Paper 137

SAS® Software MDDBs Solve Real-World Problems for 1997 Economic Census

Shirin A. Ahmed, *U.S. Bureau of the Census*
Kimberly D. Yarbrough, *U.S. Bureau of the Census*

ABSTRACT

To accommodate the conversion from the Standard Industrial Classification (SIC) to the North American Industry Classification System (NAICS), the U.S. Census Bureau implemented a warehousing solution. Known as the NAICS Database System, it became pivotal to the 1997 Economic Census. It provided corporate data, access, and analytical tools using SAS® products — primarily SAS data sets, SAS/EIS® software, SAS/MDDB™ software, SAS/AF® software, and SAS programming language.

As a warehousing project, the NAICS Database System consolidates information from legacy systems across the economic area at the Census Bureau for data analysis and production of composite publications. Unlike other data warehousing projects, it differs in three respects. First, it is integrated into production systems for the Economic Census, and refreshed weekly. Second, its use is tied to actual deliverables in the publication release of first-time economy-wide *Core Statistics Reports*. Given these deliverables, third, the system is used not merely by researchers mining data for new opportunities, but rather by production analysts who have the job of reviewing and releasing data for publication. Approximately 200 analysts use the system.

Key Words: SAS, SAS/MDDB, SAS/EIS, SAS/AF, SAS/SHARE®, SAS/ASSIST®, Economic Census, Warehousing, NAICS, Reports

BACKGROUND

The Economic Census is a snapshot of the economy, with summary data released to represent more than six million employer businesses and industries. The Census Bureau conducts an Economic Census every five years, for years ending in 2 and 7. Collection of the data occurs in subsequent years ending in 3 and 8, respectively. The recent 1997 Economic Census posed unique challenges with the introduction of a completely new way to classify businesses and industries.

Known as the North American Industry Classification System, or NAICS, its introduction meant unifying the classification of establishments under one economic concept — the "production-orientation" of the establishment.[1] This differed from the Standard Industrial Classification (SIC) system which the Census Bureau used since 1930. NAICS meant industries such as bakeries, traditionally classified as retail, would now have selected establishments classified in manufacturing for 1997.

Typically, the Economic Census is processed by subject area. That means that after data are collected from establishments in 1998, decentralized analysis and publication of data occurs by subject (or industry) area from early 1998 through mid 2001. For example, establishments in manufacturing are analyzed and published separately from establishments classified in construction. Since analysts and applications developers are organized by subject area, the on-line applications processing systems are also by subject area. These legacy systems operate on DEC Alpha machines using primarily DEC relational databases (RdBs), DECForms for interfaces, COBOL, and FORTRAN. Analysts access legacy systems through PC workstations using a communications software package.

With the introduction of NAICS there arose, for the first time in Economic Census processing, a need for corporate data across all subject areas. A freeze on hiring from 1990 to 1995, required a solution without the costs of retooling existing legacy systems. To address this need the SAS-based application of the NAICS Database System became a primary development activity for the Economic Directorate, who has responsibility for conducting the Economic Census.

ABOUT THE NAICS DATABASE SYSTEM

Like other warehousing projects, there was a strong need to see corporate data across organizational barriers. The analysts in the subject areas needed access to historic data to accommodate switches from one subject area to another. Management needed a corporate data repository to judge the overall impact of NAICS. Fragmentation of data review by subject area left management without tools to ensure complete industry coverage from reclassifying all establishments. Management could correct errors found early through revised follow-up with respondents. The NAICS Database System provided this.

The NAICS Database System, however, is not typical of other data warehousing projects.[2] It differs in three ways:

❑ First, its need arose out of *processing and production concerns.* Like other production systems in the Economic Directorate approximately 200 analysts are accessing and using this application *as part of their work* to release statistics to the general public.

❑ Second, this warehousing project crossed subject lines that do the same type of work, as opposed to other warehousing projects which tend to cross functional organizational lines. This meant that data across the subject areas had some similarities and would ease the transition to developing standard definitions.

❑ Third, and more important, the corporate warehouse had *specific deliverables tied to it.* For the first time, corporate-wide publications of statistics would be released by the Economic Directorate. These *Core*

Statistics Reports are as follows:[3]

- Advance Report gives a high-level economy-wide look at what NAICS means for businesses and industries. The Advance Report is published February 1999.

- Comparative Report shows the 1997 NAICS data recast on a SIC basis *across all subject areas* to give data users comparability with historic data, specifically the 1992 Economic Census. The Comparative Report is published March 2000.

- Bridge Report shows how SIC-based industries are distributed among NAICS-based industries and, vice versa, how NAICS-based industries are distributed among SIC-based industries. This report is published in March 2000.

The NAICS Database System comprises a number of SAS data sets, searches, analytical reports, and publication reports delivered using primarily the products of SAS/AF and multidimensional databases (SAS/MDDBs). While 200 analysts in the Economic Directorate access the system for census work, another 100 will use the system as NAICS is implemented in Current Economic Surveys, starting in Spring 1999. The configuration is Client/Server with the application running on the PC and data sets residing on the DEC ALPHA. The DEC ALPHA serves as the host. Six servers running SAS/SHARE® software allow analysts to access 28 SAS/MDDBs and several data sets.

ABOUT THE SAS DATA SETS

The NAICS Database System comprises five critical data sets, which are described below:

❑ *1992 Micro Establishment Data Set*: This data set contains establishments canvassed in the 1992 Economic Census. Records or observations to populate this data set were obtained upon completion of publications for the 1992 Economic Census (around 1995). This data set provides historic information to the analysts. It represents final micro records for which no other analytical corrections to data are applied. Hence, this file remains static. File size is 5.6 million records and 27 variables. Once created, this file served in prototyping applications and publications.

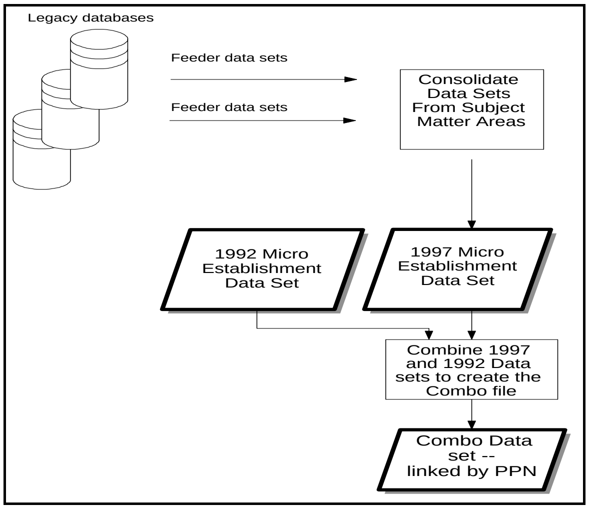
Note, for this and the other data sets, only a subset of basic data -- such as employment, payroll, and receipts -- is consolidated. Not all data collected on Economic Census questionnaires reside centrally for corporate access.

❑ *1992 Summary Data Set*: These records represent tabbed (or summary) data published for the 1992 Economic Census and, like the corresponding micro records, were obtained for this project in 1995. This data set contains historic information for input into the release of the *Core Statistics Reports*. Like its counterpart, this file remains static. Approximate file size is 15 million records and 25 variables.

❑ *1997 Micro Establishment Data Set*: This data set reflects the latest updated micro establishment records from the ongoing processing of the 1997 Economic Census. Unlike the previous two data sets, this data set changes periodically as shown in Figure 1.

Each week, the subject area programmers create feeder files from their legacy databases (RdBs). They release these feeder files — formatted in SAS data sets — for consolidation. In total, nine subject area files are released. The release occurs each week to account for corrections (transactions) analysts made to establishments on the legacy databases. Once obtained, feeder files are consolidated. The resulting file is the 1997 micro establishment data set.

Figure 1: Weekly Production Process


Legacy databases
Feeder data sets
Feeder data sets
Consolidate Data Sets From Subject Matter Areas
1992 Micro Establishment Data Set
1997 Micro Establishment Data Set
Combine 1997 and 1992 Data sets to create the Combo file
Combo Data set -- linked by PPN

The following code creates a view to the subject area feeder data sets then uses a DATA step to combine them in Permanent Plant Number (PPN) order. The PPN uniquely identifies establishments. The idea was to combine nine data sets (that have duplicate observations by PPN) in PPN order without sorting the output data set. Also, the code flags duplicate PPNs. (Note, viewing these duplicate PPNs for discrepancies and potential errors is a SAS/AF application as discussed later.)

The code is on the next page.

```
data est.vest97 / view=est.vest97;
    set naicsdiv.NAICSMIN
    naicsdiv.NAICSMAN
    naicsdiv.NAICSSER
    naicsdiv.NAICSWHL
    naicsdiv.NAICSRET
    naicsdiv.NAICSFIR
    naicsdiv.NAICSUTI
    naicsdiv.NAICSCAO
    naicsdiv.NAICSCON;
    by ppn;
run;

data est.est97(index=(ppn cfn_97));
    set est.vest97;
    by ppn;
    length dupe_97 $ 1;
    if first.ppn and last.ppn then
    dupe_97=’ ’;
    else dupe_97=’7’;
run;
```

The production process for the 1997 micro establishment data set started May 1998. At that time approximately one million records had been processed through the subject-specific (or legacy) systems. Each week, the data set grew as more establishments responded to the Economic Census and went through the traditional processing pipeline. By October 1998, the data set reached full size with 5.8 million records. Note that this data set carries 54 variables.

❑ *Combo Data Set*: Per Figure 1, once the *new 1997* micro establishment data set is available, it is combined with the *1992* micro establishment data set based on PPN. The resulting combo data set serves as the primary source for review tools. It allows analysts to review current and historic data simultaneously.

The file size is significantly more, approximately 8 million observations and 79 variables. The larger number of observations is due to unmatched records -- that is, establishments in 1992 not in business in 1997 (i.e., deaths), and new establishments in 1997 not around in 1992 (i.e., births). To meet access demand by analysts, the combo data set is indexed and sorted by classification (industry) and type of operation codes. Creation time with 11 indexes is 7 hours and 53 minutes.

❑ *1997 Summary Data Sets*: Two sources of summary data sets exist: summary data from SAS/MDDBs and summary data for publications. SAS/MDDBs are discussed in the next section.

For publications, data are extracted from the SAS/MDDBs in batch using SAS/AF SCL code developed in-house,[4] and stored in SAS data sets. Since the SASSFIO engine does not work in release 6.12 on the DEC Alpha machines, there was no better way to extract data. As discussed in the next section the summary data are derived from the combo data set.

SAS/MDDBs

Given the large read-only data sets and the need to collapse data for various tab levels, using SAS/MDDB software greatly improved the quality and efficiency of the NAICS Database System. The NAICS Database project started in Fall 1995. About one year later, the release of SAS/MDDB software with SAS 6.12 became a viable product for use with this project. Basic requirements called for numerous structured analytical reports. Additionally, analysts needed drill-down capabilities to lower tabbed cells and reach-through capabilities to review establishments defining a tab cell.

The SAS/MDDBs are built from the combo file using the MDDB procedure. They are created weekly after the combo data set is refreshed. There are three SAS/MDDBs (1992 SIC-based, 1997 SIC-based, and NAICS-based) created for each of the nine subject areas and one SAS/MDDB created for cross subject area tabulations.

Creating 28 SAS/MDDBs reduced the size of the subtables and decreased access time for the SAS/EIS reports. The SAS/MDDBs serve as the source for SAS/EIS reports. It takes 1 hour and 22 minutes to create all of these SAS/MDDBs.

A sample of the PROC MDDB code follows to show you the flexibility in using this software. For example, since the Census of Construction is a sample of establishments, you can weight the analysis variables as the MDDB is built. (Note, the estab_97 variable is weighted already from the feeder data set). In addition, you can use a where clause to build an MDDB for selected records. For example, this code only builds an MDDB for records with sictrade='9'. Finally, creation time was reduced significantly by indexing the base data set and using the (KEEP=) data set option.

```
proc mddb data=combo.combo
    (keep=sector  sictrade  naics6  naics4
    naics3
       sic_97  sic4  sic3   sic2   txton_97
    txtos_97
         st_97   estab_97   qp1_97   exp_97
    b_inv_97
      e_inv_97 purch_97 emp_97 ap_97 sr_97
      weight97)
    out=mddb.no_aux;
    class  sector  sictrade  naics6  naics4
    naics3
      sic_97 sic4 sic3 txton_97 txtos_97
      st_97 sic2;
    var  estab_97 / sum;
    var  qp1_97 exp_97 b_inv_97 e_inv_97
      purch_97 emp_97 ap_97 sr_97 /sum
      uwsum sumwgt weight=weight97;

    hierarchy sector sictrade;
    hierarchy sictrade sector;
    hierarchy sector naics4 txton_97;
    hierarchy sector naics4;
    hierarchy sector naics3 txton_97;
    hierarchy sector naics3;
    hierarchy sector;
    hierarchy sictrade;
    hierarchy naics6 sic4;
    hierarchy naics3 sic2;
       where sictrade^=’9’;
 run;
```

USER INTERFACE

Analysts access the NAICS Database System two ways. First, they can use SAS ASSIST, installed in two places — the DEC machines and their local PC networks. Once in SAS ASSIST analysts can query and produce their own reports.

In looking at the general experience level of the analysts, however, only few feel comfortable using SAS ASSIST for their tailored analysis. In building this system, as well as the legacy systems, serving the analysts’ needs required more structured interfaces. A far more appealing method for analysts is a simple SAS/EIS menu application that calls the SAS desktop and other SAS/AF applications. The high-level SAS/EIS menu of this project is in Figure 2.

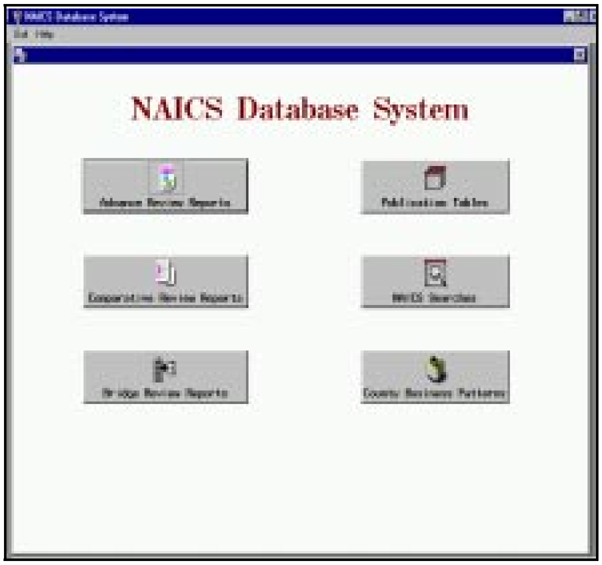
The SAS/EIS software provides easy menu building capabilities with a limited amount of effort. The desktop is perfect for displaying reports. Reports can easily be added or removed -- hence you minimize development and maintenance.

From Figure 2, the high-level menu system provides analysts with a variety of access options -- including the NAICS Review Reports (Advance, Comparative and Bridge), NAICS Searches, and Publication Tables.

ANALYTICAL AND PUBLICATION REPORTS

You access analytical reports by selecting any of the NAICS Review Reports from the high-level menu. Once you select a type of analytical report a menu asks you to select a specific subject area or the general corporate reports. A selection of reports for the construction subject area is shown in Figure 3.

Figure 2: High-level



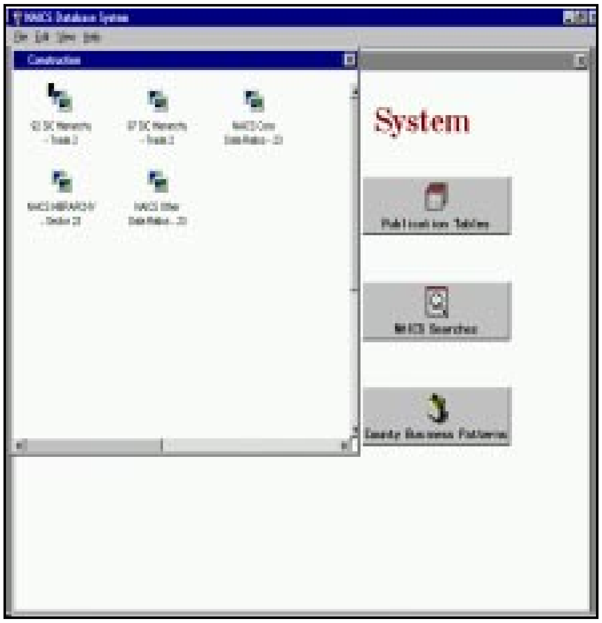
These reports, created from SAS/MDDBs, serve as the primary review tool. Data are tabbed on a current-to-historic SIC basis, on a NAICS basis, and on a geographic basis.

You can access predefined reports at various levels of classification. (Classifications at 6-digit levels define more detailed industry categories than classifications at 4-digit levels, and so on.) As an example, Figure 4 shows the Comparative Data Ratio report.

You can use the flexibility of SAS/EIS to subset data for specified classifications, create totals, remove columns not needed for your review, and so on. The coupling of the SAS/MDDBs and SAS/EIS reports provides large amounts of tabular data with countless ways to customize.

You can view the publication reports from a simple SAS/AF module containing a data table. These reports are created from SAS summary data sets, some of which are data extracted from SAS/MDDBs, as discussed earlier. These data sets feed into the existing publication process. The data in these reports are at broader publication levels. You
access these reports for final review prior to publication.

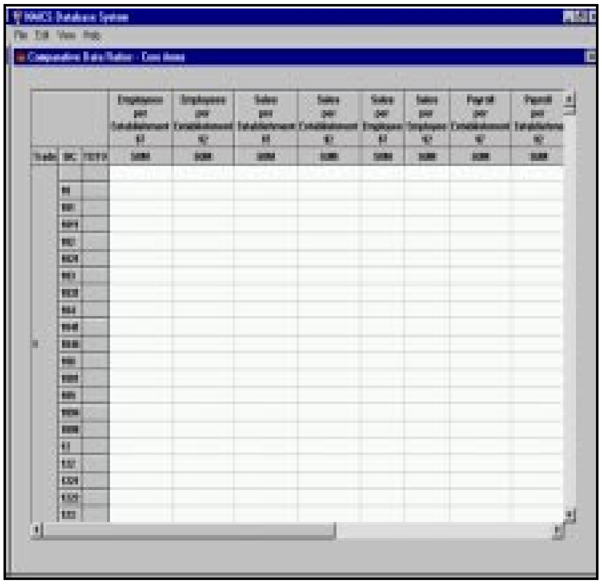
Figure 3: Analytical Reports



SAS/AF: SEARCHING[5] AND PPN DUPLICATES

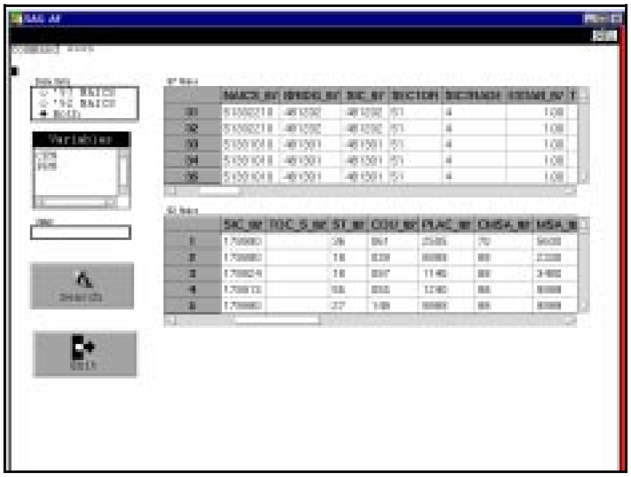
The search facility of micro establishment records is a basic SAS/AF application that displays two data tables. You can enter a PPN or Census File Number (CFN) and indicate which years of data to retrieve. Figure 5 shows

Figure 4: Sample Report



you the search screen. The search screen is used for troubleshooting inconsistencies with tabbed data.

Figure 5: Search Screen



Not shown are the PPN duplicate lists. These lists are SAS/AF frames with the data table class serving as the primary component of the frame. The data display are sorted subject area by PPN to identify duplicates. The second list is sorted by PPN only to determine subject area and other characteristics of the matching PPN. Both data tables are browse mode only.

CHALLENGES WITH THE APPLICATION

As with many large-scale projects, gathering and finalizing user requirements and overcoming technical issues became challenges in developing this application. When you work with a cross-section of subject areas on a common project, you have to address many differences in the analysis in coming up with common solutions. Negotiating agreement is time consuming. Additionally, as you prototype requirements the subject areas' needs -- while becoming better defined -- also change. Using SAS/EIS helped solve these differences by letting analysts customize displays once reports are available.

The project overcame technical issues. Although basic data are the same across subject areas, data are stored differently in legacy RdB files. Implementing standards required one year of effort, given the other competing priorities in processing the Economic Census. A second technical issue was the constraint of using existing hardware and software purchased. This application was not sized for a specific dedicated machine or supporting software. To get a workable Client/Server solution to meet access demand, a great deal of effort went into designing SAS/MDDBs that yielded subtables with <6,000 cells. The maximum cell size was determined by testing the amount of data passed across the network in 1 minute.

The last technical obstacle was that SAS was not using indexes on the base table during a "Show Detail Data" in a SAS/EIS report. Recall, accessing detail data directly from a summary cell (i.e., reach-through) was a basic requirement of this application and one reason for choosing multidimensional reports. SAS/MDDBs were advertised as quick retrieval of detail data. Upon using the "Show Detail Data" option against the 8 million record combo base table, however, it became clear that the base table indexes were not utilized. When reported to SAS, SAS technical support provided code to update the metabase so that the SAS/EIS report could utilize the indexes.

FUTURE IMPLICATIONS

The NAICS Database System reflects a growing business and technical change. Plans for the 2002 Economic Census call for consolidating the legacy RdB and supporting applications into common data structures and processes. This consolidation will enhance future warehousing and data mining initiatives.

DISCLAIMER

This paper reports the results of research, development, and implementation of a project undertaken by staff at the U.S. Bureau of the Census. It has undergone a more limited review than official publications. This report is released to inform and encourage discussion.

ABOUT THE AUTHORS

Shirin A. Ahmed
Assistant Division Chief
Economic Planning and Coordination Division
Bureau of the Census
Washington, DC 20233
sahmed@census.gov

Kimberly D. Yarbrough
Analysis and Programs Specialist
Economic Planning and Coordination Division
Bureau of the Census
Washington, DC 20233
kimberly.d.yarbrough@ccmail.census.gov

FOOTNOTES

[1] Carole A. Ambler and James E. Kristoff, "Introducing the North American Industry Classification System," *Government Information Quarterly*, Volume 15, Number 3, 1998.

[2] Kevin Strehlo, "Data Warehousing: Avoid Planned Obsolescence," *Datamation*, January 15, 1996.

[3] Mark E. Wallace, "NAICS Implementation -- Data Products," presented at the 1997 Economic Census Offsite, October 7-8, 1997. E-mail authors for copy.

[4] Mike Bretz, "Extracting SAS Data Sets from a Multi-Dimensional Database in Batch," Proceedings of the Eleventh Annual North East SAS Users Group Conference.

[5] Mike Bretz, U.S. Bureau of the Census, created the Search SAS/AF application.

TRADEMARK CITATION