A MORE 'INTELLIGENT' MERGE CAPABILITY IN SAS

Max E. Ellis
Riker Labs (218-3)
3M Company
St. Paul, MN
(612) 733-4464

ABSTRACT

Existing data manipulation functions of SAS do not provide the user with a complete set of functions required for research data management. Certain logical, data-set combining operations using the SORT and MERGE functions require excessive programming, and in some cases, extra passes of a data set(s) must be made to produce the resultant composite file.

A hierarchical logical file structure is used to illustrate how a user would combine physical SAS data sets to form output data sets corresponding to various analysis levels of the hierarchy. Household, family, person and work data sets are combined via 1) existing SAS program statements and 2) proposed new statements. By enhancing the data manipulation language of SAS, the utility of SAS can be increased and the programming now required to combine logically complex, related data sets is reduced.

INTRODUCTION

A data base in SAS is of the relational type, i.e., is comprised of data sets conforming to the relational model of data as originally presented by E. F. Codd [1]. Basically each data set corresponds to a relation or a mathematical set of related data items represented as a two dimensional table. In SAS such tables are SAS data sets where the rows of the table correspond to a SAS observation and the columns of the table are SAS variables.

A relational table has the following properties [2]:

1) no two rows are identical
2) the ordering of rows is not significant; and
3) the ordering of columns is significant.

For data manipulations involving only one data set, SAS conforms to these relational properties. However, when two or more data sets are to be combined for deriving an analysis file, the ordering of rows can become a significant factor in creating a composite file correctly and efficiently. SAS requires that data sets to be combined be sorted by common key variables. In addition to the sort requirement, SAS also requires that BY variables of a MERGE be present in all data sets.

This paper presents the concept of a more intelligent merge function and proposes that SAS be enhanced to include a more versatile merge capability, one that better conforms to the relational model. By comparing SAS programs using the existing merge capabilities with programs using the simulated new merge function, it is hoped that the benefits of this new merge algorithm will be demonstrated.

The syntax and semantics of the proposed additional SAS parameters and new statements as presented should be viewed as more of a metalanguage or a language used to describe the enhanced merge capability. Only the SAS developers can best prescribe how to integrate this capability with the existing SAS user request language and other data management functions.

A Typical Research Data Base

Most data generated and analyzed in a research environment requires that varying amounts of different types of information be combined. Consider as an example survey data. A most logical model for organizing data in the computer for this type of information is to create a SAS data set for logically related items from each section of the questionnaire. If punched cards are to be the input medium, then data from the PERSONAL INFORMATION section of the questionnaire might be punched as card types 1-3. The cards are stored in SAS as data set PERSONAL. Using this data organization model facilitates more efficient and accurate data entry and conforms most closely to the logical organization of the source document (questionnaire).

The very nature of surveys dictates that some questions may go unanswered and that entire sections of the questionnaire may not be applicable to the respondent. Therefore, not only variables used to be coded as missing but entire observations of a data set may be missing. It is the latter occurrence that presents problems for the SAS user.

Figure 1 illustrates an oversimplified questionnaire used in a survey of households. Section H is completed for every household, whether persons were home or not. Similarly, sections F, P and W are completed if the information is attainable and relevant. Figure 2 lists the data collected from the survey. Note that household, family and person identification variables have been added to enable merging data from the data sets H, F, P and W. A logical data base diagram, Figure 3, illustrates the logical paths or links possibly given this organization. The number of variables used to link the data sets, which of these variables are coded is a function of both the questions to be answered from the data and the software that will be used to manipulate and analyze the data.
The researcher collects the data assigning unique household numbers (HHNO) to observations in the household data set H, family numbers (FAMNO not appearing on the questionnaire) to family records (data set F) and person numbers to person records (data sets P and W). Such a scheme facilitates families to be uniquely identified by the variables HHNO and FAMNO and person information from data sets P and W to be uniquely identified by the variables HHNO, FAMNO and PERNO.

**Combining SAS DATA Sets**

Obviously questions that can be answered by querying only one of the data sets present no problems for the SAS user. However, if two or more data sets must be combined to create an analysis file, the SAS user must provide some thought and perhaps some clever programming to obtain the desired result efficiently.

Suppose that the question to be answered from the data base was:

"What does each working person contribute toward his/her household income and family income?"

Having asked this question the researcher decides to produce a report listing working persons' HHNO, FAMNO, PERNO, INCOME, HHINC, FAMINC, HHPER and FAMPER; where, HHPER is the percent of household income (INCOME * 100 / HHINC) and FAMPER is the percent of family income (INCOME * 100 / FAMINC). A solution using SAS is:

```sas
PROC SORT DATA = H; BY HHNO;
PROC SORT DATA = F; BY HHNO FAMNO;
PROC SORT DATA = W; BY HHNO FAMNO;
DATA F_W;
MERGE F W (IN = WORK) ; BY HHNO FAMNO;
IF WORK THEN HHINC = HHINC;
IF FAMILY THEN FAMINC = FAMINC;
IF WORK THEN GO TO OUTPUT;
RETAIN HHINC FAMINC;
RETURN;

OUTPUT: HHPER = INCOME * 100 / HHINC;
FAMPER = INCOME * 100 / FAMINC;
PUT HHNO = FAMNO = PERNO = INCOME =
   HHPER = FAMPER = ;
RETURN;
```

This solution requires only one pass of each sorted data set to perform the merge. However, more programming logic is required because the variables HHINC and FAMINC must be created so that the household and family incomes can be compared to each person's INCOME at the time a work (W) record is read during the merge process.

From these two sample solutions of a data set combining operation, one can extrapolate the degree of complex logic and the extra data manipulation that may be required to produce an analysis data set from many SAS data sets containing many "lagged" variables. Because of the SAS restriction of common BY variables in a merge and the explicit RETAIN statement that is required, the user can be overly burdened with excessive coding and increased chance for errors.

Compare these first two solutions with a more intelligent merge capability. By adding additional parameters to the DATA statement and inventing the SETn statement, an implicit merge function can be implemented. This is illustrated by the following solution based on the original data sets shown in Figure 2. Proposed new parameters and statements are italicized.

```sas
PROC SORT DATA = H; BY HHNO;
PROC SORT DATA = F; BY HHNO FAMNO;
PROC SORT DATA = W; BY HHNO FAMNO;
DATA F_W;
MERGE H_F_W (IN = WORK) ; BY HHNO FAMNO;
IF WORK THEN HHINC = HHINC;
IF FAMILY THEN FAMINC = FAMINC;
IF WORK THEN GO TO OUTPUT;
RETAIN HHINC FAMINC;
RETURN;

OUTPUT: HHPER = INCOME * 100 / HHINC;
FAMPER = INCOME * 100 / FAMINC;
PUT HHNO = FAMNO = PERNO = INCOME =
   HHPER = FAMPER = ;
RETURN;
```

As in the other solutions the new data set, H_F_W, need not be named since only the PUT output is relevant, i.e., _NULL_ could have been used. Also assume that DROP or KEEP statements could have appeared in any of these solutions so as to conserve memory. The OUT = SET3 parameter tells SAS to automatically save variables from sets 1-2 until the FAMNO and/or HHNO from SET3 W changes. The SAS program statements are executed only when a work record has been read, at which time the
last RHINC and FAMINC saved are used to calculate HHPER and FAMPER, respectively.

The SETn statements should not be confused with the existing SAS SET statement. The number following SET has been used to provide this distinction and must be present and unique among all SETn statements.

Direct Access Capability

It can be shown that although the above solution is perhaps an improvement over existing SAS capabilities, it still does not provide all the merging functions required of a complete relational data base model. To illustrate this, consider the household survey and assume the researcher had decided to identify person and work records by the single, unique variable social security number (SSNO).

Ideally one would like to answer the question: "What are the names and sex of those persons who contribute more than 50% of the income to a household or family?" Using the following SAS program:

```sas
DATA NULL (OUT=SET1 ACCESS=DIRECT);
SET1 W; BY SSNO;
SET2 P; BY SSNO FAMNO;
SET3 F; BY FAMNO HHNO;
SET4 H; BY HHNO;
IF INCOME * 100 / RHINC > 50 OR INCOME * 100 / FAMINC > 50
THEN PUT NAME SEX INCOME; RETURN;
```

Unlike the previous example, a keyed direct access method must be used to perform the necessary data set combining operations. The BY variables listed are used to construct indices which point to record locations (if such indices were not constructed in a previous SAS request). Data is retrieved by beginning with SET1 and progressing to SET2, to SET3 and finally to SET4. Note that each new data set accessed is queried via BY values derived from the previous SET. If a particular BY value is not found on the data set, missing values are generated for variables from that data set.

SUMMARY

SAS has the potential of becoming a likely candidate for a system which bridges the gap between a pure data base management system and a statistical system. By extending the data management functions of SAS to conform more with the relational model of data base organization, a more useful data processing tool can be provided the researcher.

The concept of a more user-oriented, efficient and relational-based merge capability has been introduced. If such a function were implemented in SAS the following capabilities would be required:

1) unequal number of BY variables for relating data sets being merged;
2) optional automatic RETAINing of variables specified by designating which data set triggers the output of the merged observation;
3) a keyed direct access capability to facilitate more direct merging of data sets and alleviating redundant sorts.

Acknowledgements

The staff of the SAS Institute Inc. is congratulated on their fine design and implementation of SAS. SAS' unique capability for providing the user with a means of easily integrating complex routines with the system (while maintaining the integrity of the user request language) has enabled us to simulate an 'intelligent' merge function via a user PROC.

REFERENCES

Figure 1. Household Survey Questionnaire

Section H
Household Number (HHNO) ____________________________
Household Location (HHLOC) ____________________________
Total Annual Household Income (HHINC) ____________________________

Section F
Family Name (FAMNAME) ____________________________
LAST, INITIAL ____________________________
Number in Family (NOINFAM) ____________________________
Total Annual Family Income (FAMINC) ____________________________

Section F
Name of Person (NAME) ____________________________
LAST, FIRST ____________________________
Head of Household? (HEAD) Y, N ____________________________
Sex (SEX) M, F ____________________________
Age (AGE) ____________________________
Employed (WORK) Y, N ____________________________

Section W
Social Security # (SSNO) ____________________________
Annual Income ____________________________
Figure 2. Household Survey SAS Data Base

<table>
<thead>
<tr>
<th>HHNO</th>
<th>HHLOC</th>
<th>HHINC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LAS VEGAS</td>
<td>30500</td>
</tr>
<tr>
<td>2</td>
<td>NEW YORK</td>
<td>54000</td>
</tr>
<tr>
<td>3</td>
<td>ST. PAUL</td>
<td>15000</td>
</tr>
<tr>
<td>4</td>
<td>MADISON</td>
<td>(MISSING)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HHNO</th>
<th>FAMNO</th>
<th>FAMNAME</th>
<th>NOINFAM</th>
<th>FAMINC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>HUGHES, H</td>
<td>5</td>
<td>11500</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>HUGHES, P</td>
<td>2</td>
<td>3000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>JONES</td>
<td>3</td>
<td>16600</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>STEIN</td>
<td>2</td>
<td>54000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>HANSON</td>
<td>1</td>
<td>15000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>LUCEY</td>
<td>4</td>
<td>(MISSING)</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>SMITH</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HHNO</th>
<th>FAMNO</th>
<th>PERSON</th>
<th>NAME</th>
<th>HEAD</th>
<th>SEX</th>
<th>AGE</th>
<th>WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>HUGHES, H</td>
<td>Y</td>
<td>M</td>
<td>52</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>HUGHES, A</td>
<td>N</td>
<td>F</td>
<td>49</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>HUGHES, S</td>
<td>N</td>
<td>M</td>
<td>18</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>HUGHES, E</td>
<td>N</td>
<td>M</td>
<td>16</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>HUGHES, T</td>
<td>N</td>
<td>M</td>
<td>16</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6</td>
<td>HUGHES, P</td>
<td>N</td>
<td>M</td>
<td>74</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7</td>
<td>HUGHES, G</td>
<td>N</td>
<td>F</td>
<td>74</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
<td>JONES, A</td>
<td>N</td>
<td>M</td>
<td>21</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>JONES, A</td>
<td>N</td>
<td>F</td>
<td>19</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>JONES, S</td>
<td>N</td>
<td>F</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>STEIN, V</td>
<td>Y</td>
<td>F</td>
<td>39</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>STEIN, W</td>
<td>Y</td>
<td>F</td>
<td>40</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>HANSON, B</td>
<td>Y</td>
<td>M</td>
<td>25</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>LUCEY, E</td>
<td>N</td>
<td>M</td>
<td>14</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>LUCEY, E</td>
<td>N</td>
<td>F</td>
<td>11</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>LUCEY, F</td>
<td>Y</td>
<td>M</td>
<td>42</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>LUCEY, T</td>
<td>N</td>
<td>F</td>
<td>45</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>5</td>
<td>SMITH, J</td>
<td>N</td>
<td>F</td>
<td>68</td>
<td>N</td>
</tr>
</tbody>
</table>
Figure 2. Household Survey SAS Data Base (con't)

<table>
<thead>
<tr>
<th>HHNO</th>
<th>FAMNO</th>
<th>PERNO</th>
<th>SSNO</th>
<th>INCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>111</td>
<td>11500</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>666</td>
<td>3000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>888</td>
<td>8000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>999</td>
<td>8000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1010</td>
<td>30000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1111</td>
<td>24000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1121</td>
<td>15000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1414</td>
<td>(MISSING)</td>
</tr>
</tbody>
</table>

Figure 3. Household Survey Data Base Diagram