J. Ind. Soc. Agril. Statist.
62(1), 2008 : 56-64

On-line Analytical Processing in Agriculture using Multidimensional Cubes

K.K. Chaturvedi, Anil Rai, Vipin K. Dubey and P.K. Malhotra
Indian Agricultural Statistics Research Institute, New Delhi
(Received: October 2006)

SUMMARY

The multidimensional modeling tools, which are of recent origin in the field of data warehousing, have vast potential for development of Web-enabled decision support system through deployment of Web based multidimensional cubes. In this article, an attempt has been made to present experiences of the authors related to different processes and techniques involved in design and development of multidimensional cubes in reference to agricultural sector during implementation of Central Data Warehouse (CDW) at Indian Agricultural Statistics Research Institute (IASRI), New Delhi. The features of multidimensional model from the user's perspective have been highlighted to demonstrate the power of this On-line Analytical Processing (OLAP) technology.

Key words: Analysis, Data warehouse, Modeling, Multidimensional cubes, OLAP.

1. INTRODUCTION

In most of the organizations, especially in agricultural sector, huge amount of data are being collected for planning and decision-making. However, only a fraction of these data are utilized for desired purpose. The basic reason for non-utilization of huge amount of the data may be due to its non-availability in suitable digital format and lack of powerful Graphical User Interface (GUI) for extraction of information content from this data. Ultimately, the data, which have been collected by investing huge public resources, are not being properly utilized in developmental and planning process of the country. In a developing country like India, where agriculture is one of the most important sectors of economy, this problem is quite alarming.

In the past, major efforts were made for the technological development of Relational Database Management System (RDBMS) for storing, updating and retrieving limited information. The backbone of this technology is based on Entity-Relationship (E-R) modeling (Inmon 2005, Kimball 2002). The E-R modeling is a standard technique for building an On-Line Transaction Processing (OLTP) system. An OLTP application is a system, designed for many but simple concurrent (and updating) requests. It is also useful for communicating the way data flow from one stage to the next on its way from source system to user. This system is primarily used for information updating and simple querying/reporting. In case of data analysis and its modeling where huge amount of records needs to be retrieved from the data repository, the performance of OLTP system is not satisfactory due to its technological limitations.

Technological advances in the field of Computer Science such as design and development of databases, data warehousing, networking and Web-based systems, etc. address many of these concerns. The multidimensional modeling techniques, which are of recent origin in the field of data warehousing are being used for development of Web-enabled decision support system through deployment of Web-based multidimensional cubes (Chaudhari *et al.* 1995, Zhuge *et al.* 1995). The design and development of multidimensional cubes for On-Line Analytical Processing (OLAP) system in the field of agriculture

is challenging task. Data analysis is the process of extraction of information from large databases. The data analysis using OLAP is able to provide information in understandable format about the present status and predicts future trends in the concerned domain of agriculture. Further, data analysis using OLAP is a sub-set of larger process called knowledge discovery, specifically, the steps in which advanced statistical analysis and modeling techniques are applied to the data to find useful patterns and relationships.

In this article, an attempt has been made to discuss experiences of the authors related to different processes and techniques involved in design and development of multidimensional cubes in reference to agricultural sector during implementation of Central Data Warehouse (CDW) at Indian Agricultural Statistics Research Institute (IASRI), New Delhi. This CDW has been developed under the National Agricultural Technology Project (NATP) Mission Mode sub-project entitled "Integrated National Agricultural Resources Information System (INARIS)". In this CDW, 13 different data marts related to various subjects in agriculture were designed, implemented and integrated. Initially, a brief description about the INARIS project has been provided for better understanding of the information contents available in this CDW. The problems pertaining to the design and development of multidimensional cubes along with its implementation in context of agriculture are presented in the subsequent section of this article. The features of multidimensional model from the user's perspective have also been highlighted to demonstrate the power of this OLAP technology.

2. INTEGRATED NATIONAL AGRICULTURAL RESOURCES INFORMATION SYSTEM

This Project was taken up as a sub-project under NATP. The mission set for this project was to design and develop a flexible state-of-the-art CDW of agricultural resources of the country at IASRI, New Delhi. The above project has been implemented with active collaboration and support from 13 other Institutes affiliated to Indian Council of Agricultural Research (ICAR). This CDW contains information related to soil resources, water resources, agro-meteorology, field crops, plantation crops, horticultural crops, spices, agro forestry, cropping systems, plant genetic resources, livestock resources, fish resources, agricultural implements and machinery, and socio-economic resources. Architecture of CDW developed under INARIS is presented in Fig.1.

Fig. 1. Architecture of CDW of INARIS

In all, 13 data marts were created from 59 databases of different subject matter institutions under this CDW. These databases were implemented on Oracle 9i Relational Data Base Management System whereas CDW has been implemented using Cognos products (Decision stream designer, Impromptu and Powerplay). Each of these data marts has information related to statistics, technologies and research projects of related area. The validation checks were implemented wherever possible to provide the quality data for the end users. These validation checks are of different forms such as detections of outliers, consistency checks of the data flow and its aggregations from lower level to upper level in a hierarchy, aggregation checks within the same level of hierarchy, etc.

This data warehouse provides on-line functionalities of spatial and non-spatial data analysis. This analysis includes generation of reports and graphs as per the specific requirements. It has also capability to perform on-line exploratory analysis of the data for better decision making process. In this system, subject wise information is stored to ensure better response time. This type of storage system is also known as data

marts. Table 1 provides brief description about data marts available in this system.

Table 1. Description of different data marts of INARIS

S.No.	Data Mart	Description
1.	Field Crops	Area production statistics and Crop management practices of field crops
2.	Plantation Crops	Area production statistics and Crop management practices of plantation crops
3.	Horticultural Crops	Area production statistics and Crop management practices of horticultural crops
4.	Spices Crops	Area production statistics and Crop management practices of spices crops
5.	Agro-Forestry	Area production statistics and Crop management practices of agro forestry resources
6.	Fish Genetic Resources	Species characterization statistics and habitat
7.	Animal Genetic Resources	Animal populations, Breeds and Species information
8.	Water Resources	Water resources availability, consumption and its use
9.	Soil Resources	Thematic geo-referenced soil maps
10.	Plant Genetic Resources	Germplasm information of various plant species
11.	Farm Machineries	Total number of farm equipments and their description
12.	Climatic Parameters	Climatic parameters on monthly and weekly basis
13.	Socio-Economic Resources	Socio-economic information related to statistics at national, state and district level

The dissemination of information from this data warehouse for different categories of users is through Web browser with proper authentication of the users. The Uniform Resource Locator (URL) of the project is http://agdw.iasri.res.in and the multidimensional cubes, dynamic reports, GIS maps and information systems are available to the users through this Web site.

3. DIMENSIONAL MODELING

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 2002). The data warehouse is maintained separately from operational databases. Since, data warehouses contains summarized data, perhaps from several operational databases over potentially long periods of time, they tend to be much larger in magnitude than operational databases. The organizational data warehouses are projected to hundreds of gigabytes or terabytes in size. The query, mostly adhoc in nature, can access millions of records and can perform a lot of scans, joins and aggregates (Gupta *et al*. 1995, O'Niel *et al*. 1995, 1997). In this system, query throughput and response time are more important than transaction throughput.

In case of complex analysis and visualization, the data in a data warehouse is typically modeled in multidimensionality. It requires some of the additional steps with OLTP systems. This type of storage 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 cube. If any user is retrieving only facts irrespective of any dimension, it is known as 0-D cube. This type of cube will be meaningless at most of the time. If it contains information specific to one field or dimension, then it is known as 1-D cube. Similarly, if it contains two dimensions or fields, it is known as 2-D cube. In the same way, if the table contains n-dimensions or fields for visualization of facts, it is known as multidimensional cube. For example, consider the problem of visualization of data related to agricultural production of the country through multidimensional cube for understanding the pattern of crop productivity and change in cropping pattern over the country. The development process of multidimensional cube on these aspects needs clear understanding of the crop production and the process of generation of information in this domain. Basically, it has four broad aspects (dimensions), i.e., (i) crop under consideration; (ii) locations, where it is grown; (iii) season, in which it is grown; and (iv) year(s) of interest. The administrative setup of the country has hierarchical structure. Country is divided into 28 States and 6 Union Territories (UTs) which are further divided

into several districts whereas every district is further divided into tehshils/taluks/mandals, blocks and villages at the lower level. Crop productions of different crops are dependent on season as India has three broad seasons, i.e., Rabi, Kharif and Zaid. The reference year is important to understand the trend of crop production climatic conditions. The problem of building this multidimensional cube can be represented in Fig. 2.

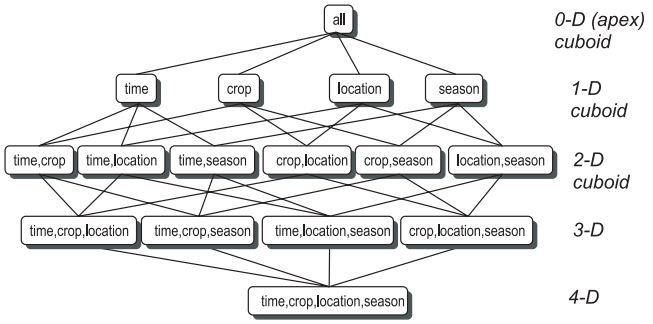

Fig. 2. Various types of dimensional view of a Cube

In case of data warehouse technology, the key performance indicators are known as facts. Facts are additive, non-additive and semi-additive. These are the numeric data items used to satisfy all calculations options that are of interest to the end user. The aggregated fact information can be viewed as a 0-D level, i.e., at the summary level, which is the top most hierarchical level. The same information can be viewed at 1-D level in context of time or location or crop or season. The table which provides the context to the fact is called dimension table. Further, fact can be analyzed in combination of two dimensions at 2-D level and with respect to three dimensions at 3-D level. The detailed information can be visualized further to the lowest grain level. Granularity of information is the lowest level of fact information. Any combinations of dimensions are possible along with slicing and dicing of dimension on various axes while browsing the OLAP cube helps deeper analysis of fact data to make better decision making.

These OLAP cubes can be designed by using the fact and dimension tables from the database. Since, these cubes are published on the 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. The drill down and roll-up functionalities of an OLAP cubes are designed along the data flow hierarchies. A hierarchy describes the organizational structure and logical parent-child relationship within the data. For example, the District is at the lowest level in the location hierarchy. The State is the next upper level in the hierarchy. All States constitute country level information which provides the data of all districts for all states (Fig. 3).

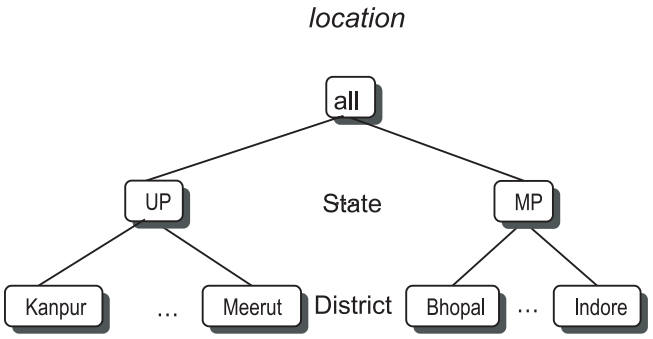

Fig. 3. Hierarchical view of Location Dimension

After deployment of these multidimensional cubes on the server, end user can perform analysis and export the desired result to desktop in any form such as, MS-Word, MS-Excel, ASCII Text file or well-known Adobe's Portable Document Format (PDF), etc. The key steps involved in designing a dimensional data model include

1. Identifying the major processes and hence the required fact tables. There is a need to identify the major aspects involved in the process. Numerical aspects can be put into fact table. The fact table also contains all the dimensions on which the facts are based.
2. Deciding the granularity and dimensions of each fact table. Identify the lowest level of granularity.
3. Defining the measures needed for reporting and analysis for each fact table including derived measures and full descriptions.
4. Aggregating the attributes, levels and hierarchy for each dimension, including all labels and full descriptions.

Schema for Dimension Modeling: The schema for a dimensional model represents relationships among fact and dimensional tables. There are mainly two types of schemas namely star schema and snowflake schema that are usually followed in the process of designing

multidimensional cubes. A schema is called a star schema if all dimension tables can be joined directly to the fact table whereas a schema is called a snowflake schema if one or more dimension tables do not join directly to the fact table but must join through other dimension tables. Generally, snowflake schema is used when the records in fact tables are less as compared to the dimension table. In case of INARIS data warehouse, this specific situation was not encountered in any of the data marts. Therefore, all data marts were designed using star schema. For example, in case of multidimensional cube related to crop area and production, the fact table stores statistics on crop area and production based on the foreign keys TimeKey (for storage of year wise historical information), CropKey (for storage of information on different crops in the fact table), SeasonKey (for storage of season wise information in the fact table) and LocationKey (for storage of information related to different administrative regions such as district or states in the fact table). This makes a star type of structure around fact table and is known as Star Schema (Fig. 4). This type of design is

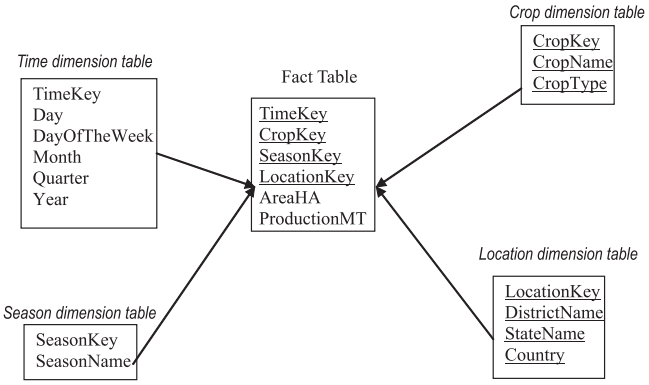

Fig. 4. Star Schema

useful when records in the fact tables are more with limited number of dimensions.

Some basic steps followed for making the star schema are

- Database design should be de-normalized as normalization deteriorates the querying performance.
- Primary or foreign key constraints should not be enabled.
- Bitmap indexes on every column of every table should be created as it improves the querying performance.

4. DESIGNING MULTIDIMENSIONAL CUBES IN AGRICULTURE

Macro level planning and decision support processes in agriculture require access to data on different resources (e.g., crops, livestock, fisheries, etc.) at varying levels of detail or aggregation. These and other requirements must translate to dimensions and fact tables of the OLAP cubes. Location, Time, and Commodity are few of the common dimensions that transcend all warehouse models. But Location and Time dimensions pose biggest problems in integrating data from the varied sources within the agricultural sector.

The problem of integration basically boils down to the four most common dimensional issues as follows

- Granularity of Location differs among different sources
- Granularity of Time varies among different sources
- Multiple overlapping time domains and
- Aggregation and disaggregation of information at different dimensional hierarchies

These dimensional issues lead to the design issues of the fact table.

Grain Level of Locations: Similar to business sectors such as retail, telecommunications, and others, the agricultural sector also uses the Location dimension extensively for its warehouse applications. In Indian agricultural sector, the Location dimension presents a number of interesting issues. Location, also known as the geography dimension, usually has a clearly defined hierarchical structure. In this, hierarchy is determined by the administrative mechanisms implemented by Government of India. The highest-level grain level of hierarchy is the entire country (National level) followed by states. Further, each state is divided into districts, which are further divided into tehsils, blocks and villages, respectively in the lower hierarchy. Generally,

information related to agriculture is collected at district or above level. Information related to livestock, agricultural census, etc., conducted in the country is available at village level which is the lowest grain level in this hierarchy.

Organizations may collect information at any or all levels of the Location hierarchy. For example, total quantity of exports and imports of different agricultural commodities such as fruits, vegetables, livestock products, tea, coffee, fish products, etc. are collected at the National level. At the National level, aggregate data are compiled through surveys/census. Apart from this other information related to the prices at the international market of these commodities is also available at this level. In the country, prices of different commodities are available on daily/weekly basis, whereas import and exports figures are available for each month but not at daily and weekly level. In contrast to the National level, the State level of the Location hierarchy supports a richer domain of sources. Information on all commodities is available at this level. Different sample surveys are conducted to get production data about some of these commodities such as fruit crops, plantation crops, etc. at the State level. At the District level, most of the information is available except which is available at the higher levels. The information available at this level is production of crops, livestock products, fisheries products, land utilization statistics, etc. Since, it provides a more detailed measure of the factors, information at this level is highly coveted by planners and decision makers. There are several fourth level attributes in the hierarchy but the most important one is the Village level data. At the Village level, the following types of data are available; data related to land utilization, census data on human, livestock, and demographic and static parameters of planning such as land ownership, employment, etc. Data on agricultural commodity trades are available at the Agricultural Market or Mandi (trading place) level. Price data from many of the important markets are collected on daily/weekly basis depending on the season of the crop/commodity. Furthermore, different agro-meteorological stations also produce information on climate and weather conditions, albeit on a daily basis and form another fourth level hierarchy attribute. Table 2 summarizes the overall picture of different levels of the Location dimensional hierarchy along with examples of availability of the information related to agricultural sector.

Table 2. Information collected at different hierarchy levels

Hierarchical Level	Hierarchy Name	Example
Level one	National	Import and export statistics (monthly), International prices (daily/weekly), Production of minor crops (annual) etc.
Level two	State	Production of major fruits and vegetables (annual), National accounts statistics (annual), Information related to various agricultural development projects at state level (annual) etc.
Level three	District	Production and area of principal crops (annual), land utilization statistics (annual) etc.
Level four	Village	The information related to land use (annual), information on different census such as human (decennial), livestock census (quinquennial) etc.

Another problem is encountered with respect to historical data is that information on production of some of the commodities is presently available at the district level but historical data is only available at the state level. Availability of resources, requisite need for information, and policies present at that time had an effect on the collection at any given level. The following issues are associated with creation of dimensions in the above situations in the development of OLAP cubes

- How many levels are required for the location?
- How to integrate the information coming from different sources (organizations) at different grain level?
- How to define fact table for these dimensions?

It is evident from Table 2 that at the fourth level of hierarchy i.e. village, there are three candidates at the lowest level. Note that data from these dimensional attributes cannot be merged together. It is also not possible to represent them with a common name as the other attribute information associated at this level is

quite different. For example with village, name of the village, name of the block/tehshil/district and name of the state is associated with the dimension. In case of agricultural market, 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, is going to be associated with the dimension. Finally, in case of agro-meteorological station, the longitude, latitude, altitude, name of the place and other agricultural parameters such as soil type etc. are associated with the dimension.

Rule of aggregations for each of the fourth level hierarchy of the Location dimension to roll up to the next higher level of hierarchy is different. In case of villages it may be a simple aggregation, whereas in case of agricultural market where the parameter is price, a simple aggregation will not work - a weighted average or any other suitable method might be applied and in case of meteorological parameters one has to apply spatial models to interpolate or extrapolate the data at the district level. It may be further noted that agricultural market or weather stations are not available in every district of the country. Availability of agricultural markets in different states for a particular commodity depends on its area, production, and consumption. Also, not every state or district in a state produces all commodities; accordingly markets are located. Hence, it may not be possible to aggregate the lower level data for the each district or in that sense for each state for the upper level of hierarchy. So in this case our hierarchical structure of the dimension will either collapse or provide misleading information to the user.

The integration of information coming from different sources, especially from various organizational sources, is also a big challenge in the design of the OLAP cubes. Data collection takes place at different levels (national, state, district, etc.) utilizing different methods (surveys, census, observations, etc.) 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. More importantly, each source and method contributes to different kinds of errors. In spite of these issues, if any information is available at the lower level, it is relatively simple to aggregate (roll up) to the higher level by taking care of all the above-mentioned factors. But when information is only available at the higher level and we have to disaggregate (drill down) to lower level, the task becomes difficult. Most of the information related to agriculture is collected through agricultural surveys or census which are designed to elicit responses at the national or state level. The surveys collect information about agricultural products and resources from respondents at the state or national levels. The regional or lower level estimates cannot be obtained from these with reliable precision. Although, sampling strategies are employed in collecting data at these levels, the assumptions of sampling design do not hold well at lower levels. So, for handling this situation, we need to have highly sophisticated statistical or mathematical modeling techniques. We must also provide estimates of errors / risk associated with the data so that users have acceptable levels of confidence with the analyses.

5. OLAP ANALYSIS

The developed cubes are being utilized for data analysis. Data analysis is the process of extraction of predictive information from large databases. The data analysis using OLAP cube may be able to provide information about the present status and predict future trends in the concerned sector of economy, which may be difficult otherwise. These can be analyzed using flexible and user friendly features of these cubes such as drill-up, drill-down, slice and dice, etc. The OLAP cubes provide user-friendly environment for interactive data analysis.

In order to see the fact data of a particular data mart say field crop, in an OLAP which has been published on Web, user needs to log on to the system and is required to provide authentication information to access these cubes for on-line analysis. A user has to open the cube with a click of mouse and default fact table view will be opened. In this, user has all the flexibilities to design the table structure through drag and drop of various dimensional parameters of interest. It has functionalities such as creating different graphs, simple statistical calculations from the table rows and columns, swapping of rows and columns of the table, option for suppression of rows and columns with nil information, it also has option for identifying major

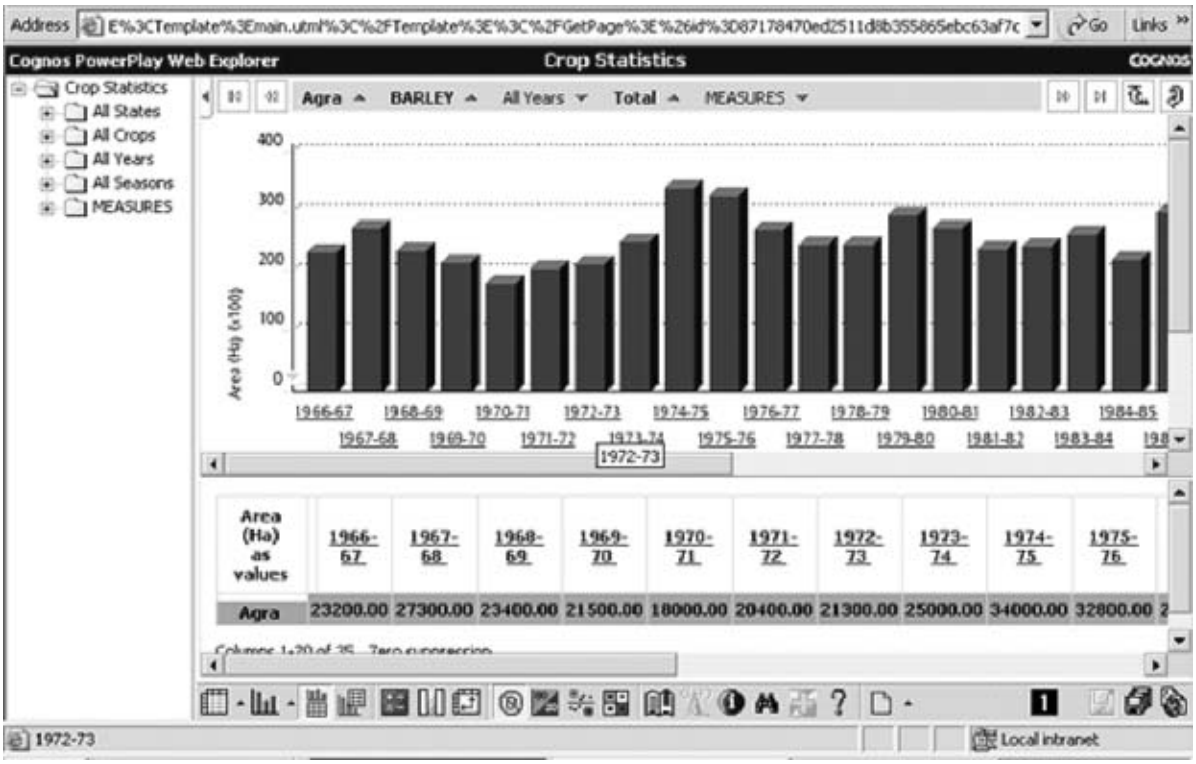

Fig. 5. Graph and Data view

contribution rows or columns in the marginal totals, user can highlight exceptional values of the table through automatically as well as user defined criterion. The standard bars and charts can be displayed on the screen by simply clicking the respective options. This includes simple bars, pie diagram, multi-line graphs, 3-D bars, cluster bars, etc. In case any user wants to display the chart as well as data together, he/she can click on that corresponding third icon on the icon bar to see data as well as graph at the same time (Fig. 5).

Data can be sorted in table element by selecting a particular row or column and selecting ascending/ descending orders. If any user wants to use the lower grain level of the data, he can click on these displayed rows/columns, the data will be displayed accordingly with the help of drill down facilities (Fig. 6).

Fig. 6. Drill Down in Cube

In case of tabular display, user has number of options for displaying the contents of the table such as percentage of row total, percentage of column total, percentage of row sub-total, percentage of column sub-total, percentage of grand total, etc. Further, this system has facility to store session history of the analysis performed by a particular user. The graphs and analysis performed through OLAP cubes can be directly exported for report preparation. Tabular data displayed on the screen can be exported to popular formats for further analysis. Further, on-line help is also available with the system to provide operational assistance to users.

6. CONCLUSION

Indian agriculture is highly diversified with respect to crops grown, climatic conditions, soils types, socio-economic and cultural conditions. These heterogeneous agricultural production systems in the country pose formidable challenges for decision makers, research managers and agricultural scientists for appropriate planning and development of agricultural research system. In order to develop on-line decision support system for management of agricultural resources of the country, an attempt was made to develop a Central Data Warehouse (CDW) under a NATP Mission Mode sub-project named as INARIS. In this project, 13 data marts pertaining to different agricultural resources were integrated and a OLAP system was developed for on-line exploratory analysis of heterogeneous agricultural information. Development of OLAP system for agricultural resources is very challenging task due to diversification and complexity of this sector in India. Since, many organizations are manually involved in data collection and compilation, the consistency, uniformity and standardization of information in homogeneous conceptual format is an important component of data modeling for integration. In the process of data modeling, it is important to resolve dimensional issues specific to agriculture such as un-balanced and un-covering hierarchies, differences in the grain levels, etc. Generally, OLAP systems are developed for research managers, therefore these systems should be designed in a simple manner with high flexibility so that most of the user requirements are fulfilled. OLAP developed in this project has all these functionalities which makes it very powerful tool for decision making. This article provides solutions to specific problems in designing multidimensional cubes in agriculture. Also, problems of dimensional modeling in the field of agriculture have been discussed in detail. Further, it also demonstrates the power and flexibility of OLAP with respect to user's perspectives.

REFERENCES

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

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

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

Kimball, R. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* Second Edition, John Wiley & Sons, New York.

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

O'Neil, P. and Quass, D. (1997). Improved query performance with variant indices. *Proceeding of ACM SIGMOD International Conference on Management of Data,* Tucson, Arizona, United States, 38-49.

Zhuge, Y., Garcia-Molina, H., Hammer, J. and Widom, J. (1995). View maintenance in a warehousing environment. *Proceeding of ACM SIGMOD International Conference on Management of Data,* San Jose, California, United States, 316-327.