

# DATA WAREHOUSING AT THE NATIONAL AGRICULTURAL STATISTICS SERVICE

## “EASY & FAST DATA ACCESS FOR EVERYONE”

**Jack Nealon and Mickey Yost**

**National Agricultural Statistics Service, U.S. Department of Agriculture**

Each year, several thousand data files containing agricultural survey data from farmers, ranchers, and agri-businesses are generated on different platforms, using different software systems, and different data definitions or metadata. This has caused an under-utilization of previous data in our survey process, which inhibits some improvements to our survey and estimation program. Therefore, NASS recently developed an easy-to-understand and easy-to-use Data Warehouse System that integrates previous survey and census data, and makes the data readily accessible to all NASS employees. This system answers the Agency's strategic need to improve the quality of our agricultural statistics program by making maximum use of previous survey and census data. Examples include: improving sampling efficiency, enhancing survey planning and administration, streamlining survey data collection, expanding analytic capability for all employees, and broadening statistical estimation opportunities.

Considerable research and evaluation were conducted to find the best Data Warehouse solution to satisfy our ambitious strategic objectives. This research, along with a great deal of senior management and user support, produced an easy-to-use and high performance Data Warehouse. About 700 NASS employees can readily access and review data from a single, integrated database, which contains over one-half billion survey responses extracted from several thousand data files. The Data Warehouse already contains data from over 80 different surveys from 1997, 1998, and 1999, and the complete set of census data (keyed, edited, and final) from the 1997 Census of Agriculture. Users of our integrated and generalized Data Warehouse must navigate only seven tables--a central data table containing all the survey and census data, and six corresponding dimension tables that provide all the necessary metadata to access the data readily. These dimension tables allow slicing and dicing of the data by any combination of rich analytic attributes, such as state, county, survey, commodity, questionnaire item, demographic characteristics, mode of data collection, sample design, and respondent. NASS's survey and census programs are benefitting from utilizing this powerful Data Warehouse System.

Three lessons learned that were critical to the success of our implementation are: (1) selecting the dimensional or star schema design for our Data Warehouse, (2) choosing a data base that is optimized for very fast data loading and data access using this dimensional design, and (3) selecting a primary data access tool that is very approachable by users using a drag and drop/point and click user interface. Currently, simple queries from our 500 million record database, such as deriving a weighted total for a census item for a county or retrieving all responses in a state for a survey item from a particular survey, are impressively resolved in less than five seconds.

Our live demonstrations will illustrate the three critical success factors mentioned previously, namely: (1) our easy-to-understand dimensional data warehouse design, (2) the rapid query response times for ad-hoc queries against our data warehouse, and (3) the ease of using the system.