Paper 72

SAS ® PROGRAM EFFICIENCY FOR BEGINNERS

Bruce Gilsen, Federal Reserve Board

INTRODUCTION

This paper presents simple efficiency techniques that can benefit inexperienced SAS ® software users on all platforms.

Efficiency techniques are frequently documented as follows.

- Describe an efficiency technique.
- Demonstrate the technique with examples.

The drawback to this approach is that it can be difficult for SAS software users to determine when to apply the techniques. This paper takes an alternate approach, as follows.

- Describe an application or data set.
- Present simple efficiency techniques for the application or data set.

This approach is designed to make it easier for SAS software users to determine when to apply the techniques to their application or data set.

SUMMARY OF PROGRAMMING TASKS

This paper presents efficiency techniques for the following programming tasks.

1. Create a SAS data set by reading long records from a flat file with an INPUT statement. Keep selected records based on the values of only a few incoming variables.
2. Create a new SAS data set by reading an existing SAS data set with a SET statement. Keep selected observations based on the values of only a few incoming variables.
3. Select only some observations from a SAS data set. The selected data are used as input to a SAS procedure, but are not otherwise needed.
4. In IF, WHERE, DO WHILE, or DO UNTIL statements, use OR operators or an IN operator to test if at least one of a group of conditions is true. In IF, WHERE, DO WHILE, or DO UNTIL statements, use AND operators to test if all of a group of conditions are true.
5. Select observations from a SAS data set with a WHERE statement.
6. In a DATA step, read a SAS data set with many variables to create a new SAS data set. Only a few of the variables are needed in the DATA step or the new SAS data set.
7. Create a new SAS data set containing all observations from two existing SAS data sets. The variables in the two data sets have the same length and type.
8. Process a SAS data set in a DATA step when no output SAS data set is needed. This could occur when a DATA step is used to write reports with PUT statements, examine a data set's attributes, or generate macro variables with the CALL SYMPUT statement.
9. Execute a SAS DATA step in which the denominator of a division operation could be zero.

1. Read long records from a flat file, keeping only selected records

Task

Create a SAS data set by reading long records from a flat file with an INPUT statement. Keep selected records based on the values of only a few incoming variables.

Technique

First, read only the variables needed to determine if the record should be kept. Test the values of the variables, and only read the rest of the record if necessary.

Example

Read 2000 byte records from a flat file. Keep the record (include it in the resulting SAS data set) if NETINC is greater than 100, and otherwise discard it.

Method 1, less efficient.

```
data income;
   infile incdata;
   input    @ 0001 bank 8.
            @ 0009 netinc 8.
            @ 0017 nextvar 8.
```

```
                .
                .
            @ 1993 lastvar 8.;
    if netinc > 100;
run;
```

Method 2, more efficient.

```
data income;
    infile incdata;
    input    @ 0009 netinc 8. @;
    if netinc > 100;
    input    @ 0001 bank 8.
             @ 0017 nextvar 8.
               .
               .
             @ 1993 lastvar 8.;
run;
```

In method 1, all 2000 bytes are read, and the current record is kept if NETINC is greater than 100. In method 2, the first INPUT statement reads only the variable NETINC. If NETINC is greater than 100, the second INPUT statement reads the rest of the current record, and the current record is kept. The trailing @ at the end of the first INPUT statement ensures that the current record is available to re-read. If NETINC is not greater than 100, the current record is discarded, and only 8 bytes are read instead of 2000.

Notes

1. The following statement is an example of a subsetting IF statement.

```
if netinc > 100;
```

Subsetting IF statements test a condition. If the condition is true, the SAS system continues to process the current observation. Otherwise, the SAS system discards the observation and begins processing the next observation. Subsetting IF statements can be distinguished from IF-THEN/ELSE statements because they do not contain a THEN clause.

The following statements are equivalent.

```
if netinc > 100;
if not (netinc > 100) then delete;
if netinc <= 100 then delete;
```

2. If all 2000 bytes have the same informat, then the following code is equivalent to method 2. Moving the informat to the end of the second INPUT statement makes the code easier to read.

```
data income;
    infile incdata;
    input    @ 0009 netinc 8. @;
    if netinc > 100;
    input    @ 0001 bank 8.
             @ 0017
             (nextvar
               .
             lastvar)
             (8.) ;
run;
```

2. Subset a SAS data set to create a new SAS data set

Task

Create a new SAS data set by reading an existing SAS data set with a SET statement. Keep selected observations based on the values of only a few incoming variables.

Technique

Use a WHERE statement instead of a subsetting IF statement.

A WHERE statement and a subsetting IF statement both test a condition to determine if the SAS system should process an observation. They differ as follows.

- A WHERE statement tests the condition before an observation is read into the SAS program data vector (PDV). If the condition is true, the observation is read into the PDV and processed. Otherwise, the observation is not read into the PDV, and processing continues with the next observation.

- A subsetting IF statement tests the condition after an observation is read into the PDV. If the condition is true, the SAS system continues processing the current observation. Otherwise, the observation is discarded, and processing continues with the next observation.

Example

Create SAS data set TWO from SAS data set ONE. Keep only observations where GNP is greater than 10 and CON is not equal to zero.

Method 1, less efficient.

```
data two;
    set one ;
    if gnp > 10 and con ne 0 ;
    more SAS statements
run;
```

Method 2, more efficient.

```
data two;
    set one ;
    where gnp > 10 and con ne 0 ;
    more SAS statements
run;
```

In method 1, all observations are read, and observations not meeting the selection criteria are discarded. In method 2, the WHERE statement ensures that observations not meeting the selection criteria are not read.

Notes

1. In Release 6.06 of the SAS system for MVS, WHERE statements performed less efficiently than a subsetting IF statement in a few cases. In subsequent releases, WHERE statements can perform less efficiently if they include SAS functions.

2. Because WHERE statements process data before they are read into the PDV, they cannot include variables that are not part of the incoming data set, such as the following.

- Variables you create in the current DATA step.
- Variables automatically created by the SAS system in the DATA step, such as FIRST. variables, LAST. variables, and _N_.

If data set ONE does not include the variable TAX, the following DATA step generates an error.

```
data two;
    set one ;
    tax = income / 2 ;
    where tax > 5 ;
    more SAS statements
run;
```

To prevent this error, use a subsetting IF statement instead of a WHERE statement, as follows.

```
if tax > 5 ;
```

3. Prior to Release 6.07 of the SAS system for MVS, Release 6.07 of the SAS system for UNIX, and Release 6.07 of the SAS system for PCs, WHERE statements could not contain SAS functions.

4. To improve efficiency for large data sets, experienced users can combine WHERE statements with indexes.

3. Subset a SAS data set for use in a PROC step

Task

Select only some observations from a SAS data set. The selected data are used as input to a SAS procedure, but are not otherwise needed.

Technique

Use a WHERE statement in the PROC step to select observations, and eliminate the DATA step.

Example

Execute PROC PRINT on observations in data set ONE in which GNP is greater than 0.

Method 1, less efficient.

```
data two;
    set one ;
    if gnp > 0 ;
run;
proc print data = two;
run;
```

Method 2, less efficient.

```
data two;
    set one ;
    where gnp > 0 ;
run;
proc print data = two;
run;
```

Method 3, more efficient.

```
proc print data = one;
    where gnp > 0 ;
run;
```

In method 1, the IF statement could be less efficient than a WHERE statement, and an intermediate data set is created. In method 2, an intermediate data set is created. In method 3, no unnecessary data sets are created.

Notes

1. The following statement is equivalent to the statements in method 3.

```
proc print data=one(where=(gnp>0));
```

4. Test multiple conditions with IN, OR, or AND operators

Task

In IF, WHERE, DO WHILE, or DO UNTIL statements, use OR operators or an IN operator to test whether at least one of a group of conditions is true.

In IF, WHERE, DO WHILE, or DO UNTIL statements, use AND operators to test whether all of a group of conditions are true.

Technique

Order the conditions to minimize the number of comparisons required by the SAS system, as follows.

For OR operators or an IN operator, order the conditions in *descending* order of likelihood. Put the condition most likely to be true first, the condition second most likely to be true second, and so on.

For AND operators, order the conditions in *ascending* order of likelihood. Put the condition most likely to be false first, the condition second most likely to be false second, and so on.

When the SAS system processes an IF, WHERE, DO WHILE, or DO UNTIL statement, it tests the minimum number of conditions needed to determine if the statement is true or false. This technique is known as *Boolean short circuiting*. Prior information is often available about data being processed. Use this information to improve program efficiency. Order the conditions in IF, WHERE, DO WHILE, and DO UNTIL statements to take advantage of Boolean short circuiting.

An IN operator and a series of OR operators are equally efficient, though the IN operator can make programs easier to read.

Examples

The examples in this section use SAS data set ONE, which has 50,000 observations. In data set ONE, CTRY is expected to be 'czech' in about 20,000 observations, 'hungary' in about 5000 observations, 'belgium' in about 2000 observations, and 'slovakia' in about 1000 observations. INV is expected to be greater than 900 in about 10,000 observations, and TAX is expected to be greater than 500 in about 1000 observations.

1. Use an IF-THEN statement to set GNP to 10,000 if CTRY is equal to one of four values. The following two examples are efficient because they order the values of CTRY from most likely to least likely. The two IF statements are equally efficient.

Using an IN operator.

```
data two ;
    set one ;
    if ctry in('czech','hungary','belgium',
        'slovakia') then gnp = 10000 ;
```

Using OR operators.

```
data two ;
    set one ;
    if ctry = 'czech' or ctry = 'hungary'
        or ctry= 'belgium' or ctry= 'slovakia'
        then gnp = 10000 ;
```

2. Use a WHERE statement to select observations in which CTRY is equal to 'czech' or 'slovakia'. The following two examples are efficient because they order the values of CTRY from most likely to least likely. The two WHERE statements are equally efficient.

Using an IN operator.

```
data two ;
    set one ;
    where ctry in('czech','slovakia') ;
```

Using an OR operator.

```
data two ;
    set one ;
    where ctry = 'czech' or ctry = 'slovakia' ;
```

3. Use a WHERE statement to select observations in which CTRY is equal to 'czech', INV is greater than 900, and TAX is greater than 500. The following example is efficient because it orders the conditions from most likely to be false to least likely to be false.

```
data two ;
    set one ;
    where tax > 500 and inv > 900
        and ctry = 'czech' ;
```

Notes

1. The following equally efficient statements set GNP to 10,000 if CODE is equal to one of several values. Boolean short circuiting for the IF statement was implemented in Release 6.08 of the SAS system for MVS, Release 6.09 of the SAS system for UNIX, and Release 6.10 of the SAS system for PCs. In earlier versions of SAS software, only the IN operator used Boolean short circuiting, so the first statement was more efficient.

```
if code in(100, 200, 300, 500)
    then gnp = 10000 ;

if code = 100   or code = 200 or code = 300
    or code = 500 then gnp = 10000 ;
```

2. If prior information about the data is not available, PROC

FREQ can provide information about the frequency of the conditions being tested.

5. Subset a SAS data set with WHERE statement operators

Task

Select observations from a SAS data set with a WHERE statement.

Technique

Previous sections of this paper demonstrated the WHERE statement. This section describes *WHERE statement operators*, several useful operators that can be used only in WHERE statements.

5.1. BETWEEN-AND operator.

The BETWEEN-AND operator is used to test whether the value of a variable falls in an inclusive range. This operator provides a convenient syntax but no additional functionality. The following three statements are equivalent.

```
where salary between 4000 and 5000 ;
where salary >= 4000 and salary <= 5000 ;
where 4000 <= salary <= 5000 ;
```

5.2. CONTAINS or question mark (?) operator.

The CONTAINS operator is used to test whether a character variable contains a specified string. CONTAINS and ? are equivalent.

The CONTAINS operator is case sensitive; upper and lower case characters are not equivalent. In the following DATA step, NAME is a character variable in data set ONE. Observations in which NAME contains the characters 'JO' are selected. For example, JONES, JOHN, HOJO are selected, but JAO or John are not selected. The CONTAINS operator is somewhat comparable to the SAS functions INDEX and INDEXC.

```
data two;
    set one ;
    where name contains 'JO';
    more SAS statements
```

The following statements are equivalent.

```
where name ? 'JO';
where name contains 'JO';
```

5.3. IS MISSING or IS NULL operator.

The IS MISSING operator is used to test whether a character or numeric variable is missing. IS MISSING and IS NULL are equivalent.

Example 1. Select observations in which the value of the numeric variable GNP is missing. The following three statements are equivalent.

```
where gnp is null ;
where gnp is missing ;
where gnp = . ;
```

Example 2. Select observations in which the value of the character variable NAME is not missing. The following three statements are equivalent.

```
where name is not null ;
where name is not missing ;
where name ne "" ;
```

The IS MISSING operator allows you to test whether a variable is missing without knowing if the variable is numeric or character, preventing the errors generated by the following statements.

- This statement generates an error if NAME is a character variable.

```
where name = . ;
```

- This statement generates an error if GNP is a numeric variable.

```
where gnp = "" ;
```

A numeric variable whose value is a special missing value (a-z or an underscore) is recognized as missing by the IS MISSING operator.

5.4. LIKE operator.

The LIKE operator is used to test whether a character variable contains a specified pattern. A pattern consists of any combination of valid characters and the following wild card characters.

- The percent sign (%) represents any number of characters (0 or more).

- The underscore (_) represents any single character.

The LIKE operator provides some of the functionality of the UNIX command *grep*. It is much more powerful than the SAS functions INDEX or INDEXC, which must be used multiple times to search for complex character patterns. The LIKE operator is case sensitive; upper and lower case characters are not equivalent.

Example 1. Select observations in which the variable NAME begins with the character 'J', is followed by 0 or more characters, and ends with the character 'n'. John, Jan, Jn, and Johanson are selected, but Jonas, JAN, and AJohn are not selected.

```
data two;
    set one ;
    where name like 'J%n' ;
    more SAS statements
```

Example 2. Select observations in which the variable NAME begins with the character 'J', is followed by any single character, and ends with the character 'n'. Jan is selected, but John, Jonas, Jn, Johanson, JAN, and AJohn are not selected.

```
data two;
    set one ;
    where name like 'J_n';
    more SAS statements
```

Example 3. Select observations in which the variable NAME contains the character 'J'. Jan, John, Jn, Johanson, Jonas, JAN, AJohn, TAJ, and J are selected, but j, Elijah, and jan are not selected.

```
data two ;
    set one ;
    where name like '%J%';
    more SAS statements
```

5.5. Sounds-like (=*) operator.

The Sounds-like (=*) operator uses the Soundex algorithm to test whether a character variable contains a spelling variation of a word. This operator can be used to perform edit checks on character data by checking for small typing mistakes, and will uncover some but not all of the mistakes.

In the following example, the WHERE statement keeps all but the last two input records.

```
data one;
    input bankname $1-20;
    length bankname $20;
    cards;
    new york bank
    new york bank.
    NEW YORK bank
    new york bnk
    new yrk bank
    ne york bank
    neww york bank
    nnew york bank
    ew york bank
    new york mets
    ;
run;
```

```
data two ;
    set one;
    where bankname=* 'new york bank' ;
run;
```

To identify cases where leading characters are omitted from a character variable, use the Sounds-like operator multiple times in a WHERE statement, removing one additional character in each clause of the WHERE statement, as follows.

```
where bankname=* 'new york bank'
or bankname=* 'ew york bank'
or bankname=* 'w york bank'
or bankname=* 'york bank' ;
```

5.6. SAME-AND operator.

The SAME-AND operator is used to add more clauses to a previous WHERE statement without reentering the original clauses. The SAME-AND operator need not be in the same PROC or DATA step as the previous WHERE statement. The SAME-AND operator reduces keystrokes during an interactive session, but can make programs harder to understand. It is not recommended for large programs or applications that have code stored in multiple files.

Example. The WHERE statement for data set FOUR selects observations in which GNP is greater than 100, INV is less than 10, and CON is less than 20.

```
data two ;
    set one ;
    where gnp > 100 and inv < 10 ;
    more SAS statements

data four ;
    set three ;
    where same-and con < 20 ;
    more SAS statements
```

The following statement is equivalent to the WHERE statement for data set FOUR.

```
where gnp > 100 and inv < 10 and con< 20 ;
```

Notes

1. See pages 498-504 in the "SAS Language Reference, Version 6, First Edition," for more information about WHERE statement operators.

6. Read a SAS data set, but need only a few of many variables

Task

In a DATA step, read a SAS data set with many variables to create a new SAS data set. Only a few of the variables are needed in the DATA step or the new SAS data set.

Technique

Use the KEEP= or DROP= option in the SET statement to prevent unnecessary variables from being read into the SAS program data vector (PDV) or the new data set.

Example

Create SAS data set TWO from SAS data set ONE. Data set ONE contains variables A and X1-X1000. The only variables needed in data set TWO are A and B.

Method 1, less efficient.

```
data two;
      set one ;
      b = a*1000 ;
run;
```

Method 2, less efficient.

```
data two;
      set one ;
      keep a b ;
      b = a*1000 ;
run;
```

Method 3, incorrect result.

```
data two (keep = b) ;
      set one (keep = a) ;
      b = a*1000 ;
run;
```

Method 4, more efficient.

```
data two ;
      set one (keep = a) ;
      b = a*1000 ;
run;
```

In method 1, 1001 variables are read into the PDV from data set ONE, and 1002 variables are written to data set TWO. 1000 of the variables are not needed.

In method 2, 1001 variables are read into the PDV from data set ONE, and two variables are written to data set TWO. The KEEP statement specifies that only A and B are written to data set TWO. Variables X1-X1000 are not written to data set TWO, but are still read unnecessarily into the PDV from data set ONE.

In method 3, one variable, A, is read into the PDV from data set ONE, and one variable, B, is written to data set TWO.

In method 4, one variable, A, is read into the PDV from data set ONE, and two variables, A and B, are written to data set TWO. This is the most efficient method.

Notes

1. If variables X1-X1000 are needed during the execution of DATA step TWO (for example, if they are used to calculate B), but are not needed in the output data set, then use method 2.

2. If A is the only variable needed during the execution of DATA step TWO, and B is the only variable needed in the output data set, then use method 3.

3. In method 2, either of the following statements are equivalent to KEEP A B ;.

```
drop x1-x1000;
data two(keep = a b);
```

4. In method 3, the following statement is equivalent to (KEEP = A);.

```
(drop = x1-x1000);
```

7. Concatenate (append) one SAS data set to another

Task

Create a new SAS data set containing all observations from two existing SAS data sets. The variables in the two data sets have the same length and type.

Technique

Use PROC APPEND instead of a SET statement. The SET statement reads both data sets. PROC APPEND reads the data set to be concatenated, but does not read the other data set, known as the BASE data set.

Example

Concatenate SAS data set TWO to SAS data set ONE.

Method 1, less efficient.

```
data one;
      set one two ;
run;
```

Method 2, more efficient.

```
proc append base = one data = two ;
run;
```

In method 1, the SAS system reads all observations in both data sets. In method 2, the SAS system reads only the observations in data set TWO.

Notes

1. Since PROC APPEND reads only the second data set, set BASE= to the larger data set if, as in the following example, the order of the data sets does not matter.

```
proc append base = one data = two ;
proc sort data = one ;
     by var1 ;
run ;
```

2. The documentation for PROC APPEND in the "SAS Procedures Guide, Version 6, Third Edition," describes how to concatenate data sets that contain different variables or variables with the same name but different lengths or types (character versus numeric).

8. Execute a DATA step, but no output SAS data set is needed

Task

Process a SAS data set in a DATA step when no output SAS data set is needed. This could occur when a DATA step is used to write reports with PUT statements, examine a data set's attributes, or generate macro variables with the CALL SYMPUT statement.

Technique

Name the data set the reserved name "_NULL_" to avoid creating an output SAS data set.

Example

Use PUT statements to write information from data set ONE. No output data set is needed.

Method 1, less efficient.

```
data two;
     set one;
     put @1 var1 @11 var2 ;
     more PUT and printing statements
run;
```

Method 2, less efficient.

```
data one;
     set one;
     put @1 var1 @11 var2 ;
     more PUT and printing statements
run;
```

Method 3, more efficient.

```
data _null_;
     set one;
     put @1 var1 @11 var2 ;
     more PUT and printing statements
run;
```

In method 1, the SAS system creates an unnecessary data set, TWO. In method 2, data set ONE is unnecessarily recreated. In method 3, an output data set is not created.

Notes

1. Using the statement DATA; instead of DATA _NULL_; causes the creation of an output data set called DATA*n*, where *n* is 1,2,3,... (the first such data set is called DATA1, the second is called DATA2, and so on).

9. Execute a SAS DATA step in which the denominator of a division operation could be zero

Task

Execute a SAS DATA step in which the denominator of a division operation could be zero.

Technique

Test the value of the denominator, and divide only if the denominator is not zero. Programs execute substantially faster if you prevent the SAS system from attempting to divide by zero.

Example

Execute a simple DATA step. Set the variable QUOTIENT to the value NUM/DENOM.

Input data set for this example.

```
data one;
     input num denom ;
     cards;
     2    1
     6    2
     10   0
     20   2
     9    0
     ;
```

```
run;
```

Method 1, less efficient.

```
data two;
    set one;
    quotient = num/denom;
run;
```

Method 2, more efficient.

```
data two;
    set one;
    if denom ne 0
        then quotient = num/denom;
        else quotient =.;
run;
```

Method 1 is less efficient because the SAS system attempts to divide by zero. Since attempting to divide by zero results in a missing value, both methods generate the same output data set, TWO, as follows.

OBS	NUM	DENOM	QUOTIENT
1	2	1	2
2	6	2	3
3	10	0	.
4	20	2	10
5	9	0	.

Notes

1. When the SAS system attempts to divide by zero, a note similar to the following is printed to the SAS log.

```
NOTE: Division by zero detected at line 329 column 11.
NUM=10 DENOM=0 QUOTIENT=. _ERROR_=1 _N_=3
NOTE: Division by zero detected at line 329 column 11.
NUM=9 DENOM=0 QUOTIENT=. _ERROR_=1 _N_=5
NOTE: Mathematical operations could not be performed at the
  following places.
      The results of the operations have been set to missing values.
Each place is given by: (Number of times) at (Line):(Column).
2 at 329:11
```

2. An example of a tremendous performance improvement from preventing the SAS system from attempting to divide by zero is as follows. In Release 6.06 of the SAS system for MVS at the Federal Reserve Board, a program processed a SAS data set with 200,000 observations and 16 variables, performing 8 divisions per observation, of which approximately half were attempts to divide by zero. Recoding the program to test for division by zero reduced the CPU time from 4 minutes to 7 seconds.

CONCLUSION

This paper presented simple efficiency techniques that can benefit inexperienced SAS software users on all platforms. Each section of the paper began with a description of an application or data set, followed by efficiency techniques for that application or data set. The author of this paper hopes that presenting the techniques this way will make it easier for users to determine when to apply these techniques.

For more information, contact

Bruce Gilsen
Federal Reserve Board, Mail Stop 171
Washington, DC 20551
202-452-2494
email: m1bfg00@frb.gov

REFERENCES

Polzin, Jeffrey A, (1994), "DATA Step Efficiency and Performance," in the Proceedings of the Nineteenth Annual SAS Users Group International Conference, 19, 1574 - 1580.

SAS Institute Inc. (1990), "SAS Language Reference, Version 6, First Edition," Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), "SAS Procedures Guide, Version 6, Third Edition," Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990), "SAS Programming Tips: A Guide to Efficient SAS Processing," Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The following people contributed extensively to the development of this paper: Donna Hill, Julia Meredith, and Steve Schacht at the Federal Reserve Board, Mike Bradicich at Vistech, and Peter Sorock. Their support is greatly appreciated.

TRADEMARK INFORMATION