TF19

SQL, HASH Tables, FORMAT and KEY= — More Than One Way to Merge Two Datasets

David Franklin

TheProgrammers Cabin.com

Introduction

- Merging data is one of the most common data manipulation task done with data
- This paper looks at four common methods
- No one method is better than the other
- Looking at only a one-to-one or one-to-many merge -- while some methods can be adapted to do a many-to-many merge, this is beyond the scope of the paper

First some data...

Dataset: PATDATA

SUBJECT	TRT_CODE
001	A
002	A
003	B
004	B

Dataset: ADVERSE

SUBJECT	EVENT
003	FEVER
002	FRACTURE
001	HEADACHE
005	FRACTURE
003	NAUSEA

- SUBJECTs 001 and 004 are not represented in dataset ADVERSE, SUBJECT 003 has multiple ADVERSE records, and SUBJECT 005 is not in dataset PATDATA.

The Most Common Way ...

```
PROC SORT DATA=patdata;          *SAS Program;
  BY subject;
PROC SORT DATA=adverse;
  BY subject;
DATA alldata;
  MERGE patdata adverse;
  BY subject;
RUN;
```

- It is the method with the most control.

Dataset Indexes

- Instead of using PROC SORT before calling the datastep, efficiency can be better is the PROC SORT calls were replaced by creating an index of the data.

```
PROC DATASETS LIBRARY=WORK NOLIST NODETAILS;
  MODIFY patdata;
     INDEX CREATE subject /UNIQUE;
  MODIFY adverse;
     INDEX CREATE subject:
QUIT;
DATA alldata;
  MERGE patdata adverse;
  BY subject;
RUN;
```

PROC SQL

- SQL is an standard industry language for database manipulation that has been around in SAS since SAS version 6.06.

```
PROC SQL;
  CREATE TABLE alldata AS
  SELECT a.*, b.trt_code
  FROM adverse a OUTER UNION JOIN patdata b
  ON a.subject=b.subject;
QUIT;
```

PROC SQL, _METHOD

- Hard to find documentation
- When using PROC SQL and the _METHOD option, SAS will show in the SAS LOG what it is doing
- Table below shows four codes that relate to merging data:

_METHOD Code	**Description**
sqxjsl	Step Loop Join (Cartesian)
sqxjm	Merge Join
sqxjndx	Index Join
Sqxjhsh	HASH Join

PROC SQL, and some MAGIC (1)

- The most common of joins within SQL is the SQXJM (MERGE) join which will usually sort the data then merge
- There is a limited (hard to find) way that you can tell SQL how to do the join though the undocumented MAGIC= option

  PROC SQL _METHOD MAGIC=101;
  Step loop join, when an equality condition is not specified, a read of the complete contents of the right table is processed for each row in the left table.

PROC SQL, and some MAGIC (2)

PROC SQL _METHOD MAGIC=102;
Merge join, when the tables specified are already in the desired sort order, resources will not need to be extended to rearranging the tables.

PROC SQL _METHOD MAGIC=103;
Hash join, when an equality relationship exists, the smaller of the tables is able to fit in memory, no sort operations are required, and each table is read only once.

Hash Tables

- First appearing in SAS version 9.1, and used by database programmers in other languages, this is considered one of the fastest ways to merge data in two datasets.

```
DATA alldata0;
  IF _n_=0 THEN SET patdata;
  IF _n_=1 THEN DO;
     DECLARE HASH _h1 (dataset: "PATDATA");
     rc=_h1.definekey("SUBJECT");
     rc=_h1.definedata("TRT_CODE");
     rc=_h1.definedone();
     call missing(SUBJECT,TRT_CODE);
  END;
  SET adverse;
  rc=_h1.find(); IF rc^=0 THEN trt_code=" "; DROP rc;
```

PROC FORMAT

- It is possible to create a format from the dataset that has unique observations, in this case the PATDATA dataset, and the TRT_CODE variable as the label

```
DATA fmt;
  RETAIN fmtname 'TRT_FMT' type 'C';
  SET patdata;
  RENAME subject=start trt_code=label;
PROC FORMAT CNTLIN=fmt;
DATA alldata0;
  SET adverse;
  ATTRIB trt_code LENGTH=$1 LABEL='Treatment Code';
  trt_code=PUT(subject,$trt_fmt.);
RUN;
```

SET KEY=

- Many options have been added to the SET statement, one of them being the KEY= option

```
DATA alldata0;
  SET adverse;
  SET patdata KEY=subject /UNIQUE;
  DO;
     IF _IORC_ THEN DO;
        _ERROR_=0;
        trt_code='';
     END;
  END;
RUN;
```

- Requires the dataset PATDATA to have an INDEX associated with it.

Just Because You Are Here …

Which is faster? This is only a **GUIDE**

	Number of Unique Subjects					
Method	**1k**	**10k**	**50k**	**100k**	**500k**	**1M**
MERGE Statement	1.00	1.00	1.00	1.00	1.00	1.00
SQL	1.71	0.29	0.25	0.26	0.26	0.26
HASH	0.43	0.10	0.12	0.15	0.21	0.23
PROC FORMAT	0.43	0.12	0.13	0.16	0.25	0.28
SET KEY=	2.43	2.05	1.95	2.34	2.56	2.51

Notes:
1. PATDATA is merged with ADVERSE - ration is 1:10
2. Number given is ratio against MERGE statement

Conclusion

- There is more than the MERGE statement to combine two datasets
- No one method is better than another, but the most control is through the MERGE statement
- Try the different methods shown here back at your own site

Contact Information

David Franklin
The Programmers Cabin.com
603-275-6809
dfranklin@TheProgrammersCabin.com
http://www.TheProgrammersCabin.com
@ThePgmrsCabin