Data Warehousing and OLAP Technologies for Decision-Making Process

Hiren H Darji
Asst. Prof in Anand Institute of Information Science,Anand

Abstract
Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Many commercial products and services are now available, and all of the principal database management system vendors now have offerings in these areas. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications. This paper provides an overview of data warehousing and OLAP technologies. Data warehousesprovide on-line analytical processing (OLAP) tools for theinteractive analysis of multidimensional data of varied granularities, which facilitates effective data mining. Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. s. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives and analysts. Data warehousing and OLAP have emerged as leading technologies that facilitate data storage, organization and then, significant retrieval. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications.

Keywords: Data Warehousing, OLAP, and Decision Making.

1.Introduction:

A data warehouse is a “subject-oriented, integrated,timevarying, non-volatile collection of Data that is used primarily in organizational decision making[2].”Typically, the data warehouse is maintained separately from the organization’s operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements. Data warehouses are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Query throughput and response times are more important than transaction throughput.

To facilitate complex analyses and visualization, the data in a warehouse is typically modeled multidimensionally. For example, in a sales data warehouse, time of sale, 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 a day-month-quarter-year hierarchy, product as a product category-industry hierarchy. Typical OLAP operations include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, slice_and_dice (selection and projection), and pivot (re-orienting the multidimensional view of data).

Data warehouses might be implemented on standard or extended relational DBMSs, called

Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and 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 are servers that directly store multidimensional data in special data structures (e.g., arrays) and implement the OLAP operations over these special data structures.

There is more to building and maintaining a data warehouse than selecting an OLAP server and 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 (e.g., customers, products, sales, assets, personnel) spanning the whole organization. However, building an enterprise warehouse is a long and complex process, requiring extensive business modeling, and may take many years to succeed. Some organizations are settling for data marts instead, which are departmental subsets focused on selected subjects (e.g., a marketing data mart may include customer, product, and sales information). These data marts enable faster roll out, since they do not require enterprise-wide consensus, but they may lead to complex integration problems in the long run, if a complete business model is not developed.

2. Data Warehousing

2.1 Definition of data warehousing

[2]A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process . So, data warehouse can be said to be a semantically consistent data store that serves as physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. The functional and performance requirements of OLAP are quite different from those of the on-line transaction processing applications traditionally supported by the operational databases.

A data warehouse is defined as a "subject-oriented, integrated, time variant, non-volatile collection of data that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. In data warehouses historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be much larger than operational databases. Most queries on data warehouses are ad hoc and are complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Due to the complexity query throughput and response times are more important than transaction throughput.

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order 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 outcomes analysis). This paper presents a roadmap of data warehousing technologies, focusing on the special requirements that data warehouses place on database management systems (DBMSs).

2.2 Architecture and End-to-End Process

Figure 1 shows a typical data warehousing architecture.

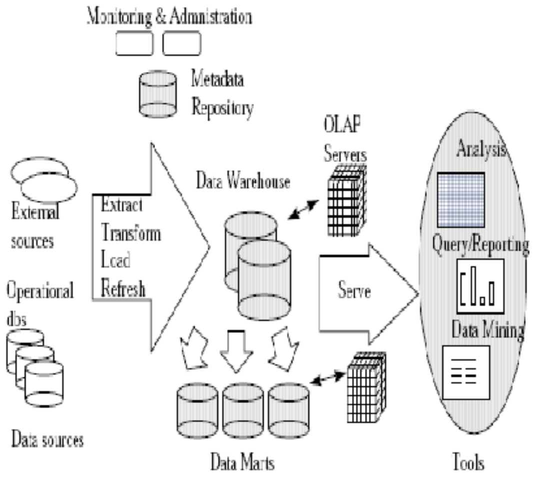

Figure 1: Data Warehousing Architecture

It includes tools for extracting data from multiple operational databases and external sources; for cleaning, transforming and integrating this data; for loading data into the data warehouse; and for periodically refreshing the warehouse to reflect updates at the sources and to purge data from the warehouse, perhaps onto slower archival storage.

In addition to the main warehouse, there may be several departmental data marts. Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front end tools: query tools, report writers, analysis tools, and data mining tools. Finally, there is a repository for storing and managing metadata, and tools for monitoring and administering the warehousing system.

Designing and rolling out a data warehouse is a complex process, consisting of the following activities [3]

- Define the architecture, do capacity planning, and select the storage servers, database and OLAP servers, and tools.
- Integrate the servers, storage, and client tools.
- Design the warehouse schema and views.
- Define the physical warehouse organization, data placement, partitioning, and access methods.
- Connect the sources using gateways, ODBC drivers, or other wrappers.
- Design and implement scripts for data extraction,
cleaning, transformation, load, and refresh.
- Populate the repository with the schema and view definitions, scripts, and other metadata.
- Design and implement end-user applications.
- Roll out the warehouse and applications.

2.3 DATA WAREHOUSING FUNDAMENTALS

A data warehouse (or smaller-scale data mart) is a specially prepared repository of data designed to support decision making. The data comes from operational systems and external sources. To create the data warehouse, data are extracted from source systems, cleaned (e.g., to detect and correct errors), transformed (e.g., put into subject groups or summarized), and loade d into a data store (i.e., placed into a data warehouse).

The data in a data warehouse have the following characteristics [5]:

- Subject oriented — The data are logically organized around major subjects of the organization, e.g., around customers, sales, or items produced.
- Integrated — All of the data about the subject are combined and can be analyzed together.
- Time variant — Historical data are maintained in detail form.
- Nonvolatile — The data are read only, not updated or changed by users.

A data warehouse draws data from operational systems, but is physically separate and serves a different purpose. Operational systems have their own databases and are used for transaction processing; a data warehouse has its own database and is used to support decision making. Once the warehouse is created, users (e.g., analysts, managers) access the data in the warehouse using tools that generate SQL (i.e., structured query language) queries or through applications such as a decision support system or an executive information system. “Data warehousing” is a broader term than “data warehouse” and is used to describe the creation, maintenance, use, and continuous refreshing of the data in the warehouse.

3. OLAP:

Data warehouse systems serve 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. These systems are called on-line analytical processing (OLAP) systems.

3.1 Need of data warehousing and OLAP

Data warehousing developed, despite the presence of operational databases due to following reasons:

• An operational database is designed and tuned from known tasks and workloads, such as indexing using primary keys, searching for particular records and optimizing ‘canned queries’. As data warehouse queries are often complex, they involve the computation of large groups of data at summarized levels and may require t he use of special data organization, access and implementation methods based on multidimensional views. Processing OLAP queries in operational databases would substantially degrade the performance of operational tasks.

• An operational database supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanisms, such as locking and logging are required to ensure the consistency and robustness of transactions. While and OLAP query often needs read-only access of data records for summarization and aggregation. Concurrency control and recovery mechanisms, if applied for such OLAP operations, may jeopardize the execution of concurrent transactions.

• Decision support requires historical data, whereas operational databases do not typically maintain historical data. So, the data in operational databases, though abundant, is always far from complete for decision-making.

• Decision support needs consolidation (such as aggregation and summarization) of data from heterogeneous sources; and operational databases contain only detailed raw data.

4. Data warehouse models [8]

There are 3 data warehouse models, according to architecture point of view

4.1 Enterprise warehouse

• Collects all of the information about subjects spanning the entire organization.
• Provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
• Typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to terabytes or beyond.
• May be implemented on traditional mainframes, UNIX super servers, or paralleled architecture platforms.

4.2 Data mart

• Contains a subset of corporate-wide data that is of value to a specific group of users, however, scope is confined to specific selected subjects.
• Are usually implemented on low-cost departmental servers that are UNIX or windows/NT –based.
• Are categorized as independent or dependent, depending on the source of data operational systems or external information providers, or from data generated locally within a particular department. But, dependent data marts are sourced directly from enterprise data warehouse.
• The data contained in data mart tend to be summarized.

4.3 Virtual warehouse

• Is a set of views over operational databases.
• Only some of the possible summary views may be materialized for efficient query processing.
• Is easy to build but requires excess capacity on operational database servers.

5. Decision making using a Data Warehouse

Data Warehouses (DW) integrate data from multiple heterogeneous information sources and transform them into a multidimensional representation for decision support applications. Apart from a complex architecture, involving data sources, the data staging area, operational data stores, the global data warehouse, the client data marts, etc., a data warehouse is also characterized by a complex lifecycle. In a permanent design phase, the designer has to produce and maintain a conceptual model and a usually voluminous logical schema, accompanied by a detailed physical design for efficiency reasons. The designer must also deal with data warehouse administrative processes, which are complex in structure, large in number and hard to code; deadlines must be met for the population of the data warehouse and contingency actions taken in the case of errors. Finally, the evolution phase involves a combination of design and administration tasks: as time passes, the business rules of an organization change, new data are requested by the end users, new sources of information become available, and the data warehouse architecture must evolve to efficiently support the decision-making process within the organization that owns the data warehouse. All the data warehouse components, processes and data should be tracked and administered via a metadata repository.

The combination of all the data warehouse viewpoints is depicted in Fig. 2.

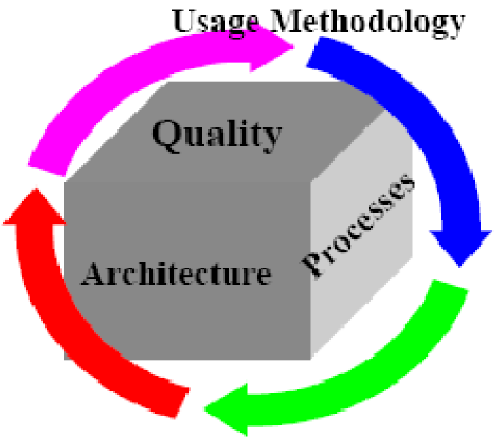

Fig. 2. The different viewpoints for the metadata repository of a data warehouse.

The framework describes a data warehouse in three perspectives: a conceptual [4] , a logical and a physical perspective. Each perspective is partitioned into the three traditional data warehouse levels: source, data warehouse and client level. On the meta model layer, the framework gives a notation for data warehouse architectures by specifying meta-classes for the usual data warehouse objects like data store, relation, view, etc. On the metadata layer, the meta model is instantiated with the concrete architecture of a data warehouse, involving its schema definition, indexes, table spaces, etc. The lowest layer in Fig. 3 represents the actual processes and data [4].

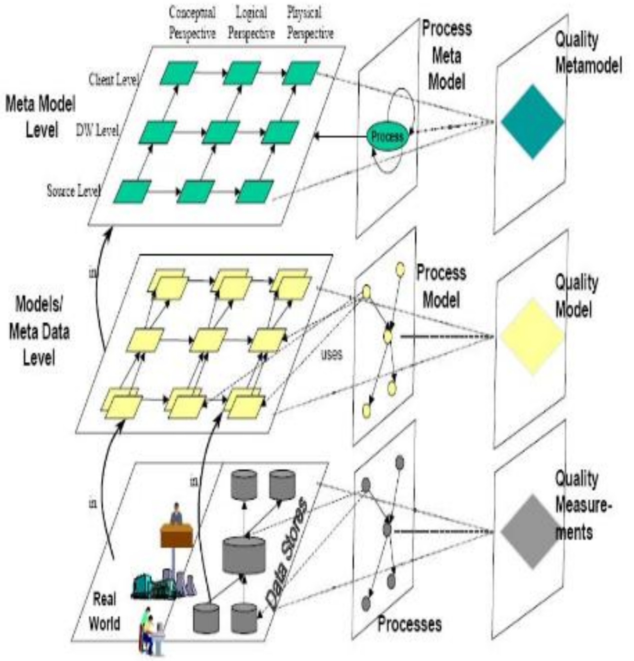

Figure 3: Frame work for Data Warehousing Architecture.

Another important issue shown in Fig. 4 is that we can observe a data flow in each of the three perspectives. In the logical perspective, the modeling is concerned with the functionality of an activity, describing what this particular activity is about in terms of consumption and production of information. In the physical perspective, the details of the execution of the process are the center of the modeling. The most intriguing part, though, is the conceptual perspective covering why a process exists. The answer can be either due to necessity reasons (in which case, the receiver of information depends on the process to deliver the data) and/or suitability reasons (in which case the information provider is capable of providing the requested information).

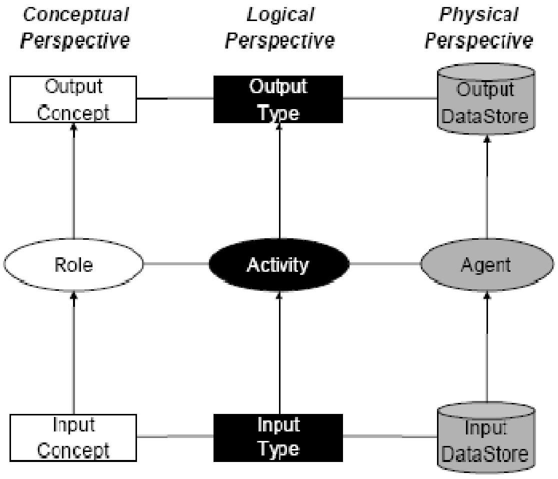

Figure 4:The reasoning behind the 3 perspectives of the process meta model.

6. **Conclusion:**

Data warehouse can be said to be a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. So, its architecture is said to be constructed by integrating data from multiple heterogeneous sources to support and /or adhoc queries, analytical reporting and decision-making. Data warehouses provide on-line analytical processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining. 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 systems serve users or knowledge workers in the role of data analysis and decisionmaking. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. OLAP applications are found in the area of financial modeling (budgeting, planning), sales forecasting, customer and product profitability, exception reporting, resource allocation, variance analysis, promotion planning, market share analysis. Moreover, OLAP enables managers to model problems that would be impossible using less flexible systems with lengthy and inconsistent response times. More control and timely access to strategic information facilitates effective decision-making. This provides leverage to library managers by providing the ability to model real life projections and a more efficient use of resources. OLAP enables the organization as a whole to respond more quickly to market demands. Market responsiveness, in turn, often yields improved revenue and profitability. And there is no need to emphasize that present libraries have to provide market-oriented services.

References:

[1] Devlin, B. & Murphy, P. (1988) An Architecture for a Business and Information System, IBM Systems Journal, 27 (1), 60-80.

[2] Inmon, W.H. (1996) Building the Data Warehouse, Second Edition, New York: John Wiley & Sons.

[3] Kimball, R. (1996) The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, New York: John Wiley & Sons.

[4] M. Jarke, M.A.Jeusfeld, C. Quix, P. Vassiliadis: Architecture and quality in data warehouses: An extended repository approach. Information Systems, 24(3): 229-253 (1999). A previous version appeared in Proc. 10th Conf. of Advanced Information Systems Engineering (CAiSE ’98), Pisa, Italy (1998).

[5] Adelman, S. & Moss, L. (2000) Data Warehouse Project Management, Boston: Addison-Wesley.

[6] Kachur, R. (2000) Data Warehouse Management Handbook, Paramus: Prentice Hall.

[7] Inmon, W.H., Building the Data Warehouse. John Wiley, 1992.

[8] Han, Jiawei and Kamber, Micheline. Data Mining: Concepts and techniques. Academic Press, a. 2001.