DOI: https://doi.org/10.53555/nncse.v1i3.521 **Publication URL: http://nnpub.org/index.php/CSE/article/view/521**

DATA WAREHOUSING AND OLAP TECHNOLOGY

Vinayak Bhardwaj[1*], Ricy Jacob[2]

*[*1]Information Technology Engineering, Maharishi Dayanand University Gurgaon, Haryana, India*

[2]Information Technology Engineering, Maharishi Dayanand University Gurgaon, Haryana, India

***Corresponding Author: -**

Email ID- vinayak.14510@ggnindia.dronacharya.info

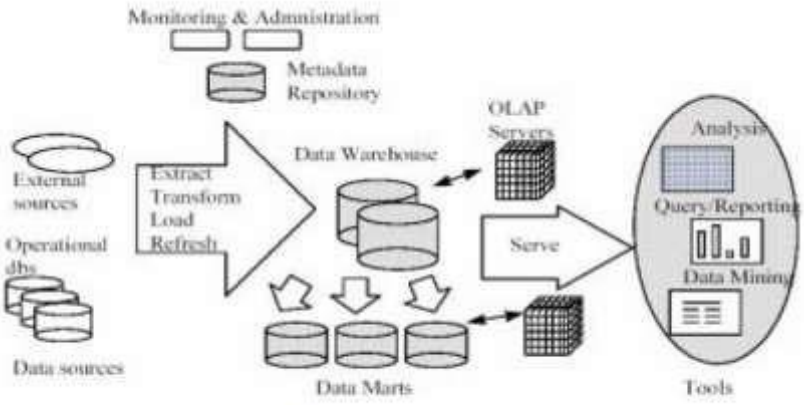
Abstract: -

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 static 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 offering in these areas. The paper introduces the data warehouse and online analysis process with an accent on their new requirements.

1. INTRODUCTION

Data Warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make a 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. Data Warehousing technologies have been successfully deployed in many industries: manufacturing (for shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcome analysis). This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs). Online Analytical Processing (OLAP) is a key feature supported by most warehousing systems. All of the data mining and OLAP are powerful tools to support decision making. An OLAP system is market oriented and is used for data analysis by knowledge workers, including mangers, executives and analysts. An OLAP system typically adopts either a star or a snowflake model and a subject-oriented database design. To facilitate complex analyses and visualization, the data in a warehouse is typically modelled multi dimensionally. In a multidimensional model, data are organized into multiple dimensions, and each dimensions contains multiple levels of abstraction defined by concept hierarchies. Typically OLAP operations include roll-up, drill-down, slice and dice, pivot (rotate). An OLAP system typically adopts either a star or a snowflake model and subject-oriented database design. Core of any OLAP system is an OLAP cube (also called a 'multidimensional cube' or a hypercube). It consists of numeric facts called measures which are categorized by dimensions. The measures are placed at the intersections of the hypercube, which is spanned by the dimensions as a Vector space. The usual interface to manipulate an OLAP cube is a matrix interface like pivot tables in a spreadsheet program, which performs projection operations along the dimensions, such as aggregation or averaging.

2. Architecture and process design



Typical Architecture of Data Warehouse

On the left 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 known as gateways. It is supported by the DBMS which is underlying and allows the client program to generate the SQL code which will be executed by the server.
The middle part is an OLAP server that is typically implemented using a ROLAP that maps operation on multidimensional data to standard relational operations or using MOLAP that is special purpose server that directly implements multidimensional data and operations.
The right part is a client which contains query and reporting tools, analysis tools, and data mining tools.

3. Back end tools and utilities

Data warehousing systems use a variety of data extraction and cleaning tools and load and refresh utilities for populating warehouses.

3.1 Cleaning

Data cleaning routines attempt to fill in the data 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 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.

3.2 Load

After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional processing may still be required: checking the integrity constraints; sorting summarization, aggregation and other computation to build the derived tables stored in the warehouse; building indices and other access paths; and partitioning to multiple target storage areas.

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 warehouse have to deal with much larger data volumes than for operational databases. There is only a small time window (usually at night) when the warehouses can be taken offline to refresh it. However, the load process now is harder to manage. The incremental load conflict with ongoing queries, so it is treated as a sequence of shorter transactions.

3.3 Refresh

Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. Refreshing techniques may also depend on the characteristics of the source and the capabilities of the database servers. Refresh techniques may also depend on the characteristics of the source and the capabilities of the databases servers. Extracting an entire source file or databases is usually too expensive but may be the only choice for legacy data sources. Refreshing can be done by using replication techniques. These two techniques are: data shipping and transaction shipping.

In **data shipping** (e.g., used in the Oracle Replication Server, Praxis OmniReplicator), a table in the warehouse is treated as a remote snapshot of a table in the source database. After_ row triggers are used to update a snapshot a log table whenever the source table changes; and an automatic refresh schedule (or a manual refresh procedure) is the set up to propagate the updated data to remote snapshot.

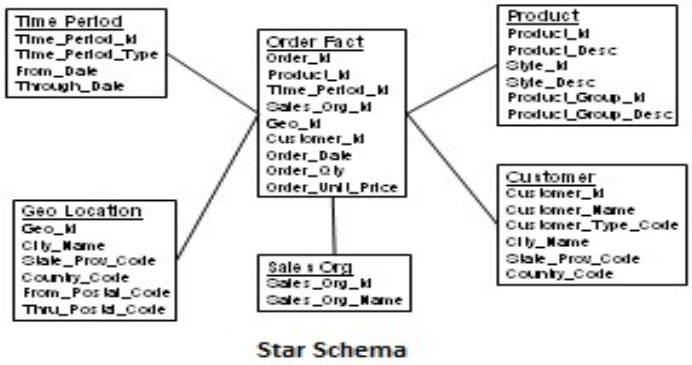
In **transaction shipping** (e.g., used in the Sybase Replication Server and Microsoft SQL Server), the regular transaction log is used, instead of triggers and a special snapshot log table.
Such replication servers have been used for refreshing data warehouses. However, the refresh cycles have to be properly chosen so that the volume of data does not overwhelm the incremental load utility.

3.4 Front- End Tools

The multidimensional data grew out of the view of business data popularized by spreadsheet programs that are extensively used by business analysts. The spreadsheet is till the most compelling front-end application for OLAP. The challenge is supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operations efficiently over large multi-gigabyte databases. 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 alternate presentation of the data. Other operations are roll-up, drill-down, slide and dice. The roll-up operation performs the aggregation on a data cube, either by climbing up the concept hierarchy for a dimensions 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 operations performs a selection on one dimensions of the cube. The dice operation performs a selection on two or more dimensions.

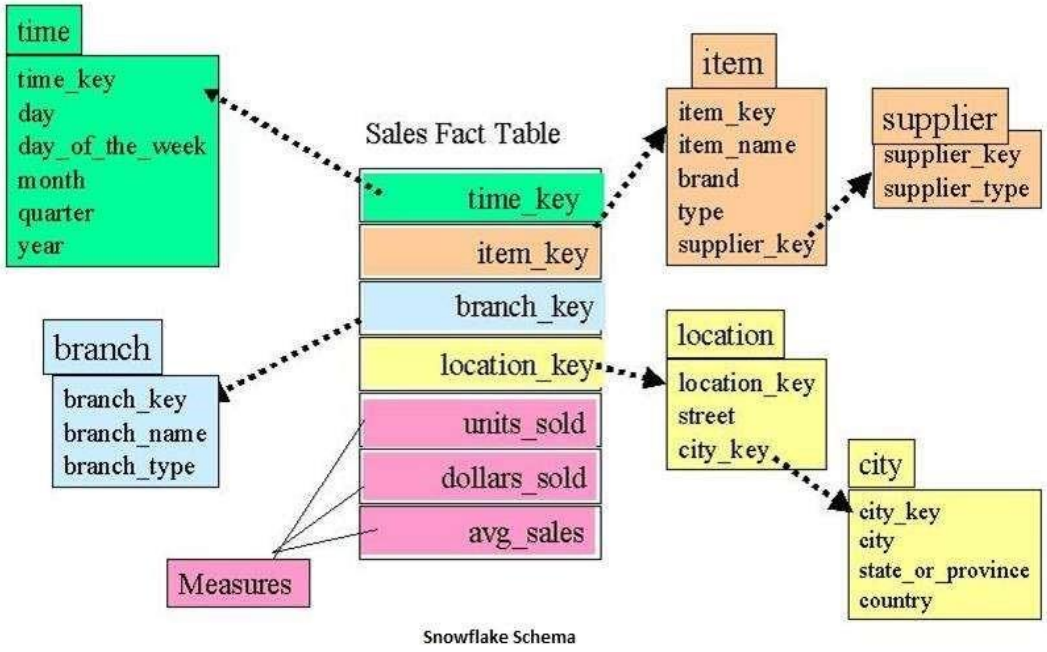
4. Database design

Most Database Warehouse use a star schema to represent the multidimensional data model. The database consist 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 dimensions that provides its multidimensional coordinates and stores the numeric measures for that coordinates. Mostly data warehouses use the star schema to represent the multidimensional data model. The database contains a single fact table and a single table for each of the dimensions. Each tuple in the fact table consist of a pointer to each of the dimensions. Below is the example of star schema.



Star Schema

Star schemas do not explicitly provide support for attribute hierarchies.
Snowflake schema is the variant of the star schema model, where some dimensions tables are normalized, thereby further splitting the data into additional tables. 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. In other words snowflake provide a refinement of star schemas where dimensional hierarchy is explicitly represented by normalizing the dimensional tables. Below is an example of Snowflake Schema



Snowflake Schema

5. Olap servers

OLAP servers makes the data readily available to the business users from data warehouses or data marts, irrespective of the inner working of how or where the data is stored. However the architecture of the OLAP servers must consider data storage problems and issues. Implementing of the warehouse server for OLAP can be done by following ways.

5.1 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. ROLAP servers include optimization for each DBMS back end, implementation of aggregation navigation logic, and additional tools and services.

5.2 Multidimensional OLAP (MOLAP) servers: These servers support multidimensional views of data through array-based multidimensional storage engines. They map multidimensional views directly to data cube array structures. The advantage of using a data cube is that it allows fast indexing to pre computed summarized data. Notice that with multidimensional data stores, the storage utilization may be low if the data set is sparse. In such cases, sparse matrix compression techniques should be explored. 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.

5.3 Hybrid OLAP (HOLAP) servers: The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the 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.

5.4 Specialized SQL servers: To meet the growing demand of OLAP processing in relational databases, some relational and data warehousing firms (e.g., Red Brick from Informix) implement specialized SQL servers that provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

6. Metadata repository

A **Metadata repository** is a database created to store metadata. Metadata itself is information about the structures that contain the actual data. 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. A well-designed metadata repository typically contains data far beyond simple definitions of the various data structures. Typical repositories store dozens to hundreds of separate pieces of information about each data structure. 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. There are many types of metadata that have to be managed. Administrative metadata includes all of the information necessary for setting up and using a warehouse: descriptions of the source databases, back-end and front-end tools; definitions of the warehouse schema, derived data, dimensions and hierarchies, predefined queries

and reports; data mart locations and contents; physical organization such as data partitions; data extraction, cleaning, and transformation rules; data refresh and purging policies; and user profiles, user authorization and access control policies. Business metadata includes business terms and definitions, ownership of the data, and charging policies. Operational metadata includes information that is collected during the operation of the warehouse: the lineage of migrated and transformed data; the currency of data in the warehouse (active, archived or purged); and monitoring information such as usage statistics, error reports, and audit trails.

7. Conclusions

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. 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].http://searchcrm.techtarget.com/tip/Types-of-OLAP-servers
[2].http://en.wikipedia.org/wiki/Metadata_repository
[3].http://en.wikipedia.org/wiki/Data_warehouse
[4].http://en.wikipedia.org/wiki/Online_analytical_processing
[5].An Overview Of Data Warehousing and OLAP Technology by Surajit Chaudhari and Umeshwar Dayal
[6].Apply On-Line Analytical Processing (OLAP)With Data Mining For Clinical Decision Support