CLOSING THE GAP: REQUIREMENT OF A CONCEPTUAL MODEL

Resmi Pillai
Pillai@csse.monash.edu.au

Bala Srinivasan
srini@csse.monash.edu.au

School of Computer Science and Software Engineering
Monash University, Australia- 3145

1 Abstract

The gap between expectations of the data warehouse user community and the solutions provided by the researchers as well as IT companies seems to be widening. The aim of this paper is to look at some of the burning issues faced by the practitioners and the way in which those issues are addressed by the researchers. The paper stresses on the need of a data model for a seamless transition from the conception stage through to the full life cycle of a data warehouses system.

Keywords*:* Conceptual model, Data warehouse, OLAP, Hierarchies.

2 Introduction

Organizations making quick and better business decisions succeed in today's competitive marketplace. Understandably, organizations seeking to improve their decision-making can be overwhelmed by the sheer volume and complexity of data available from their varied operational and production systems. Making this data available to a wide audience of business users and allowing them to extract high value information is one of the most significant challenges for today's information technology professionals.

In response, many organizations choose to build a data warehouse to unlock the information in their operational systems and understand real-world business problems. The adoption of data warehouses has helped many companies respond to their information needs. The software companies and researchers engaged in the field of Data Warehousing have come up with different technologies and concepts. Even then the data warehouse community is dissatisfied with the capabilities of DW systems as they fail to deliver their expectations.

3 Requirements/Expectations

The Data warehouse designers as well as researchers agree on the requirement of a conceptual model for expressing the semantics of the system. The requirement specification through to the design stage is not seamless because of a lack of a conceptual model and this makes it difficult to implement the expected functionalities envisioned by the users. The current data models used by the industry are not sufficient enough to make the system robust and flexible. One of the inherent properties of data warehouses is that the requirements change dynamically and the system should be flexible enough to incorporate these changes with out major re- work.

Some serious research has been done to formalize the requirements of a data model for data warehouses. According to [5] a data model should support the following properties.

1. Handling explicit hierarchies
2. Symmetric treatment of measures and dimensions
3. Support for multiple and non-strict hierarchies
4. Many to many relation between fact and dimension
5. Data with different level of granularities
6. Correct aggregation of data.
7. Uncertainty associated with data.
8. The change in data overtime

Even though it is hard to prioritize these requirements, as it changes with the type of system, aggregation of data is a key functionality of data warehouses. We discuss some of the proposed models later and analyze them in terms of their concordance with the above mentioned properties and also relating them to the demands of the user community.

4 Current tools

Broadly we can classify the existing data warehouse implementations into two, one with underlying data in relational tables with client tools/application server having the OLAP functionality (ROLAP) and the other has a dedicated multi-dimensional database (MDD) with multi-dimensional storage of data in an array like structure with client tools just having the capability to manipulate and display the summary information (MOLAP). We could also find custom developed data warehouse applications on a relational database with the client tools having some built-in logic to substitute the OLAP functionality.

Some of the popular products that use ROLAP technology are MicroStrategy's DSS server and related products, Informix's (now owned by IBM) Informix-MetaCube, Information dvantage's Decision Suite and those using MOLAP technology are AlphaBlox from AlphaBlox Corporation, BusinessObjects from BusinessObjects SA, PowerPlay by Cognos, Essbase from Hyperion Solutions (formerly Arbor Software), Oracle Corporation's Oracle Express, Sybase's IQ and Planning Sciences' Gentium.

These tools have helped companies, to an extent, to respond to an ever-shifting competitive landscape. But as requirements and business structures are evolved these tools proved to be inadequate to meet the information needs of users.

5 Research efforts

Research in Data warehousing is continuously increasing so as the literature in this area. We take some of the major works on modeling here and analyze them in terms of their practicality and richness of semantics.

In [6] the focus is on combining the features of E/R modeling and star schema. The concept of dimensions in star schema is treated as entities and an additional construct for fact sets have been proposed. The attributes are classified into stock, flow and value-per-unit and represented in the model as "S", "F" and "V". This helps to identify if an attribute can be summed or averaged. The strictness/non-strictness of dimension hierarchies is also represented in the model. The model fails to define the concepts of those new constructs proposed by them. For example, in the sample system they are trying to analyze, "repayment" is taken as a fact set. In E/R modeling it can be modeled as an entity type. The paper does not discuss the issue of aggregation where a non-strict hierarchy is modeled.

[1] Proposes a logical cube model and algebra. The cell values of the cube can be 0, 1 or an n-tuple. 1 represents that the measure value exits for that combination of dimensions and 0 corresponds to a cell with no contents. An n- tuple shows the presence of multiple measures for the dimension values. There is no definition for the dimension or their levels. This can be done only through operators. The structure of dimension is one of the important properties in OLAP applications. Even though the authors claim this as a logical model it assumes relational mapping.

[2] Model is an extension of classical multi dimensional model. Main focus of the model is easy analysis of data. The hierarchical data is categorized into two levels. The first level, called Primary Multidimensional Object (PMO), allows classification oriented analysis and the second level, called Secondary Multidimensional Object (SMO), for feature oriented analysis. According to this model different instances of the same level might have different attributes. This complicates the issue of defining a meta data.

The model proposed by [4] formalizes a multidimensional model for OLAP applications. The basic component of the model is a multidimensional cube, which consists of a number of relations (dimensions) and for each combination of dimension (coordinate) there is a scalar value. The grouping algebra and definition of aggregation operation is strong enough to treat this as a good conceptual model but the single value mapping is not good enough in OLAP applications.

MAC model by [7] starts modeling from the end user point of view. The central concept of the model is the Multidimensional Aggregation Cube (MAC), which gives a broad and flexible definition for a multidimensional cube. The MAC is equivalent to an n- way relationship relating measure values to a set of dimension values. Multiple hierarchy and non-strict hierarchy are handled by defining drilling relationships. The dimension domain allows the dimension values represent all possible properties that can be used for multidimensional analysis. Almost all multidimensional analysis is based on aggregation functions so it should be captured at the conceptual level itself. This model doesn't consider aggregation function at all.

6 The Gap

There exists a gap between the expectation of the user community and the final implemented system. This gap is caused by various reasons such as the data is not defined correctly or the data integration of source systems is not done properly or analysis views that are in the 'can't do' list for power users etc. The users are often unable to define what they need and technicians do not have the business knowledge to assess the requirements of the users. This frequency mismatch and the issues associated with them are because of a lack of conceptual model that can be used during the conceptualization of the system.

There is a gap existing between the researchers and the software companies who bring out products incorporating the research ideas. Basically the systems should have the drill down and roll up functionality and requires the lowest level and aggregated data. The amount of low level data in a data warehouse is humongous and the roll up functionality needs pre-aggregation and for those adhoc queries the users would have to wait. As there is a tendency to measure the data warehouse tool in terms of their response time the software companies tend to concentrate much of their effort in improving the response time and over look the importance of a logical model that satisfy the requirements as mentioned in section 1.

An ideal data model should be seamless during the transition from conceptual model to logical model and then from logical model to implementation model. Most of the data warehouse tools are based on either relational or multidimensional cube model. So the proposed conceptual models can not be integrated with current tools in a seamless way. Logical models should have a strong algebraic base so that an interface can be implemented to query the underlying data. That leads us to a chicken or egg problem.

7. A step towards closing the gap in Data warehouse

In order to reduce the gap here we are proposing a conceptual model which represents user requirements. This can be summarized as

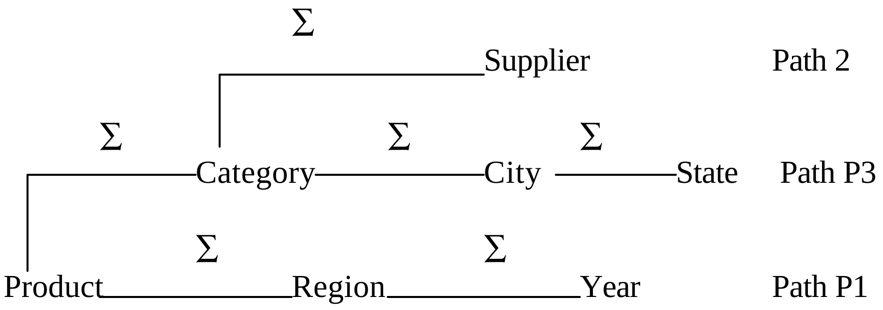
1. Requirement Specification
2. Defining aggregation
3. Translation

7.1 Requirement Specification and Aggregation

To capture user requirements we are analyzing user queries. Some of the examples are given below.

1. Give me the sales of health care Products by Region during the Year 2000.
2. Show me the category performance by supplier.
3. What is the category performance by city and state?

The concept of analysis path [7] is used here for requirement specification.



Each query is considered as an analysis path, which is represented as P1, P2 and P3. Along these paths it needs an aggregation to produce the results. The richness of the schema increases with the aggregation function. In the above examples the function is "SUM" and that also should include along the path.

7.2 Translation

The final step is to translate this schema to the operational schema. During the translation process the entire paths may not be aggregated. Aggregation along the path depends on some conditions, [3] stated this as summarizability conditions.

In Path2 product is aggregated to the category level then it is aggregated to supplier assuming that products are supplied by category wise. So it can be translated. The Path P3 in the above figure violates one of the summarizability conditions, namely disjointness of category attributes when a particular city is associated to more than one state. This situation can be overcome by adding a prefix to the names. So it is difficult to translate all the paths to operational schema and this depends on the defined aggregation function.

8. Conclusion

In this paper we looked at some of the concerns of data warehouse users and analyzed them against the efforts of researchers to alleviate them. And also proposed a data model based on user requirements. We found that some of the issues the DW community is unheard off by the researchers as well as the software companies. The most important contributor of this widening gap is the lack of a conceptual model for data warehouse. While the software companies concentrate their effort towards optimizing the underlying data structure and query mechanism the researchers should continue to look for a conceptual model to specify the user requirements.

REFERENCES

[1] Agrawal,R., Gupta, A., Sarawagi, S., " Modeling Multidimensional Databases". Int 13th Int. Conf. On Data Engineering (ICDE' 97)

[2] Lehner, Wolfgang "Modeling Large scale OLAP scenarios" 6th Int. Conf. on Extending Database Technology (EDBT) 1998

[3] Lenz,J. Hans, Shoshani,Arie " Summerizability in OLAP and Statistical Data Bases" In Proc. of SSDBM 1997

[4] Li, Chang. Wang, Sean. "A Data Model for Supporting On- Line Analytical Processing". Proc. Conf. On Information and Knowledge Management. 1996

[5] Pedersen,T.B., Jensen,C.S.," Multidimensional Data Modelling For Complex Data" Int.Conf.On Data Engineering

[6] Tryfona, N., Busborg, F., Christiansen, J. star ER: A Conceptual Model for Data Warehouse Design". In Int. Workshop on Data Warehouse and OLAP (DOLAP) 1999.

[7] Tsosis, Aris. , Karayannidis,N. " MAC: Conceptual Data Modeling for OLAP". Proc. Of the Int. Workshop on Design and Management of Data Warehouses (DMDW 2001).