Advance in Electronic and Electric Engineering.
ISSN 2231-1297, Volume 3, Number 6 (2013), pp. 669-674

http://www.ripublication.com/aeee.htm

Data Warehousing

Ritham Vashisht, Sukhdeep Kaur and Shobti Saini

M.Tech (CSE), Department of Computer Science and Engineering
Sri Sai College of Engineering and Technology
Pathankot, Punjab, India.

Abstract

DATA WAREHOUSING and Online Analytical Processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Data Warehouse provides an effective way for the analysis and statistic to the mass data and helps to do the decision making. In recent years data warehousing has become a prominent buzzword in the database industry, but attention from the database research community has been limited. Data warehouse is a booming industry with many interesting research problems. The database research community has concentrated only on few aspects. In this paper we motivate the concept of a data warehouse, we outline a general data warehousing architecture, and we propose a number of technical issues arising from the architecture that we believe are suitable topics for exploratory research. We describe back end tools for extracting, cleaning and loading data into the data warehouse, tools for metadata management and for managing the warehouse.

Keywords: Data warehouse design; multidimensional modeling; OLAP; ETL.

1. Introduction

Databases are developed on the idea that data is one of the critical materials of the information age. Information, which is created by data, becomes the bases for decision making.

Data warehousing encompasses architectures, algorithms, and tools for bringing together selected data from multiple databases or other information sources into a single repository, called a data ware-house, suitable for direct querying or analysis.

It is well known that data warehouses (DWs) are focused on decision support rather than on transaction support and that they are prevalently characterized by an OLAP work-load. Data Warehouse is a database used for reporting and analysis. It refers to the database that is maintained separately from an organization's operational databases. The data stored in the data warehouse is uploaded from the operational systems.

Data warehousing is a phenomenon that grew from the huge amount of electronic data stored in recent years and from the urgent need to use that data to accomplish goals that go beyond the routine tasks linked to daily processing. Many years ago, database designers realized that such an approach is hardly feasible, because it is very demanding in terms of time and resources, and it does not always achieve the desired results. Moreover, a mix of analytical queries with transactional routine queries inevitably slows down the system, and this does not meet the needs of users of either type of query.

Today's advanced data warehousing processes separate online analytical processing (OLAP) from online transactional processing (OLTP) by creating a new information repository that integrates basic data from various sources, properly arranges data formats, and then makes data available for analysis and evaluation aimed at planning and decision-making processes.

Traditionally, OLAP applications are based on multidimensional modeling that intuitively represents data under the metaphor of a cube whose cells store events that occurred in the business domain. Adopting the multidimensional model for DWs has a two-fold benefit. On the one hand, it is close to the way of thinking of data analyzers and, therefore, it helps users understand data; on the other hand, it supports performance improvement as its simple structure allows designers to predict users' intentions.

Multidimensional modeling requires specialized design techniques. Though a lot has been written about how a DW should be designed, there is no consensus on a design method yet. Most methods agree on the opportunity for distinguishing between a phase of conceptual design and one of logical design.

Several methods also support a phase of physical design that addresses all the issues specifically related to the suite of tools chosen for implementation such as indexing and allocation. In some cases, a phase of requirement analysis is separately considered.

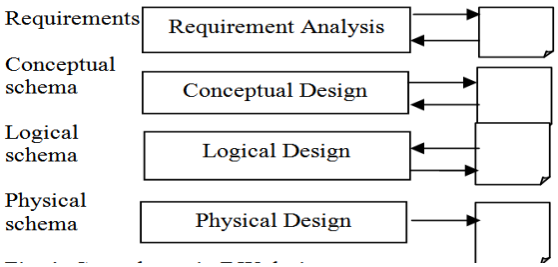

Fig. 1: Core phases in DW design

1.1 Data Warehouse

It is a collection of integrated, subject-oriented databases designed to support the DSS function, where each unit of data is non-volatile and relevant to some moment in time. DW is a repository of integrated information from operational (OLTP) and legacy system that provides data for analytical processing and decision making. The data in data warehouse is cleansed, temporal (historic), summarized and non-volatile.

1.2 Characteristics of data warehouse

Subject-Oriented: Information is presented according to specific subjects or areas of interest, not simply as computer files. Data is manipulated to provide information about a particular subject.

Integrated: All inconsistencies regarding naming convention and value representations are removed.

Non-Volatile: Stable information that doesn't change each time an operational process is executed. Information is consistent regardless of when the warehouse is accessed.

Time-Variant: Containing a history of the subject, as well as current information. Historical information is an important component of a data warehouse.

Accessible: The primary purpose of a data warehouse is to provide readily accessible information to end-users.

Process-Oriented: It is important to view data warehousing as a process for delivery of information. The maintenance of a data warehouse is ongoing and iterative in nature.

2. Multidimensional Modeling

Multidimensional data model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. Dimensions are perspectives or entities with respect to which an organization wants to keep records. Each dimension has a table associated with it, called a dimension table, which further describes the dimension. A multidimensional data model is typically organized around a a fact table. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Conceptual modeling: Conceptual modeling provides a high level of abstraction in describing the warehousing process and architecture in all its aspects, aimed at achieving independence of implementation issues. Conceptual modeling for DWs has been tackled from mainly two points of view so far:

- Multidimensional modeling
- Modeling of ETL

Logical modeling: Once the conceptual modeling phase is completed, the overall task of logical modeling is the transformation of conceptual schemata into logical schemata that can be optimized for and implemented on a chosen target system.

OLAP(on-line analytical processing)

It enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

OLAP operations:

- Drill down
- Drill up
- Slice and dice
- Pivot

OLAP Servers

An OLAP Server is a high capacity, multi user data manipulation engine specifically designed to support and operate on multi-dimensional data structure.

OLAP servers available are:

1. ROLAP server - These are the intermediate servers that stand in between a relational back-end server and client front-end tools. ROLAP severs include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services
2. MOLAP server - These servers support multidimensional views of data through array-based multidimensional storage engines.
3. HOLAP server - The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and faster computation of MOLAP.

OLAP Tools

A set of software products that attempts to facilitate multidimensional analysis. It can incorporate data acquisition, data access, data manipulation, or any combination thereof.

3. Data Warehouse Maintenance Issues

DW maintenance issues include data cleansing, transform, loading, subsequent loading (refreshing), data purging.

ECTL (Extract, cleansing, transform, and load): Refers to the methods involved in accessing and manipulating source data and loading it into the target database.

1. Extract: Some of the data elements in the operational database can be reasonably expected to be useful in the decision-making, but others are of less value for that purpose. For this reason, it is necessary to extract the relevant data from the operational database before bringing into the data warehouse.
2. Transform: The operational database developed can be based on any set of priorities, which keep changing with requirements. Therefore those who develop data warehouse based on these databases are typically faced with inconsistency among their data source.

3. Cleansing: Information quality is the key consideration in determining the value of the information. The developer of the data warehouse makes the data error-free before entering into the warehouse as much as possible. This process is known as data cleansing. It must deal with many types of possible errors. It includes missing data and incorrect data at one source, inconsistent data and conflicting data when two or more sources are involved.
4. Loading: It often implies physical movement of data from the computer storing the source database to that which will store the source data warehouse database assuming it is different.
5. Data refreshing and data purging: After the initial loading, updates at the source database should be propagated to the data warehouse. This propagation is called data refreshing.

4. Applications of Data Warehouse

- Operational and business intelligence applications
- Knowledge discovery
- Agriculture
- Biological data analysis
- Call record analysis
- Churn Prediction for Telecom subscribers, Credit Card users etc.
- Decision support
- Financial forecasting
- Insurance fraud analysis
- Logistics and Inventory management
- Trend analysis
- Health care providers
- Security agencies

5. Conclusion

In this paper, we have discussed about the construction and designing of data warehouses. The construction of data warehouses involves data cleaning and data integration. A data warehouse is a subject-oriented, integrated, time-variant and non volatile collection of data in support of management's decision making process. Data warehousing is very useful from the point of view of heterogeneous database integration. It provides an interesting alternative approach to the traditional approach of heterogeneous database integration. It employs an update-driven approach in which information from multiple, heterogeneous source is integrated in advance and stored in a warehouse for direct querying and analysis. Many people feel that with competition mounting in every industry, data warehousing is the latest must-have marketing weapon- a way to keep customers by learning more about their needs. Ad-hoc

techniques are required for dealing with the emerging applications of data warehousing and with advanced architectures for business intelligence. Besides, the need for real-time data processing raises original issues that were not addressed within traditional periodically-refreshed DWs. Thus, over-all, we believe that research on DW modeling and design is far from being dead, partly because more sophisticated techniques are needed for solving known problems, partly because of the new problems raised during the adaptation of DWs to the peculiar requirements of today's business. Data warehouse do not contain the current information. However, data warehouse brings high performance to the integrated heterogeneous database system. It can store and integrate historical information and support complex multidimensional queries. As a result, data warehousing has become very popular in industry.

References

[1] B.H¨ usemann, J.Lechtenb¨orger, and G.Vossen. Conceptual data warehouse design. InProc. DMDW, pages 3–9, 2000.
[2] M.Golfarelli, S.Rizzi, and I.Cella. Beyond data warehousing: What's next in business intelligence? In Proc. DOLAP, pages 1–6, 2004.
[3] Data Warehousing- Wikipedia
[4] Surajit Choudhary: Data Warehousing and OLAP technology.
[5] http://www.123helpme.com/data-wharehouse-paper-view.asp?id=164499