Paper 339-2009

SAS® and Teradata: Accelerating the Power to Know

R. Wayne Thompson, SAS Institute, Cary, NC
Robert S. Ray, SAS Institute, Cary, NC

ABSTRACT

The collaboration between SAS® and Teradata on enterprise class scoring has produced the SAS® Scoring Accelerator for Teradata which deploys SAS® Enterprise Miner™ scoring models in a highly scalable format to a Teradata server. Once deployed, these models have been shown to be linearly scalable, completely leveraging the power of the parallel shared-nothing architecture of Teradata. This presentation will cover the model deployment process from start to finish and include performance profiles from a multi-node Teradata server. There will also be a preview of how this technology will be integrated into SAS® Model Manager 2.2.

INTRODUCTION

Data mining is the process of extracting useful information from data to help support knowledge discovery and decision making. It includes the primary tasks of model building and model scoring. Model development typically requires preparing a historical data set(s) and applying data mining modeling algorithms to classify a known outcome or segment the cases into more homogenous groups. Numerous candidate models are often tried with the objective to train a set of functions to produce the best model.

Scoring is the process of applying a trained scoring algorithm built on historical data to new data typically stored in a production operational environment. Model deployment is the end result of data mining where the ROI from the mining process is often realized. A model is often used many times after it is created to score new data. Scoring is typically done on a periodic basis – either daily, weekly, monthly or on an event driven basis.

Many SAS customers store their operational Enterprise Data Warehouse (EDW) in Teradata. Conventional scoring of SAS models requires connecting to the Teradata server through SAS® Access to Teradata to extract rows to SAS for scoring. The scores are commonly bulk loaded back to the Teradata EDW. Many customer scoring applications require scoring large numbers of models against enormous tables. As the number of rows in the table grows over time, network latency grows because the amount of data that is fetched from the Teradata EDW to the SAS scoring process increases.

The SAS® Scoring Accelerator for Teradata is new product for publishing SAS® Enterprise Miner™ models into Teradata specific scoring functions for execution directly in Teradata. Organizations can now leverage a best of breed solution comprised of the robustness and flexibility of SAS analytics and the inherent scalability of the Teradata shared-nothing architecture. This paper is provides an introduction to the SAS Scoring Accelerator for Teradata covering these topics,

- Developing and exporting the SAS® Enterprise Miner™ model.
- Publishing and registering the scoring files using the Scoring Accelerator for Teradata.
- Executing the scoring functions.
- Performance tips and profiles.
- Integration of SAS Model Manager 2.2 with the Scoring Accelerator.

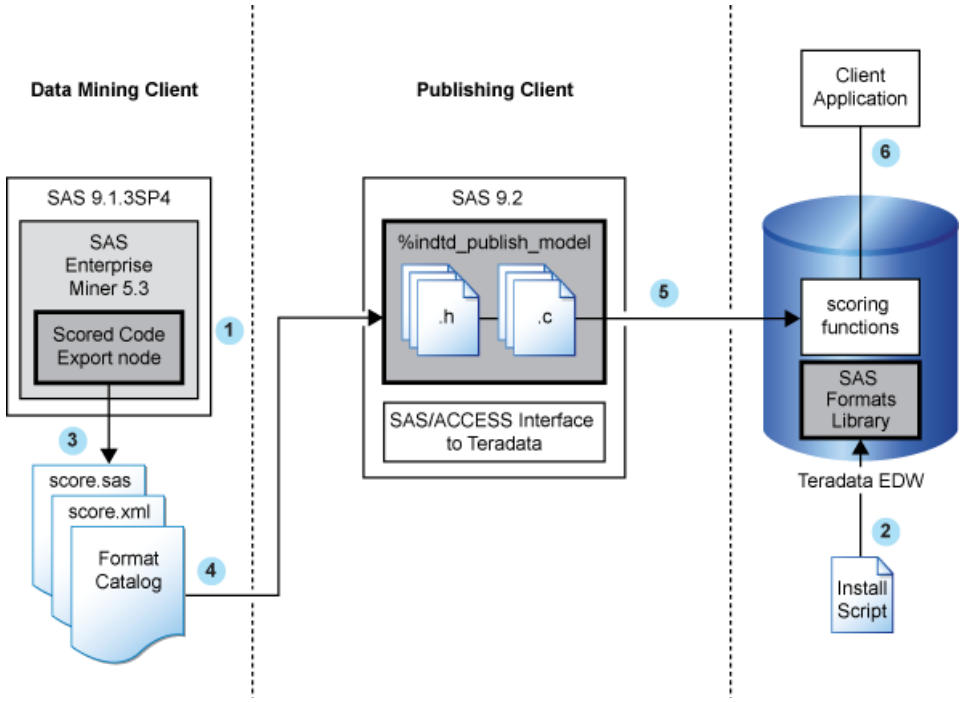


Figure 1. SAS® Scoring Accelerator for Teradata Topology.

DEVELOPING AND EXPORTING THE SAS® ENTERPRISE MINER™ MODEL

SAS® ENTERPRISE MINER™ OVERVIEW

SAS® Enterprise Miner™ is the SAS solution for data mining providing unparallel model development and deployment opportunities. Delivered as a distributed client-server system, SAS® Enterprise Miner™ is well suited for joint workgroup collaborations and large data mining applications. The process flow diagram eliminates the need for manual coding and reduces the model development time for both business analysts and power statisticians. The system is customizable and extensible; users can integrate their code and build new nodes for redistribution.

A key use case for SAS® Enteprise Miner™ is the generation of the complete scoring code from the modeling process. SAS® Enterprise Miner™ models have long been used in industries, such as banking, telecommunications, insurance, retail, pharmaceutical research and manufacturing to reduce costs, detect fraud, enhance research, and increase sales. SAS® Enterprise Miner™ generates scoring code for the model in several languages. At the core is the SAS data step scoring code which covers a full range of Enterprise Miner™ modeling functions including preliminary transformations and customized scoring code written in an Enterprise Miner SAS Code node. Enterprise Miner SAS scoring code requires only BASE SAS for deployment with the exception of a few algorithms, such as Memory Based Reasoning and Associations which also require the underlying SAS® Enterprise Miner procedure. In the early 2000s, SAS® Enteprise Miner™ delivered scoring code in the C and Java programming languages to enable application developers to embed the power of SAS models into non-SAS environments. Although several customers deploy SAS® Enterprise Miner™ C and Java scoring code in data bases and on the web, many organizations do not have dedicated C or Java programmers to create the scoring function or require a more turnkey solution. In the mid 2000s, SAS® Enterprise Miner supported producing the Predictive Modeling Markup Language (PMML) scoring, which is a standard for sharing models among members of the Data Mining Group consortium. SAS and Teradata conducted extensive compatibility tests using over 150 models to validate SAS PMML models can be deployed in Teradata via the Teradata Analytic Data Set Generator. PMML presently covers the core subset of data mining functions and is also well suited when the variable transformation and enrichment layer has been performed primarily in the DBMS. With the SAS9.2 system, SAS and Teradata through extensive research and development collaboration have delivered the SAS® Scoring Accelerator for Teradata to enable user to score a larger class of SAS® Enteprise Miner™ models than is presently covered by the PMML solution. The SAS Scoring Accelerator for Teradata also provides more automation to customers than manually creating and compiling the scoring function from SAS® Enteprise Miner™ Java and C score code. For a comparison of the SAS™ Enterprise Miner™ modeling functions supported for each scoring language, please see Chapter 2 of the SAS Scoring Accelerator 1.4 for Teradata Users Guide (SAS 2008).

The SAS Scoring Accelerator for Teradata is compatible with SAS Enterprise Miner 5.3 SAS9.1.3 and SAS Enterprise Miner 6.1 SAS9.2. SAS Enterprise Miner 5.2 and 5.1.3 projects can be used in Enteprise Miner 5.3 without any migration action. SAS Enterprise Miner 5.3 and 6.1 includes a SAS macro that enables you to convert SAS Enterprise Miner 4.3 projects.

SCORE CODE EXPORT EXTENSION NODE

The SAS® Enterprise Miner™ Score Code Export node exports score files that are used as input to the publishing macro of the SAS Scoring Accelerator. The SAS® Scoring Accelerator for Teradata media includes a zip file containing the Score Code Export node files. An extension node provides a mechanism for users to extend the functionality of a SAS® Enterprise Miner™ installation. The Score Code Export node was developed by SAS and integrates into the SAS Enterprise Miner 5.3 or 6.2 so that, from the perspective of the end user, it functions and works like any other node. The installation instructions for the Score Code Export node are provided in Chapter 3 of the Configuration Guide for SAS 9.2 Foundation for Microsoft™ Windows® (SAS Institute 2008).

To use the Score Code Export node, drag-and-drop the node onto the diagram workspace from the Utilities toolbar and connect it to a terminal Score node in your process flow. The Score node aggregates code for all nodes that generate scoring code. In this example, the score code from the Interactive Binning node used to optimally class the variable input attributes along with the final gradient boosting model is assembled (Figure 2).

Figure 2. SAS Enterprise Miner Process Flow Diagram including the Score Code Export extension node.

The score code for SAS Enteprise Miner 6.1 has been optimized to contain only the required variable inputs to score the model so for this release, set the Optimize Scoring Code property for the Score node to Yes. To define the name of the folder for the output files that are created when the Score Code Export node is executed, set the path for the output directory (Figure 3).

Property	Value
General	
Node ID	CodeXpt
Imported Data	
Exported Data	
Notes	
Train	
Rerun	Yes
Output Directory	C:\models\
Name	scorecard
Status	
Create Time	3/6/09 11:20 AM
Run Id	1326a30f-f691-48b6-8eca-d3a1
Last Error	
Last Status	Complete
Last Run Time	3/11/09 8:05 PM
Run Duration	0 Hr. 0 Min. 12.93 Sec.
Grid Host	
User-Added Node	Yes

Figure 3. Score Code Export node properties.

The Score Code Export node exports the following files to the output directory:

- The BASE SAS Scoring Code (score.sas)
- An XML file containing scoring variables and other properties (score.xml).

- Another XML containing descriptions of the final variables that are used and created by the scoring code (emoutput.xml).
- A SAS data set that is a ten-row sample of the scored data set showing typical input attributes, intermediate variables, and final output variables. This file is useful to test and debug new scoring processes.
- Another SAS data set that is also a ten-row sample of the training data showing typical cases of the input attributes used to develop the score code.
- A formats catalog if any variables in the model development data set use a user-defined format.

PUBLISHING AND EXECUTING THE SCORING FILES

At the heart of the SAS Scoring Accelerator for Teradata is the DATA Step to C translator which is driven by the surrounding publishing macro code to automatically convert the scoring code that SAS® Enterprise Miner™ originally generated to be run inside the SAS system into a form that can be run directly, in parallel inside Teradata. The model publishing macro, %INDTD_PUBLISH_MODEL, is invoked with elements from the SAS Enterprise Miner Score Code Export node, score.sas and score.xml, and other parameters which it supplies to the source translator to produce the .c and .h files necessary to register the scoring function in Teradata.. The DATA step language translator has been specifically tuned to translate the type of statements that SAS® Enterprise Miner™ generates. Once defined, the score function can be invoked directly from Teradata from a SQL statement which can run the score function in parallel across the entire Teradata grid. For example,

```
CREATE TABLE RESULTS AS SELECT
   BASEBALL_EM_EVENTPROBABLITY( COL1, COL2, COL3…COLn ) AS EVENT_PROB
   FROM INPUT TABLE.
```

Because the scoring function can reference SAS formats, even those which are user defined with PROC FORMAT, the publishing macro also inspects the SAS Enterprise Miner Score Code Export node for a format catalog reference. If found, the macro translates the format range and label data into C structures and places these into an .h file. This .h file will be included in the source that is exported to Teradata during the publishing process. Because the format definitions are embedded in the exported scoring code, format execution is extremely fast. Also, because each score UDF is “linked” to its format at publish time, it is not affected by subsequent changes to the format catalog. The score execution path is frozen at the time of export. This differs from native execution in the SAS system where the DATA Step code will reload formats at each execution and so it will use the definition it finds at runtime. The static nature of the compiled function can be a safeguard against unintentional association of different revisions of a DATA step and format at runtime.

INVOKING THE SCORING ACCELERATOR MACRO

The code sample below is an example of how the model publishing macro is typically invoked inside the SAS system. Notice that the INTDLOGON macro variable must be defined with the users account on Teradata. The %INDTDPRM macro searches the autocall library for the indtdpm.sas file. The indtdpm.sas file contains all the macro definitions that are used in conjunction with the %INDTD_PUBLISH_MODEL macro. The publishing macro will form a batch execution script which it will send to Teradata on the user’s behalf. In its more simple form, the publish macro requires only two parameters, the directory where the SAS Enterprise Miner Score Export node can be located and the name of the model within the node that is to be published.

```
%indtdpm;
%let indtdlogon = terabase/user1,open1;
%indtd_publish_model( dir=C:\SASIN\baseball1, modelname=baseball1 );
```

An example of the script that is produced when the macro runs is shown below. Notice that the function name is a combination of the model name and the specific returned statistic. For each returned statistic, a separate UDF entry point will be generated although they will share the same scoring logic.

```
CREATE FUNCTION baseball1_EM_eventprobablility
   (
   "CR_ATBAT" float,
   "CR_BB" float,
   "CR_HITS" float,
   "CR_HOME" float,
   "CR_RBI" float,
   "CR_RUNS" float,
   "DIVISION" varchar(31),
   "LEAGUE" varchar(31),
   "NO_ASSTS" float,
   "NO_ATBAT" float,
   "NO_BB" float,
   "NO_ERROR" float,
   "NO_HITS" float,
   "NO_HOME" float,
   "NO_OUTS" float,
   "NO_RBI" float,
   "NO_RUNS" float,
   "YR_MAJOR" float
   )
   RETURNS float
   LANGUAGE C
   NO SQL
   PARAMETER STYLE SQL
   NOT DETERMINISTIC
   CALLED ON NULL INPUT
   EXTERNAL NAME 'SL!"jazxfbrs"'
   '!CI!tkcsparm!c:\SASIN\baseball1\tkcsparm.h'
   '!CS!baseball1_EM_eventprobability!c:\SASIN\baseball1\EM_eventprobability.c';
```

The UDFs must be published to a particular database so the permissions that apply the containing database will be applied to the UDF as well so it will be reasonable for a DBA to publish scoring routines to a shared read-access database. The scoring accelerator macro has three primary actions ***create***, ***replace***, and ***drop***. In addition, the scoring UDF can be registered as ***protected*** or ***unprotected***. Protected mode causes Teradata to isolate the UDF execution in a separate process space from the body of the query. This is an additional safety measure against program faults in the scoring code itself but at a cost in performance. The normal life cycle of a published scoring model will be to publish it using *protected* mode for model validation and then to republish the model in *unprotected* mode for maximum production use throughput.

PEFORMANCE PROFILE

Because scoring function operates on rows of the database independently, they are perfect candidates for a parallel shared-nothing architecture. As a result, the scalability of these routines is perfectly linear with respect to the number of nodes applied to the score query as shown in the following graph. Full table scoring can now be accomplished in as little time as necessary by applying sufficient database resources.

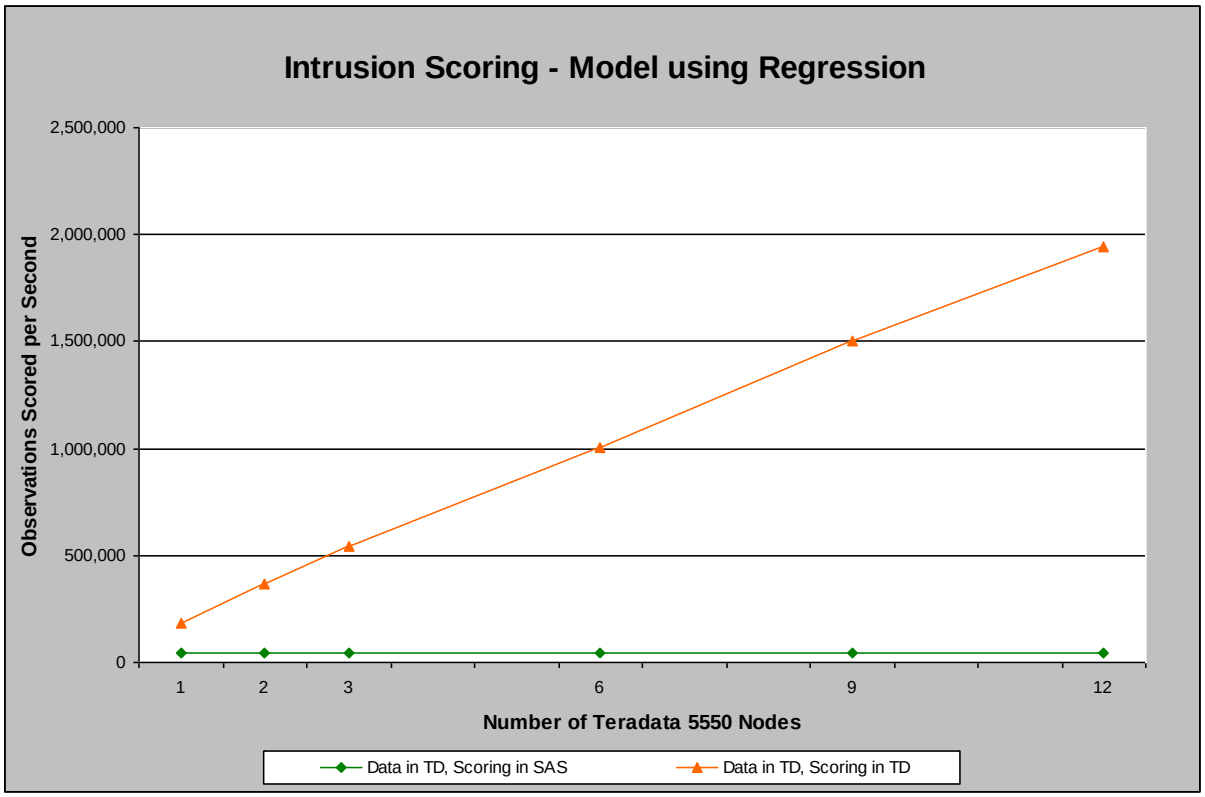


Figure 4. Plot comparing the scalability of a model (Intrusion) executed in DATA step via SAS.ACCESS versus the same code embedded in Teradata as a scoring function

SAS® MODEL MANAGER 2.2 INTEGRATION

SAS® Model Manager is designed for the selection, maintenance, and continuous enhancement of analytical models for operational decision making. SAS® Model Manager is all about enabling processes to effectively manage and deploy analytical models by delivering all the necessary functionality for each stage of the model life cycle. It includes a secure, centralized repository for storing and organizing models. Models are organized by projects that map to the business purpose or application. Each project contains the champion and challenger models along with extensive metadata, scoring code, data sources definitions, and supporting documentation. Supported models include prediction, classification, segmentation, and rules-based models developed using SAS Enterprise Miner™, SAS/STAT®, and Base SAS®. Models can be promoted from one version to another within a project as they are applied over time. Event logging and notes correspondence are also supported.

SAS Model Manager 2.2 for SAS9.2 provides integration with the SAS Scoring Accelerator for Teradata for publishing and validating SAS Enterprise Miner models as Teradata scoring functions. In this example “EM Scorecard Detail” is the champion model for the version “Q22009” within the “Application Scorecard” project.

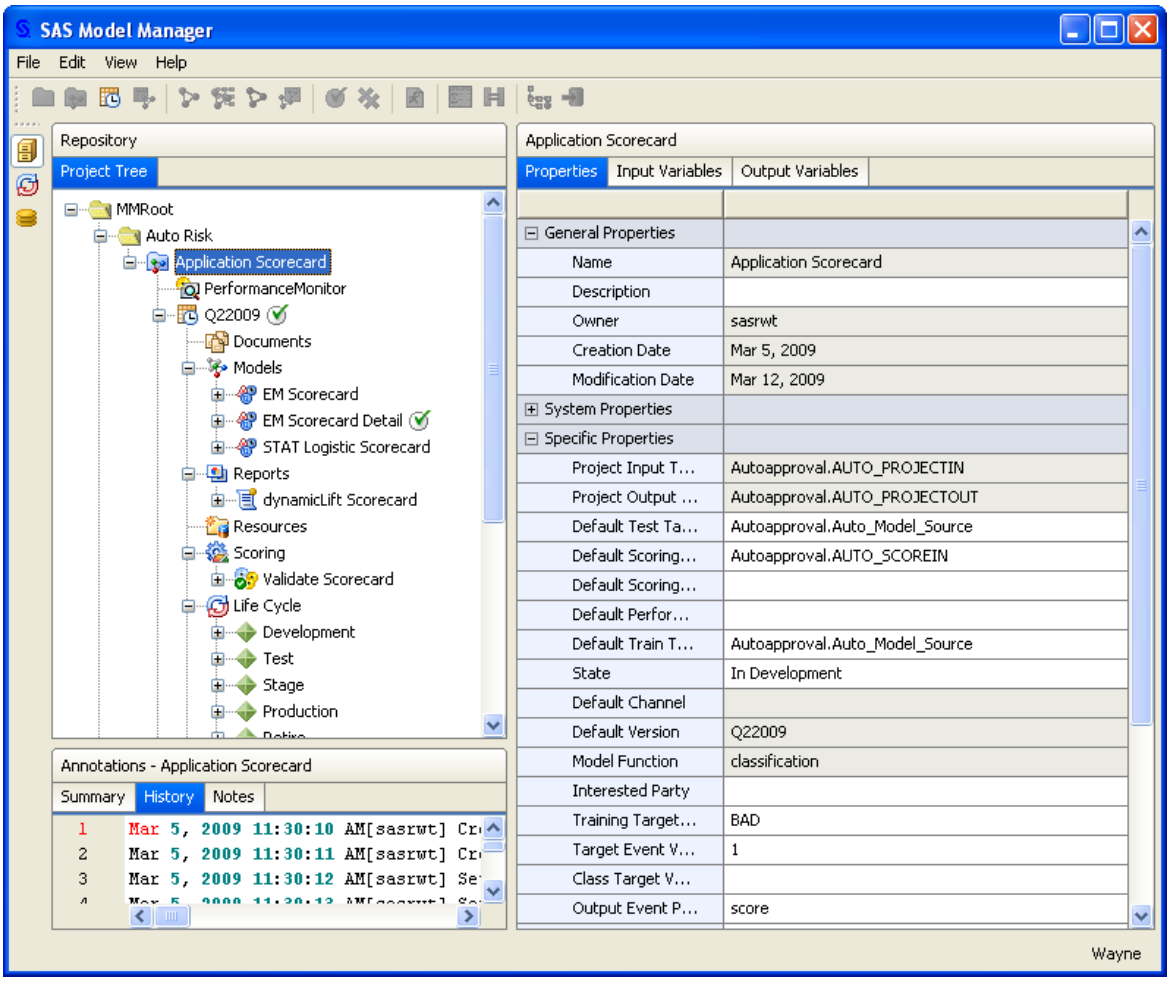
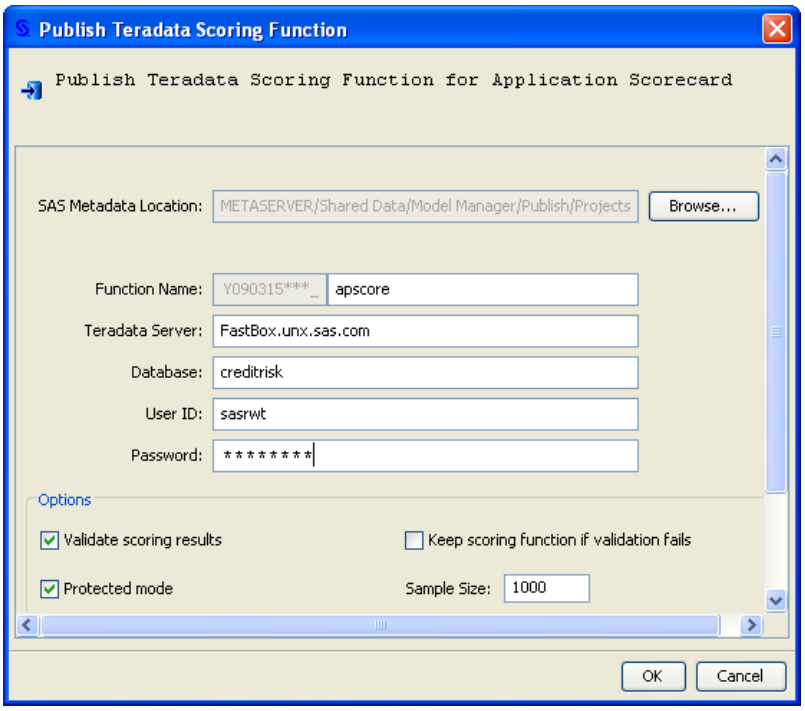


Figure 4. SAS Model Manager 2.2.Repository View.

The Publish Teradata Scoring Function feature of SAS Model Manager 2.2 enables users to export the champion model for the project to the SAS Metadata and execute the %INTD_PUBLISH_model macro of the SAS Scoring Accelerator for Teradata (Figure 5).

.



Benefits of using SAS Model Manager to publish the SAS model as a Teradata scoring function include,

- The champion model for the frozen version within the project is the only model that can be promoted. Data miners can set the champion model which will automatically be selected when the Scoring Officer publishes the model.
- A single scoring function is generated for each published model which generates a concatenated string for all of the outcome variables of the scoring code when executed.
- The scoring function name automatically includes a time stamp prefix for uniqueness.
- The scoring results are compared within Teradata and SAS using a user defined sample of the training data.
- Lifecycle signoff is supported for promoting the model from development to test to production to retire milestones.
- Model monitoring jobs can be executed to identify when the scoring results are changing over time indicating model decay.

CONCLUSION

Organizations today rely more on analytics to accurately predict their business outcomes. When the operational data is stored in an external database, such as Teradata the transfer of extremely large data sets to the SAS system for model scoring can be a performance bottleneck. The SAS® Scoring Accelerator for Teradata improves the integration between SAS and Teradata by enabling users to automatically publish and score SAS® Enterprise Miner™ models in the Teradata server. Scoring SAS models in Teradata achieves higher performance run times and faster development and processing by reducing data movement and eliminates the need to recode scoring code. The SAS Scoring Accelerator for Teradata supports a large class of SAS® Enterprise Miner™ models including the preliminary statistical transformations. Integration is also included with SAS Model Manager for those companies that need to manage a large repository of models throughout the entire development, test, production, and retire lifecycle.

REFERENCES

- SAS Institute Inc. SAS® Scoring Accelerator 1.4 for Teradata: User's Guide. Cary, NC: SAS Institute Inc. 2008
- SAS Institute Inc. Configuration Guide for SAS 9.2 Foundation for Microsoft® Windows®, Cary, NC: SAS Institute Inc., 2008.

CONTACT INFORMATION (HEADER 1)

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

Robert Ray, Senior Software Manager
SAS Institute Inc.
R1436
SAS Campus Drive
Cary, NC 27514
Email: robert.ray@sas.com

Wayne Thompson, Product Manager
SAS Institute Inc.
S6100
SAS Campus Drive
Cary, NC 27513
Email: wayne.thompson@sas.com