

Using The SAS System to Help Manage Data

Clarence W. Jackson

ABSTRACT

This paper describes how the SAS System can be used to help manage data. Actions such as adding, deleting, and updating of data requires a data manager to be alert in the management of data in order to insure data integrity, especially when that data is being used in reports, graphs, and other displays of the data to an organization operational and management staff.

INTRODUCTION

The SAS System is very useful in presenting information about data for a company's operational and management decisions. The SAS System can also be used as a data management tool. In some ways, the SAS System can present information about the data being managed that facilitates decisions about how to store, maintain, and manipulate data. This paper describes methods that can be used to help manage data, and aid in management decisions.

Sections of this paper are:

- understanding the data
- Summarize data element values

Data management applications include:

- allocate resources
- verifying data relationships and edits

- evaluating testing conditions
- defining system requirements
- defining program requirements
- writing SAS programs

UNDERSTANDING THE DATA

Understanding the data on files being managed is key to any data management function. Data management functions include controlling the storage, maintenance, and modification of data. To properly manage data, one must have an understanding of the data, how it will be used, who will use it, how it will be modified, and verified.

The data manager should also understand the flow of data within the system and throughout the organization. Data flow diagrams, Entity Relationship diagrams, database dictionaries, and organization charts could be used to help the data manager understand who does what with the data elements within the system.

Understanding the data to be managed, and it's life cycle is fundamental to effective data management.

Familiarity with the diverse data elements can be gained by reviewing copybooks, production programs, and summarized lists or tables. Also, reports requested by the organization's business

managers provide information of what data they are concerned with.

SUMMARIZE DATA ELEMENT VALUES

Summary data values can provide information about the data stores being managed. In some cases, keeping summarized reports will eliminate the need of running ad-hoc reports for redundant information. Also, by using summary reports, the data manager can define the system's requirements using actual data element values.

Tools provided by the SAS System that summarize values of data elements are numerous, just pick a PROCedure such as PROC FREquency, PROC TABULATE or any other PROC. When using PROCs to summarize different data elements (or variables), the data element values can be noted easily.

The use of summarized lists and tables can aid the data manager in many ways. Requirements for programs and processes can be structured around the data. Redundant data elements can be more easily identified.

ALLOCATING RESOURCES

Understanding your data and using summary information of data elements can aid the data manager in defining requirements for critical resources, such as coding programs, defining screen edits, and allocating file space. For instance, finding that you only have 3 observations containing a specific value in a file of 300,000,000 could lead you to process those 3 records using some other method than with programming, such as a manual operation to deal with those 3.

Example: using PROC FREquency on a billing file's payment status flags not only gives the manager of the Accounting and Collections a summary of what the statuses are, but also can alert the data manager of upcoming needs to clear the data stores of data that can be put to history data stores, give a picture of the universe of data types that need to be programmed and edited for, and other things of interest such as the total number of records. This information can then be used in determining resource allocations such as file space needed to store the data, and the rate of file growth.

VERIFYING DATA RELATIONSHIPS AND EDITS

Using Cross-table summary data can alert you to problems in edits resulting from data entry, batch updates gone wrong, and other invalid conditions. In cases where there should only be certain combinations or data relationships allowed, PROC FREquency works well, and is very simple to code. To test the data relationship between the payment status flag and the balance due flag would require a simple amount of coding:

```
DATA CHECKBAL
  (KEEP=PAYSTAT BALDUE);
INFILE BILLING;
INPUT @10 BAL PD6.2
      @20 PAYSTAT $1.;
SELECT;
  WHEN (BAL=0)
    BALDUE='PAID';
  WHEN (BAL>0)
    BALDUE='DUE';
  WHEN (BAL<0)
    BALDUE='CREDIT';
  ELSE BALDUE='INVALID';
END;
RUN;

PROC FREQ;
  TABLES PAYSTAT*BALDUE;
```

TITLE "PAYMENTS 2 BALANCE";
RUN;

to produce a table of payment status and balance due showing the values of each data element and it's related data element value. All combinations or relationships of the data elements can be seen. If any invalid data relationships are found, action can be taken to resolve the problem.

EVALUATING TESTING CONDITIONS

Creating the right test conditions can be critical to the success of a project involving data management. Defining the test data is the most critical element of testing and evaluating any data management project. There are many methods used in testing, and there are many schools of thought as to what constitutes good test data. There are those that feel that only the entire file or systems must be used so that a test can be deemed completed. There are also those that feel a subset of the data is enough to use in a testing environment.

The most effective test involves the use of each type of data that the systems will encounter. Using the entire file and/or system will accomplish testing for everything, but at such a cost that effective testing could be prohibitive. Using subsets of the file and/or system will reduce the cost, but the potential danger of something not being tested increases the cost if a failure occurs at a critical period in production.

Using the same methods already discussed, the data manager can create subsets of the file and/or system, and include all

variations of data elements. The required data could be extracted based on the summary data, and with creative programming and use of MACROS, the data manager could even allow the SAS System to build the test data stores. Or, using the summary data, the data manager can better define the test data file requirements with the actual values of data needed, instead of some random query.

The use of subsets containing representatives for the entire file would reduce the cost of testing, and ensure that testing covered all the requirements of the entire system.

DEFINING SYSTEM REQUIREMENTS

The system requirements describes how data will be processed for specific tasks. It should describe what action should be performed for each occurrence of specific data item values. By using summarized lists generated by the SAS System, the data manager can better define how the system requirements should address processing specific data, building other data stores, and other data management functions.

DEFINING PROGRAM REQUIREMENTS

Defining program requirements after this information has been gathered should be a lot easier with less unknowns about what to expect because more information is available about the data elements and volume. Program requirements can be written to more effectively process the data, instead of just coded because of the probability of data existing.

WRITING SAS PROGRAMS

Writing SAS programs using

summarized list is much easier that if no list of data elements are available. The program can be written more efficiently when structured around the data. Debugging and error checking is made much easier with benchmark numbers supplied by the lists.

The programmer can make use of efficient programming methods such as conditional processing, data transformations, and branching within the SAS program.

CONCLUSION

The SAS System is used in presentation of data for a number of reasons and for many different types of managers. The data manager can make use of these data presentations to help manage the data itself, the data stores on which the data is stored, and better define data management requirements in allocating resources. The SAS System can also aid the data manager in verifying data and it's relationship to other data, and monitor the data more effectively. Testing evaluation can be improved by using the SAS System in defining test file conditions and data elements. The SAS System provides many tools in which the data manager can enlist in managing data as it progresses through it's life cycle of being added, updated, and deleted or archived.

The SAS System can help the data manager manage the data.

the average person.

REFERENCE BOOKS USED

- SAS Language and Procedures Guide, Ver 6, 1st edition
- SAS Procedures Reference, Ver 6, 1st edition
- Database Analysis and Design, QED publications

PROC FREQ, and PROC TABULATE are part of the BASE SAS product of the SAS System. SAS and the SAS System are registered trademark of SAS Institute INC, in the USA and other countries.

AUTHOR CONTACT INFORMATION:

Clarence Jackson
2815 Williams Way Ct
Dallas, Tx 75228-4702
INTERNET:
73021.3404@COMPUSERVE.COM

ACKNOWLEDGEMENTS

Thanks to my wife, Norita Jackson, who put up with the kids, and provided feedback as