Paper MPSF-076

Tailoring the Use of SAS® Enterprise Miner™

Sascha Schubert, SAS Institute Inc., Cary, NC

ABSTRACT

A growing number of SAS users with different goals and skill levels need access to data mining functionality. The new generation of SAS® Enterprise Miner™ 5 and the SAS stored process facility provide an easy way to tailor data mining functionality to the user's needs.
The flexible architecture of SAS Enterprise Miner and the integration of Enterprise Miner into the SAS Enterprise Intelligence architecture allow the product users to create data mining projects for interactive or batch execution and share projects with other users.
The software's Extension facility allows users to build specific functions that are fully integrated into the Enterprise Miner workbench. Based on the integrated batch processing capabilities, model training and model scoring code can be easily extracted from SAS Enterprise Miner and integrated into the SAS Enterprise Intelligence Platform using the SAS stored process facility. For users of the SAS Add-In for Microsoft Office, customized data mining interfaces can be integrated into their favorite Microsoft applications.

KEYWORDS: SAS Enterprise Miner, Data Mining, Business Intelligence, Extension Nodes, Stored Processes, Web Services

ARCHITECTURE OF SAS ENTERPRISE MINER 5

Through its three-tier architecture, SAS Enterprise Miner 5 provides many ways of accessing data mining functionality on the data mining server (Figure 1).

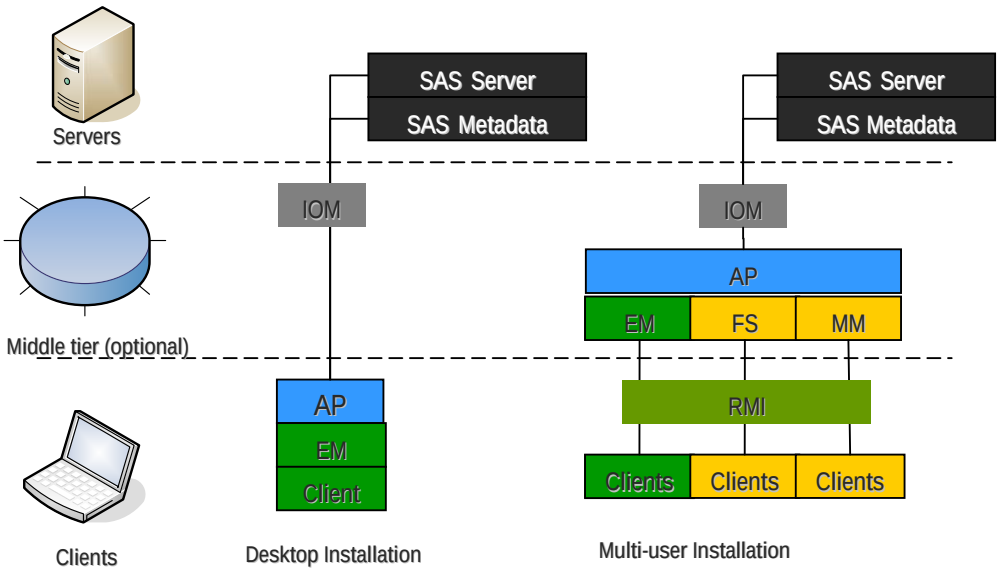

Figure 1: SAS Enterprise Miner three-tier architecture.

The architecture comprises three layers: the server layer, the middle-tier layer, and the client layer. Depending on user preferences and application requirements, there are several ways of configuring Enterprise Miner ranging from a single stand-alone desktop configuration to a multi-user configuration with three layers. The middle-tier layer – provided by the SAS Analytics Platform – provides the connectivity between the client and the server layers.

The SAS Analytics Platform provides a common application framework for analytical applications, such as SAS Enterprise Miner, SAS Forecast Server, SAS Model Manager, and SAS Inventory Policy Studio . Centralizing common middle-tier application functionality into one installable component simplifies the overall installation and administration process for these applications, especially when you take advantage of the Analytics Platform's server functionality.

Most analytics applications that use the SAS Analytics Platform require the platform to be run as a middle-tier server, which provides access to its installed applications via remote clients. However, Enterprise Miner can also run as a single-client process without a middle tier, which is appropriate for a single user environment.

In a multi-user configuration the three-tier architecture also provides asynchronous processing and batch processing support. Thus, users can be very flexible in the way they create and execute data mining projects. Here are examples of several methods to use:

1. Use the interactive Java client to build and execute data mining projects.
2. Build data mining projects in the interactive interface, start the execution on the server, disconnect and connect later from a different location to evaluate the results.
3. Prototype the Enterprise Miner analysis on the personal workstation, and use the XML save and import mechanism to promote the analysis on the production analytical server.
4. Build data mining projects using the interactive interface, save the underlying code as a SAS program and execute this code in batch mode, using scheduling applications. The results created are fully compatible with the project structure and can be opened in the interactive interface for further evaluation and modification.
5. Build projects in the interactive interface and create stored processes that can be distributed to business users throughout the organization.
6. Add functionality from other SAS modules into data mining projects for extended analytical capabilities.
7. Create customized Enterprise Miner nodes (Extension nodes) that are fully integrated into Enterprise Miner and that can be shared with other users.

The remaining sections of this paper provide examples of methods 4 to 7.

BATCH PROCESSING

SAS Enterprise Miner 5 batch processing is a SAS macro-based interface to the Enterprise Miner client / server environment that operates without running the Enterprise Miner graphical user Interface (GUI). Batch processing supports the building, running, and reporting of Enterprise Miner 5 process flow diagrams. The same diagram can be run from either the Enterprise Miner 5 GUI or from a batch job. The results can be viewed in the Enterprise Miner 5 GUI or integrated into a reporting SAS program.

Enterprise Miner 5 batch processing code is not designed to be submitted to Enterprise Miner through the Enterprise Miner GUI Program Editor. Instead, the data mining batch processing code should be submitted in a SAS batch job or submitted through the Base SAS Program Editor.

All Enterprise Miner 5 actions have batch interfaces. SAS Enterprise Miner produces SAS batch code for process flows built in the GUI, or process flow diagrams can be manually coded by experienced SAS Enterprise Miner users. The macro interface used for batch processing in SAS Enterprise Miner 5 is compatible with all SAS Enterprise Miner file structures and SAS language capabilities. These are the tools users need to automate creation and execution of a data mining analysis.

With batch processing, you can perform tasks such as these:

- Schedule processor-intensive data mining process flow diagrams for off-peak processing hours.
- Automate daily, weekly, or monthly data mining process flow diagram runs and model training.
- Automate event-driven SAS Enterprise Miner process flow diagram runs and model training.
- Automate regular data integration jobs for SAS Enterprise Miner.
- Create data mining templates for analysts and business users.

The batch processing tool is intended for use by statisticians and programmers who have strong experience in writing SAS code and building SAS Enterprise Miner models.

The SAS Enterprise Miner batch processing code can be created directly from the process flow in the workspace diagram of the workbench. From any node in the flow, the context menu includes an item called “Export Path as SAS Program”, see Figure 2. When you select this item, Enterprise Miner creates batch processing code for the entire flow up to this point.

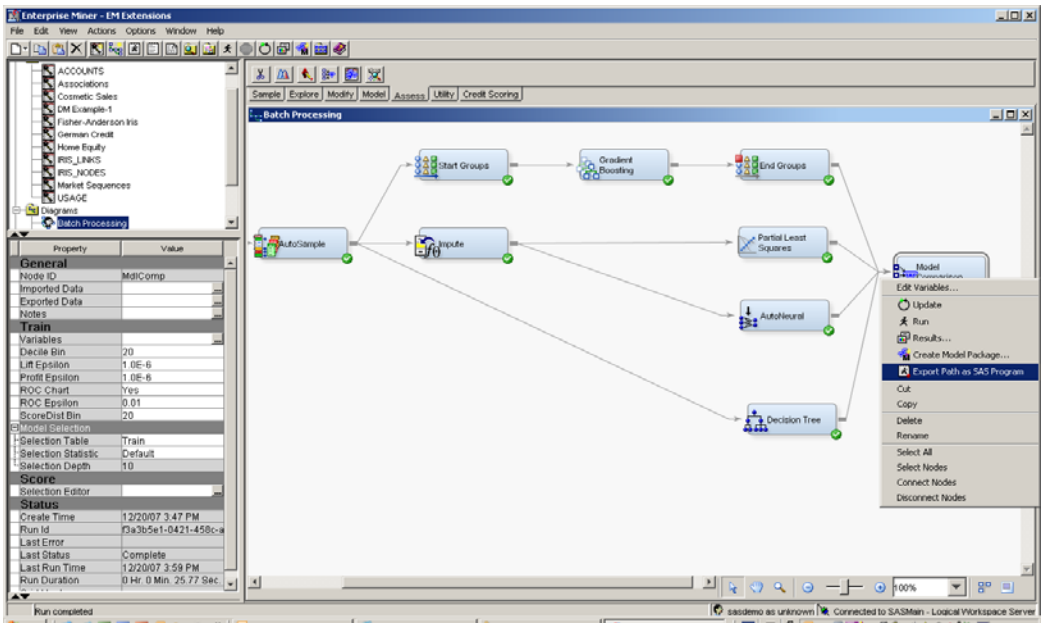

Figure 2: Export the SAS Enterprise Miner batch code from the process flow.

The batch processing code contains several components. These components configure different settings to provide SAS Enterprise Miner with the relational information and individual node settings that are required for a valid process flow diagram. The following SAS definition data sets are the required components that you will need. When the batch code is created interactively in the GUI, all these components are created automatically.

Table 1: SAS definition data sets for batch processing.

Data Source Data Set	The data source that you want to use in your process flow diagram.
Workspace Data Set	The values of the configuration properties in your workspace.
Nodes Data Set	The nodes that are used in your process flow diagram.
Actions Data Set	The actions to be taken by each node in your process flow diagram.
Connections Data Set	The connections that indicate directional data flow from predecessor nodes to successor nodes.
Node Properties Data Set	The functional properties of each individual node in the process flow diagram.

The project structure is fully compatible with the SAS Enterprise Miner project structure. That is, projects that are created in batch can be opened and modified in the interactive GUI. This way, a project can be created by batch processing in an organization based on best practices and then optimized in an interactive way using the GUI. Conversely, you can also build models in the GUI, schedule them in batch code, and then review results in the GUI as well.

The most interesting part of the batch processing code components is the Nodes Properties data set. This is where the settings of the nodes included in the process flow are defined. The code sections in Figure 3 show the settings of the Decision Tree node, where the splitting criterion has been set to

Chi-Square Probability. This is one of several available splitting criteria for the decision tree. By pre-selecting the most appropriate splitting criterion for a problem at hand, standardized best practice algorithm options can be distributed throughout the organization. These methods of setting node properties in batch will be used in the later section on data mining stored processes. For details about the structure and for examples, see the product Help.

```
*------------------------------------------------------------*;
* Create node properties data set;
*------------------------------------------------------------*;
data nodeprops;
length id $12 property $32 value $64;
id= "Tree";
property="TrainMode";
value= "BATCH";
output;
id= "Tree";
property="Criterion";
value= "PROBCHISQ";
output;
id= "Tree";
property="SigLevel";
value= "0.2";
output;
id= "Tree";
property="Splitsize";
value= ".";
output;
id= "Tree";
property="LeafSize";
value= "20";
output;
```

Figure 3: Example of SAS Enterprise Miner batch code.

Another useful way to disseminate SAS Enterprise Miner analysis templates is through Data Mining Results Packages and XML diagrams.

Every process flow diagram can be registered to the SAS Metadata server as a mining result object. This SAS package file serves as a complete record of all of the steps that were performed in the analysis, and contains all intermediate and final results. This function is useful for archiving models over long periods and for distributing models to individuals that are not SAS Enterprise Miner users, such as business managers and database administrators. Tools such as SAS Data Integration Studio and SAS Enterprise Guide can be used to access and incorporate the function logic of the model into larger processes.

Saving SAS Enterprise Miner diagrams as XML files provides another channel to share analysis templates throughout an organization. An experienced user can create a template process flow diagram to solve a specific business problem and then save this diagram as an XML file which contains all the required information. Another user can import the XML file by the push of a button, link the required data to the imported diagram, and run the flow immediately in the new environment.

THE FLEXIBILITY OF THE SAS CODE NODE

The SAS Code node can be used to incorporate new or existing SAS code into process flows. The SAS Code node extends the functionality of SAS Enterprise Miner by making other SAS procedures available in a data mining analysis. Not only the SAS DATA step but also customized scoring code can be included, in order to conditionally process data, and to concatenate or to merge existing data sets.

The SAS Code node in SAS Enterprise Miner 5.3 presents a significant upgrade from previous versions. You can use the SAS Code node to build predictive models, to format outputs, to define tables and plot views that appear on the user interface, and to modify the variables metadata. The user can edit and submit code interactively while viewing the log and output listings. The user can also run the diagram path up to and including the SAS Code node and view the results without losing the programming interface. The SAS Code node provides a full and interactive development environment within the data mining GUI.

You can use the SAS Code node to create custom nodes and share them with other users. You can also create an XML file that defines property elements to be displayed in the properties panel in the Enterprise Miner user interface. For example, you can write a SAS program that requires a percentile cutoff value and corresponding XML file that contains the property name and its valid values. The cutoff value property and a list of valid values will be displayed in the property panel.

The SAS code node is fully integrated into the data mining process flow and automatically creates the necessary metadata for reading information from upstream nodes and pushing out information to downstream nodes. The SAS Enterprise Miner Help provides detailed information about the SAS Code node functionality including getting-started examples.

INTEGRATING SAS CODE ONLY

Virtually any SAS code that is licensed at the data mining server can be run in the SAS Code node as part of the data mining process flow. Users have to decide whether the code should be run as training, scoring, or reporting code. In Figure 4, a simple correlation analysis is run as part of the training flow. The CORR procedure in Base SAS is used with its usual syntax. The SAS Code node provides integration into the SAS Enterprise Miner metadata and lists all available macro variables for dynamic assignment of options to the SAS code. The macro variables can be dragged and dropped from the macro selection list directly into the code editor. For example, the macro variable &EM_IMPORT_DATA contains the name of the data set imported into the SAS Code node. The macro %EM_INTERVAL_INPUT dynamically queries the data source metadata and imports all variables defined with the role “Input” and with the measurement “Interval”.

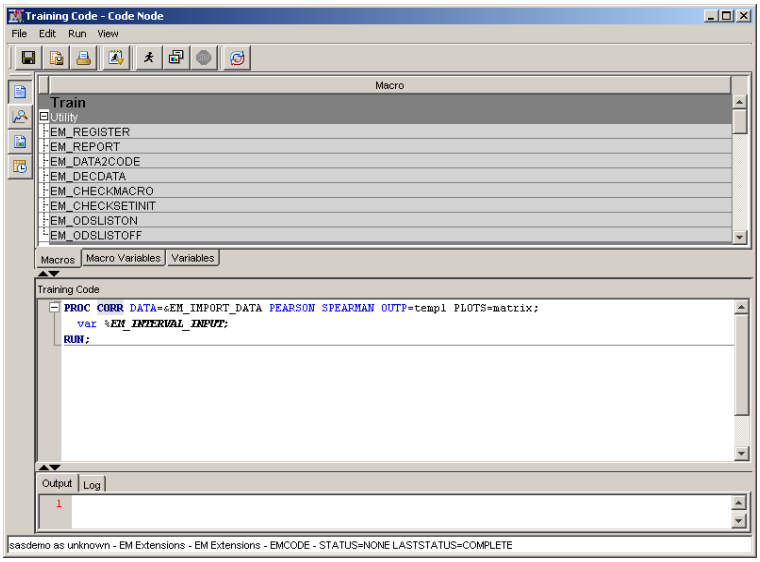

Figure 4: Running SAS procedures in the SAS Code node.

Running the SAS Enterprise Miner flow path up to the SAS Code node generates the PROC CORR output.

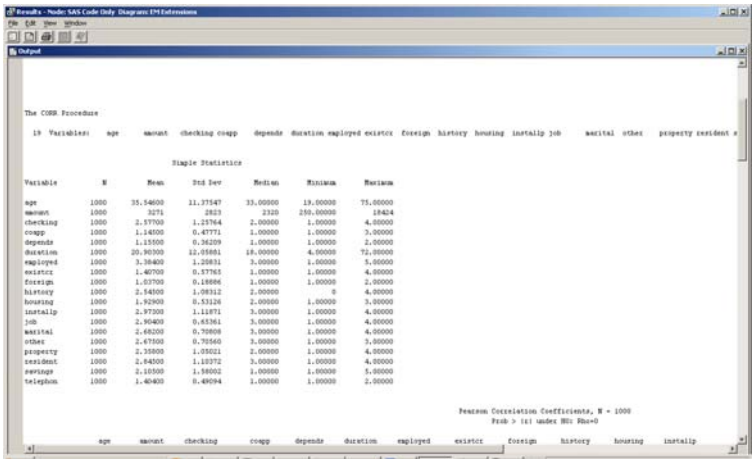

Figure 5: Output of PROC CORR in the SAS Code node results browser.

SAS CODE WITH INTEGRATED GRAPHICS

SAS Enterprise Miner also allows users to make use of the integrated graphical capabilities in customized SAS code through two macros that are part of the SAS Enterprise Miner installation: the EM_REGISTER and EM_REPORT macros.

EM_REGISTER: This macro maps a key to a file type and registers the file for use with Enterprise Miner.

EM_REPORT: This macro specifies the contents of a results window display created using a registered data set. The display contents, or view, can be a data table view or a plot view. Examples of available plots are: bar, histogram, pie, lineplot, scatterplot, lattice, matrix, density , 3-D charts, parallel axis, constellation, contour, vector, needle, and band.

When using these macros as part of the SAS code in the SAS Code node editor, you can achieve great enhancements in the layout and information relayed in the results browser of the SAS Code node. As an example, the use of EM_REGISTER and EM_REPORT is shown for the creation of the correlation plot in Figure 6. The plot uses data registered with the key “CORRPLOT” and creates a histogram of the correlation statistics for all interval input variables in the imported data set.

```
%EM_REGISTER(key=CORRPLOT, type=DATA);
%EM_REPORT(key=CORRPLOT, viewtype=HISTOGRAM, X=_X_, Y=_Y_,
FREQ=correlation, autodisplay=Y, block=Correlation,
description=Correlation Plot);
```

Figure 6: Example application of %EM_Register and %EM_Report.

Figure 7 shows the resulting graph as well as the correlation matrix table. These objects are fully integrated into the SAS Enterprise Miner environment and can use of the interactive data formatting features.

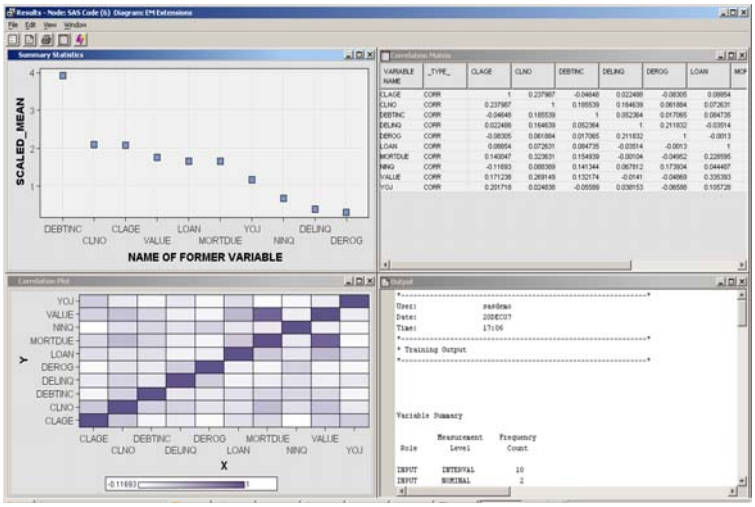

Figure 7: Customized graphics in the SAS Code node results browser.

THE SAS ENTERPRISE MINER EXTENSION FACILITY

Creating Extension nodes for SAS Enterprise Miner is a powerful tool to provide additional data mining functionality to users. Extension nodes can be seamlessly integrated into the existing SAS Enterprise Miner environment. Extension nodes go beyond the classic SAS Code node by providing macro parameters via selection lists and property sheets and by presenting graph and table output like standard SAS Enterprise Miner nodes. Default selection lists can be extended with custom developed tools written with SAS code or XML logic, which opens the entire world of SAS to data miners.

There are several steps in the process to create an Enterprise Miner Extension node. As each step builds on the results of the previous one, the results and the usage of an Enterprise Miner Extension node can be gradually refined.

An extension has two components: the code that is run when the node is executed and the property sheet that allows users to select parameters as with any other node in SAS Enterprise Miner.

Here are the steps to build EM Extension nodes:

1. Write analytics code and integrate it in SAS Code node.
2. Write reporting code and integrate it with analytics code.
3. Define and create macro variables for parameters.
4. Create icon files.
5. Create XML files for property sheets.
6. Integrate the nodes into SAS Enterprise Miner architecture.

For steps 1 and 2, any SAS code that has been integrated into the SAS Code node can be used.

3 DEFINE AND CREATE MACRO VARIABLES

The tool designer needs to define the parameters that users of the new interactive tool will be able to interactively change. The parameters should be defined in cooperation with the end-users.

The following example extends the correlation analysis from the previous section. For the sake of the exercise the Correlation code should have the following parameters:

- Define type of correlation: Pearson or Spearman
- Toggle display of summary statistics

```
%macro corr;
%if &EM_PROPERTY_CORRPears = Yes %then %do;

    proc corr data=&EM_IMPORT_DATA pearson outp=_temp;
        var %EM_INTERVAL_INPUT;
    RUN;
```

Figure 8: Syntax of macro variable CORRPears.

The code written for the correlation analysis has to be organized into a macro to allow for conditional processing based on the property selection. As an example, the syntax for the macro variable CORRPears is shown in Figure 8. This macro variable controls the calculation of the Pearson correlation statistics in the code. The prefix EM_PROPERTY_ manages the integration between the SAS code on the server and the definition of the parameters in the extension node property sheet in the SAS Enterprise Miner GUI.

Once a macro variable has been defined for every customizable property of the Enterprise Miner Extension node and integrated in the SAS Server code, the XML property file can be created.

4 CREATE ICON FILES

Using a paint program, create images and save them as gif files. Enterprise Miner needs one gif file at 16 by 16 pixels for the toolbar icon, and one gif file at 32 by 32 pixels for the node icons in the process flow diagram image. They will have the same name so they must be saved to different directories. For Enterprise Miner, they should be saved into the ext/gif16 and ext/gif32 directories respectively. If you do not create custom image files, a question mark image will appear in the user interface.

5 CREATE AN XML PROPERTY FILE

Using the templates provided with the production tools in SAS Enterprise Miner, the creation of a new property sheet is easy if you understand the basic structure of an XML file. Every node in SAS Enterprise Miner is controlled by such an XML file. The production XML property sheets are located in !SASROOT \SASAPCore\apps\EnterpriseMiner\conf\components.
For the Correlation node we have only three control properties. An example of their syntax is displayed in Figure 9..

The syntax in bold needs to be customized for the Correlation node. The SAS code that should run as part of this node needs to be made available on the workspace server at runtime. The default way for this is to integrate the SAS code into a SAS catalog and integrate this catalog into the SAS environment. In the example shown below the catalog, “SASHELP.EMEXT.CORRELATION.SOURCE” contains the source code. Also note, that the name of the parameter matches the macro variable “CORRPears” as described earlier. The parameter CORRPears can have two pre-defined values: “Yes” or “No”, with the default selection of “Yes”.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Component PUBLIC "-//SAS//EnterpriseMiner DTD Components 1.3//EN"
"Components.dtd">
<Component description="Correlation" displayName="Correlation" group="EXPLORE"
icon="correlation.gif" name="Correlation" prefix="Correlation" serverclass="EM6"
type="AF">

<PropertyDescriptors>

<Property initial="CATALOG" name="Location" type="String"/>
<Property initial="SASHELP.EMEXT.CORRELATION.SOURCE" name="Catalog"
type="String"/>

<Property description="Variable Properties" displayName="Variables" name="VariableSet"
type="String">
<Control>
<Dialog allowTyping="N" class="com.sas.analytics.eminer.visuals.VariablesDialog"
showValue="N"/>
</Control>
</Property>

<Property description="Determines if Pearson Correlation is calculated."
displayName="Pearson Correlation" initial="Yes" name="CORRPears" type="String">
<Control>
```

Figure 9: XML syntax of node properties in SAS Enterprise Miner.

Note: For illustration purposes, the syntax for only one parameter is shown. The complete code can be requested from the author.

Once the node has been installed, it will become part of the production palette of SAS Enterprise Miner.

INTEGRATION INTO THE SAS ENTERPRISE MINER ARCHITECTURE

STEP 1A: SAS Enterprise Miner 5 Full Desktop systems
Use the following steps to install the extension nodes on a system that is not running the Enterprise Miner 5 Analytics Platform (middle tier). You can check your configuration by selecting

- Help → Configuration from the main menu.

1. Close the SAS Enterprise Miner 5 client application.
2. Find the SAS root directory (for example: C:Program Files\SAS).
3. Find the SAS Analytics Platform directory (for example: C:\Program Files\SAS\SASAPCore).
4. Copy the XML file to the SAS Enterprise Miner extension subdirectory. C:\Program Files\SAS\SASAPCore\apps\EnterpriseMiner\ext.

5. Copy the GIF files to the SAS Enterprise Miner extension subdirectory.,C:\Program Files\SAS\SASAPCore\apps\EnterpriseMiner\ext\gif16 and gif32.
6. Restart the SAS Enterprise Miner 5 client application.

STEP 1B: Installation for SAS Enterprise Miner 5 Shared Platform systems

Follow these steps to install the extension nodes on the SAS Enterprise Miner 5 Shared Platform system. You do not need to update each individual end-user client. The extension nodes will be available to all users.

1. Close the SAS Enterprise Miner 5 shared platform.
2. Find the SAS root directory (for example: C:\Program Files\SAS).
3. Find the SAS Analytics Platfornm directory (for example: C:\Program Files\SAS\SASAPCore).
4. Copy the XML file to the SAS Enterprise Miner extension subdirectory, C:\Program Files\SAS\SASAPCore\apps\EnterpriseMiner\ext.
5. Copy the GIF file to the SAS Enterprise Miner extension subdirectory, C:\Program Files\SAS\SASAPCore\apps\EnterpriseMiner\ext\gif16 and gif32.
6. Restart the SAS Enterprise Miner 5 Shared Platform.

STEP 2: Install the Source Code

1. Integrate the SAS code into the catalog, “SASHELP.EMEXT”.
2. Alternatively you can create a separate catlog for your source code and integrate it into the SASHELP library definition by making the directory path available to the –SASHELP option in your SAS config file. Here is an example:.

```
-SASHELP        (
                 "!SASCFG\SASCFG"
                 "!sasroot\core\sashelp"
                 "!sasext0\dmine\sashelp"
```

Once integrated, the Enterprise Miner Extension node becomes part of SAS Enterprise Miner production node palette. Figure 10 shows the use of the Correlation Extension node as part of the data mining process flow.

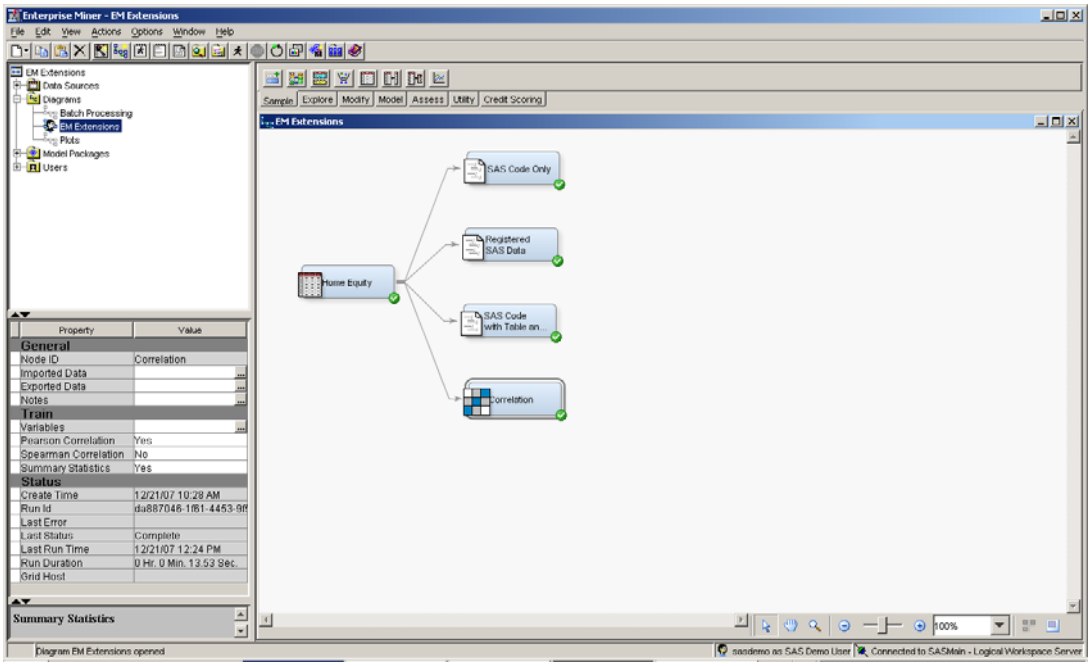

Figure 10: Enterprise Miner Extension node integrated into SAS Enterprise Miner tool palette.

Examples of Extension nodes for SAS Enterprise Miner can be obtained from *Data Preparation for Analytics Using SAS*. This book provides four nodes that have been created based on the Extension facility:

TrendRegression	calculates derived variables that describe the trend of an interval variable in up to two time intervals and creates a concatenated group variable.
Correlation	calculates derived variables that describe the correlation of values with their overall mean per time ID or with other input variables.
CategoryCount	calculates derived variables for categorical data. Aggregations like counts, distinct counts, or proportions are calculated.
Concentration	calculates derived variables that describe the concentration of an interval variable in a sub-hierachy of the analysis subject.

SAS Institute also offers a one-day education course called “Extending SAS Enterprise Miner” that enables users to develop Extension nodes to add customized functionality to the SAS data mining workbench.

DATA MINING THROUGH SAS STORED PROCESSES

The SAS Enterprise Intelligence platform offers a powerful integration of SAS capabilities with a multitude of external client front-ends through stored processes. With the SAS Add-In for Microsoft Office, stored processes can be easily made available to users of Microsoft Office front-ends, such as Microsoft Excel.

SAS offers several ways to create stored processes. For example, you can do this by using SAS Enterprise Guide or the SAS Management Console. This paper assumes that you are familiar with the creation of stored processes. For more information on SAS stored processes and the SAS Add-In for Microsoft Office, see the product documentation that is available at http://support.sas.com/documentation/index.html

Using the source code created for the batch processing above as a basis, it is easy to create a stored process in SAS Enterprise Guide. This code trains a Decision Tree on the data imported to the Tree node. Using SAS stored processes and the SAS Add-In for Microsoft Office, you can make data mining algorithms available to users in an organization in a customized way. Based on the profile of the information consumers, the interface to the data mining algorithms can provide access to the functionality they need and hide the complexity that is not needed.

Again, every parameter that users will be able to modify needs to be defined as a macro variable. Here, the structure of the SAS batch processing code is very helpful, since every parameter is clearly defined and can be easily identified in the code. As an example, see the definition of the splitting criterion in the Decision Tree node by the macro variable &_Crit (Figure 11).

```
id= "Tree";
property="Criterion";
value= "&_Crit";
output;
```

Figure 11: Dynamic definition of splitting criterion in Decision Tree node by a macro variable.

The stored process wizard in SAS Enterprise Guide allows users to create a user interface for the parameter settings interactively. When the stored process has been created and integrated into the SAS Metadata Server, it becomes available to users from external interfaces.

Figure 12 shows the example of running the Decision Tree stored process from an Microsoft Excel interface. With this, business users can easily run data mining algorithms from an envioronment they might be more familiar with. The results of the stored process can be streamed to the Microsoft Excel interface to integrate them into front-ends for further proceessing if required as shown in Figure 13.

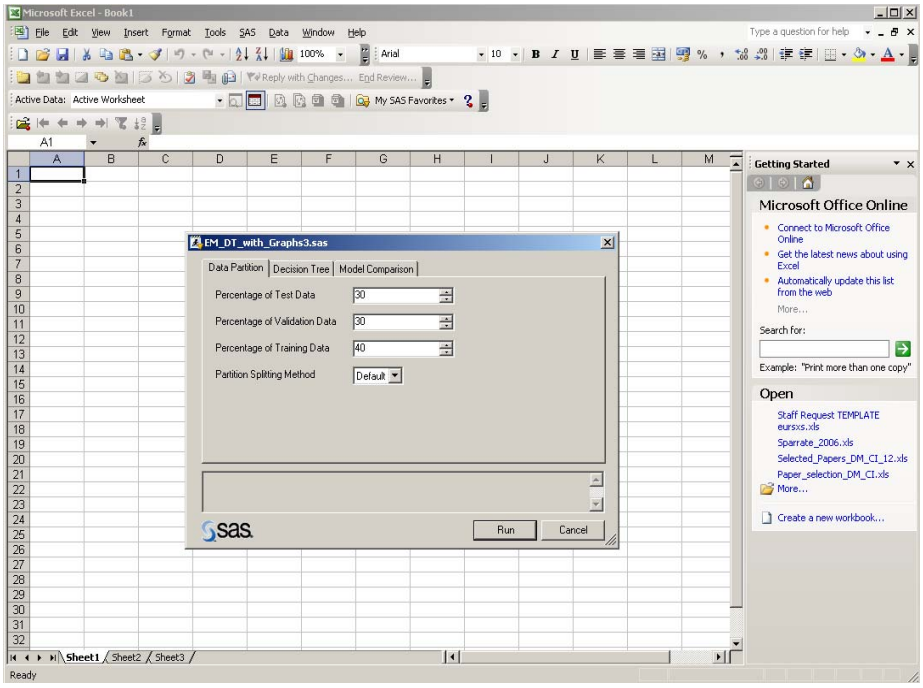

Figure 12: Dynamic parameter interface for SAS stored process in Microsoft Excel.

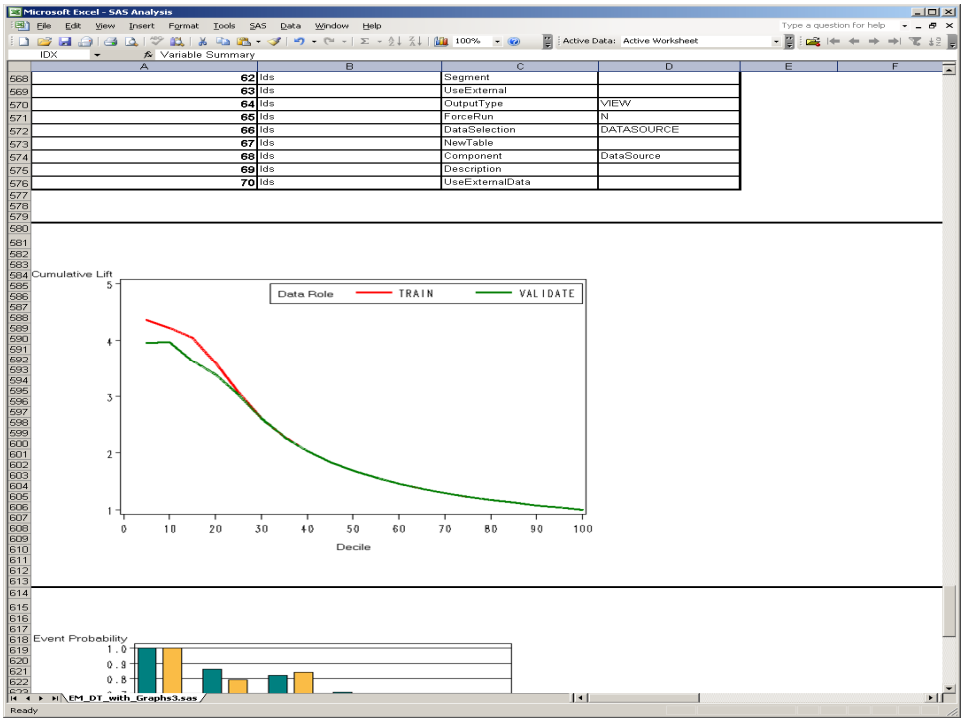

Figure 13: Results from a SAS stored process streamed into Microsoft Excel.

CONCLUSION

To give users at a variety of skill levels access to their organization's advanced analytical processing capabilities requires interfaces that fit those skill levels. Through its flexible three-tier architecture and through its integration into the SAS Enterprise Intelligence Platform, SAS Enterprise Miner 5 delivers this flexibility and can cater to users ranging from the occasional business user to the power user who needs the power to tweak every detail of the data mining process.

REFERENCES

SAS Institute Inc., "SAS Enterprise Miner 5.3 Fact Sheet," SAS Institute Inc., Cary, NC.
http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf

SAS Institute Inc., *SAS Enterprise Miner 5.3 Help*. SAS Institute Inc., Cary, NC.

SAS Institute Inc., *Administrator's Guide for SAS Analytics Platform 1.3*, Cary, NC: SAS Institute Inc., 2006.
http://support.sas.com/documentation/onlinedoc/miner/admin13.pdf

Svolba, Gerhard, *Data Preparation for Analytics Using SAS*, Cary, NC: SAS Institute Inc., 2006.

What's New in SAS Enterprise Miner 5.3, SAS Institute Inc., Cary, NC.
http://support.sas.com/documentation/whatsnew/91x/emgui53whatsnew.htm

ACKNOWLEDGMENTS

The author would like to thank the following SAS employees for their valuable contributions to this paper: Wayne Thompson, David Duling, Dominique Latour, Allen Mcdowell, and Gerhard Svolba.

CONTACT INFORMATION

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

Sascha Schubert
SAS Institute Inc.
Domaine de Grégy
Grégy-sur-Yerres
77257 Brie Comte Robert Cedex
Sascha.Schubert@sas.com