USING A DIMENSIONAL DATA WAREHOUSE TO STANDARDIZE SURVEY AND CENSUS METADATA

Mickey Yost and Jack Nealon
National Agricultural Statistics Service, U.S. Department of Agriculture

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. Many questions have gone unanswered, because the datasets were not readily available, access was limited, and information about the business metadata was inconsistent, not well defined, or simply unavailable. This paper focuses on the database modeling techniques that aid in the standardization and tracking of survey and census metadata. A generalized dimensional model is presented that can be used for any census or survey to track the full history of the data series and to standardize the metadata.

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

Background

Each year several thousand data files from hundreds of surveys are created containing agricultural survey and census data from farmers, ranchers, agri-businesses and secondary sources. These data files are generated on different hardware platforms, use different software systems, and have varying and inconsistent data definitions or metadata across surveys. This lack of data integration and standardization contributes to an under-utilization of historical data, and inhibits survey efficiencies and analysis. In late 1994, the NASS Strategic Plan formalized an initiative to develop and implement a Historical Database (Data Warehouse) that would integrate all of our survey responses and the metadata. Metadata in this context is the information describing such things as the question text, the mode of data collection, reporter attributes, sampling specifics, and survey descriptions. The requirement that data definitions or metadata be standardized was not formally stated in the NASS Strategic Plan, but was tacitly implied by the requirement that our historical survey data would be integrated into a single database. In 1996, NASS began work on an easy-to-understand and easy-to-use high performance historical Data Warehouse. Considerable research and evaluation was conducted during 1996 and 1997 to find the best Data Warehouse solution to satisfy our ambitious strategic objectives. At a minimum, the Data Warehouse would track previous survey data, and changes in specifications and metadata. It would also be readily accessible by all NASS employees, and not just a few power users. The model would need to be robust enough to handle new surveys or data sources that might arise in the future and not require a redesign effort. In 1997 that research produced an easy-to-understand and easy-to-use integrated Data Warehouse of all the major agricultural surveys conducted that year by NASS. That same year the USDA became the sole source of all Official Agricultural Statistics when Congress transferred the responsibility for the Census of Agriculture from the Commerce Department to USDA. This transfer required integrating not only our own survey data and metadata into the Data Warehouse, but the newly acquired census data and metadata as well. This was achieved in the summer of 1998 when close to 700 NASS employees were able to access, in one integrated database, both survey and census responses. Currently, the database has grown to over 500 million records covering 3 years, 80 surveys, and the 1997 Census of Agriculture.

Dimensional Database Modeling for Survey and Census Data

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 an easy-to-administer data model that would not only integrate and track historical data, but manage the metadata 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 (**N**), and over 10,000 discrete survey items being surveyed every year (**P**) did not seem possible. Tracking history using the NxP 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 user understandability test. People using the system could not navigate the hundreds of tables required for this model, and applications written to support analysis were slow to respond to new analytical requirements.

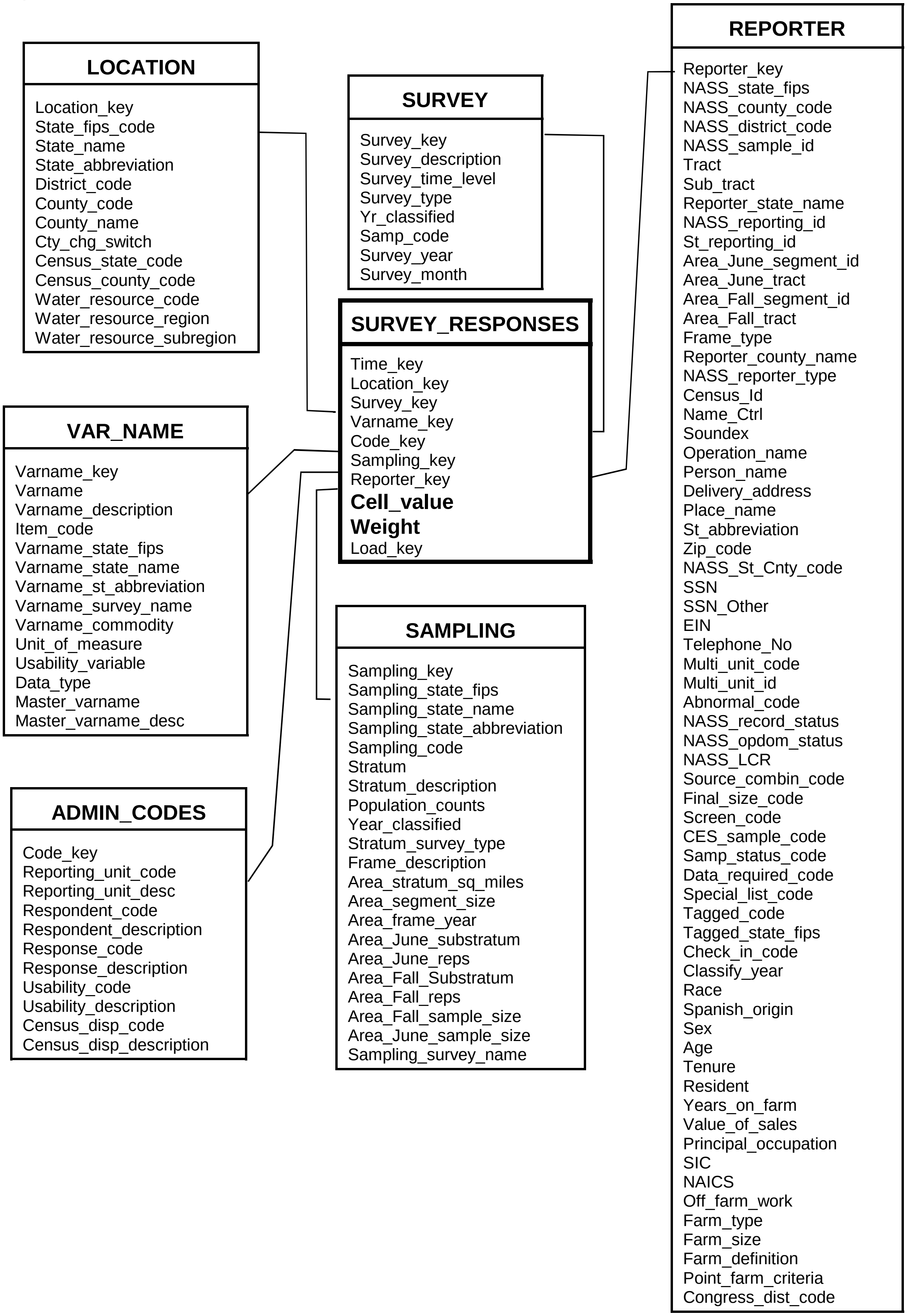
Enter the star join schema or the dimensional database model. The star join schema represents to the end user a simple and query-centric view of the data by partitioning the data into two groups of tables: facts and dimensions. Facts are the data items being tracked, and dimensions contain the metadata describing the facts. It is helpful to point out that dimensional modeling uses a top-down design approach that relies on the statistician's understanding of the survey and census processes to determine the facts and the dimensions that are in the database. 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 sample weight. The dimensionality of each fact is described by the metadata stored in the columns of the dimension tables. The **Var Name** table, for example, contains specifics about questions that are used for a particular survey or census. The **Location** table specifies in what state and county the agricultural item was produced. The **Reporter** table contains detail information on persons responding to a survey and/or census. This table, incidentally, is highly confidential, and can only be accessed by approved NASS employees. The **Survey** table describes the events by which the data were 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 to 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 or metadata 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. Adding detail, such as County from the **Location** table and Census ID from the **Reporter** table produced new break level reports by county and/or reporter. Direct comparisons of survey and census data at the individual reporter level were readily provided, because 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 for missing values.

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.

Figure 1



LOCATION

Location_key
State_fips_code
State_name
State_abbreviation
District_code
County_code
County_name
Cty_chg_switch
Census_state_code
Census_county_code
Water_resource_code
Water_resource_region
Water_resource_subregion

SURVEY

Survey_key
Survey_description
Survey_time_level
Survey_type
Yr_classified
Samp_code
Survey_year
Survey_month

SURVEY_RESPONSES

Time_key
Location_key
Survey_key
Varname_key
Code_key
Sampling_key
Reporter_key
Cell_value
Weight
Load_key

VAR_NAME

Varname_key
Varname
Varname_description
Item_code
Varname_state_fips
Varname_state_name
Varname_st_abbreviation
Varname_survey_name
Varname_commodity
Unit_of_measure
Usability_variable
Data_type
Master_varname
Master_varname_desc

ADMIN_CODES

Code_key
Reporting_unit_code
Reporting_unit_desc
Respondent_code
Respondent_description
Response_code
Response_description
Usability_code
Usability_description
Census_disp_code
Census_disp_description

SAMPLING

Sampling_key
Sampling_state_fips
Sampling_state_name
Sampling_state_abbreviation
Sampling_code
Stratum
Stratum_description
Population_counts
Year_classified
Stratum_survey_type
Frame_description
Area_stratum_sq_miles
Area_segment_size
Area_frame_year
Area_June_substratum
Area_June_reps
Area_Fall_Substratum
Area_Fall_reps
Area_Fall_sample_size
Area_June_sample_size
Sampling_survey_name

REPORTER

Reporter_key
NASS_state_fips
NASS_county_code
NASS_district_code
NASS_sample_id
Tract
Sub_tract
Reporter_state_name
NASS_reporting_id
St_reporting_id
Area_June_segment_id
Area_June_tract
Area_Fall_segment_id
Area_Fall_tract
Frame_type
Reporter_county_name
NASS_reporter_type
Census_Id
Name_Ctrl
Soundex
Operation_name
Person_name
Delivery_address
Place_name
St_abbreviation
Zip_code
NASS_St_Cnty_code
SSN
SSN_Other
EIN
Telephone_No
Multi_unit_code
Multi_unit_id
Abnormal_code
NASS_record_status
NASS_opdom_status
NASS_LCR
Source_combin_code
Final_size_code
Screen_code
CES_sample_code
Samp_status_code
Data_required_code
Special_list_code
Tagged_code
Tagged_state_fips
Check_in_code
Classify_year
Race
Spanish_origin
Sex
Age
Tenure
Resident
Years_on_farm
Value_of_sales
Principal_occupation
SIC
NAICS
Off_farm_work
Farm_type
Farm_size
Farm_definition
Point_farm_criteria
Congress_dist_code

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. As more and more surveys are loaded into the Data Warehouse, the dimension tables begin to serve as the clearing house for new survey or census metadata. For example, questionnaire design efforts can use these master tables to select appropriate and standardized questions. Survey processing systems can, likewise, make use of the common definitions for system variables that are tracked in the tables. The **Var Name** table is a prime example of this result. Stored in the Master Varname column is the variable that links all like variables from the different surveys being tracked. The original name of the variable that was used by the system of record is maintained in the column named Varname. This column preserves the history of that system. The Master Varname and Description become the standard across all surveys being tracked, and are used in all future design efforts. See **Table 1**, a report taken directly from the Data Warehouse. The column labeled Varname is the original source system variable name used for editing, imputation, analysis, and summaries. The description is also presented to show the wide variation in name and description for the same item. This is typical in non-integrated systems. The column labeled Master Varname is applied across the variations creating a standard name and definition. All of this occurs in one table so end users can easily browse and see the relationship. For the analyst doing research on data linked to the original name, that choice is available, because the original name is preserved. Hence, the mechanism at NASS to track history (the dimensional model), is the very mechanism used to standardize future surveys and censuses. 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 varnames for progress sake. 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. Thus, standardization of survey and census metadata is achieved.

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, 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 facilitates the gathering of 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 and metadata to the end user in the familiar and understandable terms of the business, and facilitates standardization of metadata across all surveys and censuses.

Table 1

TABLE SHOWING ORIGINAL SOURCE SYSTEM VARNAME AND DESCRIPTION AND NEW STANDARD MASTER VARNAME AND DESCRIPTION

Varname	Varname Description	Master Varname	Master Varname Desc
C018	CORN ALL GRAIN HARVESTED ACRES	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C133	CORN ALL GRAIN HARV AC.	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C133	CORN ALL GRAIN HARVESTED ACRES	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C202	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C203	CORN ALL GRAIN HARVESTED ACRES	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C531	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
C543	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
CCRNXXHV	ALL CORN - HARV GRAIN	**CCRNXXHV**	CORN ALL HARVESTED ACRES
CCRNXXHV	CORN ACRES HARVESTED	**CCRNXXHV**	CORN ALL HARVESTED ACRES
CCRNXXHV	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
CCRNXXHV	CORN ALL GRAIN HARVESTED ACRES	**CCRNXXHV**	CORN ALL HARVESTED ACRES
CNAHCURR	CORN ALL GRAIN HARVESTED ACRES	**CCRNXXHV**	CORN ALL HARVESTED ACRES
HARVGRN	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
IC321	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES
K67	ACRES OF FIELD CORN FOR GRAIN HARVESTED	**CCRNXXHV**	CORN ALL HARVESTED ACRES
W305	CORN ALL GRAIN HARV AC	**CCRNXXHV**	CORN ALL HARVESTED ACRES

References

Adamson, Christopher, ET AL., 1998, *Data Warehouse Design Solutions*, New York: John Wiley & Sons, Inc.

Devlin, Barry, 1997, *Data Warehouse from Architecture to Implementation*, Reading: Addison Wesley Longman, Inc.

Inmon, W. H., 1993, *Building the Data Warehouse*, New York: John Wiley & Sons, Inc.

-----------------, 1997, *Managing the Data Warehouse*, New York: John Wiley & Sons, Inc.

Kimball, Ralph, 1996, *The Data Warehouse Toolkit*, New York: John Wiley & Sons, Inc.

------------------, 1998, *The Data Warehouse Lifecycle Toolkit*, New York: John Wiley & Sons, Inc.

Poe, Vidette, 1996, *Building the Data Warehouse for Decision Support*, Upper Saddle River: Prentice Hall PTR

Red Brick Systems, 1996, *Star Schema and STARjoin^{TM} Technology*, Los Gatos: White Paper