J. Ind. Soc. Agril. Statist.
61(2), 2007 : 264-273

Data Warehousing for Agricultural Research - An Integrated Approach for Decision Making

Anil Rai, P.K. Malhotra, S.D. Sharma and K.K. Chaturvedi
Indian Agricultural Statistics Research Institute, New Delhi

SUMMARY

The growing need for the data warehousing technology in recent years has stemmed because of the technology's importance in supporting decision support processing and analysis. A specific property of data warehousing is to make efficient application processing for the development of decision support system, which need to summarize huge amounts of data. In this article an attempt is being made to share process and experience faced during development of agricultural data warehouse at Indian Agricultural Statistics Research Institute (IASRI), New Delhi (http://www.iasri.res.in) under a National Agricultural Technology Project (NATP), Mission Mode sub-project entitled "Integrated National Agricultural Resources Information System (INARIS) (http://www.inaris.gen.in)". This article also provides description about broad architecture and process of agricultural data warehouse. The problems of implementation of agricultural data warehouse including dimensional modeling are different from the data warehouse developed for business process. Further, query and reporting process of this data warehouse has been briefly outlined, which may provide readers, especially research managers, the idea of flexibility and ease of operations of this on-line decision support system. An introduction about INARIS project is also given for readers.

Key words: Data warehouse, Dimensional modeling, OLAP tools, Agricultural resources, Decision support system.

1. INTRODUCTION

A data warehouse is a "subject-oriented, integrated, time variant, non-volatile collection of data that is primarily used in organizational decision making" (Inmon 2005, Kimball 1998). The data warehouse is maintained separately from operational databases. Since, data warehouse contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be larger than operational databases. The organizational data warehouses are projected to hundreds of gigabytes or terabytes in size. A query, mostly adhoc in nature, can access millions of records and perform scans, joins and aggregates (Gupta and Mumick 1995; O'Niel and Graefe 1995; O'Neil and Quass 1997). In this system query throughput and response time are more important than transaction throughput. Kimball (1998) defines data warehouse as "a copy of information data specifically structured for query and analysis".

The growing need for the data warehousing technology in recent years has stemmed because of the technology's importance in supporting decision support processing and analysis. A specific property of data warehouse, which makes efficient application processing, is that most of the applications are decision support oriented applications, which need to summarize huge amount of data. The management of this huge amount of data and its complex analysis during queries are most important in development of a data warehouse (Neil and Graefe 1995, Zhuge *et al.* 1995, Choudhari and Shim 1995, Gupta and Mumick 1995, Neil and Quass 1997, Bonifati *et al.* 2001, Chen *et al.* 2003, Kambayashi *et al.* 2004). The growing trend in data warehouse architecture is to store the data both in the data warehouse and in several data marts, where each data mart contains the data pertaining to a particular domain of the organization's operations. An important issue that needs to be addressed is how to maintain data in both the

warehouse and the data marts in response to update the source data. Kimball and Ross (2002) and Inmon (2005) discussed data warehouse technology in detail.

In case of complex analysis and visualization, the data in a data warehouse is typically modeled in multi-dimensional environment. It requires some additional steps apart from On-Line Transaction Processing (OLTP) systems. Storage system of data warehouse database is generally used for read only purpose. The updation will be done on periodic basis. This data storage is again converted into a form of multidimensional model, known as a cube. These cubes can be designed by using fact and dimension tables from a database. Since, these cubes are deployed on Internet for on-line analysis, these are also known as On-Line Analytical Processing (OLAP) cubes. In these cubes aggregations are pre-calculated and stored in multidimensional form. After deployment of these multidimensional cubes on the server, end user can perform his analysis and export the desired result to his desktop or computer in any form such as, MS-Word, MS-Excel, ASCII text file or well-known Acrobat Reader (PDF) format etc.

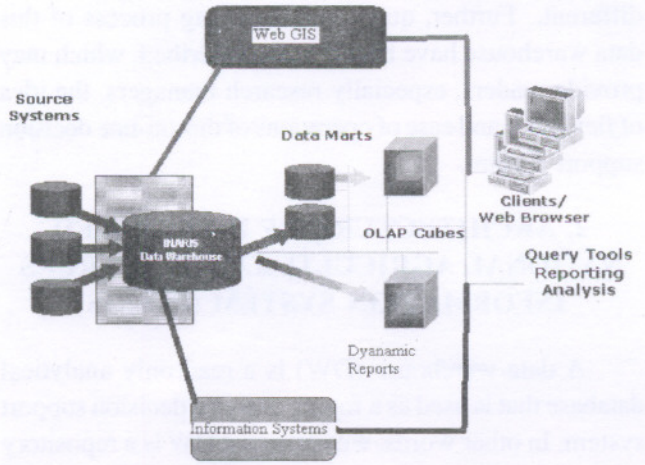
In fact, the development of an integrated information system and databases (non-spatial, spatial and bibliographic) in the field of agriculture research and education was recognized by the Indian Council of Agricultural Research (ICAR) Review Committee in 1988. The development of a computerized satellite based information network (ICARNET) was recommended by the committee with on-line terminal connectivity to ICAR scientists and State Agriculture Universities (SAUs). In 1991, the development of a computer-based Agricultural Research Information System (ARIS) network was initiated by ICAR. IASRI took up a National Agricultural Technology Project (NATP) Mission Mode project entitled "Integrated National Agricultural Resources Information System (INARIS)" in 2001. In this a State-of-Art Central Data Warehouse has been developed at IASRI, New Delhi. This provides description about agricultural data warehouse, broad architecture and developmental strategies adopted during dimensional modeling of agricultural data warehouse which is different from the data warehouse developed for business process. The main aim of development of a data warehouse in business enterprise is to maximize the profit. The profit making activities of an enterprise are well defined and structured. Moreover, the data generation and its utilization is the responsibility of the same organization. However, in case of agricultural sector in India, numbers of organizations are involved in the data collection process. Moreover, the concepts, definitions, format and purpose of this data collection mechanism followed by these organizations are entirely different from each other. Further, most of the time information user organizations are different from these information generation organizations. Therefore, data extraction, cleaning and integration mechanism of these information are entirely different. Further, query and reporting process of this data warehouse have been briefly described, which may provide readers, especially research managers, the idea of flexibility and ease of operations of this on-line decision support system.

2. ARCHITECTURE OF INTEGRATED NATIONAL AGRICULTURAL RESOURCES INFORMATION SYSTEM (INARIS)

A data warehouse (DW) is a read only analytical database that is used as a foundation of a decision support system. In other words, we can say, a DW is a repository of integrated information, available for queries and analysis. Data and information are extracted from heterogeneous sources as they are generated. INARIS project was taken up to design and develop a State-of-Art flexible Central Data Warehouse (CDW) of agricultural resources of the country at IASRI, New Delhi (lead center) and databases on different subjects at respective co-operating centres. The above project was implemented with active collaboration and support from 13 other ICAR institutions, namely National Bureau of Soil Survey and Land Use Planning (NBSSLUP), Nagpur (for soil resources); Central Research Institute on Dryland Agriculture (CRIDA), Hyderabad (for agro-meteorology); Project Directorate of Cropping Systems Research (PDCSR), Modipuram (for crops and cropping systems); National Bureau of Animal Genetic Resources (NBAGR), Karnal (for livestock resources); National Bureau of Fish Genetic Resources (NBFGR), Lucknow (for fish resources); National Bureau of Plant Genetic Resources (NBPGR); New Delhi (for plant genetic resources); National Centre for Agricultural Economics and Policy Research (NCAP), New Delhi (for socio-economic resources), Central Institute of Agricultural

Engineering (CIAE), Bhopal (for agricultural implements and machinery); Central Plantation Crops Research Institute (CPCRI), Kasargod (for plantation crops), Indian Institute of Spice 'S' Research (IISR), Calicut (for spices crops); ICAR Research Complex for Eastern Region, Patna (for water resources); National Research Centre for Agro-Forestry (NRC-AF), Jhansi (for agro forestry) and Indian Institute of Horticultural Research (IIHR), Bangalore (for horticultural crops).

Fig. 1. Architecture of CDW of INARIS



In all 59 databases on agricultural technologies generated by ICAR, research projects in operation and related agricultural statistics from published official sources at least from the year 1990 onwards at the district level were integrated into this information system. The INARIS data warehouse architecture comprises of operational source systems i.e. 59 source databases, a data staging area for structuring and formatting the information extracted from the source system, one or more conformed data marts based on different themes or subjects and a data warehouse database. The important concepts of this technology have been described in Rai *et al.* (2007). Subject-wise data marts were created, multi-dimensional data cubes developed and published on Internet/Intranet. The validation checks were implemented wherever possible. The broad architecture of CDW of INARIS is presented in Fig. 1. The information systems developed under this project are directly accessible to any general users. The Web GIS (Geographical Information System) has been independently developed from the data warehouse database under ARC-GIS environment and integrated into this CDW through ARC-IMS. Though users will have all basic GIS functionalities through normal Web browser. This CDW is being used for decision making and policy planning for sustainable growth and development of agricultural research.

This project strengthens the information system conceptualized by ICAR. An on-line decision support system of agricultural resources has been developed. This system is capable of providing information based on the interaction among the basic resources like soil, water, climate, animal and vegetation that form the prime components of the production system in the country. This data warehouse is also useful in determining the carrying capacity of the region. The project provided suitable opportunity on multi-disciplinary mode through enhanced linkages among research institutes and other development agencies by providing first hand information on problems and potential in production systems. This data warehouse is being intensively used with an ultimate aim of enhancing better quality of life of the farming community and society at large.

3. DIMENSIONAL MODELING

A dimension model constitutes logical design of a data warehouse. In order to build a data warehouse for agricultural research top down approach was followed. Initially, Bus Architecture Matrix (BAM) was build in which names of all possible data marts and all possible dimensions were identified and possible linkages/ association were established through BAM. A data mart can be thought as collection of different fact tables with in a domain. To start with, single source data marts were identified followed by multiple source data marts and then possibility of combining these data marts was worked out keeping in view the broader perspective. The rows of BAM are data marts and columns are dimensions (Table 1). Important data marts and important dimensions of INARIS are presented in BAM as Table 1 and intersections of data marts and dimensions are marked.

Table 1. Bus Architecture Matrix (BAM) of INARIS CDW

Item	Time	Location	Commodities	Species	Add. Factors
Crop Statistics	✓	✓	✓		
Crop Inputs	✓	✓	✓		
Crop Variety		✓	✓		
Crop Weed		✓			
Crop Pest		✓			
Crop Disease		✓			
Cropping System	✓	✓			
Commodity Price	✓	✓	✓		✓
Commodity Trade	✓	✓	✓		✓
Crop Management		✓			
Farm Equipment		✓	✓		✓
Farm Mechanization	✓	✓			
Forest Cover	✓	✓			
Animal Breed		✓		✓	✓
Animal Census	✓	✓		✓	✓
Livestock Infrastructure	✓	✓	✓		✓
Livestock Production	✓	✓	✓		
Livestock By-products	✓	✓	✓		✓
Water Canal	✓	✓			✓
Water Rivers	✓	✓			✓
Water Quality	✓	✓			✓
Water Rates	✓	✓			✓
Ground Water Resource	✓	✓			✓
Fish Statistics	✓	✓	✓		✓
Fish Infrastructure	✓	✓			✓
Agricultural Wages	✓	✓			✓
Crop Arrivals	✓	✓	✓		
Demography	✓	✓			✓
Employment	✓	✓			✓
Land Use	✓	✓			✓
Expenditure	✓	✓			✓
Import-export	✓	✓	✓		
Household Amenities	✓	✓			✓
Agro-climatic	✓	✓			

Table 1 is a broader presentation of bus architecture of INARIS CDW and it can be seen from BAM that location dimension is common to all the data marts. Time dimension followed by commodity dimension is also required in number of data marts. This INARIS BAM mapped all the processes which need to be taken up to get all groups to be agreed on a common definition of dimension so that these data marts can be conformed to each others. Apart from these dimensions, there are number of other dimensions in each of these data marts but those are not common to any other data marts so those are not provided in Table 1. Selection of a data mart was followed by identification of fact tables of the data mart. Initially, fact tables from single source of data were identified. Then fact tables, which were generated from multiple data sources, were identified. Keeping in view the user requirements and data availability at the source data, grain level of each fact table was decided. Declaration of grain level also provided information about the individual record level of each fact table. A good and clear declaration of grain level of each fact table made it easy to choose appropriate dimensions which can be associated with a particular fact table. Logical fact diagram of each selected fact table has been prepared. The fact diagram shows not only the specifics of a given fact table but also shows the context of the fact table in overall data mart. The fact table diagram provides information about the name of the fact table, its grain level and dimensions connected to this fact. This serves as introduction to the overall model.

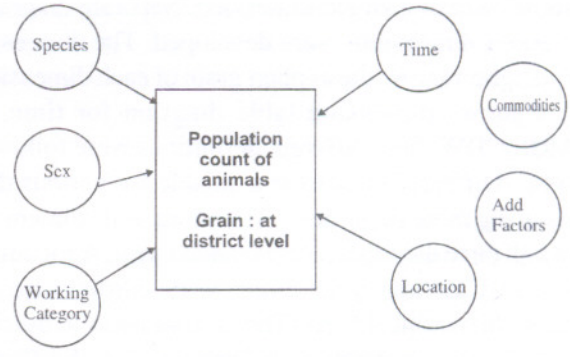


Fig. 2. Fact diagram of livestock census data mart in INARIS CDW

Fig. 2 shows the fact diagram of livestock census data mart in INARIS CDW. In this fact diagram dimensions associated with this fact are shown to be connected through arrow and important dimension of

BAM which are not relevant to this fact are written at the back and not connected to this table. The center fact table shows important description of the fact along with its grain level. The fact diagrams were made for all the fact tables of different data marts in the warehouse. The details of each fact were described in the fact table. The fact table provides a complete list of all the facts. This list includes actual facts in the physical table, derived facts and other facts that are possible to calculate from the first two. The fact table of the corresponding fact diagram (Fig. 2) is shown in Fig. 3.

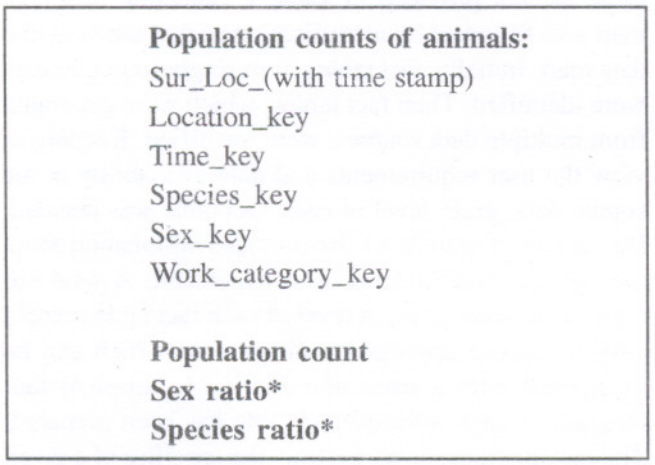


Fig. 3. Fact table diagram showing dimension keys, basic facts and derived facts (shown with asterisk)

The fact table diagrams were made for the entire fact tables in different data marts of the INARIS CDW. The second type of detail diagram is the dimension table detail diagram. This diagram shows the individual attributes within a single dimension. Separate diagrams for various dimensions were developed. The dimension table diagram shows the explicit grain of each dimension. Fig. 4 shows dimension table diagram for time. In INARIS CDW three different definitions were followed for year. Different facts were available for various data marts using these definitions. Therefore, time dimension shows all three definitions i.e. Financial year, Agricultural year and Calendar year along with cardinalities at different hierarchical level. This allows users to quickly see the multiple hierarchies and relationship of different attributes. The dimension table diagrams for all dimensions were developed for each dimensions of the data warehouse.

Some very specific situations such as many-to-many relations, slowly changing dimensions, artificial attributes etc. were identified and accordingly model has been rectified. Many-to-many relation problems were solved at logical design stage. These relations were represented as separate bridge tables. In case of slowly changing dimensions such as location, the concept of surrogate key was implemented through which historical information of the members at different levels in the dimension were gracefully preserved. In case of livestock census fact table (Fig. 3) Sur_Loc is a surrogate key for location with time stamp. In this case the name of the districts or states may change over time (years). These historical changes can be preserved along with recent name for future reference. In some situations some artificial attributes were created at different level of hierarchies which were not available at data source level to support the roll-up and drill down process in different dimensions. Two kinds of derived facts i.e. additive and non-additive were identified. Derived additive facts were calculated entirely from the other facts of the same fact table records. Some of non-additive facts such as ratios were expressed at different grain levels than base facts. Apart from this, different aggregation rules were identified based on different levels of hierarchies of the dimensions, user's perspective and technical interpretation. The rules of aggregation were different for different measures over different dimensions and the selection and implementation of these rules were done very carefully. For example in case of weekly/monthly information related to agro meteorology the measures are (i) temperature, (ii) humidity, (iii) rainfall, (iv) sunshine hours, (v) evaporation, (vi) rainy days, (vii) potential evapotranspiration etc. The dimensions are time with the grain level week/month and location with the grain level of weather stations in a district. It can be clearly seen that aggregation rules for different measures over the same dimension are different as application of same aggregation rule is quite absurd. Again, the aggregation rules for the same measure over different dimension are also different otherwise aggregated figures will be misleading. For example in case of maximum temperature, aggregation rule over different levels of location hierarchy may be maximum of the maximum temperature of the lower grain level. This may not be changed to average of the maximum temperature of the lower grain level as it becomes absurd for the users and will provide misleading information. Average of maximum temperature is better aggregation rule as compared to others in case of time dimension. Therefore, best combinations of aggregation rules were applied to get more information over different measures and

dimensions. The data from source to target was mapped which is the basic foundation of data staging area process. This ensured that proper analysis of the source data was done and all the data items were transformed as per user perspective and requirements delivered to the data warehouse. The map also served to identify difference between the source and target.

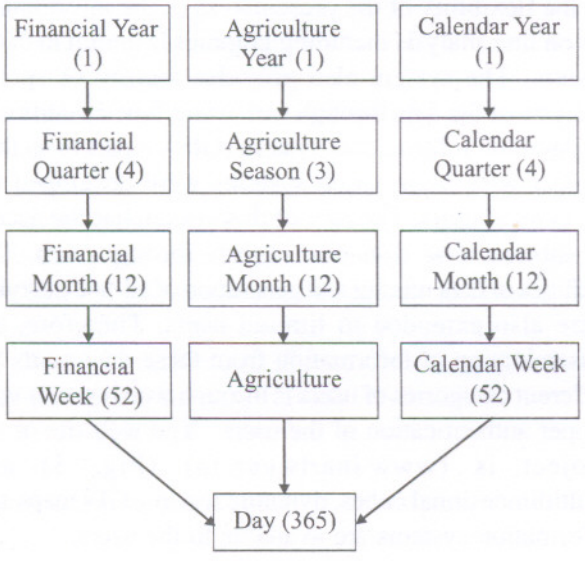


Fig. 4. Dimension table detail diagram of time dimension for different definitions of year (relative cardinalities are shown in brackets)

4. PROBLEMS IN IMPLEMENTATION

Implementation of this agricultural data warehouse had number of typical problems. Some of the problems are described in brief but details can be seen from Rai *et al.* (2007).

- **Uniformity of Information:** It has been observed that information about a particular commodity is being collected by different government/private organizations and there are wide differences. Even different government organization provides information differs so widely that users loose their confidence. The basic reasons attributed to these differences are due to non-uniformity of the definitions and methodology. The methodology for data collection by these organizations depends on the mandate of the institution, purpose, technical capabilities, available resources etc. Sometimes, it has been observed that no objective scientific methodology has been adopted for data collection. The problem of differences in the statistics becomes more prominent due to lack of coverage of the information also. These problems exists almost every sector but it is more visible in the sector where national and international trading are extensive.

- **Integration of the Information:** The integration of information related one data mart is a challenging task even sometimes not possible. The information of different parameters of a particular commodity is collected by different organizations following their own definitions and format. The grain level of information collected and published from these sources depends on their requirements and its demand on a particular period of time.

- **Aggregation Rules:** The issues of defining rules for aggregations are very pertinent in the multidimensional cubes where drill-down and drill-up facility of decision support system is the most desirable feature. In ideal conditions this problem may exists when standard data quality checks and techniques of aggregations are applied to the data before releasing the same. This problem is more dominating in the historical data generated by different government organizations as in those days there were limited options for application of above techniques due to limited computing facilities.

- **Data Quality and Data Gaps:** The data quality is very important issue for design and development of successful data warehouse. If the data qualities of the databases are not good whatever efforts are done in design and development of a warehouse is a futile exercise. The developers will loose the confidence of their users in a very short span of time and support for these developments from the management of the organization will become weak. Again, it has been found that data gaps and data quality problems are more dominant in case historical data due to lack of awareness and importance of the information generated at that time. These problems in the past were more frequent due to lack of availability of proper technical and computing support in this area. It is understood that in this era of information revolution these problems are likely to reduce very fast. Unfortunately, the information related to agricultural

sector in this area is more vulnerable as compare to other sector. The agricultural statistics are mainly collected by non-profit making government organizations and there are no direct visible losses due to data quality problems. Therefore, not much of attentions were paid in the past for data quality and data gaps. It has been seldom realized there are problems and in some cases huge losses to the country due to lack of proper planning by using these data.

Recently, the problem arises due to data quality and data gaps have gained importance due to computerization. Now, it is easy to detect and correct handle these issue after computerization. The best solution of this may be to apply check programme while computerization and development of large databases where it is easy to detect and apply corrections.

- **Selection of Hardware and Software:** Selection of hardware tools of a data warehouse tools depends on the user requirements of data storage, security, processing time, number of expected queries, number of processors and processor uptime. The cost of H/W may vary depending on the server class, vendor to vendor. The selection of data warehousing software tools depends on number of factor such as type of data, amount of data, nature and type of OLAP required by users, quality of data, heterogeneity of data sources, etc. The success of any data warehousing development process is highly related to selection of these tools. The requirement of tools varies from project to project. All developmental tools of data warehouse can be categorized in three categories i.e. (i) ETL tools, (ii) OLAP tools and (iii) reporting tools.
- **Maintenance and Updation:** The maintenance and updation of the agricultural data warehouse is a continuous process. During the maintenance there is regular change in the users demand and perspective of the user. User acquires knowledge and capacity of the technology in a very short learning curve and demand of the information changes accordingly. Further, this technology is still growing, therefore this fast growth and development of this technology put more pressure on the maintenance. The updation in case of agricultural data warehouse is not as fast as in case of commercial enterprise. In this case updation may be after three months as the agricultural data is not dynamic in nature where as in case of business sector it should be on daily basis.

5. REPORTING AND QUERYING

The information of this data warehouse is available to a user in the form of decision support system in which all the flexibility of the presentation of the information, it's on line analysis including graphics is inbuilt in to the system. The system also provides facility of spatial analysis of the data through web using functionalities of Geographic Information System (GIS). Apart from this, subject wise information systems were developed for the general users. The user of this system has the access of subject wise dynamic reports through web. The facilities of data mining and generation of ad-hoc querying were also extended to limited users. Therefore, the dissemination of information from these data marts for different categories of users is through web browser with proper authentication of the users. The web site of the project is (www.inaris.gen.in) (Fig. 5) and multidimensional cubes, dynamic reports, GIS maps and information systems are available to the users.

A multidimensional OLAP cube (Fig. 6) provides online decision support system. This online system has drag and drop option for creation of nested tables, drill up and drill down functionalities based on hierarchies of various dimensions. Simple calculation options are available on tabular data. Hide and show buttons exit to conceal/display certain rows or columns on the screen. Graphical representation options such as line, bar, chart, three-dimensional graphs etc. through single click of a button are provided. Aggregations, segregations, slicing and dicing options are available without any additional requirements.

The zero suppression option suppresses all rows/columns with no information. It also has 80/20 option, which displays major rows/columns contributing to 80% of grand total of respective row/column. This may be used to identify major contributors in the total. It has option of exceptional highlights in which well-defined exceptions can be highlighted with different text and cell colors in a displayed table. The history of a particular session of analysis is recorded and can be seen whenever it is required by single click of a button. This will help a user to look into list of operations performed

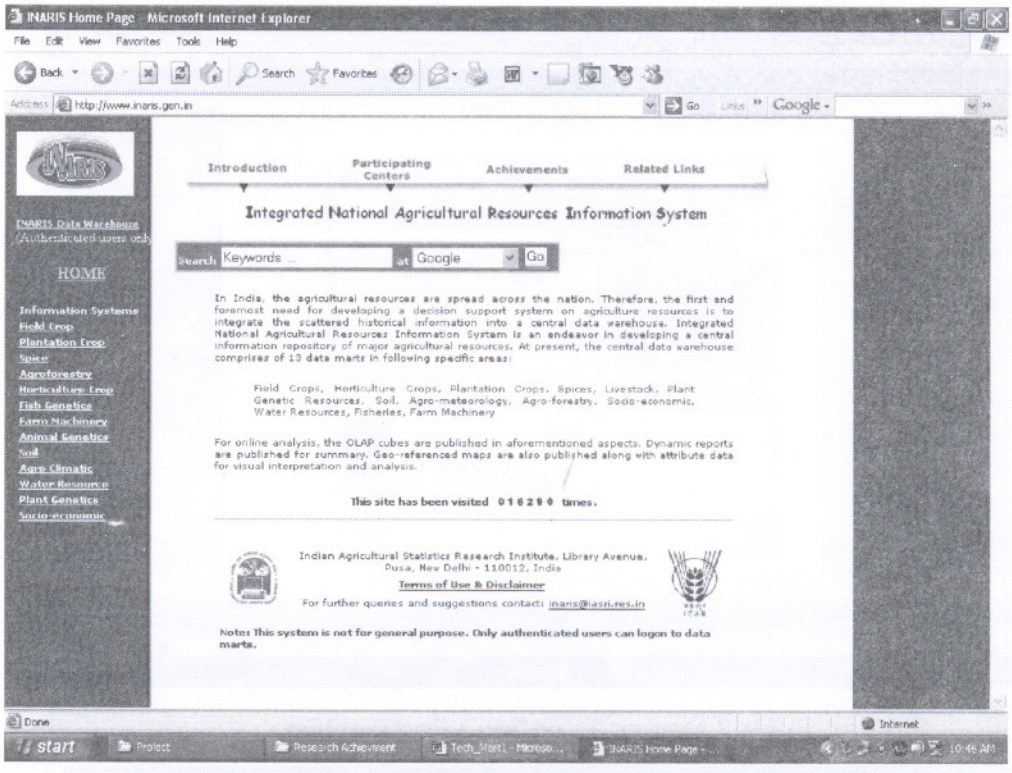


Fig. 5. INARIS Home Page

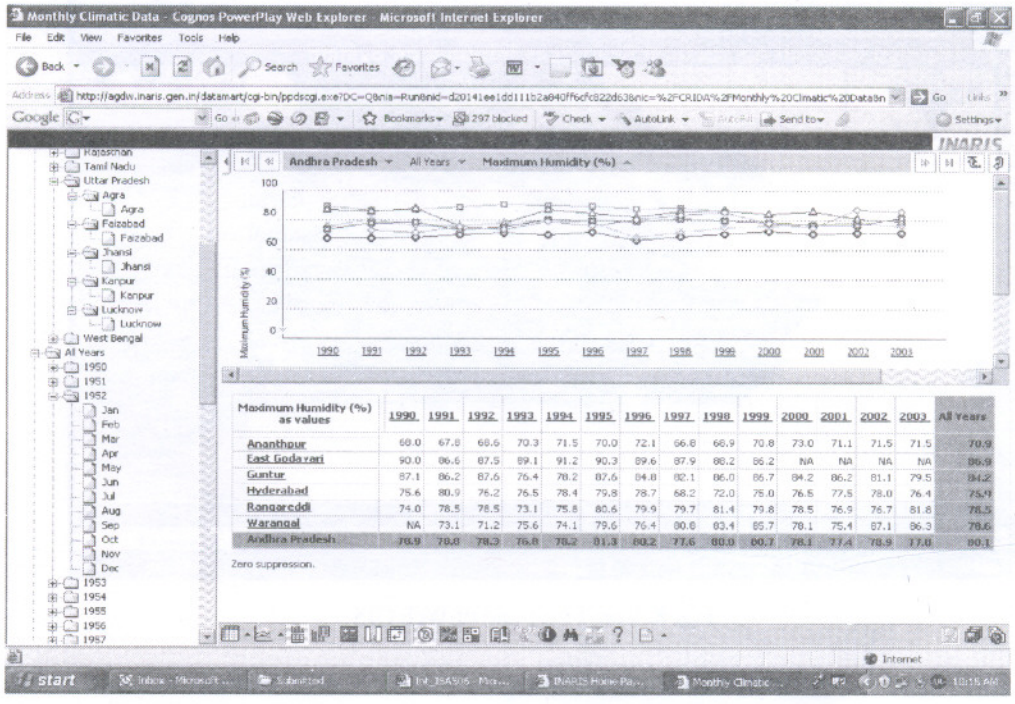


Fig. 6. Visual Representation of Multidimensional (OLAP) Cube

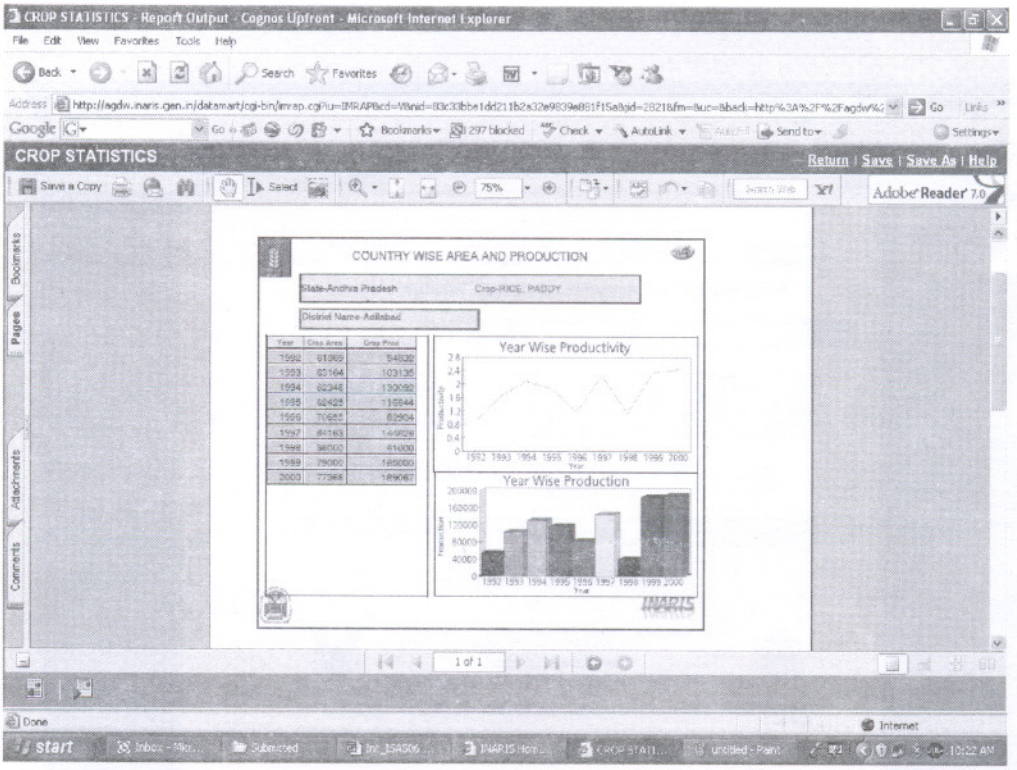


Fig. 7. Dynamic Report from Animal Data Mart

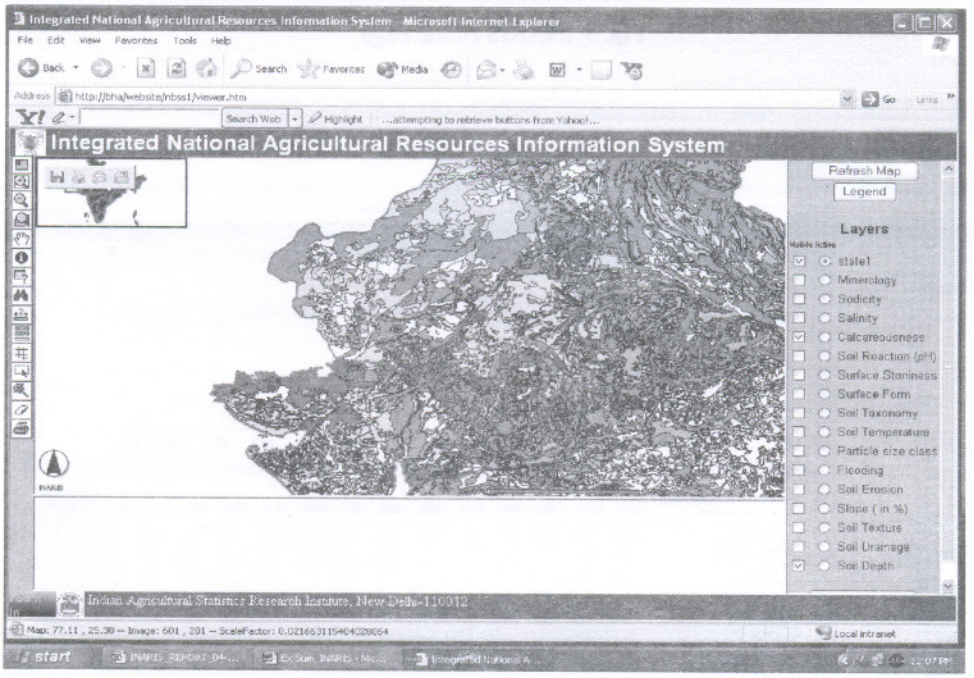


Fig. 8. User Front-end of Web GIS

by him in a particular session. The find and search option is available for locating a particular information in tabular data of a cube. It also has help option to assist user in working with this decision support system. The most important option is that any table/graphs can be exported to different formats as per requirement. Apart from this option, it has functionalities of disseminating information through dynamic reports. In dynamic reports, a query is fired through web browser to the data warehouse database and preformatted reports are generated on-line in PDF/HTML format (Fig. 7).

This decision support system also has functionality of on-line spatial analysis of information in a data mart. In this case, a user has access to web based GIS and need not have to install any GIS software. It has all simple and routine functionalities including layer analysis, spatial querying functionalities etc. (Fig. 8).

6. CONCLUSIONS

In INARIS project, a State-of-Art Central Data Warehouse (CDW) of agricultural resources of the country has been developed at IASRI, New Delhi. This provides systematic and periodic information to research scientists, planners, decision makers and developmental agencies through implementation of latest information dissemination technology i.e. On-line Analytical Processing (OLAP) decision support system. The development of agricultural data warehouse has its own inherit problems which are completely different then data warehousing in business environment. The basic different in both these system are data generation, validation and application process. In case of business process, the data generation and its utilization are mostly confined with in organization. In case of agricultural sector in India, data is being collected by different government organizations adopting different approaches and as such development of an agricultural data warehouse is a challenging task. It, therefore, requires in-depth understanding of data collection and compilation process.

REFERENCES

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A. and Paraposchi, S. (2001). Designing data marts for data warehouse. *ACM Trans. Software Engg. Methodol.*, **10(4)**, 452-483.

Chaudhuri, S. and Shim, K. (1995). An over view of cost-based optimization of queries with aggregates. *IEEE Data Engg. Bull.*, **18(3)**, 3-9.

Chen, R., Chen, C. and Cheng, C. (2003). A Web-based ERP data mining system for decision-making. *Int. J. Comp. Appl. Tech.*, **17(3)**, 156-158.

Gupta, A. and Mumick, I.S. (1995). Maintenance of materialized views: Problems, Techniques, and Applications. *IEEE Data Engg. Bull.*, **18(2)**, 3-18.

Inmon, Bill (2005). *Building the Data Warehouse.* Fourth Edition, John Wiley, New York.

Kambayashi, Y., Kumar, V., Mohania, M., and Samtania, S., (2004). Recent Advances and Research Problems in Data Warehouse. *Lecture Notes in Computer Science*, **1552**, 81-92.

Kimball, R. (1998). *The Data Warehouse Lifecycle Tool Kit.* John Wiley & Sons, New York.

Kimball, R., and Ross, M. (2002). *The Data Warehousing Toolkit.* John Wiley & Sons, New York.

O'Neil P. and Graefe G (1995). Multi-table joins through bitmapped join indices. *ACM SIGMOD*, **24(3)**, 8-11.

O'Neil, P. and Quass, D. (1997). Improved query performance, with variant indices. *ACM SIGMOD*, **26(3)**, 38-49.

Rai, Anil, Dubey, V., Chaturvedi, K.K. and Malhotra, P.K. (2007). Issues of design and development of agricultural data warehouse in India. *CSI Comm.*, **31(1)**, 43-51.

Zhuge, Y., Garcia-Molina, H., Hammer, J. and Widom, J. (1995). View maintenance in a warehousing environment. *ACM SIGMOD*, **24(2)**, 316-327.