



Differences in performance between the options KEEP and WHERE as compared to the respective statements KEEP and IF in SAS data steps.



Ruurd Bennink, OCS Consulting, Netherlands



EU Connect 2021

Introduction



In general, it is assumed that keeping variables while reading in a SAS dataset (via the KEEP option on the SET statement) is more efficient than applying a KEEP statement within the SAS data step.

It is also assumed that selecting observations while reading in a SAS dataset (via the WHERE option on the SET statement) is more efficient than applying the IF statement.

This presentation will look into the difference in performance of using the beforementioned options and statements by applying them in several scenarios. Is there really a big difference in performance? Or is it negligible?



*KEEP statement;

VAR10_1;

SET all;

KEEP cvar10 nvar10;

RUN;

*KEEP option;

DATA var10_1;

SET all (Keep=cvar10 nvar10);

RUN;



What is performance and how is it measured?

Within this presentation, the performance of a SAS program is determined by the time you need to wait until your program run is finished. This time is given in the SAS log file as 'real time'. The longer the 'real time' is, the lower the performance is. In the SAS log file the 'real time' is presented after the conduct of each data step or SAS procedure.



What is performance and how is it measured?

KEEP statement

NOTE: There were 1000000 observations read from the data set WORK.ALL.
NOTE: The data set WORK.VAR10_1 has 1000000 observations and 2 variables.
NOTE: DATA statement used (Total process time):
| real time 0.38 seconds
 cpu time 0.37 seconds

KEEP option

NOTE: There were 1000000 observations read from the data set WORK.ALL.
NOTE: The data set WORK.VAR10_1 has 1000000 observations and 2 variables.
NOTE: DATA statement used (Total process time):
 real time 0.35 seconds
 cpu time 0.36 seconds



From literature (e.g. <https://support.sas.com/resources/papers/proceedings12/357-2012.pdf>) it can be concluded that two tips to increase the performance of your SAS programs are:

- 1) Use the KEEP or DROP option on the SET statement instead of the KEEP and DROP statement in a data step.**

Using the KEEP option on the SET statement means that only the selected variables are included in the program data vector (PDV). On the other hand, when using the KEEP statement all variables will be included into the PDV. Hence it is expected that using the KEEP option on the SET statement will show a gain in performance as compared with the KEEP statement. Another possibility is to use the KEEP option on the DATA statement, however that will have the same effect as using the KEEP statement.

Background



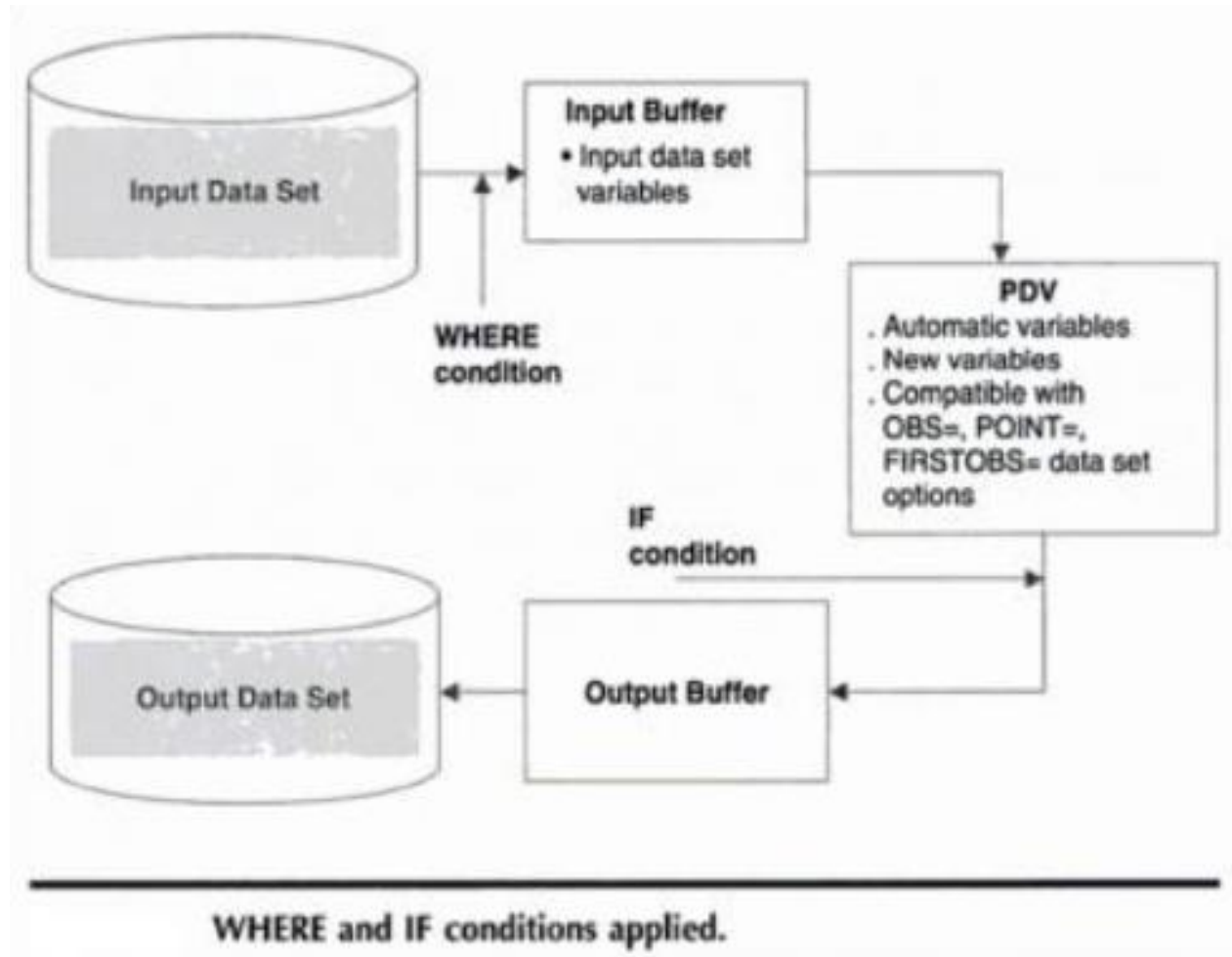
2) Use the WHERE option on the SET statement or WHERE statement instead of the IF statement in a data step.

The graph on the next slide shows that when using the WHERE condition, only the selected observations will be included in the PDV whereas the IF condition will include all observations in the PDV. Another option would be to use the WHERE statement, but this has the same effect as the WHERE option on the SET statement.

Background



2) Use the **WHERE** option on the **SET** statement or **WHERE** statement instead of the **IF** statement in a data step.





The effect of using the KEEP option instead of the keep statement

To examine this effect in detail a dataset was simulated with 50 randomly filled variables, 25 character and 25 numeric, and one million observations, keeping 2 variables:

```
DATA all;

  ARRAY cvar{25} $25;
  ARRAY nvar(25) 8;

  DO vari=1 to 25;
    DO chari=1 to 25;
      substr(cvar{vari},chari,1)=byte(floor(ranuni(354)*26)+62);
    END;
  END;

  DO vari=1 to 25;
    nvar{vari}=ranuni(354);
  END;

  DO obsi=1 to 1000000; OUTPUT; end;

RUN;
```



The effect of using the KEEP option instead of the keep statement

```
%macro keeploop;  
  
  %do i=1 %to 25;  
  
    %put KEEP statement;  
    DATA var10_1;  
      SET all;  
      KEEP cvar10 nvar10;  
    RUN;  
  
    %put KEEP option;  
    DATA var10_1;  
      SET all (keep=cvar10 nvar10);  
    RUN;  
  
  %end;  
  
%mend keeploop;  
  
%keeploop  
  
/* End of Program */
```



The effect of using the KEEP option instead of the keep statement

The presented part from the log file in the introduction presents only one run with KEEP option and KEEP statement, but the macro KEELOOP in the program above performs this cycle 25 times. After those 25 runs it is possible to make a statement about the difference in performance between KEEP option and KEEP statement.

A SAS program was created that reads in the log file and makes a comparison between the real times of the 25 KEEP statements vs. the real times of the corresponding 25 KEEP options. The summary statistics are as follows:



The effect of using the KEEP option instead of the keep statement

KEEP statement vs. KEEP option, 50 vars of which 25 numeric and 25 character, 1.000.000 observations

The MEANS Procedure

Analysis Variable : perc_dif Percentage increase in performance Option vs statement				
N	Mean	Std Dev	Minimum	Maximum
25	4.5858937	2.4661464	0	10.2564103

KEEP statement vs. KEEP option, 200 vars of which 100 numeric and 100 character, 1.000.000 observations

The MEANS Procedure

Analysis Variable : perc_dif Percentage increase in performance Option vs statement				
N	Mean	Std Dev	Minimum	Maximum
25	6.5418208	1.4967834	3.5326087	11.8090452



The effect of using THE WHERE option instead of the IF statement

To examine the difference in performance between the use of the WHERE option on the Set statement instead of the IF statement in a SAS data step a similar approach was used as for the comparison between the performance of the KEEP option vs. the KEEP statement. Almost the same simulated dataset was used (25 character variables, 25 numeric variables, but 10 million observations) and including an additional variable EXTRAVAR to select approximately 50% of the records

```
DATA all;

  ARRAY cvar{25} $25;
  ARRAY nvar{25} 8;

  DO vari=1 to 25;
    DO chari=1 to 25;
      substr(cvar{vari},chari,1)=byte(floor(ranuni(354)*26)+62);
    END;
  END;

  DO vari=1 to 25;
    nvar{vari}=ranuni(354);
  END;

  DO obsi=1 to 10000000;
    extravar=ranuni(354);
    OUTPUT;
  END;

RUN;

%macro keeploop;

  %do i=1 %to 25;
    %put IF statement;
    DATA var10_1;
      SET all;
      IF extravar > 0.5;
    RUN;

    %put WHERE option;
    DATA var10_1;
      SET all (Where=(extravar>0.5));
    RUN;
  %end;

%mend keeploop;
```



The effect of using THE WHERE option instead of the IF statement

As can be seen in the presented program, 50% of the records are selected by the IF statement and WHERE option. The performances are compared for the 25 runs and the summary statistics are:

IF statement vs. WHERE option, 50 vars of which 25 numeric and 25 character, 10.000.000 observations.
Condition: Ranuni(354) > 0.5.

The MEANS Procedure

Analysis Variable : perc_dif Percentage increase in performance Option vs statement				
N	Mean	Std Dev	Minimum	Maximum
25	-1.1782444	2.3097388	-5.6971514	5.4156171

The above table shows that the use of the WHERE option is about -1.2% more efficient on average after 25 runs with a minimum of -5.7% and a maximum of 5.4%.



The effect of using THE WHERE option instead of the IF statement

The same was redone, but now with *extravar* > 0.99 as a condition, meaning that now only approximately 1% of the records is selected instead of the 50% above. The result is then:

IF statement vs. WHERE option, 50 vars of which 25 numeric and 25 character, 10.000.000 observations.
Condition: Ranuni(354) > 0.99.

The MEANS Procedure

Analysis Variable : perc_dif Percentage increase in performance Option vs statement				
N	Mean	Std Dev	Minimum	Maximum
25	7.5739719	4.4492105	-9.2948718	17.1919771

The above table shows that the use of the WHERE option is about 7.6% more efficient on average after 25 runs with a minimum of -9.3% and a maximum of 17.2%.



Conclusion – KEEP option vs. KEEP statement

Although there is a difference in performance between the KEEP option (on the Set statement) and the KEEP statement in SAS data steps in favor of the KEEP option, the difference may not be as large as one may expect. However when the number of variables to keep is small as compared to the number of variables in the input dataset the increase in performance is increasing.

The smaller the ratio $N_{\text{keep}} : N_{\text{var}}$ is the more significant the gain in performance of the KEEP option as compared to the KEEP statement becomes. With N_{keep} is the number of variables to keep and N_{var} is the number of variables in the input dataset.



Conclusion – WHERE option vs. IF statement

The same can be said for the use of the IF statement vs. the use of the WHERE option (on the SET option) in a SAS data step. In this case there is even no difference in performance if approximately more than half of the observations is selected. However if a real small amount of observations is selected then there is some gain in performance, although, again, not as large as one may expect.

Here it can be said that the ratio $N_{\text{selected}} : N_{\text{observations}}$ determines the performance gain. The smaller this ratio is, the higher the performance gain is. With N_{selected} is the number of selected observations and $N_{\text{observations}}$ is the number of observations in the input dataset.

NOTES:

- All performances were measured using SAS LSAF.
- More details can be found in the corresponding paper (CT08).

Questions?

