Paper 205-27

The Differences in Developing an Interactive Review System in SAS®6.12 and in SAS®8.2 HOLAP

Richard A. Denby, United States Census Bureau
Lori A. Guido, United States Census Bureau

Please note: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

ABSTRACT

The Housing and Household Economic Statistics Division (HHES) of the United States (U.S.) Census Bureau evaluated different SAS technologies to find ones best suited for the development of an interactive system to review the Census 2000 long form data. Two systems using the same data but different technologies were built. The systems that were evaluated access data stored on centralized, large-scale Unix servers running Sun SOLARIS 2.8 and the data are accessed through personal computers (PCs) running Windows 95. All the systems use multidimensional databases (MDDBs), and display the same data, but each system uses distinct SAS technologies. The first system uses a client-server approach, SAS v6.12 and the MDDB report object in SAS/EIS® to create more than 2,548 commonly used reports. The second system uses a client-server approach, SAS v8.2's Hybrid on-line Analytical Processing (HOLAP) techniques to build a 'proxy' HOLAP cube in addition to the MDDB report object in SAS/EIS, but only 49 reports had to be created. This paper describes the file architecture each system used, differences in the processing flows of each system and the advantages and disadvantages of each system, including development, deployment, and performance issues. This paper is intended for an audience with intermediate SAS skills.

INTRODUCTION

The U.S. Census Bureau is best known for conducting a national census in years ending in a zero. By December 31 of a census year, the Census Bureau must provide the U. S. President with population totals for each of the 50 states and the District of Columbia (D.C.). This information determines the number of seats to which each state is entitled in the House of Representatives, which is fixed at 435 seats, with each state getting at least one representative. In this way, seats in the House of Representatives are "apportioned" among the states. Census data are also used to delineate congressional and other election districts within each state. This processing is called "redistricting." States typically have tight deadlines for completing their redistricting work in time for the 2001 or 2002 primaries and elections. The Census Bureau is required to provide redistricting data at a very detailed level of geography, to the states within one year of the census. Producing these data is a massive undertaking that involves tabulating the characteristics of more than 280 million people in 120 million housing units assigned to 39,000 governmental entities in 7.5 million census blocks.

The apportionment and redistricting data are derived from the total universe of all people. Additionally, approximately one in six households received a "long form" questionnaire containing 53 questions covering 34 subjects. The remaining five in six households get a shorter form with only seven questions. Every question in Census 2000 was required by law to manage or evaluate federal programs or was needed to meet legal requirements stemming from U.S. court decisions such as on the application of the Voting Rights Act. Over a given decade, more than 180 billion dollars for schools, employment services, housing assistance, highway construction, hospital services, programs for the elderly, and other programs are distributed based on census data. The systems HHES evaluated both used the "long form" questionnaire data.

The SAS v6.12 Review System

The first review system HHES evaluated uses a client-server approach, SAS v6.12 under Windows95 and resides on user PC desktops. Each PC has to have SAS v6.12, and the application's configuration and profile files installed on it. The data reside on Sun Unix servers running the Solaris operating system and SAS v6.12. The MDDBs and the underlying detail data sets on the Unix servers are accessed using "remote library services" via SAS/Connect®. There is one detail data set, one MDDB, one set of SAS/EIS reports and one set of application files for each of the fifty states, D.C., and Puerto Rico (a U.S. commonwealth).

One metabase file holds all of the metadata for all of the states. A metabase is a specialized SAS data set that contains information about SAS data sets and SAS data set variables. A metabase does not contain the actual tables, it contains data about the data sets registered in it. Each metabase can contain information describing one or more tables. In the SAS v6.12 Review System, only one metabase is used to hold all the information about the 416 data sets comprising the system.

The first time you invoke the SAS/EIS development environment, a default metabase named MBUSER is automatically created for you in your SASUSER library. You can use this default metabase to register data, or you can create and use other metabases in other libraries. The developers of the SAS v6.12 Review System that HHES evaluated used a metabase with a unique file name and library. They felt that a uniquely named metabase and library would be more self documenting and eliminate some of the risks and possible confusion that is inherent when using a default file name supplied by SAS.

Application databases consist of an indexed SAS data set and a catalog of the same name. An application database is used to store definitions for an application created using SAS/EIS. The developers of the SAS v6.12 Review System created one application database for each state. Each application database holds a set of the 49 reports needed for the review of that state. For this review system to be fully functional, 2,548 reports, or one set of 49 reports for each state, D.C. and Puerto Rico would have to be created.

This file architecture allowed the developers to modify the reports for one state without impacting the review of other states. Also, since one state's data is processed at a time and is made available to the review system as each state's processing is completed, this architecture fits well with the process flow. However, since the reports are stored in separate application databases, the mass correction of reports is impossible. If a modification had to be made to all states, the developers would have to make it manually, one state at a time. Old data sets for a state could also be easily replaced with a new version, as the new version became available. Whenever a new state was available for review however, another set of reports would have to be built. Sometimes a large number of states would be released at the same time and the review of a state would be delayed because the reports had not been created.

By default, SAS/EIS also stores the application definitions, in the default application database, called SASAPPL in the SASUSER library. The developers created unique file names and libraries for the application databases. They felt that a uniquely named application databases and libraries would be more self documenting and eliminate some of the risks and possible confusion that is inherent when using a default file name supplied by SAS. The application databases and the metabase file reside on a Novell server. This configuration gives the users access to the most current version of the reports without having to copy the files onto their PCs each time a report is modified or added.

The SAS v6.12 Review System does not allow users to view several data groupings, or clusters of states, at once. For example, the Washington, D.C. metro area, which includes D.C. and several counties each from neighboring Maryland and Virginia, cannot be seen using this system.

Table 1 shows the time it takes to display a variety of the summary reports and the detail data for the states of California and Vermont. The subject areas shown include Income, Housing and Ancestry.

Table 1. The SAS v6.12 Review System Performance Statistics

State - California	Summary Report Display Time	Show Detail Data Display Time /# of Obs
Income	30 seconds	18 minutes 4,434,303
Housing	1 minute	8 minutes 1,619,586
Ancestry	4 seconds	17 minutes 4,434,303
State - Vermont		
Income	5 seconds	1 minute 30 seconds 160,404
Housing	10 seconds	30 seconds 83,581
Ancestry	4 seconds	1 minute 20 seconds 160,404

THE SAS v8..2 HOLAP REVIEW SYSTEM

The SAS v8.2 HOLAP Review System was the second system HHES evaluated. It uses a client-server approach, SAS v8.2 under Windows95 and resides on a Novell server. The data reside on Sun Unix servers running the Solaris operating system and SAS v8.2. The MDDBs and the underlying detail data sets on the Unix servers are accessed using "remote library services" via SAS/Connect. There is one detail data set and one MDDB for each of the fifty states, D.C., and Puerto Rico.

The system's main feature is a technology called hybrid online analytical processing (HOLAP). This technology allows users to access data from multiple local and remote MDDBs and data sets as if the data were coming from a single data source. The use of the HOLAP technology decreases the number of reports that will have to be created, from the 2,548 required in the SAS v6.12 Review System, to a single set of 49 reports that can be used for any state. The reports can be created in advance, decreasing the time between file processing and file availability to the users. The HOLAP technology also allows users of the system to view several data groupings, clusters of states, or multiple years in one report. The desktop frame class is overridden so new SCL code that allows multiple states to be selected will execute. If a user selects D.C., Maryland and Virginia, they can view the areas comprising the Washington, D.C. metro area in one report.

Starting with SAS v7, the SAS/EIS metabase facility is stored in a central repository. This repository allows SAS/EIS software to share metadata with other SAS System products. The central repository contains all the metabase registrations. There is only one repository per path. If a system that was using SAS v6.12

had two metabases in a single path, those two metabases would have to be converted into a single repository. Since HHES historically only used one metabase per application, this was not an issue. SAS recommends that the directory used to store the repository should not be used to store other SAS files. HHES developers stored the application databases and metabases in separate directories for other review systems, so this was not a problem either.

One application database contains all the reports for the HOLAP system and resides on the Novell server. This configuration gives the users access to the most current version of the reports without having to copy files to their PCs each time a report is modified or added. This file architecture also does not fit our processing flow because it does not allow the developers to modify the reports without impacting the review. However, since there is only one set of 49 reports to maintain the creation and maintenance of the entire system was much faster and easier then the SAS v6.12 Review System. HHES found that the release of states was sped up due to the use of the HOLAP technology. Old data sets for a state can be easily replaced with a new version, as the new version became available. Whenever a new state is available for review it can be incorporated into the system quickly.

The central repository, the application's configuration and profile files are installed on a Novell server. The SAS software used to run the system also resides on a Novell server. This configuration decreases installation problems. HHES does not have to modify software on individual PCs if a software upgrade is needed.

Table 2 shows the time it takes to display a variety of the summary reports and the detail data for the states of California and Vermont. The subject areas shown include Income, Housing and Ancestry.

Table 2. The SAS v8.2 HOLAP Review System Performance Statistics.

State - California	Summary Report Display Time	Show Detail Data Display Time /# of Obs
Income	24 seconds	2 min. 30 seconds 4,434,303
Housing	55 seconds	1 min. 30 seconds 1,619,586
Ancestry	3 seconds	2 min. 30 seconds 4,434,303
State - Vermont		
Income	3 seconds	9 seconds 160,404
Housing	8 seconds	6 seconds 83,581
Ancestry	1 second	6 seconds 160,404

CONCLUSION

The SAS v8.2 HOLAP Review System worked well for the review of the long form data. The SAS v8.2 HOLAP Review System out performed the SAS v6.12 Review System. The maintenance and flexibility of the SAS v8.2 HOLAP Review System gained via the use of the HOLAP technology, out weighed the risks of storing the reports in one application database. The differences in the file architecture were minimal and, because of the way we had been using SAS v6.12, did not impact us. The SAS v8.2 HOLAP Review System should be easily adaptable for other, similar review and tabulation-style systems.

REFERENCES

SAS Institute Inc (1999), *SAS/EIS Software: Administrator's Guide*, Version 8, Cary, NC: SAS Institute Inc.

SAS Institute Inc (2000), *SAS/OLAP Server Administrator's Guide*, Version 8.1, Cary, NC: SAS Institute Inc.

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.

CONTACT INFORMATION

Richard A. Denby
U.S. Census Bureau
4700 Silver Hill Road, 8500-3
Washington, DC 20233-1912
Phone   301-457-6810
Fax   301-457-3248
Email   richard.a.denby@census.gov

Lori A. Guido
U.S. Census Bureau
4700 Silver Hill Road, 8500-3
Washington, DC 20233-1912
Phone   301-457-3204
Fax   301-457-3499
Email   lori.a.guido@census.gov