AN INTRODUCTION TO SAS® HIGH-PERFORMANCE ANALYTICS PROCEDURES

ISSA GUINDO
ANALYTICS SPECIALIST
SAS CANADA
**BIG DATA**

**VOLUME, VELOCITY AND VARIETY IN ACTION**

**Volume**
- A Boeing 737 engine generates 10 tb of data per 30 minute in-flight
- The 2 engine 737 will generate 200 tb of data on a 5 hour flight
- Assuming this plane is active for 300 days a year this single machine will generate around 60 pb of data a year
- Per Wikipedia, Scandinavian Airlines has around 90 Boeing 737 in its current fleet
- *Fun fact: Worldwide, on average about 1,000 737s are in the air at all times*

**Velocity**
- The Lotus Formula 1 race car has 240 sensors which registers around 25mb of data per lap
- A race lap is often completed in less than 2 minutes
- The data needs to be collected and analyzed so that any corrective action can be taken to optimize race performance
- Outside of race events, using car sensor data to optimize performance is highly valuable since each lap driven cost around $450,000
- *Fun fact: Kimi Raikkonen (a world champ in 2007) responding to input from his engineer (car’s head mechanic): “Leave me alone, I know how to drive!”*

**Variety**
- Recent news stories documented that the NSA captures a wide variety of the data generated by humans
- Details of phone records and conversations
- Activity on media and sites such as MSN, Twitter, Google and Facebook are stored and analyzed
- This means that the NSA has to deal with massive amounts of both structured and unstructured data
- *Fun fact: In 2009, German Malte Spitz sued his mobile company over access to his mobile records. Combining this with publically available posts on Twitter and Facebook it has been possible to map out and document most of Malte Spitz’ life - http://www.zeit.de/datenschutz/malte-spitz-data-retention/*/
WHAT PROBLEMS WILL DRIVE YOU TO REPLACE DW PLATFORM AND TOOLS?

- Can’t scale to Big Data volumes
- Inadequate data loading speed
- Poor query response
- Current platform modeled for reports & OLAP only
- Can’t score analytic models fast enough

- TDWI Best Practices Report High-Performance Data Warehousing Q4 2012
**PREDICTIVE MODELING CHALLENGES**

**Volume**
- Run-times for models increase which means that fewer model development iterations are possible. This reduces confidence that the best model has been created.
- Choices of variables to use and final model to deploy are often driven by time and technology/storage constraints and not business value.
- Models often run on aggregated data and increasingly smaller samples compared to the full volume of data that is actually available.

**Velocity**
- Due to the speed with which data is generated there is a risk that modelling will happen on “out of date” data.
- This means that there is a chance that models might already be degraded before deployment.
- In order to avoid this, model validation becomes an even more critical part of the lifecycle.
- In some instances it might be necessary to perform model retraining before deployment to pick up latest changes in data.

**Variety**
- Including additional data from both structured and unstructured data sources causes the issues mentioned under “Volume.”
- The amount of different new data sources that can add value if included in the analytical models grow but they are often handled as a silo away from structured data. Including them is a slow cumbersome task.
SAS® HIGH-PERFORMANCE ANALYTICS PRODUCTS

All of your data

Model extensively, iteratively, frequently

Better decisions all the time

Management Administration Security

Teradata Greenplum Oracle Hadoop
DOMAIN SPECIFIC HIGH-PERFORMANCE ANALYTICS OFFERINGS

- SAS® High-Performance Statistics
- SAS® High-Performance Data Mining
- SAS® High-Performance Text Mining
- SAS® High-Performance Optimization
- SAS® High-Performance Econometrics
- SAS® High-Performance Forecasting
Text Mining
- Parsing large-scale text collections
- Extract entities
- Auto. stemming & synonym detection

Data Mining
- Complex relationships
- Tree-based Classification
- Variable Selections

Statistics
- Binary target & continuous no. predictions
- Linear, Non-Linear, & Mixed Linear modeling

Forecasting
- Large-scale, multiple hierarchy problems

Econometrics
- Probability of events
- Severity of random events

Optimization
- Local search optimization
- Large-scale linear & mixed integer problems
## SAS HIGH-PERFORMANCE PROCEDURES (NEW IN SAS® 9.4M2*)

<table>
<thead>
<tr>
<th>High Performance Statistics</th>
<th>High Performance Econometrics</th>
<th>High Performance Optimization</th>
<th>High Performance Data Mining</th>
<th>High Performance Text Mining</th>
<th>High Performance Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCANDISC*</td>
<td>HPCDM*</td>
<td>OPTLSO</td>
<td>HP4SCORE</td>
<td>HPTMINE</td>
<td>HPFORECAST</td>
</tr>
<tr>
<td>HPFMM*</td>
<td>HPGENSELECT</td>
<td>Select features in OPTGRAPH*</td>
<td>HPBNET*</td>
<td>HPTMSCORE</td>
<td></td>
</tr>
<tr>
<td>HPLMIXED</td>
<td>HPCOPULA*</td>
<td>OPTLP</td>
<td>HPCLUS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLOGISTIC</td>
<td>HPCOUNTREG</td>
<td>OPTMILP</td>
<td>HPDECIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPNLMOD</td>
<td>HPPLS*</td>
<td>OPTMODEL</td>
<td>HPFOREST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPLS*</td>
<td>HPPRINCOMP*</td>
<td>OPTMODEL</td>
<td>HPREDUCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSPLIT</td>
<td>HPQUANTSELECT*</td>
<td>OPTMODEL</td>
<td>HPSVM*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HPRINCOMP*</td>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HPQUANTSELECT*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HPREDUCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HPSVM*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common Set: HPDS2, HPDMDB,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### High Performance Statistics

<table>
<thead>
<tr>
<th>High Performance Statistics</th>
<th>High Performance Econometrics</th>
<th>High Performance Optimization</th>
<th>High Performance Data Mining</th>
<th>High Performance Text Mining</th>
<th>High Performance Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCANDISC*</td>
<td>HPCDM*</td>
<td>OPTLSO</td>
<td>HP4SCORE</td>
<td>HPTMINE</td>
<td>HPFORECAST</td>
</tr>
<tr>
<td>HPFMM*</td>
<td>HPGENSELECT</td>
<td>Select features in OPTGRAPH*</td>
<td>HPBNET*</td>
<td>HPTMSCORE</td>
<td></td>
</tr>
<tr>
<td>HPGENSELECT</td>
<td>HPLMIXED</td>
<td>OPTLP</td>
<td>HPCLUS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLOGISTIC</td>
<td>HPLOGISTIC</td>
<td>OPTMILP</td>
<td>HPDECIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPNLMOD</td>
<td>HPNLMOD</td>
<td>OPTMODEL</td>
<td>HPMODEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPLS*</td>
<td>HPPANEL*</td>
<td></td>
<td>HPSVM*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPRINCOMP*</td>
<td>HPQANTSELECT*</td>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPQUANTSELECT*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPREG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSPLIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Common Set:** HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR
HIGHLIGHTS

• The new HP procedures were designed to address prediction and variable selection on Big Data
• Some HP procedures bring together functionality from existing SAS/STAT procedures for greater modeler productivity (consolidate functionality and simplify syntax)
  • For e.g. HPREG includes functionality from multiple procedures including REG, GLM, and GLMSELECT
• New functionality
• Migration from your existing SAS environment to a distributed environment is seamless
**Current Process**

One algorithm (Logistic Regression)
14 million observations, 46 variables

1 model with default properties
Took 6 hours to process model

Model with Forward Selection
(sle=1, max effects=25)
167 Hours to process model

**High-Performance Process**

One algorithm (HP Logistic Regression)
14 million observations, 46 variables

Took 37 seconds to process model

Took 70 seconds to process model
**SAS® 9.4M2**

**SAS HIGH-PERFORMANCE PROCEDURES (NEW IN SAS® 9.4M2*)**

<table>
<thead>
<tr>
<th>High Performance Statistics</th>
<th>High Performance Econometrics</th>
<th>High Performance Optimization</th>
<th>High Performance Data Mining</th>
<th>High Performance Text Mining</th>
<th>High Performance Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCANDISC*</td>
<td>HPCDM*</td>
<td>OPTLSO</td>
<td>HP4SCORE</td>
<td>HPTMINE</td>
<td>HPFORECAST</td>
</tr>
<tr>
<td>HPFMM*</td>
<td>HPCPM*</td>
<td>Select features in</td>
<td>HPBNET*</td>
<td>HPTMSCORE</td>
<td></td>
</tr>
<tr>
<td>HPGENSELECT</td>
<td>HPFMM*</td>
<td>OPTGRAPH*</td>
<td>HPCLUS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLMIXED</td>
<td>HPGENSELECT</td>
<td>OPTLP</td>
<td>HPDECIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLOGISTIC</td>
<td>HPGENSELECT</td>
<td>OPTMILP</td>
<td>HPFOREST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPNLMOD</td>
<td>HPGENSELECT</td>
<td>OPTMODEL</td>
<td>HPNEURAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPLS*</td>
<td>HPGENSELECT</td>
<td></td>
<td>HPSVM*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPRINCOMP*</td>
<td>HPGENSELECT</td>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPQUANTSELECT*</td>
<td>HPGENSELECT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPRIDGE</td>
<td>HPGENSELECT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSPLIT</td>
<td>HPGENSELECT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Common Set: HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR
HIGHLIGHTS

- Process flow for repeatability and collaboration
- Integrated model comparison and scoring
- Distributed, in-memory processing
**FINANCIAL SERVICES**

**PREDICT CUSTOMER RESPONSE TO RETAIL MARKETING**

### Current Process
- Neural Network Method (1 iteration)
- 5 hours to process model
- Model lift of 1.6%
- Limited to 1 or 2 modeling methods

### High-Performance Process
- Neural Network Method (100 iterations)
- 6 minutes to process model
- Model lift of 3.2%
- Experiment with multiple modeling methods
<table>
<thead>
<tr>
<th>High Performance Statistics</th>
<th>High Performance Econometrics</th>
<th>High Performance Optimization</th>
<th>High Performance Data Mining</th>
<th>High Performance Text Mining</th>
<th>High Performance Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCANDISC*</td>
<td>HPCDM*</td>
<td>OPTLSO</td>
<td>HP4SCORE</td>
<td>HPTMINE</td>
<td>HPFORECAST</td>
</tr>
<tr>
<td>HPFMM*</td>
<td>HPCOPULA*</td>
<td>Select features in OPTGRAPH*</td>
<td>HPBNET*</td>
<td>HPTMSCORE</td>
<td></td>
</tr>
<tr>
<td>HPGENSELECT</td>
<td>HPCOUNTREG</td>
<td>OPTLP</td>
<td>HPCLUS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLMIXED</td>
<td>HPPLS*</td>
<td>OPTMILP</td>
<td>HPDECIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLOGISTIC</td>
<td>HPRLS*</td>
<td>OPTMODEL</td>
<td>HPFOREST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPNLMOD</td>
<td>HPSSP*</td>
<td></td>
<td>HPNEURAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPLS*</td>
<td>HPRINCOMP*</td>
<td></td>
<td>HPREDUCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPQUANTSELECT*</td>
<td>HPSEVERITY</td>
<td></td>
<td>HPSVM*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPREG</td>
<td></td>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSSPLIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Common Set: HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCORR
SAS® HIGH PERFORMACE TEXT MINING

HIGHLIGHTS

• Access from the drag-and-drop interface within SAS® High Performance Enterprise Miner™

• Provides full-spectrum support for text mining
  • Document parsing
  • Term weighting and filtering
  • Term-by-document matrix creation
  • Dimensionality reduction
  • Scoring

• All in a distributed in-memory processing environment
UNITED HEALTHCARE GROUP

BUSINESS ISSUE
• Electronic medical records (EMRs) driving a data explosion
• Utilize all of the unstructured text (records, case notes, emails, transcripts, etc.)
• How to improve quality and cost of care? “Create Healthier Lives”

SOLUTION
• SAS® High-Performance Text Mining
• Greenplum Data Computing Appliance

RESULTS
• Reduce model processing time from four hours to 10 seconds.
• Reduce misclassification rates from 30% to 10%
• Historical models improved with more than 10% lift
• Can now tell that a prescription will harm a patient before you write it…
• Can tell that a customer is dissatisfied before you lose her or him…
• Can now determine that a claim is fraudulent before you pay it…

“SAS is helping make our member services the best in the industry. In less than one hour, we can load a huge table (169 million row dataset), find the best variables, compare different models and pick the best model. I would not attempt to model a dataset this large without SAS HPA.”

Mark Pitts
Former Director of Data Science, Solutions and Strategy
### SAS HIGH-PERFORMANCE PROCEDURES (NEW IN SAS® 9.4M2*)

<table>
<thead>
<tr>
<th>High Performance Statistics</th>
<th>High Performance Econometrics</th>
<th>High Performance Optimization</th>
<th>High Performance Data Mining</th>
<th>High Performance Text Mining</th>
<th>High Performance Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCANDISC*</td>
<td>HPCDM*</td>
<td>OPTLSO</td>
<td>HP4SCORE</td>
<td>HPTMINE</td>
<td>HPFORECAST</td>
</tr>
<tr>
<td>HPFMM*</td>
<td>HPCOPULA*</td>
<td>Select features in OPTGRAPH*</td>
<td>HPBNET*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPGENSELECT</td>
<td>HPCOUNTREG</td>
<td>OPTLP</td>
<td>HPCLUS*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLMIXED</td>
<td>HPQLIM</td>
<td>OPTMILP</td>
<td>HPDECIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPLOGISTIC</td>
<td>HPSEVERITY</td>
<td>OPTMODEL</td>
<td>HPFOREST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPNLMOD</td>
<td></td>
<td></td>
<td>HPNEURAL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPLS*</td>
<td></td>
<td></td>
<td>HPREDUCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPRINCOMP*</td>
<td></td>
<td></td>
<td>HPSVM*</td>
<td>HPTMSCORE</td>
<td></td>
</tr>
<tr>
<td>HPQUANTSELECT*</td>
<td></td>
<td></td>
<td>HPTSDR*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPREG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSPLIT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Common Set:</strong></td>
<td><strong>HPDS2, HPDMDB, HPSAMPLE, HPSUMMARY, HPIMPUTE, HPBIN, HPCorr</strong></td>
<td><strong>HPCORR</strong></td>
<td><strong>HPSVM</strong></td>
<td><strong>HPTSDR</strong></td>
<td><strong>HPFORECAST</strong></td>
</tr>
</tbody>
</table>

*Note: HP indicates High Performance*
SAS High-Performance Econometrics allows you to:

- Investigate new opportunities, detect unknown risks, and make the right choices.
- Test more ideas and multiple scenarios with all your data.
- Perform more model iterations.
- Generate insights at breakthrough speeds for high-value and time-sensitive decision making.
High-performance count regression

- Estimates regression models where the dependent variable represents counts
- Supports Poisson and negative binomial models.
- Supports zero-inflated Poisson and negative binomial models, and can fit separate regressors for the zero-inflated distribution.
- Estimates parameters by using the maximum likelihood method.
High-performance loss distribution/severity models

- Estimates probability distributions for the severity (magnitude) of random events
- Estimates regression models for the scale of the severity distribution.
- Provides nine different probability distributions, including the Tweedie distribution, and can automatically select the best-fitting distribution.
- Allows users to add additional probability distributions.
- Can model data truncation and data censoring
High-performance qualitative and limited independent variable models

- Estimates linear, censored and truncated regression models with heteroscedasticity.
- Estimates stochastic frontier production and cost models.
- Contains options for Bayesian estimation.
VALUE DRIVERS FOR INVESTING IN SAS HIGH-PERFORMANCE ANALYTICS TECHNOLOGY

Efficiency
• Build, test and validate more models faster resulting in reduced cycle time to identify the best setup
• Automate a greater deal of the total modeling flow in order to free up analytical resources to do more value added work

Insight
• Using more data than previously possible, organizations can now uncover unknown relationships and patterns to gain better insight
• New data becomes available since sampling is no longer needed and you can use all the data combined with the fact that you are able to leverage data from multiple sources including both structured and unstructured information

Innovation
• Having improved efficiency and thereby freed up valuable resources in combination with new valuable data sources enable organizations to conduct more experimentation
• This may lead to the identification of new segments, potential new needs and new service offerings