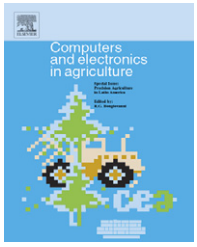
COMPUTERS AND ELECTRONICS IN AGRICULTURE 60 (2008) 263–278

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/compag



Dimensional issues in agricultural data warehouse designs☆

Sree Nilakanta[a], *Kevin Scheibe*[a,*], *Anil Rai*[b]

[a] *Iowa State University, Ames, IA 50011, United States*
[b] *Indian Agricultural Statistics Research Institute, New Delhi 110012, India*

ARTICLE INFO

Article history:
Received 13 November 2006
Received in revised form
7 September 2007
Accepted 10 September 2007

Keywords:
Data warehouse
Agriculture
Dimensional model
Warehouse architecture

ABSTRACT

Recently, the government of India embarked on an ambitious project of designing and deploying the Integrated National Agricultural Resources Information System (INARIS) data warehouse for the agricultural sector. The system's purpose is to support macro level planning. This paper presents some of the challenges faced in designing the data warehouse, specifically dimensional and deployment challenges of the warehouse. We also present some early user evaluations of the warehouse. Governmental data warehouse implementations are rare, especially at the national level. Furthermore, the motivations are significantly different from private sectors. Designing the INARIS agricultural data warehouse posed unique and significant challenges because, traditionally, the collection and dissemination of information are localized.

1. Introduction

Since the 1990s, data warehouses have been an essential information technology (IT) strategy component for medium and large sized global organizations. Data warehouses provide the basis for management reports, decision support, and sophisticated on-line analytical processing (OLAP) and data mining. A data warehouse is a repository of data taken from operational systems, aggregated and summarized to provide decision support. Data warehouses are subject-oriented, integrated, time-variant, and nonvolatile (Inmon, 1995; Inmon, 2002). Modern organizations are data-rich, but information-poor (Hoffer et al., 2005), meaning that organizations collect and store many facts, but those facts rarely translate into meaningful information. The purpose of data warehouses is to take that vast amount of data from many internal and external sources and present them in meaningful formats for making better decisions. While data warehouses meet an important business need, their design and implementation present many challenges (e.g. data quality, accessibility, and appropriate design (Agosta, 2001)). Approximately 40% of data warehouse projects fail to meet their design objectives (Kelly, 1997; Whiting, 2003). The development of data warehouses is costly with a typical project cost over $1 million in the first year (Watson and Haley, 1997). The reason for data warehouses failing is often not due to technical issues but lack of managerial support, lack of funding, lack of user involvement, and organizational politics (Watson et al., 1999; Wixom and Watson, 2001).

Data warehouses have been implemented in a variety of industries including banking and financial institutions, retail marketing of consumable and non-consumable goods and services, telecommunication services, and manufacturing (Hoffer et al., 2005; Inmon, 2002; Whiting, 2003). One area that has been slow in acceptance and implementation of data warehouses is government (Bieber, 1998; Harper, 2004; Inmon, 2005). There are many reasons for this lack of governmental data warehouse adoption. Reasons include issues over data ownership (Bieber, 1998), data privacy and security (Harper, 2004), and attitudes that are incompatible with the develop-

☆ We thank the editor and reviewers for their suggestions to improve the quality of this article.
* *Corresponding author.*
E-mail address: kscheibe@iastate.edu (K. Scheibe).
0168-1699/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2007.09.009

ment of the data warehouse (e.g. "Well that's not how we did it 10 years ago," "If I bring in a data warehouse, I am not going to need as many people," and "My tour of duty is only two years. We won't have much of a data warehouse built in that time, so it is going to hurt my chances of promotion to the next rank.") (Inmon, 2005). Even with these challenges, the government sector can benefit greatly from the adoption of data warehouse technology. Indeed, many governmental organizations are implementing such technologies. For example, the Iowa Division of Criminal & Juvenile Justice Planning has implemented a data warehouse to improve statistical and decision support information for justice system activities (Iowa Division of Criminal and Juvenile Justice Planning, 2007). The U.S. Department of Health and Human Service's Geospatial Data Warehouse (U.S. Department of Health and Human Services, 2007) provides improved health care information for underserved areas. One of the earliest developments of a governmental data warehouse is the US Department of Agriculture's National Agricultural Statistics Service. This data warehouse brought together data from agricultural surveys and census data from ranchers, farmers, agri-businesses, and secondary sources (Yost, 2000). A final example of a government data warehouse contains data on pests, pesticides, and meteorological data for the government of Pakistan (Abdullah et al., 2004).

Clearly, the government sector can benefit tremendously, from data warehouses, by supporting regional, national and global decision-making. However, as Inmon points out, government agencies have data sources and decision requirements that are significantly different than the industry (Inmon, 2003; Inmon, 2005). The primary motivation for data warehouse development in industry is increased profits or improved market share. Whereas, in order to serve and protect national interests, governments demand more accurate data, faster data access, lower costs, and better data integration.

Of particular interest to this research is the governmental sector of agriculture – specifically India's agriculture. Roughly, 70% of India's population depends on agriculture for its livelihood. India is large in size and population, and its people depend heavily upon agriculture. Improved production of agricultural resources benefits India economically and helps meet its basic food needs. Food shortages are common because of waste caused by inadequate storage and processing technologies (Tribune News Service, 2001), and while the nation financially depends upon export, there are restrictions placed by the government to ensure nationals have enough to eat (Modi, 2007). Consequently, better use of resources in tracking and monitoring agricultural production and consumption would provide not only economic benefit to India, but also food for the nation. Policy decisions within the agricultural sector not only affect individuals but also agri-business industries such as seeds, fertilizers, plant protection, livestock, etc. Because of the diversity of sources, formats, and subject areas, collecting and integrating such heterogeneous information presents a challenge for data warehouse development.

The need for sector-level data warehouses for macro economic planning and decision-making has been great, yet these types of warehouses have been scarce due to the difficulty in coordinating flow and integrating data from the disparate member organizations. Almost every government sector collects vast quantities of data, but only a fraction of the data is 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 organizations. Escalating the problems of data integration is the granular level of data that these organizations collect. For example, one may collect yearly data while another may collect weekly data. Moreover, many governmental bodies rely on other government or private organizations for data collection, and when the data are collected from a specific organization's perspective and not from a sector or national perspective, data and systems may exhibit a parochial and protectionist perspective. This data heterogeneity is a well known data warehouse problem (Bouguettaya et al., 1998; Lenzerini, 2002; Rahm and Do, 2000; Sheth and Kashyap, 1993; Widom, 1995). Therefore, data integration for sector level use is a formidable challenge (Inmon, 2003).

Aggregating data for an agricultural data warehouse is different from other industry sectors. For example, one cannot easily add together crop data from fields as if they are retail products, and averaging yields over an area is not the same as averaging sales across stores because units of measure are inconsistent across districts. Often, a district gives crop production estimates per block to the data-collecting agency, where a block may have a varying amount of cultivation. Consequently, when district data are combined they often indicate more or less production than actually occurred. Special algorithms are used to make accurate estimations such as small area estimates (Ghosh and Rao, 1994).

This research describes the development of a government-level data warehouse, and includes a description of challenges and strategies to address those challenges. Specifically, we investigate and present the design and deployment of a data warehouse for the Indian agricultural sector. The value of this research derives from the design process and observations 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. The remainder of this paper is organized as follows: Section 2 introduces some basics of data warehouse design and implementation. Section 3 describes India's agricultural sector, the need for data warehousing and the design and deployment challenges for this specific project. This is followed by our particular data warehouse architecture implemented. Section 5 details the evaluation of the Integrated National Agricultural Resources Information System (INARIS) data warehouse, including user satisfaction and usage statistics. Finally, we discuss implications of this research and conclude.

2. Data warehouses

Data warehouse design has unique characteristics compared to traditional database design. This is, in part, because data warehouse design depends upon already existing database

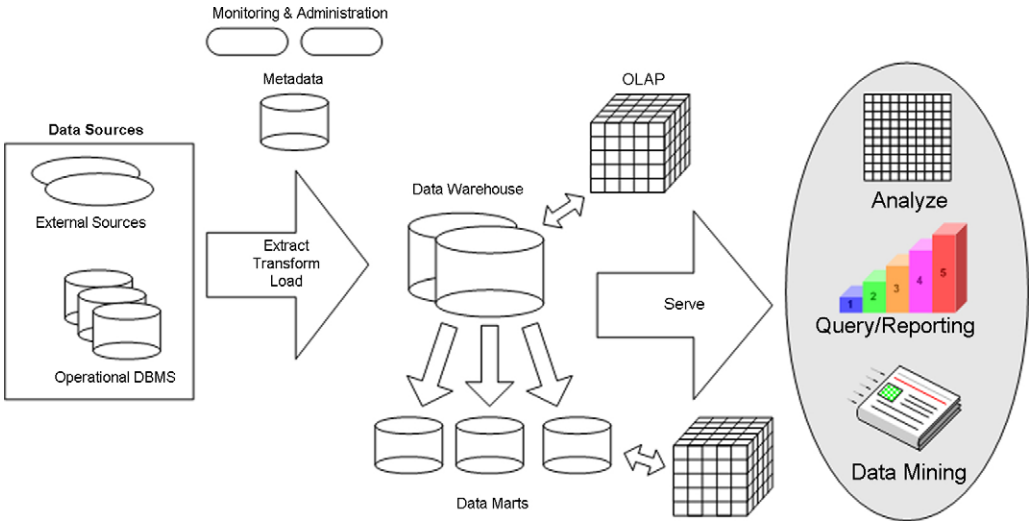


Fig. 1 – Data warehousing architecture.

systems from which data are extracted, transformed, and aggregated. Consequently, certain constraints such as data quality, amount of data, and data granularity are already present before the data warehouse design even begins. Of course, the first step in designing a data warehouse is requirements gathering, which begins with identifying the major business processes. Managing the processes requires knowledge of and access to appropriate performance metrics that translate to warehouse dimensions and facts. Granularity refers to the level of detail of the data in the data warehouse. Finer levels of granularity means more detailed data; coarser granularity mean less detail. For example, daily records of sales for a grocery store are at a finer level of granularity than monthly records of sales. Bill Inmon believes that the issue of data granularity is the single most important aspect of data warehouse design because it affects the physical amount of storage and performance requirements as well as versatility of analysis (Inmon, 2002, pp. 43–45).

Since data warehouses have existed over 20 years, there are many useful resources for their design and implementation. Almost all major database textbooks have at least one chapter devoted to the subject (e.g. Hoffer et al., 2005; Rob and Coronel, 2006). There have also been excellent research papers presenting overviews (Chaudhuri and Dayal, 1997), frameworks and current practices (Watson and Haley, 1997), failures (Watson et al., 1999), and entire books devoted to the subject of data warehouses (Inmon, 2002; Kimball, 2002; Marakas, 2003). A key component to designing a data warehouse is choosing the appropriate architecture.

2.1. *Data warehouse architecture*

Information technology architecture is a blueprint that illustrates the networking of the components of communication, planning, maintenance, learning, and reuse of an information system. IT architecture and specifically, the data warehouse architecture include different areas such as data design, technical design, and hardware and software infrastructure design. The design philosophies of data warehouse architecture are broadly classified into data mart design and enterprise-wide data warehouse design. A data mart is a smaller version of a data warehouse but is a smaller subset focused on selected subjects. The data mart follows a mixed (top-down as well as bottom-up) strategy of data design. The goal is to create individual data marts in a bottom-up fashion but in conformance with a skeleton architecture known as the "data warehouse bus." The enterprise-wide data warehouse is the union of those conformed data marts (Kimball, 2002). Common data warehouse architectures include the enterprise data warehouses, data marts, distributed warehouses, operational data stores with data marts, or any and all of these in combination. Hackney (2002) and Sen and Sinha (2005) provide descriptions of different architectures and design methodologies.

We now briefly describe a typical architecture of a data warehouse. We say typical because there are differing opinions on best practices of design and implementation. Most notably are the differences between Inmon and Kimball, but Sen and Sinha lists 15 different methodologies (Sen and Sinha, 2005). Shown in Fig. 1 is a graphical representation of a typical data warehousing architecture. This figure was adapted from Chaudhuri and Dayal's survey (Chaudhuri and Dayal, 1997). Data are identified from operational DBMS and other external sources, extracted, transformed and loaded (ETL) into the data warehouse or data marts. The ETL process provides a single authoritative data source to support decision-making. It is also the most challenging process of data warehouse development. Advanced tools are available to aid this process, but human monitoring and administration is required. Once the data exists in warehouse or data mart form, then online analytical processing (OLAP) tools provide graphical, multidimensional views for users to analyze, query, and mine the data.

A common design for data warehouses and data marts uses the star schema in which the central fact table connects to the dimensions in a star like fashion. This is the method we chose for the INARIS data warehouse. Fig. 2 shows an example star schema. The star schema consists of dimension tables and fact tables with the dimension tables containing descriptive

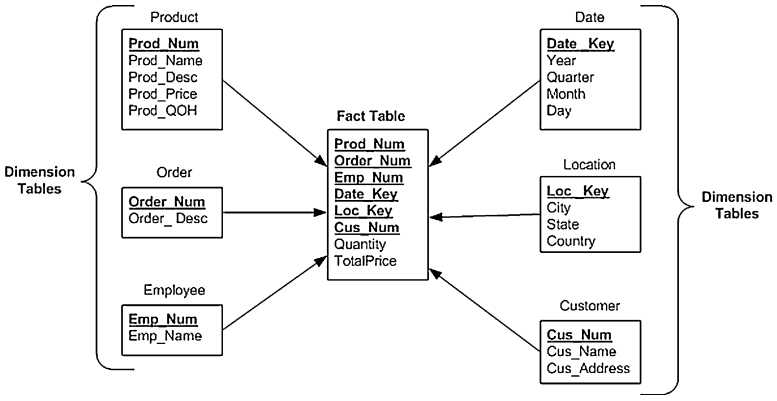


Fig. 2 – Example star schema.

data about the business subjects and the fact table containing factual or quantitative data such as number of units sold and price. In data warehouse design, fact table granularity is decided first (Kimball, 2002). In most warehouse designs, the decision is dependent on the level of detail the fact should address, namely, the business-process performance measure. The foreign key links from the fact table to the primary keys of the dimension tables yield the star configuration yielding a denormalized version of a relational data model.

Another important component of data warehouse design is the identification of the data sources. Dimensions and facts are developed based on the key performance indicators for the organization. Mapping source data to dimensions and facts require several stages of data extraction and data transformation. Cleansed data are loaded into the warehouse tables. In order to use the data warehouse, additional transformations such as building data marts, aggregations, and selections and projections of data are required. Together with OLAP applications and querying capabilities, the front end or presentation layer of the system is presented to users. Note that while each phase of the process is important for the success of the system, developing the critical dimensions of the data warehouse is crucial because the dimensions constrain understanding process key measures.

Many factors affect the architectural design for data warehouses such as the size of the business problem (Agosta, 2003), the type of business problem (Hoffer et al., 2005; Inmon, 1995, 2003, 2002; Kelly, 1997; Whiting, 2003), or fundamental data warehouse design philosophies (Armstrong, 1997; Inmon, 2002; Kimball, 2002; Marco, 2003). Consequently, there is no "one size fits all" approach to data warehouse development. Furthermore, given the strength of opinions expressed in the literature, there may always be preferred alternative methods based upon who is observing the system. Therefore, it is worthwhile viewing successful and unsuccessful implementations of data warehouses to determine best practices and avoid pitfalls (Langseth, 2004; Watson et al., 1999; Watson and Haley, 1997).

2.2. *Data warehouse quality*

Also crucial to the data warehouse development and implementation is data quality (Lenzerini, 2002; Rahm and Do, 2000; Sheth and Kashyap, 1993; Widom, 1995). The difficulty of taking separate external data sources and integrating them is not trivial and requires extensive effort from the developer. The problem of data quality is not exclusive to data warehousing, but exists any time data are pulled together from separate sources (Sheth and Kashyap, 1993). Berenguer et al. (2005) propose that data quality assurance in data warehouses metrics must be defined and adopted at the conceptual modeling stage. They present a design metrics framework in which each metric is part of a measured quality indicator. Their method defines theoretically validated metrics to measure data-quality goals. Serrano et al. (2006) also use metrics adapted from relational database design to establish data warehouse quality measures. Oliveira et al. (2005) present a taxonomy of data quality problems, derived from real-world databases. This taxonomy organizes problems at different levels of abstraction. Methods to detect data quality problems are represented as binary trees for each abstraction level.

There are intrinsic and contextual qualities of data that affect decision-making. Some intrinsic qualities that have received significant attention among researchers are accuracy, currency, and completeness. Decision makers often consider other subjective contextual factors that affect data quality (Nelson et al., 2005). Shankaranarayanan et al. (2006) tested an empirical model and determined that both data quality perceptions and the associated process metadata, when mediated by decision-making process efficiency, beneficially affect outcomes. van Vlymen et al. (2005) argue that by making the process of aggregating, processing, and cleaning data transparent, researchers can compare methods, and users can better understand the data. They suggest an eight-step process comprising: (1) design, (2) data entry, (3) extraction, (4) migration, (5) integration, (6) cleaning, (7) processing, and (8) analysis. This eight-step method provides a taxonomy enabling researchers to compare their methods of data process and aggregation. The process suggested by van Vlymen et al. (2005) is consistent with that of Winkler (2004).

Extraction, transformation, and loading are key phases of a data warehouse implementation. Data quality is assessed at each of these stages using intrinsic and contextual measures. For example, when climate data is processed for extraction, the valid range of values are examined as an intrinsic measure. Knowledge of the season and geography adds the contextual

component. For example, the winter temperature in southern India can be significantly higher than of the temperature in northern India.

3. Agricultural sector in India

3.1. *Agriculture in India*

Indian agriculture is highly diversified in climate, soil, horticultural crops, plantation crops, livestock resources, fishery resources, water resources, etc. Agriculture is the mainstay of the Indian Economy, and agriculture and allied sectors contribute nearly 25% of the Gross Domestic Production (GDP), and approximately 70% of the population is dependent on agriculture for their livelihood (Government of India, 2007). The Indian agricultural sector has great diversity in macro and micro level issues of social, economic, and cultural bases of India's vast population. Moreover, the diversity among resources generates interactions among macro and micro factors and is further complicated with their interdependencies.

The Indian agriculture infrastructure is as follows: the states maintain primary responsibility for increasing agriculture production, enhancing productivity, and exploring the sector potential. The central government supports state efforts in a catalytic way so that agricultural developments yield efficient results to the benefit of individual farmers. The Macro Management of Agriculture Scheme is a synthesis of 27 identified schemes and is being implemented in all states/union territories since 2000–2001. Under this scheme, the states have the flexibility to develop and pursue activities based on their regional priorities, but these resources need to be evaluated, monitored, and optimally allocated for balanced and sustainable development of the country.

3.2. *INARIS general objectives*

The Indian Council of Agricultural Research, under World Bank funded National Agricultural Technology Project has developed a data warehouse 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. The Indian Council of Agricultural Research is currently implementing the first two objectives, and the Ministry of Agriculture in 28 districts of 7 states is implementing objective three. It is this third objective, innovation of information technology, that is the impetus for this research.

The data warehouse is to provide systematic and periodic information to research scientists, planners, decision makers and developmental agencies via OLAP and decision support systems. Specifically, the warehouse is expected to satisfy the following goals:

- support agricultural research, management and education,
- improve the quality of research and planning,
- reduce duplication of research efforts,
- encourage dissemination of research findings,
- facilitate qualitative research supported by agricultural databases,
- help in the development of Decision Support Systems (DSS),
- use as effective tool for agricultural research and education planning,
- develop effective linkages with other national and international organizations.

To understand some of the specific data warehouse design issues it is first necessary to present an overview of the administration supporting the agricultural production and collection of information. India is divided into 28 states and 6 union territories (UT). Each state/UT is further divided into districts (elementary administrative unit) (India, 2004). Although there are further divisions of these districts into tehshils or taluks, blocks, and finally villages, a district is the basic unit of administration for all purposes.

3.3. *Sources of information*

In India, agricultural information is collected through several organizations throughout the country. For example, the National Sample Survey Organization conducts national level agricultural surveys; the National Horticulture Board and related state departments collect horticultural crops information. Similarly, 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 further compounded by a lack of common data collection standards. Consequently, the data warehouse architect has a formidable design challenge. To use the information at the macro planning and decision making level, data must be integrated and aggregated properly (Hollihan, 1982).

Thirteen organizations contributed data sources for the INARIS data warehouse. The data sources represent 59 databases and contain data collected from district level sources since 1990.

3.4. *Critical dimensions*

National level planning and decision support processes require access to data for many different resources, such as crops, livestock and fisheries, at varying levels of detail (Azad et al., 1998). Information on production (demand and supply), price levels, and population and migration statistics is also expected. These and other requirements must translate to the dimensions and fact tables. 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 three important dimensional issues, (1) granularities of location and time, (2) overlapping time domains and (3) aggregation and disaggregation of information at different dimensional hierarchies. These dimensional issues influence the fact table design and, therefore, the architecture of the data warehouse.

Table 1 – Location dimension hierarchy and type of data collected

Hierarchical level	Hierarchy name	Example
Level 1	National	Import and export statistics (monthly) International prices (daily/weekly) Production of minor crops etc. (annual)
Level 2	State	Production of major fruits and vegetables (annual) National account statistics (annual) Information about various agricultural development project at state level etc. (annual)
Level 3	District	Production and area of main crops (annual), land use statistics (annual) Production of livestock products such as milk, wool, egg, meat (annual) Information on fisheries etc. (annual)
Level 4	Village	The information on land use (annual) Information on different census such as human (decennial) Livestock census etc. (quinquennial)
Level 4	Agricultural markets	Prices of different agricultural commodities (daily/weekly depending on seasonality of the crop)
Level 4	Agro-metrological station	Information about various agro-climatic parameters of agricultural production (daily/weekly)

3.5. *Granularity of location*

Similar to industry 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 placed by the Indian government. Four levels of location hierarchies exist. Level 1 is national. Level 2 is state. India is divided into 28 states, often on a linguistic basis. Each state is further divided into districts (Level 3) which may be further divided into villages (Level 4). Although information may be collected at levels lower than villages, agricultural sector information is collected at Level 4. Agricultural surveys are the main vehicles of data collection.

Organizations may collect information at any or all levels of the location hierarchy. For example, quantity of exports and imports for different agricultural commodities such as fruits, vegetables, livestock products, tea, coffee, and fish products are collected at Level 1. Aggregate data are compiled through survey or census. International market price information for these commodities is also available at Level 1 for both location and time dimensions. For in-country use, commodity prices are available on a daily and weekly basis, while they are only available on a monthly basis for international import and export. In contrast to Level 1, Level 2 of the location hierarchy supports a richer domain of sources. Information on all commodities is available at this level. Different sample surveys acquire production figures of commodities such as fruit crops, plantation crops, etc. at Level 2. 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 2 is very important for state level planning and decision-making. Production information is available at Level 3 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. While there are several Level 4 attributes in the hierarchy, the most important is the village data. The village Level 4 has data such as land use, census data, livestock, and demographic and static parameters, e.g. land ownership and employment. Another Level 4 attribute is agricultural commodity trades available in agricultural markets or Mundi (trading place). Price data from many markets are collected daily or weekly 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 Level 4 hierarchy attribute. Table 1 summarizes the picture of different levels of the location dimensional hierarchy along with examples of information availability on the agricultural sector.

Another challenge presented with the design of the INARIS 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 level. These resources include human and financial capital and time. The following issues are associated with creation of dimensions in the development of the data warehouse:

- the number of levels required for any location,
- the integration of information from different sources (organizations) at different granular levels,
- the definition of fact tables for these dimensions.

The location hierarchy (Table 1) yields three candidates for Level 4. Note that data from these dimensional attributes cannot be merged. Additionally, it is not possible to represent them with a common name, as other attribute information associated at this level is different. For example, with village, the names of the village, block/tehshil/district and state are associated with the dimension. In agricultural markets, the name of the crop, name of the place, which may or may not be a district name, and type of market such as retail or wholesale, are associated with the dimension. Finally, in case of agro-

Table 2 – 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

meteorological stations the longitude, latitude, altitude, place name and other agricultural parameters such as soil type are associated with the dimension.

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 markets where the condition is price, a simple aggregation will not work. A weighted average or other suitable method is more appropriate, and with meteorological parameters, spatial models interpolating or extrapolating data at the District level are necessary. Furthermore, agricultural markets or weather stations are not available in every district. Availability of agricultural markets depends on its area, production, and consumption. Not every state or district within a state produces all commodities; accordingly, availability of agricultural markets and the commodities traded are local. Therefore, it may not be possible to aggregate the lower level data for each district or state to a higher level of the hierarchy. Thus, in this case, our hierarchical structure of the dimension will either collapse or provide misleading information to the user.

3.6. *Granularity of time*

Generally, the lowest granular level for the time dimension in the agricultural sector is day, but many measures are available only at a weekly, monthly, quarterly, half-yearly or yearly level. Climatic data such as rainfall, humidity, and temperature are available daily. Prices for commodities and products from different agricultural markets may be at daily, weekly or monthly periods. 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 upon the climatic, soil and water conditions. Information from human census is available every 10 years while livestock data are available after every 5 years. All other socio-economic data are available annually based on the financial year. Some of the information is collected on an *ad hoc* basis. Consequently, it is a challenge developing a data warehouse in which all sectors of agriculture integrate on a common homogeneous platform. Complexity is exacerbated when information availability for time levels follows different definitions. In India, agricultural information is available following three definitions of a year: calendar, agricultural, and financial.

Calendar year: Year starts January 1 and ends December 31. The months are in accord with the Julian calendar. The first week starts January 1 irrespective of the day, and weeks count 7 days. The last week is the 52nd week of the year and consists of 8 days to make 365 days. For leap years, the last week of February consist of 8 days. Again, quarters and half-years combine the months of the year.

Agricultural year: Year starts July 1 and ends June 30. Months are similar to the calendar year. The first week of the year starts July 1 and follows the same procedure as the calendar year. Similarly, first quarter and half-year start in July and follow the calendar years rules.

Financial year: Year starts April 1 and ends March 31. Months are as per calendar year. The first week of the year starts April 1 and it follow the same rules as a calendar year. Similarly, the first quarter and half-year of the year starts in April and continue as the calendar year does.

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 periods pose significant difficulties in integrating data. Table 2 shows the month number of each year (i.e. calendar, agricultural and financial year with respect to the months of calendar year):

A lookup table can handle integration of the information available at the granular level of months for different year types.

Let us now consider the table for quarters for each type of year (Table 3).

Integration of the information from different sources at the granular level of quarters may not be difficult in India as the definitions of different year such as calendar, agricultural and financial are offset by multiples of 3 months. However, if the offset is different, as it may be in other countries, it may not be feasible to integrate the quarterly information available for the different year types.

In case the information is available at a granular 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 6 months. Therefore, the first half-year of the calendar year corresponds to the second half of the agricultural year. This is not

Table 3 – 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

the case with financial year with these year types. Therefore, any information available at the granular level of the financial half-year may not be integrated with the information of the half-year of other year types.

The granular level of weeks presents a greater difficulty. Information available at the weekly granular level of any year type may not be integrated with weekly information of other year types. For example, in the calendar year, the first week is January 1 to January 7, but in the agricultural year, it is July 1 to July 7 and in case of financial year, 1st week is from April 1 to April 7. It can be observed that the beginning or the ending of weeks of 1-year type does not correspond to the beginning or ending of the weeks of any other year type. The weeks are off by day or two depending on normal or leap year. Thus, it is not possible to integrate the information coming from weeks of different types. Furthermore, it may be noted that within the same year type, information cannot be aggregated from the granular level of week to month, quarter and half-year, but it is possible to aggregate at the year level.

Integrating and aggregating 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, and 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 (Feijoo et al., 2003; Waichler and Wigmosta, 2003). 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 cannot be obtained from these with reliable precision. Although sampling strategies are employed in collecting data at these higher levels, the assumptions of sampling distributions do not hold well at lower levels. To address this issue highly sophisticated statistical or mathematical modeling techniques are required. Estimates of errors and risk associated with the data are also necessary for acceptable levels of confidence with the analyses.

4. INARIS warehouse

4.1. *INARIS architectural design*

We present the livestock data mart to illustrate our solution. We designed the INARIS data warehouse using Oracle Data Warehouse Builder Version 9.0. We use different fact tables for each type of dimension associated with the location dimension and its hierarchy. This arrangement results in a federation of data marts, where each data mart supports a fact table and all data marts are connected via a common dimension. In this design, each star schema configuration yields a data warehouse cube. The data warehouse system supports the information needs of research managers and planners, research scientists, and general users. OLAP tools including geographic information systems are deployed via a web interface. The system is available for authorized users at www.inaris.gen.in.

DIM_LIVESTOCK_SPECIES contains information on the species and breeds within species for livestock. DIM_LIVESTOCK_LOCATION contains information about states and their districts. DIM_LIV_TIME contains information for year only, as the information is not available below year. This is attributable to almost all the information collected through the surveys. Integrated livestock surveys are conducted yearly to collect data about the production of livestock products. The census for livestock is conducted every 5 years. Some of the information in this data mart has been collected only once.

Because we do not have all available measures at the lowest level of detail, several fact tables are necessary. For example, if measures corresponding to the location dimension are available only at the state level, then the grain of the fact table is fixed at that level. Therefore, the granularity of the dimensions results in many fact tables. Although this approach seems excessive, the alternative is unacceptable because either a fact table becomes too sparse or the measures become non-additive and misleading. Our prototype shows five separate fact tables.

LIV_ANIMAL_POPULATION (Fig. 3): This fact table is about all the information about population of the animals. Three dimensions, DIM_LIVSTOCK_SPECIES, DIM_LIVESTOCK_LOCATION, and DIM_LIV_TIME, are linked to the fact table of LIV_ANIMAL_POPULATION. The population measures are available only at the district level for each agricultural year for each species. Thus, the granularity of the measures in the fact table is limited to yearly district values. For clarity, we have expanded the name of the dimension, LOCATION to DIM_LIVESTOCK_LOCATION.

LIV_CENSUS (Fig. 4): This fact table is for information collected through livestock census after every 5 years. It is available at the lowest level of each dimensional hierarchy. In case of DIM_LIVESTOCK_SPECIES measures are available for each breed of the species. Measure about location is available at the district level. Time dimension related measures are available at the agricultural yearly grain, although collected every 5 years!

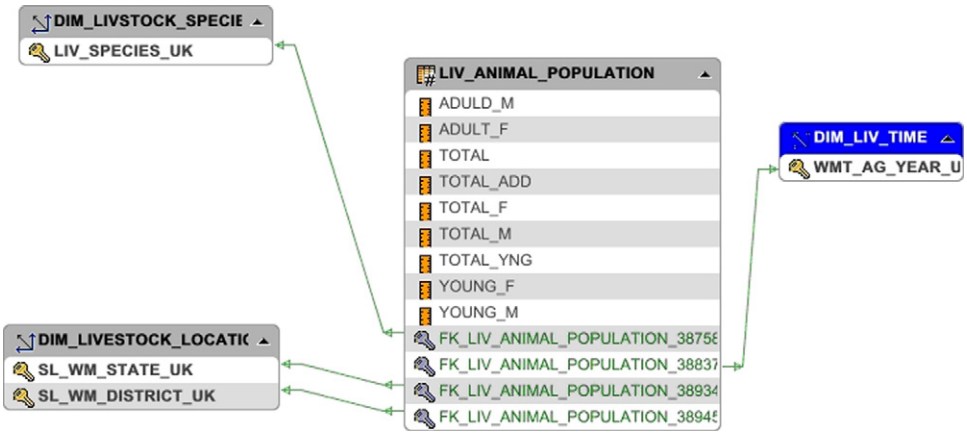


Fig. 3 – Yearly animal population data mart.

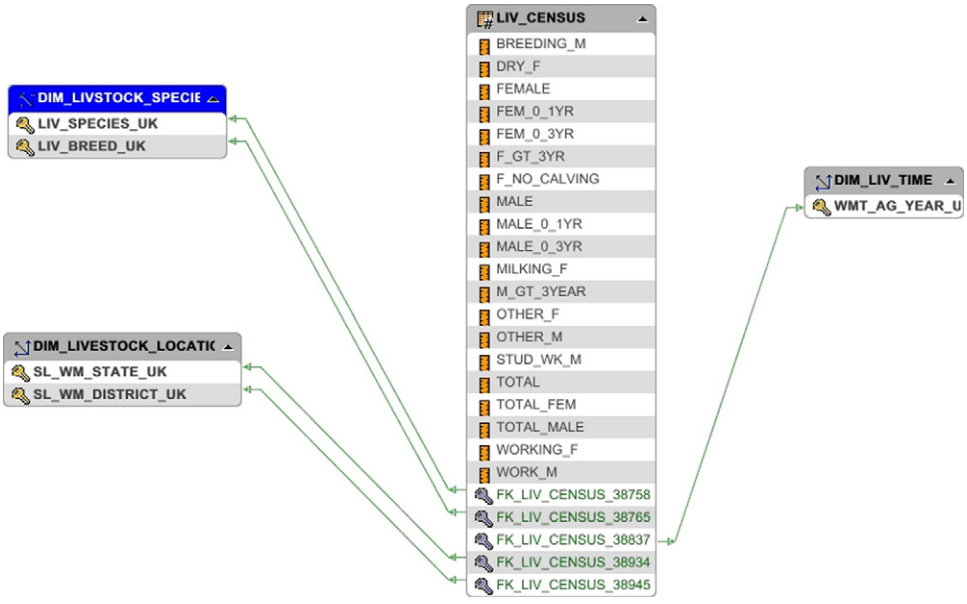


Fig. 4 – Five year animal census data mart.

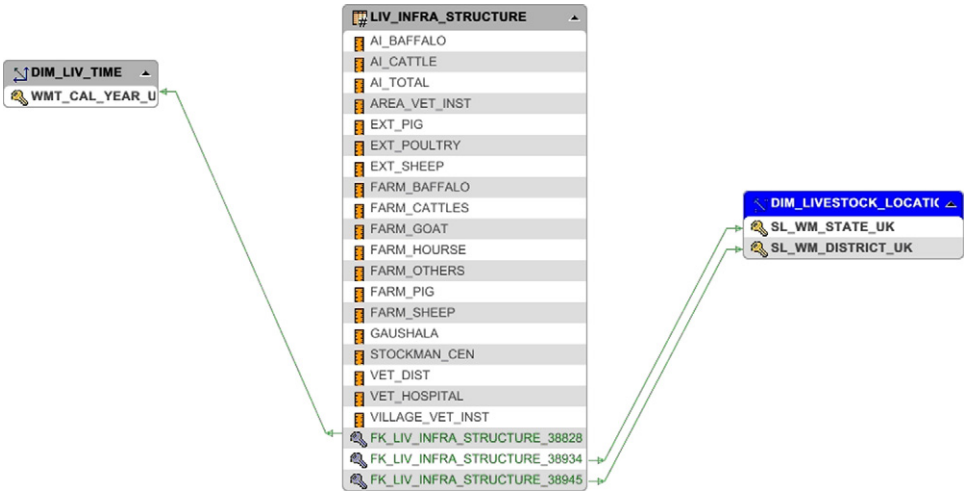


Fig. 5 – District level livestock infrastructure data mart.

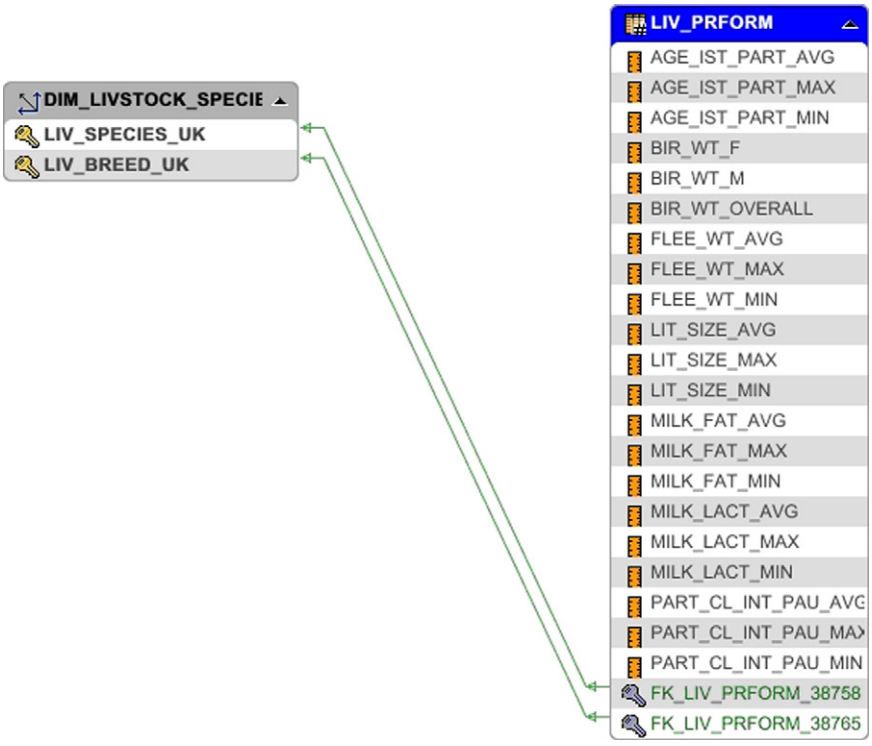


Fig. 6 – Livestock performance data mart.

LIV_INFRA_STRUCTURE (Fig. 5): This fact table provides information for infrastructure available at the district level for livestock production such as number of farms of different species, animal hospitals, artificial insemination centers, etc. This information is not collected on a regular basis. It is available at district level for some of the years. The location dimension, DIM_LIVESTOCK_LOCATION, yield measures at the district level, while the time dimension, DIM_LIVE_TIME, provides measures for calendar year for which they are available.

LIV_PERFORM (Fig. 6): In this fact table information on the production and anthropometric characteristics are stored for each breed belonging to different species. It is collected one time for the breed. The number of rows in the fact increases with increase in species and breed combination. Therefore, it is connected to surrogate keys of species and breed only in DIM_LIVESTOCK_SPECIES.

LIV_PRODUCTION (Fig. 7): This fact table is associated with production of different livestock products and by-products and is available at the state level for each agricultural year. The dimensions, DIM_LIVESTOCK_LOCATION, yielding measures at the state level and DIM_LIV_TIME, for agricultural year are linked to the fact table.

Note that because these cubes strictly follow the star schema design, it is possible to connect these cubes (drill across) through common dimensions using coordinated or conformed dimensions. This type of design not only provides flexibility and simplicity to the architecture but also provides the simplest solution to a highly complex problem (Kimball, 2002).

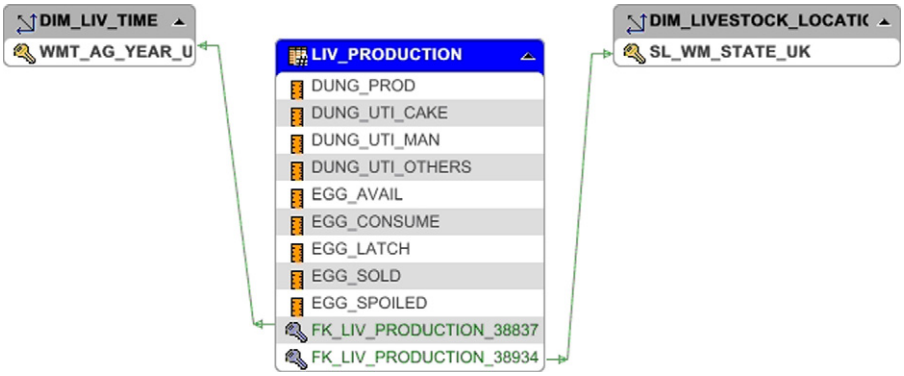


Fig. 7 – Yearly livestock production data mart.

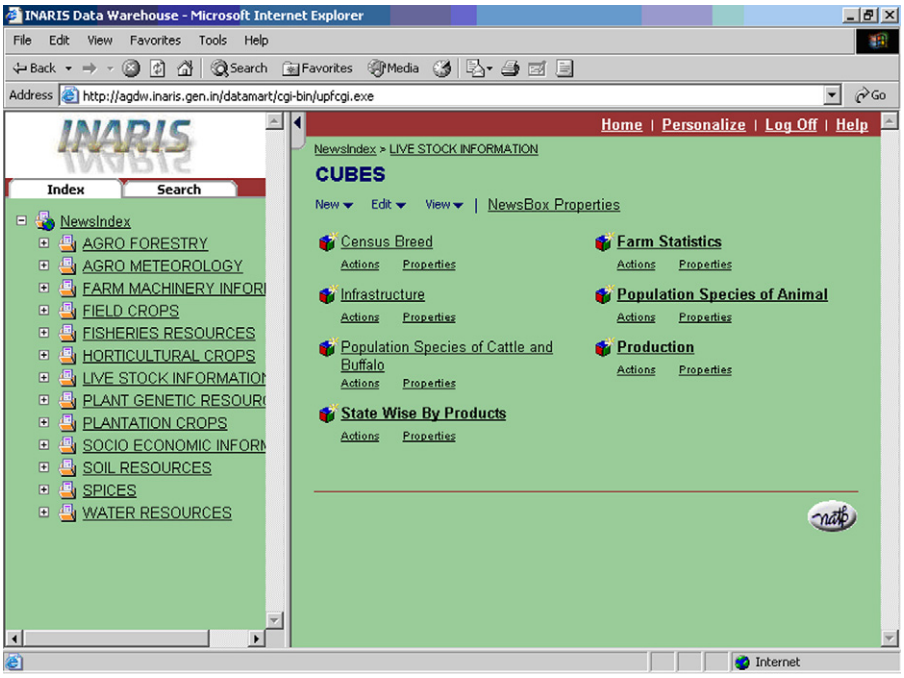


Fig. 8 – Initial screen after login.

4.2. *INARIS implementation*

The data warehouse project developed with funding from the World Bank and under the National Agricultural Technology Project (NATP) initiative has several goals and three user types. The users are (1) research managers, (2) research scientists, and (3) 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 another level, different users would have the capabilities to use various decision support capabilities through an OLAP application.

In the initial phase of the warehouse implementation, the development team and ICAR management identified 13 organizations that have been active in agricultural research and data collection. Moreover, most of these organizations have been implementing database solutions under different national initiatives. These 13 institutions have been collecting and organizing records and agreed to give data. INARIS created procedures for accessing/transmitting/sending data from these sources. Fifty-nine different databases were identified as source feeds for the data warehouse based upon participant willingness and availability. 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, district level data from 1990 onwards are integrated into this system. Many of these databases have statistical information dating to 1950. In building the central data warehouse, we started by creating subject-oriented data marts and multi-dimensional data cubes. The validation checks have been put into effect wherever possible.

The data warehouse system also provides spatial analysis through 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 www.inaris.gen.in, and the multidimensional cubes, dynamic reports, GIS maps and information systems are implemented. Fig. 8 shows several cubes relating to live stock information. Fig. 9 shows one of several graphical representations of buffalo population levels over time. Fig. 10 demonstrates the reporting capability of the data warehouse.

5. Evaluation of the INARIS data warehouse

To evaluate the satisfaction of the stakeholders of the INARIS data warehouse, we adopted the instrument proposed by Chen et al. (2000). In their research, Chen et al. developed a measure for data warehouse satisfaction in end-user support, information accuracy, format and preciseness, and overall fulfillment of end-user needs.

The INARIS data warehouse user satisfaction survey consisted of 16 items measured in a seven point Likert-type scale with 1 = strongly disagree, 4 = neutral, and 7 = strongly agree. The survey was conducted online; a copy is shown in Appendix A. Fifty-nine subjects were identified to participate in the survey, and 25 of those subjects responded yielding a 42.4% response rate. Of those who responded to the survey, 68% have used the INARIS data warehouse for more than 2 years. Twelve percent have used it from 1 to 2 years, and 20% have used it less than 1 year. When asked how often the user uses the system, 20.83% said daily, 29.17% said weekly, 4.17% said bi-weekly, 37.5% said monthly, and 8.34% said quarterly or longer. The majority of the respondents were either scientists

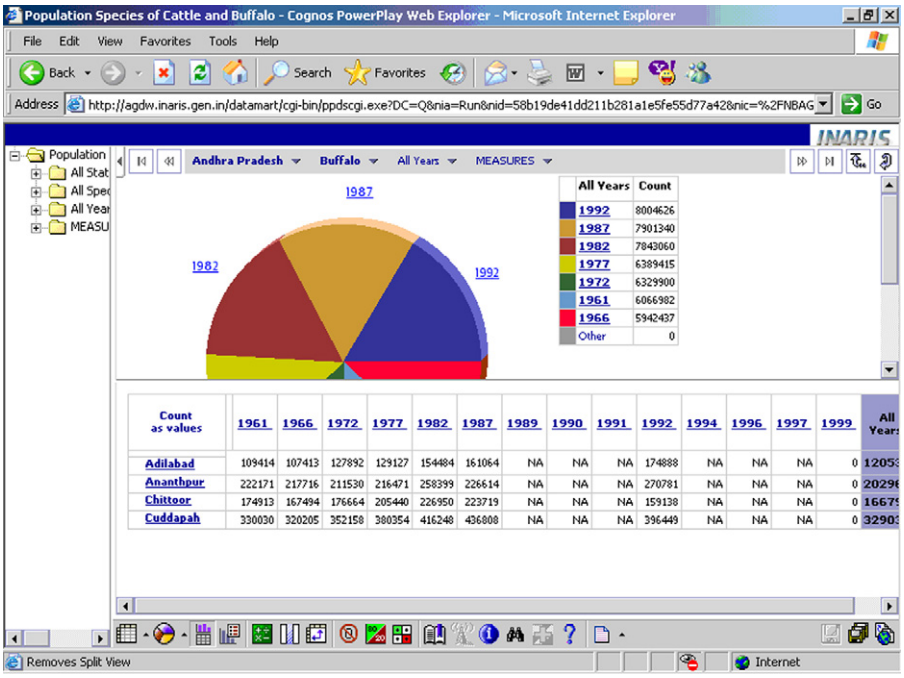


Fig. 9 – Graphical representation of buffalo over time.

or researchers. One respondent was an economist and another was a software engineer.

The information accuracy, format and preciseness factor consists of six items and has a high reliability (Cronbach's alpha (α) = 0.944), meaning that data warehouse users responded consistently to the items in the measure. The end-user satisfaction of support has six items and a reliability of α = 0.899. The overall fulfillment of end-user needs has three items and a reliability of α = 0.906. The mean response for user satisfaction of information accuracy, format and preciseness is 5.93. The mean response for user satisfaction of support in the INARIS data warehouse is 6.01. The mean response for user satisfaction for overall fulfillment of user needs is 5.58. These mean responses indicate that the users of the INARIS data warehouse report positively in their satisfaction of the system.

Since its deployment, the INARIS data warehouse has received over 500 queries per month. The major users are

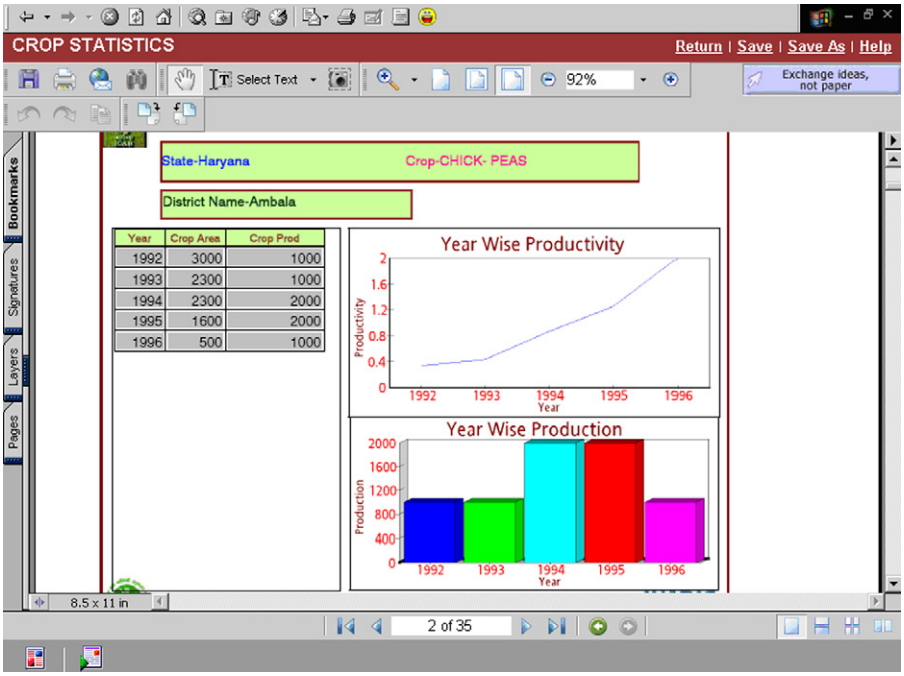


Fig. 10 – Example report of chick-pea crop statistics.

still limited to senior management personnel of the Indian Council of Agricultural Research, senior staff and Head of Divisions of Indian Agricultural Statistics Research Institute, and research scientists at the 13 participating institutes. Users are limited to these few because the project, though deployed, is still a research project and is in the first stage of deployment. Currently the data warehouse has 40 gigabytes of data and comprises approximately 115 aggregate tables. Information from the warehouse has been used to answer questions from many of the executive and legislative branches of the Indian government. For example, analysis of agricultural data including livestock, fisheries for estimation of contribution of research and development in agriculture through total factor productivity (TFP) was presented to the Honorable Prime Minister (Government of India, 2003).

In another use of the data warehouse, queries and analysis of the retrieved data is used in a district level research report on productivity and yield gap analysis of rice crop in the country. Other applications of the warehouse include the presentation of a technical report for identifying economically backward districts within different states. The report helps various developmental projects implementations through development of livelihood security index comprised of 57 district level parameters provided to the ICAR. These applications show the extent of use of the warehouse and its importance to long-range planning and macro-level decision-making.

6. Implications and conclusion

Implementing a centralized, national, agricultural data warehouse is a non-trivial task. For developers and stakeholders, the development of the INARIS data warehouse has brought several important issues to light. Good planning is essential. Because of the multi-institutional nature of the project, coordination and communication is a key planning success factor. The feasibility and utility of distributed and/or centralized data requires great care in its collection, manipulation and distribution. Issues of information ownership in databases and data warehouses must be resolved among different institutions. All the participating institutes must be convinced of the added benefits from the warehouse system to ensure participation and success. This issue harkens back to the earlier discussion of the problem of government data warehouses (Bieber, 1998; Inmon, 2003, 2005).

The data warehouse architecture choice is a crucial factor in making the INARIS data warehouse and similar projects successful. The project size and nature warrants subject-oriented data marts and data cubes using a bus architecture. Selection of hardware and software deserves full consideration and one should not compromise the system on the basis of lower expenditures, or as an old adage says, "you get what you pay for."

The quality control of the information storage is top priority as this makes the data relevant and useful to the all users. The users' confidence is dependent upon information quality. Therefore, quality is more important then quantity, or, in other words, garbage in yields garbage out. Lastly, a critical component for this project's success is management support. This last component is the most difficult to overcome in a government venture. Bureaucracy exists in every government and often kills even essential projects. Therefore, key people must be in place to champion the project to provide the opportunity to succeed.

We demonstrate that while challenges persist, it is possible to implement a successful national level data warehouse serving multiple disparate stakeholders. A major obstacle for this project is the network infrastructure. Internet bandwidth is not sufficient for many participating organizations. Therefore, it is not yet feasible to automate the collection and updates of data from the centers providing the source feeds. Currently, the centers mail the data on CD-ROMs and thus, delay information assimilation. Another challenge is that the data collection organizations are different for different sectors of agriculture and participating centers have no administrative control over them. Thus, getting current information on time is difficult. Lastly, most of the historical information is available in booklets or official records, and data quality is often poor or even inaccurate. Primary source for this information is not available.

Sector level data warehouses are rare, but increasing. Governments at national and state levels, industry groups, and regulatory organizations have begun to realize the potential of integrating data from many sources and using data warehouses to implement solutions like the INARIS data warehouse. We present some of the problems arising in integrating data collected in the Indian agricultural sector, and we discuss specific problems associated with granularity of location and time, the two key dimensions for an agricultural warehouse. While these problems exist, it is possible to overcome them, as we have demonstrated through the success of the INARIS data warehouse. This data warehouse is being used to provide responses to questions raised by legislators in the Indian Parliament about agriculture issues (Government of India, 2003).

Some of the objectives of governments should be to increase profit or market share, and reduce cost to provide national or regional benefit, and this is true of the Indian government. The national benefit of improved agricultural production elevates this type of project from simply an academic exercise to essential for the promotion of well-being of India (Maheshwar and Chanakwa, 2006; Modi, 2007; Tribune News Service, 2001).

Appendix A. Appendix A

INARIS data warehouse user satisfaction survey

Thank you for taking the time to respond to this short survey. We have 16 questions for which we would like your feedback. Your responses can help to improve the INARIS data warehouse.

All responses are completely anonymous.

1. Please rate your satisfaction with each of the following dimensions of the INARIS data warehouse.

	Strongly Disagree			Neutral			Strongly Agree
1. The data warehouse provides the precise information you need	+	+	+	+	+	+	+
2. The information content of the data warehouse meets your needs.	+	+	+	+	+	+	+
3. The data warehouse provides reports that are exactly what you need.	+	+	+	+	+	+	+
4. The data warehouse provides sufficient information for your decision-making.	+	+	+	+	+	+	+
5. The data in the data warehouse are accurate.	+	+	+	+	+	+	+
6. The output of the data warehouse is presented in a useful format.	+	+	+	+	+	+	+
7. The information extracted from the data warehouse is clear.	+	+	+	+	+	+	+
8. The data warehouse is user friendly.	+	+	+	+	+	+	+
9. You get the information you need in time from data warehouse.	+	+	+	+	+	+	+
10. The Information Systems department provides satisfactory support to users using data warehouse.	+	+	+	+	+	+	+
11. Your suggestions for future enhancement of the data warehouses will be responded by Information Systems department cooperatively.	+	+	+	+	+	+	+
12. The data warehouse applications provide proper level of on-line assistance and explanation.	+	+	+	+	+	+	+
13. The data warehouse applications provide users with adequate error-control facilities, including error prevention, error detection, error correction and error recovery.	+	+	+	+	+	+	+
14. The Information Systems department provides users with adequate level of training on using the data warehouse.	+	+	+	+	+	+	+
15. The data warehouse developers interact with users extensively during the development of data warehouse.	+	+	+	+	+	+	+
16. Overall, the INARIS data warehouse fulfills my informational needs.	+	+	+	+	+	+	+

2. How long have you used the system?

   + More than 2 years
   + 1 to 2 years
   + 6 months to 1 year
   + Less than 6 months
   + Other (please specify)
3. How often do you use the system?
   + Daily
   + Weekly
   + Bi-weekly
   + Monthly
   + Quarterly
   + Semi-annually
   + Annually
   + Other (please specify)
4. What is your primary department?
5. What is your primary job title?

REFERENCES

Abdullah, A., Brobst, S., Umer, M., Khan, M.F., 2004. The Case for an agri data warehouse: enabling analytical exploration of integrated agricultural data. In: Proceedings of the IASTED International Conference on Databases and Applications (DBA 2004), Innsbruck, Austria.

Agosta, L., December 13 2001. "Top 10 Data Warehousing Challenges," Forrester Research.

Agosta, L., 2003. Data warehouse size depends on the size of the business problem. DM Rev. 13 (16), 16–17.

Armstrong, R., 1997. A Rebuttal to the Dimensional Modeling Manifesto.

Azad, A.N., Erdem, A.S., Saleem, N., 1998. A framework for realizing the potential of information technology in developing countries. Int. J. Com. Manage. 8 (2), 121–133.

Berenguer, G., Romero, R., Trujillo, J., Serrano, M., Piattini, M., 2005. A set of quality indicators and their corresponding metrics for conceptual models of data warehouses. In: Data Warehousing and Knowledge Discovery, pp. 95–104.

Bieber, M., 1998. Data Warehousing in Government, DM Rev.

Bouguettaya, A., Benatallah, B., Elmagarmid, A.K., 1998. Interconnecting Heterogeneous Information Systems. Kluwer, Boston, MA.

Chaudhuri, S., Dayal, U., 1997. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26 (1), 64–74.

Chen, L.-d., Soliman, K.S., Mao, E., Frolick, M.N., 2000. Measuring user satisfaction with data warehouses: an exploratory study. Info. Manage. 37, 103–110.

Feijoo, S.R., Caro, A.R., Quintana, D.D., 2003. Methods for quarterly disaggregation without indicators; a comparative study using simulation. Comput. Statist. Data Anal. 43 (1), 63–78.

Ghosh, M., Rao, J.N.K., 1994. Small area estimation: an appraisal. Statist. Sci. 9 (1), 55–76.

Government of India, 2003. Lok Sabha Unstarred Question No. 3669, http://164.100.24.208/lsq/quest.asp?qref=57381.

Government of India, 2007. India Image, http://indiaimage.nic.in/.

Hackney, D., 2002. Architectures and Approaches for Successful Data Warehouses, Oracle White Paper.

Harper, F.M., 2004. Data warehousing and the organization of governmental databases. In: Digital government: Principles and Best Practices. IGI Publishing, pp. 236–247.

Hoffer, J.A., Prescott, M.B., McFadden, F.R., 2005. Modern Database Management. Pearson Education, Inc., Upper Saddle River, New Jersey.

Hollihan, M., 1982. The Relationship between Inflation and Relative Prices. Stud. Econ. Finance 6 (1), 29.

India, N.I.C., 2004. Districts of India: A Gateway to Districts of India on the web, http://www.districts.nic.in.

Inmon, W., 1995. What is a Data Warehouse? PRISM Tech Topic 1 (1).

Inmon, W., 2003. BI in the Government. DM Rev. 13 (8), 26–27.

Inmon, W., 2005. Data Warehousing for Government.

Inmon, W.H., 2002. Building the Data Warehouse. John Wiley & Sons, Inc.

Iowa Division of Criminal & Juvenile Justice Planning, 2007. Justice Data Warehouse (JDW), http://www.state.ia.us/government/dhr/cjjp/jdw/index.html.

Kelly, S., 1997. Data Warehousing in Action. John Wiley & Sons, Chichester.

Kimball, R., 2002. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Inc.

Langseth, J., 2004. Real-Time Data Warehousing: Challenges and Solutions, DSSResources.COM, http://dssresources.com/papers/features/langseth/langseth02082004.html.

Lenzerini, M., 2002. Data integration: a theoretical perspective. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Madison, Wisconsin, pp. 233–246.

Maheshwar, C., Chanakwa, T.S., 2006. Postharvest losses due to gaps in cold chain in India—a solution. In: Proceedings of the IV International Conference on Managing Quality in Chains – The Integrated View on Fruits and Vegetables Quality, Bangkok, Thailand, pp. 777–784.

Marakas, G.M., 2003. Modern Data Warehousing, Mining, and Visualization. Prentice Hall, Upper Saddle River, NJ.

Marco, D., 2003. Independent Data Marts: Stranded on Islands of Data, Part 1. DM Rev. 13 (4), 30–33.

Modi, A., 2007. Export of agri goods up 9%, Business Standard, http://www.business-standard.com/common/storypage.php?autono=288679&leftnm=3&subLeft=0&chkFlg=.

Nelson, R.R., Todd, P.A., Wixom, B.H., 2005. Antecedents of information and system quality: an empirical examination within the context of data warehousing. J. Manage. Info. Systems 21 (4), 199–235.

Oliveira, P., Rodrigues, F., Henriques, P., Galhardas, H., 2005. A taxonomy of data quality problems. In: Proceedings of the International Workshop on Data and Information Quality.

Rahm, E., Do, H., 2000. Data cleaning: problems and current approaches. IEEE Data Eng. Bull. 23 (4), 1–11.

Rob, P., Coronel, C., 2006. Database Systems: Design, Implementation, and Management. Course Technology.

Sen, A., Sinha, A.P., 2005. A comparison of data warehousing methodologies. Commun. ACM 48 (3), 79–84.

Serrano, M., Calero, C., Piattini, M., 2006. An experimental replication with data warehouse metrics. Int. J. Data Warehousing Mining 1 (4), 1–21.

Shankaranarayanan, G., Watts, S., Even, A., 2006. The role of process metadata and data quality perceptions in decision making: an empirical framework and investigation. J. Info. Technol. Manage. 17 (1), 50–67.

Sheth, A., Kashyap, V., 1993. So far (schematically) yet so near (semantically). In: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), North-Holland, pp. 283–312.

Tribune News Service, 2001.Rs 70,000 cr food 'wastage' a year, The Tribune.

U.S. Department of Health and Human Services, 2007. Geospatial Data Warehouse, U.S. Department of Health and Human Services, http://datawarehouse.hrsa.gov/default.htm.

van Vlymen, J., de Lusignan, S., Hague, N., Chan, T., Dzregah, B., 2005. Connecting medical informatics and bio-informatics. In: Proceedings of the MIE2005 – The XIXth International Congress of the European Federation for Medical Informatics, pp. 1010–1015.

Waichler, S.R., Wigmosta, 2003. Development of hourly meterological values from daily data and significance to hydrological modeling at H.J. Andrews Experimental Forest. J. Hydrometerol. 4 (2), 13.

Watson, H.J., Gerard, J.G., Gonzalez, L.E., Haywood, M.E., Fenton, D., 1999. Data warehousing failures: case studies and findings. J. Data Warehousing 4 (1), 44–55.

Watson, H.J., Haley, B.J., 1997. Data warehousing: a framework and survey of current practices. J. Data Warehousing 2 (1), 10–17.

Whiting, R., 2003. The Data-Warehouse Advantage, pp. 63–66.

Widom, J., 1995. Research problems in data warehousing. In: Proceedings of the Fourth International Conference on Information and Knowledge Management, Baltimore, MD, pp. 25–30.

Winkler, W.E., 2004. Methods for evaluating and creating data quality. Info. Systems 29 (7), 531–550.

Wixom, B.H., Watson, H.J., 2001. An emperical investigation of the factors affecting data warehousing success. MIS Quarterly 25 (1), 17–41.

Yost, M., 2000. Data warehousing and decision support at the National Agricultural Statistics Service. Soc. Sci. Computer Rev. 18 (4), 434–441.