Paper 149-27

MDDBs, HOLAP, OverRide Methods, SCL: What Worked and Did Not Work for the Review of the Census 2000 Data

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 was responsible for developing an interactive system for the review of the Census 2000 long form data. The system had to incorporate data on a flow basis, display detail data on the screen in less than 10 seconds, handle more than forty million observations contained in more than 400 files, and be developed in a short time. The system that was developed 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®, and accesses data stored on centralized, large-scale Unix servers through PCs running Windows 95.

This paper describes what factors went into our decision making process for this project, what risks we took, what successes we had, and what failures we had with the development of the Census 2000 Long Form Data Review System.

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 (DC). 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 primaries and elections that will be held in the following year. 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, either to manage or evaluate federal programs or was needed to meet legal requirements stemming from U.S. court decisions such as 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 Census 2000 Long Form Data Review System, was used to review these “long form” questionnaire data during late 2001 and early 2002.

THE CHALLENGES

- HHES had to build a system that would meet the user and technical requirements for the Census 2000 Long Form Data Review System. The system had to use on-line analytical processing techniques that allowed users to do dynamic reporting, ad-hoc analysis, and access predefined reports.
- The resulting system had to be easy to maintain and available for use by Fall 2001. HHES had to develop this review system in a very short amount of time. The analysts had very little time to review the data, so the programmers had to be able to fix any problems with the application in a short amount of time.
- The system had to allow for programming flexibility. The specifications for the system changed even as we were running the final system tests to release the system for production. Changes to the reports and formats had to be made as quickly as possible, without delaying the release of the system.
- The users requested forty-nine initial reports. These reports would cover eight topics that would be used to review the data for the 50 states, DC, and Puerto Rico (a U.S. commonwealth). Each report needed many hierarchies that allowed users to compare edited and unedited data.
- The system had to handle individual data files containing more than four million observations. For example, the person file for California contains more than 4.4 million observations. For all states, there were more than 40 million records. MDDBs ranged in size from one megabyte to one gigabyte. The size of the N-way tables in the MDDBs also varied, the largest of which was 1,844,790. The number of files was also a data management challenge: there was one detail data set and one MDDB for each subject area for each state or state equivalent.
- New data had to be incorporated into it on a “flow” basis easily, without affecting the data already in the system. We used three states as a test, but once those “test” states were approved, all the states had to be available for review. We also had to replace old data sets for states with new versions of the data sets as the new versions became available.
- The technology had to allow users to view several data groupings, or clusters of states, at once. For example, the Washington, D.C. metro area includes the District of Columbia and several counties each from neighboring Maryland and Virginia.

- The system had to be easy to deploy to several different user communities, each of whom have different PC environments and differing SAS skill sets.
- The system had to use SAS v8 since the functionality we wanted was not easily available in SAS v6.12.
- The system had to allow users to view summary data using formats, but when the users viewed the observations comprising a summary level, the unformatted data was to be shown.
- Of course, the data had to be displayed to the user as quickly as possible.

WHAT WE DID

HHES worked with SAS Consulting Services to develop a system using SAS v8.2. The system uses a traditional 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®.

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 key feature of the HOLAP approach is that a "proxy" cube is created as a template. The template MDDB stores the hierarchies, formats, and base table attributes to pass to the resulting HOLAP cube.

Users select a state of interest via a user interface that displays a gif file. The file contains a map of the United States that has hot spots for each state, regional groupings of states, and the entire U.S. Screen shots of the user interface are shown in Figures 1a through 1d.

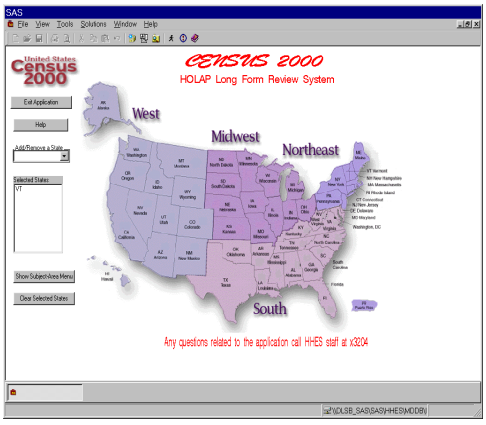

Figure 1a. The Census 2000 Long Form Review System Map User Interface

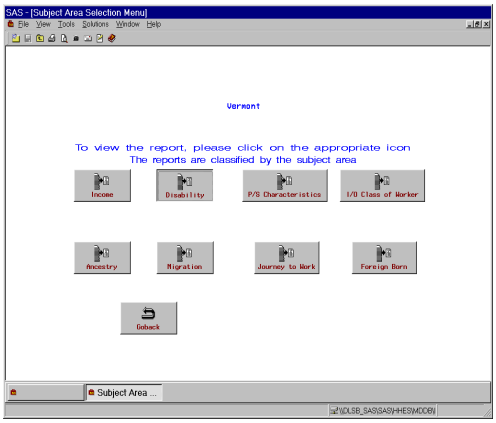

Figure 1b. The Census 2000 Long Form Review System Subject Area User Interface

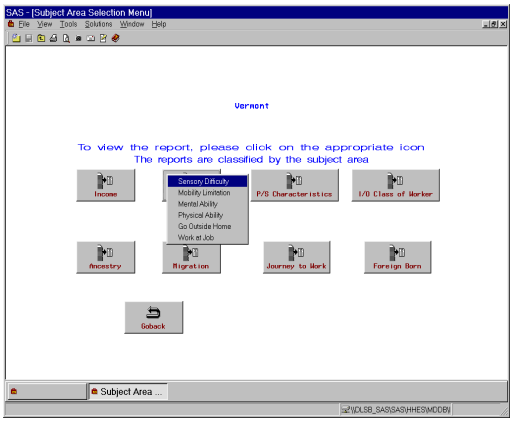

Figure 1c. The Census 2000 Long Form Review System Subject Area User Interface Continued

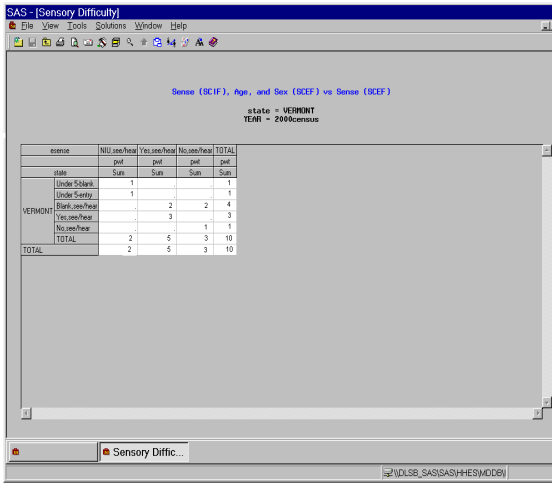

Figure 1d. The Census 2000 Long Form Review System Summary Report User Interface. Please note, the data shown does not represent actual Census 2000 data.

The desktop frame class is overridden so new SCL code that allows multiple states to be selected will execute. "Remote library services" are used to reach the MDDBs and the underlying detail data sets on the Unix servers. There is one detail data set and one MDDB for each state and subject area. The central metabase repository contains all the reports and the metabase registrations. However, the repository resides on a Novell server along with the application's configuration and profile files. This configuration gives the users access to the most current version of the reports without having to perform an installation procedure each time a report is modified or added. The SAS software used to run the system also reside on centralized servers. This configuration decreases installation problems as we do not have to modify software on individual PCs if a software upgrade is needed.

An override method captures the cell the cursor is sitting on when a "show detail data" is invoked. The SCL code determines what the cell represents, bypasses the HOLAP cube, goes directly to the data set that contains the detail data for the cell and subsets the data based on the cell's contents. This method delivers more than four million observations to the user's screen on average in 2 minutes and 20 seconds. The system takes more than 50 minutes to deliver the same four million observations to the user's screen when the override is not used.

THE RISKS

There were several risks involved in this project.

- HHES had never developed a system in SAS v8 before. The HHES programmers had only taken a "Differences Between SAS v6.12 and SAS v8" class. No one had any extensive experience working with SAS v8. HHES had never used the HOLAP technology in a production system before.

- HHES choose to develop two parallel systems. The first system did not use HOLAP and closely resembled systems we had previously developed in SAS v6.12. Programmers who were very familiar with the data were assigned to this system. Because we had substantial experience with traditional online analytical processing (OLAP) cubes, this system was easy to develop. However, because there was not sufficient time to create the 2,548 reports required to view the data manually, this was intended as a limited, backup strategy. The second system used HOLAP and programmers who were unfamiliar with the data were assigned to this system. Developing two parallel systems allowed us to create a backup system in case the HOLAP system failed, but it also had disadvantages. It spread our programming resources very thinly over two projects. The programmers who worked on the HOLAP system had to learn the intricacies of the data in addition to the HOLAP technology.

- HHES did not control the due dates for this project. The due dates were determined using information from several different areas at the Census Bureau and the dates were constantly changing. This made the project planning and scheduling the resources very difficult.

OUR RISK MITIGATION STRATEGY

HHES did several things to try to ensure the success of this project.

HHES used SAS Pilot Program to research methodologies before the actual production work began. Through this program a prototype HOLAP-enabled MDDB was developed, tested, and installed at Census. The prototype was used to determine proof of concepts, i.e., would a HOLAP application meet users' reporting and performance requirements. While the prototype was not a fully functioning system, it did allow HHES to run some benchmark tests to estimate whether or not a HOLAP system would meet the users' performance requirements in production. HHES pursued these pilot projects months before the production system. We were able to apply the lessons learned from those projects to our production system.

HHES also used SAS Consulting services as an expert resource. SAS Consulting was able to use what was learned in the pilot to identify several different methodologies HHES might use for the development of the production system. They also identified the risks involved in each methodology so HHES could make informed decisions and were able to help HHES implement the methodologies quickly and efficiently.

HHES chose to develop two parallel systems. Although it introduced other risks, we were able to develop a traditional OLAP system that could be used if the HOLAP project failed. Although HHES would not have been able to create all the reports needed in time to meet the production schedule, having a few of the first states available in another system bought the HOLAP project more development time.

The use of the HOLAP technology dramatically decreased the number of reports needed. If the OLAP system had been used, we would have had to create 2,548 reports, or one set of 49 reports for each state, DC, and Puerto Rico. In other words, there would have to be 52 copies of each report, since each report could access only one MDDB, or one state's data. If a problem was found in a report, all the copies of that report would have to be updated. There is no way to mass-correct the reports. Also, the risk of making mistakes increased when manually creating such a large number of reports.

The HOLAP approach allowed us to create a single set of 49 reports that could be used for any state. The reports can be created in advance, decreasing the time between file processing and file availability to the users. It allowed users of the system to view data for multiple states and multiple years in one report.

HHES racked the data. Racking involves partitioning a single sub-table into multiple physical structures based on the value of one or more dimension levels. HHES stored the data for each state and subject area combination in a separate base table and separate MDDB. HHES felt racking was more applicable to our business rules and processing methods. A common alternative to racking the data is stacking the data, in which each sub-table can be stored in one physical structure. Sub-tables containing non-additive dimensions are stored in separate files. If HHES had stacked the data, the data for each subject area would have been stored together in one base table and MDDB. Any time a new state was to be added to the system, HHES would have had to stop the review of all the states to add it. HHES would have also had to stop the review each time an existing state on the file had to be updated.

HHES also added some SCL code behind the map graphic that would check parameters in a SCL list that could be set to show if the entire system was not available and what states were not

available. The use of these parameters and the way we stored the data allowed us to update the data for an individual state without affecting any other states and the users' review of those other states. Figure 2 shows the message that appears when a state chosen is not available.

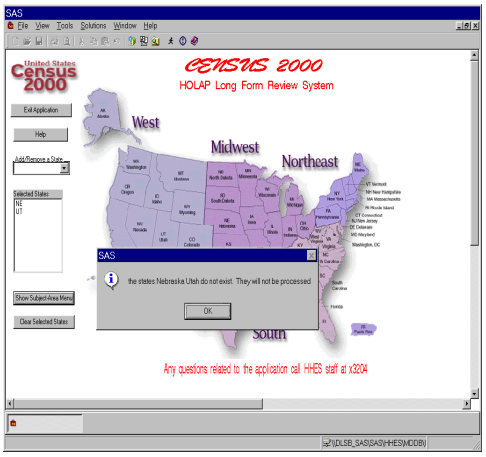

Figure 2. No Data Available for a Chosen State

RESULTS

HHES was able to deploy a working application in time for the review. Since we had developed a MDDB system in SAS 6.12 previously, most users had sufficient experience with MDDB applications. This lessened the amount of user training needed.

The application is flexible and adaptable and can probably be modified for the review of future, similar file types or comparisons to previous time series data.

The system performs better then previous systems. However, we continue to modify the application to improve its performance. Table 6 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 Census 2000 Long Form Review System Performance Statistics

State- California		
Subject	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

HHES explored and paid for alternatives that weren't needed. For example, to meet the requirement of showing formatted summarized data and unformatted detail data, HHES and the SAS consultants wrote SCL code to read in the Base SAS code used to create a format library and load the formats into the repository. Without this SCL code, HHES and the SAS consultants believed the formats would have to be manually entered into the repository. Later, HHES found that the same result could have been achieved by applying the formats to the detail data, creating the MDDBs and then removing the formats from the detail data in a subsequent data set.

CONCLUSION

The HOLAP capabilities inherent in SAS software provided us substantial flexibility and the ability to manage very large files encompassing more than 40 million records and numerous reports. Up front planning of both the system and data design elements allowed us to meet most users' functional requirements. Users' ideas were a critical part in the development of reports, hierarchical structures, and display formats. This HOLAP application works with some of the largest data stores at Census. It should be easily adaptable for other, similar review and tabulation-style systems.

REFERENCES

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