Target Driven Approach Of Data Warehouse For Early Information In Requirements Engineering.

Hanu Bhardwaj, Ronika Sirohi

Department of Computer Science and Engineering,
Manav Rachna College of Engineering(MRCE)
Faridabad 121001

1. Abstract:

A Data Warehouse is designed for query and analysis rather than transaction processing. A Data warehouse requirements engineering basically has two parts, first is an 'early information' part where the information relevant to decision making is discovered, and second is a 'late' part where the information discovered is represented in terms of facts & dimensions. Data warehouse requirements engineering can be done using different approaches like database driven, goal driven etc. Here the Target Driven Approach is being used. Early information data warehouse requirements engineering starts with targets defined as pairs. In this targets are organized in a target hierarchy. This hierarchy is a complete specification of what is to be achieved by a top level target. We associate targets with choice sets so that alternative ways of target achievement can be represented. These alternatives are also organized in a choice set hierarchy. These techniques determine early information that is to be subsequently processed in the 'late information' requirements engineering stage. Our early information requirements engineering phase is illustrated through a case study of NDDB. Once the information is discovered, we will represent it using ER Diagram. Further we will implement the algorithm available to convert ER schema to dimensional Model using one of the various algorithm (i.e. Golfarelli., Moody etc). The dimensional model will be implemented using SQL Server 2005.

Index terms- Early information, Requirement Engineering, target driven, aspects, indicators, SQL.

2. Introduction:

Data warehousing technology supports information management for decision making by integrating data from operational systems and external sources in a separate database, the data warehouse. In contrast to operational systems which store detailed, atomic and current data accessed by OLTP (on-line transactional processing) applications, data warehousing technology aims at providing integrated, consolidated and historical data for OLAP (online analytical processing) applications.

Two stages of Data Warehouse Requirement Engineering are early information and late part. Targets are organized in target hierarchy. This hierarchy is a complete specification of what is to be achieved by a top level target and associated with choice sets so that alternative ways of target achievement can be represented. Thus from the set, choice is made according to the achievements of target .Thus achievement hierarchy is maintained. Late Information is structured as facts and Dimensions. Thus Early Information is to be subsequently processed in the late information requirements engineering stage. At the first stage of late part, E-R Model is drawn.

The E-R model is a detailed, logical representation of the data for an organisation or business area. It must be flexible enough so that it can be used and understood in practically any environment where information is modelled. It is expressed in terms of entities in the business environment, the relationships (or associations) among those entities and the attributes (properties) of both the entities and their relationships The E-R model is usually expressed as an E-R diagram. Than at second stage convert this diagram into multidimensional schemas.

There are schemas like star and snowflake which are used to support multi-dimensional data representation. They offer flexibility, but often at the cost of performance because of more joins for each query required. A star/snowflake schema models a consistent set of facts (aggregated) in a fact table and the descriptive attributes about the facts are stored in multiple dimension tables. In a star schema, a single fact table is related to each

dimension table in a many-to one and is easy to understand while in Snowflake Schema single, large, central fact table and one or more tables for each dimension. Dimension tables are normalized that is split into additional table.

2.1Data Warehouse Definition [Rth09]

Data Warehouses are developed to meet the growing demand for information analysis that could not be met by operational systems. This is because the processing load of reporting affects their response time and is not optimized for strategic decision making. It enables the organization to make use of an enterprise wide data store to link information from diverse sources. The information is now accessible to decision makers for strategic analysis which includes trend analysis, forecasting, competitive analysis &targeted market research.

2.2 Importance of Data Warehouse[Rth09]

2.2.1 Immediate information delivery: It reduces the time for processing of a request for eg. the sales report can be formulated on a daily basis which enables the business analyst to exploit opportunities.

2.2.2 Integration of data: It combines the data from multiple sources into a single unit

2.2.3 Provides an insight into the future: It stores a large amount of historical information that enables the decision makers to analyze the prevailing trends in market.

2.2.4 Enable decision makers to look data in different ways: It provide tools for manipulation of data and facilitate users to drill down into detail data with click of mouse

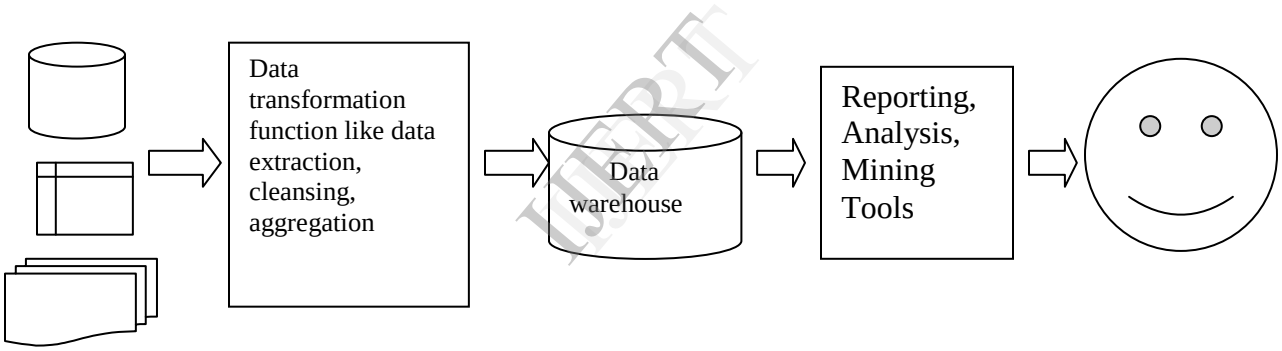

Figure 1: Data warehousing blend of many technologies

2.3 Data warehouse features: W.H.Inmonn [Imo00] defines the data warehouse as the subject oriented, integrated, and non volatile, and time variant collection of data in support of management decisions.

2.3.1 Subject Oriented data: Data warehouse is concerned with the things in the business environment that are driving day to day transaction. It stores data by subjects and not by application. These business subjects vary from one business to another e.g. For a retail company sales, products, customer may be critical business subjects. A particular subject may be involved in different types of transactions.

2.3.2 Integrated data: The integration process forms the data into a single cohesive environment. We remove all inconsistencies and errors and finally transform the data into a common format before storing into the data warehouse. The origin of data is invisible to the decision maker in the data warehouse environment. The Integration step consists of cleansing and transformation of data.

Cleansing: It is a process of removing errors and all the inconsistencies which improves the quality of data in the data warehouse. The extraction log records errors detected in the data cleansing process. The data administrator examines this log to determine the source of errors.
Data transformation: It receives the input data from the different operational systems and transforms them into one consistent format.

2.3.3 Non volatile data: We cannot update the data in the data warehouse in real time. Business transactions update the data in operational databases in real time. New records are added to the data warehouse periodically but existing records are not modified.

2.3.4 Time variant data: The decision makers can view the data across the field of time at whichever level of detail they may wish. This allows the business analysts to view the patterns and trends over time.

2.4 We can form the dimension model using schema:[Gol99]

2.4.1 Star schema: It is the simplest schema. It resembles a star radiating from a centre. The centre contains large fact table and the points of star are dimension tables. These are not joined to each other. Every dimension table is joined to the fact table using a primary to foreign key join. The dimension table is wide, demoralized has textual attributes, multiple hierarchies and has a unique identification key. The fact table contain the factual details of business events .It is deep but not wide. The star schema optimizes navigation, enhances query execution, easy to reconfigure and provides analytical flexibility. Figure 4 shows a simple star schema. It shows sales fact table in the middle and four dimension tables of customer, product, retail outlet, date.

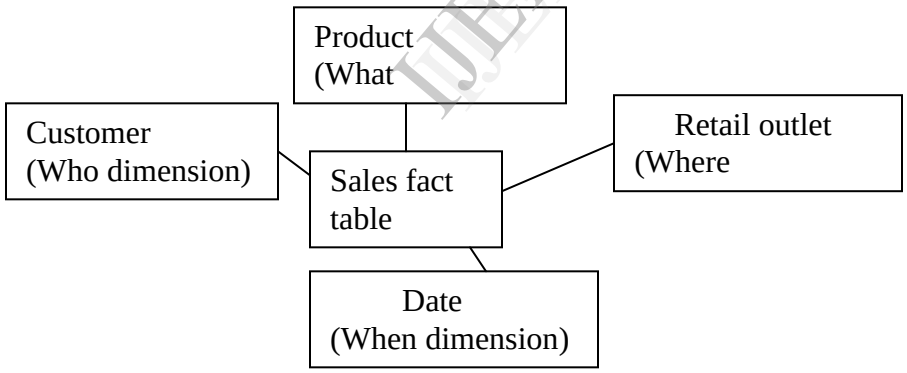

Figure-4: Star schema

2.4.2. ER schema to multidimensional schema: The design of a conceptual schema i.e. dimensional schema can also be carried by producing fact schema for each fact. Fact can be derived from an algorithmic procedure. Various algorithms are available some of them are given by: Golfarelli, Maio, and Rizzi, Husemann and Lechtenborger and Vossen.

2.5There is algorithm that specifies the process of converting ER schema to star schema [Dov05]:-

2.5.1 Golfarelli, Maio, and Rizzi proposed an algorithm in 1995 that involves following stages for conversion of an ER schema to star schema:[Gol98][Gor98]

In this approach, a fact scheme is structured as a quasi-tree whose root is a fact. A fact is represented by a box which reports the fact name and, typically, one or more measures. Dimension attributes are represented by circles. Each

dimension attribute directly attached to the fact is a dimension. Non-dimension attributes are always terminal within the quasi-tree, and are represented by lines. Subtrees rooted in dimensions are hierarchies.

2.5.2 Naveen Prakash, Hanu Bhardwaj[12] proposed two stages for data warehouse requirements engineering (i) an 'early information' part where the information relevant to decision making is discovered, and (ii) a 'late' part where this information is structured as facts and dimensions. Our focus is on the former. Early information data warehouse requirements engineering starts with targets defined as pairs of the form <A, I> where A is an aspect of an organization and I is a set of business indicators. An aspect is a work area, work unit, service or quality to be preserved in an organization. Business indicators are measures/metrics for specifying the desired performance level of aspects. Targets are organized in a target hierarchy. This hierarchy is a complete specification of what is to be achieved by a top level target. We associate targets with choice sets so that alternative ways of target achievement can be represented. These alternatives form their own hierarchy. Finally, information relevant to selection of each alternative is discovered through Ends, Means, Key Success Factor, and Outcome Feedback analysis techniques. These techniques determine early information that is to be subsequently to be processed in the 'late information' requirements engineering stage. Our early information requirements engineering phase is illustrated through a case study.

3. Objective(s)/Problem Statement:

Objective is to illustrate the Early Information and Late Information requirement engineering through a case study of NDDB (National Dairy Development Board). using Moody, Husemmann or Golfarelli for conversion of ER schema to a star schema.

Step1: Target is organized in target hierarchy.

Step2: Associate target with choice set and alternative way can be represented.

Step3: Relevant information is selected known as Early Information.

Step4: In late information draw the E-R Model of Early Information.

Step5: Convert the above E-R into Star Schema.

Step6: Implement the Schema in SQL server 2005.

4 .Methodology:

The Proposed Methodology for this is 'Target Driven approach' used to form the target hierarchy and choice set hierarchy of NDDB. The information collected from these hierarchies is represented by ER diagram. Further ER is converted into Dimensional schema using (Golfarelli, Moody etc.) algorithms. Implementation of the same in SQL Server 2005. Using Target driven Approach on the case study NDDB, we will able to understand and appreciate the methodology used for deriving information for Data Warehouse and its Implementation in SQL Server 2005. In Industry it also helps or guide for deriving the information required for developing a data warehouse.

6.Result:

The Intermediate outcomes:

- Target and choice set hierarchy is organized.
- Information is represented using ER diagram.
- ER is converted into Dimensional Model.

Finally:

- Implementation of Dimensional Model in SQL Server 2005 and analysis is done.
- Dissertation on related topic.
- A term paper.

Using Target driven Approach on the case study NDDB, we will able to understand and appreciate the methodology used for deriving information for Data Warehouse and its Implementation in SQL Server 2005. In Industry it also helps or guide for deriving the information required for developing a data warehouse.

8. References:

[Moh99] Mukesh-Mohania, John.F.Roddick, Yahiko Kambayashi, Kyoto University, University of South Australia Advances and Research directions in Data Warehousing Technology ", in AJIS Vol.7.No.1, Sept-1999.

[Pra12] Naveen Prakash and Hanu Bhardwaj "Early Information Requirements Engineering for Target Driven Data Warehouse Development" PoEM 2012,LNBIP 134, pp,188-202 , 2012.

[Sin09] Singh,Y.P,From "Early Requirements to Late Requirements Modeling for a Data Warehouse"Fifth International Joint Conference, Aug , 2009.

[Rth09] Reema tahreja, Data warehousing, Oxford,2009.

[Imo00] Inmon W.H., Building the Data Warehouse, John Wiley, New York(2nd edition ,2000)

[Gol99] Golfarelli, M., Rizzi, S," Designing the Data Warehouse: Key Steps and Crucial Issues" in Journal of Computer Science and Information Management, Vol. 2. No.3, 1999

[Pra08] Naveen Prakash, Anjana Gosain,"An approach to engineering requirement of datawarehouses, Requirement enginnering",Vol 1.No.13:49-72 ,2008

[Gor98] M. Golfarelli, D. Maio and S. Rizzi, Conceptual design of data warehouses from E/R schemes, Proc. Hawaii International Conference on System Sciences, Kona, Hawaii (1998), 334-343.