The HPSUMMARY® Procedure:

An Old Friend’s Younger (and Brawnier) Cousin

The HPSUMMARY® Procedure: An Old Friend's Younger (and Brawnier) Cousin

Anh P. Kellermann, Jeffrey D. Kromrey

University of South Florida

Introduction

- HPSUMMARY is a high-performance version of the proc SUMMARY in Base SAS, and both provide similar functionalities (e.g., calculate descriptive statistics , quantiles, confidence limits for the mean, identify extreme values, do *t*- test etc.).
- HPSUMMARY procedure can run on both a single-user machine and on a cluster system where it can be run in single-machine mode or in distributed mode.
- An experiment was conducted to examine
  - SUMMARY's processing time and memory utilization compared to HPSUMMARY's in single-machine mode
  - Scalability of HPSUMMARY

Experimental Design

- PROC SUMMARY and HPSUMMARY (v.9.4) were run against 24 simulated datasets with number of variables *k* ranging from 50, 100, 500, and 1000 and number of records ranging from 10,000, 50,000, 100,000, 500,000, 1,000,000, and 10,000,000; (data volume from 0.004Gb to 76Gb).
- Each process was replicated 10 times to obtain the average measures of interest.
- The experiments were conducted on two different machines with a large different capacity:
  - a single-user machine: Windows 7 , quad core with 2.80GHz each, and a total of 8 Gb MEM
  - a 32Gb-memory node with 16 2.6GHZ-CPU's and 20Mg cache per core in a Linux cluster
- Syntax options were set so that the two procedures generate the same output.

- HPSUMMARY was run on sinlge-machine mode on the Linux cluster (on 3 different thread counts 4, 8, 16 against 3 datasets of different sizes (.004 Gb, 38Gb, 76Gb).

HPSUMMARY vs. SUMMARY: on Single-User Windows Machine

N of Obs	Number of Variables							
	50		100		500		1000	
	HP	Sum.	HP	Sum.	HP	Sum.	HP	Sum.
10,000	✓	✓	✓	✓	✓	✓	✓	✓
50,000	✓	✓	✓	✓	✓	✓	✓	Failed
100,000	✓	✓	✓	✓	✓	Failed	✓	Failed
500,000	✓	✓	✓	✓	✓	Failed	✓	Failed
1,000,000	✓	✓	✓	✓	✓	Failed	✓	Failed
10,000,000	✓	✓	✓	✓	✓	Failed	✓	Failed

Table 1. The Completion of HPSUMMARY and SUMMARY Procedures Running against Different Datasets on Windows Machine with Limited Memory

- See HPSUMMARY vs. SUMMARY: real time & CPU time on Windows machine.
- See HPSUMMARY vs. SUMMARY: memory, real time & CPU time on Linux HP machine .
- See HPSUMMARY's real time & CPU time on different thread counts
- See conclusion/recommendation

HPSUMMARY vs. SUMMARY: on Win. Single-User Machine

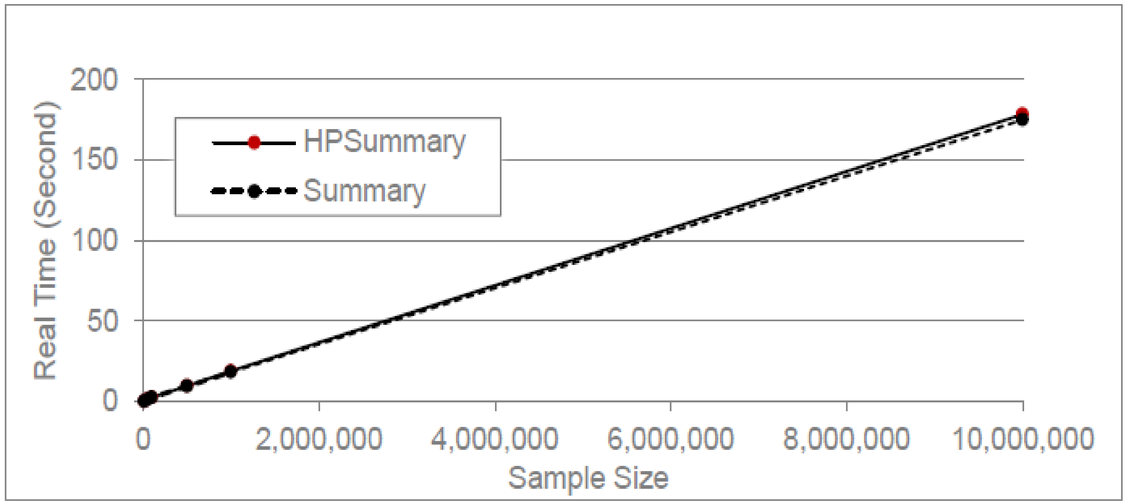

Figure 1: Average Real Time for HPSUMMARY and SUMMARY Procedure (k=50) on Windows Laptop

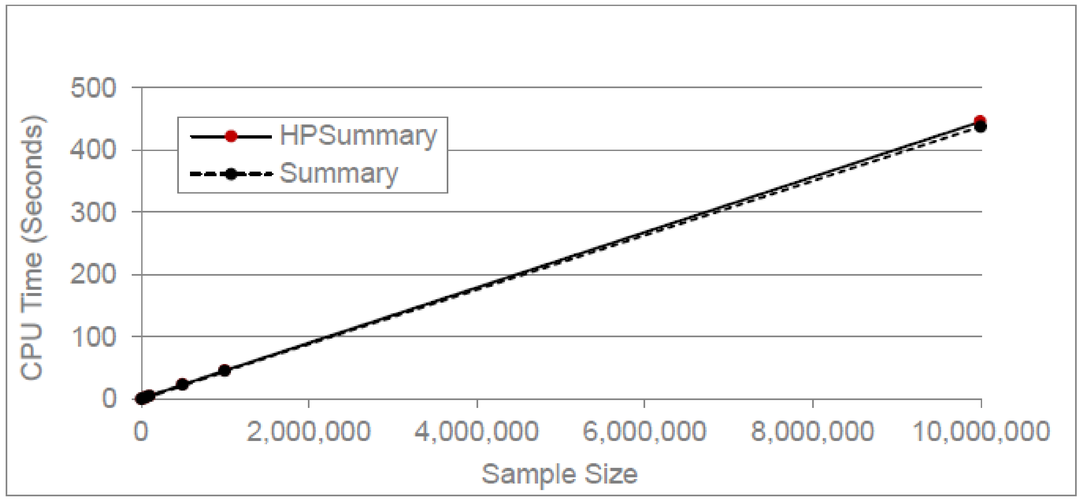

Figure 2: Average CPU Time for HPSUMMARY and SUMMARY Procedure (k=50) on Windows Laptop

- On a single-user machine with low memory capacity, when data volumes are small, HPSUMMARY's and SUMMARY's processing times are virtually the same.

HPSUMMARY vs. SUMMARY: on Linux HP Machine

❖ Memory Usage

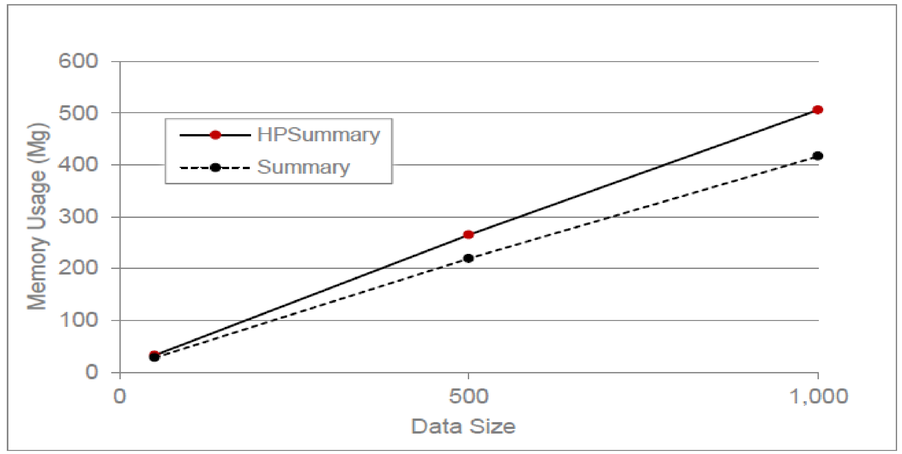

Figure 5: HPSUMMARY and SUMMARY's Memory Usage by Data Size

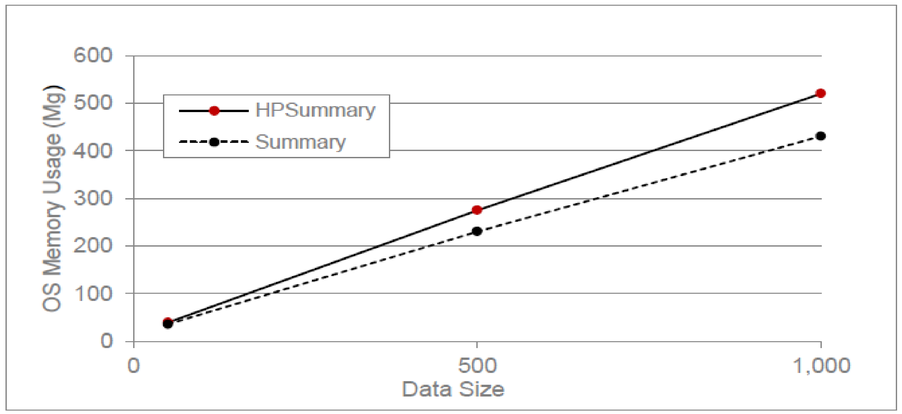

Figure 5: HPSUMMARY and SUMMARY's OS Memory by Data Size

- HPSUMMARY utilized more memory than SUMMARY
- Both reserved memory efficiently (e.g.,OS memory ≈ memory)

❖ Processing Time

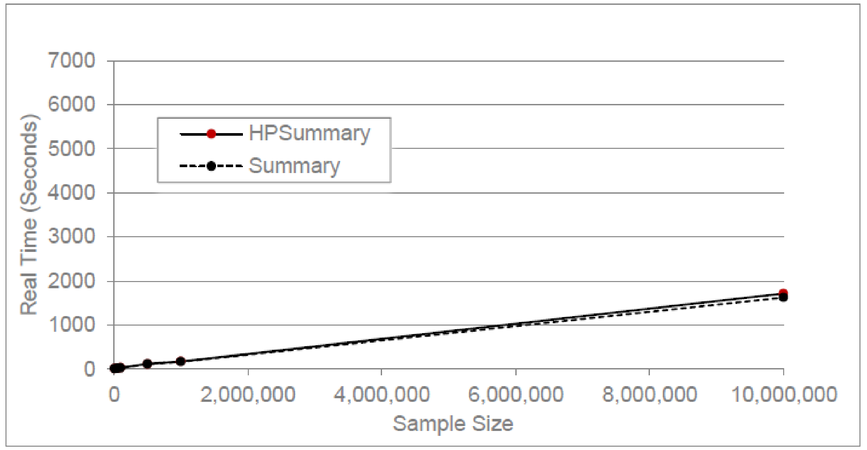

Figure 4: Real Time for HPSUMMARY and SUMMARY (k=1000) on Linux

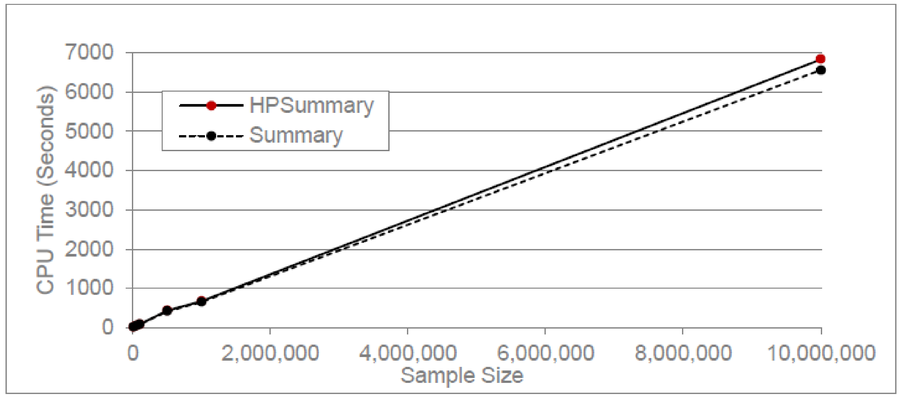

Figure 4: CPU Time for HPSUMMARY and SUMMARY (k=1000) on Linux

- On an HP machine, HPSUMMARY's & SUMMARY's real time to process large data volumes are virtually the same.

HPSUMMARY Processing Time on Different Thread Counts

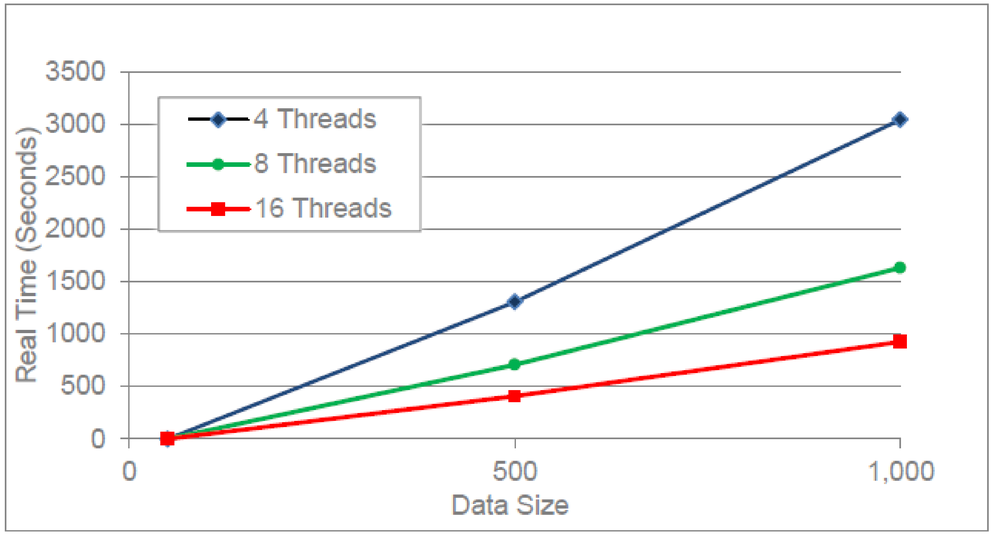

Figure 7: HPSUMMARY's Real Time for Different Thread Counts

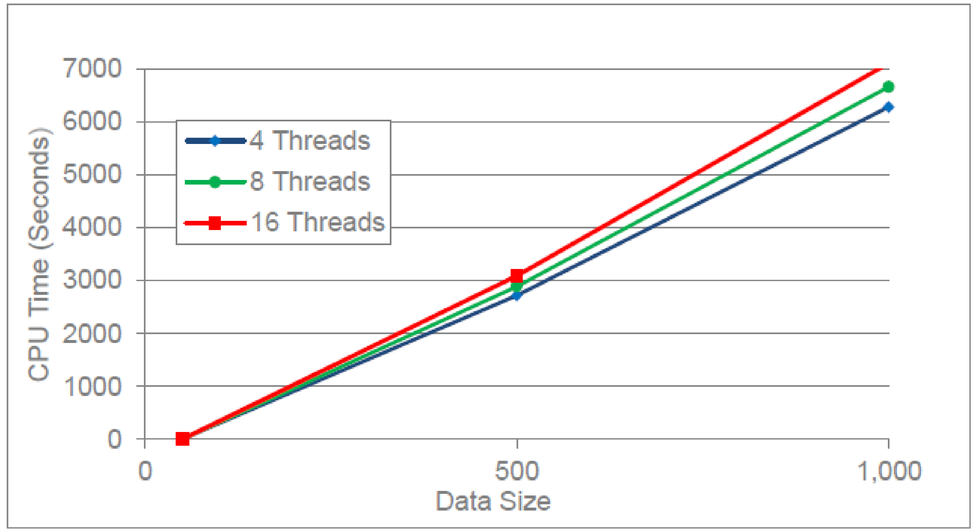

Figure 7: HPSUMMARY's CPU Time for Different Thread Counts

CONCLUSION

- PROC HPSUMMARY provides better memory management than PROC SUMMARY in a memory-limited, single-user machine.
- On an HP machine, HPSUMMARY utilized all cores available to it, substantially reducing real time.
- HPSUMMARY's real time decreasing rate is not proportional with the increasing rate of consumed resource (e.g., CPU time).
- Recommendation:
  - In a limited-memory, single-user machine environment, if available, HPSUMMARY can be a preferred choice to SUMMARY
  - In a busy, shared environment where requesting large computation resource requires a long waiting time period, trading off between waiting time and shorter real time should be taken into consideration.