haworth and Debbie Buck for their "thot-ful" and insightful comments on this manuscript and for their continued support of this program.

Biography

Greg Nelson, President and CEO of ThotWave Technologies, LLC.

Greg is a certified practitioner with over two decades of broad Business Intelligence and Analytics experience. This has been gained across several life sciences and global organizations as well as government and academic settings. He has extensive software development life cycle experience and knowledge of informatics and regulatory requirements and has been responsible for the delivery of numerous projects in private and commercial environments. Greg's passion begins and ends with helping organizations create *thinking data®* – data which is more predictive, more accessible, more useable and more coherent.

His current area of interest is helping companies take advantage of the shifting world of convergence around data and systems and how modernization and interoperability will change the way that we discover new relationships, manage change and use data and analytics to improve organizational outcomes.

Mr. Nelson has published and presented over a 150 professional papers in the United States and Europe. Mr. Nelson holds a B.A. in Psychology and PhD level work in Social Psychology and Quantitative Methods and certifications in project management, Six Sigma, balanced scorecard and healthcare IT.

Greg can be reached at greg@thotwave.com or www.linkedin.com/in/thotwave

About ThotWave

ThotWave Technologies, LLC is a Cary, NC-based consultancy and a market leader in real-time decision support, specializing in regulated industries, such as life sciences, energy and financial services. ThotWave works at the juncture of business and technology to help companies improve their operational and strategic performance and recognizes the difference between simply accessing data and making data work for business. Through products, partnerships and services, ThotWave enables businesses to leverage data for faster, more intelligent decision making.

Contact information:

Your comments and questions are valued and encouraged. Contact the authors at:

Greg Nelson greg@thotwave.com

ThotWave Technologies, LLC

Chapel Hill, NC 27517 (800) 584 2819

http://www.thotwave.com

thinking data® is registered trademark of ThotWave Technologies, LLC.

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