Proceedings of Placrosym XV (2002): 784-790

DESIGN OF DATA MARTS FOR PLANTATION CROPS

N. Ravi Kumar, K. Muralidharan, C.V. Sairam, C. Palaniswamy, R. Dhanapal S. Arulraj, Anil Rai*, Vipin Dubey* and K. K. Chaturvedi*

Central Plantation Crops Research Institute, Kasaragod

* Indian Agricultural Statistics Research Institute, New Delhi

ABSTRACT

Plantation crops occupy a significant position in the agricultural sector and play a significant role in the Indian economy. To apply the Information Technology tools, it is required to link all the relevant data sets in to a single system. To achieve this objective, the Indian Council of Agricultural Research (ICAR) has launched a mission mode project entitled "Integrated National Agricultural Resource Information System" (INARIS)' involving 14 research institutes. The efficiency of the databases at various centres could be enhanced by providing solutions to customer queries in an integrated manner, for which data marts are to be developed. In this paper, basic concepts of designing of the data marts are applied to the plantation crop sector. Three data marts viz., statistics, agro-techniques and research were identified in the field of plantation crops as a part of Centralized data warehousing of national agricultural resources. Different aspects of data mart design were delineated. The detail and summary levels of data in the data mart were documented. Developing data marts will help to evolve decision support tools in respective area of specialization without much cost involved. Finally, the data warehouse would be built-up by joining the data marts.

INTRODUCTION

Plantation crops play a significant role in agricultural sector and Indian economy. For achieving sustainable growth in this sector, long-term and vision oriented perspective planning is required. Creation of a sound database system is a basic necessity for implementation of the same. Voluminous data on various aspects of plantation crops are available at national/international levels but not in any database format. To apply the Information Technology tools, it is required to link all the relevant data sets in to a single system. To achieve this objective, the Indian Council of Agricultural Research (ICAR) has launched a mission mode project entitled "Integrated National Agricultural Resource Information System" (INARIS). Fourteen institutes are involved in the development of this information network representing field crops, horticultural crops, plantation crops, spices, Bureaus of genetic resources, natural resources management, agricultural engineering and policy research. The Lead Centre is Indian Agricultural Statistics Research Institute, New Delhi. Based on the interaction among the basic resources like soil, water, climate, animal and vegetation that form the prime components of the production system, this network will help in determining the carrying capacity of the region. This information system will be intensively used with an ultimate aim of enhancing better quality of life of the farming community and society at large.

The efficiency of the databases at various centres could be enhanced by providing solutions to customer queries in an integrated manner, for which data marts are to be developed. Data marts take explicit knowledge and helps to create new tacit knowledge and figure out what to do with it. Development in this direction will also regulate the traffic in the proposed centralized data warehousing of agricultural resources under the INARIS. In this paper, basic concepts of designing of the data marts are applied to the plantation crops.

Data Marts

A data mart is defined as a subject-specific subset of data called from the main data repositories and made available for analytical purposes. The data mart generally contains nuclear source data and often aggregations, to facilitate analysis. The data can exist at both the detail and summary levels in the data mart. The data mart can be populated with data taken directly from operational databases. Because the volume of data in a data mart is less, than the query processing as in relational database, it is often faster.

Data marts have dramatically lowered the cost for decision-support system creation and operation and has put the deployment of the decision support technology within the

affordability range of a much larger number of users.

Understanding the requirements

Participation of data providers as well as data users is essential for the development of data marts. To meet this end, a requirement analysis workshop was conducted at CPCRI. The workshop identified three major databases for the plantation crops; namely databases on Agro techniques, Statistics and Research. Agro techniques database includes subjects like cultivars/hybrids, crop production, crop protection and post harvesting technology. Region specific information will be given wherever necessary by overlying the geographical information. The Statistics database deals with area, production, quantity and value of export and import on different products and processed products. The research database gives institutes details and project information. The data items, parameters, frequency of updating and source of information were documented by Muralidharan *et. al.*, (2001). All the participating centres of the project presented the structure of respective databases together and decided to follow a common coding structure. It was decided that CPCRI should maintain data base on the following plantation crops: coconut, arecanut, cocoa, cashew, and oil palm.

Outlining the set of information needed

Based on the deliberations in the aforesaid workshops, subject matter specialists discussed the type of information to be maintained in the data marts. Accordingly, database structures were designed and fact tables and dimension table were created.

i) Dimension table

Dimension table encapsulates the attributes associated with facts and separates these attributes into logical groups. Dimensions contain fields to give lookup information. It also has a primary key to identify uniquely the records in a dimension. Obviously, a value in a primary key can appear only once in the dimension. A series of dimension tables were developed covering the these database. An illustration of dimension table with regard to price statistics is given in Fig. 1.

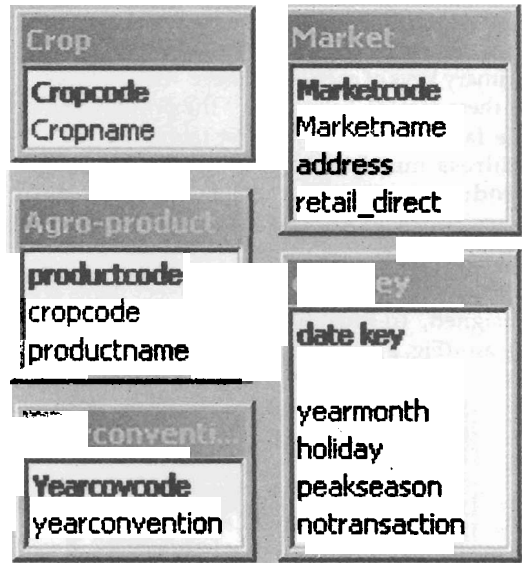

Fig. 1

In Fig. 1, there are five dimensions viz., crop, market, date, year(convention) and agro-product. The dimensions (or tables) have two or more fields. The crop dimension have two fields which - crop names and corresponding unique code. which is the primary key of this dimension. The date-key dimension is having eight fields namely date-key, date, month-year, weekly, year, holidays, no-transaction, and peak-season which store details of the data availability time. The date-key is the primary key. Market table is having four fields with market code as the primary key. Agro-product table have two fields which stores agro product code and product name. The year-convention table gives type of year followed by the source data. Reporting of the annual price data for different commodities is not uniform: Some of the commodities are reported in financial year, some in agricultural year or calendar year. Yearcode is the primary key for this table. These primary keys can be called as surrogate key in the data mart.

ii) Fact table

Fact tables were designed for the data mart identified viz., plantation crops statistics. Fact tables contain subject details for summarization. Surrogate Keys from the dimension tables are

taken for creating the primary keys of the fact table. Fact table contains a foreign key column for the primary keys of each dimension. The combination of these foreign keys defines the primary key for the fact table. Multiple fact tables are used to address multiple subject functions such as production statistics, trade statistics, crop management aspects etc. each subject item has its own fact table.

Fact tables for various subject items were designed. To illustrate, the fact table for price details (Fig. 2) is presented there.

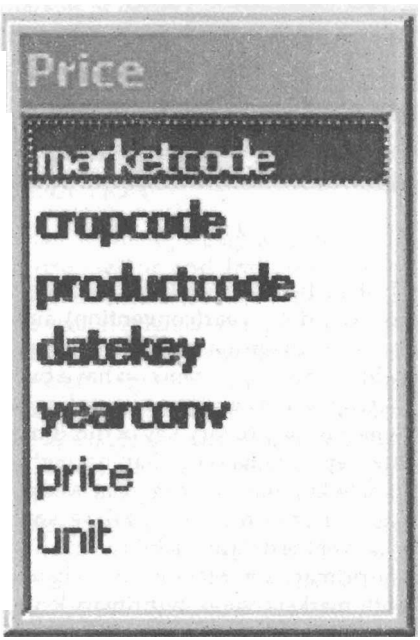

Fig. 2

In Fig. 2, six fields of the price table is shown. Here, price and unit fields are facts and other four fields are primary key of price fact table. For example, monthly price data of coconut was fed into fact table from April, 1977 to July 2002. the daily price data of coconut was fed from 01-03-1998 to 31-07-2002. Similarly, year wise data on arecanut prices was fed from 1990-1991 to 2000-2001 (Financial year).

The description of all the fields of various fact tables as well as dimensions together with definition of possible values the field can possess were documented in the data dictionary of the design.

Defining the data transformations required

The data flow for building a data mart consisting of an SQL Query (for capturing source data), a Batch loader (to specify the relational RDBMS loader to be used to load the data mart) and optionally one or more coded transforms for transforming and enhancing captured data before it is loaded into the data mart. Data flows from source (operation databases) were created for different data marts. Fig. 3 is an illustration on data transformation SQL queries with reference to the statistics data mart.

Defining schema for efficient querying and analysis

i) Design the dimensional model

User requirements and data realities drive the design of the dimensional model, which must address subject needs, grain of detail and what dimensions and facts to be included. This model design supports Online Analytical Process (OLAP) cubes to provide instantaneous query results for analysts. The dimensional model may produce a star schema or a snowflake schema.

ii) Star schema

The first step of the data flow captures data from an operational database, and at the last step loads the captured data in to the fact table of a data mart, which has been designed as a star schema. All dimension tables are joined directly to the fact table.

For the design under consideration, the star schema was used for the statistics data mart (Fig. 4).

iii) Snowflake schema

In snowflake schema, one or more dimension tables do not join directly to the fact table but must join through other dimension tables. The snowflake schema was adopted for most of the data items in the ago-tech data mart. For example, a dimension that describes pest and its control methods are separated into three tables. If user wants to access control methods of a pest of particular crop, the information is being fetched from the following four tables (Fig.5).

Fig. 3

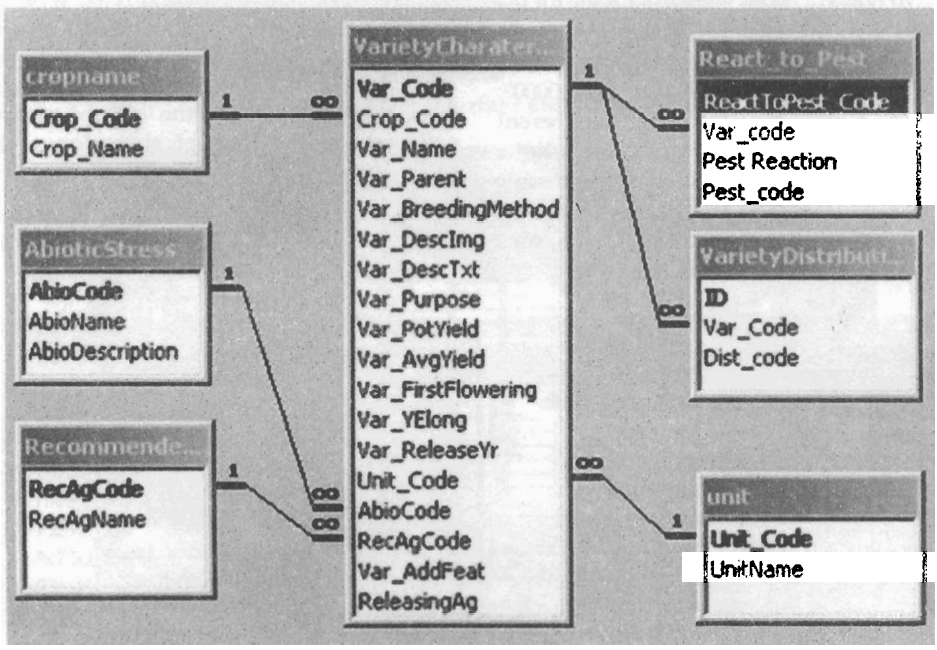
cropname
Crop_Code
Crop_Name
AbioticStress
AbioCode
AbioName
AbioDescription
Recommende...
RecAgCode
RecAgName
VarietyCharater...
Var_Code
Crop_Code
Var_Name
Var_Parent
Var_BreedingMethod
Var_DescImg
Var_DescTxt
Var_Purpose
Var_PotYield
Var_AvgYield
Var_FirstFlowering
Var_YElong
Var_ReleaseYr
Unit_Code
AbioCode
RecAgCode
Var_AddFeat
ReleasingAg
React_to_Pest
ReactToPest_Code
Var_code
Pest Reaction
Pest_code
VarietyDistributi...
ID
Var_Code
Dist_code
unit
Unit_Code
UnitName

Fig. 4

The records in a dimension table establish one-to-many relationship with the fact table. For example, there may be a number of pests to a crop or number of by products for a crop. The dimension table contains attributes associated with the dimension entry; these attributes are rich and user-oriented textual details.

Hierarchies

The data in a dimension is usually hierarchical in nature. Hierarchies are determined by the user needs to group and summarize data into usable information. For example, a time dimension often contains the hierarchy elements: all time, year, month day or year quarter, week for the production and trade statistics. A dimension may contain multiple hierarchies. A time dimension often contains agricultural, calendar or financial year hierarchies. OLAP tools depend on hierarchies to categorize data. Analysis services will create by default an entry for a hierarchy used in a data mart.

To deal various queries, proper coding scheme need to be employed. In the INARIS project, two types of codes were used. One for the internal to the use by the participating centers and another as a unified coding scheme for all the centers. Eight digit code was followed "00000000". From left to right, first two digits represent plantation crop sector, third and fourth digits represents crops, fifth and sixth digits represents various subjects and last two digits represents items. Flowing are the examples of codes for crops shown in Table1.

Crop Dimension

CPCRI Code	Value
14010000	Arecanut
14020000	Cashew
14030000	Cocoa
14040000	Coconut
14050000	Coffee
14060000	Oil palm
14070000	Palmyrah
14080000	Rubber
14090000	Tea

Queries – illustration

User interface facilitates retrieval of data from the data mart. Two types of interfaces are provided in the project viz., one for report generation and the second for displaying output on the screen. User should select options/ parameters for generating the query. The query generating process is devised as follows:

As the first step, the crop is to be selected from the list provided on the screen. On selecting, the products for which data is available would be displayed as list. Next period (year wise, month wise, date wise or all dates) for which information sought is to be entered. Finally, market is to be selected. Based on the given parameters query is to be generated. OK button is to be pressed for getting the report. The Fig 6. shows input selection screen and model report generated.

For obtaining out put on the screen, the input giving process is similar to the above. Record navigation facility is also made available.

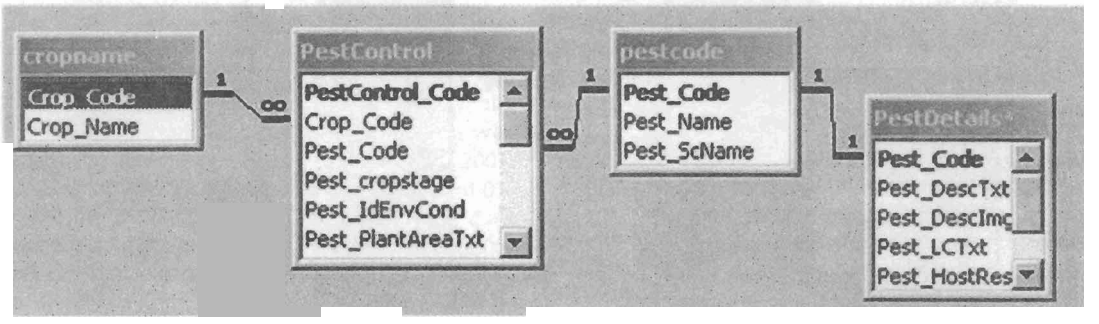

Fig. 5

Database Tuning

The data mart is designed with a star or snowflake schema in order to obtain the multidimensional access that is so critical to data analysis. Number of indexes and summary tables were created in order to avoid I/O intensive table scans against large fact tables. The indexes and summary tables proliferate in order to optimize performance for the known queries and aggregation that the users perform. Database tuning is a never ending cycle for providing better performance.

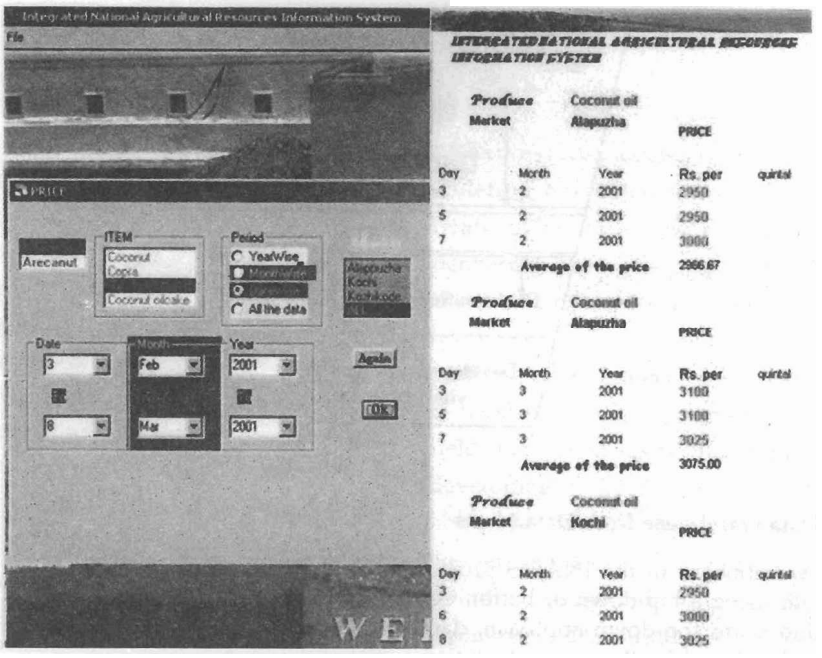

Fig. 6. User interface and Report

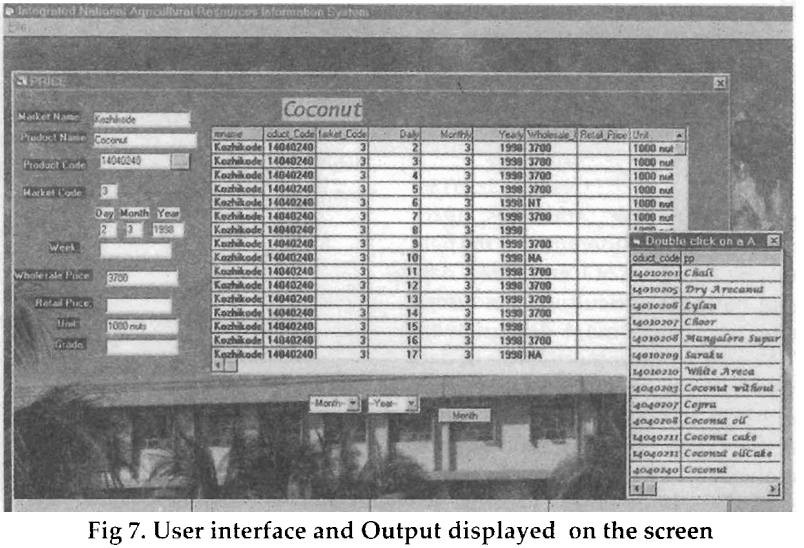

Fig 7. User interface and Output displayed on the screen

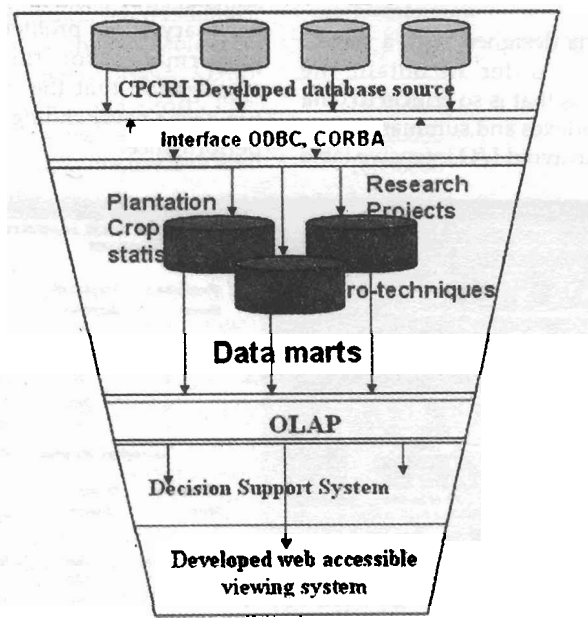

Fig. 8

Building a Data Warehouse from Data Marts

Data warehouses in the INARIS project would be built using a top-down or bottom-up approach. Under the Top-down approach, data warehouse is developed for the entire plantation crops, containing data from multiple, heterogeneous, operational sources. In the bottom-up approach, data marts for agro-climatic zones, or specific crop areas, are developed and then joined to provide the data for the entire plantation crops.

Summary

Three data marts viz., statistics, agro-techniques and research were identified in the field of plantation crops as a part of Centralized data warehousing of national agricultural resources. Different aspects of data mart design were delineated. The detail and summary levels of data in the data mart were documented. Developing data marts will help to evolve decision support tools in respective area of specialization without much cost involved. Finally the data warehouse would be built-up by joining the data marts.

REFERENCE

Barry Devlin 1997. "Data Warehouse – from Architecture to Implementation", Addison Wesley Longman

Christopher Adamson and Michael Venerable, 1998, Data Warehouse Design Solutions, John wiley & sons, 544pp.

Erik Thomson, 1997, OLAP Solutions: Building Multidimensional Information Systems, John wiley & sons, 450pp.

Muralidharan K, Ravi Kumar N, Vijayakumar K, Palaniswamy C, Dhanapal R, Sairam C. V, Jose C. T, Acharya C, Madhavan K and Arulraj S, 2001, "Plantation crops resources information system requirement analysis", Central Plantation Crops Research Institute

Olap Train and Reed Jacobson, 2000, Microsoft SQL Server 2000 Analysis Services Step by Step, Microsoft Press, 375pp.

Ralph Kimball, Laura Reeves, Margy Ross and Warren Thornthwaite, 1998, The Data warehouse lifecycle toolkit Expect methods for Designing, Developing and Deploying data warehouses: John Wiley & Sons, 800pp.

* * *