Data Warehousing and OLAP Technology for Knowledge Discovery

Aparajita Suman

Abstract

Since time immemorial, libraries have been generating services using the knowledge stored in various repositories to the users. However, now the techniques of data storage, organization, and retrieval have changed with drastic changes in media and access tools. So, the new generation of libraries has to adopt emerging tools and techniques.In this context, data warehousing and OLAP have emerged as leading technologies that facilitate data storage, organization and then, significant retrieval. This article aims at exploring the features of data warehousing and OLAP and their application in modern libraries for knowledge discovery.

Keywords : Data Warehousing, OLAP, Information Retrieval, Data Mining

0. Introduction

Building data warehouses and performing OLAP on the top of them has become critical solution for KM and business intelligence. But now, their benefits have started expanding in the areas of research activities and knowledge discovery. The idea of knowledge discovery i.e. data mining has attracted great attention in information industry in recent years. The need has been felt due to wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. So, various processes have evolved which help one in finding small sets of precious nuggets from a great deal of raw data, stored in various 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.

1. How data warehousing, OLAP and knowledge discovery are related

The interrelationship of data warehousing, OLAP and knowledge discovery gets pretty clear from data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP). However, data mining goes far beyond the narrow scope of summarization –style analytical processing of data warehouse systems by incorporating more advanced techniques for data understanding as shown in figure : 1, on following page.

 © INFLIBNET Centre, Ahmedabad

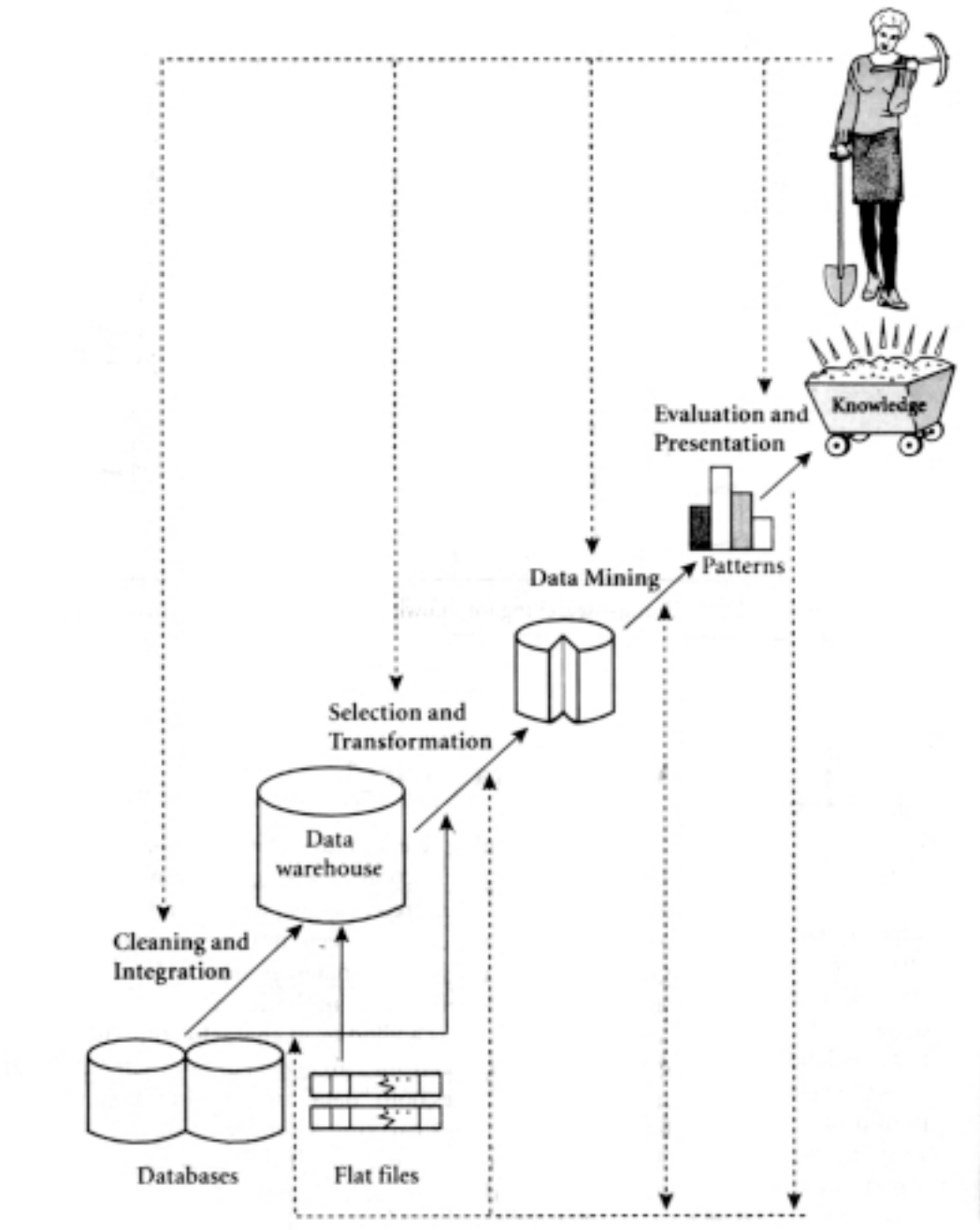

Figure 1: Data mining as a step in the process of knowledge discovery [2,pg 6]

2. Definition of data warehousing

According to W.H.Inmon [2], 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 multi-dimensional 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.

3. 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. OLAP : A component of data warehouses

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.

5. Major distinguishing features between OLTP and OLAP [1, 2]

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.

6. Data warehouse architecture

Data warehouses often adopt a 3-tier architecture.

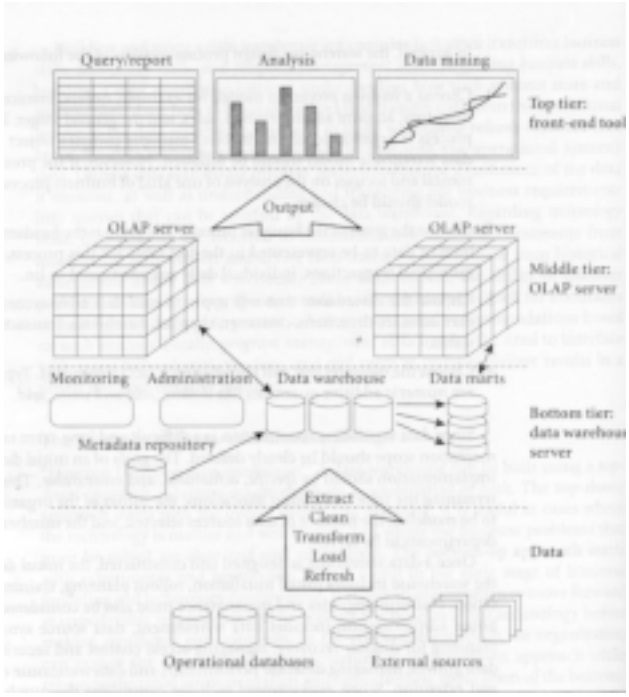

Figure 2: A 3-tier data warehousing architecture [2, pg 66]

- Here, the bottom tier is a warehouse database server that is always a relational database system. Data is extracted from this tier using application program interfaces known as gateways. The gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server.

  E.g.: ODBC (open database connection), OLE-DB (open linking and embedding for Databases) is some examples of gateways.
- Middle tier is an OLAP server which is implemented using either (a) relational OLAP (ROLAP) model, that is an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (b) a multidimensional OLAP (MOLAP) model, that is, a special –purpose server that directly implements multidimensional data and operations.
- Client forms the top tier, it contains query and reporting tools, analysis tools, and /or data mining tools(e.g. trend analysis , prediction etc.).

7. Steps for designing data warehouse [1]

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 [2]

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

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 [3]

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, which 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. Features of OLAP [4]

The key indicator of a successful OLAP application is its ability to provide information as needed i.e. its ability to provide “just in time” information for effective decision-making.

All the OLAP applications, found in divergent functional areas, have following key features:

10.1 Multidimensional views of data

- Is inherently representation of an actual business model.
- Provides more than the ability to “slice and dice”, it provides the foundation for analytical processing through flexible access to information.
- Managers must be able to analyze data across any dimensions at any level of aggregation, with equal functionality and ease.

10.2 Calculation-intensive capabilities

- Real test of an OLAP application is its ability to perform complex calculations; they must be able to do more than simple aggregation.
- OLAP software must provide a rich tool kit of powerful yet succinct computational methods, because key performance indicators often require involved algebraic equations.
- Analytical processing systems are judged on their ability to create information from data.

10.3 Time Intelligence

- Is and integral component of almost any analytical application. Time is a unique dimension because it is sequential in character. True OLAP systems should understand the sequential nature of time.
- Time hierarchy is not always used in the same manner as other hierarchies. Concepts such as year–to-date and period over period comparisons must be easily defined in an OLAP system.

In addition, they must understand the concept of balances over time.

11. OLAP applications in Library and information profession

As discussed earlier, OLAP stands for dynamic synthesis, analysis and consolidation of large volumes of multidimensional data. Needless to say here that information processing has always been the domain of library professionals. So, in the electronic era where most of the information is stored in databases, the use of this information-processing tool is very appropriate. Its application can be found in the following areas:

- OLAP generates a series of hypothetical patterns and relationships and uses queries against the database to verify them. It is complementary in the early stages of the knowledge discovery process because it can help to explore the data by focusing attention on important variables, identifying exceptions, or finding interaction.
- When we are switching over to digital and electronic libraries, then this tool can be used for conducting online user survey to ultimately make the services more useful [5]. It helps automate the feedback step in the evaluation modules.
- Its application may be found in the administrative function of library to come up with a projection of services provided, users' response to them, budget spent etc.
- As OLAP helps in discovering the pattern in the data, so, it can be useful for knowledge organization also.

  Finally, the better we understand the data, the more effective the discovery and retrieval will be.

12. Conclusion

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.

13. References

1. Chaudahari, Surajit and Dayal, Umeshwar. An overview of Data warehousing and OLAP technology. SIGMOD Record, 26(1), March 1997 Full text (pdf) available, as on 02/10/03: http://portal.acm.org/results.cfm?coll=portal&dl=ACM&CFID=13701999&CFTOKEN=8934106
2. Han, Jiawei and Kamber, Micheline. Data Mining: Concepts and techniques. Academic Press, 2001.
3. Witten, Ian H. and Frank, Eibe. Data mining: Practical machine learning tools and techniques with Java implementation. Academic Press, 2000.
4. OLAP council white paper, accessed on 15/10/03. http://www.olapcouncil.org/research/whtpaply.htm
5. OLAP application, accessed on 10/10/03. http://www.olapreport.com/Applications.htm
6. http://www.intelligententerprise.com/030320/605warehous1_1.shtml
7. http://www.cdacindia.com/html/dwh/dwhegov.asp

About Author

Ms. Aparajita Suman is ADIS STUDENT at DRTC/ISI, Mysore Road, Bangalore - 560 059, India and holds M.Sc. in Life Sciences and BLISc.
E-mail : aprajita_drtc@rediffmail.com, s_aparajitha@yahoo.co.in