Indian Agricultural Data Warehouse Design

Anil Rai, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi, India 110012; E-mail: anilrai@iasri.delhi.nic.in
Sree Nilakanta, Iowa State University, Ames, IA 50011, USA; E-mail: nilakant@iastate.edu
Kevin Scheibe, Iowa State University, Ames, IA 50011, USA; E-mail: kscheibe@iastate.edu

ABSTRACT

Data warehouse implementations at the sector levels, especially at the national agricultural level are non-existent. Designing an agricultural data warehouse poses unique and significant challenges because traditionally the collection and dissemination of information have been extremely parochial. Moreover, there has been very little adoption of information technology. Recently the Government of India has embarked on an ambitious project of designing and deploying a data warehouse for the agricultural sector with the intent of using the system for macro level planning decisions. The paper presents some of the design challenges the project has faced and solutions undertaken. It is hoped that the paper will help other such large scale, sector level warehouse designs learn from our experiences.

Keywords: Data warehouse, agriculture, dimensional model, warehouse architecture

INTRODUCTION

Since the 1990s, data warehouses have been an essential information technology strategy component for many medium and large sized, global organizations. Data warehouses provide the basis for management reports, decision support, and sophisticated on-line analytical processing and data mining. A data warehouse is a repository of data that is aggregated and summarized from operational systems to provide decision making support and is subject-oriented, integrated, time-variant, and nonvolatile [7].

Historically, data warehouses have been implemented in banking and financial institutions, retail marketing of consumable and non-consumable goods/services, and telecommunication services. The architectural designs for these types of warehouses are similar with differences usually occurring due to warehouse size and system analysis complexity [9]. A major reason for the similarity in designs of data warehouses across these industries is the ability of each organization to collect data at the finest level of granularity. Another reason for data warehouse similarity is the stability of the organizations and their respective industry sectors.

Data warehouses are often developed to address the business process requirements of a single organization or division. While many unique architectural designs exist across a myriad of companies and industries, these designs all share a common single company or division scope characteristic, and for good reason. Despite the growth in data warehouse development, there is little evidence for warehouses addressing the needs of large holding companies (having multiple organizations) or entire industry sectors and government agencies. Industry sectors and specifically, government agencies have data sources and decision requirements that are significantly different from other firms [6].

The government sector is one area that data warehouse technology can benefit tremendously to support regional, national and global decision making. Of particular interest to this research within the government sector is agriculture. Roughly 70% of the India's population depends on agriculture for its livelihood. Policy decisions within this sector not only directly or indirectly affect its people but also its agri-business industries such as seeds, fertilizers, plant protection, etc. Given the diversity of sources, formats, and subject areas, collecting and integrating such heterogeneous information poses a challenging task for data warehousing. There are few examples of data warehouses at sectoral levels. The earliest development of a sectoral level data warehouse is perhaps by the National Agricultural Statistics Service of the US Department of Agriculture. Their warehouse brought together data from agricultural surveys and census data from ranchers, farmers, agri-businesses, and secondary sources [12]. Another example of a sectoral level data warehouse contains data on pest, pesticides, and meteorological data for the government of Pakistan [13].

While the need for sector level data warehouses for macro economic planning and decision-making has been great, these types of warehouses have been almost non-existent because of the difficulty in coordinating flow and integrating data from the many member organizations. Almost every government sector collects vast quantities of data, but only a fraction of those data are used for planning and decision making. Several factors contribute to this problem; member organizations are often independent, autonomous entities with their own data requirements – namely formats, naming conventions, measurement units, etc. Furthermore, little, if any interaction exists among the different members. Escalating the problems of data integration is that these organizations may collect data at different granular levels. Moreover, many governmental bodies rely on different government, semi-government, non-government and private organizations for data collection, and when the information is collected from one organization's perspective and not from a sector or national perspective, data and systems may exhibit a parochial and protectionist perspective. Therefore, data integration for sector level use becomes a formidable challenge [6].

In this paper we investigate and present the challenges in designing and deploying a data warehouse for the Indian agricultural sector. The observations and classifications made in this research are important at two levels, (1) other sector level organizations may use the information presented to avoid pitfalls when designing data warehouses, and (2) researchers interested in large scale, multi-organizational data warehouses may use the information and actual working data warehouse to direct their research.

AGRICULTURAL SECTOR IN INDIA

Indian agriculture is highly diversified in its climate, soil, horticultural crops, plantation crops, livestock resources, fishery resources, water resources and so on. The diversity of its agricultural sector is a bane and boon to the social, economic, and cultural bases of India's vast population. Moreover, the diversity among resources generates interactions among many different macro and micro factors, and is further complicated with the interdependencies that exist among these. The Indian Council of Agricultural Research, New Delhi under World Bank funded National Agricultural Technology Project has developed a data warehouse for some of these agricultural resources to (1) improve the Indian Council of Agricultural Research's organizational and management system efficiency, (2) enhance scientific research performance and effectiveness to benefit farmers, and (3) encourage farming community participation through innovation and improved technology management. Objectives one and two are being implemented by the Indian Council of Agricultural Research, and objective three is being implemented by the Ministry of Agriculture in 28 districts of seven states. The implementation of the third objective, innovation of information technology, is the impetus for this research, and the Indian Agricultural Statistics Research Institute, New Delhi has been charged with building the data warehouse.

Sources of Information

India is divided into 28 states and six union territories (UT). Each state/UT is further divided into districts (elementary administrative unit) [5]. The district is the basic unit of administration for all purposes. The collection of Indian agricultural information is conducted through multiple organizations throughout the country. There are many national and state level boards and organizations for each agricultural sector. These information collecting agencies operate in the interest of their client organization, often specific to a region or state. Because there are many different data collection agencies and equally diverse resources for which the information is collected, there exists information heterogeneity. This problem is compounded by a lack of common standards applied to data collection. To use the information at the macro planning and decision making level, data must be integrated and aggregated properly [4].

Critical Dimensions

National level planning and decision support processes require access to data on many different resources, such as crops, livestock and fisheries, at varying levels of detail [2]. Information on production (demand and supply), price levels, and population and migration statistics is also expected. Location, Time, and Product are a few of the common dimensions that transcend all warehouse models, but Location and Time pose the biggest problems in integrating data from the varied sources in the agricultural sector. The integration problem may be categorized into one of four common dimensional issues, (1) granularity of Location differs among the different sources, (2) granularity of Time varies among the different sources, (3) several overlapping time domains and (4) aggregation and disaggregation of information at different dimensional hierarchies. These dimensional issues lead to the design the fact table and, therefore, the architecture of the data warehouse.

Location granularity: Similar to sectors such as retail and telecommunications, the agricultural sector uses the Location dimension extensively for its warehouse applications. In the Indian agricultural sector, the Location dimension presents many interesting issues. Location, also known as the Geography dimension, usually has a clearly defined hierarchical structure. In our case this hierarchy is determined by administrative mechanisms and put into effect by the Indian government. Level one is the National level. Level two is State. India is divided into 28 states, often on a linguistic basis. Each State is further divided into Districts (Level three) which may be further divided into Villages (Level four). This level is the lowest that agricultural sector information is collected through agricultural surveys.

Organizations may collect information at any or all levels of the Location hierarchy. Different sample surveys are conducted to acquire production figures of commodities such as fruit crops, plantation crops, etc. at the State level. Statistics of national accounts and different sectors of economy are mostly available at this level. Because each state is somewhat autonomous, the information collected at Level two is very important for state level planning and decision making. Production information is available at the District level for crops, livestock products, fisheries products, land use statistics, etc. Due to its detailed measure of factors, information at this level is very important to planners and decision makers at all levels. The Village level has data such as land use, census data, livestock, and demographic and static parameters such as land ownership and employment. Another Level four attribute is agricultural commodity trades available in Agricultural markets or Mundi (trading place). Price data from many important markets are collected on daily or weekly basis depending on the season of the crop or commodity. Finally, different agro-meteorological stations produce information on climate and weather conditions on a daily basis and form another fourth level hierarchy attribute.

Another challenge presented with the Indian agricultural data warehouse is historical data. Information on production of some commodities is available at the district level, but historical data are only available at the State level. Availability of resources, requisite need for information, and governmental policies present at that time affect the collection at any given level. These resources include human and financial capital and time. The following issues are associated with creation of dimensions in the development of data warehouse:

- The number of levels needed for any location.
- The integration of information coming from different sources (organizations) at different grain levels.
- The definition of fact table for these dimensions.

Aggregation rules to roll up each of the fourth level hierarchy of the Location dimension to the next higher level are different. In the case of villages, it may be a simple aggregation, but in agricultural market where the condition is price, a simple aggregation does not work. Availability of agricultural markets in different states for a commodity depends on its area, production, and consumption.

Integrating information from different sources, especially from various organizational sources, is also a big challenge in the design of the warehouse. Data collection takes place at different levels (e.g. National, State, District) using different methods (e.g. surveys, census, observations) and by different organizations, each with its own formats, procedures, and objectives. Further, definitions, concepts, and purpose are likely to be different for different parameters. Moreover, each source and method contributes to different types of errors. Despite these issues, if information is available at the lower level it is possible to aggregate (roll up) to the higher level. However, when information is only available at a higher level it is very difficult to disaggregate (drill down) to lower level [3, 11]. Most information about agriculture is collected through agricultural surveys or census, which are designed to elicit responses at the National or State level. Regional or lower level estimates can not be obtained from these with reliable precision.

PROTOTYPE

Warehouse Architecture

We propose the use of multiple dimensions to resolve the problems described above. A similar solution is also recommended in click stream data warehouses for enterprise relationship management (eRM) applications [10]. In a click stream data warehouse, the focus is on capturing the mouse clicks, sites and products visited, as well as the time and decisions taken by the consumer. We differ from the click stream warehouse implementation in that we use different fact tables for each type of dimension associated with the Location dimension and its hierarchy. We use the livestock data mart to illustrate our solution.

The data warehouse design employs the Star schema concept in which the central fact table is connected to the dimensions in a star like fashion. The foreign key links from the fact table to the primary keys of the dimension tables yield the star configuration. Each star schema configuration yields a data warehouse cube. Because the software supported only star schema the resulting cubes appear disconnected. Otherwise, they are connected through the respective common dimensions. Integrated livestock surveys are conducted every year for collecting data about the production of livestock products. The Census or complete enumeration for livestock is conducted after every five years. Some of the information in this data mart has been collected only once. Among the data collected, the period over which the data falls differs. Time at which data are collected differ because of the different calendars used. Three calendars, namely Calendar year (January 1 to December 31), agricultural year (July 1 to June 30) or financial year (April 1 to March 31) are employed. Since each time measure is important, it raises the possibility of three independent time hierarchies.

In data warehouse design, fact table grain has to be decided first [8]. In most warehouse designs, the decision is dependent on the level of detail the fact should address, namely, the business process performance measure. Because we do not have all available measures at the lowest level of detail, several fact tables have to be designed. For example, if measures corresponding to the Location dimension are available only at the State level then the grain of the fact table will be fixed at that level. So the granularity of the dimensions will result in many fact tables.

Granularity of Time

Generally, for time dimension in the agricultural sector the lowest grain level is day but many measures are available only at the weekly, monthly, quarterly, and half yearly, or yearly level. Climatic data such as rainfall, humidity, and temperature are available daily. Prices for different commodities and products from different agricultural markets of the country may be available at daily, weekly or monthly for a calendar year. Production measures of food crops, horticultural crops, and plantation crops are always available annually based on the agricultural year. Some of these crops are perennial and others are produced in one, two or three seasons in different parts of the country depending on the climatic, soil and water conditions. Information from human census is available after every ten years while for livestock it is available after every five years. All other socio-economic data are available annually based on the financial year. Keeping in view the above diversified grain level of the time dimension it is a challenging task to develop a data warehouse in

which all these sectors of agriculture are integrated on a common homogeneous platform. The complexity is further raised when availability of information for time levels follows different definitions. In India information about agriculture is available following three definitions of a year as follows:

Calendar Year
Years start from January 1 and ends on December 31. The months are in accord with the Julian calendar months. The first week starts from January 1 irrespective of name of the day and weeks are generated by counting seven days in a week. Last week, 52nd week of the year, consists of eight days to make 365 days in a year. For leap year last week of February consist of eight days.

Agricultural Year
Year starts from July 1 and ends at June 30. The first month of the year is July and the last month of the year is June. Months are assembled similar to the calendar year. The first week of the year will start from July 1 and it will be generated as per the procedure of calendar year. Similarly, first quarter and half year will start from July and generated as per the rule of calendar year.

Financial Year
Year starts from April 1 and ends March 31. The first month of the year is April and last month is March. Months are as per calendar year. The first week of the year will start from April 1 and it will be generated as per the rule of calendar year. Similarly, the first quarter and half year of the year will start from April and generated as per the rule of calendar year.

Because the three types of years have different start and end times, our data warehouse needs three independent hierarchies in the time dimension. The overlapping time periods poses significant difficulties in integrating data from the sources. The following table shows the month number of each year i.e. calendar, agricultural and financial year with respect to the months of calendar year.

Table 1. Time dimension properties

S.No.	Name	Calendar Year Number	Agricultural Year Number	Financial Year Number
1	January	1	7	10
2	February	2	8	11
3	March	3	9	12
4	April	4	10	1
5	May	5	11	2
6	June	6	12	3
7	July	7	1	4
8	August	8	2	5
9	September	9	3	6
10	October	10	4	7
11	November	11	5	8
12	December	12	6	9

Table 2. Time dimension quarterly

S.No.	Starting Month	Ending Month	Calendar year quarter No.	Agricultural Year Quarter No.	Financial year quarter No.
1	January	March	Q1	Q3	Q4
2.	April	June	Q2	Q4	Q1
3.	July	September	Q3	Q1	Q2
4.	October	December	Q4	Q2	Q3

Table 1 shows that integration of the information available at the grain level of months for different year types will not have any problem irrespective of their definitions. Let us consider Table 2 for quarters for each type of year, namely calendar, agricultural and financial year with respect to calendar year:

The integration of the information from different sources at grain level of quarters may not have problems in case of India as the definitions of different year such as calendar, agricultural and financial years are offset from each other by multiples of three months. If the offset is different, as it may be in other countries, it may not be feasible to integrate the information available at the grain levels of quarters for the different year types.

In case the information is available at grain level of half year with respect to any year type, it is possible to integrate the information of the half years of calendar year with half year of the agricultural year because as per the definition, the offset between calendar year and agricultural year is six month. Therefore, the first half year of the calendar year corresponds to the second half of the agricultural year. Any information available at the grain level of half year with respect to financial year may not be integrated with the information of the half year of other two year types.

The information available at the weekly grain level of any type of year may not be integrated between weekly information with any other week of year type. The beginning or the ending of weeks of one-year type does not correspond to the beginning or ending of the weeks of any other year type.

The Current Status of the Project
The data warehouse project was developed with funding from the World Bank and under the National Agricultural Technology Project (NATP) initiative. Several goals and users had been defined for the project. Three types of users were identified for the system. They were (i) research managers, (ii) research scientists, and (iii) general users at IASRI and other research institutes and agencies. At one level, the system was to provide systematic and periodic information about the entire agricultural sector to research scientists, planners, decision makers, and development agencies. At a different level, different users would have the capabilities to use various decision support capabilities through an on-line analytical processing (OLAP) application.

The data warehouse project has been carried out with active collaboration of 13 institutions that operate under Indian Council of Agricultural Research (ICAR), New Delhi. Each of these institutions deals with one or more areas of agriculture. Fifty-nine different databases have been used as source feeds for the data warehouse. The data in these databases are gathered from council and research projects on various agricultural technologies in operation and from published official sources (related agricultural statistics). At a minimum, data from year 1990 onwards, at the district level, are integrated into this system. Many of these databases have statistical information dating to year 1950. In building the central data warehouse, we started by creating subject-oriented data marts and multi-dimensional data cubes. These are published and now available by way of Intranet and Internet access. The validation checks have been put into effect wherever possible.

The data warehouse system also provides spatial analysis of the data with the help of a Geographic Information System (GIS). Data mining and ad-hoc querying are also extended to a small set of users. The web site of the project is already launched (www.inaris.gen.in) and the multidimensional cubes, dynamic reports, GIS maps and information systems are already available to the users.

CONCLUSION
Data warehouses at sector levels are seldom seen. Governments at national and state levels and industry groups and regulatory organizations have begun to realize the potential of integrating data from many sources and using data warehouses to implement such solutions. We have presented here some of the problems that arise in integrating data collected in the Indian agricultural sector. We discuss specific problems associated with granularity of location and time, two key dimensions for an agricultural warehouse. We then present one solution in our prototype using Oracle software.

REFERENCES

1. Agency, A.-A.T.M. NATIONAL AGRICULTURAL TECHNOLOGY PROJECT - NATP, 2001, Describes the NATP project.
2. Azad, A.N., Erdem, A.S. and Saleem, N. A Framework for Realizing the Potential of Information Technology in Developing Countries. International Journal of Commerce & Management, 8 (2). 121-133.
3. Feijoo, S.R., Caro, A.R. and Quintana, D.D. Methods for Quarterly Disaggregation without Indicators; A Comparative Study using Simulation. Computational Statistics and Data Analysis, 43 (1). 63-78.
4. Hollihan, M. The Relationship between Inflation and Relative Prices. Studies in Economics and Finance, 6 (1). 29.
5. India, N.I.C. Districts of India: A Gateway to Districts of India on the web, 2004, Districts of India.
6. Inmon, B. BI in the Government DM Review, 2003, 26-27.
7. Inmon, B. What is a Data Warehouse? PRISM, Prism Solutions, Inc., 1995.
8. Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses. John Wiley & Sons, 1998.
9. Lou, A. Data Warehouse Size Depends on the Size of the Business Problem DM Review, 2003, 16.
10. Sweiger, M., Madsen, M.R., Langston, J. and Lombard, H. Clickstream Data Warehousing. John Wiley & Sons, 2002.
11. Waichler, S.R. and Wigmosta Development of Hourly Meterological Values from Daily Data and Significance to Hydrological Modeling at H. J. Andrews Experimental Forest. Journal of Hydrometerology, 4 (2). 13.
12. Yost, Mickey, "Data Warehousing and Decision Support at the National Agricultural

Statistics Service, United States Department of Agriculture," www.advance-datatools.com/BrioUserGroup/mtg200404/3_pillars2.pdf, accessed August, 10, 2005.

13. Abdullah, A, Brobst, B., Umer, U, " The Case for an Agri Data Warehouse: Enabling Analytical Exploration of Integrated Agricultural Data," in Proceedings of IASTED International Conference on Databases and Applications, Innsbruck, Austria, Feb. 2004.

Figure 1. Yearly animal population mart

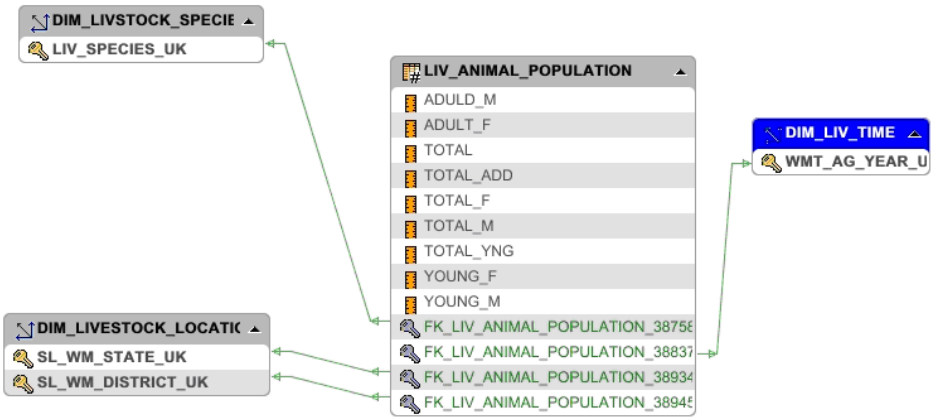

Figure 2. Graphical representation of buffalo over time