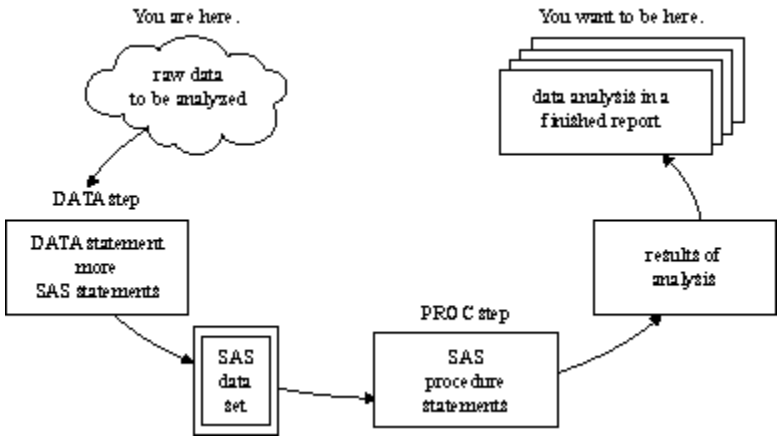
Introduction to DATA Step Processing

The SAS Data Set: Your Key to the SAS System

Understanding the Function of the SAS Data Set

SAS enables you to solve problems by providing methods to analyze or to process your data in some way. You need to first get the data into a form that SAS can recognize and process. After the data is in that form, you can analyze it and generate reports. The following figure shows this process in the simplest case.

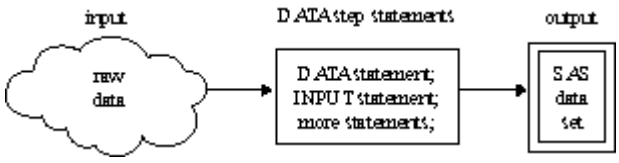
From Raw Data to Final Analysis



You begin with **raw data**, that is, a collection of data that has not yet been processed by SAS. You use a set of statements known as a **DATA step** to get your data into a SAS data set. Then you can further process your data with additional DATA step programming or with SAS procedures.

In its simplest form, the DATA step can be represented by the three components that are shown in the following figure.

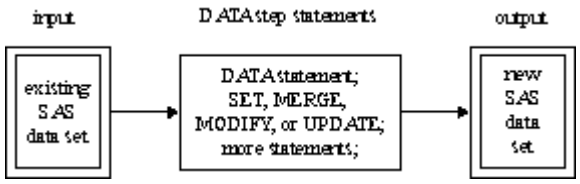
From Raw Data to a SAS Data Set



SAS processes input in the form of raw data and creates a SAS data set.

When you have a SAS data set, you can use it as input to other DATA steps. The following figure shows the SAS statements that you can use to create a new SAS data set.

Using One SAS Data Set to Create Another



Understanding the Structure of the SAS Data Set

Think of a SAS data set as a rectangular structure that identifies and stores data. When your data is in a SAS data set, you can use additional DATA steps for further processing, or perform many types of analyses with SAS procedures.

The rectangular structure of a SAS data set consists of rows and columns in which data values are stored. The rows in a SAS data set are called **observations**, and the columns are called **variables**. In a raw data file, the rows are called **records** and the columns are called **fields**. Variables contain the data values for all of the items in an observation.

For example, the following figure shows a collection of raw data about participants in a health and fitness club. Each record contains information about one participant.

Raw Data from the Health and Fitness Club

raw data

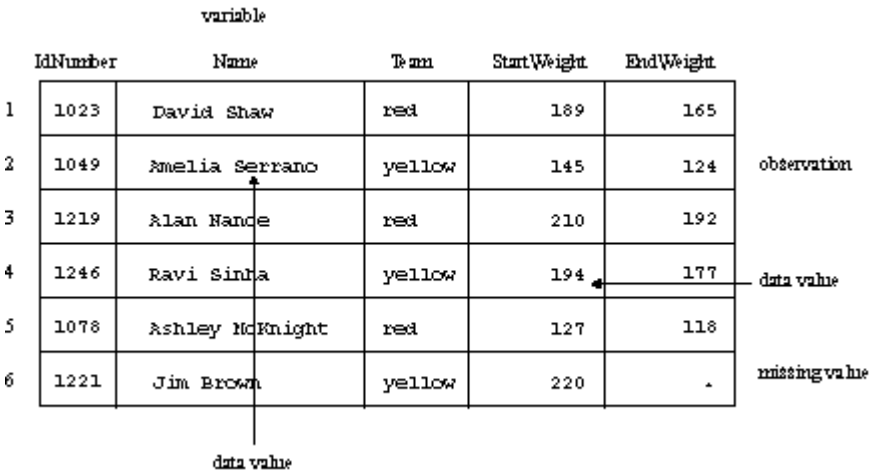
data fields

Healthand Fitness Club Data

Id	Name	Team	Starting Weight	Ending Weight
1023	David Shaw	red	189	165
1049	Amelia Serrano	yellow	145	124
1219	Alan Nance	red	210	192
1246	Ravi Sinha	yellow	194	177
1078	Ashley McKnight	red	127	118
1221	Jim Brown	yellow	220	-

The following figure shows how easily the health club records can be translated into parts of a SAS data set. Each record becomes an observation. In this case, each observation represents a participant in the program. Each field in the record becomes a variable. The variables represent each participant's identification number, name, team name, and weight at the beginning and end of a 16-week program.

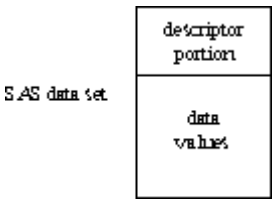
How Data Fits into a SAS Data Set



In a SAS data set, every variable exists for every observation. What if you do not have all the data for each observation? If the raw data is incomplete because a value for the numeric variable EndWeight was not recorded for one observation, then this **missing value** is represented by a period that serves as a placeholder, as shown in observation 6 in the previous figure. (Missing values for character variables are represented by blanks. Character and numeric variables are discussed later in this section.) By coding a value as missing, you can add an observation to the data set for which the data is incomplete and still retain the rectangular shape necessary for a SAS data set.

Along with data values, each SAS data set contains a descriptor portion, as illustrated in the following figure:

Parts of a SAS Data Set



The descriptor portion consists of details that SAS records about a data set, such as the names and attributes of all the variables, the number of observations in the data set, and the date and time that the data set was created and updated.

Operating Environment Information: Depending on your operating environment and the engine used to write the SAS data set, SAS may store additional information about a SAS data set in its descriptor portion. For more information, refer to the SAS documentation for your operating environment. ■

Temporary versus Permanent SAS Data Sets