Manya Sethi / International Journal of Engineering Research and Applications (IJERA)
ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.955-960

DATA WAREHOUSING AND OLAP TECHNOLOGY

Manya Sethi
MCA Final Year
Amity University, Uttar Pradesh
Under Guidance of Ms. Shruti Nagpal

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. Many commercial products and services are now available and all of the principal database management system vendors now have offerings in these areas. The paper introduces the data warehouse and the online analysis process with an accent on their new requirements. I describe back end tools for extracting, cleaning and loading data into the data warehouse, tools for metadata management and for managing the warehouse.

1. Introduction
Data Warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. It provides architecture and tools for business executives to systematically organize, understand and use their data to make strategic decisions. 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 Warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for inventory management), financial services (for credit card analysis, risk analysis, and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis). This paper presents an overview of data warehousing technologies, focusing on the special requirements that data warehouses place on Database Management Systems (DBMSs).
The four keywords subject-oriented, integrated, time variant, and nonvolatile, distinguish data warehouse from other data repository systems, such as relational data base systems, transaction processing systems and file systems.

Subject-oriented: A data warehouse is organized around major subjects such as customer, supplier, product and sales. Rather than concentrating on the day to day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures and so on.

Time-variant: Data are stored to provide information from a historical perspective. Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Nonvolatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrence control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

The data warehouse supports on line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.

OLTP covers most of the day to day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration and accounting. An OLTP system is customer oriented and is used for transaction and query processing by clerks, clients and information technology professionals. It manages the current data that, typically, are too detailed to be easily used for decision making. An OLTP system

focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. The access patterns of OLTP systems consist mainly of short, atomic transactions. Such system requires concurrency control and recovery mechanisms.

Data warehouse systems, on the other hand, are targeted for decision support. It serves users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts. These systems manages large amount of historical data, provides facilities for summarization and aggregation, and store and manages information at different levels of granularity. These features make the data easier to use in informed decision making. An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design. This system often spans multiple versions of a database schema, due to the evolutionary process of an organization. These systems also deal with information that originates from different organization, integrating information from multiple data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multi dimensionally. For example, in a sales data warehouse, time of sales, sales district, salesperson, and product might be some of the dimensions of interest. Often these dimensions are hierarchical; time of sale may be organized as day-month-quarter-year hierarchy, product as a product-category-industry hierarchy.

In a multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exists to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis. Typical OLAP operations include roll-up, drill-down, slice and dice, pivot (rotate).

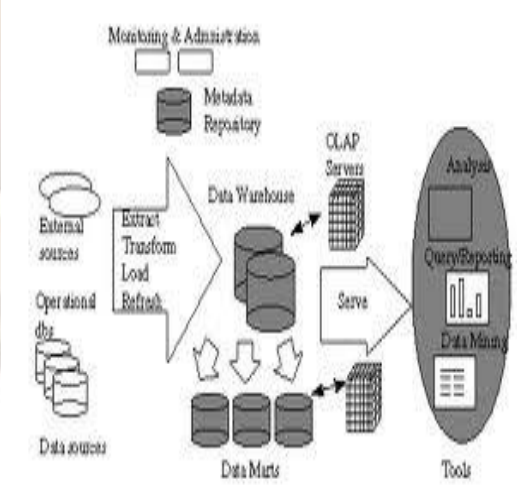
An operational database is designed and tuned from known tasks and workloads such as indexing and hashing using primary keys, searching for a particular record and optimizing 'canned' queries. On the other hand, data warehouse queries are often complex. They involve the computation of large amount of data at summarized levels and may require the use of special data organization, access and implementation methods based on multidimensional views.

Data warehouse might be implemented on Relational OLAP (ROLAP) servers. These are the intermediate servers that stand in between a relational back-end server and client front-end tools. They use a relational or extended-relational DBMS to store and manage warehouse data, and OLAP middleware to support missing pieces. They support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations. In contrast, Multidimensional OLAP (MOLAP) servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures.

There is more to building and maintaining the data warehouse than selecting an OLAP server, defining a schema and some complex queries for the warehouse. Different architectural alternatives exist. Many organizations want to implement an integrated enterprise warehouse that collects information about all subjects spanning the whole organization. However, building an enterprise warehouse is long and complex process. Some organizations are settling for data marts instead, which are departmental subsets focused on selected subjects. These data marts enable faster roll out since they do not require enterprise-wide consensus.

In the next point, we describe the architecture of the data warehouse and the process of a data warehouse design.

2. Architecture and the Process of Design



The left most tier (bottom tier) is a **warehouse database server** that is almost always a relational database system. Data from operational databases and external sources are extracted using application program interface known as gateways. A gateway is supported by the underlying

DBMS and allows client programs to generate SQL code to be executed by a server.

The middle tier is an OLAP server that is typically implemented using a ROLAP that maps operations on multidimensional data to standard relational operations or using a MOLAP that is a special purpose server that directly implements multidimensional data and operations.

The right most tier (top tier) is a client, which contains query and reporting tools, analysis tools, and data mining tools.

A data warehouse can be built using a top-down approach, a bottom-up approach, or a combination of both. The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments. In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

The design and construction of the data warehouse consists of the following steps: planning, requirements study, problem analysis, warehouse design, data integration and testing, and finally deployment of the data warehouse.

a. Choose a business process to model, for example, shipments, inventory, sales and the general ledger.
b. Choose the grain of the business process. The grain is the Fundamental, atomic level of data to be represented in the fact table for this process.
c. Choose the dimensions that will apply to each fact table record.
d. Choose the measures that will populate each fact table record.

Once a data warehouse is designed and constructed, the initial deployment of the warehouse includes initial installation, roll out planning, training and orientation.

3. Back-End Tools And Utilities

Data warehouse systems use back-end tools and utilities to populate and refresh their data.

Data Cleaning:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data

Since a data warehouse is used for decision making, it is important that the data in the warehouse must be correct. Some examples where data cleaning becomes necessary are: inconsistent field length, inconsistent descriptions, inconsistent value assignments, missing entries and violation of integrity constraints.

Load

After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints; sorting; summarization; aggregation; and other computations to build the derived tables stored in the warehouse. In addition, load utility also allows the system administrator to monitor status, to cancel, to suspend and resume a load, and to restart after failure with no loss of data integrity.

The load utilities for data warehouses have to deal with much larger data volumes than for operational databases

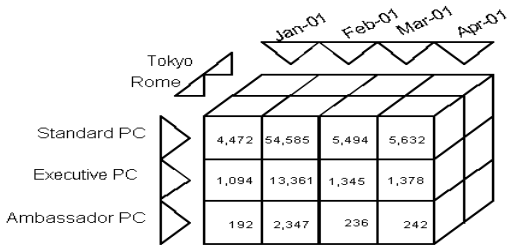
Refresh

Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: when to refresh and how to refresh. Usually, the warehouse is refreshed periodically. The refresh policy is set by the warehouse administrator, depending on user needs and traffic, and may be different for different sources. Refresh techniques also depends on the characteristics of the source and capabilities of the database servers. Replication servers can be used to refresh a warehouse when the sources change.

4. Multidimensional Data Model

Data warehouse and OLAP tools are based on a multidimensional data model. This 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 central theme. This theme is represented by 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.



Concept Hierarchies

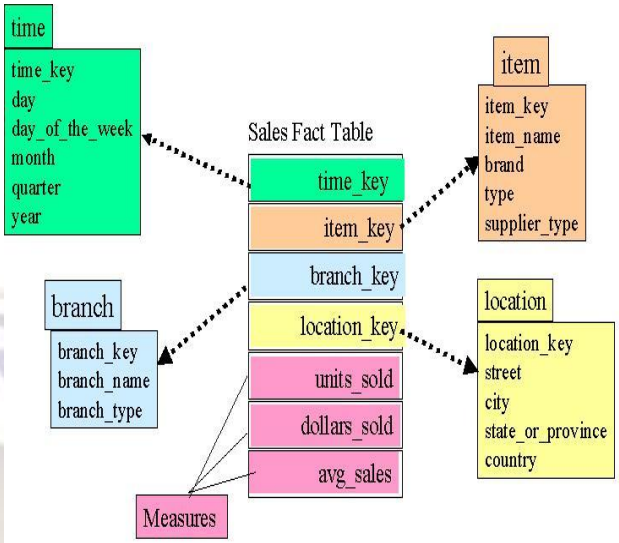
A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Many concept hierarchies are implicit within the database schema. A concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy.

Front-End Tools

The multidimensional data model grew out of the view of business data popularized by spreadsheet programs that are extensively used by business analysts. One of the popular operations that are supported by the multidimensional spreadsheet is pivoting. Pivot also called rotate, is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data. Other operations are roll-up, drill-down, slice and dice. The roll-up operation performs the aggregation on a data cube, either by climbing up the concept hierarchy for a dimension or by dimension reduction. Drill-down is the reverse of the roll-up. It navigates from less detailed data to more detailed data. The slice operation performs a selection on one dimension of the cube. The dice operation performs a selection on two or more dimensions.

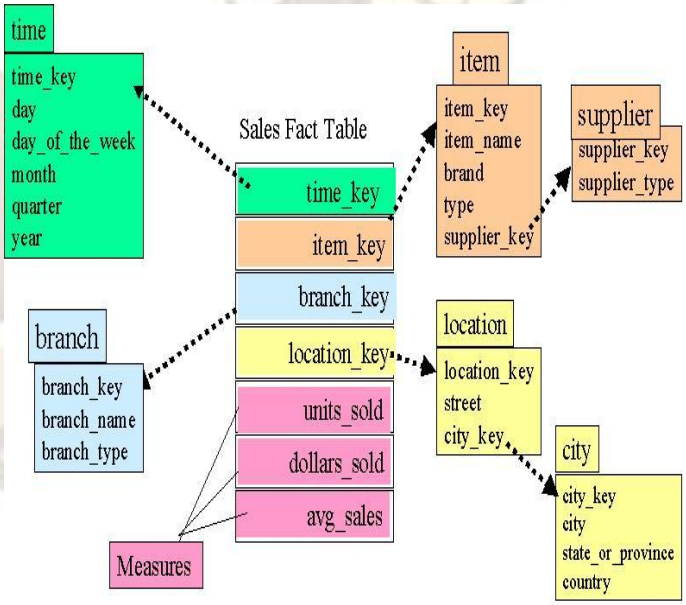
5. Database Design

Most data warehouse use a star schema to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer to each of the dimension that provides its multidimensional coordinates and stores the numeric measures for that coordinates. Each dimension table consists of columns that correspond to attributes of the dimension.



The Snowflake schema is the variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms the shape similar to a snowflake.

The major difference between the snowflake and star schema models is that the dimension table of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.



Indexing OLAP Data

To facilitate efficient data accessing, most data warehouse systems support index structures and materialized views.

The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes. The bit map index is an alternative representation of the record ID (RID) list. In the bitmap index for a given attribute, there is a distinct bit vector, Bv, for each value v in the domain of the attribute. If the domain of the given attribute consists of n values, then n bits are needed for each entry in the bitmap index. If the attribute has the value v for a given row in the data table, then the bit representing that value is set to 1 in the corresponding row of the bitmap index.

Bitmap indexing is advantageous as compared to hash and tree indices. It is especially useful for low cardinality domains because comparison, join, aggregation operations are then reduced to bit arithmetic, which substantially reduce the processing time. It leads to significant reduction in space since a string of characters can be represented by a single bit.

6. OLAP Servers

Logically, OLAP servers present business users with multidimensional data from data warehouses or data marts, without concerns regarding how or where the data stored. However, the physical architecture and implementation of OLAP servers must consider data storage issues. Implementation of a warehouse server for OLAP processing includes the following:

Relational OLAP servers - 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.

Multidimensional OLAP servers – These servers support multidimensional views of data through array-based multidimensional storage engines. Many MOLAP servers adopt a two-level storage representation to handle sparse and dense data sets: the dense sub cubes are identified and stored as array structures, while the sparse sub cubes employ compression technology for efficient storage utilization.

Hybrid OLAP servers – The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and faster computation of MOLAP. For example, a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store.

7. Metadata Repository

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Administrative metadata includes all of the information necessary for setting up and using a warehouse; description of the source databases; back-end and front-end tools. Business metadata includes business terms and definitions; ownership of the data. Operational metadata includes information that is created during the operation of the warehouse; monitoring information such as usage statistics, error reports, and audit trails.

Metadata repository is used to store and manage all the metadata associated with the warehouse. The repository enables the sharing of the metadata among tools and processing for designing, setting up, using, operating and administering a warehouse.

Metadata play a very different role than other data warehouse data, and are important for many reasons. For example, metadata are used as a directory to help the decision support system analyst to locate the contents of the data warehouse, as a guide to the mapping of the data when the data are transformed from the operational environment to the data warehouse environment. So, metadata should be stored and managed persistently.

8. Issues Of Research

Data cleaning is a problem that is a reminiscent of heterogeneous data integration, a problem that has been studied for many years. But here the emphasis is on data inconsistencies. Data cleaning is closely related to data mining, with the objective of suggesting possible inconsistencies.

The management of data warehouses also presents a new challenge. Detecting the queries,

Managing and scheduling resources are the problems that are important but that have not been well solved.

9. Conclusion

In this paper, I have discussed about the construction of data warehouses, designing of data warehouses. The construction of data warehouses involves data cleaning and data integration.

Data cleaning attempt to fill in the missing values, smooth out the noise, while identifying the outliers and remove the inconsistencies in the data.

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.

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 the process of constructing and using data warehouses. 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.
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

a. Jiawei Han and Micheline Kamber : Data Mining Concepts and Techniques

b. Surajit Choudhary: Data Warehousing and OLAP technology.

c. Umeshwar Dayal: An overview of data warehousing and technology.

d. http://www.dwinfocenter.org

e. http://www.carolla.com/wp-dw.htm

f. http://system-services.com/dwintro.asp

g. Data Warehousing- Wikipedia

h. http://research.microsoft.com/pubs/76058/sigrecord.pdf

i. http://www.cs.sunysb.edu/~cse634/presentations/DataWarehousing-part-1.pdf

j. http://lambda.uta.edu/cse6331/spring03/papers/dw1.pdf

k. http://www.123helpme.com/data-wharehouse-paper-view.asp?id=164499

l. http://www.1keydata.com/datawarehousing/dimensional.html

m. http://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm

n. http://www.peterindia.net/DataWarehousingView.html