CHAPTER

1

Introduction to the SQL Procedure

What Is SQL?

Structured Query Language (SQL) is a standardized, widely used language that retrieves and updates data in relational tables and databases.

A *relation* is a mathematical concept that is similar to the mathematical concept of a set. Relations are represented physically as two-dimensional tables that are arranged in rows and columns. Relational theory was developed by E. F. Codd, an IBM researcher, and first implemented at IBM in a prototype called System R. This prototype evolved into commercial IBM products based on SQL. The Structured Query Language is now in the public domain and is part of many vendors' products.

What Is the SQL Procedure?

The SQL procedure is the SAS System implementation of Structured Query Language. PROC SQL is part of base SAS software, and you can use it with any SAS data set (table). Often, PROC SQL can be an alternative to other SAS procedures or the DATA step. You can use SAS language elements such as global statements, data set options, functions, informats, and formats with PROC SQL just as you can with other SAS procedures. PROC SQL can

- □ generate reports
- □ generate summary statistics
- □ retrieve data from tables or views
- □ combine data from tables or views
- □ create tables, views, and indexes
- □ update the data values in PROC SQL tables
- □ update and retrieve data from database management system (DBMS) tables

□ modify a PROC SQL table by adding, modifying, or dropping columns.

PROC SQL can be used in an interactive SAS session or within batch programs, and it can include global statements, such as TITLE and OPTIONS.

Terminology

Tables

A PROC SQL *table* is the same as a SAS data file. It is a SAS file of type DATA. PROC SQL tables consist of rows and columns. The rows correspond to observations in SAS data files, and the columns correspond to variables. The following table lists equivalent terms that are used in SQL, the SAS System, and traditional data processing.

SQL Term	Base SAS Term	Data Processing Term
table	SAS data file	file
row	observation	record
column	variable	field

You can create and modify tables by using the SAS DATA step, or by using the PROC SQL statements that are described in Chapter 4, "Creating and Updating Tables and Views," on page 83 . Other SAS procedures and the DATA step can read and update tables that are created with PROC SQL.

DBMS tables are tables that were created with other software vendors' database management systems. PROC SQL can connect to, update, and modify DBMS tables, with some restrictions. For more information, see "Accessing a DBMS with SAS/ACCESS" on page 121 .

Queries

Queries retrieve data from a table, view, or DBMS. A query returns a *query result*, which consists of rows and columns from a table. With PROC SQL, you use a SELECT statement and its subordinate clauses to form a query. Chapter 2, "Retrieving Data from a Single Table," on page 11 describes how to build a query.

Views

PROC SQL views do not actually contain data as tables do. Rather, a PROC SQL view contains a stored SELECT statement or query. The query executes when you use the view in a SAS procedure or DATA step. When a view executes, it displays data that is derived from existing tables, from other views, or from SAS/ACCESS views. Other SAS procedures and the DATA step can use a PROC SQL view as they would any SAS data file. For more information about views, see Chapter 4, "Creating and Updating Tables and Views," on page 83 .

Null Values

According to the ANSI Standard for SQL, a missing value is called a *null*. It is not the same as a blank or zero value. However, to be compatible with the rest of the SAS System, PROC SQL treats missing values the same as blanks or zero values, and considers all three to be nulls. This important concept comes up in several places in this book.

Comparing PROC SQL with the SAS DATA Step

PROC SQL can perform some of the operations that are provided by the DATA step and the PRINT, SORT, and SUMMARY procedures. The following query displays the total population of all the large countries (countries with population greater than 1 million) on each continent.

```
proc sql;
    title 'Population of Large Countries Grouped by Continent';
    select Continent, sum(Population) as TotPop format=comma15.
        from sql.countries
        where Population gt 1000000
        group by Continent
        order by TotPop;
quit;
```

Output 1.1   Sample SQL Output

```
            Population of Large Countries Grouped by Continent

            Continent                               TotPop
            ----------------------------------------------
            Oceania                              3,422,548
            Australia                           18,255,944
            Central America and Caribbean       65,283,910
            South America                      316,303,397
            North America                      384,801,818
            Africa                             706,611,183
            Europe                             811,680,062
            Asia                             3,379,469,458
```

Here is a SAS program that produces the same result.

```
title 'Large Countries Grouped by Continent';
proc summary data=sql.countries;
    where Population > 1000000;
    class Continent;
    var Population;
    output out=sumPop sum=TotPop;
run;

proc sort data=SumPop;
    by totPop;
run;
```

```
proc print data=SumPop noobs;
   var Continent TotPop;
   format TotPop comma15.;
   where _type_=1;
run;
```

Output 1.2   Sample DATA Step Output

```
                Large Countries Grouped by Continent

        Continent                                      TotPop

        Oceania                                     3,422,548
        Australia                                  18,255,944
        Central America and Caribbean              65,283,910
        South America                             316,303,397
        North America                             384,801,818
        Africa                                    706,611,183
        Europe                                    811,680,062
        Asia                                    3,379,469,458
```

This example shows that PROC SQL can achieve the same results as base SAS software but often with fewer and shorter statements. The SELECT statement that is shown in this example performs summation, grouping, sorting, and row selection. It also displays the query's results without the PRINT procedure.

PROC SQL executes without using the RUN statement. After you invoke PROC SQL you can submit additional SQL procedure statements without submitting the PROC statement again. Use the QUIT statement to terminate the procedure.

Notes about the Example Tables

For all examples, the following global statements are in effect:

```
options nodate nonumber linesize=80 pagesize=60;
libname sql 'SAS-data-library';
```

The tables that are used in this book contain geographic and demographic data. The data is intended to be used for the PROC SQL code examples only; it is not necessarily up to date or accurate.

The COUNTRIES table contains data that pertains to countries. The Area column contains a country's area in square miles. The UNDate column contains the year a country entered the United Nations, if applicable.

Output 1.3 COUNTRIES (partial output)

```
                                   COUNTRIES

Name                    Capital          Population      Area  Continent          UNDate
-----------------------------------------------------------------------------------------
Afghanistan             Kabul              17070323    251825  Asia                 1946
Albania                 Tirane              3407400     11100  Europe               1955
Algeria                 Algiers            28171132    919595  Africa               1962
Andorra                 Andorra la Vell       64634       200  Europe               1993
Angola                  Luanda              9901050    481300  Africa               1976
Antigua and Barbuda     St. John's            65644       171  Central America      1981
Argentina               Buenos Aires       34248705   1073518  South America        1945
Armenia                 Yerevan             3556864     11500  Asia                 1992
Australia               Canberra           18255944   2966200  Australia            1945
Austria                 Vienna              8033746     32400  Europe               1955
Azerbaijan              Baku                7760064     33400  Asia                 1992
Bahamas                 Nassau               275703      5400  Central America      1973
Bahrain                 Manama               591800       300  Asia                 1971
Bangladesh              Dhaka              1.2639E8     57300  Asia                 1974
Barbados                Bridgetown           258534       200  Central America      1966
```

The WORLDCITYCOORDS table contains latitude and longitude data for world cities. Cities in the Western hemisphere have negative longitude coordinates. Cities in the Southern hemisphere have negative latitude coordinates. Coordinates are rounded to the nearest degree.

Output 1.4 WORLDCITYCOORDS (partial output)

```
                       WORLDCITCOORDS

         City           Country       Latitude  Longitude
         ------------------------------------------------
         Kabul          Afghanistan        35         69
         Algiers        Algeria            37          3
         Buenos Aires   Argentina         -34        -59
         Cordoba        Argentina         -31        -64
         Tucuman        Argentina         -27        -65
         Adelaide       Australia         -35        138
         Alice Springs  Australia         -24        134
         Brisbane       Australia         -27        153
         Darwin         Australia         -12        131
         Melbourne      Australia         -38        145
         Perth          Australia         -32        116
         Sydney         Australia         -34        151
         Vienna         Austria            48         16
         Nassau         Bahamas            26        -77
         Chittagong     Bangladesh         22         92
```

The USCITYCOORDS table contains the coordinates for cities in the United States. Because all cities in this table are in the Western hemisphere, all the longitude coordinates are negative. Coordinates are rounded to the nearest degree.

Output 1.5 USCITYCOORDS (partial output)

```
                              USCITYCOORDS

                  City            State  Latitude  Longitude
                  ------------------------------------------
                  Albany          NY         43        -74
                  Albuquerque     NM         36       -106
                  Amarillo        TX         35       -102
                  Anchorage       AK         61       -150
                  Annapolis       MD         39        -77
                  Atlanta         GA         34        -84
                  Augusta         ME         44        -70
                  Austin          TX         30        -98
                  Baker           OR         45       -118
                  Baltimore       MD         39        -76
                  Bangor          ME         45        -69
                  Baton Rouge     LA         31        -91
                  Birmingham      AL         33        -87
                  Bismarck        ND         47       -101
                  Boise           ID         43       -116
```

The UNITEDSTATES table contains data that is associated with the states. The Statehood column contains the date when the state was admitted into the Union.

Output 1.6 UNITEDSTATES (partial output)

```
                              UNITEDSTATES

Name               Capital        Population     Area  Continent        Statehood
--------------------------------------------------------------------------------
Alabama            Montgomery       4227437     52423  North America    14DEC1819
Alaska             Juneau            604929    656400  North America    03JAN1959
Arizona            Phoenix          3974962    114000  North America    14FEB1912
Arkansas           Little Rock      2447996     53200  North America    15JUN1836
California         Sacramento      31518948    163700  North America    09SEP1850
Colorado           Denver           3601298    104100  North America    01AUG1876
Connecticut        Hartford         3309742      5500  North America    09JAN1788
Delaware           Dover             707232      2500  North America    07DEC1787
District of Colum  Washington        612907       100  North America    21FEB1871
Florida            Tallahassee     13814408     65800  North America    03MAR1845
Georgia            Atlanta          6985572     59400  North America    02JAN1788
Hawaii             Honolulu         1183198     10900  Oceania          21AUG1959
Idaho              Boise            1109980     83600  North America    03JUL1890
Illinois           Springfield     11813091     57900  North America    03DEC1818
Indiana            Indianapolis     5769553     36400  North America    11DEC1816
```

The POSTALCODES table contains postal code abbreviations.

Output 1.7 POSTALCODES (partial output)

```
                        POSTALCODES

        Name                            Code
        -------------------------------------
        Alabama                         AL
        Alaska                          AK
        American Samoa                  AS
        Arizona                         AZ
        Arkansas                        AR
        California                      CA
        Colorado                        CO
        Connecticut                     CT
        Delaware                        DE
        District Of Columbia            DC
        Florida                         FL
        Georgia                         GA
        Guam                            GU
        Hawaii                          HI
        Idaho                           ID
```

The WORLDTEMPS table contains average high and low temperatures from various international cities.

Output 1.8 WORLDTEMPS (partial output)

```
                      WORLDTEMPS

      City            Country         AvgHigh    AvgLow
      --------------------------------------------------
      Algiers         Algeria           90         45
      Amsterdam       Netherlands       70         33
      Athens          Greece            89         41
      Auckland        New Zealand       75         44
      Bangkok         Thailand          95         69
      Beijing         China             86         17
      Belgrade        Yugoslavia        80         29
      Berlin          Germany           75         25
      Bogota          Colombia          69         43
      Bombay          India             90         68
      Bucharest       Romania           83         24
      Budapest        Hungary           80         25
      Buenos Aires    Argentina         87         48
      Cairo           Egypt             95         48
      Calcutta        India             97         56
```

The OILPROD table contains oil production statistics from oil-producing countries.

Output 1.9   OILPROD (partial output)

```
                             OILPROD

                                         Barrels
                Country                   PerDay
                ----------------------------------------
                Algeria                 1,400,000
                Canada                  2,500,000
                China                   3,000,000
                Egypt                     900,000
                Indonesia               1,500,000
                Iran                    4,000,000
                Iraq                      600,000
                Kuwait                  2,500,000
                Libya                   1,500,000
                Mexico                  3,400,000
                Nigeria                 2,000,000
                Norway                  3,500,000
                Oman                      900,000
                Saudi Arabia            9,000,000
                United States of America 8,000,000
```

The OILRSRVS table lists approximate oil reserves of oil-producing countries.

Output 1.10   OILRSRVS (partial output)

```
                             OILRSRVS

          Country                              Barrels
          --------------------------------------------------
          Algeria                        9,200,000,000
          Canada                         7,000,000,000
          China                         25,000,000,000
          Egypt                          4,000,000,000
          Gabon                          1,000,000,000
          Indonesia                      5,000,000,000
          Iran                          90,000,000,000
          Iraq                         110,000,000,000
          Kuwait                        95,000,000,000
          Libya                         30,000,000,000
          Mexico                        50,000,000,000
          Nigeria                       16,000,000,000
          Norway                        11,000,000,000
          Saudi Arabia                 260,000,000,000
          United Arab Emirates             100,000,000
```

The CONTINENTS table contains geographic data relating to world continents.

Output 1.11 CONTINENTS

```
                                   CONTINENTS

Name                 Area  HighPoint              Height  LowPoint           Depth

-----------------------------------------------------------------------------------
Africa           11506000  Kilimanjaro             19340  Lake Assal          -512
Antarctica        5500000  Vinson Massif           16860                         .
Asia             16988000  Everest                 29028  Dead Sea           -1302
Australia         2968000  Kosciusko                7310  Lake Eyre            -52
Central America         .                              .                         .
Europe            3745000  El'brus                 18510  Caspian Sea          -92
North America     9390000  McKinley                20320  Death Valley        -282
Oceania                 .                              .                         .
South America     6795000  Aconcagua               22834  Valdes Peninsul     -131
```

The FEATURES table contains statistics that describe various types of geographical features, such as oceans, lakes, and mountains.

Output 1.12 FEATURES (partial output)

```
                                   FEATURES

Name           Type        Location          Area   Height    Depth   Length

------------------------------------------------------------------------------
Aconcagua      Mountain    Argentina            .    22834        .        .
Amazon         River       South America        .        .        .     4000
Amur           River       Asia                 .        .        .     2700
Andaman        Sea                         218100        .     3667        .
Angel Falls    Waterfall   Venezuela            .     3212        .        .
Annapurna      Mountain    Nepal                .    26504        .        .
Aral Sea       Lake        Asia             25300        .      222        .
Ararat         Mountain    Turkey               .    16804        .        .
Arctic         Ocean                     5105700        .    17880        .
Atlantic       Ocean                    33420000        .    28374        .
Baffin         Island      Arctic         183810        .        .        .
Baltic         Sea                        146500        .      180        .
Baykal         Lake        Russia          11780        .     5315        .
Bering         Sea                        873000        .     4893        .
Black          Sea                        196100        .     3906        .
```