Paper RF-01-2015

Essential DATA Step Merge Techniques Using SAS® University Edition Software

Kirk Paul Lafler, Software Intelligence Corporation, Spring Valley, California
Charles Edwin Shipp, Consider Consulting Corporation, San Pedro, California

Abstract

After installing SAS Institute's free "SAS University Edition" you'll want to test drive the software. SAS University Edition includes Base SAS, SAS/STAT, SAS/IML, Designer Studio (user interface), and SAS/ACCESS for Windows, with all the powerful features found in the licensed SAS versions. To demonstrate the power found within SAS University Edition, we present conventional and unconventional DATA step merge programming techniques using Base SAS software. All SAS users are encouraged to attend and learn essential concepts, syntax and programming techniques.

Introduction

A powerful and essential DATA step programming technique that all SAS users should be aware of, and comfortable performing, is the process of merging (or combining) two or more datasets. The intent of this paper is to describe the merge process, including what a merge is, the preparation requirements for each dataset being specified in a merge, the merge syntax, and the various types of merges available to SAS users.

Example Datasets

The examples used throughout this paper utilizes two datasets. (A relational database is a collection of datasets/tables.) The data used in all the examples in this paper consists of a selection of movies that I've viewed over the years. The Movies dataset consists of six variables: title, length, category, year, studio, and rating. Title, category, studio, and rating are defined as character variables with length and year being defined as numeric variables. The data stored in the Movies dataset is depicted below.

MOVIES Dataset

	Title	Length	Category	Year	Studio	Rating
1	Brave Heart	177	Action Adventure	1995	Paramount Pictures	R
2	Casablanca	103	Drama	1942	MGM / UA	PG
3	Christmas Vacation	97	Comedy	1989	Warner Brothers	PG-13
4	Coming to America	116	Comedy	1988	Paramount Pictures	R
5	Dracula	130	Horror	1993	Columbia TriStar	R
6	Dressed to Kill	105	Drama Mysteries	1980	Filmways Pictures	R
7	Forrest Gump	142	Drama	1994	Paramount Pictures	PG-13
8	Ghost	127	Drama Romance	1990	Paramount Pictures	PG-13
9	Jaws	125	Action Adventure	1975	Universal Studios	PG
10	Jurassic Park	127	Action	1993	Universal Pictures	PG-13
11	Lethal Weapon	110	Action Cops & Robber	1987	Warner Brothers	R
12	Michael	106	Drama	1997	Warner Brothers	PG-13
13	National Lampoon's Vacation	98	Comedy	1983	Warner Brothers	PG-13
14	Poltergeist	115	Horror	1982	MGM / UA	PG
15	Rocky	120	Action Adventure	1976	MGM / UA	PG
16	Scarface	170	Action Cops & Robber	1983	Universal Studios	R
17	Silence of the Lambs	118	Drama Suspense	1991	Orion	R
18	Star Wars	124	Action Sci-Fi	1977	Lucas Film Ltd	PG
19	The Hunt for Red October	135	Action Adventure	1989	Paramount Pictures	PG
20	The Terminator	108	Action Sci-Fi	1984	Live Entertainment	R
21	The Wizard of Oz	101	Adventure	1939	MGM / UA	G
22	Titanic	194	Drama Romance	1997	Paramount Pictures	PG-13

The data stored in the ACTORS dataset consists of three variables: title, actor_leading, and actor_supporting, all of which are defined as character variables. The data stored in the Actors dataset is illustrated below.

ACTORS Dataset

	Title	Actor_Leading	Actor_Supporting
1	Brave Heart	Mel Gibson	Sophie Marceau
2	Christmas Vacation	Chevy Chase	Beverly D'Angelo
3	Coming to America	Eddie Murphy	Arsenio Hall
4	Forrest Gump	Tom Hanks	Sally Field
5	Ghost	Patrick Swayze	Demi Moore
6	Lethal Weapon	Mel Gibson	Danny Glover
7	Michael	John Travolta	Andie MacDowell
8	National Lampoon's Vacation	Chevy Chase	Beverly D'Angelo
9	Rocky	Sylvester Stallone	Talia Shire
10	Silence of the Lambs	Anthony Hopkins	Jodie Foster
11	The Hunt for Red October	Sean Connery	Alec Baldwin
12	The Terminator	Arnold Schwarzenegge	Michael Biehn
13	Titanic	Leonardo DiCaprio	Kate Winslet

The Process of Match-Merging

A traditional merge is the process of combining observations from two or more datasets into a single observation in a newly created dataset. The specific type of merge that we will examine in this paper is known as a match merge. A match-merge combines observations from two or more datasets into a single observation in a new dataset according to the values found in a common variable in each dataset. Its purpose is to bring data from two or more datasets together to explore exciting insights into data relationships. The process consists of a matching process between a dataset's observations bringing together some or all of two or more datasets contents, illustrated in Figure 1.

Figure 1. The Process of Merging Datasets

The ability to define relationships between multiple datasets and retrieve information based on these relationships is a powerful feature of the relational model. Merges are a data manipulation technique on a minimum of two datasets, where a common variable from each dataset is used for the purpose of connecting the datasets. Connecting variables should have the same column attributes and ***"like"*** values since the success of the process depends on these values.

Match-Merge Features and Requirements

1. Relevant only to the SAS System – not portable to other vendor data bases.
2. Requires common variable name in all datasets.
3. Datasets must first be sorted using common variable.
4. Duplicate matching column is automatically overlaid.
5. Results are not automatically printed.

Symmetrical Match-Merging

A traditional match-merge process consists of combining observations in a symmetrical fashion from two or more datasets. The result set from this type of matching process automatically eliminates unmatched observations and is referred to as the intersect (Movies_Actors) between the Movies and Actors datasets, as shown in the Venn diagram in Figure 2.

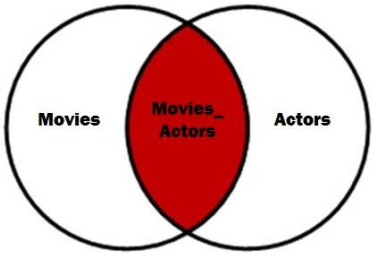

Figure 2. Venn Diagram – Match-Merge

To illustrate the match-merge process, the MOVIES and ACTORS datasets are combined together using one or more common variables. In our example, TITLE, is the common variable found in both datasets and is used to combine observations along with all selected variables (highlighted), as shown in Figure 3.

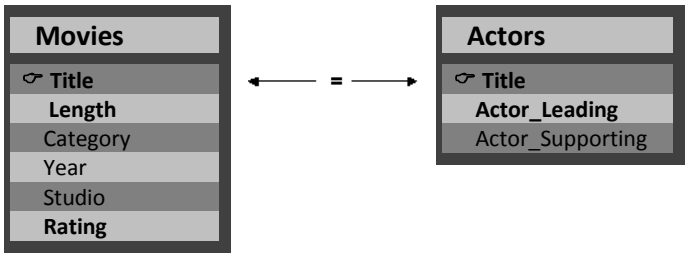

Figure 3. Match-Merge using the MOVIES and ACTORS Datasets

The code to successfully perform a match-merge using the MOVIES and ACTORS datasets is shown below.

```
PROC SORT DATA=MOVIES  OUT=SORTED_MOVIES ;
  BY TITLE ;
RUN ;
PROC SORT DATA=ACTORS  OUT=SORTED_ACTORS ;
 BY TITLE ;
RUN ;

DATA  MATCH_MERGE ;
  MERGE  SORTED_MOVIES (IN=M KEEP=TITLE LENGTH RATING)
         SORTED_ACTORS (IN=A KEEP=TITLE ACTOR_LEADING) ;
  BY TITLE ;
  IF M AND A ;
RUN ;

PROC PRINT DATA=MATCH_MERGE NOOBS ;
RUN ;
```

Results

Title	Length	Rating	Actor_Leading
Brave Heart	177	R	Mel Gibson
Christmas Vacation	97	PG-13	Chevy Chase
Coming to America	116	R	Eddie Murphy
Forrest Gump	142	PG-13	Tom Hanks
Ghost	127	PG-13	Patrick Swayze
Lethal Weapon	110	R	Mel Gibson
Michael	106	PG-13	John Travolta
National Lampoon's Vacation	98	PG-13	Chevy Chase
Rocky	120	PG	Sylvester Stallone
Silence of the Lambs	118	R	Anthony Hopkins
The Hunt for Red October	135	PG	Sean Connery
The Terminator	108	R	Arnold Schwarzenegge
Titanic	194	PG-13	Leonardo DiCaprio

Asymmetrical Match-Merging

A typical merge consists of combining observations from one dataset with observations in another symmetrically. But, occasionally observations need to be captured differently than in a typical merge. One approach, referred to as an asymmetrical type of merge, is designed to preserve unmatched observations from one or both datasets along with the matching observations.

Left Outer Merge

A Left Outer merge produces matched observations from two or more datasets while preserving all unmatched observations from the first specified (left) dataset. A ***Left Outer merge*** is illustrated by the shaded areas (Movies and Movies_Actors) in the Venn diagram, illustrated in Figure 4.

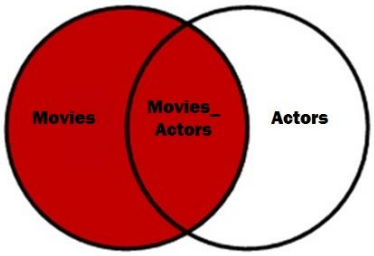

Figure 4. Venn Diagram – Left Outer Merge

The merge code, illustrated below, shows a left outer merge construct that selects “matched” movies based on their titles from the MOVIES and ACTORS datasets, plus all “unmatched” movies from the MOVIES dataset.

```
PROC SORT DATA=MOVIES  OUT=SORTED_MOVIES ;
  BY TITLE ;
RUN ;
PROC SORT DATA=ACTORS  OUT=SORTED_ACTORS ;
  BY TITLE ;
RUN ;

DATA  LEFT_OUTER_MERGE ;
   MERGE  SORTED_MOVIES (IN=M KEEP=TITLE LENGTH RATING)
          SORTED_ACTORS (IN=A KEEP=TITLE ACTOR_LEADING) ;
     BY  TITLE ;
     IF  M ;
RUN ;

PROC PRINT DATA=LEFT_OUTER_MERGE NOOBS ;
RUN ;
```

Results

Title	Length	Rating	Actor_Leading
Brave Heart	177	R	Mel Gibson
Casablanca	103	PG	
Christmas Vacation	97	PG-13	Chevy Chase
Coming to America	116	R	Eddie Murphy
Dracula	130	R	
Dressed to Kill	105	R	
Forrest Gump	142	PG-13	Tom Hanks
Ghost	127	PG-13	Patrick Swayze
Jaws	125	PG	
Jurassic Park	127	PG-13	
Lethal Weapon	110	R	Mel Gibson
Michael	106	PG-13	John Travolta
National Lampoon's Vacation	98	PG-13	Chevy Chase
Poltergeist	115	PG	
Rocky	120	PG	Sylvester Stallone
Scarface	170	R	
Silence of the Lambs	118	R	Anthony Hopkins
Star Wars	124	PG	
The Hunt for Red October	135	PG	Sean Connery
The Terminator	108	R	Arnold Schwarzenegge
The Wizard of Oz	101	G	
Titanic	194	PG-13	Leonardo DiCaprio

Right Outer Merge

A Right Outer merge produces matched observations from two or more datasets while preserving all unmatched observations from the second specified (right) dataset. A ***Right Outer merge*** is illustrated by the shaded areas (Movies and Movies_Actors) shown by the Venn diagram, illustrated in Figure 5.

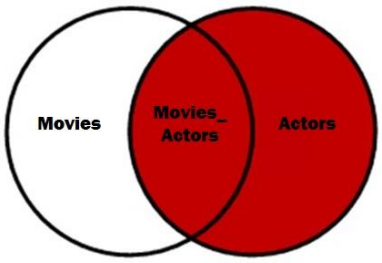

Figure 5. Venn Diagram – Right Outer Merge

The merge code, illustrated below, shows a right outer merge construct that selects "matched" movies based on their titles from the MOVIES and ACTORS datasets, plus all "unmatched" movies from the ACTORS dataset.

```
PROC SORT DATA=MOVIES  OUT=SORTED_MOVIES ;
  BY TITLE ;
RUN ;
PROC SORT DATA=ACTORS  OUT=SORTED_ACTORS ;
  BY TITLE ;
RUN ;

DATA  RIGHT_OUTER_MERGE ;
   MERGE  SORTED_MOVIES (IN=M KEEP=TITLE LENGTH RATING)
          SORTED_ACTORS (IN=A KEEP=TITLE ACTOR_LEADING) ;
     BY  TITLE ;
     IF  A ;
RUN ;

PROC PRINT DATA=RIGHT_OUTER_MERGE NOOBS ;
RUN ;
```

Results

Title	Length	Rating	Actor_Leading
Brave Heart	177	R	Mel Gibson
Christmas Vacation	97	PG-13	Chevy Chase
Coming to America	116	R	Eddie Murphy
Forrest Gump	142	PG-13	Tom Hanks
Ghost	127	PG-13	Patrick Swayze
Lethal Weapon	110	R	Mel Gibson
Michael	106	PG-13	John Travolta
National Lampoon's Vacation	98	PG-13	Chevy Chase
Rocky	120	PG	Sylvester Stallone
Silence of the Lambs	118	R	Anthony Hopkins
The Hunt for Red October	135	PG	Sean Connery
The Terminator	108	R	Arnold Schwarzenegge
Titanic	194	PG-13	Leonardo DiCaprio

Conclusion

The Base-SAS DATA step is a wonderful language for SAS users to explore and use in a variety of application situations. This paper presented a brief introduction to the exciting world of DATA step merges, what a merge is, illustrations of the various merge techniques, and how the DATA step can be used to merge two or more datasets. You are encouraged to explore these and other techniques to make your SAS experience an exciting one.

References

Lafler, Kirk Paul (2013). *PROC SQL: Beyond the Basics Using SAS, Second Edition*, SAS Institute Inc., Cary, NC, USA.

Lafler, Kirk Paul (2012), *"Exploring DATA Step Merges and PROC SQL Joins,"* Proceedings of the 2012 SAS Global Forum (SGF) Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2011), *"Exploring DATA Step Merges and PROC SQL Joins,"* Proceedings of the 2011 PharmaSUG Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2010), *"DATA Step and PROC SQL Programming Techniques,"* Ohio SAS Users Group (OSUG) 2010 One-Day Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2009), *"DATA Step and PROC SQL Programming Techniques,"* South Central SAS Users Group (SCSUG) 2009 Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2009), *"DATA Step versus PROC SQL Programming Techniques,"* Sacramento Valley SAS Users Group 2009 Meeting, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul, Advanced SAS® Programming Tips and Techniques; Software Intelligence Corporation, Spring Valley, CA, USA; 1987-2007.

Lafler, Kirk Paul (2007), *"Undocumented and Hard-to-find PROC SQL Features,"* Proceedings of the PharmaSUG 2007 Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul and Ben Cochran (2007), *"A Hands-on Tour Inside the World of PROC SQL Features,"* Proceedings of the SAS Global Forum (SGF) 2007 Conference, Software Intelligence Corporation, Spring Valley, CA, and The Bedford Group, USA.

Lafler, Kirk Paul (2006), *"A Hands-on Tour Inside the World of PROC SQL,"* Proceedings of the 31st Annual SAS Users Group International Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2005), *"Manipulating Data with PROC SQL,"* Proceedings of the 30th Annual SAS Users Group International Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2004). *PROC SQL: Beyond the Basics Using SAS*, SAS Institute Inc., Cary, NC, USA.

Acknowledgments

The authors thank Dave Foster, Rapid Fire Section Chair, for accepting our abstract and paper; Michael G. Wilson, MWSUG 2015 Academic Chair, David Bruckner, MWSUG 2015 Operations Chair, the MidWest SAS Users Group (MWSUG) Executive Board, and SAS Institute for organizing and supporting a great conference!

Trademark Citations

About the Authors

Kirk Paul Lafler has been using SAS since 1979 and is consultant and founder of Software Intelligence Corporation. He is a SAS Certified Professional, provider of IT consulting services, trainer to SAS users around the world, mentor, and sasCommunity.org emeritus Advisory Board member. As the author of six books including Google® Search Complete! (Odyssey Press. 2014) and PROC SQL: Beyond the Basics Using SAS, Second Edition (SAS Press. 2013); Kirk has written more than five hundred papers and articles; been an Invited speaker and trainer at five hundred-plus SAS International, regional, special-interest, local, and in-house user group conferences and meetings; and is the recipient of 23 "Best" contributed paper, hands-on workshop (HOW), and poster awards.

Charles Edwin Shipp is a programmer, consultant, mentor and author, and has been using the SAS and JMP software since 1980. He is credited in the original JMP manual for his roles in the early days. Charlie has written more than one hundred papers and has been an invited speaker at more than one hundred International, regional, special-interest, local, and in-house SAS and JMP conferences and meetings, and is the recipient of 13 "Best" contributed paper and poster awards. He is the co-author of several books including Google Search Complete! (Odyssey Press. 2014); and Quick Results with SAS/GRAPH Software (SAS Press. 1994). Currently, Charlie is involved as an eBook author, sasCommunity.org Advisory Board member, mentor, application developer, and consultant in SAS and JMP software.

Comments and suggestions can be sent to:

Kirk Paul Lafler
Senior SAS® Consultant, Application Developer, Data Scientist, Trainer and Author
Software Intelligence Corporation
E-mail: KirkLafler@cs.com
LinkedIn: http://www.linkedin.com/in/KirkPaulLafler
Twitter: @sasNerd
~~~
Charles Edwin Shipp
Senior SAS® and JMP® Consultant, eBook Author, Programmer, Trainer and Author
Consider Consulting Corporation
E-mail: CharlieShipp@aol.com
LinkedIn: https://www.linkedin.com/in/charlieshipp
Twitter: @ShippAhoy
~~~