PROC SQL

Is it possible to perform a merge where key values are split across two columns?

Normally when performing a horizontal join of datasets, or performing a lookup, the key values exist in a single column, however sometimes 'dirty' data may appear in different forms, and in different columns.

Submit the following code to generate a sample dataset and a lookup table:

```
*** Generate some test data which may have one field, or the other, or
 both *** ;
data ds1 (drop = gender)
     ds2 (drop = sex)
     ;
  set sashelp.class (keep = name sex) ;
    if (sex = 'M' and ranuni(0) > 0.6) then
      do ;
        gender = '1' ;
        if ranuni(0) > 0.4 then sex = '' ;
      end ;
 if (sex = 'F' and ranuni(0) > 0.7) then
      do ;
        gender = '2' ;
 if ranuni(0) > 0.4 then  sex = '' ;
      end ;
run ;

data ds3 ;
  merge ds1 ds2 ;
    by name ;
run ;

*** Generate a lookup table with all data values *** ;
data gen_lookup ;
  gen_sex = 'F' ;
  gen_type = 'Female' ;
  output ;
  gen_sex = '2' ;
  gen_type = 'Female' ;
  output ;
  gen_sex = 'M' ;
  gen_type = 'Male' ;
  output ;
  gen_sex = '1' ;
  gen_type = 'Male' ;
  output ;
run ;
```

```
Name       Sex    gender

Alfred      M       1
Alice               2
Barbara     F
Carol       F
Henry       M       1
James       M
Jane        F
Janet       F
Jeffrey     M
John        M       1
Joyce       F
Judy        F
Louise      F
Mary        F       2
Philip      M
Robert              1
Ronald      M
Thomas              1
William     M
```

There are two columns containing the 'key' values - sex and gender. In the generated dataset values may in appear in one, or the other, or both columns. Although values differ, the relationship remains constant:

M = 1 = Male

F = 2 = Female

In order to perform the lookup, the key value is the first non-missing value from either sex or gender.

While it would be possible to consolidate this data into a single column, this would require a further pass through the data, and create an inefficiency.

To perform the lookup on the 'best' value, simply use the COALESCE function on the join condition in an SQL clause:

```
*** Join the lookup table by a combination of the sex and gender colum
ns *** ;
proc sql ;
  create table ds4 as
  select  ds3.*
```

```
        ,gen_type
  from  ds3
       ,gen_lookup
  where coalesce(gender, sex) = gen_sex
  ;
quit ;
```

The COALESCE function returns the first non-missing value from a list of arguments, and produces the desired result:

Name	Sex	gender	gen_type
Alfred	M		Male
Alice		2	Female
Barbara	F		Female
Carol	F		Female
Henry		1	Male
James		1	Male
Jane	F		Female
Janet	F		Female
Jeffrey	M	1	Male
John		1	Male
Joyce	F		Female
Judy		2	Female
Louise	F		Female
Mary	F	2	Female
Philip		1	Male
Robert	M		Male
Ronald		1	Male
Thomas		1	Male
William		1	Male

Unique solution ID: #1028
Author: Alan D Rudland
Last update: 2017-05-25 11:13