

Efficient Analysis of Large Datasets

Kevin D. Brunson, University of Notre Dame, Notre Dame, Indiana
 Patricia Ledesma Liebana, University of Notre Dame, Notre Dame, Indiana

ABSTRACT

This paper will describe experiences trying to analyze and store data in a combined NT and UNIX environment constrained by the total size and number of data files extracted from the 1990 Census of Population and Housing 5% Microdata Sample. The total space required exceeded local PC disk space so that permanent storage of the data had to take place on AFS file servers. Therein, the total size was an encumbrance due to bandwidth limitations that impeded the transfer, in addition to the undesirability of manipulating files larger than about 100 Mb. in AFS.

We believe that this paper will be particularly relevant to researchers and support personnel in an academic setting because computing environments are usually heterogeneous. Often, each task is unique, and a method that works in one instance may be inadequate in another as dictated by the input data, required analysis techniques, or software and hardware configuration. In addition to describing experiences, the authors hope to provoke a discussion of various alternatives.

INTRODUCTION

When researchers are confronted with space constraints they may have to resort to "tricks" to complete their project. It really doesn't matter if the files are small or large in absolute measurements as long as the amount of available space is being exceeded. However, file size does matter when time becomes an impediment, either because of a large amount of computation time, or because other data handling tasks such as reading input files or writing output files requires large amounts of time. If a program requires an hour to complete, and the program fails in the last few minutes, at best it is quite annoying--more likely maddening--especially when a deadline approaches. Unfortunately, it is unlikely that all obstructions are considered beforehand, especially when the computing environment is dynamic as was ours, so identifying important general issues can have large benefits.

In the paper "Selectivity Bias in Studies of Wage Discrimination" (Bartholomew, Brunson, and Kořanowski) presented at the 1998 Midwest Economic Association Meetings, the authors presented results of a study that had two goals. First, two regression techniques, the linear probability approach and a probit analysis

employing the inverse-Mills ratio, were compared in order to investigate the consequences of selection bias. Second, they built a database of the 50 states plus Washington D.C. that would permit continuing analysis across states and between regions. The 5% Microdata Sample from the 1990 Census of Population and Housing was the basis for the database.

This paper will describe experiences trying to analyze and store that data in a combined NT and UNIX environment constrained by the total size and number of data files. The largest ASCII file on the CDROM's was California at 475 Mb and the smallest was Wyoming at 9.1 Mb; the total size of all 51 files was 3.9 Gb. Details of the various processes is not as vital as learning to identify key components of them.

The next section of the paper provides information about the computing environment and the section that follows it describes data extraction issues. Following that is a brief summary of the SAS® programs, which precedes concluding comments.

COMPUTING ENVIRONMENT

The desktop computer was an IBM Model 350 with a 166 Mhz Pentium processor, 64 Mb of RAM, a 1.5 Gb hard drive, and a 24X CDROM. The operating system was NT 4.0 and SAS was version 6.12. The PC was connected to a 10 mps ethernet employing the AFS file system on SUN UNIX servers.

Also utilized was a SUN Enterprise 4000 with 2-250 MHz UltraSPARC CPUs each with 4-Mbyte of external cache, 512 Mbyte memory, and a total of 60 Gb of disk storage. The system uses 12Gbyte of scratch space. The operating system was Solaris 2.x and SAS was version 6.12. This server is dedicated to statistical/research computing.

Notre Dame employs a variation of network served NT applications, called NT@ND, since August, 1997. Applications reside on UNIX servers and an NT NFS client provides the link so that applications are loaded locally at launch, but some interaction with the user's AFS file space occurs. The deployment of NT@ND in lieu of Windows 3.1 and Novell servers is one significant disruption in the way research is conducted at Notre Dame. The other development is elimination of the IBM mainframe and dependence upon SUN UNIX servers in conjunction with AFS (which has been employed for about 10 years). It's because we

have 2 new computing systems that this paper came about.

DATA EXTRACTION

The 3.9 Gb of plain text data files were contained on 7 CDs.

Because the entire analytical process was time consuming and because it was to be shared with collaborators at another school, it was advantageous in the long run to create a library of permanent datasets. It wasn't so clear how the library should be created, but it was necessary for it to reside in AFS. So, somehow the data had to be read from the CDs, analyzed with SAS, and stored on the network. These strategies are suggested by those requirements:

A-- Use the PC to read the CDs and move the data to the SUN server where it could be analyzed and stored.

B-- Read and analyze entirely on the PC and then store the output in AFS.

C-- Use SAS on NT to first produce a subset with only essential variables, transfer the output to the SUN server either as raw data or in transport format, and complete analysis and storage there.

Parameters necessary to consider are:

- 1-- How efficiently could data be transferred from the PC to the SUN server?
- 2-- How efficiently could the PC execute the SAS programs?
- 3-- How efficiently could the SUN server execute the SAS programs?
- 4-- Should data be compressed before transfer?
- 5-- When would space become insufficient on the PC?
- 6-- How could SAS programming techniques be used to improve the efficiency of the process?

SAS PROGRAM

The SAS program was a series of REG and MEANS procedures on various subsets of the data. The objective was to explore the influence of race (blacks and whites) on wages so not every variable and every observation was needed.

The first stage selected 15 variables and calculated an additional 61. PROC MEANS was next performed

on the 76 variables before performing PROC REG on 58 of them.

The output of the PROC REG was used to calculate a variable for the linear probability model, which consisted of a PROC MEANS and PROC REG separately on blacks and whites.

PROBIT estimation required calculation of a binary dependent variable, PROC PROBIT to calculate a variable needed to compute the inverse-Mill's ratio, and PROC MEANS and PROC REG on both race categories separately.

CONCLUSION

Statistics on only 3 states are reported here, the largest, California, the third smallest, Alaska, and one slightly above the median, Arizona. They were conveniently contained on the same CD. File size is in megabytes. Time format is mm:ss except for California where jobs required more than an hour of Real Time.

A-- The data could be moved by ftp or cut-and-paste:

STATE	FILE SIZE	CUT-PASTE	FTP
AK	10	00:23:00	00:18:10
AZ	62	02:24:00	01:55:50
CA	475	17:53:00	15:28:20

Since ftp could be accomplished with less oversight and intervention, it would be preferred even if it did not have a time advantage.

The data was not analyzed on the SUN server, but times would be similar to those in C, perhaps slightly longer since a larger file would have to be processed initially.

B-- Reading from CD's, performing all SAS analysis on the PC, and writing a SAS dataset to AFS.

STATE	OUTPUT FILE SIZE	TIME
AK	2.4	14:18.69
AZ	15.0	68:12.00
CA	*	*

* Size of the California file prevented processing on the PC.

C-- Using SAS on the PC to extract a subset of 15 variables and PUT a raw data file to AFS:

STATE	OUTPUT FILE SIZE	TIME
AK	1.2	0:11.23
AZ	7.1	1:17.79
CA	58.1	8:38.78

Analyze the data on the SUN server:

STATE	OUTPUT FILE SIZE	TIME
AK	2.4	02:47.94
AZ	15.0	55:06.42
CA	123.9	9:35:34.77

I did not control for the time of day or day of the week. After all, a researcher can't always choose the optimum time, and if they all did, then that time would become sub-optimum.

Once again, there was no expectation that there would be a best solution, but influential conditions were identified to guide future computing models.

ACKNOWLEDGMENTS

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. Indicates USA registration.

REFERENCES

Bartholomew, Wayne, Kevin Brunson, and Paul Kochanowski (1998), "Selectivity Bias in Studies of Wage Discrimination," paper presented at the 1998 Midwest Economic Association Meetings, Chicago, Illinois.

Bureau of the Census, "1990 Census of Population and Housing, Public Use Microdata Samples," Washington, D.C.

AUTHORS

Kevin D. Brunson, 315 College of Business, Notre Dame, IN 46556
219-631-9448 kbrunson@nd.edu

Patricia Ledesma Liebana, Flanner Hall, Notre Dame, IN 46556
219-631-8595 pledasma@nd.edu

```

/***** UNIX SAS PROGRAM *****/;

%global st;
%let st = AK;

OPTIONS NOCENTER NOSOURCE errors=1;

libname sdh2 "/scratch1/cph";

filename txth "~/Private/sas/coba/cph/rawdata/&st..txt";

title "STATE OF &st";

proc format;
  invalue schyear 1-3=0 4=2 5=7 6=9 7=10 8=11 9-10=12 11=13
    12-13=14 14=16 15=18 16=21 17=22;

data st1; infile txth;
  input
    rectype $ sex race age marital yrsch mobility migstate
    disabl1 industry occup class week89 hour89 rearning
    ;

(lines showing creation of 61 variables deleted)

proc means data = st1 n mean min max std nmiss;

/*** OLS REGRESSION ***/

proc printto log = "~/Private/sas/coba/cph/logs/cph2.log"
  print = "~/Private/sas/coba/cph/listings/cph2_reg.lst";

title "STATE OF &st";

```

```

data st2; set st1(keep = obsnum deph
yearschr yschsq ba exp expsq disabil migrant married divorced
separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui
wti rti firei brsi psi ersi prsi admin amdi ptw
ysexp yschq expq yssso yspcro ysoflo ysmo ysmi ysci ysmndi ysmdi
ystcopui yswti yrti ysfirei ysbrsi yspsi yseri ysprsi yadmin
ysamdi ysptw ysgovern);

proc reg data = st2;
&st: model deph =
yearschr yschsq ba exp expsq disabil migrant married divorced
separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui
wti rti firei brsi psi ersi prsi admin amdi ptw
ysexp yschq expq yssso yspcro ysoflo ysmo ysmi ysci ysmndi ysmdi
ystcopui yswti yrti ysfirei ysbrsi yspsi yseri ysprsi yadmin
ysamdi ysptw ysgovern
;
output out = st2 predicted = predval;

data st2; set st2(keep = obsnum predval);

/*** LPM REGRESSION **/

data st3; set st2;
pvm = predval - 1;
obsnum = _n_;
select;
when (0 lt pvm le 1) pvm1 = 0;
when (pvm lt -1) pvm1 = -1;
otherwise pvm1 = pvm;
end;
drop pvm predval;

data sdh2.&st st4; merge st1 st3;
by obsnum;

proc printto log = "~/Private/sas/coba/cph/logs/cph3.log"
print = "~/Private/sas/coba/cph/listings/cph3bw_mr.lst";

title "STATE OF &st";

data black white; set st4(where = (reaming gt 0 and age gt 15 and sex = 0));
if race = 2 then output black;
if race = 1 then output white;

proc means data = black n mean min max std;

proc reg data = black;
BLACKI: model lneph =
yearschr yschsq ba exp expsq disabil migrant married divorced
separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
firei brsi psi ersi prsi admin amdi ptw govern
;
BLACKL: model lneph =
yearschr yschsq ba exp expsq disabil migrant married divorced
separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
firei brsi psi ersi prsi admin amdi ptw govern pvm1
;

title2 'BLACK';

```

```

proc means data = white n mean min max std;

proc reg data = white;
  WHITE1: model lneph =
    yearsch yschsq ba exp expsq disabil migrant married divorced
    separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
    firei brsi psi ersi prsi admin amdi ptw govern
  ;
  WHITE1: model lneph =
    yearsch yschsq ba exp expsq disabil migrant married divorced
    separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
    firei brsi psi ersi prsi admin amdi ptw govern pvm1
  ;

title2 'WHITE';

  /*** CREATE A BINARY DEPENDENT VARIABLE ***/
data st; set st4;
  select (lneph);
    when (.) depend = 0;
    otherwise depend = 1;
  end;

  /*** PROBIT MODEL ***/
proc probit data = st order = data;
  class depend;
  PROBIT: model depend =
    yearsch yschsq ba exp expsq disabil migrant married divorced
    separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
    firei brsi psi ersi prsi admin amdi ptw govern
  ;
  output out = st2 xbeta = jhat;

  /*** COMPUTE INVERSE MILL'S RATIO ***/
data sdh2.&st st3; set st2;
  bigf = probnorm(jhat);
  if depend = 1 then do;
    lamda = ((1 / sqrt(2 * (22 / 7))) * exp(-.5 * (jhat**2)))/bigf;
  end;
  else if depend = 0 then do;
    lamda = ((1 / sqrt(2 * (22 / 7))) * exp(-.5 * (jhat**2)))/(bigf - 1);
  end;

  /*** REGRESSIONS WITH INVERSE MILL'S RATIO ***/
data black white; set st3(where = (rearning gt 0 and age gt 15 and sex =
0));
  select;
    when (race=2) output black;
    when (race=1) output white;
    otherwise;
  end;

proc means data = black n mean min max std;

proc reg data = black;
  BLACK: model lneph =
    yearsch yschsq ba exp expsq disabil migrant married divorced
    separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti
    firei brsi psi ersi prsi admin amdi ptw govern lamda

```

```
;  
  
proc means data = white n mean min max std;  
  
proc reg data = white;  
  WHITE: model lneph =  
    yearsch yschsq ba exp expsq disabil migrant married divorced  
    separate widowed mpso tsaso so ppcro oflo mo mi ci mndi mdi tcopui wti rti  
    firei brsi psi ersi prsi admin amdi ptw govern lamda  
  ;  
  
run;
```