



THE POWER TO KNOW.®

What does SAS® High-Performance Analytics do?

SAS High-Performance Analytics enables organizations to analyze big data at the lowest level of granularity and derive accurate and timely insights in minutes, not hours. This enables you to prepare, explore and model multiple scenarios using complete data, instead of only small subsets of data, and provides the quick response times needed for faster, better-informed decision making.

Why is SAS® High-Performance Analytics important?

SAS High-Performance Analytics helps organizations leverage big data and sophisticated analytics to solve difficult problems and problems that were previously thought to be unsolvable, enabling them to seize new opportunities in a rapid manner. SAS High-Performance Analytics helps you weed out unimportant variables, identify important relationships and use the best modeling techniques to improve accuracy and make targeted, high-impact decisions.

For whom is SAS® High-Performance Analytics designed?

SAS High-Performance Analytics offers analytics professionals a dedicated environment to conduct analytical data preparation, model building and model deployment. It is also suitable for IT as a purpose-built and highly scalable infrastructure for managing and processing the complex analytics needed to meet ever-changing business conditions.



SAS® High-Performance Analytics

Generate highly accurate and timely insights and solve complex problems using larger volumes of data than ever before

SAS High-Performance Analytics solves complex business problems that require sophisticated, high-end analytics and access to large data sources. It removes the limitations of trying to analyze large volumes of data with current modeling tools, and eliminates the restrictions imposed by existing computing infrastructures.

You can expect delivery of results in near-real time (typically in seconds or minutes, instead of hours or days). Analytics professionals will be able to prepare, explore and model multiple scenarios to produce the accurate and timely insights needed for improved, better-informed decision making.

SAS High-Performance Analytics uses specifically configured hardware from database appliance partners (EMC Greenplum or Teradata) to run high-performance-enabled capabilities from the following SAS solutions: Base SAS®, SAS/ETS®, SAS/STAT® and SAS® Enterprise Miner™.

By allowing complex analytical computations to run in a distributed, in-memory environment, SAS High-Performance Analytics enables you to prepare, explore and model multiple scenarios using data volumes never before possible to generate accurate and rapid insights. If you can reduce analytic processing from days or hours to minutes or seconds, you have the ability to ask a lot more what-if questions. Instead of waiting long periods of time, you can adjust models and run them again.

Key Benefits

- **React more quickly and confidently to seize new opportunities, manage risks and make the right choices.** Better, faster and more accurate analytical insights allow organizations to achieve added business value, drive new revenue and increase bottom-line savings before the window of opportunity closes.
- **Don't compromise or limit yourself to using simplified analytical approaches to solve complex problems.** Use sophisticated analytics against *all* of your data (not just subsets or aggregates) to improve accuracy and enable more targeted and focused decisions. Users can employ the best modeling techniques, perform model iterations more frequently and test new ideas to discover more accurate insights.
- **Derive insights at breakthrough speeds for high-value and time-sensitive decision making.** Shrink the time it takes to go from model inception to deployment. SAS High-Performance Analytics delivers blazing fast performance so you can evaluate numerous scenarios and quickly detect, and act on, changing market conditions.
- **Take advantage of a highly scalable and reliable infrastructure that is optimized for large volumes of data and complex computations.** For optimal performance, data is pulled and collocated within the memory of a dedicated database appliance. Requests to run new scenarios or new computations can be handled much faster because data can be pulled quickly into the memory without architecture constraints.

Product Overview

SAS High-Performance Analytics is specifically configured software that requires a database appliance from SAS partners (EMC Greenplum or Tera-data) to solve complex problems using in-memory processing resources. SAS High-Performance Analytics is the only in-memory offering on the market that processes big data and high-end analytics to produce time-sensitive insights at blazingly fast speeds.

SAS High-Performance Analytics is not just about using query, reporting and descriptive statistics techniques within an in-memory environment. For optimal performance, data is pulled and collocated within the memory of the dedicated database appliance for analytic processing. Because the data is stored locally in the database, it can be pulled quickly into memory again for future analysis in a rapid manner.

High-performance analytics environment

The SAS High-Performance Analytics environment includes a client-side SAS session (e.g., Base SAS, SAS/ETS, SAS/STAT or SAS Enterprise Miner), and it leverages the database appliance as a shared-memory computing grid. This enables massively parallel computing of SAS code and analytic procedures. It quickly moves large amounts of data in-memory, and threaded analytical algorithms are processed in-memory on collocated data. SAS High-Performance Analytics software requires a database appliance for optimized processing, and to pool large memory resources on demand, to solve complex problems in near-real time.

High-performance data exploration

SAS High-Performance Analytics provides large-scale data exploration and summarization through a series of parallelized procedures. The data exploration capabilities enable descriptive

statistics to be generated on big data in very short time frames. Information such as mean, min, max, range, measures of spread and centrality can be created along with data for cardinality, summary and levels of variables.

High-performance variable reduction

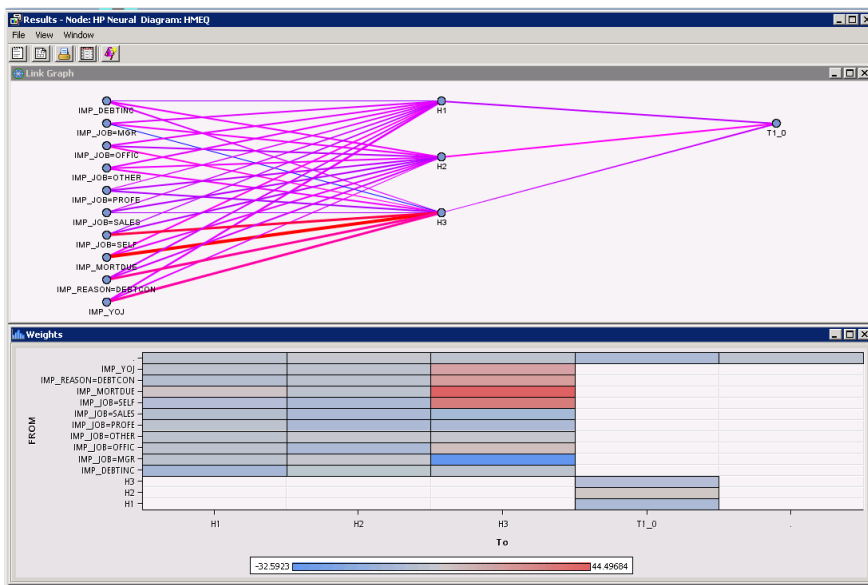
This feature performs unsupervised variable selection by identifying a set of variables that jointly explain the maximum amount of data variance. The HPREDUCE procedure conducts a variance analysis and reduces dimensionality by selecting a subset of the original variables that contribute the most to the overall data variance.

High-performance linear regression

SAS High-Performance Analytics supports ordinary least squares regression so that users can examine the relationship between a response variable and a set of explanatory variables. Regression in SAS High-Performance Analytics includes many different model selection methods and produces a variety of diagnostic measures. Its unique capabilities include effect selection and support of a CLASS statement to allow model-effect selection for data containing inherent classes.

High-performance logistic regression

Logistic regression analysis is a standard for predicting binary, binomial and multinomial outcomes. High-performance logistic regression includes modifications to maximize the algorithms while distributing the calculation over the computing grid. It also fits logistic regression models for binary, binomial and multinomial data. In the high-performance paradigm, model selection techniques complete in seconds or minutes, allowing modelers to experiment with more variables and build better models.



High-performance neural networks enable modelers to produce more training runs for significantly more lift and incremental predictive power.

High-performance nonlinear regression

This feature uses either nonlinear least squares or maximum likelihood to fit nonlinear regression models. It enables you to specify the model with SAS programming statements, which gives you greater flexibility in modeling the relationship between the response variable and independent variables.

High-performance mixed linear models

These features fit a variety of mixed linear models to data and enable you to use these fitted models to make statistical inferences about the data. A mixed linear model is a generalization of the standard linear model used in the GLM procedure – the generalization being that the data is permitted to exhibit correlation and nonconstrained variability. The mixed linear model provides you with the flexibility of modeling not only the means of your data (as in the standard linear model) but the variances and covariances of the data as well.

High-performance data mining

SAS High-Performance Analytics includes high-performance-enabled nodes from SAS Enterprise Miner (for specifics, please refer to the Key Features section). This increases the speed at which you can sample, explore, modify data and model data and assess your models. In addition, a high-performance DATABASE procedure creates summaries for a variety of input data sources.

High-performance neural networks and random forests (experimental)

High-performance neural networks take advantage of the parallel computing environment to enhance the predictive power of the algorithm. Users can produce better models with significantly more lift by enabling more runs to provide incrementally more predictive power. The SAS High-Performance Random Forest procedure creates a

Key Features

High-performance analytics

- Enables high-performance capabilities of select SAS Analytics products (i.e., Base SAS, SAS/ETS, SAS/STAT and SAS Enterprise Miner).
- Uses a platform that supports access to large data stored in nearly any format currently on the market.
- Reads input data in parallel and writes output data in parallel.
- Executes SAS Analytics procedures across a distributed computing environment in parallel.

High-performance data exploration

- Enables large-scale data exploration and summarization through a series of parallelized procedures.
- Generates descriptive statistics on a large scale, very quickly.
- Creates mean, min, max, range and measures of spread and centrality along with data for cardinality, summary and levels of variables.

High-performance variable reduction

- Reduces dimensionality by using the HPREDUCE procedure to select a subset of the original variables (variable selection) to preserve model interpretation.
- Performs unsupervised variable selection by identifying a set of variables that jointly explain the maximum amount of data variance (covariance analysis).
- Provides distributed computation and output of the CORR, COV or SSCP matrix.
- Uses the CLASS statement to support categorical inputs.
- Supports main and interaction effects with the VAR statement.
- Outputs statistics and matrix information for exploratory data analysis that can also be used as direct input for statistical procedures. This saves time by eliminating redundant matrix aggregations.

High-performance linear regression

- Supports generalized linear models and reference parameterization for classification effects.
- Supports partitioning of data into training, validation and testing roles.
- Supports a FREQ statement for grouped analysis and a WEIGHT statement for weighted analysis.
- Provides multiple effect-selection methods.

High-performance logistic regression

- Predicts binary, binomial and multinomial outcomes.
- Provides model-building syntax with the CLASS and effect-based MODEL statements.
- Provides cumulative link models for ordinal data and generalized logistic modeling for unordered multinomial data and enables model building (variable selection).
- Provides WEIGHT and FREQ statements for weighted analysis and grouped analysis, respectively.
- Provides an OUTPUT statement that produces a data set with predicted probabilities and other observation-wise statistics.

High-performance nonlinear regression

- Computes analytical derivatives of user-provided expressions for more robust parameter estimations, improving the estimations as well as making them faster.
- Evaluates user-provided expressions and their confidence limits.
- Estimates parameters by using least squares and the maximum likelihood method.

High-performance mixed linear models

- Fits a variety of mixed linear models to data and enables you to use the fitted models to make statistical inferences about the data.
- Supports multiple covariance structures.
- Provides appropriate standard errors for all specified, estimable linear combinations of fixed and random effects, and corresponding *t*-tests and *F*-tests.
- Provides special dense and sparse algorithms that take advantage of distributed and multiple-core computing environments.

Continued on next page

predictive model called a forest that consists of several decision trees. Hundreds of decision trees can be trained and run in parallel independently on different nodes.

High-performance count regression and severity models

Two procedures from SAS/ETS have been high-performance-enabled. The COUNTREG procedure models integer-valued dependent variables. Severity modeling fits probability distributions for the severity (magnitude) of random events (such as the distribution of losses claimed under insurance policies as well as events with positive impacts).

SAS® High-Performance Analytics System Requirements

To learn more about SAS High-Performance Analytics system requirements, download white papers, view screenshots and see other related material, please visit www.sas.com/hpanalytics.

Key Features (continued)

High-performance data mining

- Includes the following high-performance-enabled SAS Enterprise Miner nodes: HP Data Source, HP Explore, HP Transform, HP Variable Selection, HP Regression, HP Neural Network, HP Random Forest (experimental) and HP Impute.

High-performance data mining DATABASE procedure

- Creates summaries of key input data sources, including:
 - Number of observations.
 - Number of observations that contain a missing value.
 - Minimum observed value.
 - Maximum observed value.
 - Mean of observed values.
 - Standard deviation.
 - Measure of asymmetry.
 - Measure of the “heaviness of the tails.”
 - Sum of all non-missing observations.
 - Corrected sum of squares.
 - Sum of squares.

High-performance neural networks

- Provides automatic standardization of input and target variables.
- Provides automatic selection and use of a validation data subset.
- Provides automatic termination of training when the validation error stops improving.
- Provides the ability to weight individual observations.

High-performance random forest decision trees (experimental)

- Creates an ensemble of hundreds of decision trees to predict a single target.
- Trains hundreds of decision trees in parallel independently on different grid nodes.
- Randomly selects the input variables considered for splitting a node from all available inputs.
- Considers only a single variable that is most associated with the target for splitting.

High-performance count regression

- Fits regression models where the dependent variable represents counts.
- Supports zero-inflated Poisson and negative binomial models.
- Estimates parameters by using the maximum likelihood method.

High-performance severity models

- Fits probability distributions for the severity (magnitude) of random events.
- Fits regression models for the scale of the severity distribution.
- Provides nine different probability distributions, including the Tweedie distribution.
- Allows users to add additional probability distributions.
- Models data truncation and data censoring.