E-Business Data Warehouse Design and Implementation

Xudong Chen

Computer Department, Jiaying University, MeiZhou 514015, China
chenxd@jyu.edu.cn

ABSTRACT

E-Business have a variety of on-line transaction processing (OLTP) systems and operational database. Data Warehouse is different from operational database. A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process. The features of Data Warehouse cause the its design process and strategies to be different from the ones for OLTP Systems. This paper presents a brief description of approaches that address the data warehouse Design and Implementation for E-business.

Keywords: E-Business, Data Warehouse, Data Warehouse Design, OLAP , Dimensional models

1. INTRODUCTION

In the last decade, we witnessed numerous extraordinary business and technology innovations, such as business process reengineering, enterprise resource planning systems, and electronic business[6]. As transaction databases become more powerful, communication networks grow, and the flow of commerce expands, E-Business have accumulated a variety of on-line transaction processing (OLTP) systems and operational database.

OLTP systems (such as financial, order, work scheduling) which create operational data operational data focuses on transactional functions. This data is part of the corporate infrastructure. It is detailed, nonredundant, and updateable.

However, the rapidly changing market dynamics, competitive pressures, globalization of the commercial markets, reduced profit margins, and other similar factors forced business to review their structures, approaches and strategies[1].

A data warehouse collects, organizes, and makes data available for the purpose of analysis in order to give management the ability to access and analyze information about its business. This type of data can be called informational data. The systems used to work with informational data are referred to as online analytical processing (OLAP).

OLAP is the technology that enables client applications to efficiently access the data organized by data market and data warehouse. Data warehouse, provide a database organized for OLAP rather than OLTP, can solve OLTP problems.

Data warehousing is one of the most important strategic initiatives in the information systems field. These repositories of data play critical roles in understanding customer behavior in customer to business e-commerce, connecting trading partners along the supply chain, implementing customer relationship management strategies, and supporting comprehensive performance measurement systems[7].

2. DATA WAREHOUSING ARCHITECTURE

Bill Inmon defined the term data warehouse: "A (data) warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process."[4]. Subject-oriented: Data that gives information about a particular subject. Integrated: Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole. Time-variant: All data in the data warehouse is identified with a particular time period.

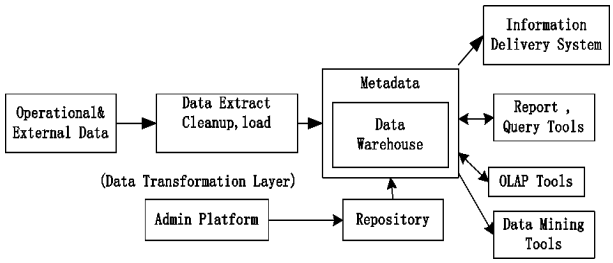

Fig. 1 Data Warehouse Architecture (Refers to Berson 1997)

Data warehouse architecture is based on a relational database management systems server that functions as the central repository for informational data (see Fig.1). Data transformation layer involves a wide range of transformations that have to be applied to source data, for example data quality control and data cleaning, data integration, and conversions that are necessary for adapting data to the DW structures. Metadata: It is data about data that describes the data warehouse, used for building, maintaining, managing, and using the data warehouse. Information delivery system: It distributes data warehouse to other data warehouse and end-user. Data Mining: It is the process of discovering meaningful new correlations, patterns, and trends by mining large amounts of data stored in data warehouse.

3. DATA WAREHOUSE DESIGN AND IMPLEMENT

3.1 Data warehouse design

Data warehouse design is different from traditional OLTP. Kimball,Ralph (1996) proposed Nine-Step Method in the Design of Data Warehouse[3]: 1. Choosing the subject matter. 2. Deciding what a fact table represents. 3. Identifying and conforming the dimensions. 4. Choosing the facts. 5. Storing precalculations in the fact table. 6. Rounding out the dimension tables. 7. Choosing the duration of the database. 8. The need to track slowly changing dimensions. 9. Deciding the query priorities and the query modes.

Most successful data warehouses design are based on a dimensional model. Dimensional models [2,5] represent data with a "cube" structure, making more compatible logical data representation with OLAP data management. The key to build data warehousing is data design. The business users know what data they need and how they want to use it. Focus on the users, determine what data is needed, locate sources for the data, and organize the data in a dimensional model that represents the business needs.

The basic concepts of dimensional modelling include: A fact, dimensions and measures. A fact is a collection of related data items, consisting of measures and context data. It typically represents business items or business transactions. A dimension is a collection of data that describe one business dimension. Dimensions determine the contextual background for the facts; they are the parameters over which we want to perform OLAP. A measure is a numeric attribute of a fact, representing the performance or behaviour of the business relative to the dimensions.

The principal characteristic of a dimensional model is a set of detailed business facts surrounded by multiple dimensions that describe those facts. When realized in a database, the schema for a dimensional model contains a central fact table and multiple dimension tables. A dimensional model may produce a star schema or a snowflake schema.

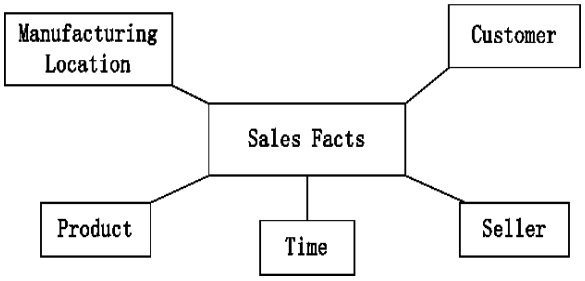

Fig. 2 The star schema example

The star schema is the basic structure for a dimensional model. It has one large fact table and a set of smaller dimensions tables arranged in a radial pattern around the central table.

In this example (see Fig. 2): Sales table is fact table. Time table, customer table, Seller table, Product table and Manufacturing table are dimensional table.

The snowflake schema is the result of decomposing one or more of the dimensions. The many-to-one relationships among sets of attributes of a dimension can separate new dimension tables, forming a hierarchy.

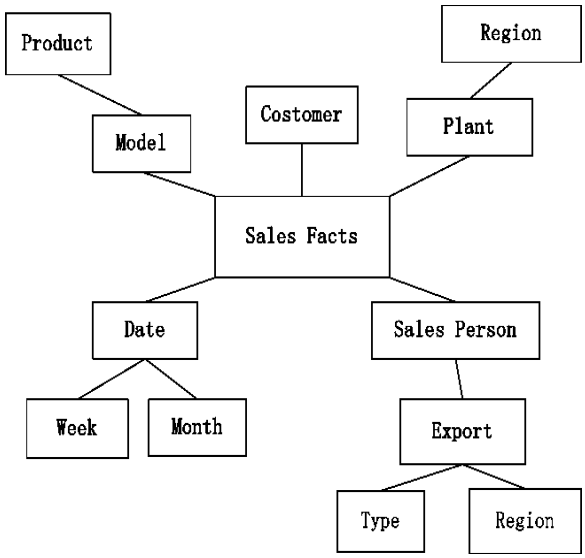

Fig. 3 The snowflake schema example

In this example(see Fig. 3), Sales table is fact table, but one or more dimension tables do not join directly to the fact table but must join through other dimension tables.

Both star and snowflake schemas are dimensional models; the difference is in their physical implementations. Snowflake schemas support ease of dimension maintenance because they are more normalized. Star schemas are easier for direct user access and often support simpler and more efficient queries.

3.2 Data Warehouse Implementation

Data Warehouse implementation must include mechanisms to migrate data into the data warehouse database. This process of data migration is generally referred to as the extraction, transformation, and loading (ETL) process (see Fig. 4).

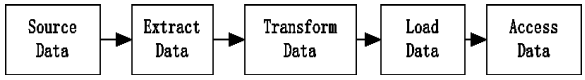

Fig. 4 The Data Warehouse Implementation Process

During the ETL process, data is extracted from an OLTP database, transformed to match the data warehouse schema, and loaded into the data warehouse database. Many data warehouses also incorporate data from non-OLTP systems, such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading.

3.2.1 Extract data from source systems

This phase is responsible for extracting data from the source system. During extraction, data may be removed from the source system or a copy made and the original data retained in the source system. It is common to move historical data that accumulates in an operational OLTP system to a data warehouse to maintain OLTP performance and efficiency.

3.2.2 Transform the data

This phase is responsible for data validation, data accuracy, data type conversion, and business rule application. Data Validation: Check that all rows in the fact table match rows in dimension tables to enforce data integrity. Data Accuracy: Ensure that fields contain appropriate values, such as only "off" or "on" in a status field. Data Type Conversion: Ensure that all values for a specified field are stored the same way in the data warehouse regardless of how they were stored in the source system. Business Rule Application: Ensure that the rules of the business are enforced on the data stored in the warehouse.

3.2.3 Load data to Data warehouse

This phase is responsible for loading transformed data into the data warehouse database. Data warehouses are usually updated periodically rather than continuously, and large numbers of records are often loaded to multiple tables in a single data load. The data warehouse is often taken offline during update operations.

3.2.4 User and application access the data

This phase rely on a suite of tools. The best way to choose this suite includes the definition of different types of access to the data and selecting the best tool for that kind of access. In general, access types include: Reporting, Ranking, Multivariable analysis, Time series analysis, Data visualization, graphing, charting, and pivorting, Statistical analysis, Artificial intelligence technique, OLAP tool, data mining tool, etc. Many tools and applications can be used to access warehouse data.

4. CONCLUSION

As transaction databases become more powerful, communication networks grow, and the flow of commerce expands, E-Business have accumulated a variety of on-line transaction processing (OLTP) systems and operational database. Data Warehouse is different from operational database. Data warehouses have been used to support business decision makers. A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process.

The key to build data warehousing is data design. The business users know what data they need and how they want to use it. Focus on the users, determine what data is needed, locate sources for the data, and organize the data in a dimensional model that represents the business needs. Dimensional modeling is the foundation of data warehouse design. A dimensional model may produce a star schema or a snowflake schema. the schema for a dimensional model contains a central fact table and multiple dimension tables.

Data Warehouse implementation include: extracting, transforming, and loading the data into the data warehouse, creating the OLAP and data mining analytical applications, developing end-user tools.

REFERENCES

[1] Berson, A., S.J. Smith, Data Warehousing, Data Mining, and OLAP, McGraw-Hill, 1997.
[2] Chaudburi, S., U.Dayal., "Database Technology for Decision Support Systems", Computer, Vol 34, No12, pp48-65, 2001.
[3] Kimball, R., The Data Warehouse Toolkit, Wiley, New York, 1996.
[4] Inmon, W.H., Building the Data Warehouse, Wiley, New York, 1992.
[5] Pedersen, T.B., C.S. Jensen, "Multidimensional Database Technology", Computer, Vol 34,No12, pp40-46,2001.
[6] Markus,M.L., "Pardigm Shifts-E-business and Business/Systems Integration", Communications of the Association for Information Systems, Vol 4, pp1-44, 2000.
[7] Watson, H.J., "Recent Developemnts in Data Warehousing", Communications of the Association for Information Systems, Vol 8, pp1-25, 2001.