International Journal of Research (IJR) Vol-1, Issue-10 November 2014 **ISSN 2348-6848**

Review on Data Warehouse, Data Mining and OLAP Technology: As Prerequisite aspect of business decision-making activity

RuchiYadav&Pramod Kumar
Dept. of Information &technology, Dronacharya College of Engineering
Farruhknagar, Gurgaon, India
Email:ruchiyadav477@gmail.com

Abstract

This paper describes the technology of data warehouse in decision making and tools for support of these technology. Data warehousing and on-line analytical processing (OLAP) are prerequisite aspects of decision support, which has increasingly become a focus of the database industry. The construction of data warehouses involves data cleaning, data integration, data transformation and as important pre-processing step for data mining. 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. The OLTP is customer-oriented which is used for transactions and query processing by clerks, clients. An OLAP is market-oriented which is used for data analysis by knowledge employees, including managers, executives and analysts. Data warehousing and OLAP have evolved as one of primary technologies that facilitate data storage, organization and, denoting retrieval. Different requirements on database technology compared to traditional on-line transaction processing applications.

KEYWORDS: *Data warehousing; OLTP; OLAP; Data Mining; Decision-making ;Decision Support.*

1. Introduction

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make a strategic decisions. Data warehouse systems are the valuable tools in today's fast-evolving world .Data warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated data for analysis. "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support management's decision making process "(William Hinson, 1992).

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. It serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions.

The data can be stored in many different types of databases. One data base architecture that has recently emerged is the "data warehouse", a repository of multiple heterogeneous data sources, organized under a unified schema at a

single site in order to facilitate management decision-making. Data warehouse technology includes data cleansing, data integration and online Analytical processing. OLAP stands for analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information from different angles.

Since the introduction of the data warehouse concept in the late 1980ies (e.g. Devlin/Murphy 1988), data warehouse systems are now an established component of information systems landscape in most companies. Due to high failure rates of data warehouse projects, several procedure models for building data warehouse systems were published considering their special requirements (e.g. Inman 1996, Kimball 1996, Gardner 1998, Simon 1998). However, these methodologies were mainly focused on technical issues, like architectural concepts and data modeling. But according to studies about critical success factors of data warehouse projects organizational, political and cultural factors are at least as important as technical ones (Frolic/Lindsey 2003, Hwang et al. 2002, Finnegan/Summon 1999). In addition, most development methodologies are lacking concepts to ensure long-term evolution and establishment of data warehouse systems (O'Donnell et al. 2002), which are both primarily organizational challenges (Meyer/Winter 2001). Up to now only few authors have adopted a mainly organizational-driven view on data warehouse systems. Kocher (2000) describes activities and organizational structures for data warehouse management

2. Data Warehousing

2.1 Definition of data warehousing

According to W.H.Inmon, a leading architect in the construction of data warehouse systems, a data warehouse is a subject-oriented, integrated, time-variant and non-volatile 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 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. 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.

Data can now be stored in many different types of databases. One type of database architecture that has recently emerged is data warehouse, which is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making (Chaudhuri&Dayal 1997; Chawatte, Garcia-Molina, Hammer, Ireland, Papakonstantinou, Ullman & Wisdom 1994; Han &Kamber 2001). Data warehouse technology includes data cleaning, data integrating, and on-line analytical processing (OLAP) that is, analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information from different angles.

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, and 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 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 loaded into a data store (i.e., placed into a data warehouse).

The data in a data warehouse have the following characteristics:

- ***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.

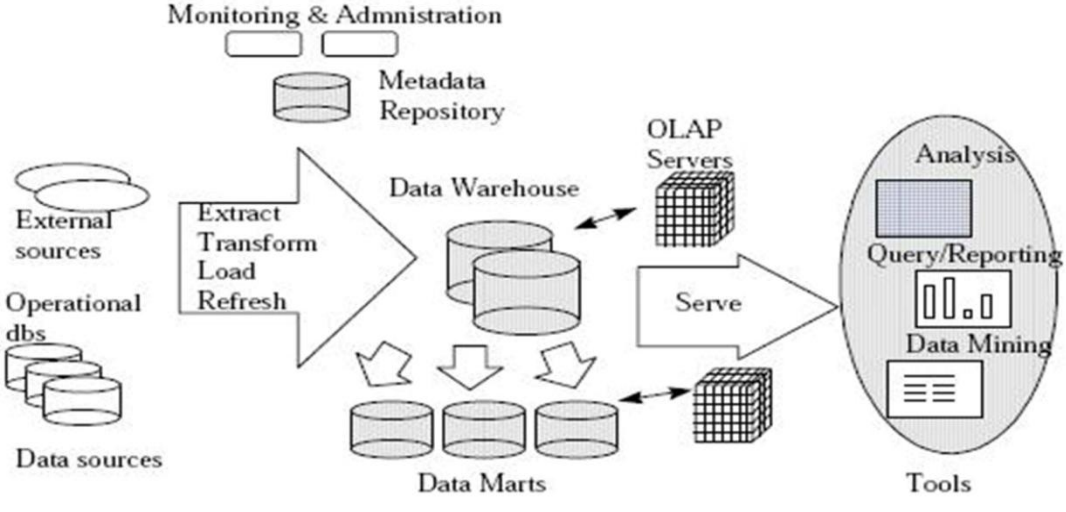
Adelman/Moss (2000) explains how to design and manage data warehouse systems focusing on project management aspects. In addition, they give an overview of organizational roles involved in a typical data warehouse project. Meyer (2000) and Meyer/Winter (2001) present organizational requirements for data warehousing and the concept of data ownership. A two-dimensional organizational structure for large financial service companies combining infrastructural competencies and content competencies is derived. Auth (2003) develops a process-oriented organizational concept for metadata management providing detailed activity chains and organizational roles. As shown above the organizational domain of data warehouse systems still lacks attention of data warehouse researchers compared to technical aspects. Therefore this paper aims at providing deeper insights in the current organizational

situation of data warehouse departments in practice. The organizational domain of companies can be divided in a structural, human resource, political, and symbolic dimension and each dimension requires different design instruments (Bolman/Deal 2003, Mueller-Stewens 2003). The structural dimension focuses on goals, formal roles and relationships. Structures are created to achieve the company's goals considering technological and environmental factors. Rules, policies, processes, and hierarchies are the design elements of the structural dimension. Drawing from psychology, the human resource dimension takes care about the needs, feelings, prejudices, and limitations of all individuals. The political dimension sees organizations as arenas. Different interest groups cause conflicts while competing for power and resources and the organizational life is characterized by bargaining, negations and compromises. The symbolic dimension abandons the assumptions of rational behavior and views organizations as some kind of theatres. Rituals, ceremonies and stories are more important than rules or managerial authority (Bolman/Deal 2003). Many studies on critical success or failure factors of data warehouse projects have been published (e.g. Wixom/Watson 2001, Little/Gibson 1999, Frolick/Lindsey 2003). Most studies comprise technical as well as organizational factors. In the organizational domain issues like management support, sponsorship, user participation, and organizational politics are often mentioned. But these studies do not explicitly distinguish the different organizational dimensions and they only provide a general overview of relevant organizational factors. However, more detailed insights are needed to design and apply appropriate instruments and measures to implement these success factors.

The OLAP Council(http://www.olapcouncil.org) is a good source of information on standardization efforts across the industry, and a paper by Codd, et al. defines twelve rules for OLAP products. Finally, a good source of references on data warehousing and OLAP is the Data Warehousing Information.

2.3 Architecture and End-to-End Process

Figure 1 shows a typical 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.

3.OLTP and OLAP

The job of earlier on-line operational systems was to perform transaction and query processing. So, they are also termed as on-line transaction processing systems (OLTP). 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 Major distinguishing features between OLTP and OLAP

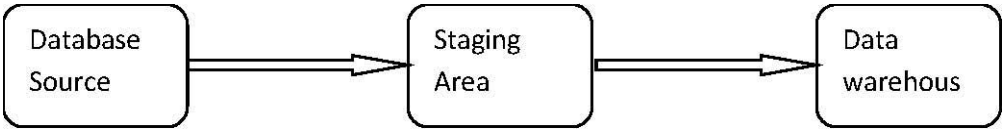
i) **Users and system orientation:** OLTP is customer-oriented and is used for transaction and query processing by clerks, clients and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives and analysts.

ii) **Data contents:** OLTP system manages current data in too detailed format. While an OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation. Moreover, information is stored and managed at different levels of granularity, it makes the data easier to use in informed decision-making. iii) Database design: An OLTP system generally adopts an entity –relationship data model and an application-oriented database design. An OLAP system adopts either a star or snowflake model and a subject oriented database design. iv) **View:** OLTP system focuses mainly on the current data without referring to historical data or data in different organizations. In contrast, OLAP system spans multiple versions of a database schema, due to the evolutionary process of an organization. Because of their huge volume, OLAP data are shared on multiple storage media. v) Access patterns: Access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency, control and recovery mechanisms. But, accesses to OLAP systems are mostly read-only operations, although many could be complex queries.

3.2 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 the 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 Flow

The steps for building a data warehouse or repository are well understood. The data flows from one or more source databases into an intermediate staging area, and finally into the data warehouse or repository (see Figure 2). At each stage there are data quality tools available to massage and transform the data, thus enhancing the usability of the data once it resides in the data warehouse.

Figure 2-Data Flow.

5. Data Mining

Data Mining is the extraction or "Mining" of knowledge from a large amount of data or data warehouse. To do this extraction data mining combines artificial intelligence , statistical analysis and database management Systems to attempt to pull knowledge from stored data . Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools . The spreadsheet is still the most compiling front-end application for Online Analytical Processing (OLAP). The challenges in supporting a query environment for OLAP can be crudely summarized as that of supporting spreadsheet operation effectively over large multi gigabytes databases. To distinguish information extraction through data mining from that of a traditional database querying, the following main observation can be mede. In a database application the queries issues are well defined to the level of what we want and the output is prcise and is a subset of the database. Beside the data used not the operational data that represent th e today transactions. 'For instance during the process of building a data warehouse the operational data is summarized over different characteristics , such as borrowings during three month period. Queries can be of type of "identify all borrowers who have similar interest "or "items a member would frequently borrow along with movies" which is not precise as list of books borrowed by a member. Users can use data mining techniques on the data warehouse to extract different kinds of information which would eventually assist the decision making process of an organization (figure 3). For example, if certain books are rarely used by member of particular library, while same books are frequently used at other libraries then it's appropriate to transfer these books to respective libraries to ensure its effective use. Such knowledge could not only be discovered through sharing experiences of librarians or by capturing the knowledge through database and integrating them as done when building data warehouses. Decision support tools assist users in discovering knowledge.

6. Decision making using a Data Warehouse

A Decision Support System (DSS) is any tool used to improve the process of decision making in complex systems. A DSS can range from a system that answer simple queries and allow a subsequent decision to be made, to a system that employ artificial intelligence and provides detailed querying across a spectrum of related datasets. Amongst the most important application areas of DSS are those complicated systems that directly "answer" questions, in particular high-level "what-if "scenario

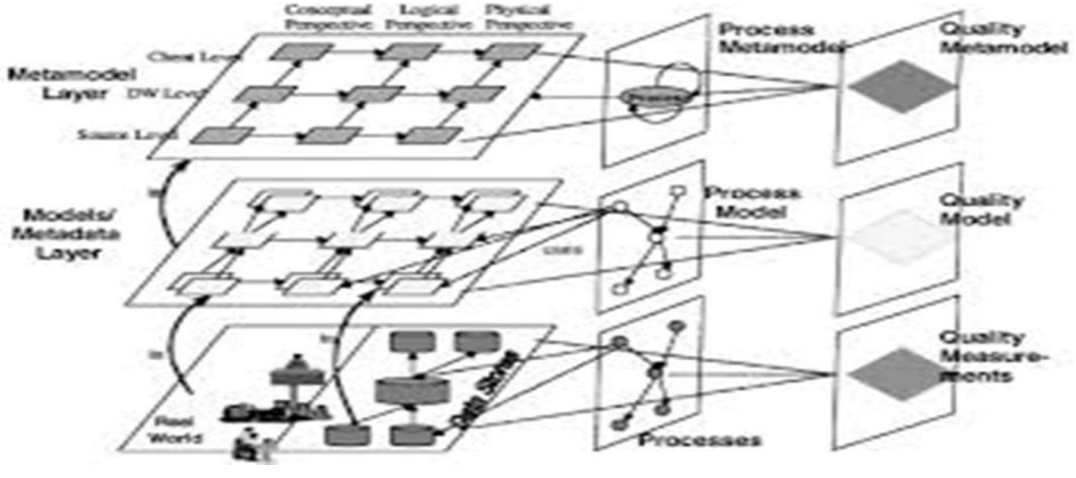
modeling . Over the last decade there was a transition to decision supporting using data warehouses (Inmon 2002).The data warehouse environment is more controlled and therefore more reliable for decision support than previous methods. The data warehouse environment supports the entire decision support requirements by providing high-quality information, made available by accurate and effective cleaning routines and using consistent and valid data transformation rules and document pre-summarization of data values. It contains one single source of accurate, reliable information that can be used for analysis.

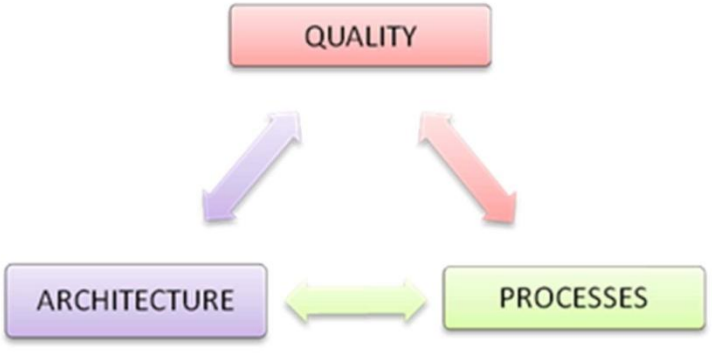
Data Warehouses (DW) integrate data from multiple heterogeneous information sources and transform them into a multidimensional representation for decision support applications. Apart from a complex architecture, involving data sources, the data staging area, operational data stores, the global data warehouse, the client marts, etc., a data warehouse is also characterized by a complex lifecycle. In a permanent design phase, the designer has to produce and maintain in 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.

The linkage of the architecture model to the quality parameters (in the form of a quality model) and its implementation in the metadata repository. Concept Base has been formally described in [M.A. Jeusfeld, C. Quix, M. Jarke, 1998]. [P. Vassiliadis, M. Bouzeghoub, C. Quix, 1999 & 2000] Presents a methodology for the exploitation of the information found in the metadata repository and the quality-oriented evolution of a data warehouse based on the architecture and quality model. In this paper, we complement these results with meta models and support tools for the dynamic part of the data warehouse environment: the operational data warehouse processes. The combination of all the data warehouse viewpoints is depicted in Figure 3.

USAGE METHODOLOGY

Fig. 3. The different viewpoints for the metadata repository of a data warehouse. In [M. Jarke, M.A.Jeusfeld, C. Quix, P. Vassiliadis, 1998 &1999] a basic meta model for data warehouse architecture and quality has been presented as in Fig. 3. The framework describes a data warehouse in three perspectives: a conceptual, 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. 4 represents the actual processes and data.





7. Steps for designing data warehouse [Chaudahari, Surajit and Dayal, Umeshwar,1997]

Designing a data warehouse is a complex process, which consists of following activities:

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, and 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. **8. Data warehouse models [**Han, Jiawei and Kamber, Micheline**, 2001]**

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

1. **Enterprise warehouse**
2. **Data mart**
3. **Virtual warehouse**

8.1 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.

8.2 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.

8.3 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.

9. Why OLAP in data warehouse [Witten, Ian H. and Frank, Eibe**, 2000]**

Simply told, a data warehouse stores tactical information that answers "who?" and "what?" questions about past events. While OLAP systems have the ability to answer "who?" and "what?" questions, it is their ability to answer "what if?" and "why?" that sets them apart from Data warehouses.

OLAP enables decision making about future actions. In contrast to Data warehouse, this is usually based on relational technology. OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis. OLAP and data warehouses are complementary. A data warehouse manages and stores data. OLAP transforms data warehouse "data" into "strategic information". It ranges from basic navigation and browsing (often known as 'slice and dice') to calculations, to more serious analysis such as time series and complex modeling.

10. 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 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 on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. OLTP is customer-oriented and is used for transaction and query processing by clerks, clients and information technology professionals. The job of earlier online operational systems was to perform transaction and query processing. 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. 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, and 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.

11. References

[1] Han, J &Kamber, M 2001, Data Mining Concepts and Techniques, Morgan Kaufmann.

[2] Inmon, WH 2002, Building the Data Warehouse, 3rd Edition, Wiley.