Using SAS® Business Intelligence Server and
SAS® Data Integration Studio with academic research data
Kevin Davidson
Texas Institute for Measurement, Evaluation, and Statistics
University of Houston, Houston, TX

ABSTRACT
This paper will review a data warehousing project at the University of Houston. A scientist in the Psychology department has a large number of studies involving the development of cognitive and academic skills in children. SAS® Business Intelligence Server and SAS® Data Integration Studio features are utilized to standardize ETL procedures across studies with the added benefit of making explicit the cleanup process for a given data source. The current and planned use of SAS® Web Report Studio for reporting and extracting analyzable data will be reviewed. The emphasis of the paper will be upon the use of SAS Business Intelligence Server and SAS Data Integration Studio as part of the warehouse process rather than on the data itself.

INTRODUCTION
The purpose of this paper is primarily to demonstrate the use of SAS Business Intelligence Server and SAS Data Integration Studio as it is being used to integrate data across a number of projects.

SAS Business Intelligence Server is a software product that became available when Version 9 of SAS was released. Three of the basic tenets of a strong Business Intelligence solution are the consolidation of metadata (information about your information, user rights and permissions, etc.), aligning the data to business needs so that critical questions can be addressed quickly, and putting the ability to query the data in the hands of the business analysts thus removing IT folks from fielding endless reporting requests. SAS Business Intelligence Server incorporates these basic tenets through the packaging of a number of software products including SAS Metadata Server, SAS Stored Process Server, BASE SAS, SAS Web Report Studio, SAS Add-in to Microsoft Office, SAS Information Map Studio, as well as several other components. For more specification information on SAS Business Intelligence Server, go to http://www.sas.com/technologies/bi.

SAS Data Integration Studio is the successor to SAS ETL Studio and SAS/Warehouse Administrator. SAS Data Integration Studio is a powerful visual design tool for the construction, execution, and maintenance of data integration projects. SAS Data Integration Studio has an easy-to-use interface, extensive built-in transformations, and powerful productivity capabilities while providing a single point of control (Hunley, Mehler, and Rausch 2008, p. 1). SAS Data Integration Studio provides a framework for extending the SAS tools and code that developers have created over the years while providing an expanding set of pre-packaged components that can cut down on code development significantly.

The Texas Institute for Measurement, Evaluation, and Statistics (TIMES) at the University of Houston is led by Dr. David Francis. For the past decade or so, TIMES has collected voluminous data for a wide range of studies including monitoring the development of reading capabilities in Bilingual children, teacher and environmental effects on reading development, and developmental abilities of children with Spina Bifida. Data has been collected in a number of different formats including paper and pencil with ensuing data entry, scannable forms, third-party scoring with spreadsheets then provided to TIMES, and web entry. The type of data collected includes things such as standardized tests (Intelligence, academic skills, cognitive skills, and behavioral traits), experimental tests, and questionnaires.

Data management for each study at TIMES was handled primarily by project specific TIMES personnel including data managers and project managers. Data analysis has predominantly been done with SAS although other analytic tools are also frequently used. As the number of projects multiplied, so did the number of data sets that needed to be maintained. Given the large number of people involved in the various projects, there were of course resultant discrepancies in how data decisions were handled leading to data incongruencies. Thus in 2004, the decision was made to create a data warehouse in order to create “one version of the truth”. The goals of the warehouse were to improve the ability to quickly analyze data, standardize the cleaning and decision making processes across projects, and document these processes so that the processes are rather independent from those that run them and thus are more easily passed to a new data manager when personnel turnover occurs.

DATA FLOW PROCESS
Simple representations of the data flow process before and after the implementation of the warehouse process at TIMES are presented below. Note that one of the main goals of the warehouse process is simplicity so that it can be understood and learned in a short period of time by a data manager with a modest to average understanding of SAS.
### BEFORE WAREHOUSE PROCESS

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Project Specific Cleanup programs</th>
<th>Project Specific Analyzable Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan Forms</td>
<td>Usually individually maintained .sas programs with some use of macro libraries</td>
<td>Often specific to a given test/form and frequently specific to a year or wave of data collection leading to many data sets</td>
</tr>
<tr>
<td>Spreadsheets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web Entry Forms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsoft Access Tables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### AFTER WAREHOUSE PROCESS

<table>
<thead>
<tr>
<th>Data Sources</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scan Forms</td>
<td>Sas Data Integration Studio jobs</td>
</tr>
<tr>
<td>Spreadsheets</td>
<td></td>
</tr>
<tr>
<td>Web Entry Forms</td>
<td></td>
</tr>
<tr>
<td>Microsoft Access Tables</td>
<td></td>
</tr>
</tbody>
</table>

**Staging Area 1**

- Creation of a standard set of editable data sets specific to a given job:
  - A “REMAP” table contains a column with the original variable names and a RENAME column allows for variable renaming consistent with the Target table structure.
  - A “BAD” data set contains violations to Integrity Constraints (out-of-range values). This data set can be edited to supply the corrected values.
  - A “DELETED” data set contains records that are flagged for deletion due to duplicate key field detection or from manual flags set from previous iterations of the job
  - A composite data set that contains all of the original source data with the REMAP names applied and any previously created corrections. This data set has an audit trail attached so that the before and after values, time of edit, and person making the edit is automatically captured the next time the job is run.

SAS Data Integration Studio jobs are rerun until the BAD data sets are clean and the generated reports indicate that the REMAP data set has properly renamed items to the desired naming conventions.

Last step within SAS Data Integration Studio jobs

**Staging Area 2**

Test/Form/Questionnaire specific data sets

Data Warehouse loading job run as needed

**Dimensional Warehouse**

Fact table with answers to a given test/form/questionnaire item. Dimensions include Date, Test, and Study.

### USING SAS DATA INTEGRATION STUDIO TO STANDARDIZE

Although data sources differed, TIMES often experienced the need to perform the same types of data cleansing over and over again. One of the first orders of business in transforming the cleaning processes was to create custom transformations within SAS Data Integration Studio that could then be used to create visual data flows. Figure 1 shows a simple example of one of these custom transformations called “Job Macros” that sets up parameters that drive the given program. Notice that all of the parameters are neatly encapsulated in this transformation. The “Job Macros” transformation is designed to be used at the
start of any data cleaning job (see Figure 3 below). Should the need to add a new parameter occur, the transformation can be updated and existing jobs redeployed (see the “Stored Processes” section below).

Figure 1: Example Custom Transformation showing the Code Options tab in SAS Data Integration Studio.

Another more complicated custom transformation is shown in Figure 2. The “Apply Audit Trail/Corrections and Do Range Checks” transformation performs a lot of different functions including: looking at the eventual target data sets to determine range restrictions to create reports on out-of-range data and reading staging table data sets to read data corrections that were made following previous runs of a given cleanup job.
A portion of what a job within SAS Data Integration Studio looks like is shown in Figure 3. Note on the left side of the figure that jobs can be neatly categorized into folders. Note on the right side that each square node in the job represents a transformation, either custom or one that ships with the product. Circles represent temporary (WORK) data sets. When the job is submitted within SAS Data Integration Studio, first the code is dynamically created from the current contents of the nodes from the metadata repository. The code is then automatically submitted to a SAS Workspace server session. Note the tabs at the bottom of right side of the figure: The Process Editor tab/window displays the visual job flow; the Source Editor window shows the SAS code that is created from the job flow; the Log window displays the results of running the job; and the Output window displays the results of any procedural output.

One perhaps obvious but critically important component to having the cleanup jobs neatly categorized in folders as seen in Figure 3 is the fact that in large part it documents the process and identifies the location of the most up-to-date programs. This single point of control can be utilized in many different ways. For example, if a critical job is highly customized and unlikely to be used again, it can still be registered as a job in SAS Data Integration Studio. One of the transformations that ships with SAS Data Integration Studio is a “User Written Code” transform that lets you either insert your code within the Source Code tab or point to a physical file.
STORED PROCESSES

A stored process is simply a SAS program whose name, location, and associated passed parameters are stored within metadata. This allows numerous clients to access and run the program. The clients include SAS Enterprise Guide, Web Report Studio, Add-in for Microsoft Office, and the Stored Process Web application. SAS Data Integration Studio can create stored processes that can then be run in any of the aforementioned clients. This allows SAS Data Integration Studio to be utilized as a development tool and many different clients can be enabled to run the resulting programs. SAS Data Integration Studio allows you to easily create a stored process from a given job. Note in Figure 3 in the left sided panel that the lines with the icon indicate jobs whereas lines with the icon indicate stored processes that have been created from jobs.

If your environment is one where programs often need iterative development due to the need for added features, the Stored Process redeployment tool within SAS Data Integration Studio is an extremely useful and time-saving tool. It will essentially rebuild all of the developed programs that utilize a given custom transformation. This of course means that you must be careful when making alterations to be sure that changes will be backwards compatible with your existing programs. Figure 4 shows an example of the “Redeploy Jobs to Stored Processes” prompt.
USING SAS BUSINESS INTELLIGENCE SERVER
At TIMES, being able to utilize the results from SAS Data Integration Studio is highly dependent upon SAS Business Intelligence Server. In particular, there are 3 components that we will highlight here: Metadata Server, Information Maps, and Web Report Studio.

METADATA SERVER AND SAS MANAGEMENT CONSOLE
SAS Management Console is a java application that allows you to manage your relevant metadata. As the view shown in Figure 5 shows, there are a number of built-in ‘Managers’. Note under BI Manager that there is a folder called “BIP Tree\ReportStudio” that has folders for “Users”, “Shared”, and “Maps”. This is the location where rights and permissions for users, shared reports, and information maps can be set.

The Data Library Manager is the location where libraries can be predefined. Note that when Workspace Servers or Stored Process Servers start up, they will automatically have the definitions predefined for use if the appropriate option has been checked when setting up the library.

The Server Manager allows you to specify the options for the various servers involved in your BI setup. This includes startup options for Workspace Servers, Stored Process Servers, Connect Servers, SHARE servers, and any associated spawners. The options that can be specified include most of the SAS options that can be put on an invocation line, including location of the configuration file and autoexec startup files.

The User Manager allows you to specify users of the BI tools. Permissions and rights are set here and Groups can easily be created.
SAS INFORMATION MAP STUDIO

Information Map Studio (IMS) provides a visual interface for capturing metadata about the contents of tables and how tables are related. The metadata information can then be utilized in a consistent way by end-users (single version of the truth). End-users do not have to write SQL code but instead select the data items that they want reports upon and the SQL code is automatically generated based on selections made.

There are similarities between Information Maps and SAS Views in that they both contain information about how to combine or subset data. However, Information Maps provide additional functionality in that they may contain filters, can be stored in metadata server, and they allow logical organization and presentation of components to the end-user using a familiar Microsoft Windows-like folder structure.

What does the developer see? Figure 6 shows the main interface of IMS with a two tabbed screen with the Presentation Tab highlighted. The left side of the page shows the physical tables that you want to be a part of your information map while the right side shows a chosen logical ordering of the fields. In this example, there are six tables inserted. At a glance you can pick out table names, table columns, and variable types.
Presented in Figure 7 is the Information Map Studio Relationships tab. On this page, you simply specify how the tables are to be joined. The graphic interface is very similar to Enterprise Guide and Microsoft Access. At a glance you can determine if it is a one-to-one or one-to-many join and if it is a Full, Inner, or Outer join.
Several important points about Information Maps:

- Information Maps are the only source that works with SAS Web Report Studio and its maps can access any tables SAS/ACCESS can reach along with OLAP cubes.
- You can now utilize Information Maps from BASE SAS, Enterprise Guide, and the Add-in for Microsoft Office. The INFOMAPS libname engine allows you to utilize Information Maps in data steps and most SAS procedures. You can even create Information Maps from BASE SAS without licensing Information Map Studio although you do lose some functionality and a lot of convenience.
- Information Maps can be utilized within AppDev Studio and SAS Information Delivery Portal.

SAS WEB REPORT STUDIO
SAS® Web Report Studio (WRS) is a business intelligent component part of the SAS Intelligence Platform that allows end-users to view, create, and share web-based reports. WRS provides reporting capabilities without end-users having to understand how to join tables and write reporting code. Here is a brief synopsis of the capabilities of WRS:

- Create reports from relational tables
- Create multiple section reports based on different sets of the data
- Build report templates that contain commonly used report layouts
- Report upon multi-dimensional cube data
- Six different graph types: bar charts, bar-line charts, line graphs, pie charts, progressive bar charts, and scatter plots.
- Link text, images, group break values, table values, and graph values to other reports or web pages
- Creating filters that can easily be reused
- Schedule report run times and automatic report distribution via e-mail
- Save reports as pdf files
- Export tables and/or reports to Microsoft Excel

WRS accesses Information Maps as its data source. In this way, end-users can create their own custom reports selecting the data they wish to see, applying appropriate filters, and saving the resulting reports. WRS can also run stored processes. At TIMES, running stored processes via WRS has proven very useful as SAS jobs can be run from any computer that has internet access. In addition, stored processes can be chained together within a single WRS report.
Running existing reports is very simple. You simply select ‘Report-Open’ and then navigate to the report or stored process that you want to run. Navigation is accomplished using a very Windows Explorer-like interface with file names, drill-down into folders, and an ‘Up One Level’ button. In addition, there is a radio button that allows end-users to search through ‘Shared’ reports that multiple people may have access to as well as a ‘My Reports’ button that shows only the reports that you want to keep private. As your number of reports grows, you can also conduct searches for reports by report name, keywords, or report descriptions.

Figure 8 gives you an idea of what the interface looks like as you create a new report. After selecting ‘Report-New’ from the WRS main menu, a blank report appears.

Figure 8: Creating a new report in Web Report Studio

From this point, there are basically 3 steps that are followed to create a report.

• Use the ‘Select Data’ option to select the Information Map that will be utilized and then to select the items from the Information Map that will be used in this particular report.
• After the report items are selected, you can specify any needed filtering either dynamically or utilizing pre-defined filters within the Information Map.
• The content of the report is specified within sections. Each section can contain a graph, a listing report, or a crosstabulation report. Once those steps are completed you are ready to generate and view the report. Headers, footers, fonts, styles, and other options can be altered as well.

EXTRACTION TOOL
After the development of the TIMES data warehouse, the ultimate utility of the warehouse is to make the cleaned data readily available to the researchers that need to analyze the data to address the scientific aims of the study. The “Extraction Tool” is a WRS report designed to fill that need. As shown in Figure 9, the Extraction Tool is a pre-defined WRS report that allows a researcher to select filtering information. There are a number of predefined parameter filters (not all are visible in Figure 9) that define relevant filtering criteria. After selecting the information of interest, a researcher simply clicks submit and the filters
are applied to the information map to create a data set that contains the requested information. We then utilize 3rd party software that automatically sends an e-mail to the end-user that contains a secure link to the newly created data set.

![Figure 9: TIMES Data Extraction Tool report within Web Report Studio](image.png)

**FUTURE DIRECTIONS**

Here is a brief summary of the identified future directions for the TIMES Data Warehouse project:

**BUSINESS USER BUY-IN**

Any data warehousing initiative is ultimately judged by whether or not it is used. At TIMES we need to continue expanding the number of projects that are included in the warehouse and make sure that data managers are fully aware of the utility and structure of the warehouse process including how current processes need to be adapted.

**CASCADE PROMPTS**

Coming with Phase 2 of the SAS 9.2 (1st quarter of 2009) is a much anticipated upgrade to the prompting capabilities so that prompts are dynamic. Currently within WRS, Enterprise Guide, and other clients that support it, the prompts are static so that any prompt values are not linked to prior selected values. This makes selecting filtering criteria somewhat clumsy for end-users.

**SAS DATA INTEGRATION STUDIO ENHANCEMENTS**

SAS Data Integration Studio 4.2 is also scheduled for release with Phase 2 of 9.2. Announced enhancements include:

- Integrated debugger
- Early detection of design errors
- Ability to submit individual steps, run even when a job is not complete, and view intermediate results
- Runtime progress indicators and status
- Detailed performance, warning, and error information
- Control over node execution order
• Transformation enhancements and additions

**QUERY SPEED IMPROVEMENTS**
SAS 9.2 contains an improved query engine that further optimizes query speed. It is anticipated that the SQL queries that are created dynamically from Information Maps will see improved efficiency.

**CONCLUSIONS**
Earlier in the paper, the goals of the warehouse were defined as: improve the ability to quickly analyze data, standardize the cleaning and decision making processes across projects, and document these processes so that the processes are rather independent from those that run them and thus are more easily passed to a new data manager when personnel turnover occurs. Although the warehouse process is still in development, each of these goals is within reach and we have seen substantial progress toward their attainment.

Both SAS Business Intelligence Server and SAS Data Integration Studio play important parts in the development of the warehouse process at TIMES. Much of the pre-existing SAS code has been utilized within these warehouse processes. Upcoming enhancements to the software will help improve the usability and performance of the warehouse.

**REFERENCES**


**CONTACT INFORMATION**
Your comments and questions are valued and encouraged. Contact the author at:

Kevin Davidson, Ph.D.
Texas Institute for Measurement, Evaluation, and Statistics
University of Houston
100 TLCC Annex
Houston, TX  77204-6022
832-842-7050
kevin.davidson@times.uh.edu

SAS and all other SAS Institute Inc. product or service names 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.