



Teradata Warehouse Miner

How to Use Teradata Warehouse Miner at the University of Arkansas

Teradata University Network (TUN) members – faculty and students—can take advantage of Teradata Warehouse Miner (TWM) housed at the University of Arkansas (UA). This document illustrates using TWM by stepping through a decision tree and an association analysis example. The purpose is not to explain decision trees or other data mining algorithms but rather to focus on how to do a decision tree and association analysis data mining tasks using TWM. Before getting started, a brief overview of data mining is provided.

Data Mining—

Data mining has many definitions and may be called by other names such as knowledge discovery. It is generally considered to be a part of the umbrella of tasks, tools, techniques etc. within business Intelligence (BI). Many corporate managers consider BI to be the heart of all the processes that support decision making at all levels. A definition of data mining typically includes large datasets, discovering previously unknown knowledge and patterns and that this knowledge is actionable. That what is discovered is not trivial but can be usefully applied. BI and its Data Mining component are receiving considerable attention and fanfare as companies utilize BI for competitive advantage.

Different authors may address the data mining tasks slightly different from each other but the following terminology provides a helpful and useful basis for discussing data mining. The data mining tasks are:

- Description
- Estimation
- Classification
- Prediction
- Association Analysis
- Clustering

Description—uses descriptive statistics to better understand and profile areas of interest. Thus a variety of well known statistical tools and methods are used for this task—including frequency charts and other graphical output, measures of central tendency and variation.

Data Mining Tasks with a Target or Dependent Variable

Estimation, **classification** and **prediction** are data mining tasks that have a target (dependent) variable. Sometimes these, are referred to as predictive analysis; however, many authors reserve the term Prediction to use of models for the future. The terms **supervised** and **directed** apply to these data mining tasks. **Estimation** data mining tasks have an interval level dependent target variable whereas **classification** data mining tasks have a categorical (symbolic) target variable. An example of an estimation data mining task would be estimating family income based on a number of attributes; whereas a model to place families into the three income brackets of Low, Medium or High would be an example of a classification data mining task. Thus, the difference between the two tasks is the type of target variable.

When either an estimation data mining task or classification task is used to predict future outcomes, the data mining task becomes one of **Prediction**. Again, estimation and classification are referred to as predictive models because that would be the typical application of models built for these data mining tasks.

In summary, the most important concept is that estimation and classification data mining tasks require a target variable. However, the difference lies in the data type of the target variable.

Data Mining Algorithms for Directed/Supervised Data Mining Tasks—linear regression models are the most common data mining algorithms for **estimation** data mining tasks. Of course, linear regression is a very well known and familiar technique. A number of data mining algorithms can be used for **classification** data mining tasks including **logistic regression**, **decision trees**, **neural networks**, **memory based reasoning (k-nearest neighbor)**, and **Naïve Bayes**.

Data Mining Tasks without a Target or Dependent Variable

Association Analysis and **Clustering** are data mining tasks that do not have a target (dependent) variable. Affinity analysis is another term that refers to association analysis and is typically used for market basket analysis (MBA) although association analysis can be used for other areas of study. MBA is essentially analyzing what purchases tend to be purchased together—that is what items tend to have an affinity with other items. **Clustering**, having no target variable, algorithms attempt to put records into groups based on the record's attributes. The critical concept is that of similarity—those within a cluster are very similar to each other and not similar with those in another cluster.

Note—because these data mining tasks do not have a target variable, their corresponding models are not generally used for prediction. Thus, they are many times exploratory in nature and their results can be used downstream in predictive models.

Data mining overview summary

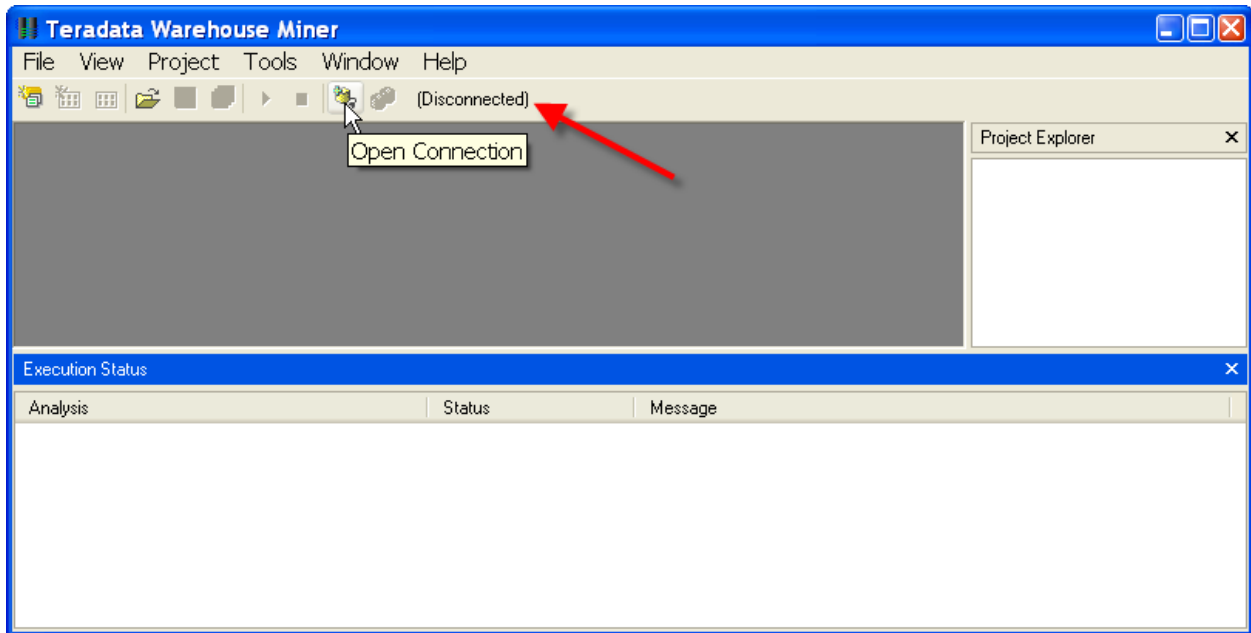
The following table summarizes the data mining tasks and associated data mining algorithms.

<i>Data mining tasks</i>	<i>Target Variable</i>	<i>Typical Data Mining Algorithm(s)</i>
Description	No	Statistics, including descriptive, & visualization
Estimation	Yes Interval Numeric	Linear Regression
Classification	Yes Categorical	Logistic Regression, Decision Trees, Neural Networks, Memory Based Reasoning, Naïve Bayes
Prediction	Yes	Estimation and Classification models for prediction
Association Analysis	No	Affinity Analysis (Market Basket Analysis)
Clustering	No	<i>k</i> -means, Kohonen Self Organizing Maps (SOM)

Teradata Warehouse Miner

To become a TUN member and details for access to the University of Arkansas Teradata resources, see the “How to Become a TUN member...” document which includes a link to remote access details.

Using the icon on the Desktop or the All Programs menu, to open Teradata Warehouse Miner 5.0. Note that TWM is Disconnected when initially opened. Click the Open Connection icon as shown to open the

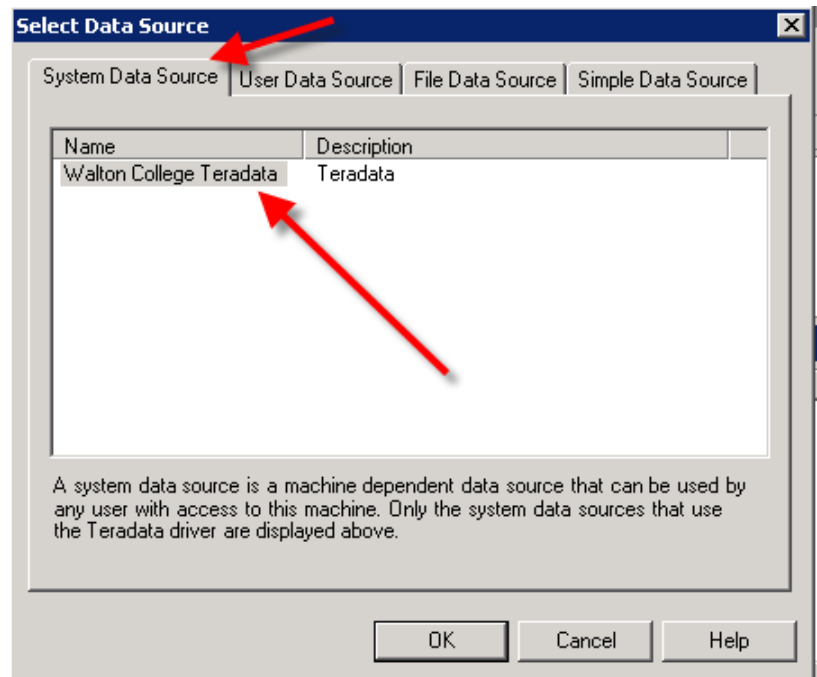


Select Data Source dialog.

The System Data Source tab should be the default tab; if not, click it.

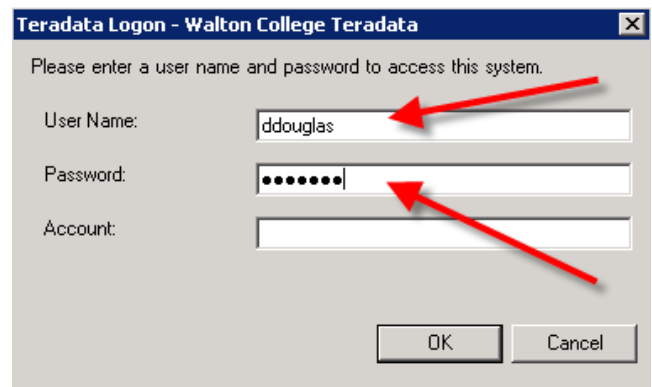
Then locate and select WaltonCollegeTeradata from the list of names—only name In this example.

Click the OK button to open the logon dialog.



Enter your User Name and Password. Click the OK button.

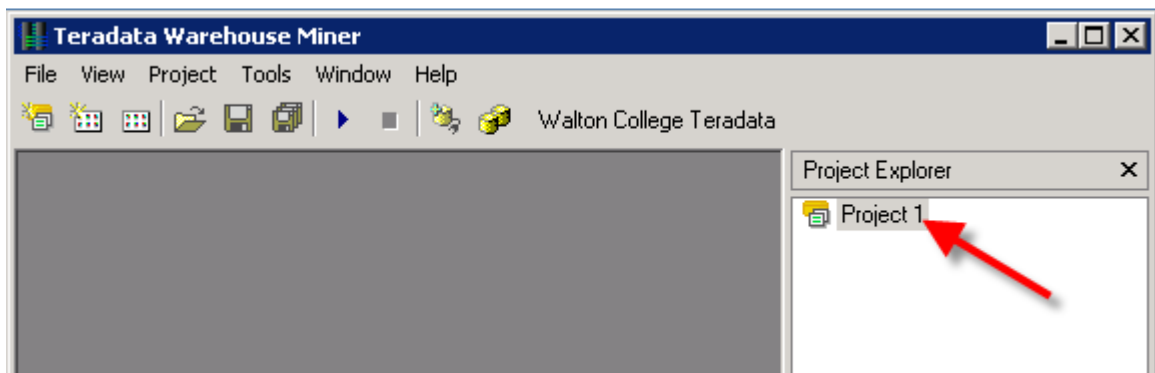
Note that the grayed out icon to the right of the Open Connection icon is now yellow and the word Disconnected has been replaced with the name of the data source to which the system is now connected. Also displayed is the tooltip for the Add New Project icon on the toolbar. The File menu can also be used to create a new project.



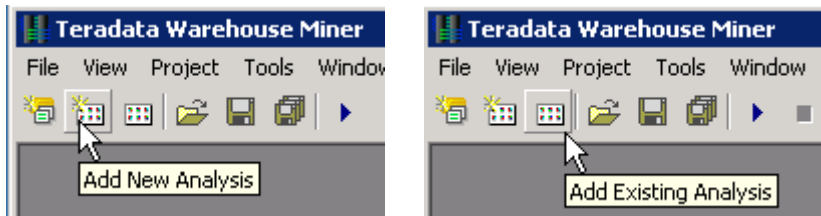


Also, note that there is a Project Explorer window to the right and an Execution Status window on the bottom.

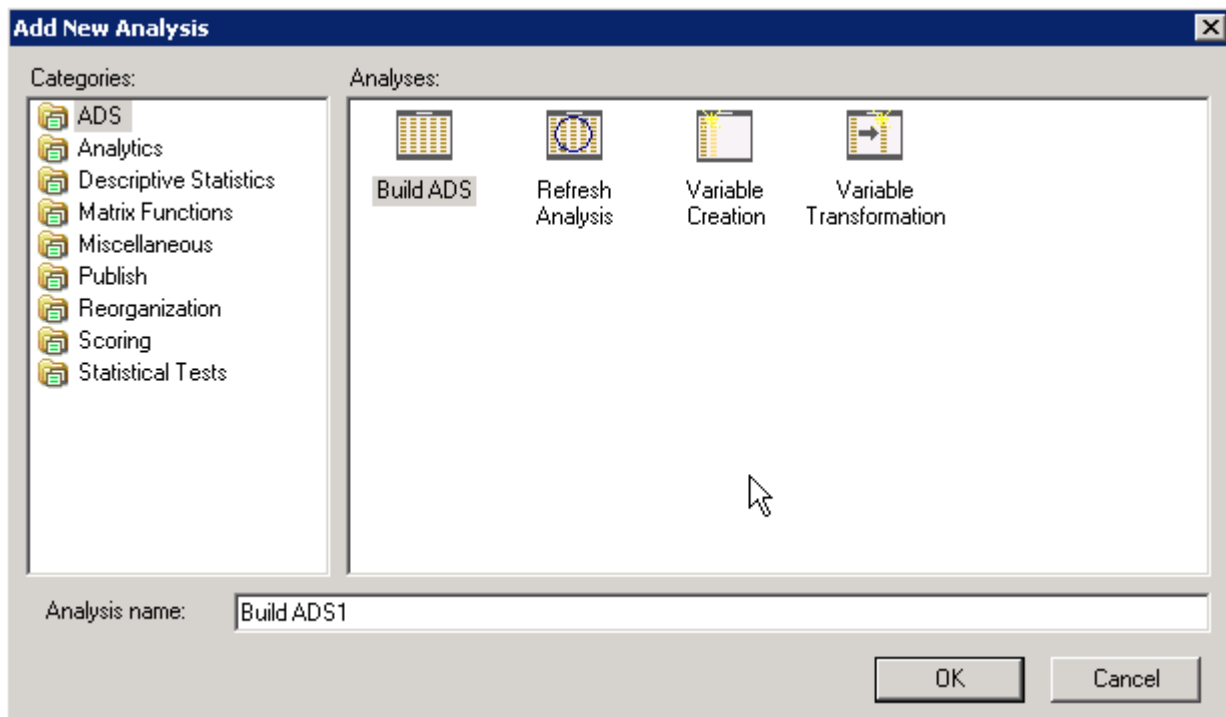
Open a new project by clicking the Add New Project icon—the default name of the new project is Project 1 and is in the Project Explorer window—you can change the project name by clicking it and then changing to a desired name.



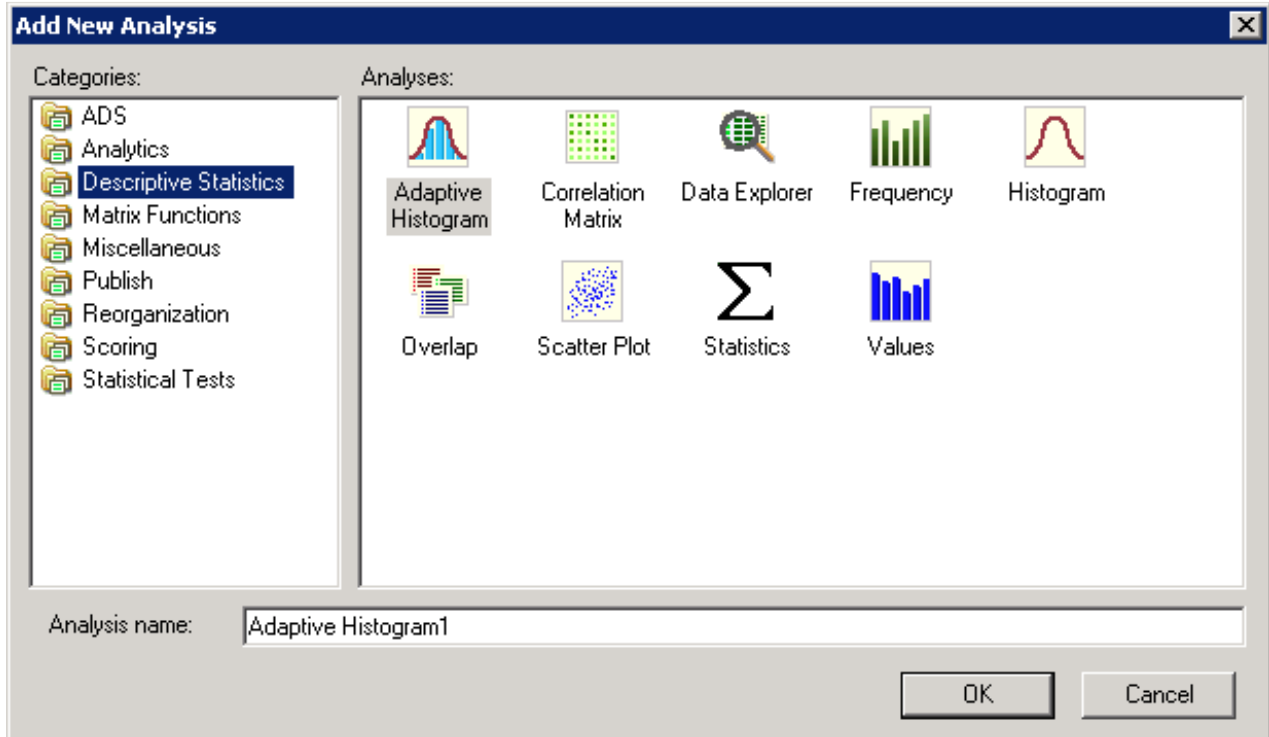
Having a project selected in the Project Explorer window activates additional icons on the toolbar—the standard windows icons for a folder, saving a file and save all. The other two icons are the Add New Analysis and Add Existing Analysis icons which are shown below.



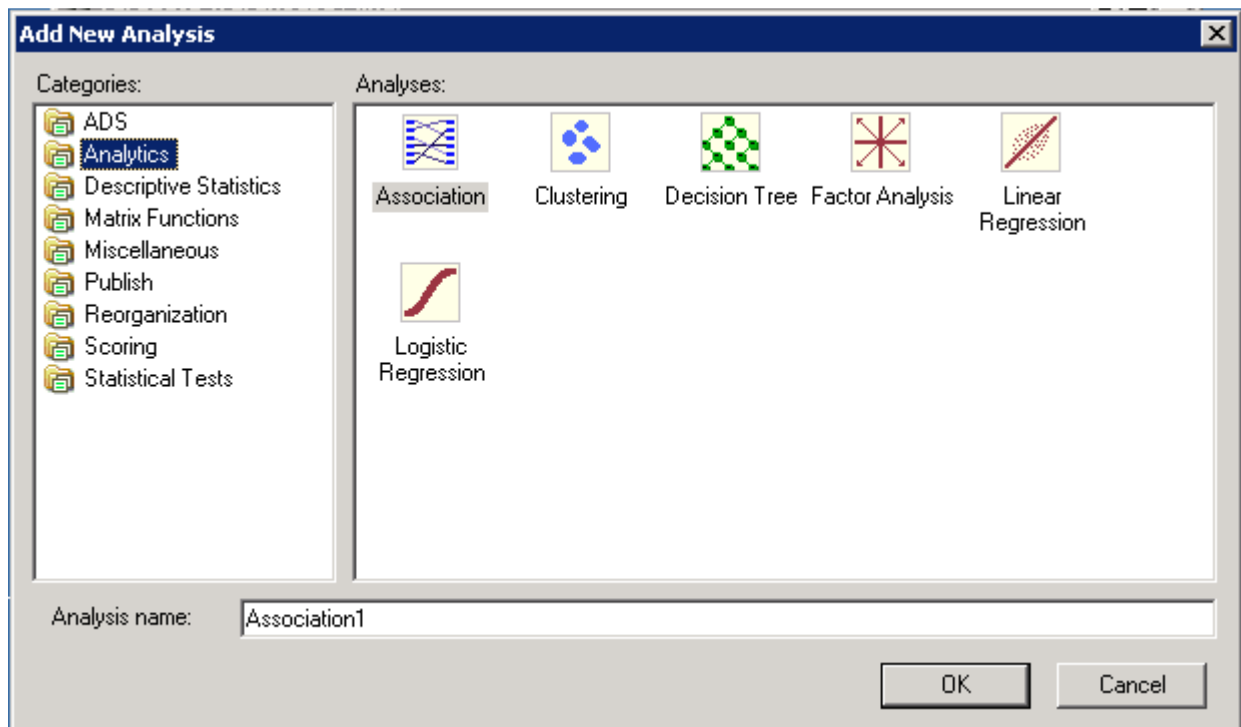
Click the Add New Analysis icon to open its dialog.



The types of analysis are listed in the Categories pane of the Add New Analysis dialog. Note the default **Analysis name** in the textbox near the bottom of the dialog. Because the type of analysis selected is ADS, the default name is BuildADS1—you can build an Analysis Data Set (ADS) if needed. Click the Descriptive Statistics category to view the types of analysis available in this category.



DM is the goal of this illustration so click Analytics in the **Categories** pane.

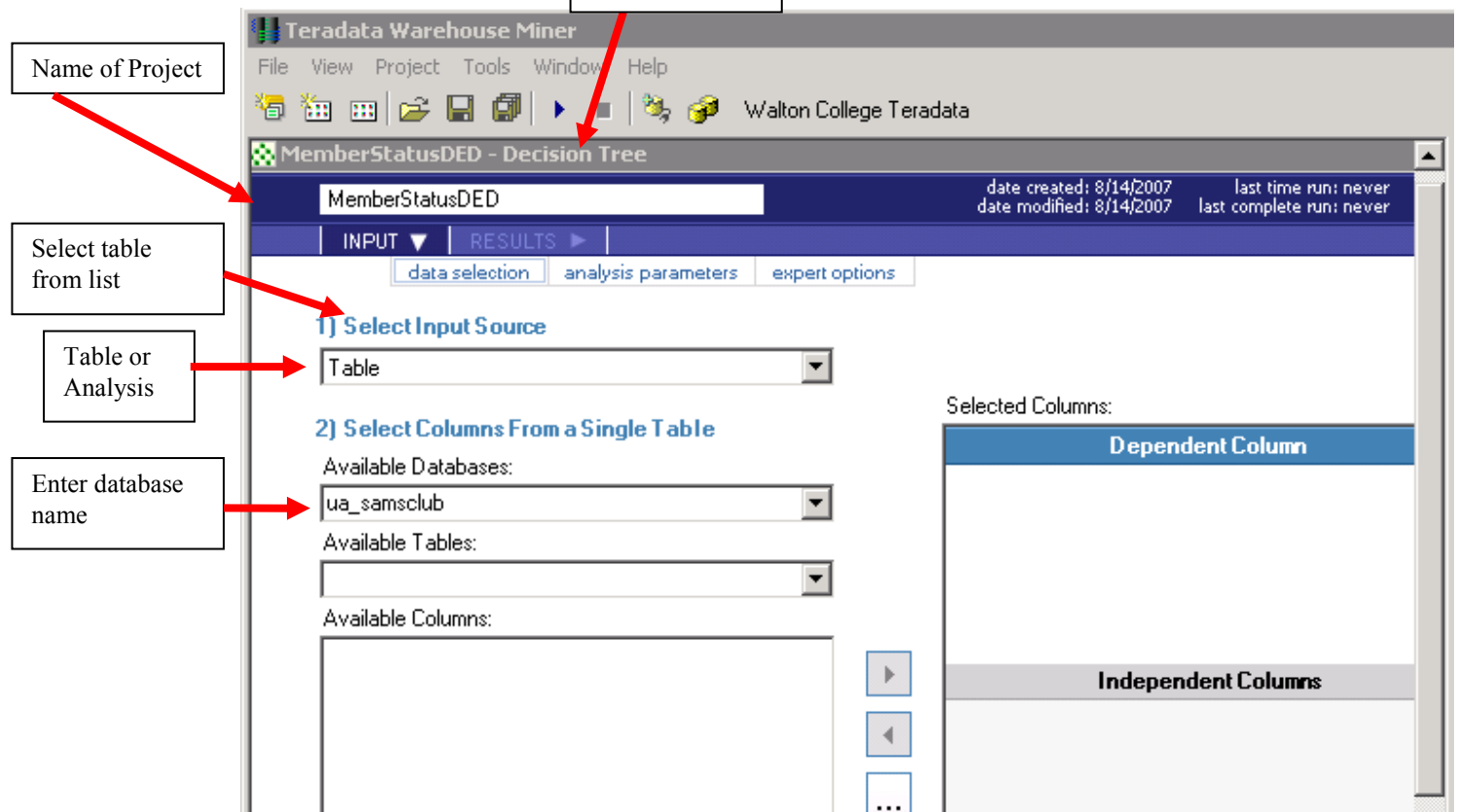


Six data mining techniques are available. Decision Trees, Linear Regression and Logistic Regression are directed or supervised data mining techniques. This means that they have a target variable (dependent variable) and thus are sometimes called predictive models. Association analysis and clustering are undirected or unsupervised data mining techniques and do not have a target or dependent variable.

The data for this example will be from a table (WAREHOUSE) in the ua_samsclub database and consists of more than 4 million rows of data. This table was created by joining desired columns from a number of tables in the ua_samsclub data. The data mining problem is to build a model to predict the member status code based on other variables in the Warehouse table.

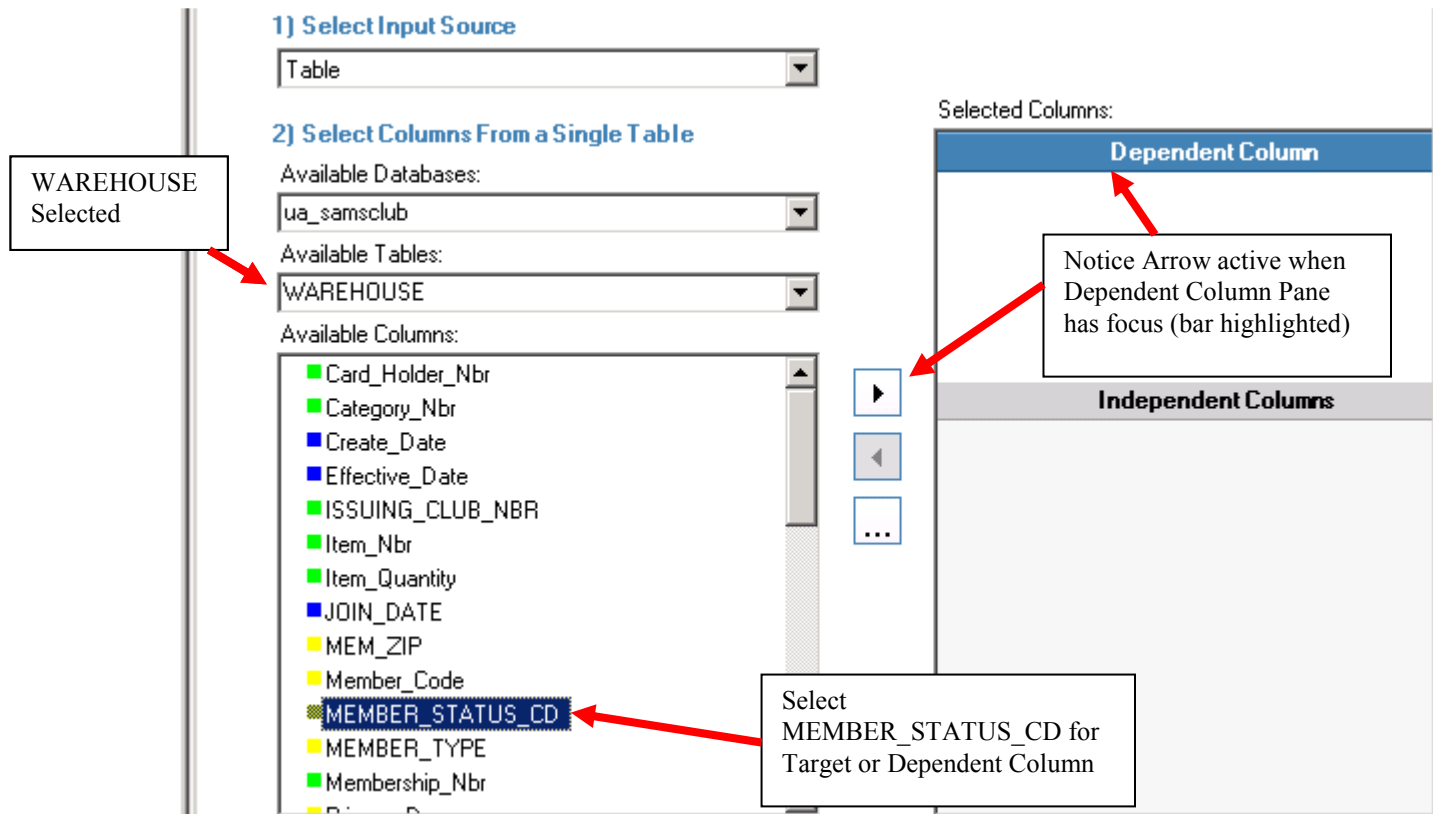
Click the Decision Tree icon in the Analyses pane and provide an Analysis name. Use the naming convention of adding your initials to the end of the Analysis name of **MemberStatusiii** if you wish. For this example, the Analysis name will be MemberStatusDED. Click the OK button.





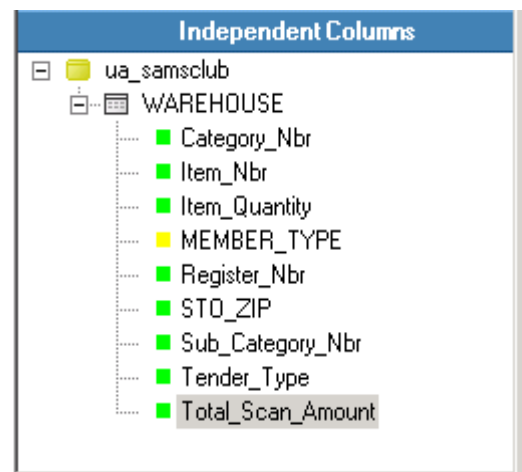
The Input source as **Table** and enter the database in the **Available Databases:** drop down list box and then press the **Enter** key. *It is important to press the Enter key as it populates the Available Tables: drop down list box with all the tables in the database.* In this example, the desired table is WAREHOUSE.

Selecting the table **WAREHOUSE** from the **Available Columns** drop down list box will populate the **Available Columns:** pane with the columns of the table. The **Selected Columns:** pane is context dependent—in this case, it contains an area for the dependent column and an area for the Independent columns because the data mining technique is a decision tree. As shown below, WAREHOUSE has been selected as the table.



After clicking the arrow to move the selected column (MEMBER_STATUS_CD) from the Available Columns pane to the Dependent Column pane, the arrow is grayed out. This is because this data mining model can only have one target or dependent variable.

The user will then need to click the Independent Columns bar so it will have the focus which allows moving selected columns from the **Available Columns** pane to the **Independent Columns** pane. *Note that the transfer arrow will not activate until selection of one or more columns from the Available Columns pane is selected.* Multiple columns can be selected and moved in a single action. Nine columns have been selected Independent Columns as shown. This completes the **data selection** process—click the **analysis parameters** tab.



The goal of decision tree analysis is to initially analyze each independent variable to determine which variable best predicts the target or dependent column (variable). Different algorithms are used to determine the predictive accuracy of variables—the default in this case is the Gain Ratio method. Click the drop down box to see that three other options are available—Gini Index, Chaid, and Regression Trees. For this example, accept the default splitting method of Gain Ratio. This process is recursive and at each node, all variables are again examined to determine which one provides the best predictive power – a variable may be split more than once at different nodes in the decision tree.

MemberStatusDED - Decision Tree

MemberStatusDED date created: 8/28/2007 last time ru
date modified: 8/28/2007 last complete ru

INPUT RESULTS

data selection analysis parameters expert options

Splitting Options

Splitting Method: Gain Ratio Chaid Significance Levels

Minimum Split Count: 2 Merging: 0.05

Maximum Nodes: 1000 Splitting: 0.05

Maximum Depth: 100

Bin Numeric Variables

Include Validation Table

Validation Table: Values...

Include Lift Table

Response Value: Values...

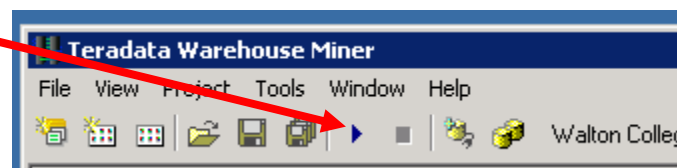
Pruning Options

Pruning Method: Gain Ratio

Gini Test Table:

For this example, change the Minimum Split Count from 2 to 50 and the Maximum Depth from 100 to 50. Accept all the other defaults. Click the expert options just to view possibilities. In this case, the only option is the amount of data to be held in memory for processing. There is no need to change the default value.

Click the run button:



Check the **Execution Status** Window at the bottom. Note that there are more than 4 million rows in the WAREHOUSE table. Also, the status is **Executing**.

Analysis	Status	Message
MemberStatusDED	Executing	Level: 7, Nodes to split: 72, 4090023 of 4109196 observati

When execution is completed, the Executing Status will change to **Complete**. Be patient as the model will take a few minutes to run—note that at each node, all the variables are evaluated as a candidate for splitting and there are more than 4 million rows for the 9 independent variables.

Execution Status		
Analysis	Status	Message
MemberStatusDED	Complete	Execution complete

After the model is complete, the RESULTS tab becomes accessible. Click the RESULTS tab.

MemberStatusDED - Decision Tree

MemberStatusDED

date created: 8/28/2007 last time run: 8
date modified: 8/28/2007 last complete run: 8

INPUT RESULTS

reports graphs

Decision Tree Report

- Variables
- Confusion Matrix

Total Observations: 4109196

Nodes Before Pruning: 117

Nodes After Pruning: 55

Model Accuracy: 92.31%

Note that the model has an accuracy of 92.31% and the default tab is the reports tab. On the left under the Decision Tree Report, you can click variables or the Confusion Matrix.

Click **Variables** to notice only 6 independent variables are listed—recall that the original model included 9 independent variables. Which variables have been dropped?

MemberStatusDED - Decision Tree

MemberStatusDED

INPUT RESULTS

reports graphs

Decision Tree Report

- Variables
- Confusion Matrix

Dependent Variable

MEMBER_STATUS_CD

Independent Variables

Category_Nbr

MEMBER_TYPE

Register_Nbr

STO_ZIP

Tender_Type

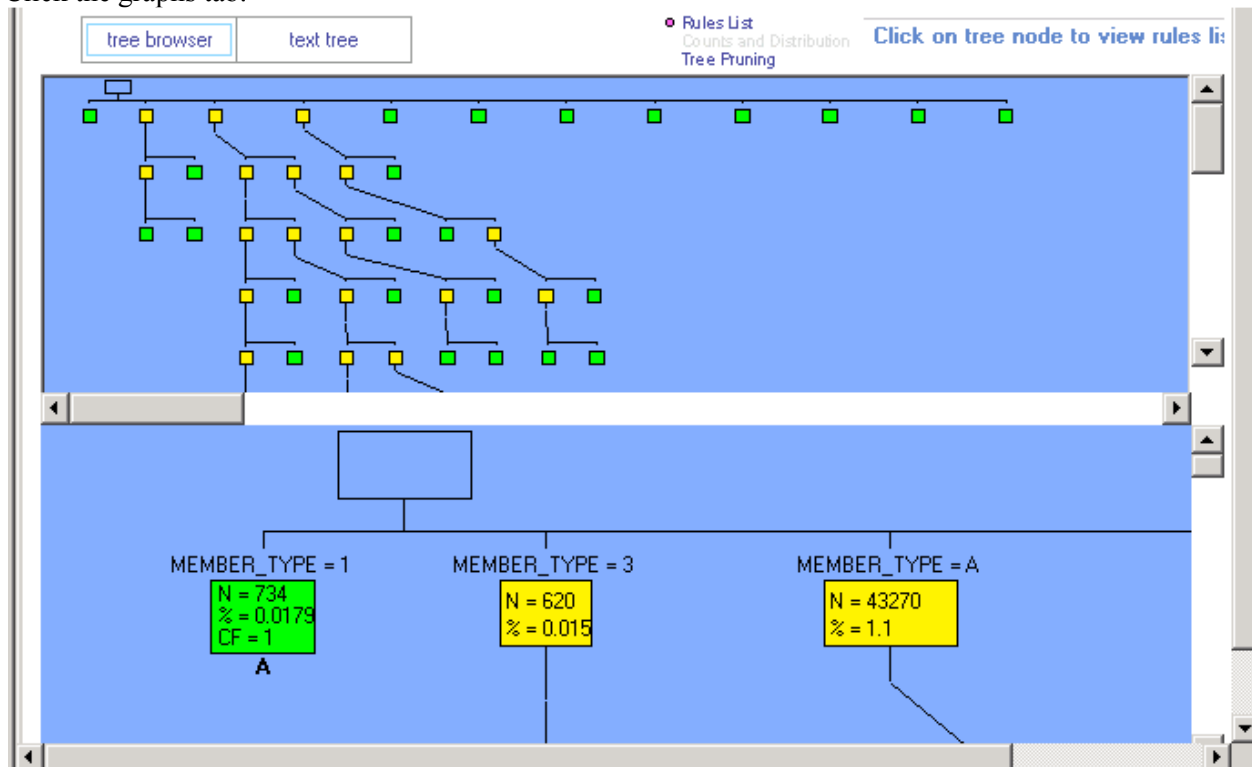
Total_Scan_Amount

Click **Confusion Matrix** and review the results.

- Decision Tree Report
 - Variables
 - Confusion Matrix**

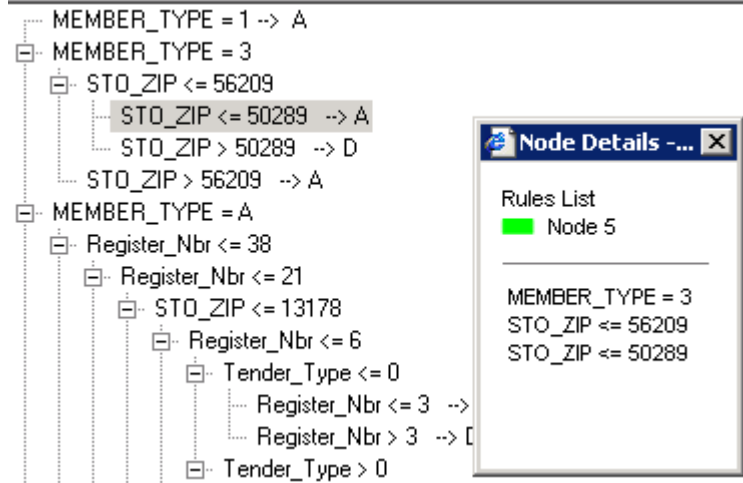
	Actual A	Actual D	Actual E	Actual P	Actual T	Correct	Incorrect
Predicted A	3790900 / 92.25%	197802 / 4.81%	117330 / 2.86%	186 / 0.00%	324 / 0.01%	3790900 / 92.25%	315642 / 7.68%
Predicted D	328 / 0.01%	1810 / 0.04%	32 / 0.00%	0 / 0.00%	0 / 0.00%	1810 / 0.04%	360 / 0.01%
Predicted E	103 / 0.00%	22 / 0.00%	359 / 0.01%	0 / 0.00%	0 / 0.00%	359 / 0.01%	125 / 0.00%
Predicted P	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%
Predicted T	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%	0 / 0.00%

Click the graphs tab.



A high-level visual is provided without any detail in the upper window. Click any box in the upper window to get detail in the lower window or just browse the lower graph. Nodes are by default colored yellow and green (green being a leaf node which means no other nodes are split from it).

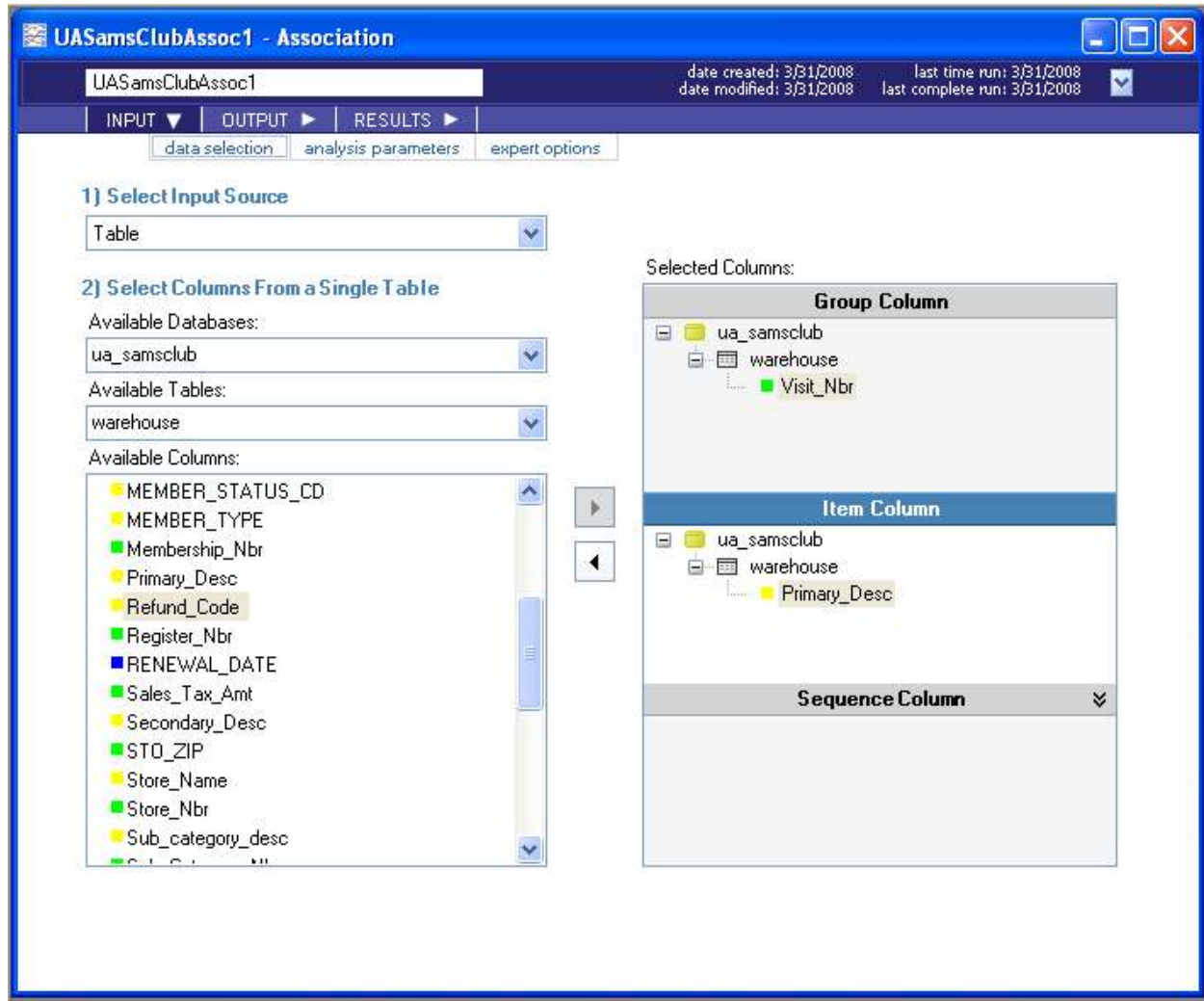
Click the **text tree** tab and then click on one of the rules—a Node Details dialog appears. It is a leaf node—green—and is node 5 if you want to reviews the data from the tree browser.



You can save the model by clicking the multiple disks on the toolbar.

Association Analysis

The following example is for association analysis that uses the **warehouse** table of ua_samsclub. The data format for **Association Analysis** is a transaction format—consisting of two columns; a customer and a product. Thus, if a customer purchases 4 items, then there would be 4 rows with the customer repeated. The equivalent of this for the **warehouse** table is visit_nbr and Primary_desc. See the setup below.



Next, click the **analysis parameters** tab toward the upper left. Note that the default setting is for one antecedent and one consequent.

UASamsClubAssoc1 - Association

UASamsClubAssoc1

date created: 3/31/2008 last time run: 3/31/2008
date modified: 3/31/2008 last complete run: 3/31/2008

INPUT OUTPUT RESULTS

data selection analysis parameters expert options

Association Combinations:

1 TO 1

Add Remove

1 TO 1

Processing Options

Perform All Steps Minimum Support: 0.01
 Perform Support Calculation Only Minimum Confidence: 0
 Recalculate Final Affinities Only Minimum Lift: 0
 Auto-Calculate group count Minimum Z-Score: 0
Force Group Count To:

Drop all support tables after execution.

Sequence Options

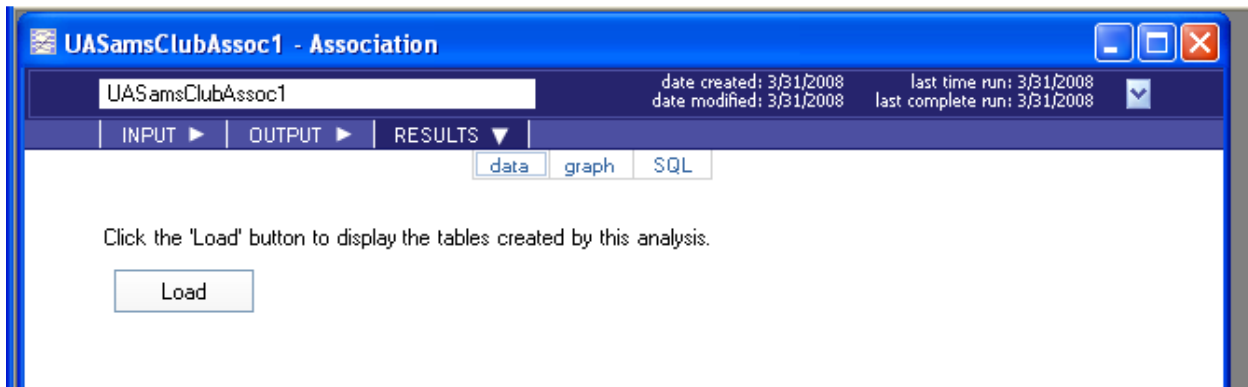
Use relaxed ordering
 Auto-Calculate ordering probability
Ordering Probability:

Click the **expert options** tab. Enter an SQL statement – in this case, only store number 15.

Optional WHERE clause text:

store_nbr = 15

Click the run button. When the run is completed, click results—a load table button should appear.



Click the Load button and review the table.

UASamsClubAssoc1 - Association

UASamsClubAssoc1

date created: 3/31/2008 last time run: 3/31/2008
date modified: 3/31/2008 last complete run: 3/31/2008

INPUT OUTPUT RESULTS

data graph SQL

OneToOne (8 rows) Edit Format Sort Export

ITEM10F2	ITEM20F2	LSUPPORT	RSUPPORT	SUPPORT	CONFIDENCE	LIFT	ZSCORE
▶ 1 % MILK	EXTRA LARG	0.06	0.06	0.01	0.22	3.51	35.87
2% MILK	EXTRA LARG	0.06	0.06	0.02	0.27	4.31	50.01
BOUNTY PAP	CHARMIN 24	0.03	0.04	0.01	0.33	9.38	67.91
CHARMIN 24	BOUNTY PAP	0.04	0.03	0.01	0.31	9.38	67.91
EXTRA LARG	1 % MILK	0.06	0.06	0.01	0.20	3.51	35.87
EXTRA LARG	2% MILK	0.06	0.06	0.02	0.28	4.31	50.01
EXTRA LARG	SKIM MILK	0.06	0.08	0.02	0.26	3.11	36.60
SKIM MILK	EXTRA LARG	0.08	0.06	0.02	0.20	3.11	36.60

Click graph.

