

Advantages and Disadvantages of Using MDDBs, HOLAP, and SAS/IntrNet® in the Development of an Interactive System

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

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. The three systems that were evaluated access data stored on centralized, large-scale Unix servers 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 3400 commonly used reports.
- The second system uses a client-server approach, SAS v8.1'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 74 reports had to be created.
- The third system uses a web-based approach, SAS v8.1's HOLAP and SAS/IntrNet® and other web products to display the data.

This paper describes the advantages and disadvantages of each technology used, including development, deployment, and performance issues, in addition to the space savings obtained via the use of formats.

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. 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 to the states, at a very detailed level of geography, 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 about 116 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 asked in Census 2000 was one that collected

information required by law to manage or evaluate federal programs or that was needed to meet legal requirements stemming from U.S. court decisions such as the Voting Rights Act. Federal 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, which will be used to review the "long form" questionnaire data, is scheduled to be in production in the fall of 2001.

SELECTION CRITERIA

HHES used the following criteria to decide which technology would be best suited for the Census 2000 Long Form Data Review system.

- The resulting system must be easy to maintain and enhance and available for scheduled use by fall 2001. HHES will have to develop a review system in a very short amount of time. The analysts will have very little time to review the data, so the programmers will have to be able to fix any problem with the application in a short amount of time.
- The technology must handle data files containing more than five million observations.
- It has to allow for programming flexibility. The specifications for the system might change even after the system is in production.
- New data will have to be incorporated into it on a "flow" basis easily, without affecting the data already in the system.
- The technology must allow users to view several groupings of data at once.
- The systems developed with this new technology will have to be easy to deploy to several different user communities, each of whom have different PC environments.
- Of course, the data have to be displayed to the user as quickly as possible.

THE SAS V6.12 PROTOTYPE

The first prototype HHES developed uses a client-server approach, SAS 6.12 under Windows95 and resides on user PC desktops. Each PC has to have SAS 6.12, and the application's configuration and profile files installed on it. The data reside on Sun Unix servers running the Solaris operating system. 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, the District of Columbia, and Puerto Rico. One metabase file holds all of the metadata for all of the states. The application files and the metabase file reside on a Novell server.

The data are displayed via a gif file that contains a map that links each state's hot spot to an icon that lists the reports for each state. This configuration only allows access to one state's data at a time. For this system, more than 3400 reports would have to

be created, with one set of 74 reports having to be created for each state. In other words, there would have to be 51 copies of each report, since each report could access only one MDDb, or one state's data. If a problem with a label or title were 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 increases when developing such a large number of reports.

It takes seven seconds to display a summary report. It takes two seconds to display 557,000 detailed observations from a data set containing 29,700,000 observations. Most of the objects used for this system are "off the shelf." No override methods are used. Minimal knowledge of Screen Control Language (SCL) was needed for the creation of this system.

HHES identified a substantial limitation in the use of SAS formats with version 6.12. Analysts want to see display formats that are more meaningful than the underlying coded values. For example, the report displays the words "male" and "female" in place of the coded values of "1" and "2". The problem occurs when analysts use the "show detail" option from the EIS multi-dimensional report object. No detail records making up the cell total are found because during the "reach-through" process, internally SAS was searching for "male," when in reality all the data values for males were stored as "1."

So, for reports where data ranges or labels have to be displayed, a separate field was created and the appropriate range was stored in it. Using another example, the "Drilldown 3 Highest Degree" report, shown in Figure 1, displays the number of people who have completed certain levels of schooling (EHIGH) by the edit parameters that were used to allocate the data if the response was blank or inconsistent (FLHIGH).

EHICH	01_000	02_No school completed	03_Nursery - 8th grade	04_5th or 6th grade	05_7th or 8th grade	06_9th	10_11th or 12th grade
FLHIGH	QWGT	QWGT	QWGT	QWGT	QWGT	QWGT	QWGT
SUM	SUM	SUM	SUM	SUM	SUM	SUM	SUM
01_Less than 3 years	25274						
02_Not changed		23108	48351	25669	34034	22110	
03_Allocated from matrix		4418	8395	3367	3753	1667	
TOTAL	25274	27526	56746	29036	37787	23777	

Figure 1. Drilldown 3 Highest Degree Report for the SAS v6.12 prototype. Please note, the data shown are from the Census Dress Rehearsal and do not represent actual Census 2000 data.

For each observation, a field called QHIGH contained a character code that stood for a certain level of education obtained. Data processing that takes place before the MDDb's are built created another field called EHICH, which contained the character string corresponding to the character code. Figure 2 shows a partial listing of the QHIGH and EHICH fields from the detail data set that goes into this report.

Obs	ehigh	QHICH
1	08_11th grade	07
2	06_9th grade	05
3	04_5th or 6th grade	03
4	08_11th grade	07
5	07_10th grade	06
6	03_Nursery - 4th grade	02
7	03_Nursery - 4th grade	02
8	14_Bachelors degree	13
9	10_High school grad.	09
10	05_7th or 6th grade	04

Figure 2. Detail observations used to create the Drilldown 3 Highest Degree Report for the SAS v6.12 prototype. Please note, the data shown are from the Census Dress Rehearsal and does not represent actual Census 2000 data.

THE HOLAP PROTOTYPE

HHES worked with SAS Consulting Services to develop the next two prototypes. The HOLAP prototype took advantage of new features found in SAS v8.1, but also retained all of the functionality contained in the SAS v6.12 prototype. The prototype's main feature is a new SAS technology called hybrid online analytical processing (HOLAP). This technology allows users to access data from multiple local and remote MDDb's and data sets as if the data were coming from a single data source.

A SAS view is created that allows concurrent access to the detail data sets. The view must be rerun each time a new state's data is added. This view is the input to a "template" MDDb. The template MDDb will store the hierarchies, formats, and base table attributes to pass to the resulting HOLAP cube. It does not have to be rerun every time a new state is added. It will need to be rerun only when the hierarchies, formats, or base tables change.

The HOLAP cube is stored in the central repository. It holds the location of each component of the logical data group and what it contains. It has to be updated each time a new state is added.

The use of the HOLAP technology decreases the number of reports that will have to be created, from the more than 3400 required in the SAS v6.12 prototype, to a single set of 74 reports that can be used for any state. The reports can be created in advance, decreasing the amount of time between file processing and file availability to the users. The HOLAP technology also allows users of the system to view data for multiple states and multiple years in one report.

The data are displayed via a modified version of the gif file used in the SAS v6.12 prototype. The file still contains the map, but for this prototype, there are hot spots for each state, regional groupings of states, and the entire U.S. The desktop frame class is overridden so new SCL code that allows multiple states to be selected will execute. It takes 12 seconds to display a summary report. Figure 3 shows how the "Drilldown 3 Highest Degree" report appears in the HOLAP prototype. The report still displays the number of people who have completed certain levels of schooling, but now the report shows data for multiple states. From this report, the data for any of the listed states can be expanded. Figure 4 shows the same report, but with data from both states expanded.

Drilldown report Allocation flag high education, uhigh, flgrade, egrade and age16 vs ehigh (SCEF)
Subset: STATEFP=California South Carolina YEAR=2000

EHIGH	01_NIU	02_No school completed	03_Nursery - 4th grade	04_5th or 6th grade	05_7th or 8th grade	06_9th grade	07_10th grade	08_
statefp	qwgt	qwgt	qwgt	qwgt	qwgt	qwgt	qwgt	qwgt
South Carolina	930777	1014060	2090093	1069887	1392660	876430	1057024	97
California	1769180	1926820	3972220	2032520	2645090	1664390	2006060	18
Total	2699957	2940880	6062313	3102407	4037750	2540820	3063084	28

Figure 3. Drilldown 3 Highest Degree Report for the HOLAP prototype. Please note, the data shown are test data created specifically to be shown to non-Census employees and does not represent actual Census 2000 data.

Drilldown report Allocation flag high education, uhigh, flgrade, egrade and age16 vs ehigh (SCEF)
Subset: STATEFP=California South Carolina YEAR=2000

EHIGH	01_NIU	02_No school completed	03_Nursery - 4th grade	04_5th or 6th grade	05_7th or 8th grade	06_9th grade
statefp	qwgt	qwgt	qwgt	qwgt	qwgt	qwgt
South Carolina	930777	1014060	2090093	1069887	1392660	876430
California	1769180	1926820	3972220	2032520	2645090	1664390
Total	2699957	2940880	6062313	3102407	4037750	2540820

Figure 4. Drilldown 3 Highest Degree Report from the HOLAP prototype with state data expanded. Please note, the data shown are test data. It was created specifically to be shown to non-Census employees and does not represent actual Census 2000 results.

"Remote library services" are also used in this system to reach the MDDBs and the underlying detail data sets on the Unix servers. There is still one detail data set and one MDDB for each state. The central repository contains all the reports and the metabase registrations. However, the repository has to reside on each user's PC, which must have SAS 8.1, and the application's configuration and profile files installed on it.

Another override method is used to correct a potential performance problem that HHES discovered. While in the "Wage/Salary Income" report, when only one state was chosen, a "show detail data" took 40 minutes to subset 269,104

observations from the 18,366,027 observations available in a test data set. SAS believes this performance problem is due to the use of the HOLAP cube, which uses a view to point to all of the detail datasets.

SAS does not use the indexes on the data sets when accessing the data via a view. Even if only one state was selected, SAS is reading all the detail data sets in the data group sequentially to pull out the selected data. The prototype now uses an override method that 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 allows the same 269,104 observations to be displayed in less than five seconds. A more in-depth knowledge of SCL is needed to maintain or update this prototype due to the use of these override methods.

The prototype also incorporated the use of reach-through of display formats to data set values so they did not have to be re-coded. This decreased disk space storage requirements with savings ranging from 32% to 63%. For example, the original education data set for Texas, which was created in SAS version 6.12 and had re-coded values stored in the data set, was 9 gigabytes. After the data set was converted to version 8.1 and the formats were applied, the file was 3.4 gigabytes, producing a 61% savings in disk space.

THE WEB-BASED PROTOTYPE

The third prototype takes advantage of new features used in the HOLAP prototype, retains most of the functionality contained in the SAS v6.12 prototype, but also takes advantage of the features of SAS/IntrNet®. This was the first attempt at the U.S. Census Bureau to build a web-enabled HOLAP system incorporating "reach-through" to the underlying detail data sets.

The Web-based prototype differs from the HOLAP prototype in how the data are displayed. The data are displayed via an HTML file that contains a map that has hot spots for each state, regional groupings of states, and the entire U.S. Javascript captures the mouse clicks and places the selections in the list box. A macro program takes the list of selected states and manipulates code to access the HOLAP cube that points to the selected data. The MDDB Report Viewer (MRV) that comes with SAS/IntrNet, is used to display the data in a report show, in figure 5.

Down: Uattend/racehisp/age16/age2 (hier)
Couple/tract/block (hier)
Flabil/citizen/yr2us/racehisp (hier)

Across: EATTEND
[Couple/tract/block (hier)]

eattend	01_NIU	02_No	03_Yes, public	04_Yes, private	TOTAL
statefp	qwgt	qwgt	qwgt	qwgt	qwgt
California	2059918	35135593	12357954	1686555	51240020
South Carolina	1221515	20838275	7325497	998482	30383769
Texas	3538360	60339720	21233100	2901220	88012400
TOTAL	6819793	116313588	40916551	5586257	169636189

Figure 5. A Summary of education data from the MRV portion of the Web-based Prototype. Please note, the data shown is test data. It was created specifically to be shown to non-Census employees and does not represent actual Census 2000 data.

Expanding the reports is time consuming. Response time varies with the size of files. For the MRV portion of the prototype, it takes seven seconds to display a summary report. It takes 80 seconds to display the first set of 50 detailed observations.

Sun's Web server is used with SAS/IntrNet[®] to access the MDDBs and the underlying detail data sets on the Unix servers. There is still one detail data set and one MDDb for each state. The HOLAP technology is still used and all the reports and the metabase registrations are still contained in a central repository. However, SAS v8.1, the application's config file, the application's profile file, and the repository reside on the Unix server. Nothing resides on the user's PC.

Setting up the SAS/IntrNet[®] server is very tedious. The auto installer did not work for this product, so the installation had to be done manually. The instructions were lengthy and written in pieces. HHES's system administrator could not get an overall picture of all the steps necessary to complete the task. Many tasks have to be performed to complete the installation. The slightest deviation from the procedure can cause the installation to fail. Knowledge of Unix system administration, web server software (Apache or Sun web server), and general installation procedures are needed to do this task. Extensive knowledge of Javascript will be required to maintain or expand this prototype.

FINDINGS

HHES compared the prototypes with each other, reviewing the benefits and disadvantages of each. HHES also evaluated the prototypes against the criteria to be used to decide which technology would be suited for the Census 2000 Long Form Data Review system. HHES could find no clear-cut winner based on this analysis. No prototype met all of HHES's criteria. All three prototypes could handle data files containing more than five million observations in them. All three prototypes can incorporate new data "flow" basis easily, without impacting the data that were already there. Only the prototypes that use the HOLAP technology allow users to view several groupings of data at once. The SAS v6.12 prototype can display the data faster than either of the SAS v8.1 prototypes. Table 1 shows various statistics HHES used when evaluating the prototypes.

	SAS 6.12 Prototype	HOLAP Prototype	Web-based Prototype
# of states that may be displayed at once	1	50	50
Time to display a summary report	7	12	5
# of obs / # of seconds to display them	557,000 / 2	269,104 / 5	89,326 / 76
Can handle new data on a "flow" basis	Yes	Yes	Yes

Table 1. Miscellaneous Statistics Used for Prototype Evaluation.

The SAS v6.12 prototype was the least flexible because of the number of reports that would have to be created. The HOLAP prototype and the Web-based prototype offer more flexibility. Changes to the report labels and titles that would affect all the states can be made quickly in the prototypes that use the HOLAP

technology, since each only has one set of reports that will access any of the states.

The SAS v6.12 prototype was the easiest to maintain, because it uses "off the shelf" SAS objects. The objects are well documented and the HHES programmers have a lot of experience using SAS v6.12.

The HOLAP prototype is the second hardest application to maintain, due to the use of the override methods and SCL code. Except for the override methods, this prototype is very similar to the SAS v6.12 prototype. The HOLAP creation process was very easy to understand, maintain and modify. There are objects in SAS v8.1 that would allow a user to choose which state to view. The use of this object would eliminate the need for the map and the need to override the desktop frame class. However, an override method had to be used to meet HHES's need for fast response time to display detail data. Organizations with more override method and SCL programming experience would be able to maintain the prototype with ease.

For HHES, the Web-based prototype is the most complicated and the hardest prototype to enhance or maintain. This assessment is based on HHES's lack of experience with Javascript, and the problems with the server installation. Organizations with more Javascript and SAS/IntrNet[®] server installation experience might be able to maintain the prototype with ease.

The Web-based prototype is the easiest application to deploy. Only a web-browser is needed on the user's PC. Both the SAS v6.12 and the HOLAP prototypes require that the appropriate version of SAS be installed on the user's PC. With a very slight modification, the prototypes could probably use a network version of SAS, but we have not tested this yet. The configuration and profile files needed by the SAS v6.12 prototype, can be copied to the user's PC via a network release package. HHES is not sure if the central repository in the HOLAP prototype can be deployed to the user's PC via a network release package. Another possible option for the HOLAP prototype deployment would be to have the central repository reside on a network server, but HHES has not yet tested the ramifications of this configuration.

CONCLUSION

HHES will have to conduct more research before drawing a final conclusion. HHES did not try tuning anything, nor did they fully load test the prototypes. HHES is leaning toward using the technology used in the HOLAP prototype for five reasons.

- There are fewer reports to create.
- SAS v8.1 is closer to the most current SAS version than SAS v6.12.
- The use of the HOLAP technology will allow the users to view groups of states instead of just individual states.
- The potential problems we might encounter with the override methods and SCL code in the HOLAP prototype will probably be overcome with additional training and internal mentoring from more experienced programmers. Unfortunately, the HHES programmers have the most SAS/IntrNet and Javascript experience of any programmers at the Census Bureau, so mentoring is not an option if we use the Web-based prototype technology.
- The benefits of using the HOLAP prototype technology outweigh the SAS v6.12 prototype's performance advantage.

Currently we are creating a SAS v8.1 HOLAP system using 1990 Census data to fully load test the technology before we go into production with Census 2000 data. It will allow us to try different tuning strategies. If these tuning strategies work on 1990 Census data, they should work on Census 2000 data. It will also allow us to research and test the other outstanding questions that we had with the prototype, while allowing the programmers to gain more override and SCL experience.

REFERENCES

SAS[®] is a registered trademark of SAS in the United States of America and in other countries

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