MULTIDIMENSIONAL SCHEMA FOR AGRICULTURAL DATA WAREHOUSE

Aditya Kumar Gupta[1], Bireshwar Dass Mazumdar[2]

[1] *Research Scholar, Department of Computer Science, Sai Nath University, Ranchi, India,*
[2] *Assistant Professor, Department of Computer Science, School of Management Sciences, Varanasi, India,*
aditya.guptas@gmail.com, bireshwardm@gmail.com

Abstract

Agriculture is one of the important issues for a nation's economy and it required technical breakthroughs in this century. With today's computerized world, the agricultural data processing is an increasing need for formers and decision makers. Agricultural data is diversified, complex and non-standard. Developing a data warehouse for agricultural is a key challenge for researchers. The objective of this work is to design a data warehouse about crops and their requirements. The proposed data warehouse may be extended to Decision Support System (DSS) combine with data mining techniques. We proposed a multidimensional data warehouse for agriculture that provides solutions for farmers and gives response of their ad-hoc quires. This multidimensional schema further promotes star schema and snowflake schema that are commonly used to design data warehouses. In our manuscript normalization is applied to store the data in to star schema and duplicate values are removed, so that space and time complexities could be minimized.

Index Terms: *Agriculture, Data Warehouse, Multidimensional Schema, Dimensional Modeling, OLAP*

----------***----------

1. INTRODUCTION

The first step in data warehouse design is to organize raw facts into suitable data model. The requirements definition of any data warehouse is completely drives the *data design* for the data warehouse. Data design consists of putting all the data structure together. The agricultural-data contains a large number of raw facts that are inter- dependent to each other. These facts required processing for converting informational data into operational data. The raw data is first classified in different attributes and then placed in a logical data model. This includes many data warehouse functions such as cleaning, summarizing and reformatting of attributes. An efficient data warehouse must be emphasis on information retrieval techniques. Here efficiency concerns both algorithms and data structured used. Logical data design includes determination of the various data elements that are needed for a particular subject. It also establishes the relationships among various data structures used for storage the data.

In our work the multidimensional modeling concepts are used to design agricultural data warehouse. It is a logical design technique to give structure to the agricultural dimensions and metrics, which are analyzed as important attributes for agriculture. We attempt to explore eight types of land, depends on combinatory percentage of various fertilizers. We also classify crops in twelve major classes and propose a classifier for weather, based on months of the calendar. These attributes are well placed in the model that systematically evaluates the performance of crops for a particular land, weather and environment domain.

2. LITERATURE REVIEW

In data warehouse literature multidimensional cube and their correspondent star and snowflake schema are used to design subject oriented databases. Dimensional modeling is the most suitable approach to design a data warehouse for the purpose of predictive data mining. According to [Kor99], the objectives of dimensional modeling are: (i) to produce database structures that are easy for end-users to understand and write queries against, and (ii) to maximize the efficiency of queries [7]. One way to look at the multidimensional data model is to view as a cube, the caveat here is that, as the number of dimensional increases, the number of cube's cells increases exponentially [13]. On the other hand, the majority of multidimensional queries deal with consolidated and high level of data. Therefore the solution to building an efficient multidimensional data base is to consolidate all logical attributes. The consolidation is especially valuable since typical dimensional are hierarchical in nature. We found this approach relevant and useful for our work.

3. DATA MINING MODELS

Data mining involves many different algorithms to accomplish different task. The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined. Data mining techniques can be used to make three kinds of models correspondent to three types of task:

descriptive model, profiling model, and predictive model. Hypothesis testing often produces descriptive models. On the other hand, both profiling model and predictive model have a goal in mind when a model is being built.

3.1 Descriptive Models

Descriptive model describe what is in the data. The output in one or more charts or numbers or graphics that explain what is going on. A descriptive model identifies patterns or relationships in data; it serves as a way to explore the properties of data examined; not to predict the new properties. Clustering, summarization, association rules, and sequence discovery are usually viewed as descriptive in nature.

3.2 Profiling Models

Profiling models are often based on demographic variables. In profiling models, the target is from the same time frame as the input. Profiling model has serious limitations. One is the inability to distinguish cause and effect, as the profiling is based on familiar demographic variables. It needs not to involve any sophisticated data analysis. Surveys are the common method of building the profiles. The distinction between profiling and prediction is that profiling has implications for the modeling methodology, especially the treatment of time in the creation of the model set.

3.3 Predictive Models

A predictive model makes a prediction about values of data using known results found from different data sources. Prediction means finding pattern in data from one period that are capable of explaining outcome in a later period. Predictive modeling may be made based on the use of other historical data, and it predicts what is likely to happen in the future. Building a predictive model requires separation in time between the model inputs and the model output, the thing to be predicted. Predictive model data mining tasks includes classification, regression, time series analysis, and prediction. In our study we use predictive approach to predict performance of crops in a particular time and particular type of soil and fertilizer set.

3.4 DIMENSIONAL MODELING

Dimensional modeling is a different way to view and interrogate data in database. This view may be used in a DSS in conjunction with data mining tasks. Decision support applications often require that information obtained along many dimension. For example, a farmer may want to predict production of any crop, in a specific geographic region, in particular month for seeding, and by crop type. This query requires three dimensions: Crop-set, Seeding Month, Land type. Each dimension is a collection of logically related attributes and is viewed as an axis for modeling the data. Within each dimension, these entities form levels, on which various DSS questions may be asked. The specific data stored is called facts and usually numeric data. Facts consist of measures and context data. The measures are the numeric attributes about the facts that are queried. DSS queries may access the facts from many different dimensions and levels. The levels in each dimension facilitate the retrieval of facts from different dimensions. The data warehouse should have summarized collection of data attributes that makes information retrieval more efficient.

4. MULTIDIMENSIONAL SCHEMA FOR AGRICULTURAL DATA WAREHOUSE

Dimensional models represent data with a "cube" structure [Kim96-1], making more compatible logical data representation with OLAP data management. The data can be queried directly in any combination of dimensions, by passing complex database queries. Multidimensional models take advantage of inherent relationships in data to populate data in multidimensional matrices called data cubes. The query performance in these multidimensional cubes is much better in comparison to relational data model. The response time of the multidimensional query depends on how many cells have to be added at each dimension. In figure 4.1 we show a three dimensional data cube, that organizes agricultural attributes crop-set, their seeding month and land type. However, a data hypercube could be produced by including additional dimensional, correspond to attributes such as irrigation levels and seed quality.

In our schema "12 months" (Jan to Dec) are the main classifier of weather and placed along x- axis. The attribute "crop-sets" are placed along y-axis and "land-type", we have placed along z-axis. Let m is the number of calendar months placed along x-axis, c is the number of crop-sets at y-axis and l is the number of land type placed at z –axis . Then total number of values in cube is given by $Vn = m \times c \times l$ i.e. $12 \times 12 \times 8 = 1152$. The detail description of cube is given in following sections.

4.1 Calendar for Seeding Crops

Every crop needs particular and specific weather for the best results. In our multidimensional schema 12 months calendar is used as a main classifier of weather. We have assumed that maximum production of any crop is possible when it is seeded in suitable month of the calendar. The weather report for every month of calendar is based on some factors, such as cold, rain, humidity and heat. Forecasting about weather is dependent to all these factors, table 4.1 shows predicted percentage of cold, rain, humidity and rain in large part on India in corresponding to months of the calendar.

4.2 Crop-Sets

There are hundreds of crops whose success level depends on some attributes such as start time, ending time, requirements of fertilizers and type of seeds. We identifying 12 major crop

sets correspond to their seeding months. These crop set consists number of crops that require similar weather and environmental needs. As displays in table 4.2, similar types of crops are put together in a crop-set and identified by the name of major crop belongs to it

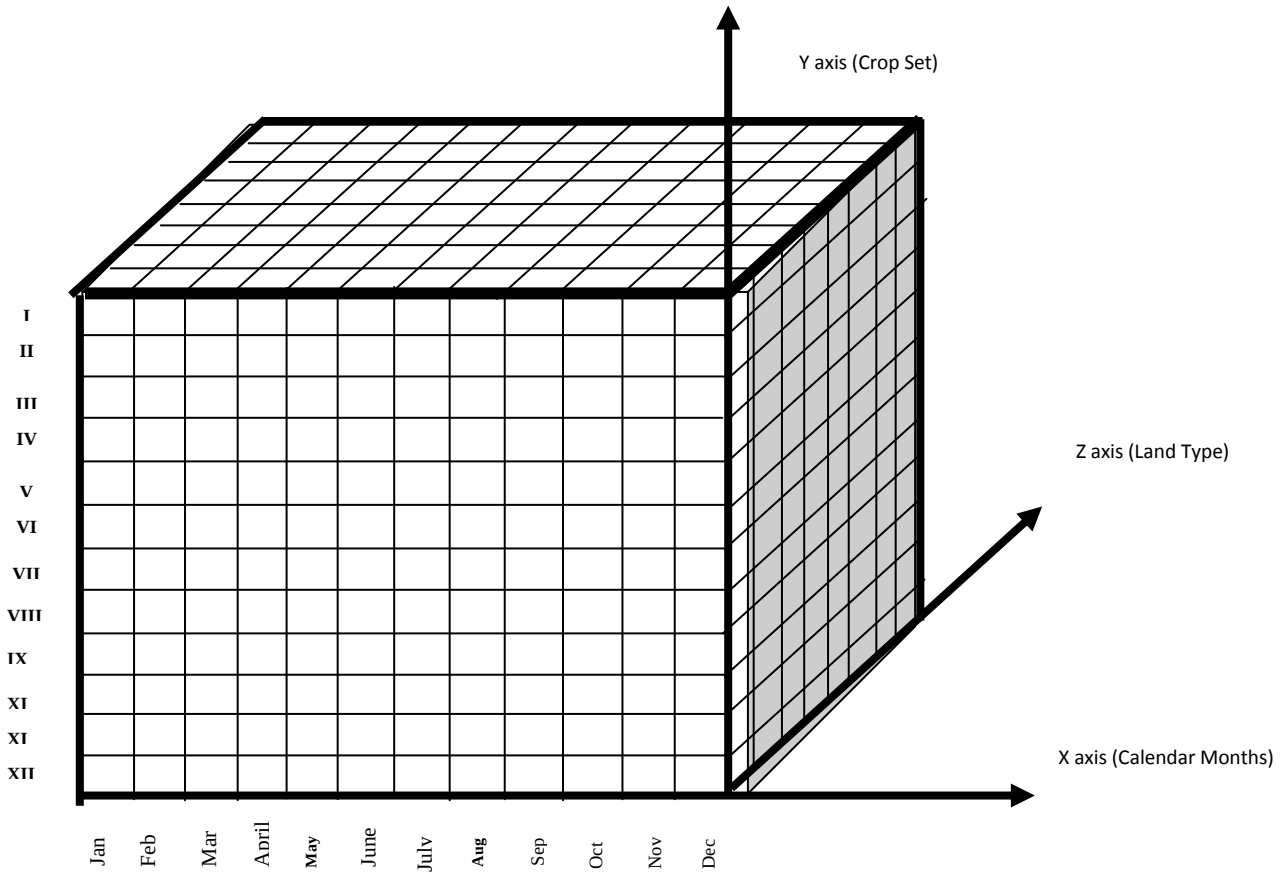

Fig-4.1: Multidimensional Cube for Agricultural Attributes

Table-4.1

	Month	% Cold	% Rain	% Humidity	% Heat
1	January	*60*	*15*	*20*	*5*
2	February	*60*	*15*	*15*	*10*
3	March	*50*	*5*	*10*	*35*
4	April	*25*	*5*	*10*	*60*
5	May	*15*	*10*	*5*	*70*
6	June	*5*	*30*	*5*	*60*
7	July	*5*	*55*	*25*	*15*
8	August	*10*	*60*	*25*	*5*
9	September	*10*	*60*	*25*	*5*
10	October	*15*	*50*	*25*	*10*
11	November	*35*	*25*	*30*	*10*
12	December	*50*	*15*	*30*	*5*

Table-4.2

Crop Set	Crop	Starting Month	Ending Month
I	Wheat	December	March
II	Potato	November	January
III	Mustered	October	March
IV	Oil-seed	September	February
V	Coffee	August	November
VI	Pulses	July	October
VII	Rice	June	October
VIII	Peppermint	May	July
IX	Sunflower	April	June
X	Cash-crops	March	June
XI	Coconut	February	August
XII	Sugarcane	January	September

Table-4.3

S No.	Land Type	Percentage of Fertilizers in Fertilizer Set								Weight of Fertilizer	Total Weight
		U	P	Fe	N	Z	S	Ca	K	(P_f)	(F_w)
1	L1	20	-	-	50	-	-	-	10	U2 + N5 + K1	8
2	L2	20	30	10	10	10	-	-	-	U2 + P3 + Fe1 + N1 + Z1	8
3	L3	20		10	20	10	20	-	-	U2 + Fe1 + N2 + Z1 + S2	8
4	L4	-	30	-	20	-	-	10	20	P3 + N2 + Ca1 + K2	8
5	L5	10	-	-	30	-	10	20	10	U1 + N3 + S1+Ca2 + K1	8
6	L6	40	10	-	10	-	20	-	-	U4 + P1 + N1 + S2	8
7	L7	-	-	40	10	20	-	10	-	I4 + N1 + Z2 + Ca1	8
8	L8	10	10	10	20	-	30	-	-	U1 + P1 + Fe1 + N2 + S3	8

4.3 Land Type

The impact of fertilizers on production of any crop is significantly measured. In a particular type of land, there are various fertilizers and minerals. We can identify many types of land based on combinatory ratio of various fertilizers in soil. In our study we identified 8 major fertilizers that are naturally available in soil. However, we can generalize the fertilizer-set, which may change with geographical regions.

1. Urea (U)
2. Phosphorus (P)
3. Iron (Fe)
4. Nitrogen (N)
5. Zinc(Z)
6. Sulfur (S)
7. Calcium (Ca)
8. Potassium (K)

Further on the basis of the ratio of above fertilizers present in soil, we have identify 8 types of land for our study; table 4.3 displays land type depends on percentage of fertilizers available in a soil. In table 3 we hypothetically demonstrate 8 different type of land based on above fertilizer, however for the generalization purpose we can identify n types of land. In the model, further we assign 8-unit weight to each type of soil; each unit of weight corresponds to 10 percent availability of a particular fertilizer in any fertilizer- set. For example land type L1 contains 50 percent of Nitrogen, 20 percent of Urea and 10 percent of Potassium (table 4.3) and, rest 20 percent includes other elements in soil. We are not including such 20 percent in the study, assuming it is non-deterministic minerals, as soil has many other impurities.

4.4 Irrigation

Irrigation is an important factor that majorly affects production of any crop. In our model irrigation factor is consider as a dimensional of the cube (not shown in figure). We have taken three possible states of irrigation: high irrigation, medium irrigation and low irrigation. Accordingly we assign some constant values to each of these levels of irrigation, for the computational purpose. The irrigation factor is also described in correspondent star schema and further extended into snowflake schema.

4.5 Seed Quality

With the development of biotechnology, researcher produces high quality of seeds; the production rate of any crop is certainly dependent on quality of seed. As we have done for irrigation, quality of seeds is also classified in three levels, high quality, medium quality and low quality. Further we have assigned some constant values to each of these levels of seed quality in our model. Like irrigation seed quality is also consider as a dimension of the cube (not shown in figure). The Seed quality factor is also described in correspondent star schema and then, in snowflake schema.

5. ARRENGING DIMENSIONS IN SCHEMA

Specialized schemas have been developed to portray multidimensional data. These include star schema and snowflake schema. A star schema shows data as a collection of two types: facts and dimensions. The star schema view can be obtained via a relation system where each dimension is a table and the facts are stored in fact table. Unlike relation

schema which is flat, a star schema is a graphical view of data. At the center of star, the data being examined, the facts are shown in the fact table. On the outside of the facts, each dimension is shown separately in dimension table. Descriptive information about the dimensions is stored in dimensions tables, which tend to be smaller. While the actual data is being accessed, are stored in fact table and thus tend to be quite large. To ensure efficiency of access, facts may be stored for all possible aggregation levels. The fact data would then be extended to include a level indicator. Snowflake schema is an extension of star schema; it facilitates more complex data views. In snowflake schema, the aggregation hierarchy is described explicitly in the schema itself. The snowflake schema can be understood as partially normalized version of corresponding star schema.

5.1 Star Schema Arrangement

The star schema proposed for multidimensional agricultural data warehouse is starts with a central fact table that corresponds to facts about agricultural technology. Data in fact table can be viewed as a regular relation with an attribute for each fact, to be stored and the key being the values, for each dimension. Each row of the central table contains some combination of keys that makes a fact unique. These keys are called dimensions.

In our model we use simplest star schema that has one fact table and three dimensional tables corresponds to weather, crop-sets and land type, two other dimensions, irrigation and seed quality is also included in the model for computation of decision coefficient .The central fact table also has other columns that typically contain information specific to each row, such as irrigation level, sequence of crop-set and seeding month and, quality of seed. For the purpose of computation we have assigned numeric values to all these information. Figure 5.1 explains the star schema arrangements of multidimensional cube given in previous section. The fact table contains major attributes such as irrigation level, crop-set, seeding month of the crop, soil type and quality of seed. A computational model is required to produce the value of decision coefficient. Decision coefficient is the success ratio of any crop based on above attributes describe in the cube. We can compute the value of decision coefficient for any crop by providing values of these attributes as input, to the computational model. There are four basic approaches to the storage of data in dimensional table [PB99], which are flattened approach, normalized approach expanded approach and levelized approach. Each dimension table can be stored in one of these manners. We have adapted normalized approach to store the data and to remove duplicate values. In this approach a table exists for each level in each dimension. Each table has one tuple for every occurrence at that level. As with normalization each lower level dimension table has a foreign key pointing to next higher level. Figure 5.2 provides queries to implement star schema for agricultural data warehouse using normalized approach.

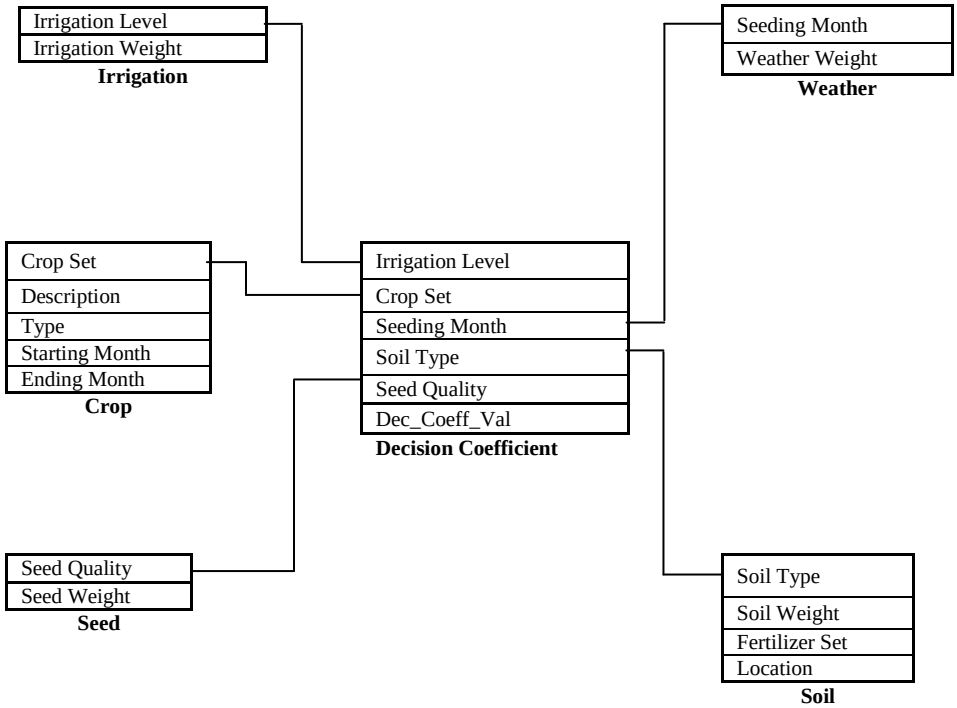

Fig- 5.1: Star Schema for Agricultural Data Warehouse

Decision Coefficient (**Irrigation Level, Crop Set, Seeding Month, Soil Type, Seed Quality,** Dec_Coeff_Val)

Irrigation (**Irrigation Level,** Irrigation Weight)

Irrigation Points (**Irrigation Weight**, Low, Mid, High)

Crop (**Crop Set**, Description, Type, Starting Month, Ending Month)

Weather (**Seeding Month, Weather Weight**)

Weather Report (**Weather Weight**, Cold, Rain, Humidity, Heat)

Soil (**Soil Type, Soil Weight, Fertilizer Set**, Location)

Fertilizers Availability (**Soil weight,** PerAvlF1, PerAvlF2, PerAvlF3, PerAvlF4, PerAvlF5, PerAvlF6, PerAvlF7, PerAvlF8, PerImpurity)

Fertilizer (**Fertilizer Set**, f1, f2, f2, f4, f5, f6, f7, f8)

Land (**Location**, State, District, Village)

Seed (**Seed Quality**, Seed Weight)

Seed Points (**Seed Weight**, Low, Mid, High)

Fig- 5.2: Implementation Details of Star Schema for Agricultural Data Warehouse

5.2 Snowflake Schema Arrangement

A snowflake schema is a logical arrangement of tables in a multidimensional database, and so the entity relationship diagram resembles as a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. In snowflake schema, dimensions are normalized into multiple related tables, where in star schema dimensions are normalized with each dimension represented by a single table. A star schema stores all attributes for a dimension into one de-normalized (flattened) table. This requires more disk space than a snowflake schema that is eventually more normalized approach. Snowflake schema normalizes the dimension by moving attributes with few distinct values into separate dimension tables that relates to core dimension table by using foreign keys. A complex snowflake shape emerges when the dimensions of a star schema are elaborated, having multiple levels of relationships, i.e. the child tables have multiple parent tables. In fig-5.3 we have elaborated weather dimension into weather report, irrigation into irrigation points, and seed into seed points. The soil dimensions is extended into three relationships i.e. land, fertilizer availability and fertilizer weight. Snowflake is used to improve query performance against low cardinality attributes. Business intelligence applications that use OLAP architecture may perform better, when the snowflake schema is used for the purpose of data warehousing.

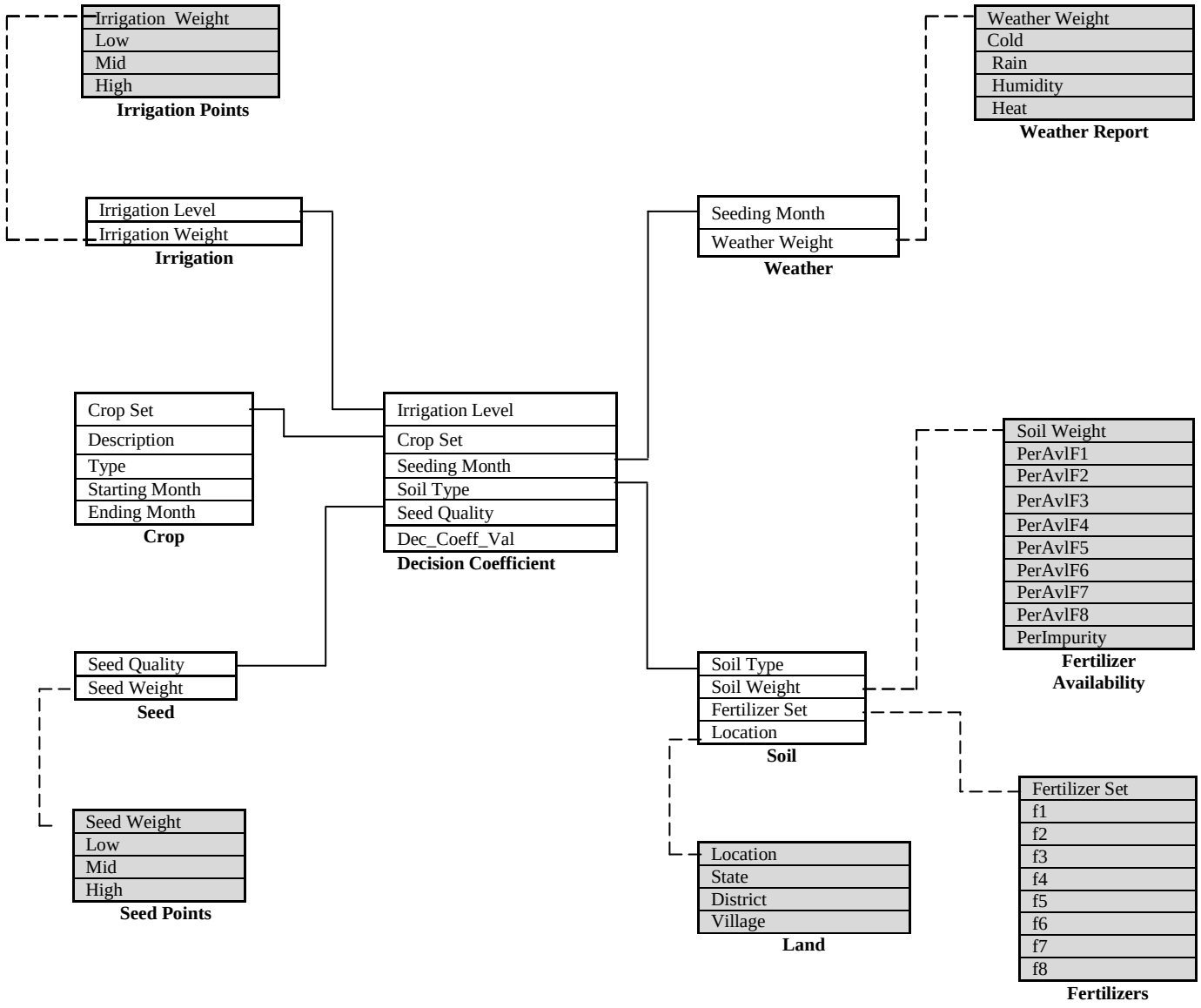

Fig- 5.3: Extension of Star Schema into Snowflake Schema for Agricultural Data Warehouse

6. OLAP FOR AGRO DATA WAREHOUSE

Online analytical processing system are targeted to provide more complex query results than online transaction processing system (OLTP) or relational database system. The OLAP system provides extra analysis of data as well as more imprecise nature of queries. This basically differentiates an OLAP application from traditional database or OLTP application. The multidimensional view of data is fundamental to OLAP applications. The complex nature of OLAP application requires a multidimensional view of data, and the type of data accessed is often a data warehouse. OLAP tools can be classified as relational OLAP or multidimensional OLAP (MOLAP). The data warehouse systems often contain multiple OLAP cubes, and the power of OLAP arises from practice of sharing dimensional across different cubes. In MOLAP system, data are modeled, viewed, and physically stored in a multidimensional schema. MOLAP tools are implemented by specialized DBMS and software systems capable of supporting the multidimensional data. With MOLAP the cube view of data is stored as an n- dimensional array, this approach require extremely high storage space and then indices may be used to speed up processing of data. The parallel computing can be used to overcome this limitation, and then such cube requires less processing time.

In our agricultural data warehouse, we purposefully use multi dimensional cube to store the data. Here every cube is particularly valuable because they make possible to drill-down and roll-up operations with other cubes. To assist roll-up and drill down operations aggregation is used, and the values that

are pre-computed are stored in the model. In section 5.1 we have mention the requirement of a computational model, which computes value of decision coefficient given in star schema. A simple query may look at a single cell within the cube; however queries may stated in multidimensional terms. There are several types of OLAP operations that also support to response more complex queries.

Slice: This operation performed by selecting values on one dimension. Such as, values of decision coefficient for all crops in a particular month. The SQL statement may be given as.

SELECT ALL (Dec_Coeff_Val) FROM Decision_Coefficient WHERE Seeding_Month = 'July';

Dice: This can be performed by slice on one dimension and then rotating the cube to select on second dimension i.e.

SELECT Dec_Coeff_Val FROM Decision_Coefficient WHERE Seeding_Month IN('Jan', 'Feb', 'March' , 'April', 'May', 'June' , 'July', 'Aug', 'Sep', 'Oct', 'Nov' , 'Dec');

Roll-up: This allows asking queries that moves up an aggregation hierarchy. Instead of looking one fact we look at all the facts.
SELECT ALL (Dec_Coeff_Val) FROM Decision_Coefficient WHERE Seeding_Month IN('Jan', 'Feb', 'March' , 'April', 'May', 'June' , 'July', 'Aug', 'Sep', 'Oct', 'Nov' , 'Dec');

Drill-down: This operation allows user to navigate lower in the aggregation hierarchy. In this user get more specified results. Such as, value of decision coefficient for a particular crop in a particular month.

SELECT Dec_Coeff_Val FROM Decision_Coefficient WHERE Seeding_Month = 'July';

The other OLAP operations may be used to support more analytical results of agricultural facts. We can combine each of above OLAP operations to response more analytical queries.

CONCLUSIONS

In this manuscript we have applied concepts of dimensional modeling. We have identified major agricultural attributes and place these attributes in the correspondent dimensions of the cube. Further we developed a star schema for multidimensional data cube and identified the primary keys to connect each of the dimensions with central fact table of the star schema. The fact table collects information from each of the dimensions and then, the values of decision coefficient is computed and stored in the fact table, correspondent to each cell of the cube. We have also extended the star schema in to snowflake schema by applying hierarchical aggregation of each dimension. In snowflake schema, the microscopic view of factors, affecting each dimension of the cube is clearly explained. In the last section of the work we have applied OLAP operations to analyze the facts stored in the multidimensional database. The OLAP tools help farmers and decision makers to predict production of crops, to find the lack of any fertilizer available in soil or to guess the level of irrigation required, during growth of the crop.

The real strength of the paper lies in the adoption of predictive data mining techniques. Many data mining tools and algorithm may be applied to this agro database, which probably responds to more complex queries. For example a farmer might discover that a particular crop produce better results, when it is irrigated at a particular time, under a particular weather conditions. The decision maker may found some pattern in crop sequence for better results such as farmers of Chhattisgarh, Orissa and parts of Bihar could found that rice-wheat system is more suitable than rice-pulses system. Similarly other complex and specific queries can respond using data mining. This paper strongly advocates the use of predictive data mining techniques to retrieve more sophisticated information from agricultural data warehouse. We will find our work meaningful if an agent based software could be developed for the purpose of warehousing and mining of agricultural attributes.

REFERENCES

[1] Aditya K Gupta, M.H. Khan, "Clock Based Model for Cropping System: A Frame work for Agricultural Data Warehouse". National Conference, CSI, Delhi Chapter. Feb 2007.
[2] Agricultural Research Data Book (2002), IASRI, New Delhi.
[3] Agricultural Statistics at a Glance (2001), Directorate of Economics and Statistics, Department of Agriculture and Cooperation, Ministry of Agriculture, Govt. of India.
[4] Alex Berson and Stephen J. Smith , Data Warehousing , Data Mining , and OLAP. Tata McGraw-Hill 2004.
[5] C. Adamson, M. Venerable. Data Warehouse Design Solutions. J. Wiley & Sons, Inc.1998.
[6] E. Thomsen. OLAP Solutions. Building Multidimensional Information. John Wiley & Sons, Inc., 1997.
[7] Kimball, R., Reeves, L., Ross, M. and Thornthwaite, W. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, developing, and Deploying Data Warehouses. John Wiley & Sons, 1998
[8] L. Silverston, W. H. Inmon, K. Graziano. The Data Model Resource Book. J. Wiley & Sons, Inc. 1997
[9] M. Blaschka. FIESTA: A Framework for Schema Evolution in Multidimensional Information Systems. Proc. of 6th. CAISE Doctoral Consortium, 1999, Heidelberg, Germany.
[10] M. Golfarelli, Stefano Rizzi. A Methodological Framework for Data Warehouse Design. DOLAP 1998.
[11] Margaret H. Dunhan, Data Mining Introdouctry & Advanced Topics. Pearson Education 2003.

[12] Michale J.A. Berry Gorden S. Linoff, Data Mining Techniques For Marketing, Sales and CRM. Wiley Publishing Inc. 2004.
[13] Paulraj Ponniah , Data warehousing Fundamentals. J. Wiley & Sons, 2005.
[14] R. Agrawal, A. Gupta, S. Sarawagi. Modeling Multidimensional Databases. ICDE 1997
[15] R. Kimball. The Data Warehouse Lifecycle Toolkit. J. Wiley & Sons, Inc. 1998.
[16] Ramez Elmasri and Shamkant B. Navathe. Fundamentals of Database System, 3rd edition Addison Weseley, 2000.
[17] S. Chaudhuri, U. Dayal. An overview of Data Warehousing and OLAP Technology. SIGMOD Record 26(1). 1997.
[18] W. H. Inmon. Building the Operational Data Store. John Wiley & Sons Inc., 1996.

BIOGRAPHIES:

Aditya Kumar Gupta has expert hands over DBMS and Data Warehouse technologies. After completing MCA in 2002; presently he is pursuing Ph.D. in computer science. His objective of research is to design data warehouse for Indian agriculture. He has more than eight years of rich experience in academia and industry. He has presented his work in national conference held at New Delhi and that was organized by department of science and technology government of India, and Computer society of India.

Bireshwar Dass Mazumdar has earned his Ph.D. from IIT - BHU Varanasi. He has completed MCA in 2004 from UP Technical University. Presently he is working as Assistant Professor in, School of Management Sciences, Varanasi. Dr. Mazumdar has seven years of core experience in teaching in reputed academic institutions. He has expertise knowledge of subjects like Multi Agent Systems, Data Warehousing, Data mining, Artificial Intelligence, and Software Quality Engineering. He has a pool of Research and Publications in reputed journals. He has seven international publications and three national publications to his credit.