



»» TRANSFORMING PROMISING IDEAS INTO COMMERCIAL REALITY

Big Data – Step towards High Performance Analytics

Author: Dhiraj Dabhi





Agenda

- Introduction
- Basics of High Performance Analytics(HPA) procedures
- What is HPSUMMARY?
- Comparison of PROC SUMMARY and PROC HPSUMMARY in Single machine mode
- Advantages of HPA procedures
- Other programming tips and techniques



Introduction

What is Big data?

- Big data describes a massive volume of structured and unstructured data that is so large that it's difficult to process using traditional database techniques.
- Gartner defines Big Data as high volume, velocity and variety (3Vs) information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
- At SAS® two additional dimensions when thinking about big data: Variability and Complexity



Basics of High Performance Analytics(HPA) procedures

- In order to deal with the challenges of massive data, SAS has introduced HPA procedures for faster processing and decision making.
- HPA procedures run in 2 different modes
 - Single-Machine mode - HPA procedures work on a PC, a Server and a Grid but will only use the cores of the machine assigned to execute it.
 - Distributed mode - A cluster of nodes with SAS High-Performance Analytics Infrastructure which makes the HPA procedures very powerful. In this mode the data will be loaded into the memory distributed across multiple nodes. The execution involves the cores of all the multiple nodes.



Basics of High Performance Analytics(HPA) procedures

Contd...

- SAS HPA procedures can execute in symmetric multiprocessing (SMP) or massively parallel processing (MPP) mode.
- Symmetric multiprocessing (SMP)
 - In SMP mode the procedures support multithreading on client machine. The procedure uses the number of CPUs on the machine to determine the number of concurrent threads.
- Massively parallel processing (MPP)
 - In MPP mode the analysis occurs on several nodes of distributed computing environment.
 - In MPP mode the HPA performs the analytics on an appliance that consists of cluster of nodes. This appliance can be:
 - 1) a database management system (DBMS) appliance on which the SAS High-Performance Analytics infrastructure is installed.
 - 2) a cluster of nodes that have the SAS High-Performance Analytics infrastructure installed but no DBMS software installed.



Basics of High Performance Analytics(HPA) procedures

Contd...

- HPA procedures has 3 main components:
 - > SAS Grid Computing - shared environment for large jobs and support a growing number of users efficiently.
 - > SAS In-Database - executes logic in the database itself. It uses a MPP (Massive Parallel Processing) for faster execution of tasks.
 - > SAS In-Memory Analytics - eliminates need for disk based processing for faster analysis. Divides processes into pieces with computation distributed across nodes.



List of HPA procedures

SAS High Performance Statistics	SAS High Performance Econometrics	SAS High Performance Optimization	SAS High Performance Data Mining	SAS High Performance Text Mining	SAS High Performance Forecasting
HPLOGISTIC HPREG HPNLIN HPMIXED HPNLMOD HPSPLIT HPGENSELECT HPCORR	HPCOUNTREG HPSEVERITY HPQLIM	HPSLO Select features in OPTMILP OPTLP OPTMODEL	HPREDUCE HPNEURAL HPFOREST HP4SCORE HPDECIDE HPSVM	HPTMINE HPTMScore	HPFORECAST
Common set includes: HPSUMMARY, HPSAMPLE, HPDS2, HPDMDB, HPIMPUTE, HPBIN, HPCORR					



What is HPSUMMARY?

- HPSUMMARY procedure computes the basic descriptive statistics of the variables.
- provides the similar functionality and syntax as traditional SAS SUMMARY procedure
- In general, PROC HPSUMMARY does the following:
 - > Calculates descriptive statistics based on moments
 - > Calculates and estimates quantiles, which includes the median
 - > Calculates CI for the mean
 - > Identifies extreme values
 - > Performs t test



What is HPSUMMARY? Contd...

Syntax:

```
PROC HPSUMMARY <options> <statistics-keywords>;  
    CLASS variables </ options>;  
    FREQ variable;  
    OUTPUT <OUT = SAS-dataset> <output-statistic-specifications> </ AUTONAME>;  
    PERFORMANCE performance-options;  
    TYPES requests;  
    VAR variables </ WEIGHT=weight-variable>;  
    WAYS list;  
    WEIGHT variable;  
  
RUN;
```



Comparison between HPSUMMARY and SUMMARY

In the below example we have PROC HPSUMMARY and PROC SUMMARY code run on a single machine mode with multithreading enabled with number of threads = 4.

PROC HPSUMMARY	PROC SUMMARY
Code: <pre>PROC HPSUMMARY data = dm; CLASS trt age race sexc; VAR height weight; TYPES () trt*age trt*race trt*sexc; OUTPUT OUT = demog_stat2 n=n mean=mean std=std median=median min=min max=max; RUN;</pre>	Code: <pre>PROC SUMMARY data = dm; CLASS trt age race sexc; VAR height weight; TYPES () trt*age trt*race trt*sexc; OUTPUT OUT = demog_stat2 n=n mean=mean std=std median=median min=min max=max; RUN;</pre>
In Single Machine Mode	
Log: <pre>206 proc hpsummary data = dm; 207 class trt age race sexc; 208 var height weight; 209 types () trt*age trt*race trt*sexc; 210 211 output out = demog_stat2 n=n mean=mean std=std median=median min=min max=max; 212 run; NOTE: Writing HTML Body file: sashtml2.htm NOTE: The HPSUMMARY procedure is executing in single-machine mode. NOTE: There were 100000000 observations read from the data set WORK.DM. NOTE: The data set WORK.DEMOG_STAT2 has 169 observations and 12 variables. NOTE: PROCEDURE HPSUMMARY used (Total process time): real time 4:46.24 user cpu time 3:57.57 system cpu time 3.20 seconds memory 11083.29k OS Memory 27164.00k Timestamp 09/04/2014 09:53:21 PM Step Count 12 Switch Count 0</pre>	Log: <pre>213 proc summary data = dm; 214 class trt age race sexc; 215 var height weight; 216 types () trt*age trt*race trt*sexc; 217 218 output out = demog_stat1 n=n mean=mean std=std median=median min=min max=max; 219 run; NOTE: There were 100000000 observations read from the data set WORK.DM. NOTE: The data set WORK.DEMOG_STAT1 has 169 observations and 12 variables. NOTE: PROCEDURE SUMMARY used (Total process time): real time 5:56.97 user cpu time 4:21.53 system cpu time 10.28 seconds memory 7778.31k OS Memory 28180.00k Timestamp 09/04/2014 10:14:13 PM Step Count 13 Switch Count 0</pre>

NOTE: In the above single machine mode example although the performance increase is not significant, the actual benefits of HPA are visible when we work on distributed machine with multiple nodes (MPP mode). In certain scenarios depending on the environment in which we are running, HPA procedures can run 15 to 20 times faster than the traditional SAS procedures.



Advantages of HPA procedures

- Enables the users to run on a single machine mode where SAS is installed.
- Enables the users to run on distributed mode to distribute data and computation on different cluster of machines.
- HPA procedures are multithreaded and can exploit all the cores available, whereas not all MVA procedures are multithreaded.
- In order to leverage the benefits of HPA procedures, in SAS 9.4 the procedures in the SAS HPA products are included with the associated MVA product, and we can run these procedures in single-machine mode without licensing the high-performance product.



Other programming tips and techniques

- Size of data
 - > Efficient programming
 - > Compress Datasets
 - > Delete unwanted datasets and Drop unwanted variables
 - > Variable lengths
- Sorting large datasets
 - > Working with SORT options: SORTSIZE, MEMSIZE and TAGSORT.
- Merging datasets
 - > PROC SQL vs. Merge
 - > Merging using format
- Other techniques
 - > Syntax check before running datasets
 - > Avoid I/O error (SGIO processing)



Conclusions

- High Performance HPA procedures are developed for high performance statistical modeling while working with massive data.
- In single machine mode HPA procedures takes the advantages of the multi-threading capabilities of the system.
- In distributed mode the HPA procedures exploits all the 3 core components while working with massive data for faster performance.
- Also as the syntax are similar to traditional SAS procedures its ease to learn and adapt.
- Furthermore along with the increasing technology advantages, the basic efficient steps help to overcome few of the memory constraints and processing time.

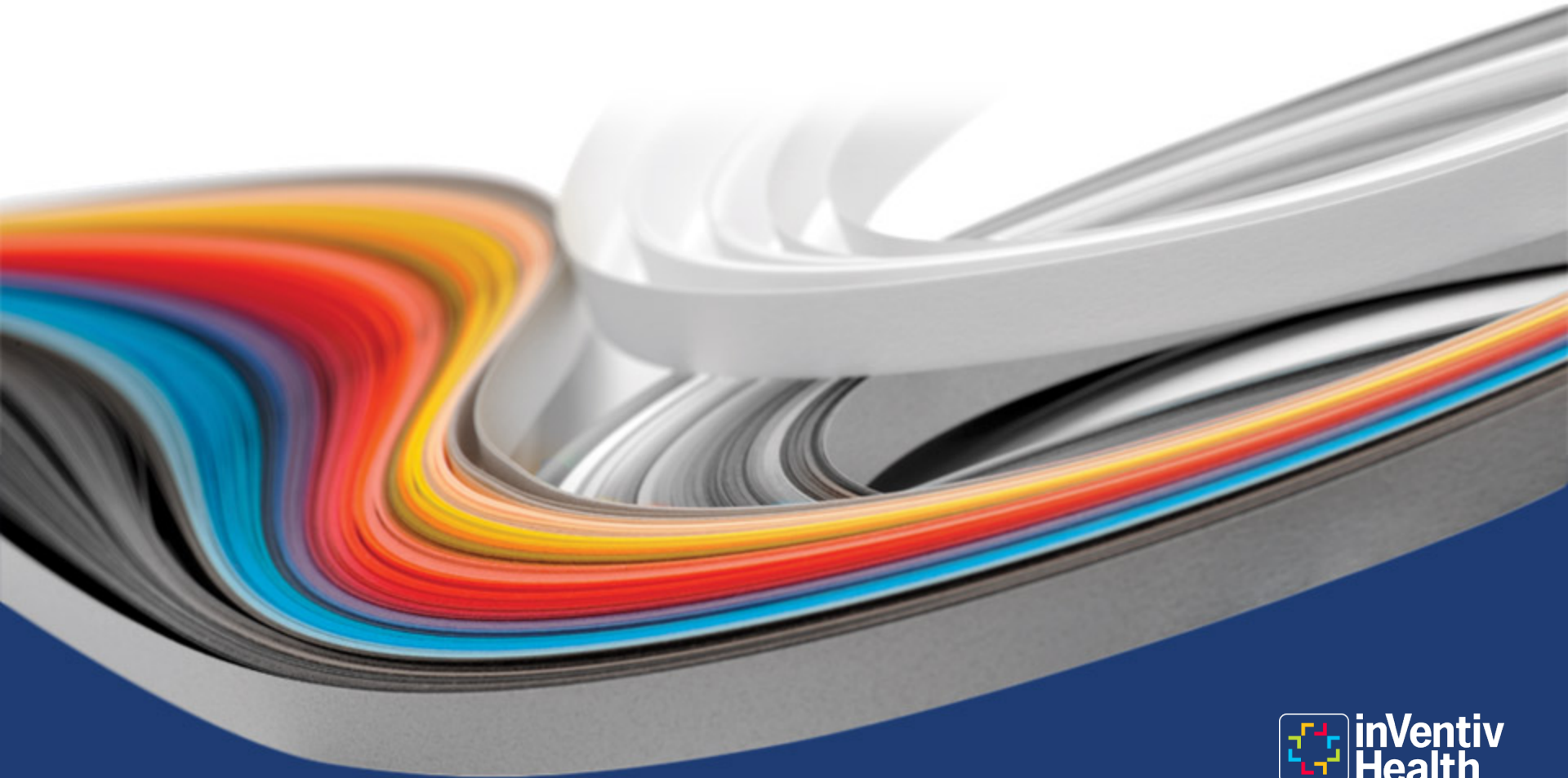


Questions / Comments/ Suggestions?





Thank You





Name: Dhiraj Dabhi

Enterprise: inVentiv Health Clinical

Address: Ground Floor, Marwah Centre,
Andheri (E)

City, State ZIP: Mumbai, Maharashtra, India – 4000 72

Work Phone: +91 (22) 4095 7363

Fax: +91 (22) 4095 7399

E-mail: dhiraj.dabhi@inventivhealth.com