Delving into the Data Warehouse using SAS: Levels of Analysis for a Three-Tier Survey Analysis System.

David J. Ross
True North Technologies
P.O. Box 1075 Morton Grove, IL 60053
Phone: 708.901.3083
FAX: 708.390.6774

Abstract
The implementation of a three-tier, enterprise wide data warehouse solution using SAS requires a clear framework of analysis for deployment and utilization of information. This is accomplished by defining a metadata repository (information about the data), procedures (to access and organize information), and levels of analysis (for storage considerations, and accessibility for novice and knowledgeable users). A case study using survey analysis is presented for a warehouse deployment in a client/server environment, demonstrating SAS file structures, appropriate SAS procedures, and methods. Finally, suggestions are made for Wrapping the components of the system within clear business objectives.

1. Introduction
The path of data warehouse development often goes from conceptualization to concrete implementation at full impulse power. This is due to the ever pressing needs of companies and institutions to understand and organize what they have and simplify the means and tools for analysis. The three major thrusts of data warehouses are:

Transforming data from disparate and sometimes contradictory sources into a consistent form.

Distributing information in a form which accurately reflects the business process and not the complexity of storage or retrieval method.

Opening up the information in a timely and concise way for access to novice and knowledgeable users.

Implementing a data warehouse in SAS requires a clear framework of analysis for deployment and ways to avoid the pitfalls associated with an enterprise activity. Furthermore, the appropriate containment in a client/server deployment requires an understanding of the applicability of each level in a three-tier enterprise system.

This paper focuses on three necessary elements. First, a metadata repository is established which defines information about the data: the source of information, transformations and computations, attributes, and formats. This results in a common view of information which is supported outside of the actual information storage methods.

Next, Procedures are defined as the windows of information which
bridge each of the Levels of Analysis. These views are composite pictures of information written as a DATA step or PROC SQL.

Finally, there is a concise definition of levels of analysis which there are four: Distinct, Event, Aggregate, and Enterprise. These elements define both the form of the information retained, the targeted use of the information, and the location of the information across a client/server environment.

Each level of analysis is composed of metadata definitions and procedures which permit inquiry within a level and across all levels. In this way, all of the distinct parts of analysis can be deployed and accessed across a three-tier enterprise system.

2. Definition of Metadata

SAS presents a wealth of information in each file structure. Variable Name, Label, Format, Source, Date, Engine, and so on are available with a PROC CONTENTS. By simply performing a PROC CONTENTS OUT=, a separate file can be generated and enhanced to document the metadata repository needed for managing the information. At each of the levels of analysis, a file is generated and joined with the central metadata file. Thus, consistency of information and common elements can be viewed, extracted, and reported upon.

The metadata repository can also contain other elements for managing SAS data sets. These are FORMAT libraries and SAS CATALOG entries (such as programs, views, and MACRO statements) which can be deployed across a client/server environment. Unfortunately, this does involve extensive manual tracking and revisions but can form a concise framework for executing queries and computations at each level of analysis.

3. Procedures

Procedures are segments of SAS code (DATA Steps, MACRO statements, or SQL) which are used to compose information at different levels of analysis. The goal is to aggregate or summarize information so that an knowledge of the underlying complexities of the database is not required. Depending on their general applicability, these may reside as flat files in a directory (client), elements of a SAS catalog (client, server, or mainframe), or as managed elements of a code management repository.

4. Level of Analysis: Distinct

The first task is to establish a warehouse backbone of information, which is consistent throughout. At this level, no aggregation or summarization has occurred. Data collected from discordant sources has been transformed into common, concise elements stored for efficient means of extraction. Consequently, it is a singular instance of information which is used to populate the data warehouse.

In survey analysis, questionnaires typically consist of a series of uniquely identified questions (Gender) and a single response to
that question (Male). Questionnaires may also contain uniquely identified questions (What do you do for fun) which contain multiple responses (camp, play golf, jog, cook).

Storage of information occurs in a column row sequence where each row is represented by a respondent and each column by a unique question. When a variable number of responses may be given by a respondent, a column would need to be reserved for the maximum number of responses ever given by any respondent. Consequently, for most respondents who give less than the maximum, considerable space needs to be set aside.

For a data warehouse, information is normalized to its component elements: that is, for every response which occurs for every question, for every respondent, a row is created. The existence of the response rather than the maximum possible responses are recorded.

File Structure in SAS
A single row in the table for the 23rd respondent to questionnaire 43, answered "Male" to question 57 is:

SRVY_ID QUES_ID RSPN_ID RSPN_CD
43 57 23 Male

Multiple rows in the table for the 23rd respondent answered "camp, jog, cook" to question 45 is:

SRVY_ID QUES_ID RSPN_ID RSPN_CD
43 45 23 camp
43 45 23 jog
43 45 23 cook

Where
SRVY_ID=Unique Questionnaire Identification
QUES_ID=Question Number
RSPN_ID=Respondent Number
RSPN_CD=Response

SAS Commands and Procedures
- LABEL or ATTRIB for column structure and description
- PROC FREQ for distribution or statistical differences
- PROC UNIVARIATE for distribution

5. Level of Analysis: Event

The next task is to define a specific framework of distinct occurrences, which has both meaning and reproducability outside of the repository. This level may have been established by imposing structure on the Distinct Level of Analysis or can exist as the common structure of information and become a distinct data warehouse.

The event level of analysis is the more traditional way of viewing questionnaire information. It is all the responses from a single respondent's questionnaire.

File Structure
Information is stored at the case level. A single row in the table for the 23rd respondent to questionnaire 43, answered (3) Disagree to question 57, (1) Under 25 to Question 009, and (2) Female to Question 007.

The 25th respondent to questionnaire 43, answered (3) Disagree to question 57, (5)55 or Older to Question 009, and (1)Male to Question 007.

SRVY_ID RSPN_ID Q57 Q009 Q007
43 23 3 1 2
43 25 3 5 1

Where
SRVY_ID=Unique Questionnaire ID_QUES_ID=Question Number RSPN_ID=Respondent Number Q57=Health Care Concerns Q009=Age Q007=Sex

**SAS Commands and Procedures**

PROC TRANSPOSE to convert Distinct Level to Event Level PROC FREQ for distribution or statistical differences. PROC UNIVARIATE for distribution. PROC TABULATE for cross tabulation of results.

6. **Level of Analysis: Aggregate**

Based upon the Event level, we can encapsulate information into a self contained object for clear comprehension by the knowledgeable user. As such, it usually resides in the departmental warehouse or accessed through procedures.

Aggregation uses the tools of summarization for clear reporting and reducing storage requirements.

**File Structure**

Information is summarized up to the question - response for each questionnaire. For example, rows in the table for question 57 collected across all respondents for questionnaire 43 are:

<table>
<thead>
<tr>
<th>SRVY_ID</th>
<th>QUES_ID</th>
<th>RSPN_CD</th>
<th>P%</th>
<th>COUNT</th>
<th>QN</th>
<th>QMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>5.2</td>
<td>26</td>
<td>497</td>
<td>3.18</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>12.0</td>
<td>60</td>
<td>497</td>
<td>3.18</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>41.2</td>
<td>205</td>
<td>497</td>
<td>3.18</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>41.4</td>
<td>206</td>
<td>497</td>
<td>3.18</td>
</tr>
</tbody>
</table>

Where

P% = Percent answering based on all answering a Question
COUNT = Number of Respondents in category

QN = Denominator for calculation of PERCENT
QMEAN = Arithmetic Mean based upon QN

**SAS Commands and Procedures**

PROC FREQ for with the OUT option.
PROC SUMMARY with the OUT option.
PROC SQL with SUM(…) as NEWVAR1, COUNT(…) as NEWVAR2

7. **Level of Analysis: Enterprise**

Finally, we can deploy a view that transcends aggregates of information, to form a complete picture for the novice or knowledgeable user. The boundaries of such will not be restricted to a single area of expertise but will often cross defined structures of the enterprise to understand business processes or dynamics.

Information is summarized to a cell across a group of questionnaires. The information is generated as:

**File Structure**

<table>
<thead>
<tr>
<th>SRVY_ID</th>
<th>QUES_ID</th>
<th>RSPN_CD</th>
<th>P%</th>
<th>QN</th>
<th>QMEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>5.2</td>
<td>26</td>
<td>497</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>12.0</td>
<td>60</td>
<td>497</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>41.2</td>
<td>205</td>
<td>497</td>
</tr>
<tr>
<td>43</td>
<td>Q57</td>
<td></td>
<td>41.4</td>
<td>206</td>
<td>497</td>
</tr>
</tbody>
</table>

I understand our medical coverage

<table>
<thead>
<tr>
<th>RESPNLBL</th>
<th>RESP_NO</th>
<th>SURV_NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Response</td>
<td>812</td>
<td>38</td>
</tr>
<tr>
<td>Hardly at all</td>
<td>1231</td>
<td>38</td>
</tr>
<tr>
<td>2</td>
<td>2509</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>6576</td>
<td>38</td>
</tr>
<tr>
<td>All the time</td>
<td>6746</td>
<td>38</td>
</tr>
</tbody>
</table>

Where

PR% = Avg % answering a response, of all answering a Question, based on SRVY_NO
SMEAN = Arithmetic Mean based upon SRVY_NO
LABEL40=40 character question label
RSPNLBL=Response label (20 characters)
RSPN_NO=Total number of respondents in a category
SRVY_NO=Total number of questionnaires

SAS Commands and Procedures
PROC SUMMARY with the OUT option.
PROC SQL with SUM(NEWVAR1) as ENTVAR1, COUNT(NEWVAR2) as ENTVAR2

8. Warehouse Deployment in a Client/Server Environment

Given the framework that has been developed, we can now show how each level of analysis fits into a three-tier enterprise wide system.

Distinct: metadata repository defined by SAS data and index files or a relational database product (Oracle, DB2, Informix) connected to SAS through SASS/ACCESS. Resides on a mainframe or common server.

Event: metadata repository defined similarly to Distinct, or as a transposition of Distinct. Resides on a mainframe or common server.

Aggregate: intermediate or departmental warehouse, typically defined by SAS data and index files and residing on a common or department server.

Enterprise: departmental warehouse or client query, defined by SAS data files and residing on a department server or client workstation.

For example, the Enterprise Level may consist of a departmental warehouse of directly accessible information which resides on a client workstation or departmental server. A query which goes against the data need not require a detailed understanding of the methods used to formulate the information. A user can 'drill down' the path of Levels of Analysis, though the different storage and structural locations, with the perception that what they are viewing resides in the client.

Procedures in this framework often need to 'skip a step': access and transform events directly into the Enterprise Level from the Aggregate, Event, or Distinct Level of Analysis. There may be a common set of procedures as well as custom implementations developed by users. In either case, the full power of the SAS language is available for complex processing and reporting.

9. Wrapping for Business Knowledge Delivery

By managing the environment with the metadata repository and developing clear and useful procedures, a transactional hierarchy between each level is established. By developing an object oriented 'front end' on the client side, a useful and integrated platform for both novice and knowledgeable users can be created. SAS FRAMES, Oracle
Power Objects, PowerBuilder or VisualBasic fit this first tier role.

Powerful middleware solutions exist which use SAS CONNECT (and ODBC drivers) as well as Digital Pathworks, Oracle SQL*NET, and CORBA elements. Finally, the server and mainframe implementation of SAS, Oracle7 (with Parallel Query, Replication, and Scalable SMP support), DB2, or Informix (with Parallel Data, Partitioning) forms a solid top tier.

10. **Conclusion**

As the data warehouse has become a necessity in companies and institutions, SAS is well positioned to meet the demands of a three-tier enterprise wide solution. The ability to implement a single source solution across diverse hardware platforms, coupled with a solid layer of connectivity, makes it the system of choice.