Asian Journal of Information Technology, 4(11): 1093-1097, 2005

ESCM: Entity-Subject Conceptual Modeling for Data Warehouse

M. K. Shahzad, J.A. Nasir, M. A. Pasha
Intelligent Information Systems Research Group (RG-IIS)
Punjab University College of Information Technology
University of the Punjab, Lahore

Abstract: Traditional Data Warehousing systems have static structure of their schemas and relationships between data; therefore they are not able to support any dynamics in their source structure and contents, consequent inconsistent analytical results. A novel modeling technique Semi-Star Schema was proposed to accommodate such dynamics. Multiple attempts have been made for conceptual modeling of star-oriented schemas of data warehouse but not for Semi-Star Schema. In this study we squabble that, in order to accurately reflect analytical business requirements into an error-free, understandable and easily extendable data warehouse schema our proposed conceptual model namely Entity-Subject Conceptual Model (ESCM) should be used. Which is (a) capable of interpreting business analytical requirements into abstraction (b) customized for tracing inter-attributes, entities, facts and dimension relationships. (c) formal and complete, so that it can be transformed into next logical schema with out ambiguities. (d) constructed in a customizable and extensible manner, so that the designer can enrich it.

Key words: Data warehouse, semi-Star, transactional entities, analytical entities, facts, dimensions, modeling, content changes, schema changes

INTRODUCTION

The most accepted definition of Data Warehouse (DW) states; a DW is subject-oriented, time-variant, non-volatile collection of integrated data. Researchers of the domain are trying to devise reliable strategies for maintaining consistent information in DW that facilitates user's decision-making and business forecast skills. Conventional DW systems are unable to support any change in their source structures and contents. In case of source change inconsistent analytical results are generated.

Modeling DW is not a trivial task as it directly deals with internal structures and implementation aspects. Several schemas have been developed using star-oriented approach like Star, Star-flake, Snowflake[1,2] and Semi-Star (SS)[1]. Attempts have been made for conceptual modeling of star-schema[4,5] but no attempt is being reported regarding the conceptual modeling of SS which has different requirements. We believe that in order to reflect user requirements into an error-free and understandable SS schema, special attention should be given to the conceptual modeling phase as it gives many benefits including effective requirement gathering, early error detection, schemas extendibility etc.

The scope of this study is restricted to understand SS modeling requirements from user point of view. Wepropose a new model, Entity-Subject Conceptual Model (ESCM), for conceptual SS Schema. Although the core of model is same as entity-relationship, but to meet SS requirements some new constructs are being added. Also Pedersen's criteria are used to evaluate the model. The derived results demonstrate that the model is formal, sound and complete.

SEMI-STAR SCHEMA

Conceptual modeling is an important phase in designing successful application. In this phase, user requirements are translated into abstract representations, which are independent of implementation details, nevertheless are understandable, formal and complete and can be transformed into the next logical schema without ambiguities. In this section we point out the special modeling needs of SS required at conceptual phase. Based on these, we proposed new constructs for designing a practical and efficient schema for Semi-Star model.

Consider the following example taken from a business environment. A company has sale points in multiple cities, which are grouped into administrative regions. Each category contains more than one product. Total amount can be obtained by multiplying the sold quantity of product with unit price. For analysis, sales are inspected in various locations at certain time.

Corresponding Author: M. K. Shahzad, Intelligent Information Systems Research Group (RG-IIS) Punjab University College of Information Technology University of the Punjab, Lahore

For modeling this business scenario, at conceptual phase for SS we need to perform followings:

Represent facts and their properties: Facts are actual metrics of the real world and are characterized by properties. Fact properties are usually numerical data and can be summarized in various ways in order to extract further information. For this reason, they are usually called as summary properties, aggregates, measures or calculated attributes.

Connect Time dimension to facts: For a better decision making users must be able to drill down and roll up along the time dimension. Connect time dimension to facts allows users to analyze and view results in a variety of ways.

Represents analytical objects and capture their properties: In analytical objects, aspects are connected to facts, which make analysis process more effective. For example, city object is connected to identification number CityID. In addition, three special types of associations among objects exist: i) Specialization/ generalization ii) Aggregation iii) Membership

Identify entities attributes and entity sets: An entity is a distinguishable object in real world described using a set of attributes. Entities are collected and maintained with relationship among them for transactional purpose. A collection of similar entities is called an entity set. Entities sets need not to be disjoint. All entities in a given entity set have same attributes. For example, City is an entity; also region and product are amongst entities.

Entity relationship and relationship set: A relationship is an association among two or more entities. As with entities, similar relationships can also be collected to form a "relationship set". For example, one product can be sold at one or more than one cities.

Record association between objects and facts: Facts are semantically connected to objects. For example, how much 'amount' is earned by selling a 'product' i.e. object. So product is associated with a fact 'amount'.

Distinguish dimensions and categorize them: Objects that are connected via associations to facts are called dimensions and they are usually the focus of the data warehouse analysis. SS transactional tables also act as dimensions, which have two types static behavior dimensions (SBD) and dynamic behavior dimensions (DBD). This association is of great importance as its help to retrieve and analyze data in terms of various aspects. For example, analysis between sales-amount, repayment and city shows the connection between the amounts earned from the city.

Distinguish transactional and analytical objects: At conceptual phase transactional and analytical objects should be identified, along with shared ones. For example, in sales company, product and city are transactional and analytical objects.

MODELING Entity-Subject Conceptual Model (ESCM)

Once all the requirements have been collected and analysed, the next step is to create a conceptual schema for the database using a high-level conceptual data model. This step is called conceptual design. Conceptual model cannot be represented without certain constructs that represent each structure uniquely. ESCM addresses the modeling requirements of SS, from the user point of view and offer following unique constructs:

- **Fact set /Aggregations:** Rrepresents a set of real world metrics. Each fact item, aggregate or measurement goes into fact table as an attribute. It is represented by circle e.g. amount

Amount 

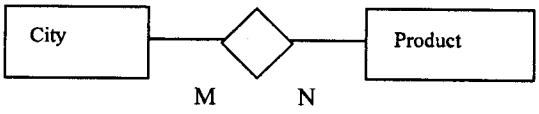
- **Entity (Object):** Represents a set of real world object with similar properties, about which information is gathered and to be maintained e.g. product

Product 

It has two types, strong entity and weak entity, the representation used for weak entity is

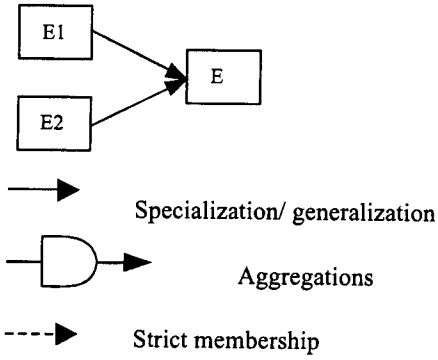
Weak entity 

- **Relationship between the entities:** This set represents a set of relationship possible between entities, it can be one to many (1:M), many to many (M:M) and many to one (M:1) e.g. one city can have more than one product and vise versa.

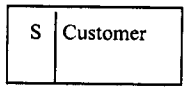


- **Attributes:** are the properties of entity sets, relationship set, facts and dimensions e.g. City ID and City Name for City, Product Name of Product. Various types of attributes are represents by making few modifications. Attributes leveled by attributes called multi-valued.

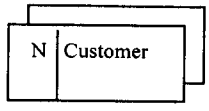
- Generalization/ specialization:



- **Shared Analytical Entities:** Selected Transactional Entities (tables) also act as dimensions. Due to their presence schema up-gradations should be avoided. For example, in our example product is shared entity (dimension). Here S indicates level.



- **Non-Shared Analytical Entities:** Also it is required to generate new entities for analytical purpose and efficiency, called non-shared entities. For example, the dimension of time in example.



- **Surrogate Key:** Shared dimensions are of two types called Static Behavior Dimensions (SBD) and Dynamic Behavior Dimensions (DBD). In SBD lower level of dimension table is fixed with upper level, while in DBD lower level may change. In SS different levels of DBDs are identified by surrogate key. An auto-generated primary key. For example, City_Key, as dimension of location is considered to be DBD.

- **Derived Attributes:** Some of the attributes can be derived by formulating facts, although these attributes are not maintained but at conceptual level they are required to be present. They are represented by hexagon. For example, profit.

EXAMPLE OF USE

Let us consider an example of company's SS, which stores data about company's sales. Company has sale points in multiple cities. Whole area is divided into regions, each region have more than one city. Multiple products are present for sale in each city; also products are grouped into categories. Sales are inspected in various locations at certain time. Temporal analysis is always the requirements while product can also be used as an dimension for analysis.

SS schema for such a company is show in Fig.1.which has two types of tables. As in example rest of constructs can be used for designing SS conceptual model for sales company.

ESCM EVALUATION

Pedersen's evaluation[6] criteria are used to evaluate proposed model. The criteria allow us to check the correctness, modeling power, efficiency in information capturing, design independency. The criteria are:

- **Explicit hierarchies in dimension:** The hierarchies in dimension should be captured explicitly, so that the user has available the relation between the different hierarchical level. ESCM supports explicit hierarchies, in non-shared dimensions e.g. time dimension.

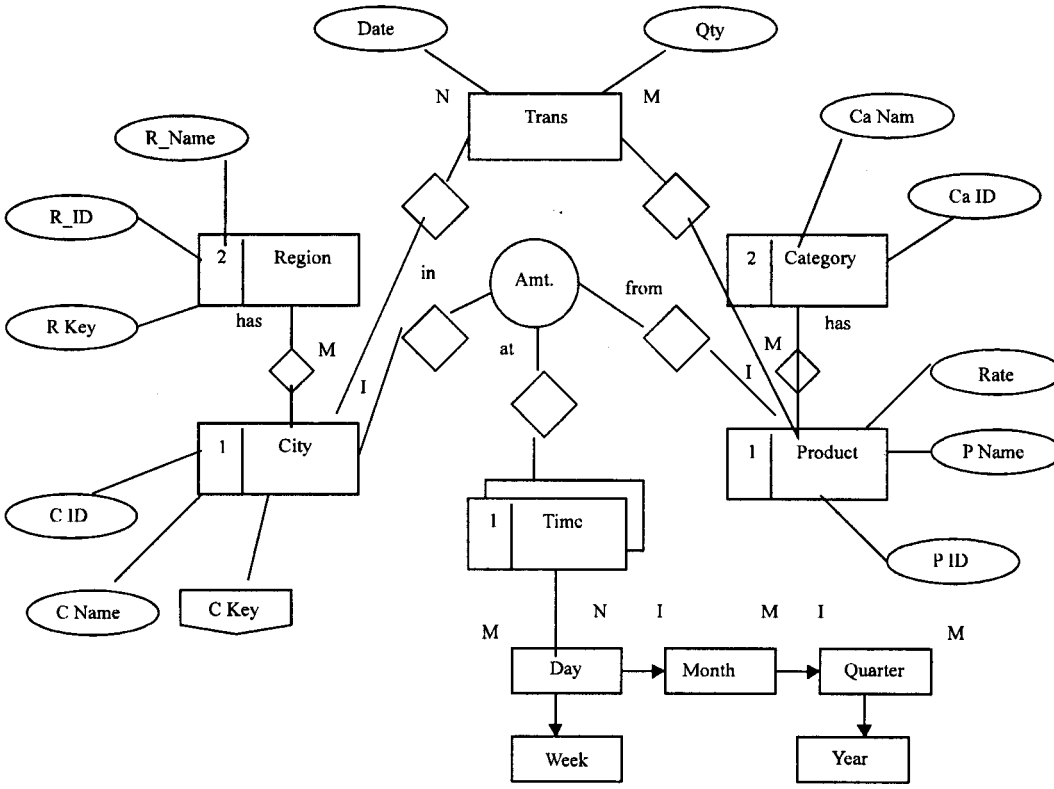


Fig. 1:Example for use of Semi-Star Schema

- **Symmetric treatment of dimensions and measures:** The model should allow measures to be treated as dimension and vice versa. In ESCM constructs, we have made a conceptual distinction between dimension and measure, but we do not restrict the user from modeling a dimension as a particular summary property in order to add extra feature for analysis. In our example, the 'amount' is the calculated attribute, to allow for computations such as summing up values, profit etc. also explicit hierarchies can be maintained in shared dimensions.

- **Multiple hierarchies in each dimension:** Any dimension can have more than one hierarchy. In our example of the time dimension days can roll-up to months and days can also roll-up to weeks, months, quarters and years.

- **Aggregation support:** The data model should give meaningful summaries or aggregations to user. ESCM provides more support aggregations and derived attributes. For example, 'amount' and 'profit'.

- **Support of non-strict hierarchies:** ESCM supports such hierarchies via the relationship cardinality.

- **Support of many-to-many relationships between facts and dimensions:** ESCM have full support for such tasks.

- **Handling different levels of granularity at summary properties:** ESCM uses membership hierarchy this purpose. For example 'amount' can be summarized per year, following the granularity of the 'time' dimensions.

CONCLUSIONS

In this study, through an example, we have illustrated a set of modeling requirements from the user's point of view. These requirements reveal a set of concepts that need to be included into conceptual models for designing an efficient DW. The focus of our research is to define special modeling needs of the schema, drawn from theoretical and practical experience. These requirements

reveal a set of concepts that need to have included into conceptual model for efficient design, so we anticipate the basic modeling constructs. A case study is formed to show the example of use. Finally, to evaluate the proposed model Pedersen's criteria are used. The results illustrate that our model gives complete conceptual model for SS. This study has only cover one aspect of SS. Transformation process in SS, its physical design issues, versioning needs and semantic modeling are some other potential areas of research.

REFERENCES

1. Paulraj Phonniah, 2002 . Data Warehousing Fundamentals. John Wiley and Sons.
2. Morzy, T. and R. Wrembel, 2003 Modeling a Multiversion Data Warehouse: A formal Approach, Proceedings of ICEST.,
3. Pasha, M.A., J.A. Nasir and M.K. Shahzad, 2004 Semi-star schema for managing data warehouse consistency. Proceedings of IEEE-NCET, 1999.
4. Tryfona, N., F. Busborg and J.G Borch, 1999. StarER: A Conceptual Model for Data Warehouse. Proceedings of DaWak.,
5. Datta, A. and H.Thomas, 1997 A Conceptual model and an algebra for Online Analytical Processing In Data Warehouses. Proceedings of WITS,
6. Pedersen, T.B. and C.S. Jensen, 1999. Multidimensional data modeling of complex data, Proceedings of ICDE.