SAS High-Performance Analytics
From Desktop to Massively Parallel System

Oliver Schabenberger
Lead Developer and Architect
High Performance Analytics
SAS High Performance Computing

- The intersection of
  - High Performance Analytics (HPA)
    - algorithms
    - hardware
    - compute parallelization
  - High Performance Data (HPD)
    - data distribution
    - storage; hardware
    - data parallelization

- HPC = HPA + HPD
- HPC = Big Analytics + Big Data
SAS High Performance Computing

- Worrying about software performance is not a new concept at SAS

- What is New?
  - Dedicated high-performance software
  - Accelerated development

- Why Now?
  - Customer needs
  - Blade systems have proven viable platforms for high-performance computing
  - New computing paradigms
  - Partnerships with MPP database vendors
SAS High-Performance Analytics

What Is It?

- New product available in Q4 2011
  - EA program starts earlier
- High-end, high-performance analytics
  - Tools → PROCs
  - Data management strategies

Motivation: You

- Experience performance issues with execution in the SAS language
- Have dedicated analytic processes (model building, scoring)
- Asked for a high-performance programming environment
- Want to work within familiar framework—SAS 4GL
SAS High-Performance Analytics

What Is It?

- A collection of SAS procedures for
  - Descriptive statistics and summarization
  - Descriptive modeling
  - Predictive modeling
  - Optimization

- Extends SAS software
  - SAS In-database
  - SAS Grid Manager

- Provides programming environment
# Analytical Tiers and HPA Procedures

<table>
<thead>
<tr>
<th>Tier</th>
<th>Examples</th>
<th>Class</th>
<th>SAS Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindsight</td>
<td>Descriptive statistics, summarization</td>
<td>Class</td>
<td><strong>HPSUMMARY, MEANS, RANK, UNIVARIATE</strong></td>
</tr>
<tr>
<td></td>
<td>Cross-tabulation</td>
<td></td>
<td><strong>FREQ</strong></td>
</tr>
<tr>
<td></td>
<td>Reporting</td>
<td></td>
<td><strong>REPORT, TABULATE</strong></td>
</tr>
</tbody>
</table>
## Analytical Tiers and HPA Procedures

<table>
<thead>
<tr>
<th>Tier</th>
<th>Examples</th>
<th>Class</th>
<th>SAS Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hindsight</strong></td>
<td>Descriptive statistics, summarization</td>
<td></td>
<td><strong>HPSUMMARY, MEANS, RANK, UNIVARIATE</strong></td>
</tr>
<tr>
<td></td>
<td>Cross-tabulation</td>
<td></td>
<td><strong>FREQ</strong></td>
</tr>
<tr>
<td></td>
<td>Reporting</td>
<td></td>
<td><strong>REPORT, TABULATE</strong></td>
</tr>
<tr>
<td><strong>Insight—descriptive modeling</strong></td>
<td>Correlation analysis</td>
<td>Relationships among variables</td>
<td><strong>REG, CORR, VARCLUS FACTOR PRINCOMP HPREG, HPREDUCE</strong></td>
</tr>
<tr>
<td></td>
<td>Variable clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Factor analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Principal component analysis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Analytical Tiers and HPA Procedures

<table>
<thead>
<tr>
<th>Tier</th>
<th>Examples</th>
<th>Class</th>
<th>SAS Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hindsight</strong></td>
<td>Descriptive statistics, summarization</td>
<td></td>
<td>HPSUMMARY, MEANS, RANK, UNIVARIATE</td>
</tr>
<tr>
<td></td>
<td>Cross-tabulation</td>
<td></td>
<td>FREQ</td>
</tr>
<tr>
<td></td>
<td>Reporting</td>
<td></td>
<td>REPORT, TABULATE</td>
</tr>
</tbody>
</table>
| **Insight—descriptive modeling** | Correlation analysis  
Variable clustering  
Factor analysis  
Principal component analysis | Relationships among variables | REG, CORR, VARCLUS  
FACTOR  
PRINCOMP  
HPREG, HPREDUCE |
| **Foresight—predictive modeling** | Linear models  
Generalized linear models | Linear elements | HPREG, HPLOGISTIC |
|                     | Nonlinear least-squares and maximum likelihood | Nonlinear elements | HPNLIN |
|                     | Neural networks                                                        |                                 | HPNEURAL                        |
|                     | Linear mixed models                                                    | Random effects                  | HPLMIXED                        |
|                     | Decision methods                                                       |                                 | HPFOREST                        |
| **Optimization**    | Optimization                                                            |                                 | TBD                             |
SAS High-Performance Analytics

- HPREG: linear regression and variable selection
- HPLOGISTIC: logistic regression and variable selection
- HPLMIXED: linear mixed models
- HPNEURAL: neural nets
- HPNLIN: nonlinear regression and maximum likelihood
- HPPREDUCE: covariance/correlation analysis, variable reduction
- HPDMDB: summarization
- HPSUMMARY: descriptive statistics
- HPFOREST: predictive modeling based on decision trees
- HPDS2: next-generation data step
SAS Procedures

Then and Now

Single-threaded
Not aware of distributed computing environment
SAS/ACCESS for data read
Runs on client
Brings distributed data to client
Large I/O

Multi-threaded
Aware of distributed computing environment
SAS/ACCESS for parsing support
Runs on client or DBMS appliance
Runs alongside distributed data source
In-Memory Analytics

proc logistic data=TD.mydata;
  class A B C;
  model y(event='1') = A B B*C;
run;

proc hplogistic data=TD.mydata;
  class A B C;
  model y(event='1') = A B B*C;
run;
What to Look For

- LOBs that use statistical modeling with
  - Millions of rows
  - Hundreds to thousands of variables
  - Variable selection

- Long-running analysis steps
  - Take hours or days
  - High value of reducing run-time to seconds or minutes
  - Initial focus is on large data, not many small By groups
Platform

- EMC Greenplum and Teradata analytic appliances
- Provides
  - MPP database
  - MPP computing environment
- Client-side operation from standard SAS session
SAS/HPA Alongside-Greenplum

proc hplogistic data=GPLib.MyTable;
class A B C D ;
model y = a b c b*d x1-x100;
output out=GPLib.logout pred=p;
run;

DCA passes data to SAS/HPA

SQL query

Instructions to HPS head node

红点 = SAS High Performance Analytics
SAS/HPA Alongside-Greenplum

```sas
proc hplogistic data=GPLib.MyTable;
    class A B C D ;
    model y = a b c b*d x1-x100;
    output out=GPLib.logout pred=p;
run;
```

SAS/HPA writes scores locally to DCA
SAS/HPA Procedures

- Operate in SMP and/or MPP mode
- Can work with any data format available to the SAS session
- Recognize an alongside-the-database environment
  - Minimize data movement
  - Can read and write data in distributed form
- ODS tables are brought to client
- User can affect
  - Distribution mode for analytics and data
  - Degree of multi-threading
SAS/HPA Procedure Modes

proc hpreg data=one;
  class a b c;
  model y = a b c x1|x2|x3|x4|x5@2;
run;

proc hpreg data=one;
  class a b c;
  model y = a b c x1|x2|x3|x4|x5@2;
  performance nodes=10 host="cda.lob.com";
run;

libname gplib greenplm server=cda.lob.com database=customer user=oliver;
proc hplogistic data=gplib.SomeTable;
  class a b c;
  model y = a b c x1|x2|x3|x4|x5@2;
  performance host="cda.lob.com";
  output out=gplib.logout pred=p;
run;

Analysis on client box
SMP mode (=multi-threaded)

Analysis on Appliance
Using 10 nodes and multi-threading on each node
Data is "farmed" on 10 nodes

Analysis on Appliance
Alongside Greenplum
Distributed read of data
Using all nodes of Greenplum DCA
SAS/HPA Procedure Highlights

- **PROC HPREDUCE**
  - Correlation analysis
  - Covariance analysis
  - Variable reduction

- To find associations among many variables
- To reduce a large number of variables quickly
  - From 10,000 to 1,000
  - Then feed reduced set to next modeling steps
SAS/HPA Procedure Highlights

- **PROC HPREG**
  - High-performance combination of REG and GLMSELECT
  - Supports
    - classical variable selection techniques
    - modern variable selection techniques (LAR, LASSO)
  - CLASS variables
  - GLM and reference parameterizations
  - SELECTION statement
SAS/HPA Procedure Highlights

- **PROC HPNLLIN**
  - High-performance combination of NLIN and NLP/NLMIXED
  - Supports
    - Classical nonlinear least squares (Levenberg-Marquardt)
    - Maximum likelihood for built-in distributions
    - Maximum likelihood for general, user-specified obj. functions
    - Boundaries, linear equality/inequality constraints
  - ESTIMATE statement for arbitrary linear/non-linear functions of parameters
  - PREDICT statement for predicting arbitrary data-dependent functions
SAS/HPA Procedure Highlights

- **PROC HPLMIXED**
  - High-performance version of PROC MIXED
  - Not to be confused with HPMIXED procedure in SAS/STAT
  - Supports
    - `RANDOM` statements
    - `REPEATED` statement
    - Covariance structures from PROC MIXED
  - Sparse MMEQs with > 40,000 unknowns
    - Impossible in MIXED
    - 12 hours in HPMIXED
    - 3 minutes in HPLMIXED
SAS/HPA Procedure Highlights

- **PROC HPDS2**
  - HPA implementation of next-generation data step (DATA step 2)
  - DS2 program is executed in parallel on appliance
  - Efficient distributed scoring
  - Efficient method of moving data into the appliance

```sas
proc hpds2 data=mydata
   out = gplib.table1(distributed_by='distributed randomly');
   performance host="cda.lob.com" commit=1000000;
   data DS2GTF.out;
   method run();
   set DS2GTF.in;
   end;
enddata;
run;
```
SAS/HPA and SAS Grid Manager

- Fully integrated products
- Grid Manager provides
  - Access to SAS sessions
  - Workload management
  - Distribution at the task (PROC, DATA) level

```sas
data one; set sasuser.baseData;
   < do stuff >
proc hpreg data=one;
run;

data two; set Mystuff.CustomerData;
proc hpreduce data=two;
run;

proc hpnlin data=GreenPlum.MyTable;
run;
```
SAS Grid Manager and SAS/HPA Alongside Greenplum

```sas
data one; set sasuser.baseData;
< do stuff >
proc hpreg data=one;
run;
```

```sas
data two; set Mystuff.CustomerData;
proc hpreduce data=two;
run;
```

```sas
proc hpnlin data=GreenPlum.MyTable;
run;
```

= SAS High Performance Suite  
= SAS Grid Manager
It should be called SAS High Performance Suite