Tips to Use Character String Functions in Record Lookup

Anjan Matlapudi
Pharmacy Informatics, PerformRx,
The Next Generation PBM, 200 Stevens Drive, Philadelphia, PA 19113

ABSTRACT
This paper gives you a better idea of how and where to use the record lookup functions to locate observations where a variable has some characteristic. Various related functions are illustrated to search numeric and character values in this process. Code is shown with time comparisons. I will discuss possible three ways to retrieve records using in SAS DATA step, PROC SQL and Perl regular expression. Real and CPU time processing issues will be highlighted when comparing to retrieve records using these methods.

Program was written for the PC using SAS 9.2 on Windows XP 62 bit environment. All the tools discussed are in BASE SAS®. The typical attendee or reader will have some experience in SAS, but not a lot of experience dealing with large number of data.

INTRODUCTION
A common task is to look for a particular field in databases ranging in size from less than a million to more than a billion. SAS Character functions are required to do this. I ask myself whether all of these functions would work in DATA step, PROC SQL and Perl Regular Expression. In this paper, I will introduce the code in several tables solving different problems for each function. In some cases PROC SQL is omitted because conditional use of the functions. This way many of us would be aware of how best we can use some of the commonly used functions in several ways.

To make the tests I used the following code to randomly generate 10 million observations with character and numeric variables using RANUNI function. DO LOOP =1 to 10e7 will generate 10 million records, but this code can generate any number of records by changing log exponential. I tested most of the functions to find one or many records out of the 10 million and noted processing time it takes to read, modify and write character variables.

CREATE TEST RECORDS USING RANUNI FUNCTION:

*----Test data-----*;
data TenMillionRecs;
  format FirstName LastName $10. MiddleName $1. Name $25.  SSN 9.;
  string1="abcdefghijklmnopqrstuvwxyz";
  string2="ABCDEFGHIJKLMNOPQRSTUVWXYZ";
  string3=reverse("abcdefghijklmnopqrstuvwxyz");
  strnglen=length(string1);
  do i=1 to 1e7;
      random=ranuni(95959);
      FirstName =''; LastName = '' ; MiddleName = '';
      length=int(ranuni(0)*3)+8; *int truncate decimal point ;
      do j=1 to length;
          pick=int(ranuni(0)*strnglen)+1;
          FirstName=substr(string1,pick,1)||FirstName;
          MiddleName = substr(string2,1,1);
          LastName=substr(string3,pick,1)||LastName;
          Name = propcase(FirstName||" "||MiddleName||" "||LastName);
          SSN = input(compress(put(1e9*random,z9.)),9.);
      end;
      output;

DATA STEP AND PROC SQL:

SAS® is an excellent tool to accommodate many functions in many ways; it has flexibility to use these functions in SAS for programmers with different sets of skills. SAS implemented SQL (PROC SQL) in version 6.0 and Perl Script in version 9.0 for better flexibility to retrieve information. In this paper, I will touch base some of the functions for records look up and also highlight the real and CPU (Central Processing Unit) time taken to run each functions in the same environment using DATA step and PROC SQL. Most of the function names itself describe the role of each function; however I highlighted purpose of each function in the tables. If you want to know more details, you can always approach SAS help or online documentation.

The code below shows some of the highlighted functions in DATA step and PROC SQL and next column for each function represent value return while running each function real and CPU time in seconds is noted for each processing. All these functions are tested on same environment, each statement run several times for processing the real time comparison.

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>RETURNS</th>
<th>REALTIME/CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIKE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose: Search option for specific string of characters.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data Like out;</td>
<td>Anjamvfk</td>
<td>A</td>
</tr>
<tr>
<td>set TenMillonRecs;</td>
<td>Zmqzneup</td>
<td></td>
</tr>
<tr>
<td>where name like 'Anjam%';</td>
<td>440019853</td>
<td></td>
</tr>
<tr>
<td>run;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proc sql;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>create table Like_tbl as</td>
<td></td>
<td></td>
</tr>
<tr>
<td>select * from TenMillonRecs</td>
<td>Anjamvfk</td>
<td>A</td>
</tr>
<tr>
<td>where name like 'Anjam%';</td>
<td>Zmqzneup</td>
<td></td>
</tr>
<tr>
<td>quit;</td>
<td>440019853</td>
<td></td>
</tr>
<tr>
<td>SUBSTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose: Extracts part of the string specified by the start and length parameters.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data SubStr_out;</td>
<td>Anjamvfk</td>
<td>A</td>
</tr>
<tr>
<td>set TenMillonRecs;</td>
<td>Zmqzneup</td>
<td></td>
</tr>
<tr>
<td>where substr(name,1,5)= 'Anjam';</td>
<td>440019853</td>
<td></td>
</tr>
<tr>
<td>run;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proc sql;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>create table SubStr_tbl as</td>
<td></td>
<td></td>
</tr>
<tr>
<td>select * from TenMillonRecs</td>
<td>Anjamvfk</td>
<td>A</td>
</tr>
<tr>
<td>where substr(name,1,5)= 'Anjam';</td>
<td>Zmqzneup</td>
<td></td>
</tr>
<tr>
<td>quit;</td>
<td>440019853</td>
<td></td>
</tr>
</tbody>
</table>
FIND

**Purpose:** Locate substring with in a string.

``` Sas
data Find_out;
set TenMillonRecs;
where find(name,'Anjamvfk   A
Zmqzneup')=1;
run;
```

```
Anjamvfk   A
Zmqzneup
440019853
3.93/2.29
```

``` Sas
proc sql;
create table find_tbl as
select * from TenMillonRecs
where find(name,'Anjamvfk   A
Zmqzneup')=1;
quit;
```

```
Anjamvfk   A
Zmqzneup
440019853
4.87/2.45
```

INDEX

**Purpose:** To locate starting portion of substring of a string.

``` Sas
data Index_out;
set TenMillonRecs;
where index(name,'Anjam')=1;
run;
```

```
Anjamvfk   A
Zmqzneup
440019853
4.75/2.53
```

``` Sas
proc sql;
create table Index_tbl as
select * from TenMillonRecs
where index(name,'Anjam')=1;
quit;
```

```
Anjamvfk   A
Zmqzneup
440019853
5.01/2.64
```

SCAN

**Purpose:** Extracts specified word from a character string.

``` Sas
data Scan_out;
length FirstName LastName $10.
MiddleName $1.;
set TenMillonRecs;
FirstName = scan(name,1,’’);
MiddleName = scan(name,2,’’);
LastName = scan(name,3,’’);
/*where name ='Anjamvfk   A
Zmqzneup';*/
run;
```

```
10000000 observations and 5 variables
32.12/6.71
```

``` Sas
proc sql;
create table Scan_tbl as
select name,ssn,
scan(name,1,’ ‘) as firstName,
scan(name,3,’ ‘) as MiddleName,
scan(name,2,’ ‘) as LastName
from TenMillonRecs
/*'where name ='Anjamvfk   A
Zmqzneup';*/
quit;
```

```
10000000 rows and 5 columns.
4.06.03/30.92
```

1 We can test all statements to read, modify and write one record using where clause option or we can retrieve 10 million records by removing where clause
**TRANSLATE**

**Purpose:** To exchange on character value to another.

```plaintext
data translate_out (keep=result);
set TenMillionRecs;
result=
  translate(name,'Anjan','Anjamvfk');  
Anjan   A
  ZmqzneupA
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sql
proc sql;
create table translate_tbl as
select translate(name,'Anjan','Anjamvfk') 
as result
from TenMillionRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

```plaintext
Anjan      A
ZnqzneupA 0.87/0.85
```

**CAT**

**Purpose:** Combine two strings.

```plaintext
data Cat_out (keep=result);
set TenMillionRecs;
result=
cat(name,'is' ,' funny name');
Anjamvfk   A
Zmqzneup is funny name
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sql
proc sql;
create table Cat_tbl as
select cat(name,'is' ,' funny name') 
as result
from TenMillionRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

```plaintext
Anjamvfk   A
Zmqzneup    is funny name
0.87/0.87
```

**VERIFY**

**Purpose:** Returns the position of the first character in a string that is not in any of several other strings.

```plaintext
data Verify_out (keep=result);
set TenMillionRecs;
result= verify(name,'Anjam');
6 1.30/0.95
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sql
proc sql;
create table Verify_tbl as
select verify(name,'Anjam') as result
from TenMillionRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

```plaintext
6 1.49/0.92
```

---

Coders’ Corner

NESUG 2011
TRIM

**Purpose:** Removes trailing blanks from a character string.

```sas
data trim_out;
set TenMillonRecs;
name= trim(name);
run;

proc sql;
create table trim_tbl as
select trim(name) as name, ssn
from TenMillonRecs;
quit;
```

10,000,000 observations and 2 variables

---

STRIP

**Purpose:** To strip leading or trailing blanks from character string.

```sas
data strip_out(keep=result);
set TenMillonRecs;
if name= 'Anjamvfk   A Zmqzneup'
then
name1 = '          Anjamvfk';
result = strip(name1);
where name = 'Anjamvfk   A Zmqzneup';
run;

Proc SQL is omitted.
```

---

RIGHT

**Purpose:** Align right side of character string.

```sas
data Right_out(keep= name1 result);
set TenMillonRecs;
if name= 'Anjamvfk   A Zmqzneup'
then
name1 = '               Anjamvf';
result = right(name1);
where name = 'Anjamvfk   A Zmqzneup';
run;

Proc SQL is omitted.
```

---

LEFT

**Purpose:** Align left side of character string.

```sas
data Left_out(keep= name1 result);
set TenMillonRecs;
if name= 'Anjamvfk   A Zmqzneup'
then
name1 = '               Anjamvf';
result = left(name1);
where name = 'Anjamvfk   A Zmqzneup';
run;
proc print data =left_out;
run;

Proc SQL is omitted.
```
### COMPRESS
**Purpose:** Remove specified character value (including blanks) from a string.

```sas
data compress out;
set TenMillonRecs;
name= compress(name);
run;
```

1000000 observations and 2 variables

```sas
proc sql;
create table compress_tbl as
select compress(name) as name, ssn
from TenMillonRecs;
quit;
```

1000000 rows and 2 columns

### COMPBL
**Purpose:** Replace two or more blanks with single blank.

```sas
data Compbl_out;
length result $100.;
set TenMillonRecs;
result= compbl(name);
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sas
proc sql;
create table Compbl_tbl as
select compbl(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

### UPCASE
**Purpose:** Convert all letters to upper case.

```sas
data Upcause_out(keep=result);
set TenMillonRecs;
result= upcase(name);
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sas
proc sql;
create table Upcause_tbl as
select upcase(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

### LOWCASE
**Purpose:** Converts all the letters to lower case.

```sas
data Lowcause_out(keep=result);
set TenMillonRecs;
result= lowcase(name);
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```sas
proc sql;
create table Lowcause_tbl as
select lowcase(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```
proc sql;
create table Lowcause_tbl as
select lowcase(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;

PROPCASE

**Purpose:** Capitalize the first letter of each word in a string.

data Propcase_out(keep=result);
set TenMillonRecs;
result = propcase(name); Anjamvfk A 0.92/0.89
where name = 'Anjamvfk   A Zmqzneup';
run;

proc sql;
create table Propcase_tbl as
select propcase(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;

ANYSPACE

**Purpose:** To locate first occurrence of white space.

data Anyspace_out (keep=result);
set TenMillonRecs;
result = anyspace(name); 9 0.87/0.87
where name = 'Anjamvfk   A Zmqzneup';
run;

proc sql;
create table Anyspace_tbl as
select anyspace(name) as result
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;

FIRST

**Purpose:** Extracts the first character from a string.

data First_out(keep=result);
set TenMillonRecs;
result = first(name); A 0.90/0.89
where name = 'Anjamvfk   A Zmqzneup';
run;

proc sql;
create table First_tbl as
select first(name) as name, ssn
from TenMillonRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
**TRANWRD**

**Purpose:** Substitute one or more words in a string.

data Tranwrd_out (keep=result);
  set TenMillionRecs;
  result = tranwrd(name,'Anjamvfk A Zmqzneup','Anjan Matlapudi');
  where name = 'Anjamvfk A Zmqzneup';
run;

proc sql;
  create table Tranwrd_tbl as
  select tranwrd(name,'Anjamvfk A Zmqzneup','Anjan Matlapudi') as result
  from TenMillionRecs
  where name = 'Anjamvfk A Zmqzneup';
quit;

**LENGTH**

**Purpose:** Determine length of a character value (not counting trailing blanks).

data Length_out;
  set TenMillionRecs;
  result = length(name);
  where name = 'Anjamvfk A Zmqzneup';
run;

proc sql;
  create table Length_tbl as
  select name,ssn,length(name) as result
  from TenMillionRecs
  where name = 'Anjamvfk A Zmqzneup';
quit;

**REVERSE**

**Purpose:** Reverse the order of character string.

data Reverse_out(keep=result);
  set TenMillionRecs;
  result = reverse(name);
  where name = 'Anjamvfk A Zmqzneup';
run;

proc sql;
  create table Reverse_tbl as
  select reverse(name) as result
  from TenMillionRecs
  where name = 'Anjamvfk A Zmqzneup';
quit;
**REPEAT**

**Purpose:** Make several copies of a string.

```plaintext
data Repeat_out(keep=result);
set TenMillionRecs;
result = repeat(name,3);
where name = 'Anjamvfk   A Zmqzneup';
run;
```

```plaintext
proc sql;
create table Repeat_tbl as
select repeat(name,3) as result
from TenMillionRecs
where name = 'Anjamvfk   A Zmqzneup';
quit;
```

**SPEDIS**

**Purpose:** Computes spelling distance between words.

```plaintext
data Spedis_out;
set TenMillionRecs;
if name = 'Anjamvfk   A Zmqzneup'
then
Name = 'Knowledge';
if spedis(name,'nowledge') le 27
then output;
run;
```

**LAG**

**Purpose:** To obtain previous value from the current character variable.

```plaintext
data Lag_out(keep=result);
set TenMillionRecs;
result = lag(name);
run;
```

Note: LAG does not work in Proc SQL.

**COUNT**

**Purpose:** Counts number of times in a given substring in a string.

```plaintext
data Count_out(keep=result);
length name $100.;
set TenMillionRecs;
if name = 'Anjamvfk   A Zmqzneup'
then
name = 'Random generated Random Name in Random data';
result = count(Name,'Random')
where name = 'Anjamvfk   A Zmqzneup';
```
Proc SQL is omitted.

**CHOOSEC**

**Purpose:** Returns a character value that represents the results of choosing from a list of arguments.

data Choosec_out(keep=result);
set TenMillionRecs;
if name= 'Anjamvfk A Zmqzneup'
then
name1 = 'Anjan';
else if name = 'Kxlgoqhma A Pcotljsnz' then name2 = 'Matlapudi';
else if name = 'Taosezswy A Gzlhvahdb' then name3 = 'Anjan Matlapudi';
result = Choosec(3,name1,name2,name3) ;
where name in('Anjamvfk A Zmqzneup','Kxlgoqhma A Pcotljsnz','Taosezswy A Gzlhvahdb');
run;

Proc SQL is omitted.

**PERL REGULAR EXPERSION:**

Perl Regular Expression in SAS has wide variety of functionality while working with matching patterns, text manipulation including validation and text replacement. Each variable holds 1-32676 bytes long as a charter string. I have tested some of the character string functions using 32676 bytes long character variable and I can able to successfully returned values of each function (data not included). PRXPARSE and other functions have great deal while we are working with text manipulation. Below mentioned prx functions show how to retrieve records in DATA step and PROC step.

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>RETURNS</th>
<th>REALTIME/CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRXMATCH</td>
<td>PRXMATCH</td>
<td></td>
</tr>
<tr>
<td><strong>Purpose:</strong></td>
<td>Searches for a pattern match and returns the position at which the pattern is found.</td>
<td></td>
</tr>
</tbody>
</table>

data PrxMatch_out;
  set TenMillionRecs;
  if prxmatch("m/Anjamvfk/oi",name)> 0
  then found =1;
    else found =0;
run;

Proc sql;
create table PrxMatch_tbl as
  select name,ssn,
    prxmatch("m/Anjamvfk/oi",name) as found
  from TenMillionRecs;
quit;
**PRXPARSE AND PRXSUBSTR**

Purpose: Perl regular expression (PRX) can be used for substring of character string matching using PRXPARSE.

```plaintext
data PrxSubstr_out;
set TenMillonRecs;
if _n_=1 then do;
   retain re;
   re = prxparse('/\w \w.+/');
   if missing(re) then do;
      stop;
   end;
   end;
   call prxsubstr(re,name,start,length);
   FirstName = substrn(name,start -11, length -4);
   LastName = substrn(name,start+2, length +1);
   if start > 0 then output;
run;
```

**RECORD LOOKUP BY CHARACTER STRING vs. NUMBER FIELD:**

I further demonstrated records retrieval using character and number fields as mentioned below. Real time and CPU time is noted to read one record out of 10 million.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>RETURNS</th>
<th>REALTIME/CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CHARACTER</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proc print data = TenMillonRecs; where name = 'Anjamvfk   A Zmqzneup'; run;</td>
<td>Anjamvfk    A Zmqzneup</td>
<td>440019853</td>
</tr>
<tr>
<td>proc sql; select * from TenMillonRecs where name = 'Anjamvfk   A Zmqzneup'; quit;</td>
<td></td>
<td>440019853</td>
</tr>
<tr>
<td><strong>NUMERIC</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proc print data = TenMillonRecs; where ssn =440019853; run;</td>
<td>Anjamvfk    A Zmqzneup</td>
<td>440019853</td>
</tr>
<tr>
<td>proc sql; select * from TenMillonRecs where ssn =440019853; quit;</td>
<td></td>
<td>440019853</td>
</tr>
</tbody>
</table>
CONCLUSION

So far I have demonstrated possible ways for a record look up using DATA step, PROC SQL and Perl regular expression. I further generalize some tips using these functions for record look

- If you already know SQL, you will be pleased to know that you can use most of the functions in PROC SQL to create, read and modify variables in SAS data sets.
- Flexibility of character string functions available while we are working with small to a large scale data. We can use all these functions in DATA step and PROC SQL. All these functions would also work in Oracle, SQL server and Access database with slight syntax modification in code.
- Real time processing will give best suitable option to choose some of the functions. Based on the above queries, data step processing has been taken less time when compare PROC step.

By now you may have some clue that all most all the functions work in DATA step and PROC SQL and if you are dealing with large scale data, you can easily pick up best suitable function in terms of time taken to run each function. Also, I have included most commonly used function at one place; some of you may take advantage instead of spending more time finding them.

REFERENCES


ACKNOWLEDGMENTS

I like to acknowledge Ian Whitlock for his valuable suggestions. I would also like to acknowledge Mr. Shimels Afework, Senior Director, PerformRx.

PerformRx, LLC provides pharmacy benefit management (PBM) services through proactively managing escalating pharmacy costs while focusing on clinical improvement and financial results.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. Contact the author at

Name Anjan Matlapudi
Senior Pharmacy Analyst
Address PerformRx, The Next Generation PBM
200 Stevens Drive
Philadelphia, PA 19113
Work Phone: (215)937-7252
Fax: (215)863-5100
E-mail: anjan.matlapudi@performrx.com
anjanmat@gmail.com

SAS® is a registered trademark or trademark of SAS® Institute, Inc. in the USA and other countries. IBM, OS/390, and MVS are registered trademarks of International Business Machines Inc.