

Kirk's Korner



Quick & Simple Tips

Kirk Paul Lafler, Software Intelligence Corporation

Tip: Manipulating Data with PROC SQL

INTRODUCTION

PROC SQL provides numerous arithmetic, statistical, and summary functions to manipulate numeric data. With one numeric data type to represent numeric data, the NUMERIC or NUM column definition is automatically assigned a default length of 8 bytes, even if the column is created with a numeric length less than 8 bytes.

The variety of functions and operators available to PROC SQL can be used for character data manipulation including words, text, and codes using the CHARACTER or CHAR data type. The CHARACTER or CHAR data type allows ASCII or EBCDIC character sets and stores fixed-length character strings consisting of a maximum of 32K characters. This paper will illustrate how columns based on the numeric and character data types are defined; how string functions, pattern matching, phonetic matching techniques, and a variety of other techniques are used with numeric and character data.

PROC SQL OPERATORS AND FUNCTIONS

PROC SQL users have a number of ways to accomplish their objectives, particularly when the goal is to manipulate data. The SELECT statement is an extremely powerful statement in the SQL language. Its syntax can be somewhat complex because of the number of ways that columns, tables, operators, and functions can be combined into executable statements. There are several types of operators and functions in PROC SQL: 1) comparison operators, 2) logical operators, 3) arithmetic operators, 4) character string operators, and 5) summary functions. This tip will illustrate concatenation, character alignment, and phonetic and pattern matching character-string operations in manipulating data.

Character String Operators and Functions

Character string operators and functions are typically used with character data. Numerous operators are presented to acquaint users to the power available with the SQL procedure. You'll see a number of operators including string concatenation and character alignment.

Concatenation and Character Alignment

The default alignment for character data is to the left, however character columns or expressions can also be aligned to the right. Two functions are available for character alignment: LEFT and RIGHT. The next example combines the concatenation operator "||" and the TRIM function with the LEFT function to left align a character expression while inserting blank spaces and a dash "-" between two character columns to subset "PG-rated" movies.

PROC SQL Code

```
PROC SQL;
  SELECT LEFT(TRIM(title) || " - " || category) AS Concatenation_Alignment
  FROM movies
  WHERE rating = "PG";
QUIT;
```

Results

Concatenation_Alignment

Casablanca - Drama
Jaws - Action Adventure
Poltergeist - Horror
Rocky - Action Adventure
Star Wars - Action Sci-Fi
The Hunt for Red October - Action Adventure

Phonetic Matching (Sounds-Like Operator =*)

A technique for finding names that sound alike or have spelling variations is available in PROC SQL. Although not technically a function, the sounds-like operator "=* " searches and selects character data based on two expressions: the search value and the matched value. Anyone that has looked for a last name in a local telephone directory is quickly reminded of the possible phonetic variations. To illustrate how the sounds-like operator works, we will search on the movie title in the MOVIES table using the string "Rucky" for any and all phonetic variations related to the movie title "Rocky".

PROC SQL Code

```
PROC SQL;  
  SELECT title, rating, category  
  FROM movies  
  WHERE title =* "Rucky";  
QUIT;
```

Results

| Title | Rating | Category |
|-------|--------|------------------|
| Rocky | PG | Action Adventure |

Finding Patterns in a String (Pattern Matching % and _)

Constructing specific search patterns in string expressions is a simple process with the LIKE predicate. The % acts as a wildcard character representing any number of characters, including any combination of upper or lower case characters. Combining the LIKE predicate with the % (percent sign) permits case-sensitive searches and is a popular technique used by savvy SQL programmers to find patterns in their data. The next example finds patterns in the movie category containing the uppercase character 'D' in the first position followed by any number of characters in the CATEGORY column.

PROC SQL Code

```
PROC SQL;  
  SELECT title, rating, category  
  FROM movies  
  WHERE category LIKE 'D%';  
QUIT;
```

Results

| Title | Rating | Category |
|----------------------|--------|-----------------|
| Casablanca | PG | Drama |
| Forrest Gump | PG-13 | Drama |
| Michael | PG-13 | Drama |
| Dressed to Kill | R | Drama Mysteries |
| Ghost | PG-13 | Drama Romance |
| Titanic | PG-13 | Drama Romance |
| Silence of the Lambs | R | Drama Suspense |

CONTACT INFORMATION

If you would like more information or have any questions about this tip, please contact: Kirk Lafler, Software Intelligence Corporation at KirkLafler@cs.com. Kirk, a SAS Certified Professional® and SAS Alliance Partner® (1996 - 2002) with more than 25 years of SAS software experience, provides consulting services and hands-on SAS training around the world. Kirk has written four books including PROC SQL: Beyond the Basics Using SAS by SAS Institute, Power SAS and Power AOL by Apress, and more than one hundred articles in professional journals and SAS User Group proceedings. His popular SAS Tips column appears regularly in the BASAS, HASUG, SANDS, SAS, SESUG, and WUSS Newsletters and websites. Kirk can be reached at:

Kirk Paul Lafler
Software Intelligence Corporation
P.O. Box 1390
Spring Valley, California 91979-1390
Voice: 619-277-7350
E-mail: KirkLafler@cs.com
Web: www.software-intel.com



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.