Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Designing multidimensional model using relational schema

Swati Hira *, Anita Bai, P. S. Deshpande

Department of Computer Science & Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form
16 May 2016
Accepted 18 May 2016

Keywords:
Data warehouse
Multidimensional model
Relational schema
Automation
OLAP

ABSTRACT

Manually creation of multidimensional schemas for data warehouses involves complex mapping which takes a long time and increases risk of failure, if number of relations in the schema is large in numbers. Multidimensional modelling involves identifying dimensions, hierarchy in the dimensions and measures in the data. In this paper we have proposed novel way of multidimensional modelling by analysing data characteristics. Our method identifies probable dimensions and hierarchies based on data characteristics and provide multiple multidimensional models. User can choose or modify these models depending on his or her requirements. We have done exhaustive experimentation on various relational schemas and find that our method saves lot of man hours in constructing multidimensional models. This method is illustrated using the sales data and various multidimensional queries. It helps to create a foundation for business intelligence tools and generate multidimensional model views for data warehouse.

1. Introduction

Multidimensional data is the data which records information of the facts which are related with various entities. These entities are commonly called as dimensions. For example when customer purchase items from some retailer in retailer chain then this information is recorded as customer X has purchased item Y at retailer Z on date W with an amount A. Here the purchase event relates information of customer X, retailer Z, item Y and time W and is having value A. Here customer, retailer, item and time forms dimensions and value is called measure. The measure in this case can be value in terms of rupees or quantity. Most of the events in life are multidimensional in nature. For example if some medical representative visits to doctor and explain the advantage of medicine then the event can be recorded on dimensions like time, doctor, medicine and sales person. It helps in management's decision making, valuable in today's world. In recent years, development in the domain of database and information management has increased. The importance of data warehouse depends on its high efficiency for various database applications such as ERP (Bollen, 2016), banking, insurance and retail etc. It is an approach to provide an integrated, uniform source of data for use in data analysis and business decision making. The most popular data model for a data warehouse is a multidimensional model (Peralta, 2003).

1.2. Dimensions

Dimensions are entities or axis which is used for analysing data. For example sales manager of the retailer chain may be interested in finding out top three locations where sale of tomato sauce is maximum in Jan-2009. This information may be useful in deciding location for advertising campaign of tomato sauce. Here the sales manager is analysing information on region, time and product axis. So the dimensions are region, product and time. Selecting proper dimensions in the data is very crucial in multidimensional analysis.

1.2. Multidimensional data analysis

Dimensions are entities or axis using which data is analysed. The dimensions are chosen such that business queries should be satisfied. Dimensions should be carefully chosen because same data, many dimensions can be obtained. Due to multidimensional nature of data and different characteristics of data, the analysis of data requires different treatment. This analysis is used to exploit multidimensional nature and used to provide mathematical model of the business process. All types of data exhibits certain characteristics. Main characteristics which are used for the business queries are hierarchy in the data, sequential

* Corresponding Author.
Email Address: dt11cse076@cse.cnit.ac.in (S. Hira)

properties in data and dependant relationships in the data. The multidimensional tools are designed in such as way so that decision making information can be obtained by exploiting these data characteristics. Most of the business intelligence tools operate on multidimensional data model for satisfying business queries. Building multidimensional model for data having large number of relations is complex task and involves lot of manual data analysis.

In this paper, we have suggested method to automate multidimensional model building. Our main contribution is to provide framework to extract relationships in the data which is required for multidimensional data model and suggesting to the user probable multidimensional models which helps to expand the result extraction capabilities using queries. This approach involves three steps. First, a common schema is constructed from the given relational schema. In the next step, multidimensional components as dimensions, measures and hierarchies are identified from this common schema. At the end, on the basis of multidimensional components, multidimensional model schema is constructed to provide various suitable multidimensional data model for online analytical processing (OLAP).

Section 2 discusses brief overview of related work in data warehouse design and techniques. Section 3 presents and illustrates an automated algorithm to construct multidimensional model views for a data warehouse from a relational schema. Section 4 describes results and discusses about various business queries solved by multidimensional models. Section 5 presents the conclusion and future work.

2. Related works

In data warehousing, the transformation of relational model to other architectures is an essential task. The identification of the model is based on the requirements, access tools and the team preferences. This could be identified as an important research area as most of the business organization right now looking for business intelligence or decision support system. Specially, the old day companies don’t store their data in traditional database system. When these companies look for an intelligent system they need to build the data warehouse from the semi-structured data. Due to this reason a lot of research works are expected in this area.

Data warehousing takes significant time and resources to implement the solution for complex problems. However, automating requirement management is not an easy task since it requires formalizing the end-user requirements (Husemann et al., 2000). On the other hand, some approaches have been discovered to automate multidimensional model design process from other sources. This design process helps to generate multidimensional schema such as the relational model, OLAP, ER-model, XML web data, requirement-driven methodology etc. Further, background of some works are describes in this section.

Multidimensional schemas are derived in multidimensional normal form (MNF) using requirement-driven methodology by satisfying some set of constraints. This approach generates a logical schema from ER diagrams and produces multidimensional schemas in terms of multidimensional arrays. A hybrid approach is also use to generate logical schemas from SER (Structured Entity Relationship) which visualizes existence dependencies between objects.

XML is used to create user-defined document types. It provides a robust, non-proprietary, persistent, and standard file format for the data storage and transmission. Data could come from different heterogeneous sources such as (Tseng and Chen, 2005). Some semi-automatic approaches are also developed to generate DW logical schema from XML schemas for example paper (Vrdoljak et al., 2003) proposed the method to integrate XML data in the data warehouse. DTD is an integrating approach used by some of the researchers (Hummer et al., 2003). In this paper presents a framework for multidimensional databases based on a logical data model.

Relational schemas are also used as sources to generate multidimensional schemas. Phipps and Davis (2002) described an approach for automatic DW conceptual schema development and evaluation. In which they uses an enterprise schema of an operational database as input for data warehouse schema design. To validate the DW schema they use the queries and also provide a guideline to expand the hierarchy level according to user requirement queries. Author (Jensen et al., 2004) presents supply-driven methodology to derive snowflake schemas and discover functional and inclusion dependencies. They applied data-mining techniques over the database instances from relational databases. (Romero and Abello, 2006) uses SQL queries and relational models to derive conceptual multidimensional schemas. This approach is fully automatic and follows a hybrid paradigm. On the other hand it does not analyse the whole data sources but those concepts closely related to the end-user requirements. Finally, they derived multidimensional schemas after applying validation process. (Giorgini et al., 2005) presented a hybrid approach to derive a conceptual multidimensional schema where both ER diagrams and relational schemas are used as inputs to describe the data sources.

From above literature we observed that limited research has been done for generation of multidimensional schemas from relational schemas. So we introduced a methodology to automate the design process of multidimensional model views (schema). In this method we work over relational schemas (i.e. at a logical level). In this approach, our main objective is to convert relational schemas to a common schema, next identify various measure, hierarchy and dimensions and finally generate all possible multidimensional model views.

3. Proposed method multidimensional schema creation

Our approach aims to automatically convert the relational schema in multidimensional schema to satisfy user requirements. This approach involve six steps for creating multidimensional model that results in multidimensional schema as a core for business processes, with relationships to other entities and their attributes to form the dimensions which describe these processes. This schema is use to generate various views of multidimensional schema for multidimensional models. Table 1 describes abbreviations used in algorithm.

Input to the algorithm is relational schema represented by table data structures. The steps of the algorithm are organized as follows:

1. Create a common table by applying join between relational tables.
2. Find attributes with numerical values and assign them under measure modelling component. Numerical values should not follow any sequence.
3. Find attributes having region information and assign them under region dimension. Region information as city, state, country and zone-wise area.
4. Find attributes with date and time entries and assign them under time dimension.
5. Identify hierarchy level among attributes of region and time dimensions, if exists. For example:
   Region: Country→State→City
   Time: Year→Month→Day
   Find other dimensions and hierarchy.

Table 1: Abbreviation used

T_i	i^{th} table of schema, i=1 to n
α_{ij}	j^{th} attribute of table i
CT	Common table
G_i	Group of attributes belong to a table T_i
D_i	i^{th} dimension of group, Each group indicate one temporary dimension (D_i)
$\alpha_{numeric}$(CT)	Numeric attribute of common table
α_{region}(CT)	Region attribute of common table
α_{date}(CT)	Date attribute of common table
α_{time}(CT)	Time attribute of common table

Algorithm

```
Initialization:
$T_i = 1\ (First\ table\ of\ schema)$
1. $CT = T_1 \bowtie T_2 \bowtie T_3 \ldots\ldots\ldots \bowtie T_n$
2. $Measure\ (M) = \alpha_{numeric}(CT)$
3. $Region\ (R) = \alpha_{region}(CT)$
4. $Time\ (T) = \alpha_{date}(CT) \vee \alpha_{time}(CT)$
5. $Region_{hierarchy}(CT)\ and\ Time_{hierarchy}(CT)$
6. Identification of other dimensions and hierarchies
   //Dimension identification//
   For each table $T_i$
       Find $G_i = \{\alpha_{i1}, \alpha_{i2}, \ldots\ldots, \alpha_{in}\}$
       such that $\{\alpha_{i1}, \alpha_{i2}, \ldots\ldots, \alpha_{in}\} \notin \{M, R, T\}$

   // Hierarchy identification among attributes of Table $T_i$//
    For i=1 to n in $G_i$
       $\alpha_{ij}C = select\ count(distinct\ \alpha_{ij})\ from\ table\ G_i$
    End for
    $If\ (\alpha_{ij} > \alpha_{ij+1})\ and$
          $(\forall\ distinct\ (\alpha_{ij}C)\ values\ \exists\ similar\ (\alpha_{ij+1}C)\ values\ )\ or$
           $(\forall\ distinct\ (\alpha_{ij+1}C)\ values\ \exists\ similar\ (\alpha_{ij}C)\ values)$
               $hierarchy = \alpha_{ij} \rightarrow \alpha_{ij+1}\ or\ \alpha_{ij+1} \rightarrow \alpha_{ij}$
   // Similarly for all attributes of table $T_i$ hierarchy will be identified, if exists//
       End for
```

Explanation:

Suppose we have two tables say T_1 (α_1, α_2, α_3) and T_2 (α_1, α_4) with same attribute name α_1 and hierarchies as

T_1: α_1→ α_2→ α_3
T_2: α_1→ α_4

It indicates that all attribute (α_1, α_2, α_3, α_4) belongs to one dimension.

From hierarchy we can see that α_1 is at higher level but the question comes when the α_4 will come at hierarchy level or not. From above algorithm count of each attribute and their relationships with other attribute will be calculated which can present any one of below hierarchy relation.

1. α_1→ α_2→ α_3
α_1→ α_4
2. α_1→ α_2→ α_3→ α_4
3. α_1→ α_4→ α_2→ α_3
4. α_1→ α_2→ α_3→ α_4

This way dimension and hierarchy between tables can be found.

4. Results and discussions

We are using sales data to describe our algorithms. Result of each steps are explain as follows: Fig. 1 represents relationship between tables as class diagram. After applying algorithm following results will be generated. This section describes the result of each step. In our figures ellipse and diamond represents attributes and measure or dimensions respectively.

Step 1: It generates the common table by applying join query which shows all attributes. Fig. 2 represents the common table.

Step 2: Here we extract measure attributes, region and time dimension attributes from common table. In our sales example, the only numeric attributes are of type integer and decimal. For example measure attributes are Sell-quantity, Buy-quantity, Sell-price, Buy-price, ID-quantity, and Transportation-cost shown in Fig. 3.

Step 3 and 4: (see Fig. 4)

Region hierarchy are as follows:

- Outlet-state→ Outlet-city,
- Owner-state→ Owner-city,

Time hierarchy are as follows:

- Year→month→day

Step 5:

Other Dimensions and attributes (see Fig. 5):

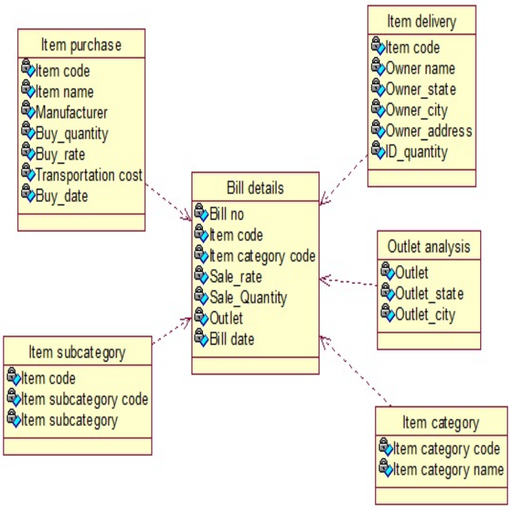

Fig. 1: Sales data relationship diagram

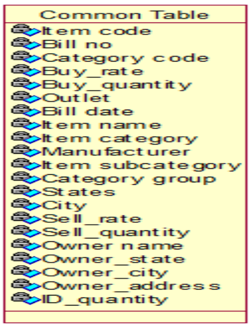
Common Table
Item code
Bill no
Category code
Buy_rate
Buy_quantity
Outlet
Bill date
Item name
Item category
Manufacturer
Item subcategory
Category group
States
City
Sell_rate
Sell_quantity
Owner name
Owner_state
Owner_city
Owner_address
ID_quantity

Fig. 2: Common table

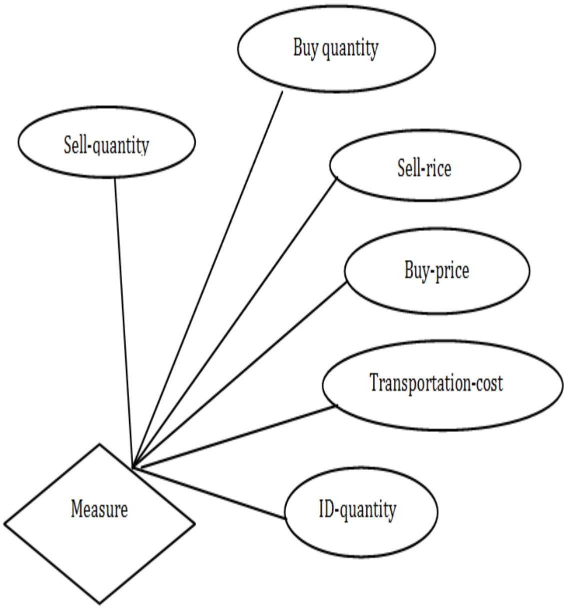

Fig. 3: Measures of sales data

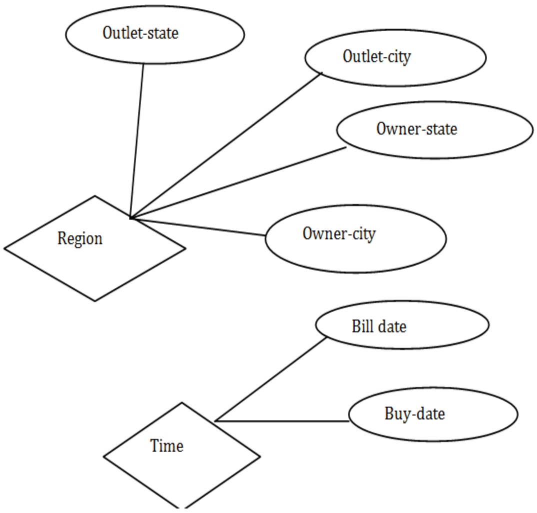

Fig. 4: Region and time dimensions

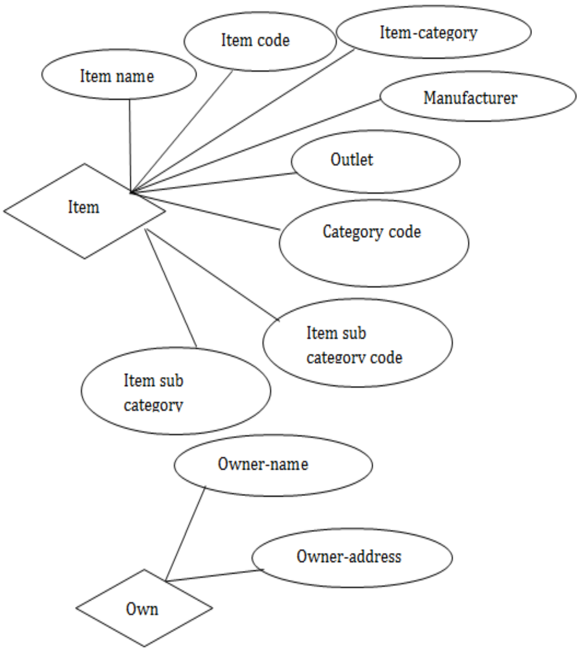

Fig. 5: Item and owner dimensions

4.1. Multidimensional models for a practical example: Sales

In this section, we present various multidimensional models using sales data. Various models can be generated on sales data, some of them as follows:

4.1.1. Model I

The hierarchies and dimensions for this model are as follows:

Dimensions and Hierarchies:

- Region→All→Outlet-state→Outlet name
- Product→All→Item-Category→Item Sub Category→ Item name
- Owner→All→Owner-state→Owner-name
- Time→All→Year→Quarter→Month

Variables (Measures):

- Sell-Quantity
- Buy- Quantity
- Sell-Price
- Buy-Price
- Transportation-cost
- ID-quantity

Queries: In this model, "Manufacturer" is ignored; therefore any query which refers to "Manufacturer" cannot be answered. The analysis is done on monthly basis. The model is able to answer various business queries, such as:

a) What will be top 3 items sold in Maharashtra for the month of October, November, and December 2010? (see Table 2)
b) Which state is performing better in Oct 2009 for an item HII_tea 500 Gms? (see Table 3)
c) What is the position of country of the company as compared to previous month and previous year?
d) What will be the locations where outlet space or outlet numbers can be increased?
e) What is optimum inventory level of each item at each outlet?

Table 2: Top 3 items sold in Maharashtra

Items	State	Quantity		
		July. 2010	Aug. 2010	Sep. 2010
HII-tea 500 Gms	Maharashtra	35710.0	31230.0	32410.0
HII-tea 1 kg	Maharashtra	30331.0	32415.0	30442.0
HII-washing Powder ½ kg	Maharashtra	25311.0	24332.0	23615.0

Table 3: HII-tea in Oct 2009

Items	State	Quantity
		Oct. 2009
HII-tea 500 Gms	Assam	35710.00
HII-tea 500 Gms	Maharashtra	30331.00
HII-tea 500 Gms	Kerala	25311.00

There, once a model is built, relevant information can be extracted according to hierarchical and sequential characteristics. Some queries are also shown in Appendix I.

4.1.2. Model II

In this model, the product hierarchy is changed by adding the Manufacturer level and replace state by city:

Dimensions and Hierarchies:

- Region→All→Outlet-city→Outlet name
- Product→All→Item-Category→Manufacturer→ Item name
- Owner→All→Owner-city→Owner-name
- Time→All→Year→Quarter→Month

Variables (Measures):

- Sell-Quantity
- Buy- Quantity
- Sell-Price
- Buy-Price
- Transportation-cost
- ID-quantity

Queries: By adding "Manufacturer" in the data hierarchy, some additional business requirements can be satisfied as follows:

a) Which are the predominant manufacturers' in Jan 2010 in Maharashtra for Washing Powder (see Table 4)?
b) In which city of Maharashtra number of deliverable owners can be increased for Parle-Chocolate 100 Gms (see Table 5)?

Table 4: Washing Powder predominant manufacturer

Rank	Description	Manufacturer	Measure	Market Share
1	Washing Powder 1/2 Kg	HLL	24118.00	5.134
2	Washing Powder 1/2 Kg	P&G	18246.00	4.502
3	Washing Powder 1 Kg	HLL	16320.00	3.910
4	Washing Powder 1 Kg	P&G	16106.00	3.823

Table 5: Parle-Chocolate deliverable owner

City	Quantity	No of deliverable owner
Mumbai	15407.00	545
Pune	13325.00	478
Nagpur	10345.00	450

Here we show the results as a sum of previous 3 years Quantity (for Quantity) and Number of deliverable owner (for owner).

c) What will be the total supply by manufacturers in category X across all outlets?

d) Which manufacturers are intermittent suppliers or are not reliable for outlet Y?

e) What is the transportation cost of outlet Y in Maharashtra?

f) Can more discounts be availed from manufacturer X?

g) What will be the impact on an outlet if a certain manufacturer is not able to supply in category X?

4.1.3. Model III

The dimensions are the same as model II; however the data hierarchy is altered:

Dimensions and Hierarchies:

- Region→All→Outlet-city→Outlet name
- Product→All→Manufacturer→Item Category→Item Sub Category → Item name
- Owner→All→Owner-city→Owner-name
- Time→All→Year→Quarter→Month

Variables (Measures):

- Sell-Quantity
- Buy- Quantity
- Sell-Price
- Buy-Price
- Transportation-cost
- ID-quantity

Queries: In Model III, Manufacturer is analysed; based on a specified category; therefore, it is unable to answer queries which involve only manufacturer. A query such as “which are the top 5 manufacturer in outlet X” cannot be answered in the previous model. In the current model, following queries can be answered:

a) Which were the top 5 manufacturers for Chocolate 20 Gms in Jan 2009? (see Table 6)

b) Which are the top manufacturers for retailer chain?

c) Which manufacturer contributes 80% of total sale?

d) The supply of manufacturer M is more in which outlets?

e) Manufacturer M contributes more in which season?

Table 6: Top 5 manufacturer for Chocolate

Manufacturer	Item Sub Category	Item Name	Jan 2009
Amul	Chocolate	Chocolate 20 Gms	6821.00
Cadburys	Chocolate	Chocolate 20 Gms	6795.00
Nestle	Chocolate	Chocolate 20 Gms	6890.00
Parle	Chocolate	Chocolate 20 Gms	6712.00

4.1.4. Model IV

In this model manufacturer is represented as separate dimensions and hierarchy level is increased by adding outlet and owner state in Model III:

Dimensions and Hierarchies:

- Region→All→Outlet-state→Outlet-city→Outlet-name
- Product→All→Item-Category→Item Sub Category→ Item name
- Owner→All→Owner-state→Owner-city→Owner-name
- Time→All→Year→Quarter→Month
- Manufacturer→All→ Manufacturer

Variables (Measures):

- Sell-Quantity
- Buy-Quantity
- Sell-Price
- Buy-Price
- Transportation-cost
- ID-quantity

Queries: This model is able to answer queries from both model II and III.

4.1.5. Model V

In this model product dimension is represented to the Item level category only:

Dimensions and Hierarchies:

- Region→All→Outlet-state→Outlet-city→Outlet-name
- Product→All→Item Category
- Time→All→Year→Quarter→Month
- Owner→All→Owner-state→Owner-city→Owner-name
- Manufacturer→All→ Manufacturer

Variables (Measures):

- Sell-Quantity
- Buy- Quantity
- Sell-Price
- Buy-Price
- Transportation-cost
- ID-quantity

Queries: This model is able to answer all business queries provided by model II and III, which are not related with item name, Item subcategory and Item sub category name.

5. Conclusion

This work presents various multidimensional views (schema) to generate data warehouses from relational tables using a novel method to automate the multidimensional modelling process to satisfy business user requirements. The contribution of our approach is to identify measures, dimensions and hierarchies and generate all possible multidimensional models. This method is applicable for any relational model where the data types can be partitioned into numeric, date/time, and textual data types. Analyst will be able to choose the most suitable multidimensional model view that meets the business requirements. Efficiency is validated by applying various business queries in generated models. In future this approach can be used to build multidimensional models on heterogeneous database. Multidimensional models can also be used as a basic framework to develop data warehouse from relational tables.

Appendix I. information extraction according to hierarchical and sequential characteristics

- Which outlets are performing better and which are not?
- Why are some outlets not performing better?
- Are some items losing their popularity in selected outlets or as a whole?
- What will be the locations where outlet space or outlet numbers can be increased?
- What will be the forecast?
- In which season the sale of items will be less?
- In the subsequent month. Sale of which item will increase or decrease?
- What will be top 10% items in 2010?
- What is the Moving Annual Total (MAT) and Year-To-Date (YTD) sale of item HII_tea 500 Gms in all states in 2011? (see Table 7)
- What is the sale pattern of item HII_tea 500 Gms, in a Chhattisgarh? (see Table 8)

Table 7: MAT and YTD sale of item HII_tea 500 Gms

Item	State	Quantity								
		Year 2011								
		Oct	Nov	Dec	Oct		Nov		Dec	
					Year Rolling Sum	Year To Date	Year Rolling Sum	Year To Date	Year Rolling Sum	Year To Date
HII_tea 500 Gms	All	94770	96195	93500	1133625	949940	1138055	1046135	1139635	1139635

Table 8: Sale pattern of item HII_tea 500 Gms

Name	Year 2010											
	Quantity											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
HII_tea 500 Gms	109215	75215	112285	90460	92035	90485	91790	96475	97210	94770	96195	93500

References

Bollen P (2016). Fact-based declarative business rule modeling for the static and dynamic perspectives in ERP applications. In Multidimensional Views on Enterprise Information Systems, Springer International Publishing: 123-131.

Giorgini P, Rizzi S and Garzetti M (2005). Goal-oriented requirement analysis for data warehouse design. In Proceedings of the 8th ACM international workshop on Data warehousing and OLAP, ACM: 47-56.

Hümmer W, Bauer A and Harde G (2003). XCube: XML for data warehouses. In Proceedings of the 6th ACM international workshop on Data warehousing and OLAP, ACM: 33-40

Husemann B, Lechtenbörger J and Vossen G (2000). Conceptual data warehouse design. Universität Münster, Angewandte Mathematik und Informatik, Sweden: 1-6.

Jensen MR, Holmgren T and Pedersen TB (2004). Discovering multidimensional structure in relational data. In Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg: 138-148.

Peralta V (2003). Data warehouse logical design from multidimensional conceptual schemas. Universidad de la República, Uruguay.

Phipps C and Davis KC (2002). Automating data warehouse conceptual schema design and evaluation. In DMDW, 2: 2-2.

Romero O and Abelló A (2006). Multidimensional design by examples. In Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg: 85-94.

Tseng FS and Chen CW (2005). Integrating heterogeneous data warehouses using XML technologies. Journal of Information Science, 31(3): 209-229.

Vrdoljak B, Banek M, and Rizzi S (2003). Designing web warehouses from XML schemas. In Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg: 89-98.