ISSN: 2321-7782 (Online)
Volume 3, Issue 8, August 2015
International Journal of Advance Research in
Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Implementation of Star Schema Using XML Document as Source in Data Warehouse

Jyoti[1]
CBS Group of Institution,
Fathepuri, Jhajjar

Rajiv Munjal[2]
Assistant CSE, Department,
CBS Group of Institution,
Fathepuri, Jhajjar

Upasna[3]
Ganga Institute of technology and management,
Jhajjar

Abstract: Data warehousing is a complicated task which involves knowledge of business processes and familiarity with the operational databases. 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. We have seen various sources of data for a data warehouse but we have chosen XML Document as our source. We have implemented two algorithms for converting XML Document to Star schema and made comparisons on the basis of their working, principle.

Keywords: XML, OLAP, OLTP.

I. Introduction to Data Models

In most database environments, users perform two basic types of tasks: modification (inserting, updating, and deleting records) and retrieval (queries). Modifying records is generally known as online transaction processing (OLTP). Data retrieval is referred to as online analytical processing (OLAP) or decision support, because the information is often used to make business decisions. This section describes these data models and their structural requirements.

When database records are modified, the most important requirements are update performance and data integrity. These needs are addressed by the entity relation model of organizing data. Entity relation schemas are highly normalized. This means that data redundancy is eliminated by separating the data into multiple tables.

The process of normalization results in a complex schema with many tables and joins paths. When database records are retrieved, the most important requirements are query performance and schema simplicity. These needs are best addressed by the dimensional model. Another name for the dimensional model is the star schema.

A diagram of a star schema resembles a star, with a fact table at the centre. The following figure is a sample star schema. A fact table usually contains numeric measurements, and is the only type of table with multiple joins to other tables. Surrounding the fact table are dimension tables, which are related to the fact table by a single join.

Dimension tables contain data that describe the different characteristics, or dimensions, of a business. Data warehouses and data marts are usually based on a star schema. In a star schema, subjects are either facts or dimensions.

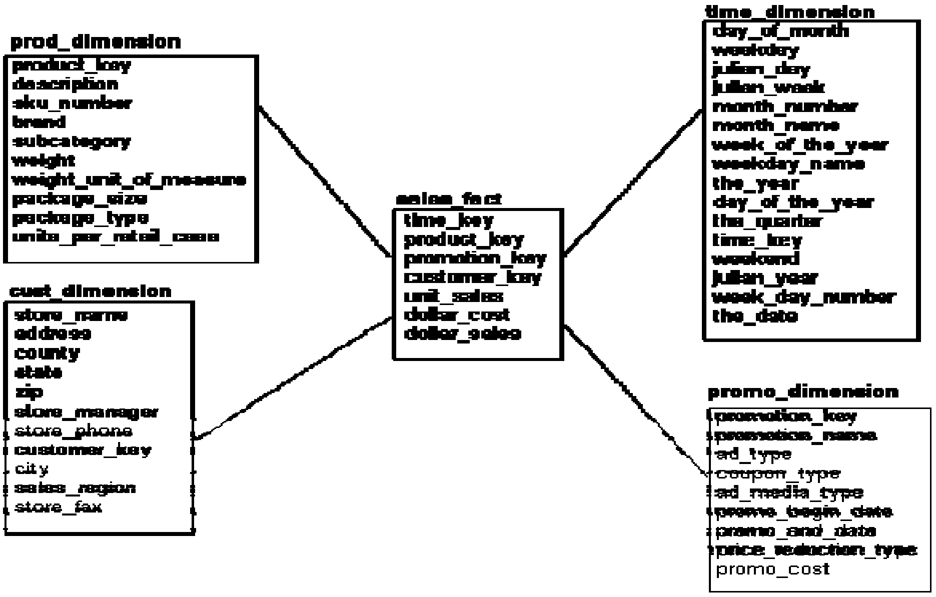


Fig 1. Diagram of a Star Schema

II. Star Schema Key Structure

The join constraints in a star schema define the relationships between a fact table and its dimension tables. In the star schema diagram at the beginning of the chapter, the product key is the primary key in the product dimension table. This means that each row in the product dimension table has a unique product key. The product key in the fact table is a foreign key drawn from the product dimension table. Each row in a fact table must contain a primary key value from each dimension table. This rule is called referential integrity and is an important requirement in decision-support databases. The reference from the foreign key to the primary key is the mechanism for verifying key values between the two tables. Referential integrity must be maintained to ensure valid query results. The primary key of a fact table is a combination of its foreign keys. This is called a concatenated key. The join cardinality of dimension tables to fact tables is one-to-many, because each record in a dimension table can describe many records in the fact table. A star schema database uses very few joins, and each join expresses the relationship between the elements of the underlying business. The join between the product dimension table and fact table represents the relationship between the company's products and its sales.

III. Star Schema Advantages

Star schemas are easy for end users and applications to understand and navigate. With a well-designed schema, users can quickly analyze large, multidimensional data sets. The main advantages of star schemas in a decision-support environment are:

- » **Query performance:** Because a star schema database has a small number of tables and clear join paths, queries run faster than they do against an OLTP system. Small single-table queries, usually of dimension tables, are almost instantaneous. This design feature enforces accurate and consistent query results.
- » **Load performance and administration:** Structural simplicity also reduces the time required to load large batches of data into a star schema database. By defining facts and dimensions and separating them into different tables, the impact of a load operation is reduced. Dimension tables can be populated once and occasionally refreshed. You can add new facts regularly and selectively by appending records to a fact table.
- » **Built-in referential integrity:** A star schema has referential integrity built in when data is loaded. Referential integrity is enforced because each record in a dimension table has a unique primary key, and all keys in the fact tables are legitimate foreign keys drawn from the dimension tables.

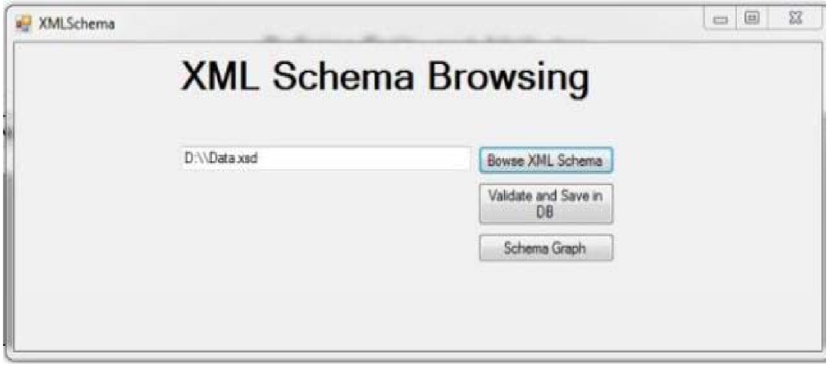
» **Easily understood:** A star schema is easy to understand and navigate, with dimensions joined only through the fact table. These joins are more significant to the end user, because they represent the fundamental relationship between parts of the underlying business.

IV. DESIGNING A MODELING STARS USING XML

A data warehouse (DW) is a collection of data from many sources. The data is used for reporting, analysis, issue resolution, and predictive modeling. In other words, a DW is a special database. As with other databases, a design of any DW requires certain modeling stages. Since DW prepares the data for analytical processing, which requires multidimensionality, it seems to be suitable to use a multidimensional model. Usually multidimensional modeling for DW makes a separate design stage in its live cycle. The conceptual stage in DW design covers the aspects of data which express its associations to real world objects, while the dimensional view reflects dimensional requirements on data. We preferred a dimensional modeling (DM) based on dimension and fact tables. Recently, XML has significantly influenced building databases [4]. XML data is generated by applications and it can be consumed by applications. It is not too hard to imagine that some data sources in the enterprise are repositories of XML data or that they are viewed as XML data independently on their inner implementation. In this case we could try to build a DW over XML data.

V. EXPERIMENTAL RESULTS

On pressing that button labeled as Schema graph, a Schema graph of given XML Schema is shown on the screen. Consequently, on selecting the option of Star Schema, a Star Schema is built against the Schema graph.



XMLSchema

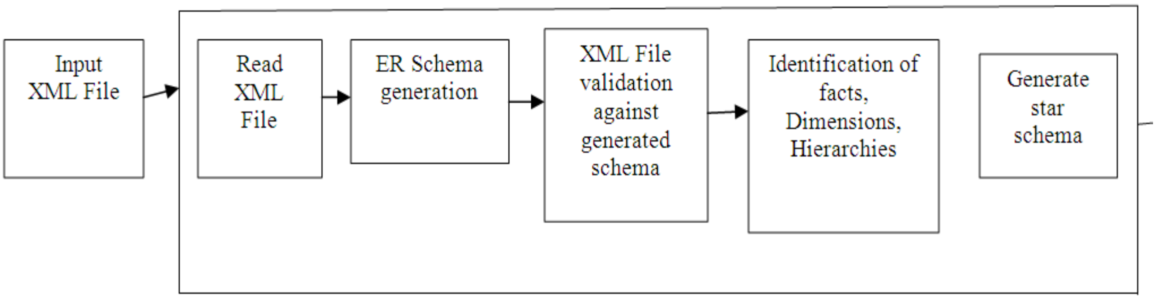
XML Schema Browsing

D:\\Data.xsd

Fig 2 XML Schema Browsing

VI. IMPLEMENTATION OF DESIGN OF TOOL

Most of the development methods utilized in today's software development include such a conceptual design phase. For instance, in relational database development, the conceptual design is presented using Entity Relationship (ER) diagrams [ER] , and in software model development, the conceptual model is presented using data flow diagrams (DFD's), and in Object-Oriented design, the conceptual model is presented using the Unified Modeling Language (UML), and so forth. Unfortunately, the area of conceptual design with XML has not been explored significantly in literature or in practice.



Input XML File
Read XML File
ER Schema generation
XML File validation against generated schema
Identification of facts, Dimensions, Hierarchies
Generate star schema

Fig3. Design of Tool

VII. Conclusion

XML is chosen as the data source to be transported over the internet between various Web applications and here specifically facilitated the transferring of data in XML format from various heterogeneous data sources to the Warehouse Schema. Data present in these heterogeneous data sources in the form of XML Schema and after going through the ETL phase. The Warehouse kept the contents of XML Schema in the form of their Schema. In order to form the Warehouse Schema, the CASE Tool in this paper will be used to convert the XML Schema into the appropriate Warehouse Schema. This approach can be helpful in the business intelligence in organizations to build the Warehouse Schema in short duration as the proposed methodology cab be applied on to the changes in the XML sources whenever needed and quickly build the Warehouse Schema corresponding to that change.

References

1. Reema Thareja, Data warehousing, Oxford,2009.
2. Inmon W.H., Building the Data Warehouse, John Wiley, New York(2nd edition ,2000)
3. Rilson,F., and Freire, J., "DWARF: AN Approach for Requirements Definition and Management of Data Warehouse Systems", Proceeding of the 11th IEEE International Requirements Engineering Conference 1090-9, September 08 – 12,2003.
4. Husemann, B., Lechtenborger, J., Vossen, G., "Conceptual Data Warehouse Design" Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) Stockholm, Sweden, June 5-6, 2000
5. 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
6. Naveen Prakash, Anjana Gosain,"An approach to engineering requirement of datawarehouses, Requirement enginnering",Vol 1.No.13:49-72 ,2008
7. 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
8. Moody, D.L.and Kortink, M.A.R.,"From ER Models to Dimensional Models: Bridging theGapbetween OLTP and OLAP Design", Journal of Business Intelligence, 8, 2003.
9. Moody, D.L. and Kortink, M.A.R.," From ER Models to Dimensional Models: Advanced Design Issues", Journal of Business Intelligence, 8, 2003.
10. Chen,P.P.,"The entity relationship model :toward a unified view of data", ACM transactions on database system,1,1,march 1976,9-37
11. Dov dori,Roman Feldman,Arnon Sturnm," Transforming an Operational System Model to a Data Warehouse Model: A Survey of Techniques", Proceedings of the IEEE International Conference on Software - Science, Technology & Engineering .