Running SAS® Enterprise Miner™ 13.1 High-Performance Procedures in Alongside Mode Using SASHDAT Data

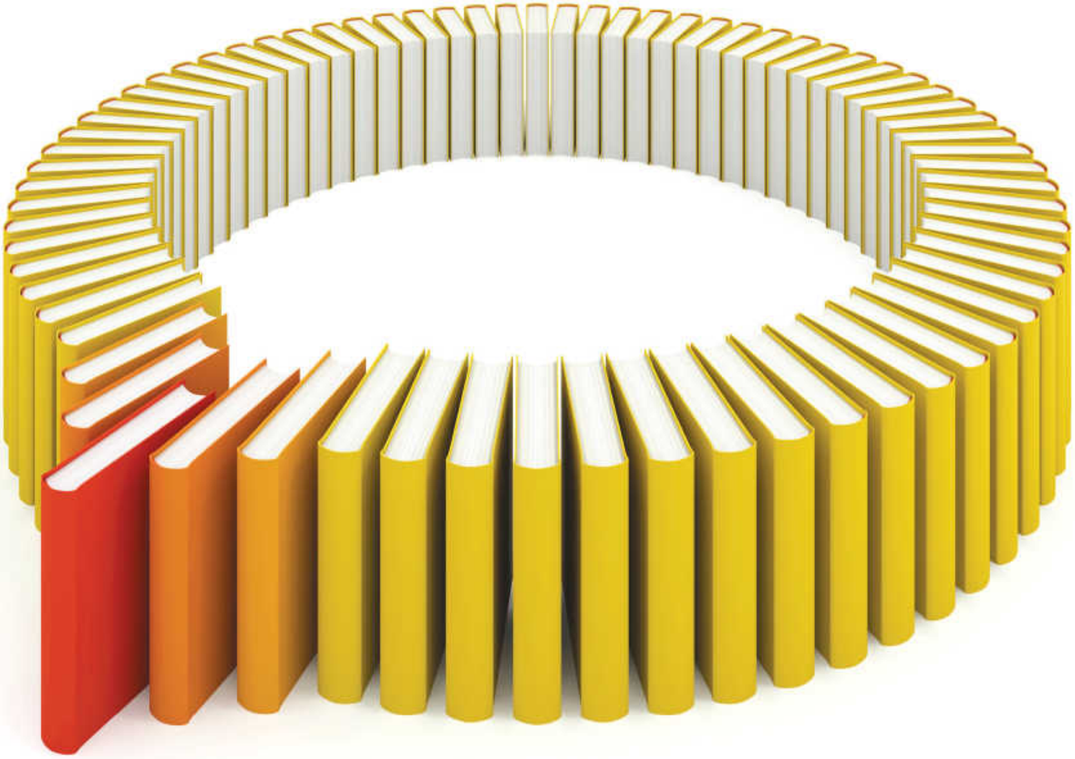
SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *Running SAS® Enterprise Miner™ 13.1 High-Performance Procedures in Alongside Mode Using SASHDAT Data*. Cary, NC: SAS Institute Inc.

Running SAS® Enterprise Miner™ 13.1 High-Performance Procedures in Alongside Mode Using SASHDAT Data

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

Printing 1, December 2013



Contents

1

The Hadoop File System and SASHDAT

Introduction to Hadoop and SASHDAT

Hadoop is a distributed file system supported by the Apache Foundation. It is most commonly used to manage large files in a distributed environment. That is, the file system spans multiple computers. Hadoop uses the MapReduce distribution mechanism. A file is mapped into smaller pieces, distributed to active nodes on the grid, and then results are combined and reduced.

The SASHDAT engine is available as part of SAS High-Performance. SASHDAT can refer to both the name of a specialized file type that is developed for distributing SAS data sets through Hadoop and the LIBNAME engine to access those files. SASHDAT is not the same as the Hadoop LIBNAME engine available through SAS/ACCESS. The LIBNAME statement is the primary way to access SASHDAT data sets.

Differences between Hadoop Data Sets and Teradata and Greenplum Data Sets

There are several differences between the Hadoop file system and both Teradata and Greenplum that necessitate different implementations. The two most important differences are that you cannot create a view of a SASHDAT data set and you cannot merge SASHDAT data sets. You cannot create a view of a SASHDAT data set because SASHDAT libraries cannot be downloaded to the SAS client. As a result, you cannot browse SASHDAT data. Put another way, SASHDAT files are not portable to the client.

Because of these differences, the HP Data Partition node is implemented differently for SASHDAT data sources and Teradata or Greenplum data sources. When used with the HP Data Partition node, Teradata and Greenplum data sources require a key variable. They produce a database table that contains the key variable and a partition indicator variable. This database table is merged with the input data source using the key variable. The key variable must be unique for all observations and this requirement is checked for in the Input Data node. Hadoop data sources do not require a key variable. Instead, Hadoop data sources are copied to the grid with an additional partition indicator variable.

Note: The partition indicator variable _partind_ is a binary variable that indicates whether the observation is in the training data set or the validation data set. Observations with a value of 1 are in the training data set, and observations with a value of 0 are in the validation data set.

Other differences between Hadoop data sources and Teradata or Greenplum data sources include the following:

- The HP Text Miner node is not supported for SASHDAT data sources. An error message is displayed if you try to use the HP Text Miner node with a SASHDAT data source. The HP Text Miner node is experimental for Greenplum and Teradata data source.

- Greenplum and Teradata database tables have a maximum number of columns, but Hadoop database tables have no maximum number of columns. Also, there are different requirements on the length of variable names and reserved names.
- Currently, you cannot run SAS Rapid Predictive Modeler on Hadoop data sources in a grid environment because SAS Rapid Predictive Modeler does not use any of the high performance data mining nodes. You can run SAS Rapid Predictive Modeler with Teradata and Greenplum. However, the process uses the SAS/ACCESS engine to download the data to the SAS Server, and does not use the high performance alongside mode.

Similarities between Hadoop Data Sets and Teradata and Greenplum Data Sets

When comparing Hadoop with Teradata and Greenplum, you find certain similarities are found in all three systems. The most important are as follows:

- All training and scoring data sets that are created on the grid are transient. That is, they are deleted after a node runs. This applies to all nodes except the HP Data Partition node.
- When you delete an HP Data Partition node from a process flow diagram, the data set on the grid that is associated with the HP Data Partition node is deleted.
- The scoring data sets from modeling procedures contain the predicted variables, target variables, and other relevant variables. These data sets should contain a relatively small number of variables.
- The results should be identical in Hadoop, Greenplum, and Teradata. However, the representation of the partition table is different between Hadoop and Teradata or Greenplum.
- When a data source is created, a sample is downloaded to the SAS client. That sample is used by all non-HPDM nodes.