An Overview of Data Warehousing, Data mining, OLAP and OLTP Technologies

Ashish Gahlot , Manoj Yadav
Dronacharya college of engineering
Farrukhnagar, Gurgaon ,Haryana

Abstract- **Data warehousing , Data Mining, OLAP, OLTP technologies are essential elements of decision support, which has increasingly become a focus of the database industry. 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.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 ,Data Mining, OLAP, OLTP technologies with an emphasis on their new requirements. We describe back end tools for extracting, cleaning and loading data into a data warehouse; multidimensional data models typical of OLAP; front end client tools for querying and data analysis; server extensions for efficient query processing; and tools for metadata management and for managing the warehouse.**

I. 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. The past three years have seen explosive growth, both in the number of products and services offered, and in the adoption of these technologies by industry. According to the META Group, the data warehousing market, including hardware, database software, and tools, is projected to grow from $2 billion in 1995 to $8 billion in 1998. Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory), 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).

A data warehouse is a "subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making. (Inmon, W.H., 1992) 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 of which are quite different from those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases.
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. Inmon 1996, Kimball 1996, Gardner 1998, and 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 (Frolick/Lindsey 2003, Hwang et al. 2002, Finnegan/Sammon 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. Kachur (2000) describes activities and organizational structures for data warehouse management.

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.

II. 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. Data warehouse technology includes data cleaning, data integrating, and online 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.

2.3 Architecture and End-to-End Process

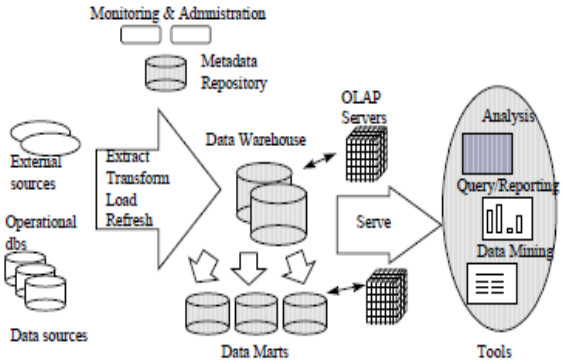

Figure 1. Data Warehousing Architecture

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 519 managing metadata, and tools for monitoring and administering the warehousing system.

The warehouse may be distributed for load balancing, scalability, and higher availability. In such a distributed architecture, the metadata repository is usually replicated with each fragment of the warehouse, and the entire warehouse is administered centrally. An alternative architecture, implemented for expediency when it may be too expensive to
construct a single logically integrated enterprise warehouse, is a federation of warehouses or data marts, each with its own repository and decentralized administration. Designing and rolling out a data warehouse is a complex process, consisting of the following activities5.

□ □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.

III. 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 form 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 made. In a database application the queries issued are well defined to the level of what we want and the output is precise and is a subset of operational data. In data mining there is no standard query language and the queries are poorly defined. Thus the output is not precise (fuzzy) and do not represent a subset of the database. Beside the data used not the operational data that represents the to day transactions. For instance during the process of building a data warehouse the operational data are summarized over different characteristics, such as borrowings during 3 month's period. Queries can be of the type of "identify all borrowers who have similar interest" or "items a member would frequently borrow along with movies", which is not a precise as the list of books borrowed by a member. The nature of the database and the query result in extracting nonsubset of data. In supermarkets such relationships have already been identified using data mining. Thus related items such as "bread and milk' or "beer and potato chips" would be kept together. Mobile companies decide on peak hours, rates and special packages based similar market research. 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 2). For example, if certain books are a particular library, while the same books are frequently books to respective libraries to ensure its effective use. Such building data warehouses. Decision support tools assist users in discovering knowledge.

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 allows a subsequent decision to be made, to a system that employ artificial intelligence and provides 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.

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.

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 M.A.Jeusfeld, C. Quix, P. Vassiliadis,1998 & 1999], we presented a metadata modeling approach which enables the capturing of the static parts of the architecture of a data Warehouse. The linkage of the architecture model to quality Parameters (in the form of a quality model) and its implementation in the metadata repository. Concept Base has been formally described in 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 Fig. 2

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

In a basic Meta model for data warehouse architecture and quality has been presented as in Fig. 2. 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 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

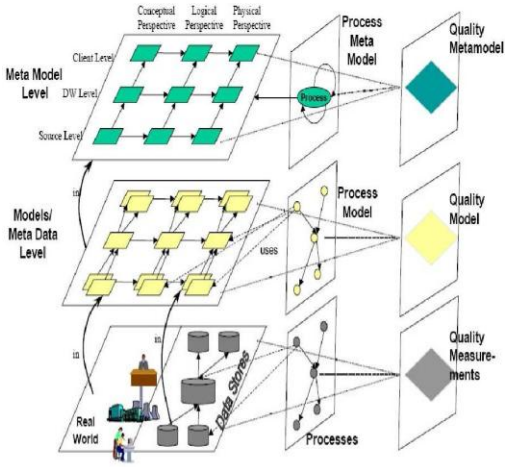

Figure 3. Frame work for Data Warehousing Architecture.

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).

IV. 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.

4.1 Major distinguishing features between OLTP and OLAP

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.

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. 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. 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. 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.

4.2 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.

V. 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 4). 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.

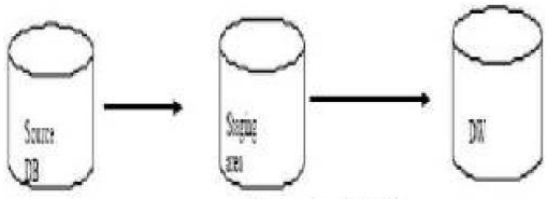

Fig4. data flow

VI. DATA WAREHOUSE MODELS

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

6.1 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.

6.2 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.

6.3 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.

VII. WHY OLAP IN DATA WAREHOUSE

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.

VIII. 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 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 on-line 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.

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.

http://pwp.starnetinc.com/larryg/articles.html

[5] Kimball, R. *The Data Warehouse Toolkit*. John Wiley, 1996.

[6]Barclay, T., R. Barnes, J. Gray, P. Sundaresan, "Loading Databases using Dataflow Parallelism." *SIGMOD Record*, Vol. 23, No. 4, Dec.1994.

[7] Hwang, H.-G., Kua, C.-Y., Yenb, D. C., & Chenga, C.-C. (2002) Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan, Decision Support Systems, In Press, Corrected Proof.

[8] Finnegan, P. & Sammon, D. (1999) Foundations of an Organisational Prerequisites Model for Data Warehousing, Proceedings of the 7th European Conference on Information Systems (ECIS 1999), Copenhagen, June.