

Data Warehousing and Decision Support at the National Agricultural Statistics Service, United States Department of Agriculture

Mickey Yost

Abstract

Easy access to large collections of historical survey and census data, and the associated metadata that describes it, has long been the goal of researchers and analysts. Solutions to problems such as understanding the behavior of current survey data, respondent burden, improved statistical techniques, and data quality are often found in the careful analysis of historical data. Many questions have gone unanswered, because the data was not readily available, access was limited, metadata was not well defined, or query performance was intolerably slow. This paper describes the database modeling techniques that permit end users fast and easy access to large amounts of micro-level data contained in different data systems and from different time frames. Also, techniques for tracking metadata changes and standardization are also discussed. A generalized dimensional model is presented that can be used for any census or survey to track the full history of the data series.

Keywords: Star Schema, Dimensional Model, Metadata, Integrated Data Sources, Database Design.

Background

The National Agricultural Statistics Service (NASS) administers the United States Department of Agriculture's program for collecting and publishing timely national and state agricultural statistics. In 1862 Isaac Newton, the first Commissioner of the newly formed Department of Agriculture, established a goal to "collect, arrange, and publish statistical and other useful agricultural information." A year later, in July 1863, the Department's Division of Statistics issued the Nation's first official Crop Production report.

Today, several thousand data files containing agricultural survey and census data from farmers, ranchers, agri-businesses and secondary sources are generated each year. These data files reside on different platforms, using different software systems, and different data definitions or metadata. This lack of data integration has created an under-utilization of historical data, inhibiting improvements to our survey and analytical procedures. From the beginning, NASS recognized the critical need for direct access to its historical data for analysis. Several reports and documents have been published, within NASS, recommending the use of historical data and the need to implement historical databases for analytical purposes. In late 1994, the NASS Strategic Plan formalized a new initiative to develop and implement a Historical Database (Data Warehouse) containing census and survey responses. In doing so, the plan acknowledged the critical role of historical data in improving customer service, reducing respondent burden, expanding analytical capabilities, enhancing sampling and estimation techniques, and improving

data quality. In 1996 NASS began work on an easy-to-understand and easy-to-use high performance historical Data Warehouse. At a minimum, the Data Warehouse would track previous survey and census data, including changes in specifications and metadata, and be readily accessible by all NASS employees, not just a few power users. It would also answer the Agency's strategic need to continually improve statistical analysis, survey and census procedures. Examples include: linking census, survey, and administrative data, improving sampling efficiency, enhancing survey management and administration, streamlining survey data collection, improving data quality, expanding analysis capabilities for all employees, and broadening statistical estimation methodology. Considerable research and evaluation was conducted during 1996 and 1997 to find the best Data Warehouse solution to satisfy our ambitious strategic objective. In April 1998 that research along with a great deal of support from senior management, and input from end users produced an easy-to-understand and easy-to-use Data Warehouse. Three months later, close to 700 NASS employees were accessing the Data Warehouse and reviewing, in one integrated database, survey responses from 35 different surveys extracted from over 1300 data files, and the complete set of census responses from the 1997 Census of Agriculture conducted during 1998. Currently, the database has grown to over 2 billion records covering over 602 surveys beginning with the January 1997 National Cattle Survey up to the most recent addition, the March 2004 Florida Citrus Survey.

1. A Brief Statement of the Problem

Internal reports articulating the strategic need for historical data, while making a powerful case, did nothing to show how such an endeavor might be accomplished. Indeed, when members of the original working groups that published these reports were interviewed, they said, in effect, the whole idea was a “pie in the sky”. The problem was one of understanding how the original data sets could be organized into a robust data model that would not only store historical data, but track all changes made to survey and census programs over time. Thinking up to this point had been rectangular. Each single data set had **N** number of observations by **P** number of variables. To combine these data sets into a rectangular model with over 1.5 million farms on the NASS list frame, and over 10,000 discrete survey items being surveyed every year was not possible. Tracking history using the $N \times P$ model, besides being very difficult to administrator and slow to query, quickly developed a severe sparsity problem. Not every farm produces the same commodity. Other database models were investigated including the standard entity/relationship (E/R) model. This model performed well for transaction processing, but could not support ad-hoc decision support queries. Ad-hoc queries were essential to understand current farm trends against historical farm trends for many different commodities. The E/R model also failed the database understandability test. People using the system could not navigate the hundreds of tables required in this model, and applications written to support analysis did not work for the ever-changing ad-hoc query.

2. The Star Join Schema

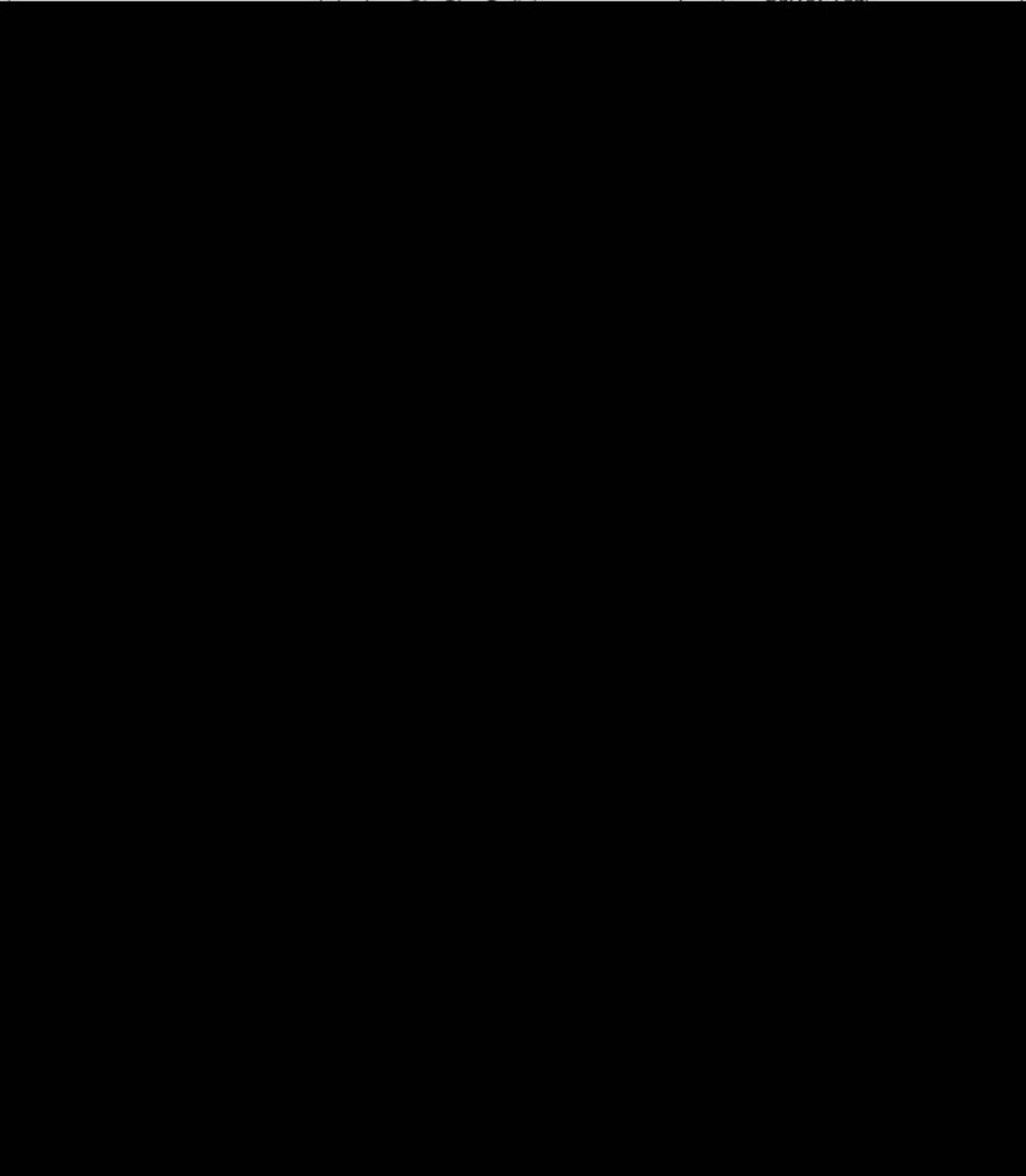
Enter the star join schema or the dimensional database model. The star join schema represents to the end user a simple and query-centric view of the data by partitioning the data into two groups of data base tables: facts and dimensions. Facts are the data items being tracked, and dimensions contain the metadata describing the facts. In the NASS model shown in **Figure 1**, facts are stored in the **Survey Responses**

table in the two columns labeled Cell Value and Weight. The cell value is the individual data response for a particular question from a particular survey or census, and the weight is the value used to adjust the cell value for such things as non-response, and the sample weights. The dimensionality of each fact is described by the words or metadata stored in the columns of the dimension tables. The keys make the link between the metadata (dimensions) and the facts. Notice the join lines connecting the keys in **Figure 1**. The dimension tables were built from the business rules that define a survey or census. The **Var Name** table, for example, contains the individual questions and their attributes for a particular survey or census, and is built prior to the loading of any data. It is necessary to pre-build each dimension table in order to have the correct combination of keys for each fact at load time. The **Location** table contains State and County names, and describes where the agricultural item was produced. Notice State is also in the **Reporter** table, but is the State of residence for the reporter, not where the commodity was produced. State is also in the **Var Name** table, but in this case it is an attribute of the question and is used to help end users navigate the thousands of questions being tracked in the table. The **Survey** table describes the event by which the data was collected, i.e. “1999 March Hog Survey”, or “1997 Census of Agriculture”. The **Sampling** table contains information on stratification, counts, and frames used to collect the data. The **Admin Codes** table contains information on the mode of data collection, the respondent, usability information, and the type of agricultural operation. The dimensions were chosen because they describe the business rules that govern the NASS survey and census programs, and are the “by” statements (slicing and dicing variables) for counts, sums, and ratios.

Using the dimensional attributes contained in the tables, data can be summed, counted, and analyzed by any of these attributes. For example, during the 1997 Census of Agriculture, data was being loaded into the **Survey Responses** table on a weekly basis. Reports were then produced that gave counts, sums, and ratios for the major agricultural items at the State and National level. Including other attributes, such as County from the **Location** table and Census ID from the **Reporter** table also produced reports by county and/or reporter. Direct comparisons of survey and census data at the individual reporter level were also possible, because the individual reports from both survey and census data sets from 1997 were stored in the Data Warehouse. Another example of using the Data Warehouse occurred during the 1999 June Agricultural Survey. Some of the questionnaires were returned with missing data. In a matter of seconds, all historical information on a respondent was retrieved and used to impute the missing values. Currently there are over 2 billion rows (cell values) covering state and national surveys from 1997 to the present.

In summary the E/R model is designed to model data relationships, and the star schema or dimensional model is designed to model the business rules. The fact table row with its unique set of keys on every row serves as a large business rule cross-reference table of metadata events that intersect at the cell value. When disparate data sets are combined in such a model, comparisons across and within surveys and censuses are quickly possible. Many different data comparisons using any of the rich dimensional attributes can be formulated and presented. Reporter classification and profiling becomes multi-dimensional. Many different aggregations can be calculated and displayed by summing to different dimensional levels.

Figure 1



The implications of this model are compelling:

- The dimension tables store the metadata about the cell value in terms familiar and understandable to the end user. Codes and their descriptions can be placed together, as well as comments and documentation about the data item.
- The dimension tables are attribute rich and hierarchical. Analysis can shift from a high vantage point with a broad set of attributes, to a very specific and narrow range of attributes depending on the study requirements.
- The dimension tables track additions and changes over time in all aspects of the survey program. New program content and questions, as well as small attribute changes, are tracked easily by adding additional rows to the appropriate dimension tables, rather than adding new columns to the fact table.
- The dimension tables are conforming. Dimension tables created for other roles or uses in new applications can retain the keys assigned in the original or master table. This way data from many sources can be combined together by any combination of key relationships.
- The cell value column in the fact table contains heterogeneous data. By transposing the column variables in the old rectangular data structure into individual rows in a dimension table, the fact table row now contains data on the complete range of attributes from all surveys. This effectively removes all sparsity from the fact table.
- The fact table stores the data at the lowest level of granularity – the cell value. This allows drilling down into any level of data.
- New fact tables can be created for any level of granularity or aggregation.
- New fact tables can reuse any dimension table keys that span the facts being tracked, and use new dimensions created for those facts.

The star join schema, therefore, represents the model of choice for on-line integrated data access organized by dimensions that end users can understand, remember, and easily navigate.

3. Metadata and the Dimensional Model

The dimensional model is an elegant relational database model for organizing and accessing survey metadata. These tables serve the needs of end users by providing, among other things, on-line access to survey and questionnaire specifications, reporter profiling, data classification, and interviewing practices. The metadata is rich and organized visually and in tables that reflect the way the business of the Agency is actually conducted. No attempt is made to remove redundancy within the table, as is done with a modeling process known as normalization. In the E/R model, the need to normalize is critical. This speeds transactional updates and saves disk storage. An E/R modeler would look at the **Location** table and immediately create at least three more tables: one table for state names, one for district codes, and one for counties. In the dimensional model, the **Location** table has a row for each state, district, and county combination. This is a problem if your design is to make updates to state, district, or county since it slows down the update process. Our **Location** table only requires updating if a new state or county is added to the United States. In a data warehouse, the design is successful if end users can navigate the tables, pose

business queries, and get results quickly. The redundancy aids browsing the dimension tables, and discovering the appropriate attributes for analysis.

In the **Var Name** table, for example, a new row is added for each change in state code, survey source, data type, or varname. This is the principle method of tracking changes in our survey program specifications. If Texas, for example, wants to start tracking data on ostrich, a new row is added to the **Var Name** table for the new question including the state name of Texas and all other appropriate attributes. The column named Varname Master is the variable that links all like variables from the different surveys being tracked.

This is a very important result, because the prospect of standardizing the metadata is now entirely possible. Attempts at standardization in the past were doomed to fail, because the data administrator could not accommodate both the original nomenclature, and the new standard. End users were asked to give up their original names in the name of progress. This model treats the original source name as an attribute of the new standard Varname Master, thus allowing its retention in the dimension table. End users wanting to do analysis using their old and familiar naming conventions may do so, while analysis across data sources can use the Varname Master to link all like variables together.

The **Reporter** table and its attributes are handled analogously. For example, the decision was made to add a new row to the **Reporter** table when a change occurred to the attribute Classify Year. Classify Year refers to the year a reporter was classified into a particular stratum for sampling purposes. For example, a reporter classified as a farm in 1997 has a row in the **Reporter** table with Classify Year = 1997. If that same reporter is classified as a farm in 1998, a new row is added to the **Reporter** table with Classify Year = 1998. Major attributes such as the ID, and certain demographics are carried forward to the new record. In this way minor changes to the name, address, certain demographics, and other non-major changes can be tracked for the particular ID. We have uncovered occurrences of a reporter's name changing completely from one classify year to the next, due to major updates being made to the name, and the ID remaining unchanged.

4. Issues with the Star Schema Model

As stated above, by converting all of the columns in the old flat file databases into rows in a dimension table, the result was a narrow two data column table with very little sparsity. Every row in the fact table has a Cell Value, and for more recent data, a weight. The model is also completely generalized. If a new variable, or a new reporter, or a new survey is needed, all that is required is the addition of a new row to the appropriate dimension table.

There are potential problems with this approach because of several issues associated with using a star schema, such as:

- How to handle comparisons among data in the same column. This is not a new problem with the star schema, because of the nature of the tables and their intended use. This tends to create fact tables that are long and narrow as the double column design demonstrates.
- Dimensional ad hoc analysis against a star schema requires multi-table joins between the dimensions and the fact table. Most query optimizers designed for transaction processing will default to a pairwise join strategy, or the joining of two tables at a time, on all related tables being queried for analysis. Of course, this will probably involve the very large fact table, because it is related to all the other dimensions. This can have a very limiting influence on the analysis if the intermediate result set

is too large or must be sampled.

- Referential integrity of the data being loaded into the fact table and the dimension tables. Referential integrity refers to the forced requirement that fact table data must have valid dimension table keys that reference that data back to the dimensions. If data are loaded into a billion row fact table, and there is a missing reference back to the dimension tables that is not enforced at load time, the data item or items loaded into the fact table will be forever lost.

There are other important issues in using a star join schema such as indexing strategies, load times, and disk drive segmentation by time periods. There are currently database engines on the market that solve many of these problems. NASS purchased one of these engines when they acquired Red Brick Data Warehouse System. Informix Data Systems recently purchased this company. Oracle worked from 1996 to 1998 developing an optimizer that would recognize the star schema. Other companies have small niches in the Data Warehouse market space such as Sybase and Microsoft, but are not considered up to the task of handling the extreme issues of very large databases.

5. Conclusion

Since the first official Crop Production Report, NASS statisticians have grappled with the need to understand their data. There are many influences on the data used to set official agricultural estimates and opportunities for error, both sampling and non-sampling errors. It is the tracking of these influences and the potential for modeling them against the estimates that give the data warehouse its true appeal. Every aspect of the business of creating official estimates, from planning and conducting surveys to statistical methodologies, and data analysis, will be influenced by this new technology. Productive and efficient analysis requires knowledge of the inputs that produce a given output. Data alone does not fulfill this requirement, because it does not carry along the information or metadata about the inputs and how they interrelate. This information and knowledge, in the past, have been separated from the data. It may have been available, but only in other disparate data sources, or in manuals and E-mail, or in programs, or in the hip pocket of an analyst. The star join schema represents a relational database model that gathers a great deal of this information and knowledge about the data, stores it, organizes it, and then relates it directly to the factual data being analyzed.

The richness of this information was not available in the transaction models. The emphasis there was on data, not on information. The end user or analyst was dependent on the Information Technology professional or power user to get at the data and report it in such a way that analysis could be performed. If further analysis was required, the process was repeated. The relational star join schema, on the other hand, simplifies the transaction model greatly and is designed for information gathering by the end user. It is an elegant software solution that presents data to the end user in the familiar and understandable terms of the business.

As information from the data warehouse is used in analysis and decision making, there will be a strong influence on all the processes that create our end product, the official estimates of U.S. agriculture. Operational systems that are choked with both operational and historical data will be freed up to operate more efficiently. As these systems are freed of excess data, re-engineering for efficiencies and quality will be less of a challenge. This re-engineering of tasks and procedures will occur, not because the warehouse needs it that way, but because the warehouse will help uncover data errors and inconsistencies resulting

from these tasks and procedures. Perhaps, and most importantly, the information in the warehouse will be used strategically to help carry out the long range goals of the Agency, which is the real reason the data warehouse is a key element in the NASS Strategic Plan.

References

Adamson, C., Venerable, Michael (1998). Data Warehouse Design Solutions, New York: John Wiley & Sons, Inc.

Kimball, R. (1996) The Data Warehouse Toolkit. New York: John Wiley & Sons, Inc.

Nealon J. (2000) Improving Our Agricultural Statistics Program Through Easy and Fast Access By Employees to Previous Survey and Census Data. Staff Report, National Agricultural Statistics Service, 2000

Yost, M. (2003) The Impact of a Data Warehouse on the Survey Process. Proceedings of the Association for Survey Computing -- Forth International Conference, 2003

----- (2000) Using a Dimensional Data Warehouse to Standardize Survey and Census Metadata. Proceedings of The 1999 Federal Committee on Statistical Methodology Research Conference, 2000

About the Author

Mickey Yost is currently a Mathematical Statistician with the National Agricultural Statistics Service in Washington, D.C., and currently serves as the Head of the Data Services Group in the Information Technology Division. He graduated from the University of California with a BS in Agricultural Economics, and did graduate work in mathematical statistics at the University of Texas. He is recognized as a leader in database designs for data warehousing, and has conducted numerous training and consulting seminars on tracking survey and census data using data warehousing architectures. He can be reached at 1400 Independence Ave., SW, Room 5847-S, Washington, D. C. 20250, Mickey.Yost@usda.gov.